M-estimation and Median of Means applied to statistical learning

Timothée Mathieu

université
PARIS-SACLAY

M-estimation and Median of Means applied to statistical learning

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°574 : mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées
Unité de recherche: Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France.
Référent: : Faculté des sciences d'Orsay

Thèse présentée et soutenue à Orsay, le 13 janvier 2021, par

Timothée MATHIEU

Composition du jury:

Gilles Blanchard Professeur, Université Paris Saclay	Président
Christophe Biernacki Professeur, Université Lille 1	Rapporteur & Examinateur
David Donoho Professeur, Université de Stanford	Rapporteur & Examinateur
Olivier Catoni Directeur de recherche, ENSAE	Examinateur
Po-Ling Loh Professeur associé, Université du Wisconsin - Madison	Examinatrice
Elvezio Ronchetti Professeur, Université de Genève	Examinateur
Matthieu Lerasle Chargé de recherche, HDR, ENSAE	Directeur de thèse
Guillaume Lecué Professeur, ENSAE	Co-directeur de thèse

Thèse de doctorat

NNT: 2021UPASM002

Contents

Remerciements

Je remercie les rapporteurs ainsi que les membres du jury pour le temps qu'ils ont accordé à la lecture de ce manuscrit.

Merci à mes directeurs de thèse, Guillaume et Matthieu, pour m'avoir laissé libre de poursuivre mes idées même si elles pouvaient paraître farfelues aux premiers abords. Merci à Stas et Zoltan pour une collaboration fructueuse, vous avez été très ouverts aux idées que je proposais et c'est ce qui a rendu l'expérience enrichissante. Merci aussi aux développeurs de scikit-learn-extra chkoar et rth, pour m'avoir aidé à améliorer mon code de non-informaticien.

Je souhaite aussi remercier Elvezio Ronchetti qui m'a permis de passer une année très enrichissante à Genève où j'ai beaucoup appris et grâce à qui j'ai trouvé les idées derrière au moins deux de mes articles. Merci aux doctorants de Genève qui m'ont très bien accueillis malgré mon cursus de maths théorique.

Une thèse se fait en grande partie seul mais ce n'est pas pour cela qu'on ne peut pas avoir de l'aide de temps à autre et même pour un solitaire comme moi la compagnie a été bienvenue. Merci à mon co-bureau Alexandre pour les cookies et les chocolats mais aussi plus généralement pour la compagnie, les discussions et les jeux du midi. Merci à Nicolas d'avoir bien voulu relire mon intro de thèse un peu en dernière minute. Merci à mes autres collègues de Cachan Antoine, Vianney, julie, Schmuel que j'ai retrouvé quelques fois pendant ce doctorat soit pour une partie de coinche soit pour parler de choses et d'autres. Merci aussi à Perrine et Etienne, même si on ne s'est pas vu à l'ENS on a rattrapé un peu ça pendant cette dernière année de thèse.

Merci à Marie-Anne pour avoir été très ouverte sur mes propositions parfois maladroite d'amélioration des cours de L3 MINT et M1 IA, à Christine pour tous ses conseils et à Nathalie mais aussi à tous mes co-chargés de TD Raphaël, Raphaëlle, Zacharie, Hedi et Gauthier. Mes missions d'enseignements ont été une bonne expérience et m'ont donné envie de peut être continuer l'enseignement ce qui n'était pas gagné d'avance et cela je vous le dois sûrement.

Parmi tous mes professeurs de Maths, j'aimerais remercier Mme Cambrai et Mr Delbecque qui ont surement été les premiers à me donner l'envie d'aller plus loin en maths. Je remercie aussi mes professeurs de l'ENS Cachan et entre autres Vianney Perchet, Nicolas Vayatis et Alain Trouvé dont j'ai beaucoup apprécié les cours.

Merci à ma très nombreuse famille, à mes parents qui m'ont soutenu dans mes études depuis le début sans me pousser à partir dans une voie ou dans une autre mais en me laissant parcourir mon chemin. À mes grands-parents pour leur présence bienveillantes. À Alice sans qui la prépa aurait été plus dure. Merci à Loic, Magalie, Marieke, Camille, Benjamin et tous leurs plus un. Merci aussi à tous mes neveux et nièces sans qui j'aurais peut-être eu plus de calme mais peut-être que le bruit en vaut la peine. Merci aussi au reste de ma grande famille.

Chapter 1

Introduction en Français

Les statisticiens ont observé que les bases de données réelles contenaient des données anormales (outliers en anglais, nous utiliserons le terme anglais) et ces données anormales peuvent avoir un effet négatif sur l'analyse statistique de la base de donnée. Par données anormales, nous entendons ici des données qui sont difficiles à modéliser ou des données que l'on devrait ignorer dans notre analyse, et à cause de la présence de ce type de données, nous devons utiliser des méthodes dites robustes. Ce problème est déjà assez vieux, Newcomb en 1882 [Gut01] et Laplace avant lui étaient déjà conscients de l'effet de outliers sur l'analyse de données. Fisher avait aussi souligné le problème en inférence statistique [FR22]. Il y a en ce moment un regain d'intérêt pour ce genre de technique, en grande partie à cause des besoins pour l'apprentissage statistique robuste.

Cette thèse a donné lieu à l'écriture de cinq articles : l'article [LLM20] qui a été publié dans le journal *Machine Learning*, l'article [LSML19] qui a été publié comme proceeding de la conférence *ICML* en 2019, l'article [MM19] qui a été accepté à la publication dans *Information and Inference: A Journal of the IMA* et les deux articles [Mat20b, Mat20a] que j'ai écris seul et qui seront bientôt proposés à la publication. Dans ces articles, nous étudions les statistiques robustes et en particulier leurs applications à l'apprentissage statistique. Cette thèse a aussi donné lieu à une contribution de code pour l'apprentissage statistique robuste dans la librairie python scikit-learn-extra.

Dans un premier temps, je présente quelques résultats de base sur les statistiques robustes et la formulation mathématique de la robustesse. Ensuite, je présente l'état de l'art en termes d'estimation robuste de la moyenne. En effet, l'estimation robuste de la moyenne est un problème central dans cette thèse et ces résultats sont le point de départ des travaux présentés ici.

Dans un second temps, je présente succinctement nos contributions aux statistiques robustes. Premièrement avec des bornes de déviations non-asymptotiques pour l'estimation de la moyenne en utilisant des résultats de [Mat20b] et [Mat20a], dans un contexte de distribution à queue lourde et/ou de base de données corrompue. Ensuite, je présente les contributions faites à l'apprentissage statistique robuste, notamment à travers les trois articles [LLM20, LSML19, MM19] avec des applications en régression, en classification et en méthode à noyaux.

Pour une lecture peut-être plus ludique de ce travail, les lecteurs sont encouragés à exécuter le notebook que j'ai préparé pour illustrer cette thèse: https://colab.research.google.com/drive/1yyGCgmif1EXBNLBgMODaZvPLyHuJW8zf?usp=sharing.

1.1 La robustesse selon Huber et les estimateurs sous-Gaussiens

1.1.1 Distributions à queues lourdes et corruption de base de données

De façon informelle, un estimateur est dit robuste si un petit changement dans les hypothèses que l'on suppose sur la base de données ne change pas beaucoup le résultat de l'estimation. Pour préciser cette définition, il faut définir ce que l'on entend par "un petit changement des hypothèses" et par "ne change pas beaucoup".

Considérons un exemple. Supposons que l'on a accès à n observations $X_1, \dots, X_n$ dans $\mathbb{R}$. Il se trouve que cet échantillon est corrompu : $X_1, \dots, X_{n-1}$ sont tirés i.i.d, suivant une distribution gaussienne $\mathcal{N}(\mu, \sigma^2)$ et X_n est égal à une constante $M \gg \mu$ (X_n est alors appelé outlier). La moyenne empirique devient alors $\frac{1}{n}\sum_{i=1}^{n-1} X_i + \frac{M}{n}$, ce qui peut être arbitrairement loin de la moyenne μ des points non anormaux si M est très grand comparé à $\sum_{i=1}^{n-1} X_i$. La moyenne empirique n'est donc pas robuste, notre but dans ce problème serait alors de trouver un estimateur qui ne présente pas ce problème comme cela est illustré dans la Figure 2.1.

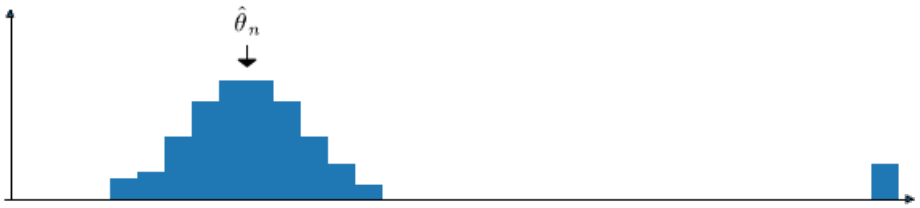


Figure 1.1: Histogramme d'une base de donnée corrompue par des outliers. Dans cette figure, $\widehat{\theta}_n$ désigne un estimateur robuste de la moyenne.

Nous considérons trois différents cadres de travail : les distributions à queue lourde, les bases de données corrompues par des outliers et ce qu'on appelle un voisinage de corruption (voir ci-après pour la définition). Ces trois cadres de travail sont représentés dans la Figure 1.2.

Distributions à queue lourde. Les résultats non asymptotiques classiques ne sont bien souvent valides que dans le cas de données i.i.d $X_1, \dots, X_n$ ayant pour loi commune une distribution Gaussienne ou une distribution à queue légère. Typiquement, pour obtenir un résultat de concentration en apprentissage statistique, on aura besoin d'une hypothèse sous-gaussienne sur les données. Cependant, les données réelles sont bien souvent plus proches d'une loi à queue lourde. Une sorte de déviation du cas idéal est donc de considérer le cas des distributions à queue lourde où X n'a typiquement qu' un second moment fini. C'est dans ce contexte que les deux articles [Cat12, DLLO16] ont été écrits.

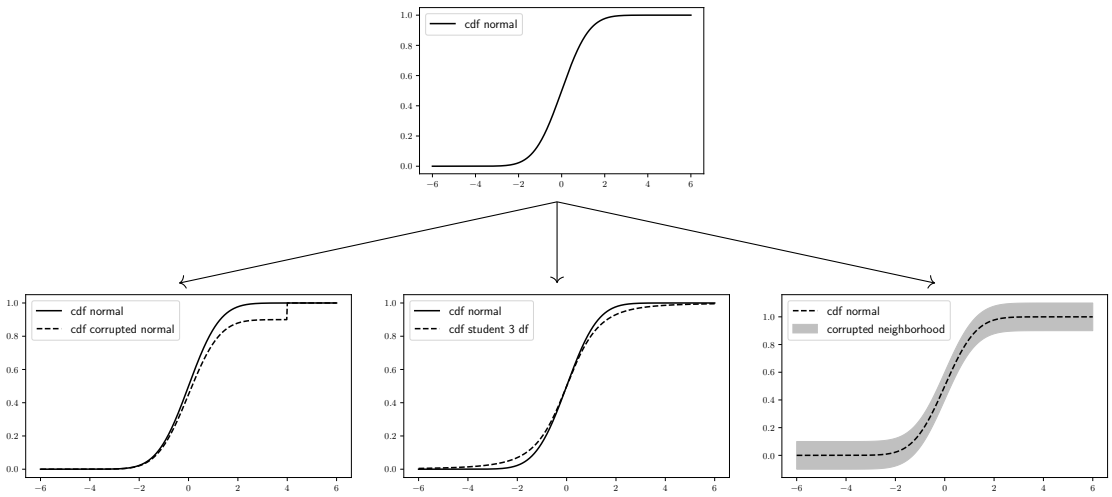


Figure 1.2: Tracé du graphe de la fonction de répartition dans différents cadres de corruption. Le graphique du haut représente le cas non corrompu (Gaussienne standard), celui d'en bas à gauche un modèle de contamination de Huber, celui d'en bas au milieu un cas de distribution à queue lourde et celui en bas à droite un voisinage de corruption (distance de Kolmogorov).

Données corrompues, cadre de travail $\mathcal{I} \cup \mathcal{O}$. Soient $X_1, \dots, X_n$ des variables aléatoires. Soit $\mathcal{I}, \mathcal{O}$ une partition de $\{1, \dots, n\}$ en deux ensembles disjoints, l'ensemble $\mathcal{I}$ des inliers et l'ensemble $\mathcal{O}$ des outliers (avec $|\mathcal{O}|$ petit comparé à $|\mathcal{I}|$). Certaines hypothèses seront faites sur $(X_i)_{i\in\mathcal{I}}$, typiquement des hypothèses de moment fini, mais par contre nous ne supposerons rien sur $(X_i)_{i\in\mathcal{O}}$. Les ensembles $\mathcal{I}$ et $\mathcal{O}$ sont inconnus du statisticien. Ceci est une généralisation du contexte des distributions à queue lourde dans lequel les données ne sont plus supposées i.i.d. Ce cadre de travail a été utilisé dans [LL20, LSML19].

Base de données corrompue par des outliers, contamination de Huber. Le second scénario de corruption considéré est ce qui est appelé dans la littérature la contamination de Huber. Dans ce modèle, $X_1, \dots, X_n$ sont i.i.d avec pour commune loi un mélange de distribution $(1-\varepsilon)P + \varepsilon H$ où ε est petit. P est supposé connu, alors que nous ne faisons aucune hypothèse sur H. H joue le rôle de la distribution des outliers. Ce cadre de travail est très proche du cadre $\mathcal{I} \cup \mathcal{O}$ présenté précédemment, mais contrairement au cadre $\mathcal{I} \cup \mathcal{O}$, la contamination de Huber contient des données i.i.d ce qui peut donc être vu comme un cadre non-adversarial. La contamination de Huber est définie dans [Hub64, HR09, ZJS19] et a entre autres été utilisée dans [CGR$^+$18].

Voisinage de corruption. Comme cela a été dit précédemment, nous pouvons définir une déviation du cas idéal en disant que $X_1, \dots, X_n$ sont i.i.d avec une distribution commune qui est voisine de la distribution P, par exemple cette notion de voisinage peut être définie par le biais de la contamination de Huber. Plus généralement, soit d une distance entre distributions de probabilité (i.e. distance de Kolmogorov, distance de variation totale, distance de Wasserstein $\dots$) et suppose que $X_1, \dots, X_n$ est tiré selon une distribution Q telle que $d(P, Q) \leq \varepsilon$. Cela généralise la corruption de Huber, en effet si $Q = (1-\varepsilon)P + \varepsilon H$ est une version corrompue de P,

on a que la variation totale entre P et Q est inférieure à ε. Utiliser les voisinages d'une distance est plus générale que la contamination de Huber et le type d'outliers considéré va dépendre en grande partie de la distance entre distributions choisies pour définir les voisinages. Les voisinages de corruptions ont été en premier utilisés dans [HR09, Ham71].

Dans tous les cas de corruptions présentés, le but est d'utiliser des méthodes qui donnent des résultats similaires quand on les utilise sur la base de donnée idéale (i.i.d suivant une loi gaussienne par exemple) que quand on utilise la même méthode sur une base de donnée corrompue. Ceci constitue une définition informelle de ce que l'on considère comme méthode robuste dans cette thèse et à chaque fois que l'on énonce un résultat, on prend soin de préciser la situation de corruption dans laquelle on se trouve.

On peut identifier deux principaux champs de recherches en statistiques robustes théoriques : un champ de recherche asymptotique et un non-asymptotique. Au début des statistiques robustes, il y a eu beaucoup de résultats asymptotiques et les théorèmes principaux étaient en terme de normalité asymptotique avec une variance asymptotique optimale [Hub64]. Récemment, beaucoup de résultats non-asymptotiques sont apparus dans ce champ de recherche notamment en utilisant des techniques telles que les inégalités de concentration et les processus empiriques avec pour motivation première d'utiliser ces méthodes en apprentissage statistique. Dans cette thèse, nous donnons à la fois des résultats asymptotiques et non asymptotiques en contribution à la théorie des statistiques robustes.

1.1.2 Robustesse et inégalités de concentration

La théorie de la robustesse a connu un renouveau récemment dû aux besoins de l'apprentissage statistique. Dans les applications en apprentissages statistiques telles que la régression ou la classification linéaire, il est assez facile de trouver des exemples où l'on constate la non-robustesse des estimateurs (regarder par exemple Figure 2.3). Soient $\mathcal{X}$ et $\mathcal{Y}$ deux ensembles, typiquement

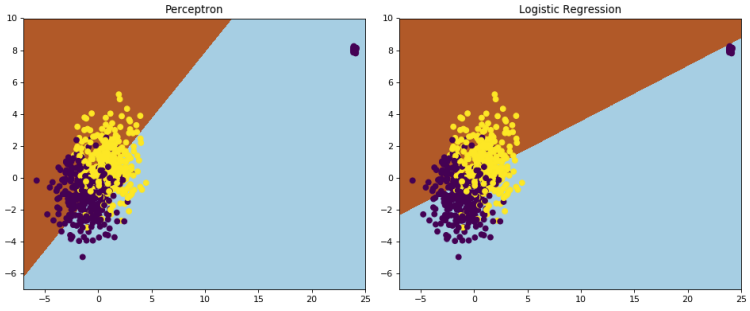


Figure 1.3: Tracé de la ligne séparatrice apprise par plusieurs algorithmes de classification sur la base de donnée représentée par les points tracés dans la même figure.Il y a un groupe de 30 outliers parmi les 300 points de la base de données, les outliers se situent dans le coin supérieur droit et perturbent l'estimation de la frontière séparatrice entre les deux classes.

$\mathcal{X} \subset \mathbb{R}^d$ et $\mathcal{Y} \subset \mathbb{R}$. Soit $\mathcal{F}$ un ensemble de fonctions $f : \mathcal{X} \to \mathcal{Y}$. Le but est de trouver f^* tel que $f^*(X)$ soit une bonne approximation de Y. La qualité de l'approximation est quantifiée par le

risque $R(f)$ d'une fonction f qui est définie par

$$R(f) = \mathbb{E}[\ell(f(X), Y)],$$

où $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ est une fonction de perte (i.e. une fonction à valeur positive, et telle que $\ell(y, y) = 0$). Par exemple, en régression on peut utiliser $\ell(f(x), y) = (f(x) - y)^2$ et le risque $R(f)$ est alors le risque quadratique moyen. Le problème se reformule donc en un problème d'optimisation, on cherche f qui minimise $R(f)$ sur l'ensemble $\mathcal{F}$. La classe de fonctions $\mathcal{F}$ peut par exemple être la classe des fonctions linéaires en régression linéaire ou alors l'ensemble de tous les réseaux de neurones possibles en apprentissage profond. Pour simplifier, on va supposer qu'il existe $f^* \in \mathcal{F}$ qui réalise le minimum sur $\mathcal{F}$:

$$f^* \in \operatorname*{argmin}_{f \in \mathcal{F}} R(f). \tag{1.1.1}$$

Soit $\widehat{f}$ un estimateur de f^* basé sur $(X_1, Y_1), \dots, (X_n, Y_n)$, des copies i.i.d du couple (X, Y). Pour étudier l'efficacité de $\widehat{f}$, nous allons principalement utiliser des inégalités oracles. Une inégalité oracle donne une borne supérieure sur l'excès de risque $R(\widehat{f}) - R(f^*)$. Remarquez que la première espérance dans la définition de l'excès de risque est aléatoire, en effet nous n'intégrons que par rapport à l'aléa de (X, Y), et $R(\widehat{f})$ est donc toujours aléatoire. Le type de résultat recherché est donc de trouver $\Delta_{n,\delta}(\mathcal{F})$ tel que pour tout $\delta \in (0, 1)$,

$$\mathbb{P}\Big(R(\widehat{f}) - R(f^*) \leq \Delta_{n,\delta}\Big) \geq 1 - \delta.$$

Pour obtenir un tel résultat, l'outil principal que nous allons utiliser est la théorie des inégalités de concentration. Les inégalités de concentration nous permettent de contrôler la moyenne empirique, en particulier elles permettent de borner la probabilité que la moyenne empirique est éloignée de plus de $t > 0$ de la moyenne théorique, quantifiant ainsi de façon non-asymptotique la vitesse de convergence de la moyenne empirique vers la moyenne théorique. Typiquement, pour borner l'excès de risque, on aura besoin de contrôler le risque empirique défini par

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

parce que la construction de $\widehat{f}$ donne souvent des informations sur $\widehat{R}(\widehat{f})$. C'est ici que les problèmes arrivent puisque les inégalités de concentration reposent le plus souvent sur une hypothèse de queue légère de $\ell(f(X), Y)$, $\ell(f(X), Y)$ doit être sous-gaussienne ou au moins sous-exponentielle (cf [BLM13]) et en général les inégalités de concentration ne sont pas valides dans le cas des distributions présentées dans la Section 2.1.1. Notre but est de trouver un estimateur ayant de bonnes propriétés de concentration même quand les données sont à queues lourdes ou corrompues. Les articles qui sont à la base de ce courant de pensée sont [Cat12] et [DLLO16].

Un des problèmes classiques dans ce champ de recherche est de trouver un estimateur robuste de la moyenne (ceci peut être vu comme une minimisation d'un risque où ℓ est la perte quadratique et $\mathcal{F}$ est l'ensemble des fonctions constantes). Nous cherchons un estimateur qui aurait les mêmes garanties que la moyenne empirique dans le cas gaussien et qui garderait ces garanties aussi quand les données ne sont pas gaussiennes. De façon informelle, si $\widehat{\mu}$ est un estimateur robuste de la moyenne basé sur $X_1, \dots, X_n$ i.i.d ayant un second moment fini, soient $W_1, \dots, W_n$ des variables i.i.d avec pour commune loi $\mathcal{N}(\mathbb{E}[X], Var(X))$, on cherche $C > 0$ tel que pour tout $\delta > 0$

$$\mathbb{P}(|\widehat{\mu} - \mathbb{E}[X]| > \delta) \leq C\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} W_i - \mathbb{E}[X]\right| > \delta\right). \tag{1.1.2}$$

Quand nous voulons prouver l'efficacité d'un estimateur, on cherchera à prouver une inégalité similaire à l'Équation (2.1.2).

1.1.3 Aspect infinitésimal de la robustesse

Il n'y a pas une définition de la robustesse qui soit acceptée par tous, Hampel a proposé d'utiliser la fonction la fonction d'influence pour quantifier la robustesse d'un estimateur. La fonction d'influence est définie comme la dérivée de Gâteau de T dans la direction de la distribution de Dirac (voir [Fer83]). Soit $D_x(t) = \mathbb{1}\{t \geq x\}$ la fonction de répartition de la distribution de Dirac pour $x \in \mathbb{R}$, la fonction d'influence est définie pour tout $x \in \mathbb{R}$ par,

$$IF(x, T, F) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon D_x) - T(F)}{\varepsilon}. \tag{1.1.3}$$

Hampel [Ham71] propose de quantifier la robustesse de T en F en utilisant $\sup_x |IF(x, T, F)|$. Une des raisons de ce choix est qu'il permet la formulation suivante du développement de Taylor de T (en supposant que T est suffisamment régulier) : soient F et G des fonctions de répartitions,

$$T(F) = T(G) + \int IF(x, T, F) d(G - F)(x) + R(F, G)$$

où le terme de reste $R(F, G)$ est négligeable comparé aux autres termes si certaines conditions sont vérifiées sur T, F et G. Ainsi, si $\sup_x |IF(x, T, F)| < \infty$, on peut alors montrer que $|T(F) - T(G)|$ est petit dès que $d(F, G)$ est petit (pour d la distance de Kolmogorov).

1.2 État de l'art dans l'estimation robuste de la moyenne

1.2.1 M-estimateurs et la fonction d'influence

Dans cette section on s'intéresse à l'estimation d'un paramètre de localisation. Soit $X \sim P$ pour P une distribution de probabilité sur $\mathbb{R}^d$, soit ρ une fonction croissante de $\mathbb{R}_+$ à valeur dans $\mathbb{R}_+$, $T(P)$ est alors définie par le problème d'optimisation suivant:

$$T(P) \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \mathbb{E}[\rho(\|X - \theta\|)], \tag{1.2.1}$$

où $\|\cdot\|$ est la norme euclidienne. Alternativement, si ρ est suffisamment régulière (ce qui sera le cas dans cette thèse), on définit $T(P)$ par

$$\mathbb{E}\left[\frac{X - T(P)}{\|X - T(P)\|} \psi(\|X - T(P)\|)\right] = 0, \tag{1.2.2}$$

où $\psi = \rho'$ est appelée la fonction de score. Pour $\psi(x) = x$, on retrouve la moyenne $T(P) = \mathbb{E}[X]$ et pour $\psi(x) = 1$, la médiane. Si ψ est bornée, $T(P)$ peut être vue comme une médiane géométrique lissée [Min15, CG17].

L'estimateur obtenu par plug-in de la densité empirique $\widehat{P}_n$ dans l'équation (2.2.2) est appelé M-estimateur associé à ψ, et sera noté $T(\widehat{P}_n)$. On calcule $T(\widehat{P}_n)$ à partir d'un échantillon i.i.d

sample $X_1, \dots, X_n$ en utilisant l'équation suivante :

$$\sum_{i=1}^{n} \frac{X_i - T(\widehat{P}_n)}{\|X_i - T(\widehat{P}_n)\|} \psi(\|X_i - T(\widehat{P}_n)\|) = 0. \tag{1.2.3}$$

Pour $\psi(x) = x$ on obtient $T(\widehat{P}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$ et pour $\psi(x) = 1$, $T(\widehat{P}_n)$ est la médiane géométrique empirique. Selon le choix de ψ que l'on fait, le M-estimateur qui en résulte peut être plus robuste aux outliers et aux distributions à queues lourdes que la moyenne empirique, et plus efficace que la médiane dans un cas non-corrompu. Les M-estimateurs sont particulièrement intéressants puisque leurs fonctions d'influence ont une forme relativement simple :

$$\mathrm{IF}(x, T, P) = M_{P,T}^{-1} \frac{x - T(P)}{\|x - T(P)\|} \psi(\|x - T(P)\|), \tag{1.2.4}$$

où $M_{P,T}$ est une matrice inversible qui ne dépend pas de x (la formule explicite de $M_{P,T}$ existe et peut être trouvée par exemple dans [HRRS86, Eq 4.2.9, Section 4.2C.] cependant elle ne sera pas utilisée dans cette étude). En particulier, nous allons étudier les trois fonctions ψ suivantes (représentées dans la Figure 2.4):

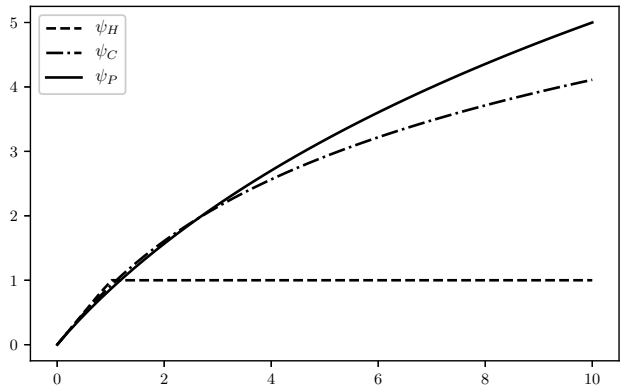


Figure 1.4: Tracé de ψ_H etψ_C pour $\beta = 1$. ψ_P est tracé pour $\beta = 10$ et $p = 5$.

Estimateur de Huber. Soit $\beta > 0$. Pour tout $x \geq 0$, on définit

$$\psi_H(x) = x\, \mathbb{1}\{x \leq \beta\} + \beta\, \mathbb{1}\{x > \beta\}. \tag{1.2.5}$$

En dimension 1, le M-estimateur associé à ψ_C est appelé estimateur de Huber [Hub64]. On note T_H la fonctionnelle associée.

Estimateur de Catoni. Soit $\beta > 0$. pour tout $x \geq 0$, on définit

$$\psi_C(x) = \beta \log\left(1 + \frac{x}{\beta} + \frac{1}{2}\left(\frac{x}{\beta}\right)^2\right). \tag{1.2.6}$$

En dimension 1, le M-estimateur associé à ψ_C est l'un des estimateurs considérés par Catoni dans [Cat12]. On note T_C la fonctionnelle associée.

Estimateur Polynomial. Soit $p \in \mathbb{N}^*$, $\beta > 0$. Pour tout $x \geq 0$, on définit

$$\psi_P(x) = \frac{x}{1 + \left(\frac{x}{\beta}\right)^{1-1/p}}. \tag{1.2.7}$$

On note T_P la fonctionnelle associée.

Le point de rupture de l'estimateur de Huber tend vers 1/2 en dimension 1 alors que celui de l'estimateur de Catoni et l'estimateur Polynomiale tendent tous deux vers 0. On verra par la suite que l'estimateur de Catoni et l'estimateur Polynomiale ne sont pas robustes au sens traditionnel du terme mais on peut tout de même définir une notion plus faible de robustesse que ces deux estimateurs vérifient. Plus généralement, le comportement de $\psi(x)$ quand x tends vers l'infini sera l'indicateur principal de la robustesse de $T(\widehat{P}_n)$ (voir Théorème 10 et Corollaire 2) alors que le comportement de ψ proche de 0 contrôle la distance $\|T(P) - \mathbb{E}[X]\|$ quand P est une distribution asymétrique (voir Lemme 2).

1.2.2 Estimateur de la médiane des moyennes

Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d ayant un second moment fini. Soit $K \in \mathbb{N}$ et suppose que n est divisible par K. Fixe $B_1, \ldots, B_K$ une partition de $\{1, \ldots, n\}$ et $b \in \mathbb{N}^*$ tel que

$$\forall k \in \{1, \ldots, K\}, \quad |B_k| = b, \quad \forall k \neq j, \quad B_k \cap B_j = \emptyset \quad \text{et} \quad \cup_{k=1}^K B_k = \{1, \ldots, n\}$$

Pour tout $B \subset \{1, \ldots, n\}$, dénote la moyenne empirique sur le bloc B par

$$P_B(X_1^n) = \frac{1}{b} \sum_{i \in B} X_i,$$

l'estimateur de médiane des moyennes, qui date de [NY83, AMS99, JGV86], est définit par

$$\mathrm{MOM}_K(X_1^n) = \mathrm{Med}(P_{B_k}(X_1^n), 1 \leq k \leq K). \tag{1.2.8}$$

La médiane des moyennes est une interpolation entre la moyenne empirique et la médiane empirique et le paramètre K indique le niveau de robustesse de l'estimation puisque moins de $K/2$ outliers ne peuvent corrompre au plus que $K/2$ blocs, ce qui implique que la médiane dans l'Équation (2.2.8) est toujours égale à la moyenne empirique d'un bloc non corrompu. Si n n'est pas divisible par K, on peut toujours utiliser des blocs de tailles différentes mais la théorie est plus facile à énoncer avec des blocs de même taille. Le point de rupture de la médiane des moyennes est de $\frac{1}{n}\lceil K/2 \rceil$. Pour plus d'information sur les déviations de la médiane des moyennes, voir [DLLO16, MS17, LCB19, LSC20, Min18]. Pour des résultats asymptotiques, voir [Min20]. La médiane des moyennes a aussi été adaptée dans d'autres contextes, par exemple dans [BAM20] est définie une médiane des moyennes qui est "differentially private" (i.e. sensé respecter la vie privée dans une application statistique) nous verrons aussi dans la suite son application à l'apprentissage statistique.

Le principe de la médiane des moyennes est le suivant : on commence par utiliser une méthode non robuste (comme la moyenne empirique, ou les moindres carrés en régression) sur des blocs de données disjoints, et ensuite on agrège les résultats en utilisant un estimateur robuste (la médiane). Ce principe peut être utilisé pour résoudre de nombreux problèmes de manière robuste.

1.3 Contributions à l'étude des déviations d'estimateurs robustes

1.3.1 Concentration d'estimateurs robustes de la moyenne en dimension 1

En dimension 1, pour un échantillon Gaussien, on a les déviations suivantes pour la moyenne empirique [BLM13] : Soient $X_1, \dots, X_n$ i.i.d de loi $\mathcal{N}(\mu, \sigma^2)$, alors pour tout $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \sigma\sqrt{\frac{t}{2n}}\right) \leq e^{-t}. \tag{1.3.1}$$

Dans cette section, on montre que la médiane des moyennes et les M-estimateurs vérifient des bornes de déviations similaires à l'Équation (2.3.1) même quand $X_1, \dots, X_n$ ne sont pas tirés selon une Gaussienne.

Pour construire des inégalités de concentration dans un contexte de robustesse, on commence toujours par montrer que la concentration de l'estimateur peut être approchée par la concentration de la moyenne empirique d'une variable transformée plus facile à contrôler. En dimension 1, nous avons les résultats suivants, connus pour la médiane des moyennes et originaux pour les M-estimateurs.

Médiane des moyennes. En interprétant la médiane, on obtient

$$\mathbb{P}(|\mathrm{MOM}_K(X_1^n) - \mathbb{E}[X]| > \varepsilon) \leq \mathbb{P}\left(\sum_{k=1}^{K} \mathbb{1}\{|P_{B_k}(X_1^n) - \mathbb{E}[X]| > \varepsilon\} \geq \frac{K}{2}\right).$$

On a changé le problème pour se ramener au cas i.i.d. Par exemple, en utilisant l'inégalité de Hoeffding, on obtient le théorème suivant.

Theorem 1 (Déviations de la médiane des moyennes)**.** *Soient $X_1, \dots, X_n, X$ des variables aléatoires réelles i.i.d, avec une variance finie σ^2. Alors, pour tout $K \in \{1, \dots, n\}$,*

$$\mathbb{P}\left(|\mathrm{MOM}_K(X_1^n) - \mathbb{E}[X]| > 2\sigma\sqrt{\frac{K}{n}}\right) \leq e^{-K/8}. \tag{1.3.2}$$

Cette borne de déviation peut être comparée à celle obtenue par la moyenne empirique dans le cas Gaussien dans l'Équation (2.3.1)mais il y a cependant des différences notables. L'estimateur dépend d'un paramètre K qui représente le nombre de blocs, et ce nombre de blocs intervient aussi dans le niveau de confiance, selon le niveau de confiance voulu, on n'utilisera pas le même estimateur. Une autre différence avec l'Équation (2.3.1) est que tous les niveaux de confiances ne sont pas accessibles. Le terme de droite dans l'Équation (2.3.2) ne peut être fixé arbitrairement petit, on ne peut aller que jusqu'à des probabilités d'ordre e^{-n}. Ceci est en fait inévitable, il a été montré dans [DLLO16] qu'en général on ne peut avoir un estimateur qui a une concentration sous-gaussienne autour de la moyenne pour tous les niveaux de confiance en même temps.

Ainsi, la médiane des moyennes est appropriée pour estimer la moyenne même quand les données sont à queues lourdes (second moment fini). On peut aussi montrer des bornes de déviations similaires dans un contexte $\mathcal{I} \cup \mathcal{O}$ (voir [Ler19]).

M-estimateur. Dans le cas des M-estimateurs, on montre le théorème suivant qui contrôle les déviations d'un M-estimateur.

Theorem 2 ([Mat20b]). *Soient $X_1, \dots, X_n, X$ des variables aléatoires réelles i.i.d de loi P, soit ψ l'une des trois fonctions de scores définies dans la Section 2.2.1, on suppose que $T(P)$ et $T(\widehat{P}_n)$ existent et sont uniques. On note $\psi_{odd}(x) = \text{sign}(x)\psi(|x|)$, le résultat suivant est alors vérifié.*

- *Pour tout $\lambda > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\psi_{odd}(X_i - T(P))\right| > 3\lambda\right) \leq \mathbb{P}\Big(\Big|T(\widehat{P}_n) - T(P)\Big| > \lambda\Big).$$

- *Si de plus $V = \mathbb{E}[\psi(|X - T(P)|)^2] \leq \psi(\beta/2)^2/2 < \infty$,alors pour tout $\lambda \in (0, \beta/2)$,*

$$\mathbb{P}\Big(\Big|T(\widehat{P}_n) - T(P)\Big| > \lambda\Big) \leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\psi_{odd}(X_i - T(P))\right| > \frac{\lambda\gamma}{4}\right) + e^{-n\gamma^2/8}. \tag{1.3.3}$$

où $\gamma = 1$ si $\psi = \psi_h$, $\gamma = 4/5$ si $\psi = \psi_C$ et $\gamma = 1/4$ si $\psi = \psi_P$.

Équation (2.3.3) montre que les déviations de $T(\widehat{P}_n)$ sont contrôlées à travers celles de $\psi_{odd}(X - T(P))$ et que le paramètre de variance est $V = \mathbb{E}[\psi(|X - T(P)|)^2]$. Ayant que ψ est concave sur $\mathbb{R}_+$ et $\psi(0) = 0$,on obtient que $\psi_{odd}(X - T(P))$ a une queue de distribution plus légère que X et par exemple si ψ est bornée on a que $\psi_{odd}(X - T(P))$ est sous-gaussien. Cela rend le contrôle des déviations de $T(\widehat{P}_n)$ facile puisque une somme de variables à queue légère i.i.d peut être gérée en utilisant des inégalités de concentration classiques. Une inégalité de concentration pour M-estimateurs dans un cas plus général (multivarié et en contexte corrompue) est présenté dans la sous-section suivante. Pour plus d'information sur la dimension 1, voir [Mat20b].

1.3.2 Estimation robuste de la moyenne multivariée

Dans cette section, nous étudions l'estimation robuste de la moyenne en dimension $d > 1$. Le théorème suivant, conséquence de l'inégalité d'Hanson-Wright, nous donne les déviations de la moyenne empirique dans le cas gaussien.

Theorem 3 ([HW71]). *Soit $X \sim \mathcal{N}(0, \Sigma)$ et $X_1, \dots, X_n$ des copies i.i.d de X, avec Σ une matrice définie positive. Alors, pour tout $t > 0$,*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}X_i\right\|^2 > \frac{2Tr(\Sigma)}{n} + \frac{9t\|\Sigma\|_{op}}{n}\right) \leq e^{-t},$$

où $\|\cdot\|_{op}$ est la norme d'opérateur associée à la norme euclidienne $\|\cdot\|$.

Les vitesses de convergence exhibées dans le Théorème 12 est notre objectif quand on estime la moyenne. Remarque que le Théorème 12 nous donne des déviations d'ordre $O(1/\sqrt{n})$ similaire au cas $d = 1$ mais au-delà de ça, nous devons aussi faire attention au numérateur, $\|\Sigma\|_{op}$ peut être beaucoup plus petit que $Tr(\Sigma)$ (par exemple, si $\Sigma = I_d$ on $\|\Sigma\|_{op} = 1$ et $Tr(\Sigma) = d$).

Pour pouvoir utiliser un estimateur de la moyenne dans $\mathbb{R}^d$, on doit faire attention à ce que la dimension n'ai pas d'effet non voulu sur l'erreur d'estimation. En particulier, en estimation robuste, l'erreur due à la corruption ne doit pas augmenter avec la dimension. Dans ce contexte, il y a eu de nombreuses propositions d'estimateurs. Premièrement, il y a les estimateurs qui ont de bonnes garanties théoriques mais qui ne sont pas calculables en pratique, par exemple les estimateurs basés sur l'agrégation d'estimateurs uni-dimensionnel comme dans [Ler19, Theorem 44] ou des estimateurs basés sur la profondeur, voir [DG92, CGR+18], par exemple avec la médiane de Tukey. Une seconde famille d'estimateurs sont ceux qui sont calculables en pratique mais dont les garanties théoriques ne sont pas suffisantes, par exemple la médiane coordonnée par coordonnée ou la médiane géométrique [Min15]. Récemment, il y a eu quelques propositions d'algorithmes qui se disent en même temps calculables et minimax [DKP20, DL19, Hop20] mais en pratique la plupart de ces algorithmes sont trop long à exécuter.

Dans le cas de la médiane de Tukey, [CGR+18] montre que dans un contexte de données corrompues où $X_1, \dots, X_n$ sont i.i.d de loi $(1-\varepsilon)P + \varepsilon H$, si P est une gaussienne et $\varepsilon \leq 1/\sqrt{n}$, alors on peut estimer la moyenne efficacement (avec une vitesse minimax, i.e. les mêmes vitesses que l'inégalité de Hanson-Wright). Au contraire, si P a un second moment fini et aucun moment d'ordre supérieur à 2 n'est fini, on doit demander à ce que $\varepsilon \leq 1/n$ pour récupérer les vitesses minimax. Cela peut être interprété comme une sorte de point de rupture : quelle doit être la valeur de ε pour que l'ordre de grandeur de la vitesse de convergence ne change pas (voir [LL20]).

Pour continuer notre étude, une version multivariée du théorème 10 est montrée dans [Mat20b], cependant il manque encore un ingrédient à notre analyse puisque pour l'instant nous n'avons la maitrise que de la distance $\|T(\widehat{P}_n) - T(P)\|$ et pas de la distance $\|T(P) - \mathbb{E}[X]$.

Lemma 1. *Suppose que ψ est C^k avec une dérivée k^{eme} bornée, $\psi'(0) = 1$ et pour $2 \leq j \leq k-1$, $\psi^{(j)}(0) = 0$. Soit X une variable aléatoire sur $\mathbb{R}^d$ telle que $\mathbb{E}[\|X\|^k] < \infty$, alors,*

$$\|\mathbb{E}[X] - T(P)\| \leq \frac{2\|\psi^{(k)}\|_\infty}{\gamma k! \beta^{k-1}} \mathbb{E}\big[\|X - T(P)\|^k\big] \tag{1.3.4}$$

où $\gamma = 1$ si $\psi = \psi_h$, $\gamma = 4/5$ si $\psi = \psi_C$ et $\gamma = 1/4$ si $\psi = \psi_P$.

De plus, on peut montrer que dans le cas de l'estimateur de Huber, on a que $\|T_H(P) - \mathbb{E}[X]\|$ est d'ordre $O(1/\beta^{q-1})$ où q est le nombre de moments finis de P. On peut montrer que cet ordre de grandeur est optimal en β dès que la distribution P est asymétrique (Si P est symétrique $\|T_H(P) - \mathbb{E}[X]\| = 0$ et il n'y a pas besoin d'une telle borne).

Cette séparation de l'effet des déviations et de celui de la distance $\|T(P) - \mathbb{E}[X]\|$ peut être comparé au compromis biais variance souvent constaté en statistiques.

Dans le cas de l'estimateur de Huber (dont on rappelle que la fonction de score est définie pour $x \geq 0$ par $\psi_H(x) = x \wedge \beta$) on est ramené à contrôler la somme de variables aléatoires bornées i.i.d.

Dans [Mat20a] on montre le résultat suivant, que l'on présente ici sous forme informelle. Soient $X_1, \dots, X_n$ i.i.d de loi $(1-\varepsilon)P + \varepsilon Q$ où P est une distribution ayant q moments finis, alors si l'on suppose quelques hypothèses sur ψ et n, il existe une constante absolue $C > 0$ telle que, pour

tout $0 < \lambda \lesssim n$, avec probabilité plus grande que $1 - 5\exp(-\lambda/8)$, on a

$$\left\|\mathbb{E}[X] - T_H(\widehat{P}_n)\right\| \lesssim \frac{\sqrt{Tr(\Sigma)} + \sqrt{\|\Sigma\|_{op}\lambda}}{\sqrt{n}} \bigvee \mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q} \varepsilon^{1-1/q} g\left(\lambda, \frac{\varepsilon^{1/2-1/q}}{M}, \frac{1}{M^{\frac{q}{2}} n^{\frac{q-2}{4}}}\right). \tag{1.3.5}$$

Où $\lesssim$ représente le fait d'être plus petit à une constante multiplicative près, $M = \frac{\sqrt{Tr(\Sigma)}}{\mathbb{E}[\|X-\mathbb{E}[X]\|^q]^{1/q}}$, et $g : \mathbb{R}^3 \to \mathbb{R}_+$ est tel que $g(\lambda, x, y) = 1 + o(\lambda) + o(x) + o(y)$ pour (λ, x, y) qui tends vers 0.

La dépendance en le nombre de moments finis fait le lien entre les deux contextes déjà connus : quand P a deux moments finis, la borne est d'ordre $\sqrt{\varepsilon}$ comme dans [DL19, DKP20], on a alors besoin de $\varepsilon \leq 1/n$ pour récupérer des vitesses en $1/\sqrt{n}$ similaires à l'inégalité d'Hanson-Wright. Si on a que P est gaussien, la dépendance en ε devient linéaire ce qui revient à demander $\varepsilon \leq 1/\sqrt{n}$ pour récupérer les vitesses voulues. l'Équation (2.3.6) interpole entre ces deux extrêmes.

Une conséquence peut être surprenante de cette séparation de l'effet du biais et de la variance de $T(\widehat{P}_n)$ est que l'on peut avoir une vitesse de l'ordre de ε quand les inliers (points non outliers) sont symétriques, il n'y a pas besoin que P soit Gaussien et ceci est vrai même quand le second moment de P n'est pas fini.

Une généralisation possible de la médiane des moyennes est de remplacer la médiane empirique par un estimateur de huber et prendre ainsi l'estimateur de huber des moyennes sur les blocs. On peut montrer que dans ce cadre, il n'est utile de faire des blocs que si les données proviennent d'une distribution très asymétrique, voir [Mat20a].

1.4 Contributions aux résultats asymptotiques de statistiques robustes : une notion plus faible de la robustesse pour des estimateurs plus efficaces

1.4.1 Continuité des M-estimateurs asymptotiques

Prenant naissance dans le travail de Hampel [Ham71], une définition de la robustesse est donnée comme une continuité de T si l'estimateur est un estimateur de type plug-in : T est continue à une distribution de probabilité P si pour toute distribution de probabilité Q,

$$\forall \varepsilon > 0, \exists \delta > 0, \text{tel que} \quad d(P, Q) \leq \delta \Rightarrow |T(P) - T(Q)| \leq \varepsilon.$$

Ceci est une propriété importante de T qui se traduit en français en disant qu'un petit changement dans la loi P ne devrait causer que de petits changements dans la valeur de $T(P)$. Par exemple, on peut montrer en dimension $d = 1$ que si la fonction de score ψ est continue et bornée (la fonction d'influence est alors bornée), alors T est continue pour la distance de Kolmogorov, ce qui signifie que T est robuste dans un voisinage de corruption pour la distance de Kolmogorov comme cela a été défini dans la Section 2.1.3. Il y a aussi des travaux qui considèrent d'autres distances comme la distance de Prokhorov, la distance en variation totale [HR09], et plus récemment la distance de Hellinger avec les travaux sur la rho-aggregation [BBS17, BGH14]. Toutes ces distances entre distributions sont insensibles à des outliers arbitraires, c'est à dire que pour toute distribution

P, Q, on a $d(P, (1-\varepsilon)P+\varepsilon Q) \leq \varepsilon$. Un tel voisinage peut donc contenir des outliers arbitrairement loin de l'origine.

Si au lieu de prendre en compte des outliers arbitraires, on suppose qu'ils vérifient des conditions faibles, il se peut que l'ensemble des fonctions T continues soit plus grand. C'est ce qui motive la définition d'une nouvelle famille de distances entre distributions.

Soit $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ et on note $\mathcal{P}_\psi = \{P \in \mathcal{P} : \mathbb{E}_P[\psi(\|X\|)] < \infty\}$. Pour tout $P, Q \in \mathcal{P}_\psi$, on définit

$$W_\psi(P, Q) = \sup_{h \preceq \psi} \left\{ \int h(x) dP(x) - \int h(x) dQ(x) \right\}, \tag{1.4.1}$$

où $h : \mathbb{R}^d \to \mathbb{R}$ vérifie $h \preceq \psi$ si et seulement si pour tout $x, y \in \mathbb{R}^d$, $|h(x) - h(y)| \leq \psi(\|x - y\|)$. W_ψ n'est pas une distance inconnue jusqu'alors puisque c'est en fait une distance Wasserstein-1 dans l'espace métrique $(\mathbb{R}^d, d_\psi)$ où $d_\psi(x, y) = \psi(\|x - y\|)$. Si ψ est l'identité, on récupère la distance Wasserstein-1 usuelle et dans le cas extrême où ψ est la fonction constante égale à 1 on récupère la distance de variation totale et on peut se convaincre que W_ψ est une distance plus faible que la distance en variation totale puisqu'on doit en demander plus à la distribution Q pour que $W_\psi((1-\varepsilon)P + \varepsilon Q, P) \xrightarrow[\varepsilon \to 0]{} 0$. Le théorème suivant est central dans l'article [Mat20b].

Theorem 4. *Soit ψ une des trois fonctions ψ_H, ψ_C ou ψ_P. Soit T la fonctionnelle construite à partir de ψ par le biais de l'Équation (2.2.2), et soit $P \in \mathcal{P}_\psi$. On suppose que $\psi(+\infty) > \mathbb{E}_P[\psi(\|X\|)]$ et que $\|X\|$ est presque sûrement finie.*
Alors, T est continue en P pour la distance W_ψ sur $\mathcal{P}_\psi$. Ce qui signifie que pour tout $Q \in \mathcal{P}_\psi$,

$$\|T(P) - T(Q)\| \xrightarrow[W_\psi(P,Q) \to 0]{} 0.$$

Le Théorème 13 nous dit que le choix de la fonction ψ nous donne la distance W_ψ pour laquelle T est continue et de façon informelle, cela définit aussi la corruption à laquelle T peut résister. Dans la Section 2.4.2, nous précisons cette remarque et on montre l'usage que l'on peut faire de W_ψ.

1.4.2 Stabilité asymptotique des M-estimateurs

Le choix de la fonction de score ψ influe beaucoup sur la robustesse du M-estimateur associée à cette fonction de score. En particulier, si ψ est bornée, alors l'estimateur est robuste au sens de Hampel (voir [HR09]) et si ψ est proche de l'identité au voisinage de 0,alors le biais de l'estimateur associé est petit (voir Lemme 2). Si au contraire ψ n'est pas borné, cela ne signifie pas forcément que l'on perd toute la robustesse de l'estimateur mais nous devons faire des hypothèses supplémentaires sur les outliers pour que l'estimateur soit encore consistant dans un contexte corrompu. Plus précisément, on a le résultat suivant.

Corollary 1. *Supposons que l'on est dans le contexte $\mathcal{I} \cup \mathcal{O}$ où $(X_j)_{j \in \mathcal{I}}$ sont i.i.d suivant P et $(X_j)_{j \in \mathcal{O}}$ sont tous égaux à $g(n)u$ pour un certain $u \in \mathbb{R}^d$, $u \neq 0$ et $g : \mathbb{N} \to \mathbb{R}$ croissante. On note $|O| = k_n$ le cardinal de l'ensemble des outliers. Le résultat suivant est vérifié.*

Éstimateur de Huber *Soit P une distribution de probabilité sur $\mathbb{R}^d$ et suppose que $\mathbb{E}[\psi_H(\|X\|)] <$*

$\beta < \infty$, *alors*

$$\left(\frac{k_n}{n} \xrightarrow[n\to\infty]{} 0\right) \Rightarrow \left(T_H\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right) \xrightarrow[n\to\infty]{P} T_H(P)\right).$$

Éstimateur de Catoni *Soit P une distribution de probabilité telle que $X \sim P$, $\mathbb{E}[\psi_C(\|X\|)] < \beta < \infty$, alors*

$$\left(\frac{k_n \log(g(n))}{n} \xrightarrow[n\to\infty]{} 0\right) \Rightarrow \left(T_C\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right) \xrightarrow[n\to\infty]{P} T_C(P)\right).$$

Éstimateur polynomiale *Soit P une distribution de probabilité telle que $X \sim P$, $\mathbb{E}[\psi_P(\|X\|)] < \beta < \infty$, alors*

$$\left(\frac{k_n g(n)^{1/p}}{n} \xrightarrow[n\to\infty]{} 0\right) \Rightarrow \left(T_P\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right) \xrightarrow[n\to\infty]{P} T_P(P)\right).$$

Ce corollaire nous dit à quel point les outliers peuvent être grands et en quel nombre, pour qu'ils n'affectent pas le comportement de l'estimateur. Pour l'estimateur de Huber, il n'y a pas de restriction sur g tant que le nombre d'outliers $k_n = o(n)$. Pour l'estimateur de Catoni, si k_n est borné par exemple (i.e. le nombre d'outliers est fini), alors $g(n)$ soit être négligeable par rapport à $\exp(n)$ pour conserver des bonnes propriétés d'estimations. Pour l'estimateur Polynomiale, si k_n est borné alors $g(n)$ doit être négligeable comparé à n^p. En pratique, si les hypothèses de convergence de l'estimateur de Catoni sont vérifiées par exemple, alors celui-ci sera plus efficace que l'estimateur de Huber sur ces mêmes données. Ce résultat nous donne des informations sur comment construire des M-estimateurs quand on a une idée de l'échelle à laquelle se placent les outliers.

1.5 Contributions à l'apprentissage robuste

Dans cette section nous sommes intéressés dans l'estimation de fonctions. Nous commençons par la classification et la régression (voir [Kol11, DGL96, MRT12] pour des références sur le sujet) et ensuite nous présentons une application dans les méthodes à noyaux pour le maximum mean discrepancy (MMD) introduit dans [GBR+12]. Notre but est d'avoir des estimateurs robustes des quantités considérées.

1.5.1 Classification et régression utilisant le principe de la médiane de moyenne

En classification et en régression, nous utilisons les M-estimateurs et MOM (médiane des moyennes) pour rendre robuste un estimateur déjà existant. En particulier, nous nous sommes intéressés à la régression/classification linéaire à travers la régression logistique et les moindres carrés ordinaires, mais nous avons aussi appliqué nos méthodes au cas non-linéaire avec des méthodes à noyaux en classification. Le principe de la médiane des moyenne a été appliqué en apprentissage statistique dans plusieurs travaux notamment pour les tournois MOM [LM19a, LM19c] qui sont efficaces

théoriquement mais incalculable en pratique, il y a aussi les estimateurs minmax [LL18, LL20] et les travaux en grande dimension et sparsité [CLL19a, LM+18].

Dans l'Équation (2.1.1), nous cherchions f^* définie par

$$f^* \in \operatorname*{argmin}_{f\in\mathcal{F}} R(f) = \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}[\ell(f(X), Y)]. \tag{1.5.1}$$

où ℓ est une fonction de perte. Selon le principe de minimisation du risque empirique [Vap98] on cherche à approcher $R(f)$ en remplaçant l'espérance par la moyenne empirique, on estime ensuite f^* en utilisant $\widehat{f}$, un minimiseur du risque empirique,

$$\widehat{f} \in \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

Parmis les exemples classiques de fonction de perte ℓ il y a la perte logistique $\ell(f(x), y) = \log(1 + \exp(-f(x)y))$ en classification ou la perte quadratique $\ell(f(x), y) = (f(x) - y)^2$ en régression. Le problème vient du fait que l'estimation n'est pas bonne dès que la loi de $\ell(f(X), Y)$ est à queue lourde.

Pour décrire le comportement d'un estimateur de f^*, nous avons besoin d'une notion de complexité de $\mathcal{F}$, l'estimation est plus difficile si $\mathcal{F}$ est très complexe (i.e. réseaux de neurones). Il y a plusieurs notions standard de complexité comme la dimension VC ou l'entropie d'un espace de fonction. Dans ce travail, nous utilisons la complexité Rademacher (présentée par exemple dans [MRT12, p34]). On suppose que X appartient à un espace $\mathcal{X}$. Soit $\mathcal{G}$ un espace de fonctions $f : \mathcal{X} \to \mathbb{R}$ et on note $(\varepsilon_i)_{1\leq i\leq n}$ des variables i.i.d de loi de Rademacher, les ε_i sont supposés indépendants de $(X_i)_{1\leq i\leq n}$. La complexité de Rademacher de l'espace $\mathcal{G}$ est définie par

$$\mathrm{Rad}(\mathcal{G}) = \mathbb{E}\left[\sup_{f\in\mathcal{G}} \sum_{i=1}^{n} \varepsilon_i f(X_i)\right].$$

Si la complexité de Rademacher est grande, l'espace $\mathcal{G}$ sera considéré complexe et il sera alors plus compliqué d'estimer les fonctions de $\mathcal{G}$. Par exemple, pour les fonctions linéaires dans $\mathbb{R}^d$ dont le coefficient directeur a une norme euclidienne bornée par θ_2, on peut montrer que $\mathrm{Rad}(\mathcal{G}) \leq \theta_2\sqrt{nd}$.

Le principe de la minimisation du risque empirique robuste est la suivante: au lieu d'estimer l'espérance dans (2.5.1) en utilisant la moyenne empirique, on utilise un estimateur robuste de la moyenne

$$\widehat{f}_{rob} \in \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{E}(\ell(f(X_i), Y_i), 1 \leq i \leq n)$$

où $\widehat{E}$ est un estimateur robuste de la moyenne, typiquement un M-estimateur ou la médiane des moyennes. Premièrement, regardons le cas de la médiane des moyennes. On définit

$$\widehat{f}_{MOM,K} \in \operatorname*{argmin}_{f\in\mathcal{F}} \mathrm{MOM}_K(\ell(f(X_i), Y_i), 1 \leq i \leq n).$$

Theorem 5. *Soit $f \in \mathcal{F}$, $\theta_2 := \mathbb{E}[f(X)^2] < \infty$, on suppose que $\mathrm{Rad}(\mathcal{F}) < \infty$ et que la fonction de perte est Lipschitz dans le sens qu'il existe $L > 0$ tel que pour tout $(x, y) \in \mathcal{X} \times \mathcal{Y}$, et tout $f, f' \in \mathcal{F}$,*

$$|\ell(f(x), y) - \ell(f'(x), y)| \leq L|f(x) - f'(x)|.$$

On suppose que $n > K > 4|\mathcal{O}|$ *et on note* $\Delta = 1/4 - |\mathcal{O}|/K$. *Alors, avec probabilité plus grande que* $1 - 2e^{-2\Delta^2 K}$,

$$R(\widehat{f}_{MOM,K}) \leq \inf_{f\in\mathcal{F}} R(f) + 4L\max\left(\frac{4\mathrm{Rad}(\mathcal{F})}{n}, 2\theta_2\sqrt{\frac{K}{n}}\right)$$

L'inégalité du Théorème 14 atteint la vitesse de convergence optimale (quand on ne suppose pas de condition de marge) alors que nous avons supposé que $\mathcal{F}$ est bornée dans L^2 et que nous avons affaire à une base de données corrompue. Dans les résultats classiques d'apprentissages statistiques, $\mathcal{F}$ est supposé borné dans L^∞. Dans beaucoup de cas (classification linéaire, SVM, ...) on a $\mathrm{Rad}(\mathcal{F}) \leq O(\sqrt{n})$ et c'est pourquoi on va souvent dire que la vitesse optimale est de l'ordre de $O(1/\sqrt{n})$.

Dans l'article [MM19], nous sommes allés plus loin en incluant une condition de marge qui nous permet de récupérer des vitesses plus rapides que $1/\sqrt{n}$. La condition de marge se présente comme suivant : il existe des constantes $D > 0, \delta_B > 0$ telles que

$$Var(\ell(f(X),Y) - \ell(f^*(X),Y)) \leq D^2(R(f) - R(f^*)) \tag{1.5.2}$$

dès que $R(f) - R(f^*) \leq \delta_B$.

Pour obtenir des vitesses rapides, nous considérons un estimateur plus général dans lequel l'opérateur de médiane empirique est remplacé par un M-estimateur similaire à l'estimateur de Huber. On commence par définir un estimateur de $R(f)$ noté $\widehat{R(f)}_{\beta,K}$ et défini par

$$\sum_{k=1}^{K} \psi\left(\sqrt{b}\frac{P_{B_k}(\ell(f(X_i),Y_i)_{1\leq i\leq n}) - \widehat{R(f)}_{\beta,K}}{\beta}\right) = 0. \tag{1.5.3}$$

Où ψ est supposé impaire et continûment dérivable cinq fois. On suppose aussi que si $|x| \leq 1$, alors $\psi(x) = x$, si $|x| \geq 2$ alors ψ est contant et $x - \psi(x)$ est croissant. ψ est essentiellement une version plus régulière de ψ_H.On propose alors l'estimateur suivant de f^* :

$$\widehat{f}_{\mathrm{HOME},K} \in \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{R(f)}_{\beta,K}.$$

Avec cet estimateur,on montre que l'on peut obtenir une vitesse $O(1/n^{3/4})$ si la condition de marge est vérifiée (on montre aussi la vitesse lente en l'absence des conditions de marge). Cette vitesse est plus rapide que $O(1/\sqrt{n})$ mais même si on constate en pratique une vitesse en $O(1/n)$ nous n'avons pas réussi à montrer une telle borne. Nous avons donc défini un estimateur alternatif basé sur le principe min-max, et cet estimateur achève une vitesse optimale en $O(1/n)$ si la condition de marge est vérifiée. Ces théorèmes sont quelque peu techniques et nous présentons ici une version informelle simplifiée. Voir [MM19] pour des résultats détaillés sur ces estimateurs.

Theorem 6 (Informal)**.** *Si pour tout* $f \in \mathcal{F}$, $\mathbb{E}[\ell(f(X),Y)^4] < \infty$ *et* $\mathrm{Rad}(\mathcal{F}) < \infty$. *Alors pour* K *et* β *choisi de façon appropriée, il existe des constantes* $c_1, c_2 > 0$ *telles que avec probabilité plus grande que* $1 - e^{-s}$ *pour* $0 < s \leq c_1 K$, *on ait*

$$R(\widehat{f}_{HOME,K}) \leq \inf_{f\in\mathcal{F}} R(f) + c_2\left(\frac{\mathrm{Rad}(\mathcal{F})}{n} + \frac{s}{n^{3/4}} + \left(\frac{|\mathcal{O}|}{n}\right)^{3/4}\right).$$

Théorème 15 montre que dans un contexte sans condition de marge, tant que l'on a un quatrième moment fini, l'erreur additionnelle due à la corruption est de l'ordre de $O(1/n^{3/4})$ ce qui est négligeable comparé à $\mathrm{Rad}(\mathcal{F})/n$. On atteint donc la vitesse optimale. On présente ensuite le cas où la vitesse rapide peut être atteinte.

Theorem 7 (Informal). *Si $f \in \mathcal{F}$, $\mathbb{E}[\ell(f(X),Y)^4] < \infty$ et $\mathrm{Rad}(\mathcal{F}) < \infty$, si de plus la condition de marge (2.5.2) est vérifiée. Alors, il existe un estimateur $\widehat{f}_{fr}$ tel que avec probabilité plus grande que $1-e^{-s}$ pour $s \leq s_{max}$ où $s_{max} \xrightarrow[n\to\infty]{} \infty$, on ait*

$$R(\widehat{f}_{fr}) \leq \inf_{f\in\mathcal{F}} R(f) + c_2\left(\frac{\mathrm{Rad}(\mathcal{F}-f^*)}{n} + \frac{s}{n} + \frac{|\mathcal{O}|}{n}\right)$$

où $\mathcal{F}-f^* = \{x \mapsto f(x)-f^*(x), f\in\mathcal{F}\}$.

En supposant quelques conditions sur $\mathcal{F}$, on peut montrer que la borne du Théorème 16 est d'ordre $O(1/n)$ ce qui est minimax optimal pour ce problème quand la condition de marge est vérifiée. La description de l'estimateur $\widehat{f}_{fr}$ est plus compliquée que celle de $\widehat{f}_{HOME,K}$ et le lecteur est encouragé à lire [MM19] pour la description précise de cet estimateur.

1.5.2 Méthode à noyaux et estimation robuste en dimension infinie.

Soit $\mathcal{X}$ un ensemble sur lequel $\mathcal{K}$ est une fonction de noyau, on peut représenter une densité de probabilité P sur $\mathcal{X}$ comme une moyenne dans le RKHS $\mathcal{H}_\mathcal{K}$.

$$\mu_P = \int_\mathcal{X} \varphi(x)dP(x), \quad \varphi(x) := \mathcal{K}(\cdot,x).$$

Ce procédé est appelé plongement de la moyenne et en utilisant ce plongement, on introduit la distance entre distribution suivante appelée "maximum mean discrepancy" (MMD),

$$MMD(P,Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_\mathcal{K}} = \sup_{f\in\mathcal{B}_\mathcal{K}} \langle \mu_P - \mu_Q, f\rangle_{\mathcal{H}_\mathcal{K}},$$

où $\mathcal{B}_\mathcal{K} = \{f\in\mathcal{H}_\mathcal{K} : \|f\|_{\mathcal{H}_\mathcal{K}} \leq 1\}$. Les méthodes à noyaux sont très performantes pour travailler avec des données structurées, l'ADN par exemple, parce qu'on peut construire un noyau adapté à la structure. Le MMD peut ensuite être utilisé par exemple pour construire un test de comparaison entre deux populations, c'est ce que nous avons fait pour illustrer notre méthode dans [LSML19].

Dans le calcul du MMD, nous devons calculer une moyenne en dimension infinie et nous voyons ceci comme un problème d'estimation robuste de la moyenne. On suppose que l'on a accès à $X_1,\dots,X_n$ i.i.d de loi P et $Y_1,\dots,Y_n$ i.i.d de loi Q. On note $P_{B,x} = \frac{1}{|B|}\sum_{i\in B}\delta_{x_i}$ la mesure empirique associée à $(x_i)_{i\in B}$, on note $B_1,\dots,B_K$ une equi-partition de $\{1,\dots,n\}$. On propose l'estimateur suivant du MMD.

$$\widehat{MMD}_K(P,Q) = \sup_{f\in\mathcal{B}_\mathcal{K}} \mathrm{Med}\Big(\langle f, \mu_{P_{B_k},x} - \mu_{P_{B_k},y}, \quad 1\leq k\leq K\Big).$$

Cet estimateur présente des avantages théoriques en terme de robustesse mais aussi des avantages pratiques en terme de temps de calcul (voir Section 2.6.4). Nous avons les garanties suivantes en terme de concentration.Soit $(e_i)_{i\in I}$ une base orthonormale dénombrable de $\mathcal{H}_\mathcal{K}$ (qui existe car $\mathcal{H}_\mathcal{K}$ est séparable), on définit $\|A\|_1 = \sum_{i\in I}\langle (A*A)^{1/2}e_i, e_i\rangle_{\mathcal{H}_\mathcal{K}}$ où A^* est l'opérateur adjoint de A et $\|A\|$ la norme d'opérateur de A.

Theorem 8. *Suppose que Σ_P est un opérateur linéaire sur $\mathcal{H}_K$ avec $\|\Sigma_P\|_1 < \infty$. Suppose que la base de données $(x_i, y_i)_{i\leq n}$ est corrompue par n_c outliers dans le cadre de travail $\mathcal{I} \cup \mathcal{O}$ décrit dans la Section 2.1.1 (i.e. Il peut y avoir n_c couples outliers $(x_{i_1}, y_{i_1}), \dots, (x_{i_{n_c}}, y_{i_{n_c}})$ sur lesquels on ne fait aucune hypothèse). Soit $\delta \in (0, 1/2]$ tel que $n_c \leq K(1/2 - \delta)$. Alors, pour tout $\eta \in (0,1)$ avec $K = 72\delta^{-2}ln(1/\eta)$ tel que $K \in \left(\frac{n_c}{1/2-\delta}, \frac{n}{2}\right)$, avec probabilité plus grande que $1-\eta$, on a*

$$\left|\widehat{MMD}_K(P,Q) - MMD(P,Q)\right| \leq \frac{12}{\delta}\max\left(\sqrt{\frac{(\|\Sigma_P\| + \|\Sigma_Q\|)ln(1/\eta)}{\delta n}}, 2\sqrt{\frac{Tr(\Sigma_P) + Tr(\Sigma_Q)}{n}}\right).$$

Le Théorème 17 montre que l'estimateur $\widehat{MMD}_K$ atteint la vitesse $O(1/\sqrt{n})$ qui est optimale pour ce problème. Il indique aussi que notre estimateur est robuste aux outliers avec un point de rupture proche de $K/2$. La vitesse du Théorème 17 est très similaire à celle de l'inégalité d'Hanson-Wright, cela peut être vu comme une extension infinie-dimensionnelle de l'inégalité d'Hanson-Wright.

1.6 La robustesse en pratique, quelques contributions aux algorithmes robustes et illustrations numériques

1.6.1 Algorithmes robustes pour l'apprentissage supervisé et non-supervisé

Dans cette section, on étudie les deux estimateurs $\widehat{f}_{MOM,K}$ et $\widehat{f}_{HOME,K}$ et on illustre leurs performances sur la base de données représentée dans la Figure 2.7. Le second estimateur a une implémentation simple utilisant une descente de gradient sur le risque défini par l'Équation (2.5.3), on peut en effet prendre la dérivée de cette expression pour obtenir le gradient de $\widehat{R}_{K,\beta}$ par rapport à f. Par contre, la médiane des moyennes utilisée pour définir $\widehat{f}_{MOM,K}$ n'est pas différentiable, ce n'est pas véritablement un problème car ce risque empirique robuste est en réalité différentiable presque partout. Soit B_{Med} ce que l'on va appeler le bloc médian dans le sens que l'on a

$$\begin{aligned}\frac{1}{b}\sum_{i\in B_{\text{Med}}} \ell(f(X_i), Y_i) = P_{B_{\text{Med}}}(\ell(f(X_i), Y_i)) &= \text{Med}\{P_{B_k}(\ell(f(X_i), Y_i)), 1 \leq k \leq K\} \\ &= \text{MOM}_K(\ell(f(X_i), Y_i), 1 \leq i \leq n).\end{aligned}$$

On montre dans l'article [LLM20] que prendre la dérivée du critère MOM revient à prendre la dérivée seulement sur le bloc médian B_{Med}. On a presque partout (dans un sens qui est rendu rigoureux dans [LLM20]) que

$$\frac{d}{df}\text{MOM}_K(\ell(f(X_i), Y_i), 1 \leq i \leq n) = \sum_{i\in B_{\text{Med}}} \frac{d}{df}\ell(f(X_i), Y_i). \tag{1.6.1}$$

Maintenant que l'on a un gradient, on peut donc utiliser un algorithme de descente de gradient. Le problème est que la fonction objectif n'est plus convexe puisque l'on fait des blocs et il peut donc y avoir des minima locaux. Les blocs sont construits arbitrairement et sont supposés fixes en théorie. En pratique, il est plus efficace de changer les blocs à chaque étape de gradient, autrement dit on mélange les données à chaque itération. Un des effets de ce mélange à chaque itération

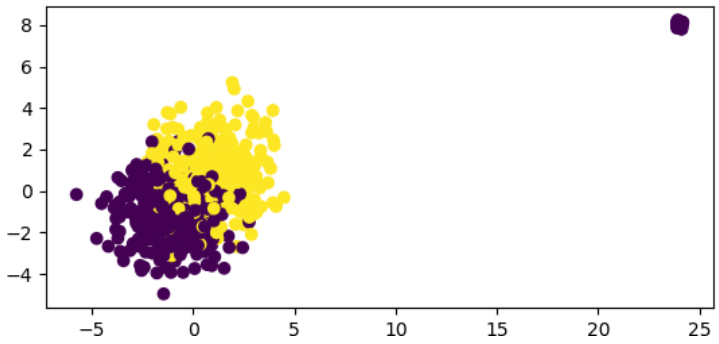


Figure 1.5: Représentation d'une base de donnée corrompue pour la classification. La base de données est composée de deux nuages Gaussiens de 300 points chacun et de 30 outliers situés en haut à droite de la figure.

est l'introduction de bruit avec la même idée que la descente de gradient stochastique où ici la partie stochastique vient de la permutation aléatoire utilisée pour mélanger les données. Il n'y a pas de preuve de convergence de l'algorithme mais des résultats théoriques sur les U-Statistics (dans l'article [MM19] et des observations empiriques semblent montrer que l'algorithme est bien construit.

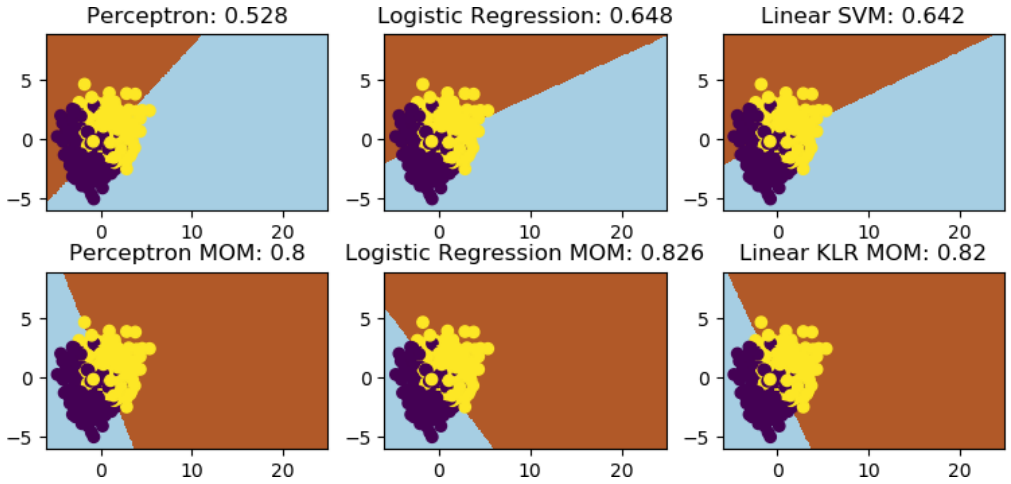


Figure 1.6: Tracé de la frontière de décision calculée pour plusieurs classifieurs tous entraînés sur la base de données de la Figure 2.7.

Par exemple, dans la Figure 2.8 on trace le résultat de l'estimation de plusieurs classifieurs entraînés sur la base de données de la Figure 2.7, le classifieur MOM a été entraîné en utilisant une descente de gradient selon l'Équation (2.6.1) et les classifieurs non-robustes viennent de la librairie scikit-learn. La Figure 2.8 montre que nos méthodes sont robustes alors que les méthodes usuelles ne le sont pas. On obtient le même genre de résultat en régression. Il est plus compliqué d'interpréter des méthodes non-linéaires parce qu'il est compliqué de concevoir ce qu'est un outlier pour une telle méthode et c'est donc difficile à simuler. On peut se rendre compte de ce qu'est un

outlier pour une méthode spécifique comme SVM ou QDA mais en général c'est plus compliqué: qu'est-ce qu'un outlier pour un réseau de neurones ?

Le choix de K peut être complexe, l'intuition nous dit de choisir K un peu plus grand que $|\mathcal{O}|/2$ mais le problème est que $|\mathcal{O}|$ est inconnu et même si il l'était (on estime que dans la plupart des bases de données réelles il y a entre 5% et 10% d'outliers) ce n'est qu'une règle intuitive et la valeur optimale de K dépend en réalité aussi des inliers. Nous pensons donc que la marche à suivre est d'utiliser la validation croisée pour choisir K.

Pour étudier les performances de notre méthode, on regarde une tâche de classification de base de donnée réelle en utilisant $\widehat{f}_{HOME,K}$. La base de données considérée est "Communities and Crime Unnormalized Data Set" que l'on peut télécharger sur UCI Machine Learning Repository. La base de données contient 2215 observations tirées d'un recensement et des données des forces de l'ordre. La tâche que nous nous sommes fixée est de prédire l'activité criminelle (représentée par le nombre d'incidents) en utilisant les caractéristiques suivantes : la population, le salaire par personne,le salaire médian par famille, le nombre de maisons vides et la superficie de l'endroit considéré. Le choix de cette base de données en particulier est motivé par le fait qu'elle contient très certainement une quantité non négligeable d'outliers dus à la nature des données et à l'absence de pré-traitement, ce qui nous permettra d'illustrer l'avantage de nos méthodes.

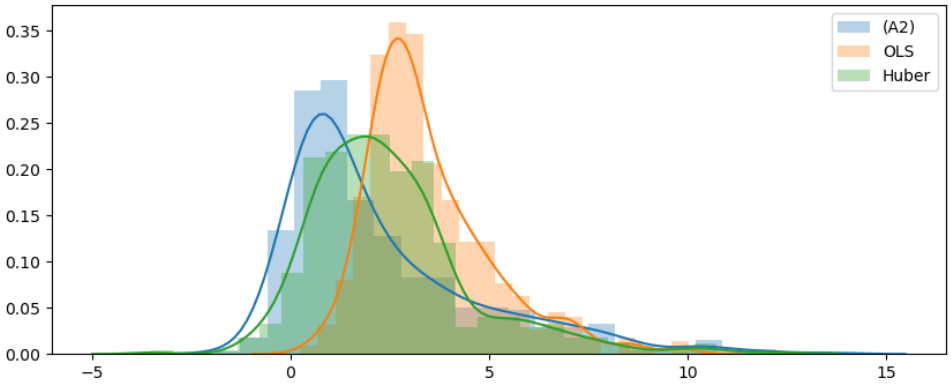


Figure 1.7: Histogrammes des densités du logarithme de la MSE (erreur quadratique moyenne) pour différentes méthodes (bleu clair correspond à notre approche, orange aux moindres carrés ordinaires et vert pour une régression de perte Huber).

Nous comparons le regresseur linéaire $\widehat{f}_{HOME,K}$ (appelé A2 dans la figure) utilisant la perte quadratique, les moindres carrés ordinaires (OLS) et un estimateur qui minimise la perte Huber (HuberRegressor dans scikit-learn, il faut cependant faire attention que HuberRegression utilise la perte Huber de telle façon à être robuste en la variable réponse y mais il n'est pas choisi robuste en le vecteur de caractéristiques x). Les paramètres sont choisis en utilisant une version robuste de la validation croisée utilisant 500 blocs de données. Nous obtenons ainsi une MSE $\simeq e^{4.2}$ pour notre approche, la MSE de OLS est $\simeq e^{22.1}$ et la MSE de l'approche utilisant la perte Huber est $\simeq e^{8.9}$. La densité de la MSE sur les différents blocs est représentée dans la Figure 2.9 (nous avons pris le logarithme du MSE pour rendre la figure plus lisible). Nous voyons ainsi qu'il semble en effet y avoir des outliers dans la base de données comme prévu et que ces outliers ne sont pas seulement en la variable réponse y mais aussi dans le vecteur de caractéristiques x (ce qui

explique que notre approche soit plus efficace que HuberRegressor).

Une version simplifiée de ces algorithmes a aussi été implémenté comme module dans la librairie python scikit-learn-extra que l'on peut trouver à https://scikit-learn-extra.readthedocs.io/en/latest/modules/robust.html. On peut aussi utiliser ces algorithmes pour la détection d'outliers. Voir par exemple l'article [LLM20] pour plus d'information.

1.6.2 MOM algorithmes pour l'estimation du MMD et méthodes à noyaux

On rappelle que notre estimateur robuste du MMD a la formulation suivante :

$$\widehat{MMD}_K(P,Q) = \sup_{f\in\mathcal{B}_\mathcal{K}} \text{Med}\left\{ \frac{1}{b}\sum_{i\in B_k} f(x_i) - \frac{1}{b}\sum_{i\in B_k} f(y_i); \quad 1\leq k\leq K \right\}. \tag{1.6.2}$$

Par le théorème de représentation, la fonction f optimale peut être exprimée par

$$f(a,b) = \sum_{i=1}^n a_i \mathcal{K}(\cdot, x_i) + \sum_{i=1}^n b_i \mathcal{K}(\cdot, y_i), \tag{1.6.3}$$

où $a = (a_i)_{1\leq i\leq n} \in \mathbb{R}^n$ et $b = (b_i)_{1\leq i\leq n} \in \mathbb{R}^n$. On note $c = [a;b] \in \mathbb{R}^{2n}$ la concaténation de deux vecteurs, et $\mathbf{K}$ la matrice de noyau définie par $\mathbf{K} = [\mathbf{K}_{xx}, \mathbf{K}_{xy}; \mathbf{K}_{yx}, \mathbf{K}_{yy}] \in \mathbb{R}^{2n\times 2n}$, $\mathbf{K}_{xx} = [\mathcal{K}(x_i,x_j)]_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$, $\mathbf{K}_{xy} = [\mathcal{K}(x_i,y_j)]_{1\leq i,j\leq n} = \mathbf{K}_{yx}^* \in \mathbb{R}^{n\times n}$, $\mathbf{K}_{yy} = [\mathcal{K}(y_i,y_j)]_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$.

Équation (2.6.5) peut être réécrite,

$$\widehat{MMD}_K(P,Q) = \max_{c\in\mathbb{R}^{2n}: c^T\mathbf{K}c\leq 1} \text{Med}\left\{ \frac{1}{b}[1_k;1_k]^T\mathbf{K}c; \quad 1\leq k\leq K \right\},$$

où 1_k est l'indicatrice du bloc B_k. De façon similaire à la Section 2.6.1 nous utilisons itérativement une étape d'optimisation sur le bloc médian, la différence ici est que l'on n'utilise pas de descente de gradient puisque l'on est en réalité confronté à une optimisation linéaire avec contraintes quadratiques. Il existe donc une solution analytique au problème d'optimisation sur le bloc médian :

$$\underset{c\in\mathbb{R}^{2n}: c^T\mathbf{K}c\leq 1}{\text{argmax}} \frac{1}{b}[1_{k_{\text{Med}}};1_{k_{\text{Med}}}]^T Kc = \frac{[1_{k_{\text{Med}}};1_{k_{\text{Med}}}]}{\|\mathbf{L}^T[1_{k_{\text{Med}}};1_{k_{\text{Med}}}]\|_2}$$

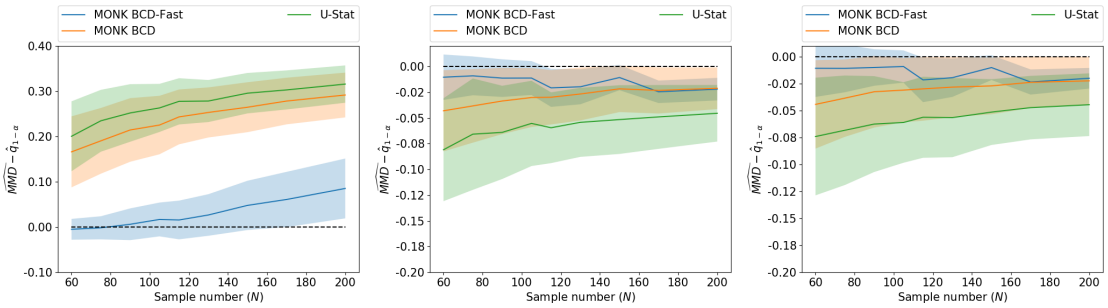
où L est la matrice de décomposition de Cholesky de $\mathbf{K}$ ($\mathbf{K} = \mathbf{L}\mathbf{L}^T$) et k_{Med} est tel que $B_{\text{Med}} = B_{k_{\text{Med}}}$. Les observations sont mélangées à chaque itération. Cet algorithme (appelé MONK BCD pour Median Of meaNs Kernel Block Coordinate Descent) a une complexité $O(n^3)$ ce qui peut être gênant quand la taille de l'échantillon n est grande. Nous proposons donc un autre algorithme appelé MONK BCD-Fast qui approche MONK BCD dans lequel la somme $\sum_{i=1}^n$ dans Équation (2.6.6) après l'avoir insérée dans Équation (2.6.5) est remplacée par $\sum_{i\in B_k}$ (cela correspond à calculer la matrice de noyau seulement comme une matrice par bloc). Nous comparons nos algorithmes à l'état de l'art qui est une approche U-Statistique qui correspond au cas où seulement un bloc est considéré. Quand un seul bloc est considéré, une simplification permet en effet de se réduire à calculer une U-Statistique.

Table 1.1: Complexité du calcul des estimateurs de MMD. n: taille de l'échantillon, K: nombre de blocs, T: nombre d'itérations.

U-Stat	$O\big(n^2\big)$
MMD BCD	$O\big(n^3 + T\big[n^2 + K\log(K)\big]\big)$
MMD BCD-Fast	$O\Big(\frac{n^3}{K^2} + T\Big[\frac{n^2}{K} + K\log(K)\Big]\Big)$

La distance MMD est utilisée par exemple pour les tests de comparaison, l'application numérique que l'on propose est sur une base de donnée biologique. Nous avons choisi une base de données d'ADN tiré du UCI repository, la base de données Molecular Biology (Splice-junction Gene Sequences). La base de données est composée de 3190 échantillons d'une chaîne de 60-caractères qui décrivent une petite partie de l'ADN. Le problème est de reconnaître, à séquence d'ADN donnée, les frontières entre exons (les parties de l'ADN conservées après épissage) et les introns (les parties de l'ADN éliminées par l'épissage). La tâche peut se subdiviser en deux sous-problèmes, identifier les frontières exon/intron (notées EI) et les frontières intron/exon (notées IE). Nous avons pris 1532 échantillons en sélectionnant 766 observations des deux classes EI et IE (La classe des parties d'ADN n'étant ni EI ni IE est plus hétérogène et ont été enlevés pour cette étude), et nous avons étudié le pouvoir de discrimination entre EI et IE. L'ADN est représentée par une suite de caractères, le noyau $\mathcal{K}$ que l'on a choisi est le "String Subsequence Kernel" pour calculer le MMD, et nous avons ensuite utilisé cette estimation du MMD pour faire des tests de comparaison des deux populations en utilisant les trois estimateurs MONK BCD, MONK BCD-Fast et U-Stat.

Les valeurs agrégées de $\widehat{MMD}(EI, IE) - \widehat{q}_{1-\alpha}$, $\widehat{MMD}(EI, EI) - \widehat{q}_{1-\alpha}$ et $\widehat{MMD}(IE, IE) - \widehat{q}_{1-\alpha}$ sont résumées dans Figure 2.13 où $\widehat{q}_{1-\alpha}$ est une estimation du $(1-\alpha)$ quantile via $B = 150$ permutations bootstrap. Dans le cas idéal, $\widehat{MMD} - \widehat{q}_{1-\alpha}$ est positif (resp. négatif) dans le cas inter-classe (resp. intra-classe). Comme le montre la Figure 2.13, les trois techniques sont capables de résoudre la tâche autant pour le cas inter-class (quand l'hypothèse nulle est fausse) que dans le cas intra-class (l'hypothèse nulle est vraie) et ils convergent vers une bonne performance. Un autre avantage majeur de notre méthode, en plus d'être robuste, est la complexité. Le temps de calcul est souvent un point déterminant dans les méthodes à noyaux quand la base de donnée devient grande et le noyau est long à calculer (ce qui est le cas du noyau considéré ici). Dans ce contexte, MONK BCD-Fastt. Par exemple, quand on prend tous les échantillons (n=766) et $K = 15$, calculer MONK BCD-Fast prend $32s$ alors que U-Stat est calculé en $1m28s$ sur le même ordinateur.



(a) Inter-class: EI-IE (b) Intra-class: EI-EI (c) Intra-class: IE-IE

Figure 1.8: Estimateurs du MMD intra-classe et inter-classe comme fonction de la taille de l'échantillon comparé au $(1-\alpha)$-quantile bootstrap. Nous avons tracé moyenne±std. Remarquez que l'ordre de grandeur est différent dans le cas intra-classe et dans le cas inter-classe.

Chapter 2

Introduction in english

Statisticians have observed that real datasets contain anomalous data (also called outliers) and those can have a serious impact on the results of a statistical analysis. Anomalous data are data difficult to model or data that we should ignore when doing inference. Hence the need for robust methods that can deal with these anomalous data. This problem is very old, Newcomb in 1882 [Gut01] and Laplace before him were already aware of outliers and their effects on data analysis. Fisher also highlighted the problem of outliers when one wants to do statistical inference [FR22].

This dissertation contains five original articles: the article [LLM20] which has been published in the Machine Learning Journal, the article [LSML19] which has been published as proceeding of 2019 ICML conference, the article [MM19] which has been accepted for publication in Information and Inference: A Journal of the IMA and the two articles [Mat20b, Mat20a] which will soon be submitted for publication. In these articles, we investigate robust statistics with a particular accent on robust machine learning and empirical processes. I also present at the end of the introduction the work I did to implement robust machine learning algorithms in the python library scikit-learn-extra.

The dissertation is composed of an introduction to the subject and to the contributions of this thesis. Then, the articles written during the phD are presented one after the other.

In the first part of the introduction, I will present some basic facts about robust statistics and the formulation of robust statistics in mathematical terms. Then, I present the state of the art in robust mean estimation, the problem of robust mean estimation is central in this dissertation and these results were the basis of this PhD. The second part of the introduction contains our contributions to robust statistics. First are non-asymptotic deviation bounds in the problem of estimating the mean using results from Chapter 4 and Chapter 3, then we look at asymptotic results such as consistency in a corrupted setting in the same problem using results from Chapter 3. Then, I explain our contributions to robust machine learning and in particular classification, regression and kernel methods using results from Chapter 5, Chapter 6 and Chapter 7. Finally, I present the algorithms and practical work I did to implement robust Machine Learning algorithms.

For a more interactive reading, the reader is encouraged to run the notebook prepared by the author at https://colab.research.google.com/drive/1yyGCgmif1EXBNLBgM0DaZvPLyHuJW8zf?usp=sharing. In this notebook, we show illustrations of robust mean estimation and robust Machine Learning algorithms.

2.1 From Huber's view of robustness to sub-Gaussian estimators

2.1.1 Corrupted and Heavy-tailed distributions

To represent real-life data, statisticians use models or assume simplifying hypotheses. These hypotheses are often considered as the ideal case, and a major question is how to handle deviations from this ideal case. Informally, an estimator is said to be robust if a small change in the hypothesis does not change the estimation by too much. To be more precise, one has to define what deviations from the hypothesis are considered and what we mean by "does not change by too much".

For example, suppose we have access to n data $X_1, \dots, X_n$. Suppose that this sample is corrupted: $X_1, \dots, X_{n-1}$ are i.i.d from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ and X_n is equal to a constant $M \gg \mu$ (X_n is called an outlier). Then, the empirical mean becomes $\frac{1}{n}\sum_{i=1}^{n-1} X_i + \frac{M}{n}$. The empirical mean can be arbitrarily far from the mean of the inliers μ of the non-outliers points because M can be very large compared to $\sum_{i=1}^{n-1} X_i$. The empirical mean is thus non-robust and using robust statistics, we aim at finding robust estimators that do not have this problem, see Figure 2.1.

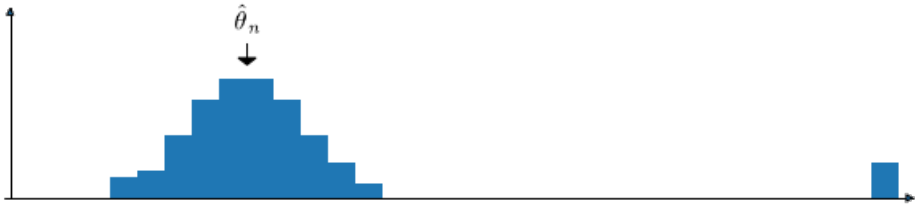


Figure 2.1: Histogram of a dataset with outliers. In this figure $\widehat{\theta}_n$ is some robust estimator of the mean.

In practice, situations where corruption arises are not rare. For instance, one may think of quality control, where one owns several machines. The majority of the machines work fine with potential errors that follow a centered Gaussian law with small variance, yet one of the machine is defective and has an error that is some order of magnitudes more variable (i.e. heavy-tailed) than the normal machines, this means that outliers will come out of the defective machine. Another possibility is for the defective machine to have an error that is Gaussian but not centered in 0 (systematic error). In this case, the objects produced by this machine are not functional and cannot be used. We may want to detect which machine is defective to repair it.

Let us consider another example. In clinical checkups, respiratory exercises would give results that would be very different if the patient is a smoker or if he is not. Whether the patient smokes or not may not be accessible to the clinician and the statistician: typically teenagers may not say that they are smoking even to their physician. Human errors or captor errors are another source of corruption and outliers arise also naturally in a lot of datasets, for instance in biological datasets or demographic datasets (see Chapter 7 and Chapter 6). Depending on the corruption and the task at hand, robust methods may be needed.

In this dissertation, we consider three different settings. The first one setting is heavy-tailed distributions, the second one is a dataset corrupted by outliers and the third is what we call a corruption neighborhood.

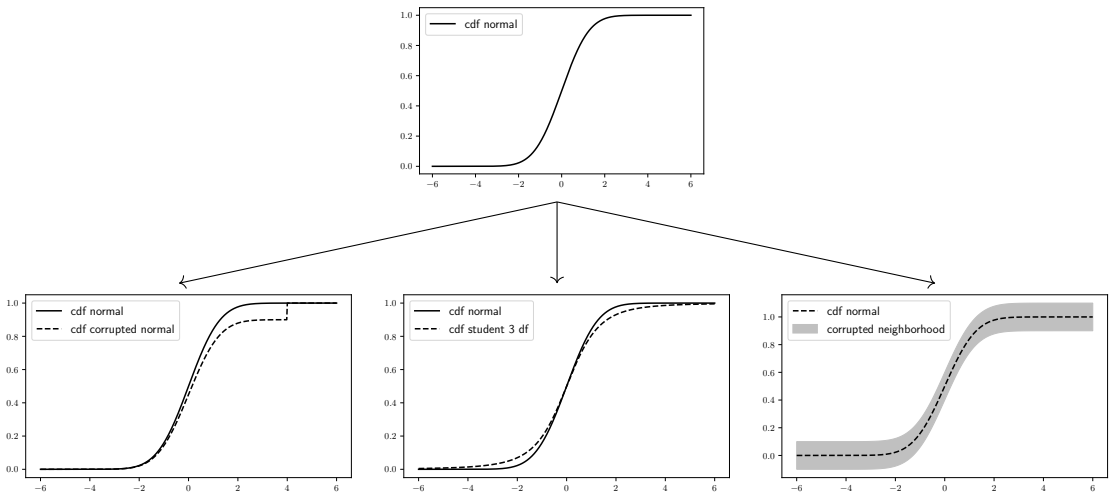


Figure 2.2: Plot of the c.d.f for different corruption settings. The plot on top corresponds to the un-corrupted case (standard Gaussian), the bottom-left plot is a case of Huber contamination model, the bottom-middle plot is a case of Heavy-tailed distribution and the bottom-right plot is a case of corrupted neighborhood where a corrupted c.d.f has to be comprised in the shaded area.

Heavy-tailed distribution. Classical non-asymptotic results in statistics often hold for i.i.d data $X_1, \dots, X_n$ following a Gaussian distribution or a distribution with light tail. Typically, to derive concentration results for Machine Learning applications, we will need a sub-Gaussian behavior of the random variables that we study but in reality, there are a lot of datasets that can't be modeled by light-tailed densities. Hence, a deviation from the assumptions that we want to consider is the case of heavy-tailed distributions where one does not suppose that $X_1, \dots, X_n$ follow a Gaussian or sub-Gaussian distribution but only that X has a finite number of finite moments. Typically we only suppose two finite moments. This is the setting in [Cat12, DLLO16].

Dataset corrupted by outliers, $\mathcal{I} \cup \mathcal{O}$ framework. Let $X_1, \dots, X_n$ be random variables. Let $\mathcal{I}, \mathcal{O}$ be a partition of $\{1, \dots, n\}$ into two sets, the set $\mathcal{I}$ of inliers and the set $\mathcal{O}$ of outliers (with $|\mathcal{O}|$ small compared to $|\mathcal{I}|$). We make some assumptions on $(X_i)_{i \in \mathcal{I}}$, typically finite moments assumptions whereas we don't assume anything on $(X_i)_{i \in \mathcal{O}}$. The sets $\mathcal{I}$ and $\mathcal{O}$ are unknown to the statistician. This is a generalization of the Heavy-tailed distribution setting and the

data are not supposed i.i.d. This is sometimes called adversarial corruption, this is the setting in [LL20, LSML19].

Dataset corrupted by outliers, Huber contamination framework. The second corruption scenario is the so-called Huber contamination where $X_1, \dots, X_n$ are i.i.d from a mixture of distribution $(1-\varepsilon)P + \varepsilon H$ with ε small. We make assumptions on P but not on H and H plays the role of outlier distribution. This setting is very close to the $\mathcal{I} \cup \mathcal{O}$ framework presented above except that in Huber contamination neighborhood the outliers can't depend on the inliers and the data are i.i.d, this could be said to be a non-adversarial corruption setting. This setting comes from [Hub64, HR09, ZJS19] and has been used for instance in [CGR+18].

Corruption neighborhood. As said previously we can define deviations from the usual hypothesis saying that $X_1, \dots, X_n$ are i.i.d from a distribution that is not far from some model distribution P, an example is Huber's contamination neighborhood. More generally, let d be a distance between probability distributions (i.e. Kolmogorov distance, Total variation distance, Wasserstein distance...) and suppose that $X_1, \dots, X_n$ come from a distribution Q such that $d(P, Q) \leq \varepsilon$. This is a generalization of Huber contamination framework as it can be seen that if $Q = (1-\varepsilon)P + \varepsilon H$ is a corrupted version of P, then the total variation distance between P and Q is smaller than ε. However the framework is more general than the Huber contamination neighborhood and difference distances between distribution will greatly change the type of outliers or deviation from P. This setting has been used in [HR09, Ham71].

In the three cases of deviation from usual assumptions, the goal is to use methods that give results that are not very different from what we would observe in some ideal cases (i.i.d Gaussian setting, uncorrupted setting,...). This is the informal definition of robustness that we consider here and each time we state a result we will first explain which deviation from the usual hypothesis we are considering.

A reasonable question is: why can't we use an outlier removal scheme and deal with the "cleaned dataset" as though there were no outliers? This is the old problem of robustness vs diagnostics, see [Hub91] for further information. The first and maybe most compelling reason is that robust algorithms work better (see Figure 2 in the introduction of [HRRS86]). One of the reasons for this is that when we perform outlier removal, we don't recover the inlier distribution but instead we recover a trimmed version of it. For example, in the case of a mixture $H(x) = (1-\varepsilon)\Phi(x) + \varepsilon\delta_M$ where δ_M is a Dirac distribution in M, we will not recover a standard Gaussian but instead we will have a trimmed Gaussian. Another reason to construct robust statistics is that we need robust algorithms in order to do outlier detection algorithms: if we want to detect outliers, we can't be influenced by these outliers. Yet another way of dealing with outliers is transformation. We often use the logarithm to transform a skewed distribution or a Box-Cox transformation to handle non-Gaussian random variables, but it has been shown that this does not necessarily solve outlier problems (see for example in the context of testing [Ras89, DW83]). Another problem with transformation is that there is sometimes no way to go back to the un-transformed variable and this is a problem if one wants to have an interpretable statistical procedure.

We may identify two streams of research in theoretical robust statistics: an asymptotic and a non-asymptotic one. At the beginning of robust statistics, there has been a large amount of work on the asymptotic properties of robust estimators and typical theorems were that an estimator is asymptotically normal with an optimal asymptotic variance [Hub64]. On the other hand, there were very little non-asymptotic results at the time. Among the non-asymptotic results, one can cite the large Princeton Monte-Carlo study [AH15] that compared 68 robust estimators

using Monte-Carlo techniques or techniques such as small-sample asymptotics [FRR90] or the breakdown point [DH83, Hub84] that we recall below in Section 2.1.3. In this thesis, we give both asymptotic and non-asymptotic results with an emphasis on non-asymptotic results such as convergence notions that hold with high probability.

2.1.2 Robustness and concentration inequalities with heavy-tailed distributions

In recent years, robustness theory has known a rebirth with the motivation of robust machine learning algorithms. In machine learning applications such as regression and classification using linear models, one can easily see on examples (Figure 2.3) that the usual techniques such as least-squares estimators, SVM or logistic regression are not robust to outliers.

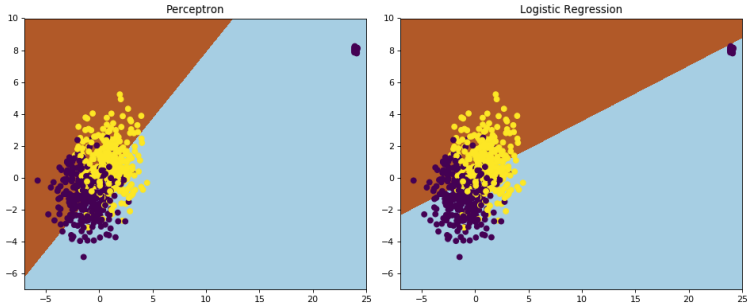


Figure 2.3: Plot of the separation line of a classifier (perceptron on the left and logistic regression on the right) trained on the dataset represented by the scatter plot present on the same figure. There is a group of 30 outliers among the 300 dataset points, the outliers are in the top right corner and they mislead the classifiers.

Formally, the problem is often stated as follows. Let $\mathcal{X}$ and $\mathcal{Y}$ be two sets, typically $\mathcal{X}$ is a subset of $\mathbb{R}^d$ and $\mathcal{Y}$ a subset of $\mathbb{R}$. Let $\mathcal{F}$ be a set of functions $f : \mathcal{X} \to \mathcal{Y}$. Our task is to find f such that $f(X)$ is a good approximation of Y, and we quantify this using the risk $R(f)$ of a function f which is defined by

$$R(f) = \mathbb{E}[\ell(f(X), Y)],$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a so-called loss function. For instance, in regression $\ell(f(x), y) = (f(x) - y)^2$ and the risk $R(f)$ is then the mean squared error. Then, when we say that we want f to be such that $f(X)$ is a good approximation of Y, we mean that f must minimize $R(f)$ over $f \in \mathcal{F}$. The class $\mathcal{F}$ can be a family of linear functions in the case of linear regression for example, or the set of all possible neural networks in deep learning. To simplify, we will suppose that there exists f^* in $\mathcal{F}$ that minimizes the risk among all the functions in $\mathcal{F}$.

$$f^* \in \operatorname*{argmin}_{f \in \mathcal{F}} R(f). \tag{2.1.1}$$

Let $\widehat{f}$ be an estimator of f^* based on $(X_1, Y_1), \ldots, (X_n, Y_n)$, i.i.d copies of (X, Y). Results to assess the efficiency of $\widehat{f}$ come mainly under the form of an oracle inequality which which is an

upper bound on the excess risk $R(\widehat{f}) - R(f^*)$. Remark that the expectation in the definition of the risk being with respect to the couple (X, Y), $R(\widehat{f}) - R(f^*)$ is in fact a random variable, its randomness coming from the data $(X_1, Y_1), \dots, (X_n, Y_n)$, hence bounding $R(\widehat{f}) - R(f^*)$ from above means finding a $\Delta_{n,\delta}(\mathcal{F})$ such that for all $\delta \in (0, 1)$,

$$\mathbb{P}\Big(R(\widehat{f}) - R(f^*) \leq \Delta_{n,\delta}\Big) \geq 1 - \delta.$$

To obtain such a result, the principal building blocks are concentration inequalities on the empirical mean used to bound the probability that an estimator deviates from the parameter it wants to estimate by more than some value $t > 0$. Typically, to bound the excess risk, we will first control the empirical risk defined by

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

because the construction of $\widehat{f}$ often gives us information on $\widehat{R}(\widehat{f})$. This is where we may have an issue because concentration inequalities on sums of i.i.d random variables generally rely on strong concentration properties of the data, typically we need $\ell(f(X), Y)$ to be sub-Gaussian or sub-Exponential (see [BLM13]) and in general concentration inequalities are not valid in any of the settings presented in Section 2.1.1. Our goal is to find estimators with good concentration properties even though the data are heavy-tailed. Important articles developing this line of thought are [Cat12] and [DLLO16].

One of the classical problems is to find a robust estimate of the mean of some random variable X (this can be set in a risk minimization framework by taking ℓ to be the squared loss and $\mathcal{F}$ the set of constant functions). We want the estimator to exhibit the same concentration as the empirical mean would on a Gaussian sample even when the data are not Gaussian. Informally, if $\widehat{\mu}$ is a robust estimator of the mean based on a sample $X_1, \dots, X_n$ of i.i.d data with finite second moment, let $W_1, \dots, W_n$ be i.i.d $\mathcal{N}(\mathbb{E}[X], Var(X))$, we ask that there exists $C > 0$ such that for all $\delta > 0$

$$\mathbb{P}(|\widehat{\mu} - \mathbb{E}[X]| > \delta) \leq C\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} W_i - \mathbb{E}[X]\right| > \delta\right). \tag{2.1.2}$$

In this dissertation, when we want to prove the efficiency of an estimator we will always try to prove a deviation inequality similar to Equation (2.1.2) that shows that our estimator is as good on a heavy-tailed dataset as the standard method would be if data were light-tailed and uncorrupted.

2.1.3 Infinitesimal aspect of robustness and breakdown point

There is no universally accepted definition of robustness, but one of the early and most widely accepted attempts to do so was by Hampel in his seminal paper of 1971 [Ham71]. Let T_n be a sequence of estimators using a sample $X_1, \dots, X_n$. Hampel's definition of robustness says that T_n is robust at a c.d.f F if and only if for all $\varepsilon > 0$, there exists $\delta > 0$ such that for any c.d.f G and any $n \in \mathbb{N}$ we have

$$d(F, G) < \delta \Rightarrow d(\mathcal{L}_F(T_n), \mathcal{L}_G(T_n)) < \varepsilon,$$

where $\mathcal{L}_F(T_n)$ (resp $\mathcal{L}_G(T_n)$) is the probability distribution of T_n when $X_1, \dots, X_n$ are i.i.d distributed according to the c.d.f F (resp G). In the case where T_n is in fact a function of the empirical c.d.f, $T_n = T(\widehat{F}_n)$, this definition can be seen as an equi-continuity of T, this is one possible mathematical translation of our robustness paradigm "small changes in the hypothesis cause only small changes in the result of the estimation". The choice of the distance d is very important in defining the corruption to which we want to be robust to, this is linked to the corruption neighborhood defined in Section 2.1.1. If $d(F,G) = \sup_x |F(x) - G(x)|$ is the Kolmogorov distance, we can show that the empirical median is robust while the empirical mean is not robust. The choice of the distance is very important to define what is a corruption and what is a robust estimator.

This definition of robustness is only qualitative and in order to quantify robustness, Hampel used the tool of the influence function. The influence function is defined as the Gateaux derivative of T in the direction of the Dirac distribution (also called the Von-Mises derivative, see [Fer83]). Let $D_x(t) = \mathbb{1}\{t \geq x\}$ be the c.d.f of a Dirac distribution in some $x \in \mathbb{R}$, the influence function is defined for all $x \in \mathbb{R}$ by,

$$IF(x, T, F) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon D_x) - T(F)}{\varepsilon}. \tag{2.1.3}$$

Hampel [Ham71] proposes to quantify the robustness of T in F using $\sup_x |IF(x, T, F)|$. One reason for this choice is a functional Taylor expansion that says (provided that T is smooth enough) for all F, G c.d.f,

$$T(F) = T(G) + \int IF(x, T, F) d(G - F)(x) + R(F, G)$$

where the term $R(F,G)$ can be shown to be negligible compared to the other terms, under some assumptions on T, F and G. Then, if $\sup_x |IF(x, T, F)| < \infty$, we can show that $|T(F) - T(G)|$ is small as long as $d(F,G)$ is small (for d the Kolmogorov distance). This motivates Hampel's definition of B-robustness [HRRS86] saying that an estimator is B-robust if its influence function is bounded. More generally, as its name indicates the influence function measures the influence that a data point placed in x has on the value of T and we want this influence to be bounded.

All these definitions of robustness are infinitesimal and this does not give us information on how many outliers the estimator can handle. A maybe more practical measure of robustness is the breakdown point introduced in [DH83]. For an estimator $T(X_1, \dots, X_n)$ invariant by permutation of $X_1, \dots, X_n$, we define the breakdown point by

$$\varepsilon_n^* = \min\left\{ \frac{m}{n}, m \in \{1, \dots, n\} : \sup_{X_1', \dots, X_m'} |T(X_1, \dots, X_n) - T(X_1', \dots, X_m', X_{m+1}, \dots, X_n\}| = \infty \right\}.$$

For example, one can compute that for the empirical mean, the breakdown point is $\varepsilon_n^* = 1/n$ while for the empirical median the breakdown point is $\varepsilon_n^* = \lfloor \frac{n}{2} \rfloor \frac{1}{n}$: we need only one outlier to make the empirical mean arbitrarily large whereas for the empirical median we need to corrupt more than half of the data if we want to affect the result arbitrarily . The breakdown point is the proportion of outliers that an estimator can handle before "breaking down" in the sense of taking arbitrary large values. This constitutes another measure of robustness of an estimator.

We will see in what follows that we can use other measures of robustness for an estimator but it will often be only variations of the influence function or the breakdown point.

2.2 State of the art in robust estimation of the mean

2.2.1 M-estimators and influence function

In this section, we are interested in the problem of estimating a location parameter meant to exhibit a central tendency of the data. Let $X \sim P$ for some P probability on $\mathbb{R}^d$, let ρ be an increasing function from $\mathbb{R}_+$ to $\mathbb{R}_+$, let $T(P)$ be defined by the following optimization problem

$$T(P) \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \mathbb{E}[\rho(\|X - \theta\|)], \tag{2.2.1}$$

where $\|\cdot\|$ is the euclidean norm. Alternatively, if ρ is smooth enough (which will be the case here), we define $T(P)$ by

$$\mathbb{E}\left[\frac{X - T(P)}{\|X - T(P)\|}\psi(\|X - T(P)\|)\right] = 0, \tag{2.2.2}$$

where $\psi = \rho'$ is called the score function. For $\psi(x) = x$, we recover the mean $T(P) = \mathbb{E}[X]$ and for $\psi(x) = 1$, we recover the median. If ψ is bounded, $T(P)$ can be seen as smoothed geometric median estimators, see [Min15, CG17].

The empirical estimator obtained by plugging the empirical density $\widehat{P}_n$ in equation (2.2.2) is called M-estimator associated with ψ, it is denoted $T(\widehat{P}_n)$ and computed from an i.i.d sample $X_1, \dots, X_n$ using the following equation:

$$\sum_{i=1}^{n} \frac{X_i - T(\widehat{P}_n)}{\|X_i - T(\widehat{P}_n)\|}\psi(\|X_i - T(\widehat{P}_n)\|) = 0. \tag{2.2.3}$$

For $\psi(x) = x$ we obtain $T(\widehat{P}_n) = \frac{1}{n}\sum_{i=1}^n X_i$ and for $\psi(x) = 1$ $T(\widehat{P}_n)$ is the geometrical median. A careful choice of the function ψ yields estimators that are more robust to outliers and heavy-tailed data than the empirical mean and more efficient than the median. M-estimators are especially nice to work with because their influence function (introduced in equation (2.1.3)) takes a very simple form:

$$\mathrm{IF}(x, T, P) = M_{P,T}^{-1}\frac{x - T(P)}{\|x - T(P)\|}\psi(\|x - T(P)\|), \tag{2.2.4}$$

where $M_{P,T}$ is a nonsingular matrix which does not depend on x (an explicit formula for $M_{P,T}$ exists and may be found for instance in [HRRS86, Eq 4.2.9, Section 4.2C.] however we will not use it for our study). In particular, we will study three different functions ψ (see in Figure 2.4):

Huber's score and estimator. Let $\beta > 0$. For all $x \geq 0$, let

$$\psi_H(x) = x\,\mathbb{1}\{x \leq \beta\} + \beta\,\mathbb{1}\{x > \beta\}. \tag{2.2.5}$$

In dimension 1, the M-estimator constructed from this score function is called the Huber's estimator [Hub64]. Call T_H the associated functional.

Catoni's score and estimator. Let $\beta > 0$. For all $x \geq 0$, let

$$\psi_C(x) = \beta \log\left(1 + \frac{x}{\beta} + \frac{1}{2}\left(\frac{x}{\beta}\right)^2\right). \tag{2.2.6}$$

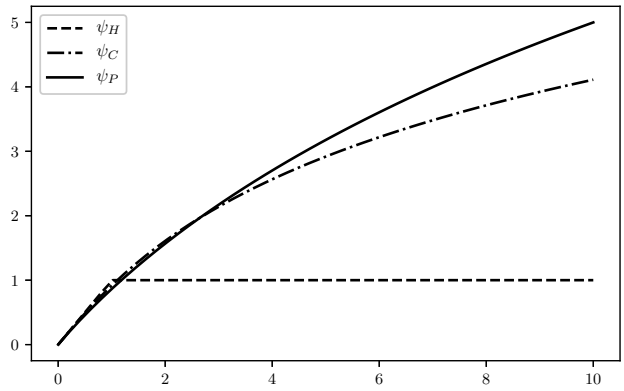


Figure 2.4: Plot of ψ_H and ψ_C for $\beta = 1$. ψ_P is plotted for $\beta = 10$ and $p = 5$.

The associated M-estimator in dimension 1 is one of the estimators considered by Catoni in [Cat12]. Call T_C the associated functional.

Polynomial score and estimator. Let $p \in \mathbb{N}^*$, $\beta > 0$. For all $x \geq 0$, let

$$\psi_P(x) = \frac{x}{1 + \left(\frac{x}{\beta}\right)^{1-1/p}}. \tag{2.2.7}$$

Call T_P the associated functional.

The breakdown point for Huber's estimator tends to $1/2$ in dimension 1 whereas the breakdown point for both Catoni and Polynomial estimators go to 0. We will see that Catoni's and Polynomial estimators will be robust in a weaker sense than Hampel's definition of robustness. The behavior of $\psi(x)$ when x goes to infinity dictates the robustness of the estimator $T(\widehat{P}_n)$ (see Theorem 10 and Corollary 2) while on the other hand the behavior of ψ near 0 will control the distance of the location parameter $T(P)$ to the mean $\|T(P) - \mathbb{E}[X]\|$ when P is a skewed distribution (see Lemma 2).

2.2.2 Median of means estimators

Let $X_1, \ldots, X_n$ be i.i.d from a distribution with finite second moment, let $K \in \mathbb{N}$ and suppose that K divides n. Let $B_1, \ldots, B_K$ be a partition of $\{1, \ldots, n\}$ and $b \in \mathbb{N}^*$ be such that

$$\forall k \in \{1, \ldots, K\}, \quad |B_k| = b, \quad \forall k \neq j, \quad B_k \cap B_j = \emptyset \quad \text{and} \quad \cup_{k=1}^{K} B_k = \{1, \ldots, n\}$$

For all $B \subset \{1, \ldots, n\}$, define the empirical mean over block B by

$$P_B(X_1^n) = \frac{1}{b} \sum_{i \in B} X_i,$$

the median of mean estimator, which dates back to [NY83, AMS99, JGV86], is defined as

$$\mathrm{MOM}_K(X_1^n) = \mathrm{Med}(P_{B_k}(X_1^n), 1 \leq k \leq K). \tag{2.2.8}$$

The median of means estimator interpolates between the empirical mean and the empirical

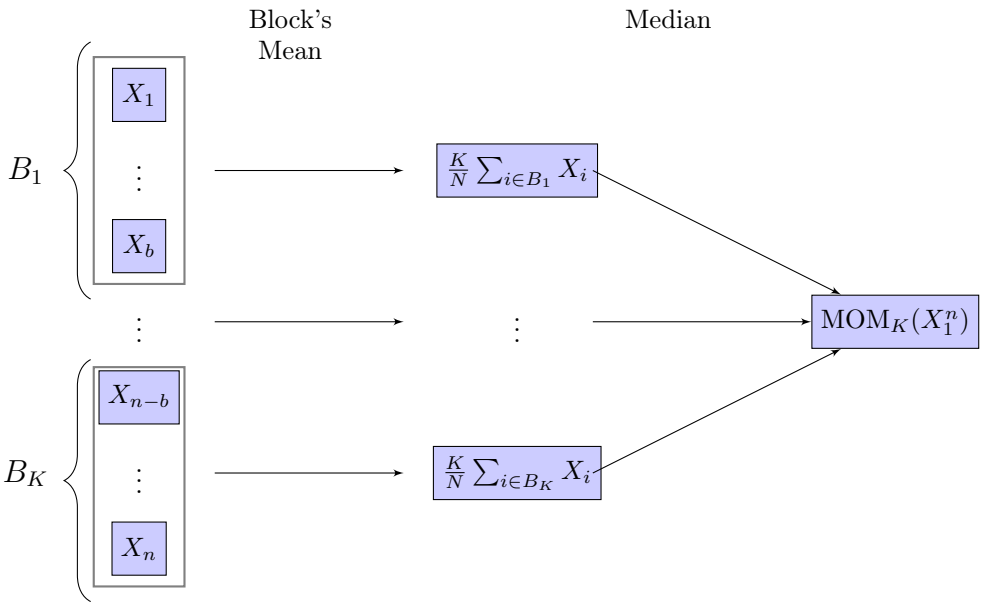


Figure 2.5: Construction of the median of means.

median with parameter K that indicates how robust the estimation is, since less than $K/2$ outliers may corrupt at most $K/2$ blocks, leaving the median in Equation (2.2.8) equal to an empirical mean of a block with uncorrupted data. Remark that if K does not divide n, in practice we can use blocks of different sizes but it is easier for the theory to consider blocks with the same size. The breakdown point of the median of means is $\frac{1}{n}\lceil K/2 \rceil$. For more information on MOM's deviation bounds, one can see [DLLO16, MS17, LCB19, LSC20, Min18], see [Min20] for asymptotic results on median of means estimators. The median of means has also been adapted to other setting, see for instance [BAM20] for a differentially private median of means estimator or Chapters 5,6,7 for its use in Machine Learning.

The median of means principle is that we begin by using classical methods (such as empirical mean, or ordinary least squares when in regression) on blocks of data, and then we aggregate the results of the blocks using a robust estimator. This principle can be used to solve numerous problems in a robust fashion. The median of means can be generalized, for instance we can use a Huber estimator or any M-estimator instead of the median, as it is done in Chapter 6 this makes the resulting estimator more stable to the choice of the number of blocks K and this is often more efficient than median of means. We can use median of means principle for kernel method estimators as in Chapter 7 to obtain robust kernel method estimators, and finally the median of means principle can also be used in learning algorithms as in Chapters 5 and 6 to obtain robust machine learning algorithms.

2.3 Contributions to the study of deviations of robust estimators

2.3.1 Concentration of robust estimators of the mean in dimension 1

In dimension 1, for a Gaussian sample, it is well known [BLM13] that the empirical mean has the following deviations. Let $X_1, \dots, X_n$ be i.i.d with law $\mathcal{N}(\mu, \sigma^2)$, then for all $t > 0$,

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \sigma \sqrt{\frac{t}{2n}} \right) \leq e^{-t}. \tag{2.3.1}$$

In this section we show that the median of means and M-estimators achieve similar deviation bounds even when $X_1, \dots, X_n$ are not drawn from Gaussian distribution.

To construct concentration inequality in a robust context, we will always show that the concentration of the estimator can be approximated by the concentration of some empirical mean of transformed variables which are easier to control. In dimension 1, we have the following known result for Median of Means and our contribution for M-estimators.

Median of Means. By interpreting what it means for a median to be larger than some constant, we get

$$\mathbb{P}(|\mathrm{MOM}_K(X_1^n) - \mathbb{E}[X]| > \varepsilon) \leq \mathbb{P}\left(\sum_{k=1}^{K} \mathbb{1}\{|P_{B_k}(X_1^n) - \mathbb{E}[X]| > \varepsilon\} \geq \frac{K}{2} \right).$$

We changed the problem into an easier problem where we can use usual i.i.d concentration inequalities. For example via Hoeffding's inequality, we obtain the following theorem.

Theorem 9 (Deviation Median of Means)**.** *Let $X_1, \dots, X_n, X$ be i.i.d real-valued random variables, with finite variance σ^2. Then, for all $K \in \{1, \dots, n\}$,*

$$\mathbb{P}\left(|\mathrm{MOM}_K(X_1^n) - \mathbb{E}[X]| > 2\sigma \sqrt{\frac{K}{n}} \right) \leq e^{-K/8} \tag{2.3.2}$$

This deviation bound can be compared to the Gaussian case from equation (2.3.1) but there are notable differences. The estimator of the mean depends on K the number of blocks, and this number of blocks also intervene in the deviation bound: we don't use the same estimator for all the confidence levels. Another difference with equation (2.3.1) is that we don't have a deviation for all level $t > 0$, i.e. the right-hand-side of equation (2.3.2) cannot be arbitrarily small we can only go until probabilities of order e^{-n}. It is in fact unavoidable, it has been shown in [DLLO16] that we can't have a sub-Gaussian concentration around the mean for all confidence levels at the same time.

Then, the median of means is suitable to estimate the mean even when the data are heavy-tailed (finite second moment). We can also show deviations very similar in a $\mathcal{I} \cup \mathcal{O}$ setting (see [Ler19]).

M-estimator. In the case of M-estimator, in Chapter 3 we show the following theorem that controls the deviation of a M-estimator.

Theorem 10 ([Mat20b]). *Let $X_1, \dots, X_n, X$ be i.i.d real-valued random variables with law P, let ψ be one of the three score functions defined in Section 2.2.1 and suppose that $T(P)$ and $T(\widehat{P}_n)$ exist and are unique. Define $\psi_{odd}(x) = \text{sign}(x)\psi(|x|)$, then the following holds.*

- *For all $\lambda > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\psi_{odd}(X_i - T(P))\right| > 3\lambda\right) \leq \mathbb{P}\Big(\Big|T(\widehat{P}_n) - T(P)\Big| > \lambda\Big).$$

- *If moreover $V = \mathbb{E}[\psi(|X - T(P)|)^2] \leq \psi(\beta/2)^2/2 < \infty$, then for all $\lambda \in (0, \beta/2)$,*

$$\mathbb{P}\Big(\Big|T(\widehat{P}_n) - T(P)\Big| > \lambda\Big) \leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\psi_{odd}(X_i - T(P))\right| > \frac{\lambda\gamma}{4}\right) + e^{-n\gamma^2/8}. \tag{2.3.3}$$

  where $\gamma = 1$ if $\psi = \psi_h$, $\gamma = 4/5$ if $\psi = \psi_C$ and $\gamma = 1/4$ if $\psi = \psi_P$.

Equation (2.3.3) shows that the deviations of $T(\widehat{P}_n)$ are controlled through the deviations of $\psi_{odd}(X - T(P))$ and that the variance parameter is $V = \mathbb{E}[\psi(|X - T(P)|)^2]$. Having that ψ is concave on $\mathbb{R}_+$ and $\psi(0) = 0$, we obtain that $\psi_{odd}(X - T(P))$ has a lighter tail than X and for instance if ψ is bounded we have that $\psi_{odd}(X - T(P))$ is sub-Gaussian. This makes it very easy for us to control the deviations of $T(\widehat{P}_n)$ because a sum of i.i.d light-tailed random variables can be handled through classical concentration inequalities, we give an example in Theorem 11 for Huber's estimator.

We also show the following lemma in Chapter 4 that controls the distance of $T(P)$ to the mean (the Lemma in Chapter 4 holds for more general score functions ψ than just the three presented here).

Lemma 2. *Suppose that ψ is C^k with bounded k^{th} derivative, $\psi'(0) = 1$ and for $2 \leq j \leq k-1$, $\psi^{(j)}(0) = 0$. Let X be a random variable such that $\mathbb{E}[\|X\|^k] < \infty$, then,*

$$\|\mathbb{E}[X] - T(P)\| \leq \frac{2\|\psi^{(k)}\|_\infty}{\gamma k! \beta^{k-1}}\mathbb{E}\big[\|X - T(P)\|^k\big] \tag{2.3.4}$$

where $\gamma = 1$ if $\psi = \psi_h$, $\gamma = 4/5$ if $\psi = \psi_C$ and $\gamma = 1/4$ if $\psi = \psi_P$.

Moreover, we can also show that for Huber score function, $\|T_H(P) - \mathbb{E}[X]\|$ is of order $O(1/\beta^{q-1})$ where q is the number of finite moments of P. We can show that this bound is tight in its dependency on β as soon as the distribution is asymmetric (in the symmetric case, $\|T_H(P) - \mathbb{E}[X]\| = 0$ and there is no need for such a bound).

This separation of the effect of the deviation and the effect of the distance $\|T(P) - \mathbb{E}[X]\|$ is similar to a bias-variance trade-off and it allows some in-depth analysis of M-estimators (see Chapter 4), to our knowledge this separation in variance term and bias term was not known before this result.

For example in the case of Huber's estimator (whose score function is defined for $x \geq 0$ by $\psi_H(x) = x \wedge \beta$) we are again reduced to controlling the concentration of a sum of i.i.d bounded random variable. The following theorem for Huber's estimator is proved in Chapter 4 and is a consequence of Theorem 10 and Lemma 2.

Theorem 11. *Let X be a real-valued random variable with $\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] < \infty$. For all $t \in (0, n/16)$, with probability greater than $1 - 2e^{-t} - e^{-n/8}$,*

$$\left|T_H(\widehat{P}_n) - \mathbb{E}[X]\right| \leq 8\sigma\sqrt{\frac{2t}{n}}. \tag{2.3.5}$$

Equation (2.3.5) shows that we recover the rates of Equation (2.3.1) but in a corrupted setting with limitations that are similar to what we obtained with MOM on the level: the probability with which Equation (2.3.5) holds with probability that can't be larger than the order $1 - e^{-n}$.

A maybe surprising corollary of Theorem 10 is that when the distribution is symmetric, we don't need to use Lemma 2 and a direct consequence of Theorem 10 when $\psi = \psi_H$ is bounded is that the estimator $T_H(\widehat{P}_n)$ has a rate of convergence to the expectation of order $O(1/\sqrt{n})$. This is surprising because we know (from [DLLO16]) that the rate of convergence of an estimator of the mean towards $\mathbb{E}[X]$ cannot be faster than $\Omega(1/n^{\delta/(1+\delta)})$ in general when the distribution has a finite moment $1 + \delta$ for some $\delta \in [0, 1]$. It shows that when we restrict ourselves to symmetric distribution, we can achieve a faster rate of convergence than in the general case.

In these theorems, we didn't specify the parameters β for M-estimators or K for Median of Means. One way of choosing the parameter is via cross-validation which we briefly describe here. In a learning framework, our task is to minimize a risk as in equation (2.1.1) and we have available an estimator of the risk that we can compute. When doing cross-validation, we try to estimate the generalization risk by training our algorithm on one part of the data, and computing the estimated risk on the second part of the data. This gives us an estimate of the generalization risk of our model and we want to find the hyper-parameters (in our case, β or K) that minimizes the estimated generalization risk. This technique is used in Chapter 5, Chapter 6 and Chapter 7.

If cross-validation is not available (if we are not in a learning framework), we can use Lepski's method as it is done for example in [Loh18] to estimate a parameter of scale in Huber loss regression or in [DLLO16] with median of means estimators, this method allows us to choose parameters but at the cost of a rather high computational cost which makes this method impractical in numerous applications. A description of Lepski's method and illustration in the case of M-estimator is available in Chapter 4.

We presented in this section several estimators and some of their theoretical guarantees. In the next subsections, we compare these estimators and give some indications on which estimator to use in practice.

2.3.2 Multivariate robust mean estimation

In this section, we study the estimation of the mean in dimension $d > 1$. First, let us see what the concentration of the empirical mean is in $\mathbb{R}^d$. The following theorem, consequence of Hanson-Wright inequality, holds.

Theorem 12 ([HW71])**.** *Let $X \sim \mathcal{N}(0, \Sigma)$ and $X_1, \ldots, X_n$ i.i.d copies of X, with Σ a positive definite matrix. Then, for any $t > 0$,*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\right\|^2 > \frac{2Tr(\Sigma)}{n} + \frac{9t\|\Sigma\|_{op}}{n}\right) \leq e^{-t},$$

where $\|\cdot\|_{op}$ *is the operator norm with respect to the euclidean norm* $\|\cdot\|$.

Theorem 12 will be our objective when estimating the mean: we aim to obtain the same convergence rates when the data are not Gaussian. Remark that Theorem 12 tells us that the deviations are of order $O(1/\sqrt{n})$ but we also have to be careful about the numerator because $\|\Sigma\|_{op}$ can be a lot smaller than $Tr(\Sigma)$ (a typical case is for $\Sigma = I_d$ for which $\|\Sigma\|_{op} = 1$ and $Tr(\Sigma) = d$).

In order to be able to use a robust estimator of the mean in $\mathbb{R}^d$, we have to be careful that the dimension does not have an unwanted effect on the estimation error. In particular, in robust estimation the error due to corruption should not increase with the dimension. In this context, there have been numerous propositions of robust estimators in high dimension. First, there are estimators that have strong theoretical guarantees but that are intractable, for example one can see estimators based the aggregation of one-dimensional estimators, see [Ler19, Theorem 44] and reference therein or estimators based on depth [DG92, CGR+18], for example Tukey's median. On the other hand, there are tractable algorithms but whose theoretical guarantees are lacking, for example the coordinate-wise median or the geometrical median [Min15]. Recently there have been several propositions of algorithms that are said to be at the same time tractable and minimax, see [DKP20, DL19, Hop20] but in practice most of these algorithms take too long to run.

Let us give an example of theoretical results. In the case of Tukey's median, [CGR+18] shows that in a corrupted framework where $X_1, \dots, X_n$ are i.i.d with law $(1-\varepsilon)P + \varepsilon H$. If P is a Gaussian and $\varepsilon \leq 1/\sqrt{n}$, we can estimate the mean efficiently (with minimax rate, i.e. same rates as Hanson-Wright's inequality). On the other hand, if P has a finite second moment and no finite higher moments, we have to ask for $\varepsilon \leq 1/n$ to recover minimax rates. This could be interpreted as a sort of breakdown point for Machine learning: which ε are sufficiently small so that the corruption does not change the rate of convergence (see [LL20]).

In Chapter 4 we show that informally if $X_1, \dots, X_n$ are i.i.d from a mixture distribution $(1-\varepsilon)P + \varepsilon Q$ with P having q finite moments, then, under some assumptions on ψ and n there exists an absolute constant $C > 0$ such that, for all $0 < \lambda \lesssim n$, with probability larger than $1 - 5\exp(-\lambda/8)$,

$$\left\|\mathbb{E}[X] - T_H(\widehat{P}_n)\right\| \lesssim \frac{\sqrt{Tr(\Sigma)} + \sqrt{\|\Sigma\|_{op}\lambda}}{\sqrt{n}} \bigvee \mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q} \varepsilon^{1-1/q} g\left(\lambda, \frac{\varepsilon^{1/2-1/q}}{M}, \frac{1}{M^{\frac{q}{2}} n^{\frac{q-2}{4}}}\right). \tag{2.3.6}$$

Where $\lesssim$ is $\leq$ up to some constant, $M = \frac{\sqrt{Tr(\Sigma)}}{\mathbb{E}[\|X-\mathbb{E}[X]\|^q]^{1/q}}$, and $g : \mathbb{R}^3 \to \mathbb{R}_+$ is such that $g(\lambda, x, y) = 1 + o(\lambda) + o(x) + o(y)$ for (λ, x, y) that tend to 0.

The dependency in the number of finite moments links the two common settings: when P has two finite moments, the bound will be in $\sqrt{\varepsilon}$ as in [DL19, DKP20] in which case we need $\varepsilon \leq 1/n$ to recover minimax rates similar to Hanson-Wright inequality, while if P is Gaussian the dependency in ε is linear which we already seen in the fact that we must have $\varepsilon \leq 1/\sqrt{n}$ in order to preserve minimax rates of convergence. Equation (2.3.6) interpolates between these two extremes.

A surprising consequence of the fact that we separate the effects of the bias and of the variance is that, to achieve rates of order ε (see Corollary 10 in Chapter 4), we don't need the inliers to be

Gaussian, we only need them to be symmetric, and this is true even when the second moment of P is not finite.

We don't achieve minimax rates with this estimator but, in my opinions, this is only because we use the same β for all coordinates. If instead of multiplying by $1/\beta$, we multiply by a $d \times d$ matrix B, then I believe that the estimator can be minimax and wI made some simulations that support this claim in Section 4.

2.3.3 Is it useful to make blocks? — from Huber estimator to HOME estimator

A generalization of Median of Means estimator is to replace the empirical median by a Huber estimator giving birth to HOME (Huber Of Means Estimator) denoted by $\text{HOME}_{K,\beta}(X_1^n)$ when K blocks are used and β is the parameter of Huber's score function. Then, we have to choose how many blocks to use and it is not clear that we really need to use blocks because, as we saw before, Huber estimator is already robust and efficient and the natural question is: when is it useful to make blocks? We made several experiments on synthetic datasets and observed that skewed distribution may be problematic for Huber's estimator and the theory validates this empirical observation at least in the case of stable distributions.

In the following results we explicitly put β in the subscripts of the quantities considered as it may change from one line to the other.

Denote $V_{H,\beta} = \mathbb{E}[\psi_{H,\beta}(|X - T(P)|)^2]$. Suppose $n, K, b \in \mathbb{N}$ with $n = Kb$, there exists an asymmetric distribution P with mean 0 such that for some $\beta_1 > 0$ there exists $\beta_2 > 0$ such that for $X_1, \dots, X_n$ i.i.d with law P, we have the following results.

Huber's estimator : suppose $8V_{H,\beta_2} \leq \beta_2^2$. For all $t > 0$ such that $4\sqrt{2V_{H,\beta_2}t/n} + 4\beta_2 t/n \leq \beta_2/2$, with probability greater than $1 - 2e^{-t} - e^{-n/8}$,

$$\left|T_{H,\beta_2}(\widehat{P}_n) - \mathbb{E}[X]\right| \leq 4\sqrt{\frac{2V_{H,\beta_2}t}{n}} + 4\frac{\beta_2 t}{n} + |T_{H,\beta_2}(P)|. \tag{2.3.7}$$

HOME : suppose $8V_{H,\beta_2} \leq \beta_2^2 b^{-(\alpha-1)/\alpha}$. For all $t > 0$ such that $4\sqrt{2V_{H,\beta_2}tb^{\frac{\alpha-1}{\alpha}}/K} + 4\beta_2 t/K \leq \beta_2/2$, with probability greater than $1 - 2e^{-t} - e^{-K/8}$,

$$|\text{HOME}_{K,\beta_1}(X_1^n) - \mathbb{E}[X]| \leq 4\sqrt{\frac{2V_{H,\beta_2}t}{Kb^{\frac{\alpha-1}{\alpha}}}} + 4\frac{\beta_2 t}{Kb^{\frac{\alpha-1}{\alpha}}} + \frac{1}{b^{\frac{\alpha-1}{\alpha}}}|T_{H,\beta_2}(P)|. \tag{2.3.8}$$

Then, in Equation (2.3.8), we see that increasing b (hence decreasing K) will decrease the value of the bias term $|T_{H,\beta_2}(P)|$ while increasing the value of the other terms of the right hand side of Equation (2.3.8). This verifies that if the bias $|T_{H,\beta_2}(P)|$ of Huber's estimator is large and n is large, then it can be interesting to use HOME instead of Huber's estimator.

Remark that replacing the median in MOM by Huber's estimator slows down the computation of the estimator and for practical purposes, it can be interesting to use MOM even in cases where Huber's estimator would have been better for the task from a theoretical point of view.

2.4 Contributions to asymptotic results in robust estimation — A weaker notion of robustness for a more efficient estimation

2.4.1 Continuity of asymptotic M-estimators

Beginning with Hampel's work [Ham71], an alternative definition of robustness different but close to the one in Section 2.1.3 is the continuity of the functional T: T is continuous at a probability distribution P if we have for all probability distribution Q,

$$\forall \varepsilon > 0, \exists \delta > 0, \text{s.t.} \quad d(P,Q) \leq \delta \Rightarrow |T(P) - T(Q)| \leq \varepsilon.$$

This is an important robust property for T as it says that a small change in the probability P should only cause a small change in the value of $T(P)$. For instance it can be shown in dimension $d = 1$ that if the score function ψ is continuous and bounded (bounded influence function), then T is continuous for Kolmogorov distance, which means that T is robust to a Kolmogorov corruption neighborhood as defined in Section 2.1.3. We interpret this as saying that T is insensitive to small corruption according to Kolmogorov distance and it can be seen that the continuity of T will depend a lot on the distance that we consider. There has been some work to define corruption using other distances like Prokhorov distance, total variation distance, Bounded Lipshitz distance [HR09] and more recently Hellinger distance with the works on rho-aggregation [BBS17, BGH14] which takes another point of view of the problem. All these distances between distributions are insensitive to arbitrary outliers meaning that for any probability distribution P, Q, we have $d(P, (1-\varepsilon)P + \varepsilon Q) \leq \varepsilon$. Such neighborhoods can contain arbitrary outliers (outliers that are arbitrarily large).

On the other hand, if we are allowed to make some assumptions on the distribution of outliers then maybe there will be a larger set of continuous functionals T with respect to this distance and the associated estimators may be more efficient in practice on uncorrupted datasets. This motivates the definition of a new family of distances on distributions.

Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ and let $\mathcal{P}_\psi = \{P \in \mathcal{P} : \mathbb{E}_P[\psi(\|X\|)] < \infty\}$. For all $P, Q \in \mathcal{P}_\psi$, let

$$W_\psi(P,Q) = \sup_{h \preceq \psi} \left\{ \int h(x) dP(x) - \int h(x) dQ(x) \right\}, \tag{2.4.1}$$

where $h : \mathbb{R}^d \to \mathbb{R}$ verifies $h \preceq \psi$ if and only if for all $x, y \in \mathbb{R}^d$, $|h(x) - h(y)| \leq \psi(\|x - y\|)$. W_ψ is not a very unusual distance, it is in fact a Wasserstein-1 distance in the metric space $(\mathbb{R}^d, d_\psi)$ where $d_\psi(x,y) = \psi(\|x-y\|)$. If ψ is the identity, we recover the usual Wasserstein-1 distance and in the extreme case for which ψ is the constant function equal to 1 we recover the total variation distance and it can be understood that W_ψ is a weaker notion of distance than total variation in the sense that we will have to ask some assumptions on Q so that $W_\psi((1-\varepsilon)P + \varepsilon Q, P) \xrightarrow[\varepsilon \to 0]{} 0$. Now that W_ψ has been defined, we can state the following theorem which is central in [Mat20b].

Theorem 13. *Let ψ denote one of the three functions ψ_H, ψ_C or ψ_P. Let T be the functional constructed from ψ according to Equation (2.2.2), and let $P \in \mathcal{P}_\psi$. Suppose that $\psi(+\infty) > \mathbb{E}_P[\psi(\|X\|)]$ and that $\|X\|$ is almost surely finite.*
Then, T is continuous at P for the distance W_ψ over $\mathcal{P}_\psi$. In other words, we have for all $Q \in \mathcal{P}_\psi$,

$$\|T(P) - T(Q)\| \xrightarrow[W_\psi(P,Q) \to 0]{} 0.$$

From Theorem 13, we see that the choice of the function ψ decides the distance W_ψ to which T is continuous and the notion of neighborhood with respect to W_ψ and informally this also defines the corruption that T can handle. In Section 2.4.2, we precise this remark and show a consequence of the use W_ψ.

2.4.2 Asymptotic stability of M-estimators — comparison of M-estimators with different score functions

The choice of the score function ψ has a big impact on the robustness of the resulting M-estimator. In particular, if ψ is bounded then the estimator is robust in the sense of Hampel (see [HR09]) and if ψ is close to the identity near 0, then the bias of the resulting estimator when we estimate the mean is small (see Lemma 2). On the other hand, if ψ is not bounded it does not necessarily mean that we are loosing all the robustness of the estimator but we have to make some assumptions on the outliers for the estimator to still be consistent in a corrupted setting. We have the following result.

Corollary 2. *Suppose we are in a $\mathcal{I} \cup \mathcal{O}$ framework where $(X_j)_{j\in\mathcal{I}}$ denote an i.i.d sample from P, and let $(X_j)_{j\in\mathcal{O}}$ all be equal to $g(n)u$ for some $u \in \mathbb{R}^d$, $u \neq 0$ and $g : \mathbb{N} \to \mathbb{R}$ increasing. Denote by $|O| = k_n$ the cardinal of the set of outliers. The following results hold true.*

Huber's estimator *Let P be a probability distribution on $\mathbb{R}^d$ and suppose that $\mathbb{E}[\psi_H(\|X\|)] < \beta < \infty$.*

$$\left(\frac{k_n}{n} \xrightarrow[n\to\infty]{} 0\right) \Rightarrow \left(T_H\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right) \xrightarrow[n\to\infty]{P} T_H(P)\right).$$

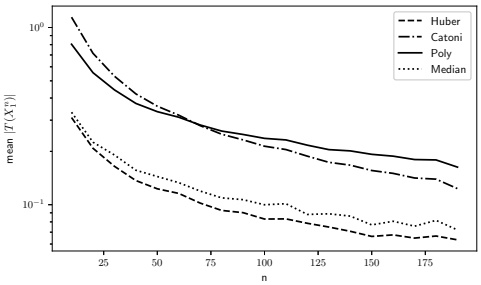
Catoni's estimator *Let P be such that for $X \sim P$, $\mathbb{E}[\psi_C(\|X\|)] < \beta < \infty$.*

$$\left(\frac{k_n \log(g(n))}{n} \xrightarrow[n\to\infty]{} 0\right) \Rightarrow \left(T_C\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right) \xrightarrow[n\to\infty]{P} T_C(P)\right).$$

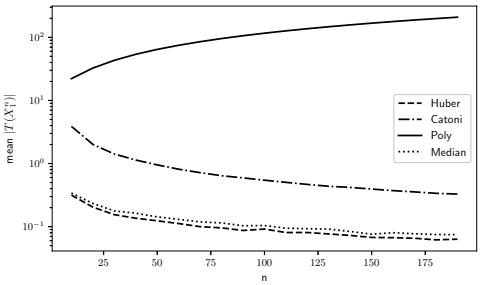
Polynomial estimator *Let P be such that for $X \sim P$, $\mathbb{E}[\psi_P(\|X\|)] < \beta < \infty$. We have,*

$$\left(\frac{k_n g(n)^{1/p}}{n} \xrightarrow[n\to\infty]{} 0\right) \Rightarrow \left(T_P\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right) \xrightarrow[n\to\infty]{P} T_P(P)\right).$$

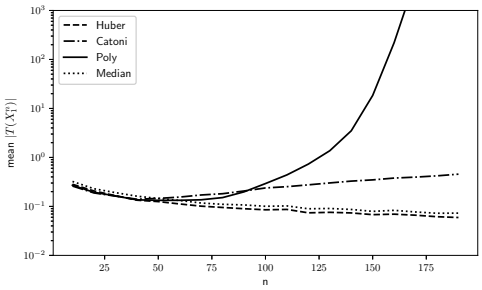
This corollary tells us how big outliers can be and how many of them the estimator can handle before it affects the behavior of this estimator. For Huber's estimator there is no restriction on g as long as the number of outliers $k_n = o(n)$. For Catoni's estimator, if k_n is bounded (finite number of outliers), $g(n)$ must be negligible compared to $\exp(n)$ and for the Polynomial estimator, if k_n is bounded, $g(n)$ must be negligible compared to n^p. Figure 2.6 illustrates this behavior. In practice, if the outliers satisfy the hypothesis for Catoni's estimator (respectively Polynomial estimator) to converge, then Catoni's estimator (respectively Polynomial estimator) will be a bit more efficient than Huber estimator (see Chapter 3). This result gives us some rules to design M-estimators when we have an idea of the scale of the outliers.



(a) One outlier in n^2



(b) One outlier in n^5



(c) One outlier in $\exp(n^2/1000)$

Figure 2.6: Plots of the mean absolute distance to 0 (log-scale) for various estimators on Gaussian datasets corrupted with one outlier as a function of the sample size n. The polynomial estimator uses $p = 3$.

2.5 Contributions to robust Machine Learning

In this section we are interested in infinite dimensional estimation. We begin with classification and regression framework (see [Kol11, DGL96, MRT12] for background on these subjects) and then we present kernel space embedding of probability distributions with application in maximum mean discrepancy (MMD) estimation (see [GBR+12] for background on MMD). Our goal is to get robust estimators of the quantities considered.

2.5.1 Classification and regression using Median of Means principle

In classification and regression, we will use M-estimators and MOM estimators to make existing estimators robust. In particular, we will be interested in linear classification/regression through logistic regression and ordinary least square but we will also apply this to non-linear classification with kernel methods. Median of Means principle has been applied to Machine Learning problem in several works and among them, one can cite the MOM tournaments methods [LM19a, LM19c] that are theoretically efficient but intractable in practice, see also minmax MOM estimators [LL18, LL20] and works on sparse recovery and high dimension Machine Learning [CLL19a, LM+18].

Recall the goal of equation (2.1.1) where we search for

$$f^* \in \operatorname*{argmin}_{f\in\mathcal{F}} R(f) = \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}[\ell(f(X), Y)]. \tag{2.5.1}$$

with ℓ some loss function. The empirical risk minimization [Vap98] framework dictates to estimate the risk $R(f)$ by replacing the expectation by an empirical mean and then we estimate f^* using $\widehat{f}$, a minimizer of this empirical risk,

$$\widehat{f} \in \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

Classical examples of loss functions ℓ are the logistic loss $\ell(f(x), y) = \log(1 + \exp(-f(x)y))$ in classification or the squared loss $\ell(f(x), y) = (f(x) - y)^2$ in regression. The problem is that empirical risk minimization is not robust when $X_1, \dots, X_n$ are not light-tailed.

To describe the behavior of an estimator of f^*, we will need a notion of complexity of $\mathcal{F}$ and the more complex $\mathcal{F}$ is the harder the estimation is. There are several standard notions of complexity like the VC dimension or the Entropy. In this work, we will use the Rademacher complexity (see [MRT12, p34]). Suppose that the features X are in a set $\mathcal{X}$. Let $\mathcal{G}$ be a set of functions $f : \mathcal{X} \to \mathbb{R}$ and let $(\varepsilon_i)_{1\le i\le n}$ denote i.i.d Rademacher random variables independent from $(X_i)_{1\le i\le n}$. The Rademacher complexity of $\mathcal{G}$ is defined by

$$\mathrm{Rad}(\mathcal{G}) = \mathbb{E}\left[\sup_{f\in\mathcal{G}} \sum_{i=1}^{n} \varepsilon_i f(X_i)\right].$$

The larger the Rademacher complexity is, the more complex $\mathcal{G}$ is and the harder it is to estimate functions in $\mathcal{G}$. For example for linear classifiers on $\mathbb{R}^d$ whose coefficient has an Euclidean norm bounded by θ_2, we can show that $\mathrm{Rad}(\mathcal{G}) \le \theta_2\sqrt{nd}$.

The principle of robust empirical risk minimization is as follows: instead of estimating the expectation in (2.5.1) using the empirical mean, we use a robust estimator of the mean.

$$\widehat{f}_{rob} \in \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{E}(\ell(f(X_i), Y_i), 1 \le i \le n)$$

where $\widehat{E}$ is some robust estimator of the mean, typically a M-estimator or Median of Means estimator. First, let us describe the result when using the Median of Means estimator. Let

$$\widehat{f}_{MOM,K} \in \operatorname*{argmin}_{f\in\mathcal{F}} \mathrm{MOM}_K(\ell(f(X_i), Y_i), 1 \le i \le n)$$

Theorem 14. *Suppose that for all $f \in \mathcal{F}$, $\theta_2 := \mathbb{E}[f(X)^2] < \infty$, suppose $\mathrm{Rad}(\mathcal{F}) < \infty$ and that the loss function is Lipshitz in the sense that there exists $L > 0$ such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and all $f, f' \in \mathcal{F}$,*

$$|\ell(f(x), y) - \ell(f'(x), y)| \le L|f(x) - f'(x)|.$$

Assume $n > K > 4|\mathcal{O}|$ and denote $\Delta = 1/4 - |\mathcal{O}|/K$. Then, with probability greater than $1 - 2e^{-2\Delta^2 K}$,

$$R(\widehat{f}_{MOM,K}) \le \inf_{f\in\mathcal{F}} R(f) + 4L\max\left(\frac{4\mathrm{Rad}(\mathcal{F})}{n}, 2\theta_2\sqrt{\frac{K}{n}}\right)$$

The inequality in Theorem 14 attains the minimax rates of convergence (when we don't suppose any margin assumption) even though we only supposed a finite L^2 norm and we are in a corruption framework; whereas typical machine learning results will suppose $\mathcal{F}$ bounded in L^∞. There are a lot of class of functions (e.g. linear classifiers, SVM, ...) for which $\mathrm{Rad}(\mathcal{F}) \leq O(\sqrt{n})$ and this is why we may say sometimes that the optimal rate is in $O(1/\sqrt{n})$.

Then in Chapter 6 we have gone further to include margin condition which allow us to recover rates that are faster than $1/\sqrt{n}$ (for this part, we change the notations compared to Chapter 6 to match the ones presented until now in the introduction). The margin condition is the following assumption: there exists constants $D > 0, \delta_B > 0$ such that

$$Var(\ell(f(X), Y) - \ell(f^*(X), Y)) \leq D^2(R(f) - R(f^*)) \tag{2.5.2}$$

whenever $R(f) - R(f^*) \leq \delta_B$.

We consider a more general estimator where the median operator in the median of means principle is replaced by a M-estimator similar to Huber's estimator. First we define an estimator of $R(f)$ denoted by $\widehat{R(f)}_{\beta,K}$ and defined as

$$\sum_{k=1}^{K} \psi\left(\sqrt{b}\frac{P_{B_k}(\ell(f(X_i), Y_i)_{1\leq i\leq n}) - \widehat{R(f)}_{\beta,K}}{\beta}\right) = 0. \tag{2.5.3}$$

Where the assumptions on ψ are that ψ is odd and five times continuously differentiable, that if $|x| \leq 1$, then $\psi(x) = x$ and if $|x| \geq 2$ then ψ is constant and $x - \psi(x)$ is non-decreasing. ψ is essentially a smoothed version of ψ_H. Then, the proposed estimator of f^* is given by

$$\widehat{f}_{\mathrm{HOME},K} \in \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{R(f)}_{\beta,K}$$

For this estimator, we show that we can get a rate $O(1/n^{3/4})$ under margin conditions (we also prove the slow rates when we don't have the margin condition). This rate is faster than the usual rate of $O(1/\sqrt{n})$ but even though we witness an optimal rate of $O(1/n)$ in practice we didn't succeed in proving this bound. On the other hand, we also propose another estimator based on the min-max principle, and this estimator achieves optimal rates $O(1/n)$ under margin conditions. These theorems are a bit technical and we present here only an informal version, see Chapter 6 for details.

Theorem 15 (Informal)**.** *Suppose that for all $f \in \mathcal{F}$, $\mathbb{E}[\ell(f(X), Y)^4] < \infty$ and suppose $\mathrm{Rad}(\mathcal{F}) < \infty$. Then for K and β appropriately chosen, there exists constants $c_1, c_2 > 0$ such that with probability greater than $1 - e^{-s}$ for all $0 < s \leq c_1 K$*

$$R(\widehat{f}_{HOME,K}) \leq \inf_{f\in\mathcal{F}} R(f) + c_2\left(\frac{\mathrm{Rad}(\mathcal{F})}{n} + \frac{s}{n^{3/4}} + \left(\frac{|\mathcal{O}|}{n}\right)^{3/4}\right).$$

Theorem 15 shows that in the context without margin condition, as long as we have a fourth moment condition, the additional error due to the corruption is of order $O(1/n^{3/4})$ which is negligible compared to $\mathrm{Rad}(\mathcal{F})/n$. We attain optimal rates. Then, we also present the case where fast rates can be proven.

Theorem 16 (Informal). *Suppose that for all $f \in \mathcal{F}$, $\mathbb{E}[\ell(f(X), Y)^4] < \infty$ and suppose $\mathrm{Rad}(\mathcal{F}) < \infty$, if the margin assumption from equation (2.5.2) is verified. Then there exists an estimator $\widehat{f}_{fr}$ such that with probability greater than $1 - e^{-s}$ for all $s \leq s_{max}$ where $s_{max} \xrightarrow[n\to\infty]{} \infty$, we have*

$$R(\widehat{f}_{fr}) \leq \inf_{f \in \mathcal{F}} R(f) + c_2 \left(\frac{\mathrm{Rad}(\mathcal{F} - f^*)}{n} + \frac{s}{n} + \frac{|\mathcal{O}|}{n} \right)$$

where $\mathcal{F} - f^* = \{x \mapsto f(x) - f^*(x), f \in \mathcal{F}\}$.

Under some conditions on $\mathcal{F}$, one can show that the bound in Theorem 16 is of order $O(1/n)$ which is minimax optimal for the problem at hands when the margin condition is verified. The description of $\widehat{f}_{fr}$ is much more involved than $\widehat{f}_{HOME,K}$ and is described in Chapter 6.

2.5.2 Kernel method and concentration in infinite dimensional space

Let $\mathcal{X}$ be a set on which $\mathcal{K}$ is a kernel, we can represent a probability P on $\mathcal{X}$ as a mean in the RKHS $\mathcal{H}_{\mathcal{K}}$.

$$\mu_P = \int_{\mathcal{X}} \varphi(x) dP(x), \quad \varphi(x) := \mathcal{K}(\cdot, x).$$

This is called a mean embedding and using this embedding, we introduce the following distance between distributions called the maximum mean discrepancy,

$$MMD(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_{\mathcal{K}}} = \sup_{f \in \mathcal{B}_{\mathcal{K}}} \langle \mu_P - \mu_Q, f \rangle_{\mathcal{H}_{\mathcal{K}}},$$

where $\mathcal{B}_{\mathcal{K}} = \{f \in \mathcal{H}_{\mathcal{K}} : \|f\|_{\mathcal{H}_{\mathcal{K}}} \leq 1\}$. Kernel methods are very efficient when dealing with structured data, DNA for example, because we can design a kernel adapted to the structure of the data. Then, the MMD can be used for instance to construct two sample tests to compare two populations, this is what has been done as an illustration of our method and this is presented in Section 2.6.4.

In the computation of MMD, we have to compute a mean in infinite dimension and we view this problem as a robust estimation of the mean. Assume that we have access to $X_1, \ldots, X_n$ i.i.d with law P and $Y_1, \ldots, Y_n$ i.i.d with law Q. Denote $P_{B,x} = \frac{1}{|B|} \sum_{i \in B} \delta_{x_i}$ the empirical measure associated with $(x_i)_{i \in B}$, denote by $B_1, \ldots, B_K$ an equi-partition of $\{1, \ldots, n\}$, we propose the following estimator.

$$\widehat{MMD}_K(P, Q) = \sup_{f \in \mathcal{B}_{\mathcal{K}}} \mathrm{Med}\Big(\langle f, \mu_{P_{B_k},x} - \mu_{P_{B_k},y} \rangle, \quad 1 \leq k \leq K \Big).$$

This estimator presents theoretical advantages in terms of robustness but also practical advantages in terms of computation (See 2.6.4). We have the following guarantees in terms of concentration. Denote $(e_i)_{i \in I}$ a countable orthonormal basis of $\mathcal{H}_{\mathcal{K}}$ (which exists because $\mathcal{H}_{\mathcal{K}}$ is separable), define $\|A\|_1 = \sum_{i \in I} \langle (A * A)^{1/2} e_i, e_i \rangle_{\mathcal{H}_{\mathcal{K}}}$ where A^* is the adjoint operator of A and $\|A\|$ the operator norm of A.

Theorem 17. *Assume that Σ_P is a linear operator on $\mathcal{H}_K$ with $\|\Sigma_P\|_1 < \infty$. Suppose the dataset $(x_i, y_i)_{i \leq n}$ is corrupted by n_c outliers in the $\mathcal{I} \cup \mathcal{O}$ framework described in Section 2.1.1 (i.e. there can be n_c outlier couples $(x_{i_1}, y_{i_1}), \ldots, (x_{i_{n_c}}, y_{i_{n_c}})$ on which we make no hypothesis). Let*

$\delta \in (0, 1/2]$ be such that $n_c \leq K(1/2 - \delta)$. Then, for any $\eta \in (0, 1)$ such that $K = 72\delta^{-2}ln(1/\eta)$ satisfies $K \in \left(\frac{n_c}{1/2-\delta}, \frac{n}{2}\right)$, with probability at least $1 - \eta$,

$$\left|\widehat{MMD}_K(P, Q) - MMD(P, Q)\right| \leq \frac{12}{\delta} \max\left(\sqrt{\frac{(\|\Sigma_P\| + \|\Sigma_Q\|)ln(1/\eta)}{\delta n}}, 2\sqrt{\frac{Tr(\Sigma_P) + Tr(\Sigma_Q)}{n}}\right)$$

Theorem 17 shows that the estimator $\widehat{MMD}_K$ attains the rate $O(1/\sqrt{n})$ which is optimal for this problem, it also shows that our estimator is robust to outliers with breakdown point almost $K/2$. The rates in Theorem 17 are very similar to Hanson-Wright inequality, this can be seen as an infinite dimension extension of Hanson-Wright inequality.

2.6 Robustness in practice, some contributions to robust algorithms and empirical studies

2.6.1 Robust algorithms for supervised and unsupervised learning

In this section we look into the two estimators $\widehat{f}_{MOM,K}$ and $\widehat{f}_{HOME,K}$ and we illustrate the computation of these estimators on the dataset represented in Figure 2.7. The second estimator has a straightforward implementation with gradient descent because using equation (2.5.3) we can take the derivative and get the gradient of $\widehat{R}_{K,\beta}$ with respect to f. On the other hand, the Median of Means used to define $\widehat{f}_{MOM,K}$ is not differentiable, it is however not an issue because it is in fact differentiable almost everywhere. Let B_{Med} be the median block in the sense that it verifies

$$\begin{aligned}\frac{1}{b}\sum_{i \in B_{\text{Med}}} \ell(f(X_i), Y_i) = P_{B_{\text{Med}}}(\ell(f(X_i), Y_i)) &= \text{Med}\{P_{B_k}(\ell(f(X_i), Y_i)), 1 \leq k \leq K\} \\ &= \text{MOM}_K(\ell(f(X_i), Y_i), 1 \leq i \leq n).\end{aligned}$$

We show in Chapter 5 that taking the derivative of the median of means criterion is the same as taking the derivative only on B_{Med}. We have almost everywhere (in a sense that is made clear in Chapter 5) that

$$\frac{d}{df}\text{MOM}_K(\ell(f(X_i), Y_i), 1 \leq i \leq n) = \sum_{i \in B_{\text{Med}}} \frac{d}{df}\ell(f(X_i), Y_i). \tag{2.6.1}$$

Now that we have a gradient, we can make a gradient descent algorithm. The problem is that the objective function is not convex due to the blocks and there can be local minima in the objective function. Recall that the blocks are constructed arbitrarily and are supposed fixed in theory. In practice we see that it is better to change the blocks at each gradient step, or said differently we shuffle the data at each step. This has the effect of introducing noise with an idea similar to stochastic gradient descent but with the stochastic part being the permutation of the data (we could rewrite the problem as saying that instead of minimizing the MOM of the losses, we minimize the mean on all permutations of the MOM's where we shuffle the data according to the permutation). There is no proof of convergence of this algorithm but theoretical results from

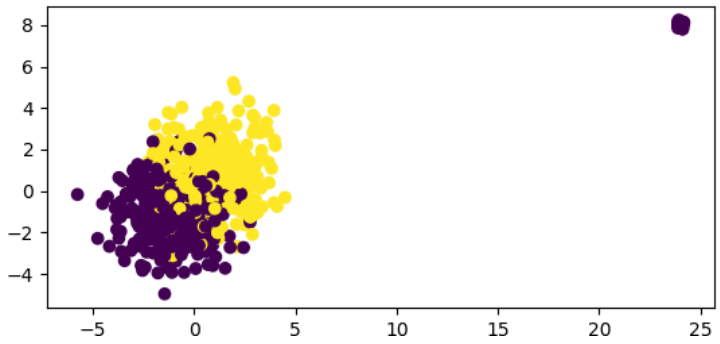


Figure 2.7: Plot of a corrupted dataset for classification purpose. It is composed of two Gaussian blobs of 300 points each and 30 outliers in the upper right corner of the figure.

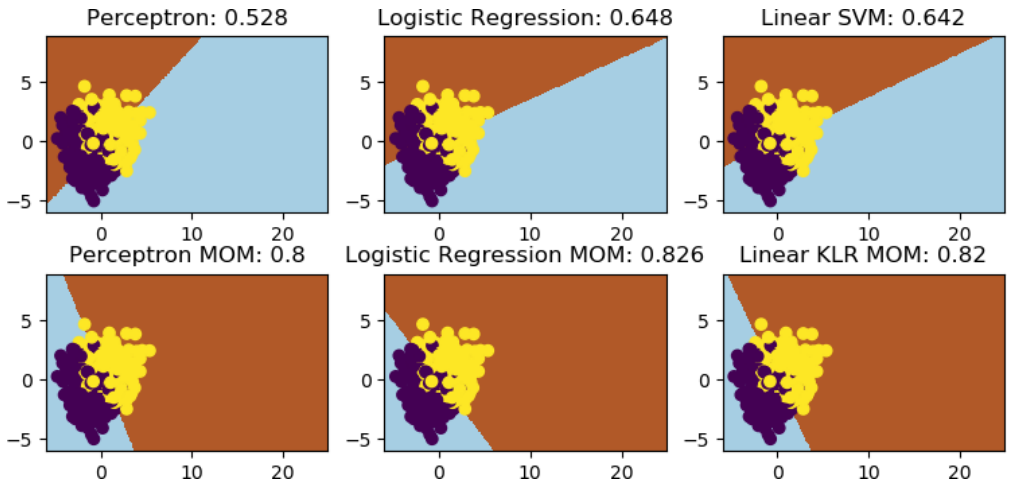


Figure 2.8: Plot of the decision boundary computed by several classifiers trained on the dataset from Figure 2.7.

U-Statistics (see Chapter 6) and empirical observations support this algorithm and we obtain good performances in practice.

For instance, in Figure 2.8 we plot the result of several linear classifiers when trained on the dataset from Figure 2.7, the MOM classifiers have been trained using gradient descent with the gradient from equation (2.6.1) and the non-robust algorithms are from python library scikit-learn. Figure 2.8 could be said to be just a sanity check: our methods are robust and the usual ones are not. We get similar results in regression with linear methods. Interpreting non-linear methods is harder because it is difficult to conceive what is an outlier for such a method and hence it is difficult to simulate, we can understand on specific examples like SVM or QDA what is an outlier with respect to such classifiers but in general it is harder.

The choice of K can be tricky, the intuition is that we should choose K a little larger than $|\mathcal{O}|/2$ but the problem is that $|\mathcal{O}|$ is not known to the statistician and even if it was (we estimate

that in most real datasets, there is between 5% and 10% outliers) this is only a rule of the thumb and the choice of optimal K depends also on the inliers. Hence we think that the best course of action is to use cross-validation to tune K.

To study the performance of our method, we look at the classification of a real dataset using $\widehat{f}_{HOME,K}$. The dataset considered is "Communities and Crime Unnormalized Data Set" and is available through the UCI Machine Learning Repository. These data contain 2215 observations from a census and law enforcement records. The task we devised was to predict the crime activity (represented as the count of incidents) using the following features: the population of the area, the per capita income, the median family income, the number of vacant houses, and the land area. The choice of this specific dataset was motivated by the fact that it likely contains a non-negligible number of outliers due to the nature of the features and the fact that the data have not been pre-processed, hence the advantages of proposed approach could be highlighted.

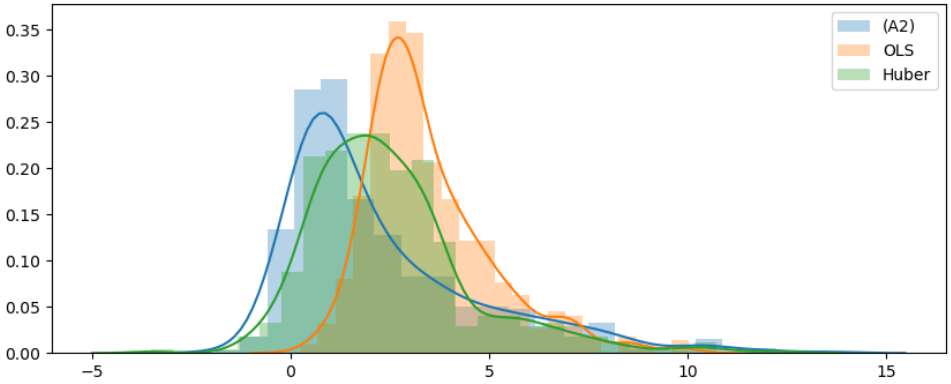


Figure 2.9: Histogram of densities of the logarithm of the MSE for the different methods (light blue corresponds to our approach, orange to the standard least squares regression, and green to Huber's loss regression).

We compare the linear regressor $\widehat{f}_{HOME,K}$ (called A2 in the figure) with respect to the squared loss, the ordinary least square estimator (OLS) and the estimator that uses the Huber loss (HuberRegressor in scikit-learn, be careful that even though HuberRegression use the Huber loss, this is not a robust algorithm according to our definition of robust estimator). The parameters are tuned using a robust version of a 500 fold cross-validation. We obtain an MSE $\simeq e^{4.2}$ for our approach, the MSE of OLS is $\simeq e^{22.1}$ and the MSE of Huber loss approach is $\simeq e^{8.9}$. The density of the MSE on the different folds is represented in Figure 2.9 (we took the logarithm to make things readable). Then, we see that indeed there seems to be outliers in the dataset as expected and that the outliers are not only in the response variable (to which Huber loss approach is robust) but also in the features because our algorithm gives the better results.

2.6.2 Outlier detection using Median of Means

In Section 2.6.1, $\widehat{f}_{MOM,K}$ is constructed via a gradient descent on the median block B_{Med}. We also said that we shuffle the blocks at each step. Then, it can be interesting to consider how many times a given point appears in the median block and this is interpreted as some degree of

how much of an outlier a point is, a score of some sort. If the point is a very bad outlier it is intuitive that it will almost never be selected in B_{Med}. For example, if we plot this score when computing the Logistic Regression MOM on the toy dataset from Figure 2.7 we get Figure 2.10, we colored the bar according to whether the point was an outlier (in the upper right corner of Figure 2.7) or not.

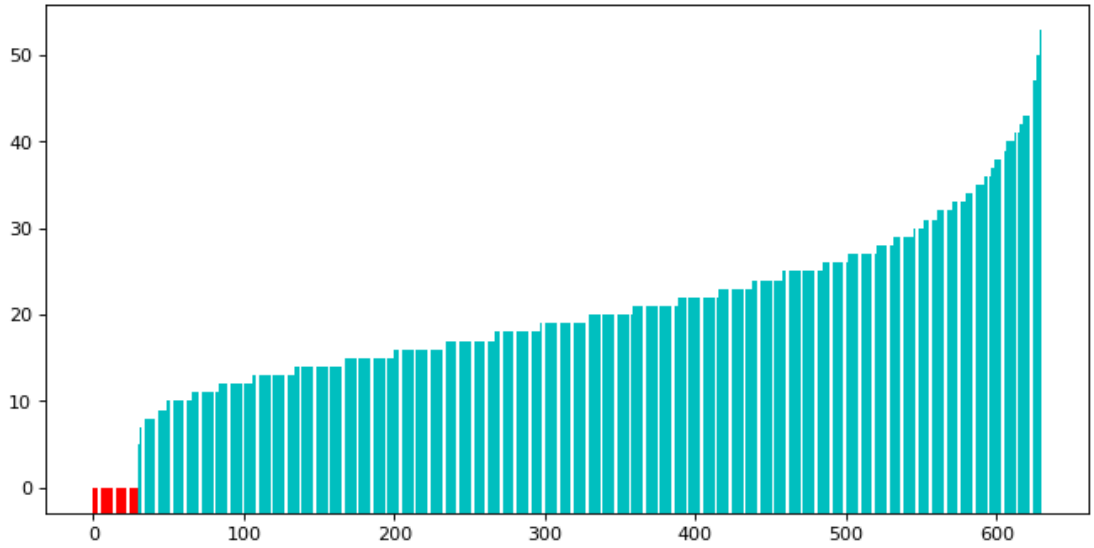


Figure 2.10: Sorted Histogram of the score (number of times a data belongs to the selected median block) of each points in a Logistic Regression MOM algorithm on a toy dataset. Red is an outlier and blue is an informative sample. $K = 120$ blocks and the number of iterations is 2000.

In Figure 2.10 we see that indeed we recover the outliers as the points with low score. This outlier detection algorithm is peculiar because it only detects outliers with respect to the classification (or regression) problem which means that in Figure 2.11 the points on the bottom left corner are not considered outliers, this is very different from the behavior of unsupervised algorithms like one-class SVM or isolation forest. We have injected some knowledge: the learning task to be considered.

Compared to a learning task, the choice of K for outliers detection is done with a different criterion. In this application, it will most of the time pay to choose K a lot larger than $|\mathcal{O}|/2$. Indeed, if K is large, then one outlier point will more easily make the mean of a group very large and then its score would be low. The parameter K becomes a measure of how robust we want to be, or said differently it is a measure of how anomalous a point has to be to detect it as an outlier.

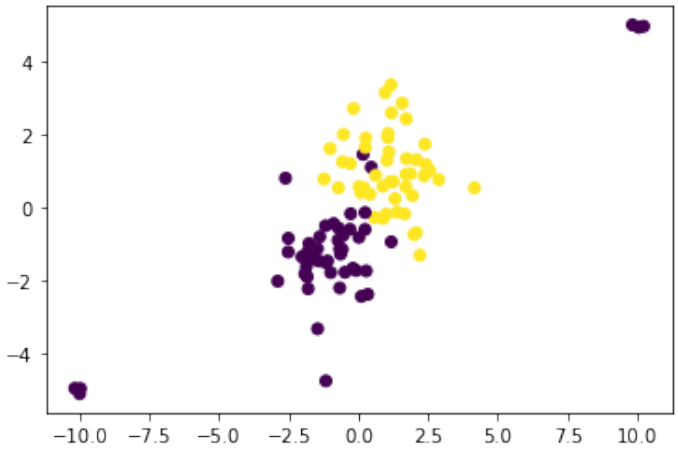


Figure 2.11: Scatter plot of a dataset that present outliers on the upper right corner and non-outliers points on the bottom left corner.

2.6.3 A unified view of robust algorithms for empirical risk minimization — implementation in scikit-learn-extra library

We explain here the principle behind the algorithms implemented by the author as a part of the python library scikit-learn-extra (accessible via pip). Scikit-learn-extra is a Python module for machine learning that extends scikit-learn [PVG+11]. It includes algorithms that are useful but do not satisfy the scikit-learn inclusion criteria, for instance due to their novelty or lower citation number.

The regressors and classifiers that we considered are all based on the robust empirical risk minimization principle:

$$\widehat{f}_{MOM,K} \in \operatorname*{argmin}_{f\in\mathcal{F}} \mathrm{MOM}_K(\ell(f(X_i),Y_i), 1\leq i\leq n)$$

or

$$\widehat{f}_{\mathrm{HOME},K} \in \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{R(f)}_{\beta,K}.$$

As a unified view of this, we see that in fact a robust estimator can be expressed as the minimizer of a weighted sum whose weights depend on the data.

$$\widehat{f}_w \in \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{n} w_i \ell(f(X_i),Y_i) \tag{2.6.2}$$

where w_i is a positive number that can depend on all the data and f. For example, using the Median of Means, $\widehat{f}_{MOM,K} = \widehat{f}_w$ with

$$w_i = \mathbb{1}\left\{ i\in B : \quad \frac{1}{b}\sum_{i\in B} \ell(f(X_i),Y_i) = \mathrm{MOM}_K(\ell(f(X_i),Y_i), 1\leq i\leq n)\right\} = \mathbb{1}\{i\in B_{\mathrm{Med}}\},$$

we consider only the points that are in the "median block" with respect to the loss function (i.e. the block that realizes the median of means).

Consider the following equation, equivalent to (2.5.3) in the extreme case where all the blocks are of size 1.

$$\sum_{i=1}^{n} \frac{\ell(f(X_i),Y_i)) - \widehat{R(f)}_\beta}{\ell(f(X_i),Y_i)) - \widehat{R(f)}_\beta} \psi\left(\frac{\ell(f(X_i),Y_i)) - \widehat{R(f)}_\beta}{\beta}\right) = 0. \tag{2.6.3}$$

that we rewrite

$$\widehat{R(f)}_\beta = \sum_{i=1}^{n} \Big(\ell(f(X_i),Y_i)) - \widehat{R(f)}_\beta\Big) \left(\frac{\psi\left(\frac{\ell(f(X_i),Y_i)) - \widehat{R(f)}_\beta}{\beta}\right)}{\sum_{i=1}^{n} \ell(f(X_i),Y_i)) - \widehat{R(f)}_\beta}\right) \tag{2.6.4}$$

this is also a weighted sum (this can be extended to the case where the blocks have a size larger than 1) where

$$w_i = \frac{a_i}{\sum_{i=1}^{n} a_i} \quad \text{where} \quad a_i = \beta \frac{\psi\left(\left|\frac{\ell(f(X_i),Y_i)) - \widehat{R(f)}_\beta}{\beta}\right|\right)}{\ell(f(X_i),Y_i)) - \widehat{R(f)}_\beta}.$$

This rewriting as a weighted sum is famous in the context of M-estimators and it gives rise to the iterative reweighted algorithm which consist in iteratively optimizing the weighted sum while considering the weight fixed (with a gradient descent typically) and then we recompute the weights and then loop back to optimizing the weighted sum. This algorithm has been implemented in the library scikit-learn-extra. It exhibits the same performances on classification and regression than the ones shown in Section 2.6.1 and it can also be used for example for k-means clustering. More generally, the same algorithm can also be used with the libraries xgboost and keras because both support the sample_weight parameter however these implementations are not part of scikit-learn-extra library. For example of results in the case of the Diabetes UCI dataset is given in Figure 2.12, the robust algorithm is both less variable and more efficient than the non-robust scikit-learn algorithm SGDClassifier (stochastic gradient descent on hinge loss). Further examples and documentation on the usage of scikit-learn-extra can be found at https://scikit-learn-extra.readthedocs.io, see in particular https://scikit-learn-extra.readthedocs.io/en/latest/modules/robust.html and examples in https://scikit-learn-extra.readthedocs.io/en/latest/auto_examples/index.html.

2.6.4 MOM algorithm for MMD estimation and kernel method

Recall that the robust MMD estimator is given by

$$\widehat{MMD}_K(P,Q) = \sup_{f\in\mathcal{B}_\mathcal{K}} \text{Med}\left\{\frac{1}{b}\sum_{i\in B_k} f(x_i) - \frac{1}{b}\sum_{i\in B_k} f(y_i); \quad 1\le k\le K\right\}. \tag{2.6.5}$$

By the representer theorem, the optimal f can be expressed as

$$f(a,b) = \sum_{i=1}^{n} a_i \mathcal{K}(\cdot,x_i) + \sum_{i=1}^{n} b_i \mathcal{K}(\cdot,y_i), \tag{2.6.6}$$

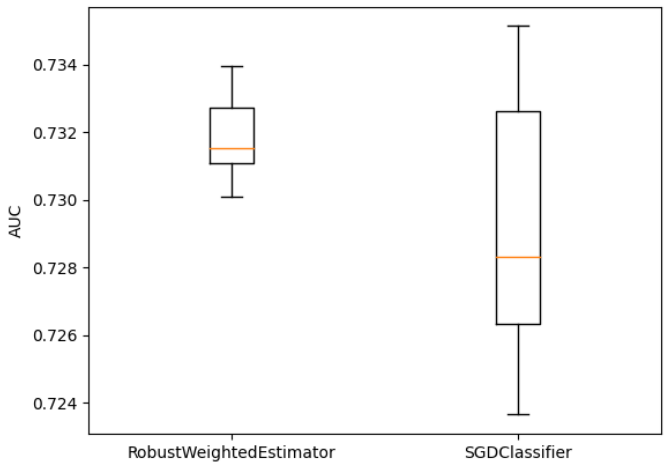


Figure 2.12: Boxplot of the AUC of the two estimators on the Diabetes dataset. The AUC is computed through 10-fold cross validation, this gives us one value and we repeat this experience 10 times to obtain 10 scores that constitute this boxplot. The algorithms are random and we use the boxplots to assess the variability of the results of the two algorithms.

where $a = (a_i)_{1\leq i\leq n} \in \mathbb{R}^n$ and $b = (b_i)_{1\leq i\leq n} \in \mathbb{R}^n$. Denote $c = [a; b] \in \mathbb{R}^{2n}$ the concatenation of the two vectors and $\mathbf{K}$ the kernel matrix defined by $\mathbf{K} = [\mathbf{K}_{xx}, \mathbf{K}_{xy}; \mathbf{K}_{yx}, \mathbf{K}_{yy}] \in \mathbb{R}^{2n\times 2n}$, $\mathbf{K}_{xx} = [\mathcal{K}(x_i, x_j)]_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$, $\mathbf{K}_{xy} = [\mathcal{K}(x_i, y_j)]_{1\leq i,j\leq n} = \mathbf{K}_{yx}^* \in \mathbb{R}^{n\times n}$, $\mathbf{K}_{yy} = [\mathcal{K}(y_i, y_j)]_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$.

Equation (2.6.5) can be rewritten as

$$\widehat{MMD}_K(P, Q) = \max_{c\in\mathbb{R}^{2n}: c^T\mathbf{K}c\leq 1} \text{Med}\left\{\frac{1}{b}[1_k; 1_k]^T\mathbf{K}c; \quad 1 \leq k \leq K\right\},$$

where 1_k is the indicator vector of the block B_k. Similarly to Section 2.6.1 we do iteratively an optimization step on the median block except that in this case we don't have to use gradient descent because on the median block, we are confronted to a quadratic constrained linear optimization problem which has an analytic solution:

$$\operatorname*{argmax}_{c\in\mathbb{R}^{2n}: c^T\mathbf{K}c\leq 1} \frac{1}{b}[1_{k_{\text{Med}}}; 1_{k_{\text{Med}}}]^T Kc = \frac{[1_{k_{\text{Med}}}; 1_{k_{\text{Med}}}]}{\|\mathbf{L}^T[1_{k_{\text{Med}}}; 1_{k_{\text{Med}}}]\|_2}$$

where L is the Cholesky factor of $\mathbf{K}$ ($\mathbf{K} = \mathbf{L}\mathbf{L}^T$) and k_{Med} is such that $B_{\text{Med}} = B_{k_{\text{Med}}}$. The observations are shuffled at each iteration. Notice that this algorithm (called MONK BCD for Median Of meaNs Kernel Block Coordinate Descent) has a complexity of $O(n^3)$ which can be prohibitive for large sample size. Hence, we propose another algorithm called MONK BCD-Fast that approximate MONK BCD where the summation $\sum_{i=1}^n$ in (2.6.6) after plugging it into (2.6.5) is replaced by $\sum_{i\in B_k}$ (this correspond to computing the kernel matrix only as a block matrix). We compare our algorithms to the state of the art U-statistic approach corresponding to the case

Table 2.1: Computational complexity of MMD estimators. n: sample number, K: number of blocks, T: number of iterations.

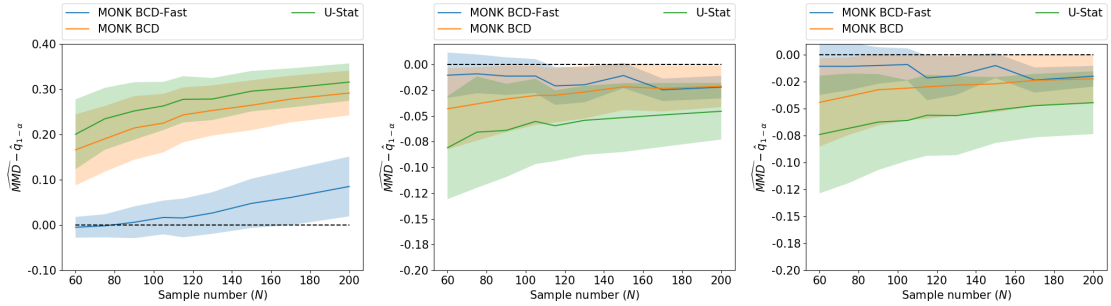
U-Stat	$O\big(n^2\big)$
MMD BCD	$O\big(n^3 + T\big[n^2 + K\log(K)\big]\big)$
MMD BCD-Fast	$O\Big(\frac{n^3}{K^2} + T\Big[\frac{n^2}{K} + K\log(K)\Big]\Big)$

where only one block is considered in which case some simplifications in the optimization reduce the problem to the computation of a U-statistic.

The MMD distance is used for example in two sample testing, the numerical application that we propose is on a biological dataset. We chose a DNA benchmark from the UCI repository, the Molecular Biology (Splice-junction Gene Sequences) Data Set. The dataset consists of 3190 instances of 60-character long DNA sub-sequences. The problem is to recognize, given a sequence of DNA, the boundaries between exons (the parts of the DNA sequence retained after splicing) and introns (the parts of the DNA sequence that are spliced out). This task consists of two sub-problems, identifying the exon/intron boundaries (referred to as EI sites) and the intron/exon boundaries (IE sites). We took 1532 of these samples by selecting 766 instances from both the EI and the IE classes (the class of those being neither EI nor IE is more heterogeneous and thus we dumped it from the study), and investigated the discriminability of the EI and IE categories. We represented the DNA sequences as strings, chose $\mathcal{K}$ as the String Subsequence Kernel to compute MMD, and performed two-sample testing based on MMD using the MONK BCD, MONK BCD-Fast and U-Stat estimators.

The aggregated values of $\widehat{MMD}(EI, IE) - \widehat{q}_{1-\alpha}$, $\widehat{MMD}(EI, EI) - \widehat{q}_{1-\alpha}$ and $\widehat{MMD}(IE, IE) - \widehat{q}_{1-\alpha}$ are summarized in Figure 2.13 where $\widehat{q}_{1-\alpha}$ is the estimated $(1-\alpha)$ quantile via $B = 150$ bootstrap permutations. In the ideal case, $\widehat{MMD} - \widehat{q}_{1-\alpha}$ is positive (resp. negative) in the inter-class (resp. intra-class) experiments. As Figure 2.13 shows all three techniques are able to solve the task both in the inter-class (when the null hypothesis does not hold) and in the intra-class experiment (null holds) and they converge to a good stable performance.

One other major advantage of our method, in addition to being robust is the complexity. The time of computation is often a bottleneck in kernel methods when the dataset gets big and the kernel is long to compute (which is the case of the sub-sequence string kernel). In this context, MONK BCD-Fast is especially adapted because it performs faster than the U-Stat approach. For example, taking all the samples (n=766) in the DNA benchmark with $K = 15$, computing MONK BCD-Fast takes $32s$ while U-Stat takes $1m28s$ on the same computer.



(a) Inter-class: EI-IE (b) Intra-class: EI-EI (c) Intra-class: IE-IE

Figure 2.13: Intra-class and inter-class MMD estimates as function of the sample size compared to the bootstrap estimated $(1-\alpha)$-quantile. Are plotted mean±std. Notice the different scales in the inter-class and intra-class experiments.

Chapter 3

Robustness to outliers and concentration of M-estimators by means of influence function

Abstract

We present a new analysis of M-estimators of locations parameters in $\mathbb{R}^d$ using their influence function and we investigate in particular the robustness of M-estimators whose influence function is not bounded. First we control the deviations of an M-estimator using the deviations of its influence function obtaining concentration inequalities for M-estimators, then we show that in a Huber contamination setting, under mild assumptions on the outliers distribution, we still have a consistency even when the influence function is unbounded (this extend Hampel's result [Ham71]). Finally, we illustrate this theory on numerical examples and in particular, we exhibit examples in which M-estimators with unbounded influence function are more efficient than Huber's estimator in a given corrupted setting.

3.1 Introduction

One of the first tasks considered in robustness theory has been to compute so-called locations estimators meant to exhibit a central tendency of the data. Let $X \sim P$ for some P probability on $\mathbb{R}^d$, let ρ be an increasing function from $\mathbb{R}_+$ to $\mathbb{R}_+$, we are interested in estimating the location parameter $T(P)$ defined by

$$T(P) \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \mathbb{E}[\rho(\|X - \theta\|)] = 0, \tag{3.1.1}$$

or alternatively, if ρ is smooth enough (which will be the case in this article), we define $T(P)$ by

$$\mathbb{E}\left[\frac{X - T(P)}{\|X - T(P)\|}\psi(\|X - T(P)\|)\right] = 0, \tag{3.1.2}$$

where $\psi = \rho'$ is called the score function. The empirical estimator obtained by plugging the empirical density $\widehat{P}_n$ in equation (3.1.2) is called M-estimator associated with ψ, it is denoted $T(\widehat{P}_n)$ and computed from an i.i.d sample $X_1, \dots, X_n$ using the following equation:

$$\sum_{i=1}^{n} \frac{X_i - T(\widehat{P}_n)}{\|X_i - T(\widehat{P}_n)\|} \psi(\|X_i - T(\widehat{P}_n)\|) = 0. \tag{3.1.3}$$

This way of estimating $T(P)$ is taken from empirical risk minimization theory and a particular case of $T(P)$ is obtained when choosing $\psi(x) = x$ in which case $T(P) = \mathbb{E}[X]$ and $T(\widehat{P}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$, however it is well known that the empirical mean is not robust. A careful choice of the function ψ yield estimators that are more robust to outliers and to heavy-tailed data (see [Cat12]).

The subsequent problem is to see how the properties of ψ impact the robustness and efficiency of $T(\widehat{P}_n)$ when estimating $T(P)$. From robust statistics theory it is known that ψ is strongly related to the influence function of the associated M-estimator. The influence function is a classical tool used to quantify the robustness of an estimator, see for example [Ham74, HRRS86, HR09, Ron97] in which are derived properties such as the asymptotic variance or the breakdown point of the estimator $T(\widehat{P}_n)$ using the influence function. The influence function is the Gâteaux derivative of T evaluated in the Dirac distribution in a point $x \in \mathbb{R}^d$ and in the case of M-estimators, from [HRRS86, Eq 4.2.9 in Section 4.2C.], the influence function takes the following simple form:

$$\mathrm{IF}(x, T, P) = M_{P,T}^{-1} \frac{x - T(P)}{\|x - T(P)\|} \psi(\|x - T(P)\|), \tag{3.1.4}$$

where $M_{P,T}$ is a non-singular matrix whose explicit formula is not important for our application (an explicit formula can however be found in [HRRS86, Eq 4.2.9 in Section 4.2C.]).

The general idea is that, if the estimator is smooth enough, for example if it is Fréchet or Hadamard differentiable, see [Fer83], then one can write the following expansion

$$T(P) = T(Q) + \int_{\mathbb{R}^d} \mathrm{IF}(x, T, Q) d(P - Q)(x) + R(P, Q), \tag{3.1.5}$$

where the remainder term $R(P, Q)$ is controlled. For example, if we apply equation (3.1.5) to $Q = \hat{P}_n$ the empirical distribution, the influence function provides a first order approximation for the difference between the estimator $T(\hat{P}_n)$ and its limit $T(P)$. This technique of approximating the estimator by its influence function is also linked to the Bahadur decomposition, see [Bah66] and [HS96] for applications to M-estimators. The influence function of M-estimators is usually chosen bounded in robust statistics, in particular from [Ham71, HR09] we have that if ψ is bounded, then the influence function is bounded and T is qualitatively robust (i.e. the estimator $T(\widehat{P}_n)$ is equi-continuous, c.f. [HR09]) and have asymptotic breakdown point $1/2$. On the other hand if ψ is unbounded, then $T(\widehat{P}_n)$ is not qualitatively robust, the influence function is not bounded and the asymptotic breakdown point is zero. From Hampel's Theorem [HR09, Theorem 2.21] we also have that ψ is bounded if and only if T is a continuous functional with respect to the Levy metric. More generally, the influence function has been used in a lot of works on asymptotic robustness, see [HRRS86, HR09] or [Ham74, Ron97].

The influence function has also been used recently in Machine Learning literature in order to have a model selection tool specialized in robustness, see for example [DHS08], [KL17] and the closely related tool of leave one out error [EP02]. The field of Robustness in Machine Learning has been very active in the last few years, in particular after several works by Olivier Catoni and

co-authors in [Cat12, CG17], the goal being to prove non-asymptotic deviation bounds when the data are more heavy-tailed than what is usually considered in classical Machine Learning. This line of thought has been continued in a number of articles, in particular [DLLO16] introduced some general concept of sub-Gaussian estimators that have been then used successfully in other applications, see [CLL19b, DK19, Cat12, LM19a, ZBFL18, MM19]. See also some comprehensive lecture notes on the subject in [Ler19].

It is interesting to note that contrary to works from classical robust theory from the 70's, the influence functions of the M-estimators used by Catoni are not bounded. In this article, we initiate the analysis of the effect of unbounded influence function on the robustness of M-estimators, Huber [Hub64] told us that the influence function must be bounded while Catoni use unbounded influence function and he still shows robust properties for this type of estimator, the difference is in their vision of what is a robust estimator.

There will be two parts in our analysis of this problem, first we develop Catoni's non-asymptotic analysis of M-estimators as we analyze M-estimators with more general influence functions using the properties of the influence function. Second we investigate asymptotic results and we show that under mild assumptions on the outliers, we still have consistency of the estimator even when the influence function is not bounded. Finally, we present numerical experiments showing the performance of M-estimators with unbounded influence function as well as some advances in choosing the scaling parameter of M-estimators.

More precisely, In Section 3.3, we show that concentration inequalities for M-estimators derive from concentration inequalities on the influence function by showing roughly that

$$\|T(\widehat{P}_n) - T(P)\| \simeq \left\| \frac{1}{n}\sum_{i=1}^{n} \frac{X_i - T(P)}{\|X_i - T(P)\|}\psi(\|X_i - T(P)\|) \right\|. \tag{3.1.6}$$

From equation (3.1.4), the right hand side of equation (3.1.6) can be interpreted as the deviation probability of the influence function. The right hand side of equation (3.1.6) can be controlled under classical assumptions, for example in dimension 1, if ψ is bounded by $\beta > 0$ (huber estimator), we can use Hoeffding or Bernstein inequality to get a control on $\|T(\widehat{P}_n) - T(P)\|$. Using Hoeffding inequality, we obtain a concentration rate similar to the rate of the empirical mean on Gaussian data.

$$\mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^{n} \psi(X_i - T(P)) \right| \geq \frac{\beta}{\sqrt{n}}\lambda \right) \leq e^{-2\lambda^2}.$$

Remark that this gives us a concentration around $T(P)$, but when the density is symmetric (which implies that $T(P) = \mathbb{E}[X]$), our result implies that the concentration of $T(\widehat{P}_n)$ around the mean is as fast as the concentration of the influence function, see Corollary 3. In particular, for Huber estimator this result implies that $T(\widehat{P}_n)$ can concentrate around $\mathbb{E}[X]$ at a $O(1/\sqrt{n})$ rate even if P does not have a finite second moment. It is known that a finite second moment is necessary in general to obtain $O(1/\sqrt{n})$ concentration around the expectation [DLLO16], but this result shows that the symmetry of P allows to overcome this limit and a finite second moment is not necessary if we want to control the deviations when the distribution is symmetric.

In a second part (Section 3.4), we investigate asymptotic Robustness results. Robustness is about dealing with deviations from the hypothesis usually supposed in statistics. One way to define deviations from hypothesis is to use a distance between distribution and say that a sample

is corrupted if instead of coming from a distribution P that is easy to handle (like a Gaussian distribution), it comes from a distribution Q such that $d(P,Q)$ is small. Depending on which distance d we use, we get different definitions of acceptable deviations from the hypothesis. The usual choice for d in robust statistics is the Total Variation distance or the Prokhorov distance. We say that T is robust if it is continuous with respect to such a distance. However, for Prokhorov distance for example, Hampel Theorem [HR09, Theorem 2.21] implies that if T is continuous, then its influence function is bounded and this is a rather conservative choice of robust estimators.

There has already been some works defining corruption using other distances like the Wasserstein distance, as in the concept of resilience introduced in [SCV17]. In this paper, we define a new family of distances to model different types of data corruption. Once these distances are defined, we study the continuity of operators with respect to these distances and this allows us to establish a class of acceptable outliers for an estimator. We show that a well-chosen Wasserstein distance can be used to derive continuity properties on M-estimators with unbounded influence function and therefore prove asymptotic distributional robustness properties for such estimators. A corollary of this result (Corollary 9) is the following, based on a sample $X_1,\dots,X_n$ containing k_n outliers and $n-k_n$ i.i.d random variables with common distribution P, let ψ be a bijection, then if the outliers are located (i.e. sampled from a dirac distribution) at $g(n)u$ for some function $g:\mathbb{N}\mapsto\mathbb{R}$ and $u\in\mathbb{R}^d$, $\|u\|=1$, then

$$\frac{k_n\psi^{-1}(g(n))}{n}\xrightarrow{n\to\infty}0 \quad\Rightarrow\quad \left\|T\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right)-T(P)\right\|\xrightarrow[proba.]{n\to\infty}0. \tag{3.1.7}$$

This is a condition on the amplitude of the outliers and their number for the estimator to converge in probability and this result is more general than hampel result in the sense that we don't limit ourselves to bounded ψ function but we also state the result for unbounded ψ for which we have to ask the outliers to verify a mild condition (right hand side of equation (3.1.7)). For example, if ψ grows like a logarithm at infinity (this is the case of Catoni's estimator), then as long as there are a finite number of outliers that are smaller than $o(\exp(n))$ then Catoni's estimator will converge. In practice, unbounded score functions are a bit more efficient than bounded score function M-estimators when the outliers verify these mild assumptions.

In Section 3.5, we present a short numerical study of M-estimators and particularly M-estimators with unbounded score function on corrupted and heavy-tailed datasets. Our algorithm is very fast but does not give, to our knowledge, minimax optimal solutions in high dimensions. In this part, we are interested in particular in three M-estimators corresponding to three different ψ function: ψ bounded (Huber's estimator), ψ with logarithmic growth (Catoni's estimator) and finally ψ with a growth like $x^{1/p}$ for some $p>1$ that we call Polynomial estimator. For all these estimators we have to tune a scale hyper-parameter such as β for Huber's estimator and p for the Polynomial estimator. We present an algorithm that allows us to automatically choose these hyper-parameters in an adaptive fashion. In particular, we show that provided that the hyper-parameter are well tuned, the Polynomial estimator can adapt to the distribution with a higher p when the distribution is very heavy-tailed and a small p if the distribution has a lot of finite moment, we exhibit a dataset on which the Polynomial estimator is better than both Huber's and Catoni's estimator.

3.2 Setting and Notations

3.2.1 Setting

We consider the functional T defined by

$$\mathbb{E}\left[\frac{X - T(P)}{\|X - T(P)\|}\psi(\|X - T(P)\|)\right] = 0, \tag{3.2.1}$$

for some $\psi : \mathbb{R}_+ \to \mathbb{R}_+$, the existence and unicity of T is discussed in Lemma 3. We are interested in the behavior of the associated M-estimator $T(\widehat{P}_n)$ defined by

$$\sum_{i=1}^{n} \frac{X_i - T(\widehat{P}_n)}{\|X_i - T(\widehat{P}_n)\|}\psi(\|X_i - T(\widehat{P}_n)\|) = 0. \tag{3.2.2}$$

Assumptions 1. *ψ is continuous, derivable, non-decreasing, concave on $\mathbb{R}_+$, $\psi(0) = 0$, and there exist $\beta, \gamma > 0$ such that*

$$\forall x \geq 0, \qquad 1 \geq \psi'(x) \geq \gamma \mathbb{1}\{x \leq \beta\}.$$

where $\mathbb{1}$ is the indicator function.

By concavity, if ψ is not identically zero, there are always a couple of positive constants β, γ such that Assumptions 1 holds. For our results to hold we will ask that β and γ are not too small. A first result that can be derived from Assumptions 1 and some additional assumptions is that our problem is well defined. This is formalized in the following lemma whose proof is in Section 3.7.2.

Lemma 3. *Let ψ satisfy Assumptions 1, define $\rho : x \mapsto \int_0^x \psi(t)dt$ and let X satisfy $\mathbb{E}[\rho(\|X - \mathbb{E}[X])\|)] < \rho(\beta)$, then $T(P)$ defined by equation (3.2.2) exists and is unique.*

In the whole article, we will suppose that $T(P)$ is unique, we do not necessarily suppose the assumptions of Lemma 3 as they are not minimal assumptions for unicity and existence of $T(P)$.

Assumptions 2. *$T(P)$ defined by equation (3.2.2) and the associated empirical estimator $T(\widehat{P}_n)$ exist and are unique.*

Assumptions 1 and Assumptions 2 will be supposed true. The behavior of ψ at 0 allows us to control the deviations of the estimator using the influence function, see Section 3.3 and it is also important to control the bias of the resulting estimator, see Section 3.3.1. On the other hand, the growth rate of ψ at $+\infty$ is central to derive concentration bounds of $T(\hat{P}_n)$, as will become clear all along Section 3.3 and Section 3.4. Assumptions 1 do not always apply to M-estimators, for example the sample median is not an estimator derived from a function ψ satisfying these assumptions. On the other hand, we provide three examples of score functions satisfying Assumptions 1, with three different growth rates when x goes to infinity.

Huber's estimator. Let $\beta > 0$. For all $x \geq 0$, let

$$\psi_H(x) = x\, \mathbb{1}\{x \leq \beta\} + \beta\, \mathbb{1}\{x > \beta\}. \tag{3.2.3}$$

In dimension 1, the M-estimator constructed from this score function is called the Huber's estimator [Hub64].

Catoni's estimator. Let $\beta > 0$. For all $x \geq 0$, let

$$\psi_C(x) = \beta \log\left(1 + \frac{x}{\beta} + \frac{1}{2}\left(\frac{x}{\beta}\right)^2\right). \tag{3.2.4}$$

The associated M-estimator is one of the estimators considered by Catoni in [Cat12]. We call the resulting M-estimator Catoni's estimator.

Polynomial estimator. Let $p \in \mathbb{N}^*$, $\beta > 0$. For all $x \geq 0$, let

$$\psi_P(x) = \frac{x}{1 + \left(\frac{x}{\beta}\right)^{1-1/p}}. \tag{3.2.5}$$

We call Polynomial estimator the M-estimator obtained using this score function.

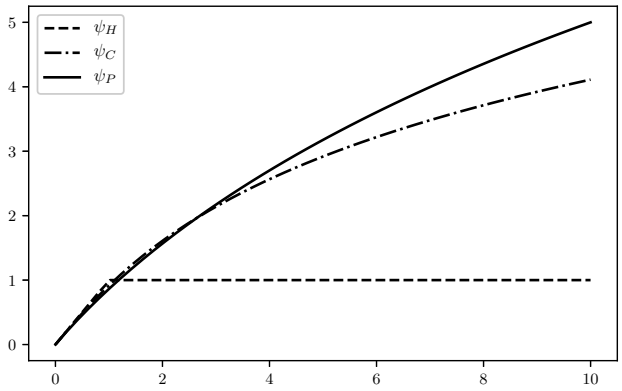


Figure 3.1: Plot of ψ_H and ψ_C for $\beta = 1$. ψ_P is plotted for $\beta = 10$ and $p = 5$.

The following result shows that the score functions from the previous three examples satisfy Assumptions 1.

Lemma 4. *For all $x \geq 0$, we have*

$$\psi_H'(x) = \mathbb{1}\{x \leq \beta\},$$

$$\psi_C'(x) \geq \frac{4}{5}\mathbb{1}\{x \leq \beta\},$$

$$\psi_P'(x) \geq \frac{1}{4}\left(1 + \frac{1}{p}\right)\mathbb{1}\{x \leq \beta\}.$$

The proof of Lemma 4 is postponed to Section 3.7.2.

3.2.2 Notations

Let $\mathcal{P}$ denote the set of probability distributions on $\mathbb{R}^d$, $S^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ where $\|\cdot\|$ is the Euclidean norm. For any $\psi : \mathbb{R}_+ \to \mathbb{R}_+$, let $\mathcal{P}_\psi = \{P \in \mathcal{P} : \mathbb{E}_P[\psi(\|X\|)] < \infty\}$.

Let $X, X_1, \ldots, X_n$ denote i.i.d random variables such that $X \sim P \in \mathcal{P}$. Let $\widehat{P}_n$ denotes the empirical distribution given by $\widehat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ where, for any $x \in \mathbb{R}^d$, δ_x is the Dirac distribution in x.

For any $h : \mathbb{R}^d \to \mathbb{R}$ and $\psi : \mathbb{R}_+ \to \mathbb{R}_+$, we denote $h \preceq \psi$ if

$$\forall x, y \in \mathbb{R}^d : \; h(x) - h(y) \leq \psi(\|x - y\|). \tag{3.2.6}$$

Let T_H, T_C and T_P denote the functionals such that, for ψ_P,ψ_H and ψ_C defined respectively in (3.2.3), (3.2.4) and (3.2.5),

$$\mathbb{E}\left[\frac{X - T_H(P)}{\|X - T_H(P)\|}\psi_H(\|X - T_H(P)\|)\right] = 0 \quad \text{and} \quad \mathbb{E}\left[\frac{X - T_C(P)}{\|X - T_C(P)\|}\psi_C(\|X - T_C(P)\|)\right] = 0$$

$$\text{and} \quad \mathbb{E}\left[\frac{X - T_P(P)}{\|X - T_P(P)\|}\psi_P(\|X - T_P(P)\|)\right] = 0.$$

Define the following variance terms

$$V_H = \mathbb{E}[\psi_H(\|X - T_H(P)\|)^2], \quad \sigma_H^2 = \left\|\mathbb{E}\left[\frac{(X - T_H(P))(X - T_H(P))^T}{\|X - T_H(P)\|^2}\psi_H(\|X - T_H(P)\|)^2\right]\right\|_{op}.$$

and similarly for V_C, σ_C^2, V_P and σ_P^2. These variance terms are to be compared with $Tr(\Sigma)$ and $\|\Sigma\|_{op}$ in the Gaussian setting, Hanson-Wright inequality tells us that $Tr(\Sigma)$ and $\|\Sigma\|_{op}$ describe the spread of the empirical mean in high dimension. Here we are not in a Gaussian setting and for example in the case of Huber's estimator, V_H and σ_H^2 will describe the spread of the influence function of Huber's estimator.

3.3 Tail probabilities of M-estimator and Influence function

3.3.1 Concentration inequalities on T using the influence function, case $d = 1$

Main result

For simplicity, we begin with a description of our results in dimension 1, the multidimensional case is treated in Section 3.3.2. The first main result of the paper compares the tail probabilities of $|T(\widehat{P}_n) - T(P)|$ and those of the influence function.

Theorem 18. *Let ψ satisfies Assumptions 1 and suppose Assumptions 2 are satisfied. Define $\psi_{odd}(x) = \text{sign}(x)\psi(|x|)$, then the following holds.*

- *For all* $\lambda > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\psi_{odd}(X_i - T(P))\right| > 3\lambda\right) \leq \mathbb{P}\Big(\Big|T(\widehat{P}_n) - T(P)\Big| > \lambda\Big).$$

- *If moreover* $V = \mathbb{E}[\psi(|X - T(P)|)^2] \leq \psi(\beta/2)^2/2 < \infty$, *then for all* $\lambda \in (0, \beta/2)$,

$$\mathbb{P}\Big(\Big|T(\widehat{P}_n) - T(P)\Big| > \lambda\Big) \leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\psi_{odd}(X_i - T(P))\right| > \frac{\lambda\gamma}{4}\right) + e^{-n\gamma^2/8}. \tag{3.3.1}$$

The proof of this result is given in Section 3.7.1. In Theorem 18 we managed to control the deviations of $T(\widehat{P}_n)$ using the deviations of a sum of i.i.d random variables and moreover these random variables are given using the function ψ_{odd} which is in general smaller than the identity at infinity. For example, in the case of Catoni's estimator, $\psi_{odd}(x)$ is logarithmic when x goes to infinity, hence even if X is heavy tailed, $\psi_{odd}(|X - T(P)|)$ might be light tailed. Then, Theorem 18 allows us to easily derive sharp concentration inequalities for M-estimators that are defined implicitly.

Theorem 18 holds if $V \leq \psi(\beta/2)^2/2$, this is a condition on the variance of the points with low error, indeed for example with Huber's estimator;

$$V = \mathbb{E}\big[\min(|X - T(P)|^2, \beta^2)\big]. \tag{3.3.2}$$

This choice of β is closely linked to robust scale estimators (see [HR09]) and in particular Huber's second proposal [Hub64] which also makes use of ψ^2 to find β. In this article, we limit ourselves to the simple estimation of location estimator and not of simultaneous estimation of scale and location as it is done for Huber's second proposal, doing so could be an extension of this work.

Notice that although our work is on location estimators, it can be extended to scale estimators by finding the location estimator of $\log((X - T(P))^2)$ as suggested in [Hub64].

Examples

Theorem 18 allows us to study the deviations of $T(\widehat{P}_n)$ using the deviations of a sum of i.i.d centered random variables. We illustrate that we can then easily get sub-gaussian concentration as defined in [DLLO16] for Huber and Catoni's estimator. For simplicity's sake, we don't always search for the best constants we can find in the bounds, better bounds are found and discussed in Section 3.3.2. We use concentration inequalities from [BLM13] and in particular Bernstein's inequality (Theorem 2.10 of [BLM13] reminded in Section 3.7.3 for completness).

Huber's estimator: in the case of Huber's estimator, ψ is bounded and from Lemma 4, $\gamma = 1$.
Because ψ_H is bounded by β, we have directly from Bernstein inequality for all $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\text{sign}(X - T_H(P))\psi_H(|X - T_H(P)|)\right| > \sqrt{\frac{2V_H t}{n}} + \frac{\beta t}{n}\right) \leq 2e^{-t}.$$

Then, by Theorem 18, if $V_H \leq \beta^2/8$, for all $t > 0$ such that $4\sqrt{2V_H t/n} + 4\beta t/n \leq \beta/2$,

$$\mathbb{P}\left(\left| T_H(\widehat{P}_n) - T_H(P) \right| > 4\sqrt{\frac{2V_H t}{n}} + 4\frac{\beta t}{n} \right) \leq 2e^{-t} + e^{-n/8}. \tag{3.3.3}$$

Remark that choosing $\beta = \sqrt{V_H}$ gives us a Sub-Gaussian concentration around $T_H(P)$, this is similar to the concentrations inequalities introduced in [DLLO16] except that we concentrate around $T_H(P)$ instead of $\mathbb{E}[X]$. Remark also that the condition $V_H \leq \beta^2/8$ is rather weak because we already have $V_H \leq \beta^2$, the condition asks that there is enough weight in the interval $[-\beta, \beta]$.

Catoni's estimator: in the case of Catoni's estimator, from Lemma 4, $\gamma = 4/5$. We use the following elementary inequality: for all $x > 0$ and $q \in \mathbb{N}$,

$$\log(1+x)^q \leq q!x.$$

To prove this, one can proceed by induction on q and use the variations of the function. Then, we have access to bounds on the moments of $\psi_C(|X - T_C(P)|)$, we have for all $q \in \mathbb{N}$,

$$\mathbb{E}[\psi_C(|X - T_C(P)|)^q] \leq q!\beta^q \mathbb{E}\left[\frac{|X - T_C(P)|}{\beta} + \frac{(X - T_C(P))^2}{2\beta^2} \right]$$

Hence, by Bernstein inequality, denoting $v = \mathbb{E}\big[\beta|X - T_C(P)| + (X - T_C(P))^2/2\big]$

$$\mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^{n} \operatorname{sign}(X - T_C(P))\psi_C(|X - T_C(P)|) \right| > \sqrt{\frac{2vt}{n}} + \frac{\beta t}{n} \right) \leq 2e^{-t}.$$

Then, by Theorem 18, if $V_C \leq \beta^2 \log(13/8)^2/2$, for all $t > 0$ that satisfy $5\sqrt{2vt/n} + 5\beta t/n \leq \beta/2$, we have

$$\mathbb{P}\left(\left| T_C(\widehat{P}_n) - T_C(P) \right| > 5\sqrt{\frac{2vt}{n}} + 5\frac{\beta t}{n} \right) \leq 2e^{-t} + e^{-2n/25}. \tag{3.3.4}$$

Remark that once again choosing $\beta = \sqrt{v}$ gives us a Sub-Gaussian concentration around $T_C(P)$.

Polynomial estimator: in the case of the Polynomial estimator, from Lemma 4, $\gamma = (1 + 1/p)/4 \geq 1/4$. Then, by Chebychev inequality, we have for all $t > 0$,

$$\mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^{n} \operatorname{sign}(X - T_P(P))\psi_P(|X - T_P(P)|) \right| > t \right) \leq \frac{V_P}{nt^2}.$$

Hence, By Theorem 18, if $V_P \leq \beta^2/(2(1 + 2^{1/p-1})^2)$, for all $\lambda \in (0, \beta/2)$, we have

$$\mathbb{P}\Big(|T_P(\widehat{P}_n) - T_P(P)| > \lambda \Big) \leq 256\frac{V_P}{n\lambda^2} + e^{-n/128}.$$

This result is not optimal and we stated it like that for simplicity, for a more detailed analysis, see Section 3.3.3.

Bias of Huber's estimator in 1D and link with previous results

In most applications in Machine Learning, we don't want to estimate $T(P)$, instead we want to estimate $\mathbb{E}[X]$, then $T(\widehat{P}_n)$ is a biased estimator and in addition to Theorem 18 we may want to bound the bias. This would also allow us to compare our results to the results obtained in other articles as most of them try to estimate $\mathbb{E}[X]$.

For simplicity, we consider only the case of Huber's estimator in dimension 1. From [ZBFL18, Proposition A.1.], we have a control of the bias of Huber's estimator if X has a finite variance σ^2. If $\beta \geq 8\sigma/\log(2)^2$, we have

$$|T_H(P) - \mathbb{E}[X]| \leq 2\frac{\sigma^2}{\beta}.$$

This bound tends to 0 when β tends to infinity and from equation (3.3.3), for all for all $t > 0$ such that $4\sqrt{2V_H t/n} + 4\beta t/n \leq \beta/2$,

$$\mathbb{P}\left(\left|T_H(\widehat{P}_n) - T_H(P)\right| > 4\sqrt{\frac{2V_H t}{n}} + 4\frac{\beta t}{n}\right) \leq 2e^{-t} + e^{-n/8}.$$

Hence,

$$\mathbb{P}\left(\left|T_H(\widehat{P}_n) - \mathbb{E}[X]\right| > 4\sqrt{\frac{2V_H t}{n}} + 4\frac{\beta t}{n} + 2\frac{\sigma^2}{\beta}\right) \leq 2e^{-t} + e^{-n/8}.$$

Similarly to [Cat12], take $\beta = \sigma\sqrt{n/t}$ which yields

$$\mathbb{P}\left(\left|T_H(\widehat{P}_n) - \mathbb{E}[X]\right| > 4\sqrt{\frac{2V_H t}{n}} + 6\frac{\sigma\sqrt{t}}{\sqrt{n}}\right) \leq 2e^{-t} + e^{-n/8}.$$

This rate of convergence is similar to [Cat12, Proposition 2.4] because it can be shown that V_H is smaller than σ^2 (using equation (3.3.2)). Our result decouples the effect of the bias and the effect of the spread of the estimator.

Notice that the bound $|T_H(P) - \mathbb{E}[X]| \leq 2\frac{\sigma^2}{\beta}$ does not take into account the symmetry of X. Actually, if X is symmetric then $|T_H(P) - \mathbb{E}[X]|$ is zero for all β and equation (3.3.3) implies the following result.

Corollary 3. *Assume that X has a finite first moment and is symmetric around $\mathbb{E}[X]$ and that $8\mathbb{E}[\psi_H(\|X - \mathbb{E}[X])\|)^2] \leq \beta^2 < \infty$.*
For all $t > 0$ such that $4\sqrt{2V_H t/n} + 4\beta t/n \leq \beta/2$,

$$\mathbb{P}\left(\left|T_H(\widehat{P}_n) - \mathbb{E}[X]\right| > 4\sqrt{\frac{2V_H t}{n}} + 4\frac{\beta t}{n}\right) \leq 2e^{-t} + e^{-n/8}.$$

Corollary 3 shows that M-estimators concentrate at rate $O(1/\sqrt{n})$, when P is symmetric, even if it does not have a finite second moment. It is known that no estimator concentrates around the expectation at rate $O(1/\sqrt{n})$, for all distributions $P \in \mathcal{P}$ with infinite second moment (see the result in [DLLO16] reminded in Section 3.7.3), from Corollary 3, it is not surprising to read in [DLLO16] that the proof of this result relies on a very asymmetric distribution.

3.3.2 Concentration inequalities on T using the influence function, case $d \geq 1$

In this section, we study the same problem as Section 3.3.1 but for a dimension greater than 1.

Definition 1. *We call t_T and t_{IF} the tail probability functions defined by*

$$t_T(\lambda) := \mathbb{P}\Big(\|T(\widehat{P}_n) - T(P)\| \geq \lambda\Big) \tag{3.3.5}$$

$$t_{IF}(\lambda) := \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{X_i - T(P)}{\|X_i - T(P)\|}\psi(\|X_i - T(P)\|)\right\| \geq \lambda\right). \tag{3.3.6}$$

The main theorem of Section 3.3 is the following.

Theorem 19. *If ψ satisfies Assumptions 1 and Assumptions 2 are satisfied, then the following holds.*

- *For all $\lambda > 0$,*
$$t_{IF}(3\lambda) \leq t_T(\lambda).$$
- *If moreover $V = \mathbb{E}[\psi(\|X - T(P)\|)^2] \leq \psi(\beta/2)^2/2 < \infty$, then for all $\lambda \in (0, \beta/2)$,*
$$t_T(\lambda) \leq t_{IF}(\lambda\gamma/4) + e^{-n\gamma^2/8}. \tag{3.3.7}$$

The proof of this result is given in Section 3.7.1.

Because of the factor 3, the lower bound on t_T is not tight. With a careful analysis of the proof, one could make this factor close to 1, but as this inequality is not the most important of the theorem, it is presented in this simplified form. Remark also that in the upper bound on t_T, we could weaken the condition to $\mathbb{E}[\psi(\|X - T - P\|)] \leq \psi(\beta/2)/2$ by weakening the Markov inequality used in equation (3.7.9) In practical examples, to get some concentration on a sum of i.i.d random variables, we need a finite second moment at least and hence we ask that V be finite to get a sharp bound on t_{IF}.

3.3.3 Examples

The results we show in this section are not optimal, a more careful analysis would be necessary to obtain the correct sub-Gaussian rates similar to [CG17]. Our goal is to illustrate the use of the influence function and particularly Theorem 19 for an easy derivation of concentration for M-estimators. These examples also illustrate an interesting phenomenon derived from Theorem 19 by showing that the concentration of $T(\widehat{P}_n)$ around $T(P)$ can be much faster than the concentration of $T(\widehat{P}_n)$ around $\mathbb{E}[X]$, and as such even though the rates are not optimal, they can be much faster than usual sub-gaussian rates in [CG17] because the variance term is not $Tr(\Sigma)$ but V which can be a lot smaller than $Tr(\Sigma)$. We use the following corollary of [A+08, Theorem 4] recalled in Section 3.7.3.

Corollary 4. *Let $Y_1, \dots, Y_n$ be i.i.d random variables taking values in $\mathbb{R}^d$, centered with covariance matrix Σ, and such that the Orlicz norm of Y is finite:*

$$\|Y\|_{\psi_1} = \inf\{\lambda > 0 : \mathbb{E}[\exp(\|Y\|/\lambda) - 1] \leq 1\} < \infty.$$

There exists an universal constant $C > 0$ such that, for all $t \geq 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n Y_i\right\| \geq \frac{3}{2}\sqrt{\mathbb{E}\left[\|Y\|^2\right]n} + 2\sqrt{nt\|\Sigma\|_{op}} + Ct\| \max_{1\leq i\leq n} \|Y_i\|\|_{\psi_1}\right) \leq 4\exp(-t). \tag{3.3.8}$$

Proof. From [A+08, Theorem 4] and because $\|Y\| = \sup_{\|u\|=1}\langle Y, u\rangle$, there exists an absolute constant C_1 such that, for all $t \geq 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n Y_i\right\| \geq \frac{3}{2}\mathbb{E}\left[\left\|\sum_{i=1}^n Y_i\right\|\right] + t\right) \leq \exp\left(-\frac{t^2}{4n\sigma^2}\right) + 3\exp\left(-\frac{t}{C_1\|\max_{1\leq i\leq n}\|Y_i\|\|_{\psi_1}}\right).$$

where $\sigma^2 = n\sup_{u\in S^{d-1}}\mathbb{E}[\langle X, u\rangle^2]$. Remark that σ^2 can be rewritten

$$\sigma^2 = n\sup_{u\in S^{d-1}} u^T\mathbb{E}[XX^T]u = n\|\Sigma\|_{op}. \tag{3.3.9}$$

By Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\left\|\sum_{i=1}^n Y_i\right\|\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^n Y_i\right\|^2\right]^{1/2} = \sqrt{n}\mathbb{E}\left[\|Y\|^2\right]^{1/2}.$$

■

The last term in equation (3.3.8) can be handled using [vdVW96, Lemma 2.2.2] from which we get that there exists an absolute constant $K > 0$ such that

$$\left\|\max_{1\leq i\leq n}\|Y_i\|\right\|_{\psi_1} \leq K\log(n)\|Y_i\|_{\psi_1}. \tag{3.3.10}$$

However, note that Hanson-Wright's inequality for Gaussian random variables shows that this logarithm factor is not optimal. This extra logarithm factor can be removed if Y is bounded, which will be the case when we apply this result to Huber's estimator but not for the two other estimators.

In the rest of the section, we prove concentration inequalities for the estimators featured in Section 3.2 using Corollary 4 applied to $Y = \frac{X-T(P)}{\|X-T(P)\|}\psi(\|X - T(P)\|)$.

Huber's estimator

Let $\beta > 0$, and, for all $x \geq 0$, let $\psi_H(x) = x\,\mathbb{1}\{x \leq \beta\} + \beta\,\mathbb{1}\{x > \beta\}$. From Lemma 4, Assumptions 1 hold in this example with $\gamma = 1$. As ψ_H is bounded by β, Hoeffding's lemma (see [BLM13, Section 2.3]) shows that $\max_{1\leq i\leq n}|\psi_H(\|X_i - T(P)\|)\|_{\psi_1} \leq \beta$.

Hence, from Corollary 4, for all $t > 0$,

$$t_{\mathrm{IF}}\left(\frac{3V_H^{1/2}}{2\sqrt{n}} + 2\sigma_H\sqrt{\frac{t}{n}} + \frac{C}{n}t\beta\right) \leq 4e^{-t}.$$

From Theorem 19, we deduce the following corollary.

Corollary 5. *If $8\mathbb{E}[\psi_H(\|X - T(P)\|)^2] \leq \beta^2 < \infty$, there exists an absolute constant $C > 0$ such that, for all $\lambda \in (0, \lambda_{max})$, with probability larger than $1 - 4\exp(-\lambda) - \exp(-n/8)$,*

$$\left\|T_H(P) - T_H(\widehat{P}_n)\right\| \leq 6\frac{V_H^{1/2}}{\sqrt{n}} + 8\sigma_H\sqrt{\frac{\lambda}{n}} + \frac{C}{n}\lambda\beta. \tag{3.3.11}$$

Where λ_{max} is such that

$$\frac{3V_H^{1/2}}{2\sqrt{n}} + 2\sigma_H\sqrt{\frac{\lambda_{\max}}{n}} + \frac{C}{n}\lambda_{\max}\beta \leq \frac{\beta}{2}.$$

Remark that the condition on λ_{max} implies that λ is at most of order n. Therefore, in this result, the additional deviation probability $\exp(-n/8)$ is asymptotically negligible.

Catoni's estimator

Let $\beta > 0$ and, for all $x \geq 0$, let

$$\psi_C(x) = \beta\,\log\left(1 + \frac{x}{\beta} + \frac{x^2}{2\beta^2}\right).$$

From Lemma 4, ψ_C satisfies Assumptions 1 with $\gamma = 4/5$. This function satisfies the following property.

Lemma 5. *If X satisfies $\mathbb{E}[\|X\|^2] < \infty$, then, for all $q \in \mathbb{N}^*$,*

$$\mathbb{E}[\psi_C(\|X - T_C(P)\|)^q] \leq q!(s\beta)^q,$$

where

$$s = \max\left(e, \log\left(1 + \frac{\mathbb{E}[\|X - T_C(P)\|])}{\beta} + \frac{\mathbb{E}[\|X - T_C(P)\|^2]}{2\beta^2}\right)\right).$$

The proof of Lemma 5 is postponed to Section 3.7.2. Then, using the power series expansion of the exponential function, we get that, for all $t > \beta s$,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{\psi_C(\|X - T_C(P)\|)}{t}\right)\right] &= \sum_{q=0}^{\infty}\frac{\mathbb{E}[\psi_C(\|X - T(P)\|^q]}{t^q q!} \\ &\leq \sum_{q=0}^{\infty}\frac{\beta^q s^q}{t^q} = \frac{1}{1 - \beta s/t}.\end{aligned}$$

Choosing $t = 2\beta s$ shows that $\|\psi_C(\|X - T_C(P)\|)\|_{\psi_1} \leq 2\beta s$. From Corollary 4 and equation (3.3.10), it follows that, for all $\lambda > 0$,

$$t_{\mathrm{IF}}\left(\frac{V_C}{\sqrt{n}}\frac{3}{2} + 2\sigma_C\sqrt{\frac{t}{n}} + \frac{2Cs}{n}t\log(n)\beta\right) \leq 4e^{-t}.$$

By Theorem 19, we deduce the following corollary.

Corollary 6. *If $\mathbb{E}[\psi_C(\|X - T_C(P)\|)^2] \leq \beta^2 \ln(13/8)^2/2$, then there exists an absolute constant $C > 0$ such that, for all $\lambda \in (0, \lambda_{max})$, with probability larger than $1 - 4\exp(-\lambda) - \exp(-2n/25)$,*

$$\left|T_C(\widehat{P}_n) - T_C(P)\right| \leq \frac{15V_C^{1/2}}{2\sqrt{n}} + 10\sigma_C\sqrt{\frac{\lambda}{n}} + \frac{Cs}{n}\lambda\log(n)\beta, \tag{3.3.12}$$

where s is defined in Lemma 5 and λ_{max} satisfies

$$\frac{3V_C^{1/2}}{2\sqrt{n}} + 2\sigma_C\sqrt{\frac{\lambda_{max}}{n}} + \frac{Cs}{n}\lambda_{max}\log(n)\beta \leq \frac{\beta}{2}.$$

Compared to the result in dimension 1, we get a faster concentration because in most cases V_C and σ_C^2 will be of lower order of magnitude compared to the variance term used in equation (3.3.4). However, we also have that the rate of convergence is slower by a factor $log(n)$ on the last term that is due to the use of Corollary 4.

Polynomial estimator

Let $\psi_P(x) = \frac{x}{1+(x/\beta)^{1-1/p}}$. The following lemma applies.

Lemma 6. *Let $n \in \mathbb{N}^*$, suppose $X_1, \ldots, X_n$ are i.i.d. Let $q \in \mathbb{N}^*$ and suppose $\mathbb{E}[\|X\|^q] < \infty$. There exists an absolute constant $K > 0$ such that*

$$t_{IF}(\lambda) \leq \frac{\mathbb{E}[\|X - T_P(P)\|^q]}{\beta^q}\left(\frac{Kpq\beta}{\sqrt{n}\lambda}\right)^{qp}.$$

The proof is postponed to Section 3.7.2.
From Lemma 4, ψ_P satisfies Assumptions 1 with$\gamma = \frac{1}{4}\left(1 + \frac{1}{p}\right) \geq 1/4$. Hence, by Theorem 19 and Lemma 6, for all $\lambda \in (0, \beta/2)$,

$$\mathbb{P}\left(\left\|T_P(\widehat{P}_n) - T_P(P)\right\| > \lambda\right) \leq \frac{\mathbb{E}[\|X - T_P(P)\|^q]}{\beta^q}\left(\frac{16Kpq\beta}{\sqrt{n}\lambda}\right)^{qp} + \exp\left(-\frac{n}{128}\right),$$

under the condition that

$$V_P = \mathbb{E}[\psi_P(\|X - T_P(P)\|)^2] \leq \psi_P(\beta/2)^2/2 = \frac{\beta^2}{2(1 + 2^{1/p-1})^2}.$$

For example, if $p = \frac{\sqrt{n}\lambda}{16eKq\beta}$ (recall that p is a tuning parameter of the estimator and as such we can choose it as we want provided that the condition on V_P is verified) which is greater than 1 for n large enough, we get the following corollary.

Corollary 7. *If X has more than $q \geq 2$ moments and $V_P \leq \beta^2/(2(1+2^{1/p-1})^2)$, then there exists an absolute constant $C > 0$ such that for $p = \frac{\sqrt{n}\lambda}{Cq\beta}$ and for all $\lambda \in (0, \beta/2)$,*

$$\mathbb{P}\Big(\Big\|T_P(\widehat{P}_n) - T_P(P)\Big\| > \lambda\Big) \leq \frac{\mathbb{E}[\|X - T_P(P)\|^q]}{\beta^q} \exp\Big(-\frac{\sqrt{n}\lambda}{C\beta}\Big) + \exp\Big(-\frac{n}{128}\Big).$$

From Corollary 7 we get that $\Big\|T_P(\widehat{P}_n) - T_P(P)\Big\|$ is of order $O(\beta/\sqrt{n})$ with high probability, this bound is comparable to the bound of Huber's estimator because $\mathbb{E}\big[\psi_H(\|X - T_H(P)\|)^2\big] \leq \beta^2$ however this bound does not seem to be optimal as it is known that the asymptotic variance of $T_P(\widehat{P}_n)$ is $\mathrm{Var}(\mathrm{IF}(X, T, P))$ which can be smaller than β.

3.4 Some asymptotic properties derived from the influence function

In Section 3.3, we have shown results for finite sample robustness to heavy-tailed distributions. We used the influence function to show that the concentration of $T(\widehat{P}_n)$ around $T(P)$ can be bounded by the concentration of a sum of i.i.d random variables. In this section, we show asymptotic results for $T(\widehat{P}_n)$ and provide sufficient conditions on the sample to have that $\big|T\big(\frac{1}{n}\sum_{i=1}^n \delta_{X_i}\big) - T(P)\big| \to 0$.

In the classical theory of robustness, it is usually assumed that the influence function is bounded. A reason is that it is necessary and sufficient for T to be continuous with respect to the Levy metric for example. In this section we study the continuity of T with respect to other distances and we relax the boundedness assumption on the influence function.

Let $P^\varepsilon = (1-\varepsilon)P + \varepsilon H(\varepsilon)$ where ε is small and $H(\varepsilon)$ is a distribution different from P modeling the distribution of outliers. It is easy to check that the total variation distance $TV(P^\varepsilon, P) \leq \varepsilon$ for any distribution $H(\varepsilon)$ and any $\varepsilon \in [0, 1]$, from this it could be said that the total variation is blind to a small portion of outliers. No matter how far from P the distribution $H(\varepsilon)$ is, P^ε is close to P with respect to the Total variation distance if ε is close to 0.

3.4.1 Definition of a family of distance between probabilities

Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ and let $\mathcal{P}_\psi = \{P \in \mathcal{P} : \mathbb{E}_P[\psi(\|X\|)] < \infty\}$. For all $P, Q \in \mathcal{P}_\psi$, let

$$W_\psi(P, Q) = \sup_{h \preceq \psi}\Big\{\int h(x)dP(x) - \int h(x)dQ(x)\Big\}. \tag{3.4.1}$$

If ψ is non-decreasing, sub-additive, continuous, increasing in a neighborhood of 0 and $\psi(0) = 0$, then $d_\psi = \psi(\|x - y\|)$ is a distance. In this case, W_ψ is the Wasserstein-1 distance in the metric space $(\mathbb{R}^d, d_\psi)$. In particular if ψ satisfies Assumptions 1 then W_ψ is a distance. If ψ is the identity function, W_ψ is the usual Wasserstein-1 distance on $(\mathbb{R}^d, \|\cdot\|)$. If ψ is constant equal to 1, W_ψ is close to Total Variation distance (for total variation we take the supremum on all functions h such that $\sup h - \inf h \leq 1$).

We recall the following result from optimal transport theory.

Theorem 20. *Let ψ satisfy Assumptions 1, W_ψ metrizes the weak convergence in $\mathcal{P}_\psi$. In other words, if $(P_k)_{k\in\mathbb{N}}$ is a sequence of probability measures in $\mathcal{P}_\psi$ and $P \in \mathcal{P}$, then the following statements are equivalent:*

$$P_k \xrightarrow[k\to\infty]{law} P \quad \text{and} \quad W_\psi(P_k, P) \xrightarrow[k\to\infty]{} 0.$$

Refer, for example, to [Vil09, Theorem 6.9] for a proof. This theorem implies the following result.

Lemma 7. *Let ψ satisfy Assumptions 1 and let $P \in \mathcal{P}_\psi$. If $X_1, \dots, X_n$ are i.i.d with distribution P, we have*

$$W_\psi(\widehat{P}_n, P) \xrightarrow[n\to\infty]{a.s.} 0.$$

Lemma 7 follows from Theorem 20 and Glivenko-Cantelli theorem. We are now in position to use the distance W_ψ to study the asymptotic properties of T.

3.4.2 Continuity of M-estimators

We define the stability of a distance.

Definition 2. *Let $\mathcal{Q}$ be a subset of the set $\mathcal{P}$ of all probabilities. Let $H : [0,1] \to \mathcal{Q}$. A distance d is $(H, \mathcal{Q})$-stable if for any probability measure $P \in \mathcal{Q}$ we have*

$$d((1-t)P + tH(t), P) \xrightarrow[t\to 0]{} 0.$$

Then, we have the following theorem.

Theorem 21. *For all ψ satisfying Assumptions 1, W_ψ is $(H, \mathcal{P}_\psi)$-stable if*

$$t\mathbb{E}_{H(t)}[\psi(\|X\|)] \xrightarrow[t\to 0]{} 0.$$

Moreover, let $g : [0,1] \to \mathbb{R}^d$, if $H(t) = \delta_{g(t)}$ is a Dirac distribution at $g(t)$, W_ψ is $(H, \mathcal{P}_\psi)$-stable if and only if

$$t\psi(\|g(t)\|) \xrightarrow[t\to 0]{} 0.$$

This result is proved in Section 3.7.1. In dimension 1 the Kolmogorov distance and the the Total Variation (two distances usually used in robustness theory) are $(H, \mathcal{P})$-stable for any H taking values in $\mathcal{P}$, in dimension greater than 2, we can't use the Kolmogorov distance anymore but the Total variation distance is still $(H, \mathcal{P})$-stable for any H taking values in $\mathcal{P}$.

Theorem 22. *Let ψ denote a function satisfying Assumptions 1 and suppose Assumptions 2 are satisfied. Let T be the M-estimator constructed from ψ, let $P \in \mathcal{P}_\psi$ and suppose that $\psi(+\infty) > \mathbb{E}_P[\psi(\|X\|)]$ and $\|X\|$ almost surely finite.*
Then, T is continuous at P for the distance W_ψ over $\mathcal{P}_\psi$. In other words, we have for all $Q \in \mathcal{P}_\psi$,

$$\|T(P) - T(Q)\| \xrightarrow[W_\psi(P,Q)\to 0]{} 0.$$

This result is proved in Section 3.7.1. The conditions of Theorem 22 are weak. Indeed, as ψ is non-decreasing and strictly increasing near 0, if $\|X\|$ is not constant, then $\psi(+\infty) > \mathbb{E}_P[\psi(\|X\|)]$. Remark that in our definition P and Q must belong to $\mathcal{P}_\psi$ but we could still define a notion of distance $W_\psi(P,Q)$ if for $X \sim P, Y \sim Q$ and X, Y independent, we have $\mathbb{E}[\psi(\|X-Y\|)] < \infty$

Theorem 21 gives a condition on H for W_ψ to be stable which gives us a condition on H for $(1-t)P + tH(t)$ to converge to P. Theorem 22 studies the continuity of T with respect to W_ψ and this gives us the condition on H for the convergence of $T((1-t)P + tH(t))$ to $T(P)$ which is interpreted as an infinitesimal robustness of T.

3.4.3 Consistency of $T(\widehat{P}_n)$ using W_ψ in $\mathcal{I} \cup \mathcal{O}$ corruption setting

In this section, we look for statistical asymptotic properties of T.

Let $\mathcal{I}$ and $\mathcal{O}$ denote unknown subsets of $\{1, \dots, n\}$, with $\mathcal{I} \cup \mathcal{O} = \{1, \dots, n\}$ and $\mathcal{I} \cap \mathcal{O} = \emptyset$. Let $(X_j)_{j \in \mathcal{I}}$ denote an i.i.d sample from P, and let $(X_j)_{j \in \mathcal{O}}$ denote random variables with distribution $H_n \in \mathcal{P}$ (not necessarily i.i.d and not necessarily independent to the other X_i's). Denote by $|O| = k_n$ the cardinal of the set of outliers. We say that $X_1, \dots, X_n$ are sampled according to the design $(\mathcal{I}, P, H_n, k_n)$.

We have the following lemma.

Lemma 8. *Let ψ satisfy Assumptions 1 and suppose Assumptions 2 are satisfied. Let T denote the M-estimator defined in Eq (3.2.2), let $P \in \mathcal{P}_\psi$. Let $X_1, \dots, X_n$ be sampled according to the design $(\mathcal{I}, P, H_n, k_n)$. Denote by $X'_1, \dots, X'_n$ random variables such that $X'_i = X_i$ for $i \in \mathcal{I}$ and $(X'_i)_{i \in \mathcal{O}}$ are i.i.d random variables with law P, independent of $X_1, \dots, X_n$, we have*

$$W_\psi\left(\frac{1}{n}\sum_{i=1}^n \delta_{X_i}, \frac{1}{n}\sum_{i=1}^n \delta_{X'_i}\right) \leq \frac{1}{n}\sum_{i \in \mathcal{O}} \psi(\|X_i - X'_i\|). \tag{3.4.2}$$

In particular, let $O_n \sim H_n$. If

$$\frac{k_n}{n}\mathbb{E}[\psi(\|O_n\|)] \xrightarrow[n\to\infty]{} 0, \tag{3.4.3}$$

then,

$$W_\psi\left(\frac{1}{n}\sum_{i=1}^n \delta_{X_i}, P\right) \xrightarrow[n\to\infty]{prob.} 0.$$

Lemma 8 is proved in Section 3.7.2. The last part of this lemma holds even if the exact value of k_n is unknown provided that Eq (3.4.3) is satisfied. Then, the following corollary follows from the continuity of T.

Corollary 8. *Let ψ satisfy Assumptions 1 and suppose Assumptions 2 are satisfied. Let T be the M-estimator associated to ψ, let $P \in \mathcal{P}_\psi$ and suppose that $\psi(+\infty) > \mathbb{E}_P[\psi(\|X\|)]$. Let $H_n \in \mathcal{P}$. Let $X_1, \dots, X_n$ be sampled according to the design $(\mathcal{I}, P, H_n, k_n)$. Denote $d_n = \mathbb{E}_{H_n}[\psi(\|O_n\|)]$, then*

$$\left(\frac{k_n d_n}{n} \xrightarrow[n\to\infty]{} 0\right) \Rightarrow \left(T\left(\frac{1}{n}\sum_{i=1}^n \delta_{X_i}\right) \xrightarrow[n\to\infty]{P} T(P)\right).$$

Corollary 8 follows from Lemma 8 and Theorem 22. An informal interpretation of this corollary is that $T(\widehat{P}_n)$ is robust to up to k_n outliers distributed as H_n as long as $\frac{k_n}{n}\mathbb{E}_{H_n}[\psi(\|O_n\|)] \xrightarrow[n\to\infty]{} 0$.

3.4.4 Examples

We apply Corollary 8 and we particularize to the case where $H_n = \delta_{g(n)}$ for some $g : \mathbb{N} \mapsto \mathbb{R}$ to get the following lemma.

Corollary 9. *For all functions $g : \mathbb{N} \to \mathbb{R}$ increasing, for all $u \in S^d$, the following results hold.*

Huber's estimator *Let $P \in \mathcal{P}$ and suppose that $\mathbb{E}_P[\psi_H(\|X\|)] < \beta$. Let $X_1, \dots, X_n$ be sampled according to the design $(\mathcal{I}, P, \delta_{g(n)u}, k_n)$. We have*

$$\left(\frac{k_n}{n} \xrightarrow[n\to\infty]{} 0\right) \Rightarrow \left(T_H\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right) \xrightarrow[n\to\infty]{P} T_H(P)\right).$$

Catoni's estimator *Let $P \in \mathcal{P}_{\psi_C}$ and suppose that $\mathbb{E}_P[\psi_C(\|X\|)] < \infty$. Let $X_1, \dots, X_n$ be sampled according to the design $(\mathcal{I}, P, \delta_{g(n)u}, k_n)$.*

$$\left(\frac{k_n \log(g(n))}{n} \xrightarrow[n\to\infty]{} 0\right) \Rightarrow \left(T_C\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right) \xrightarrow[n\to\infty]{P} T_C(P)\right).$$

Polynomial estimator *Let $P \in \mathcal{P}_{\psi_P}$ and suppose that $\mathbb{E}_P[\psi_P(\|X\|)] < \infty$. Let $X_1, \dots, X_n$ be sampled according to the design $(\mathcal{I}, P, \delta_{g(n)u}, k_n)$.*

$$\left(\frac{k_n g(n)^{1/p}}{n} \xrightarrow[n\to\infty]{} 0\right) \Rightarrow \left(T_P\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}\right) \xrightarrow[n\to\infty]{P} T_P(P)\right).$$

A consequence of Corollary 9 is that Huber's estimator converges when the number of outliers is $o(n)$. This was already known as Huber's estimator has a non-zero asymptotic breakdown point. On the other hand, Corollary 9 shows that some assumptions on the outliers are required if we want that Catoni's estimator and the Polynomial estimator converge. This is not surprising, since both Catoni's estimator and Polynomial's estimators have a zero asymptotic breakdown point, see [HR09, Theorem 3.6].

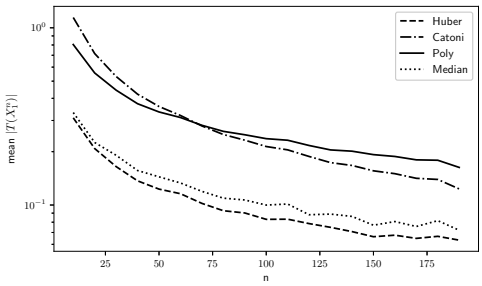
3.5 Simulations and the value of knowing the scale of outliers

To compute M-estimators, we use the iterative re-weighting algorithm, see [HR09, Section 7.8]. This famous algorithm is often used to compute M-estimators. It usually converges in a few iterations when properly initialized. For example, it can be initialized with a coordinate-by-coordinate sample median. In these simulations, we compare four estimators: sample median, Catoni's estimator, the Polynomial M-estimator and Huber's estimator.

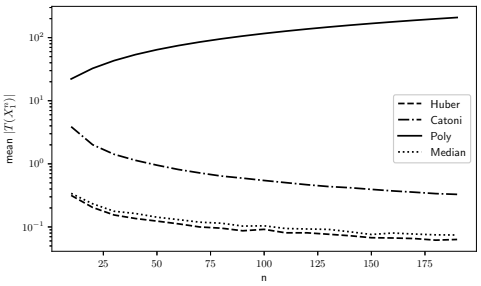
3.5.1 Gaussian corrupted simulated dataset

First, we study the behavior of our estimators when the dataset is constituted of $X_1, \dots, X_{n-1}$ i.i.d $\mathcal{N}(0,1)$ to which we add one outlier situated in respectively n^2, n^5 and $\exp(n^2/1000)$.

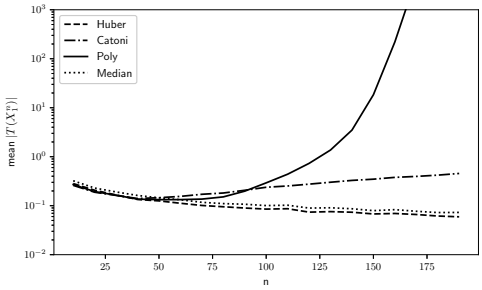
The theory predicts that for outliers of scale n^2, all estimators converge to 0, for outliers of scale n^5 the polynomial estimator with $p = 4$ does not converge but the others do and finally for outlier of scale $\exp(n^2/1000)$, only Huber's estimator and the median converges to 0, this is what is illustrated in Figure 3.2 in which we plot the Monte-Carlo mean error estimate as a function of n for the three corrupted datasets considered.



(a) One outlier in n^2



(b) One outlier in n^5



(c) One outlier in $\exp(n^2/1000)$

Figure 3.2: Plots of the mean absolute distance to 0 (log-scale) for various estimators on Gaussian dataset corrupted with one outlier as a function of the sample size n.

In these figures, the parameters are not tuned, $\beta = 1$ for all estimators.

3.5.2 Performance on heavy-tailed dataset and a corrupted simulated dataset

In this section we consider four datasets.

- Dataset 1: $X_1, \dots, X_n$ i.i.d with $X_i \sim \mathcal{N}(0,1)$ with probability $1-\varepsilon$ and $X_i = 42$ with probability ε, this is a corrupted Gaussian distribution with $\varepsilon = 0.05$.
- Dataset 2: $X_1, \dots, X_n$ i.i.d $\sim T(3)$, this is a heavy tailed symmetric distribution with $T(3)$ being a student distribution with 3 degrees of freedom.
- Dataset 3: $X_1, \dots, X_n$ i.i.d such that $X_i \sim T(10)$ with probability $1-\varepsilon$ and $5X_i \sim T(10)$ with probability ε with $\varepsilon = 0.05$.
- Dataset 4: $X_1, \dots, X_n$ i.i.d from $Pareto(5,1)$ a pareto distribution with shape parameter 5 and scale parameter 1. This is a heavy-tailed and asymmetric distribution.

Datasets 1, 2 and 3 the inliers are symmetric, hence we will compare our estimator to the center of symmetry of the inliers which is 0. Dataset 4 is asymmetric and it is meant to model a common occurrence in Machine Learning in which one wants to compute a robust estimation of the mean of some loss random value. The fact that it comes from a loss (with positive values) dictates that it must be asymmetric and we represent that with a Pareto and in a robust setting we study a Pareto with a small number of finite moments. To assess the accuracy on Dataset 4, we compare our estimators to the theoretical mean.

Because this is a toy example and we can sample at will from the initial distribution, we use Monte-Carlo simulation. Let $n = 100, M = 500$ and for a sample $(X_{i,j})_{\substack{1\leq i\leq n\\ 1\leq j\leq M}}$ we use the mean absolute error defined by

$$MAE = \frac{1}{M}\sum_{j=1}^{M}\left|T\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_{i,j}}\right) - \mathbb{E}[X]\right|,$$

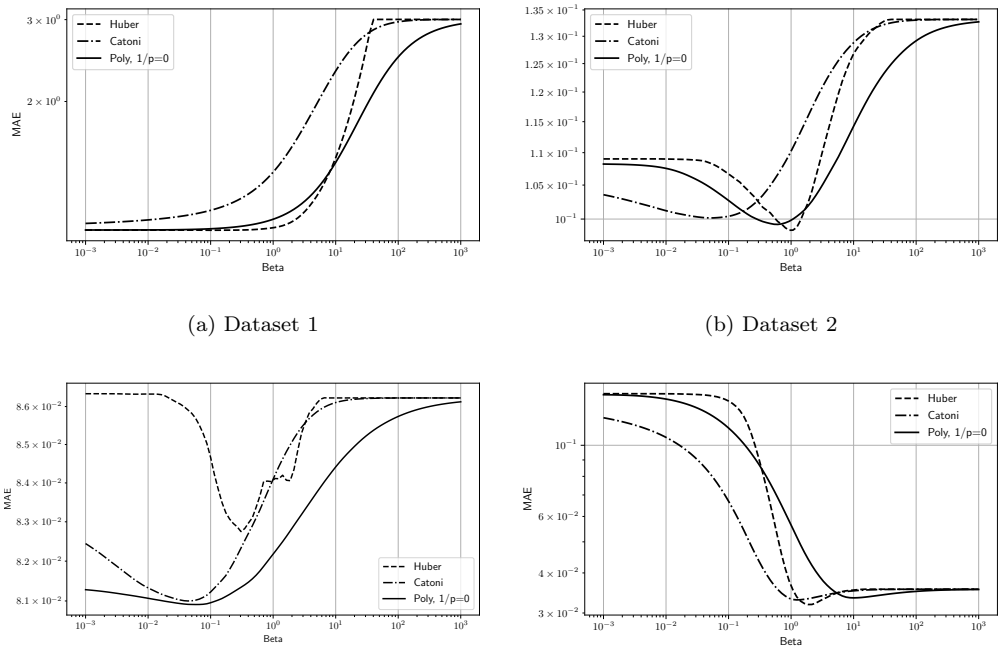
and we quantify the precision with the associated standard error:

$$SAE^2 = \frac{1}{M}\sum_{j=1}^{M}\left(T\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_{i,j}}\right) - MAE\right)^2$$

β and p are chosen optimally by computing the MAE on a grid of values of β and p, see figure 3.3 which depict the MAE as a function of β in our Datasets (for the optimal value of p when computing Polynomial estimator).

In these examples, the choice of β is through a grid-search on potential values. Catoni in his article recommended that for a distribution with variance σ^2, $\beta \simeq \sigma\sqrt{n}$. In practice, most of the time the optimal value of β to minimize the MAE will be much smaller than $\sigma\sqrt{n}$. We plot the MAE as a function of β for the different estimators and the different datasets in Figure 3.3, recall that for our datasets, $\sigma \simeq 1$ and $n = 100$ and observe in Figure 3.3 that most of the time indeed, the optimal choice for β is much smaller than $\sigma\sqrt{n} = 10$, this is due to our datasets being corrupted by outliers.

Then, using the values of β dictated by Figure 3.3, we can study the performances of our estimators. In the table Figure 3.4 we see that there is no estimator that is better in all the cases, of particular interest for our application is Dataset 3 for which the optimal parameter p for the Polynomial estimator is not infinity and the two estimators with unbounded influence function are more efficient than Huber's estimator. On the other hand, on the heavily corrupted Dataset 1 and the heavy-tailed Dataset 2 and Dataset 3, Huber's estimator outperforms the other estimators.



(a) Dataset 1

(b) Dataset 2

(c) Dataset 3

(d) Dataset 4

Figure 3.3: Log-log scale

3.6 Annex

3.7 Main Proofs

3.7.1 Proof of Theorems

Proof of Theorems 18 and 19

Theorem 18 is a particular case of Theorem 19, we will only prove the latter.
The proof is divided into two parts.

Lower bound on t_T.

By definition (Equation (3.2.2)), we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{X_i - T(\widehat{P}_n)}{\|X_i - T(\widehat{P}_n)\|}\psi(X_i - T(\widehat{P}_n)) = 0.$$

		Median	Mean	Polynomial	Catoni	Huber
Dataset 1	MAE	**1.06E+00**	3.00E+00	**1.06E+00**	1.11E+00	**1.06E+00**
	SAE	1.31E-01	8.58E-01	1.29E-01	**1.23E-01**	1.27E-01
	Parameters	NAN	NAN	1.00E-02 $1/p$ =0	1.00E-02	1.02E-01
Dataset 2	MAE	1.09E-01	1.33E-01	9.92E-02	1.00E-01	**9.84E-02**
	SAE	8.13E-02	1.00E-01	**7.46E-02**	7.49E-02	7.49E-02
	Parameters	NAN	NAN	β = 6.14E-01 $1/p$ =0	β =4.98E-02	β =9.33E-01
Dataset 3	MAE	8.63E-02	8.62E-02	**8.09E-02**	8.10E-02	8.28E-02
	SAE	6.46E-02	6.82E-02	6.17E-02	**6.16E-02**	6.30E-02
	Parameters	NAN	NAN	β = 6.58E-02 $1/p$ =2.97E-01	β = 4.33E-02	β = 2.85E-01
Dataset 4	MAE	1.45E-01	3.56E-02	3.34E-02	3.30E-02	**3.19E-02**
	SAE	2.90E-02	2.93E-02	2.51E-02	2.47E-02	**2.33E-02**
	Parameters	NAN	NAN	β =1.00E+01 $1/p$ =0	β =1.23E+00	β =1.87E+00

Figure 3.4

Define $\phi : \theta \mapsto \sum_{i=1}^n \frac{X_i-\theta}{\|X_i-\theta\|}\psi(\|X_i-\theta\|)$, by Taylor inequality on ϕ,

$$\left\|\frac{1}{n}\sum_{i=1}^n \frac{X_i - T(P)}{\|X_i - T(P)\|}\psi(X_i - T(P))\right\| \leq \|T(P) - T(\widehat{P}_n)\|\frac{1}{n}\sup_{\theta\in\mathbb{R}^d}\|d_\theta\phi\|_{\mathrm{op}}. \tag{3.7.1}$$

Where $d_\theta\phi(h)$ is the differential of ϕ in θ applied to h and $\|\cdot\|_{\mathrm{op}}$ the operator norm with respect to the Euclidean norm. Let us control the differential of ϕ. For all θ and $h \in \mathbb{R}^d$ such that $\|h\| \leq 1$,

$$\begin{aligned} d_\theta\phi(h) = \sum_{i=1}^n \left(\frac{-h}{\|X_i-\theta\|} + \frac{\langle h, X_i-\theta\rangle(X_i-\theta)}{\|X_i-\theta\|^3}\right)\psi(\|X_i-\theta\|) \\ - \frac{\langle h, X_i-\theta\rangle(X_i-\theta)}{\|X_i-\theta\|^2}\psi'(\|X_i-\theta\|). \end{aligned}$$

Then, by triangular inequality,

$$\|d_\theta\phi(h)\| \leq \sum_{i=1}^n \left(\frac{1}{\|X_i-\theta\|} + \frac{|\langle h, X_i-\theta\rangle|}{\|X_i-\theta\|^2}\right)\psi(\|X_i-\theta\|) + \frac{|\langle h, X_i-\theta\rangle|}{\|X_i-\theta\|}\psi'(\|X_i-\theta\|). \tag{3.7.2}$$

Now, by Cauchy-Schwarz inequality,

$$\|d_\theta\phi(h)\| \leq \sum_{i=1}^n 2\left(\frac{\psi(\|X_i-\theta\|)}{\|X_i-\theta\|} + \psi'(\|X_i-\theta\|)\right) \leq 3n. \tag{3.7.3}$$

To obtain the last inequality, we used $\psi'(x) \leq 1$ and $\psi(x) \leq x$ for all $x \in \mathbb{R}_+$. Plugging (3.7.3) in (3.7.1) yields the result.

Upper bound on t_T.

For all $n \in \mathbb{N}^*$, $\lambda \in \mathbb{R}$ and $u \in S^{d-1}$, the unit sphere in dimension d, let

$$f_{n,u}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \frac{\langle X_i - T(P) - \lambda u, u\rangle}{\|X_i - T(P) - \lambda u\|}\psi(\|X_i - T(P) - \lambda u\|).$$

We have the following lemma.

Lemma 9. *Suppose that the hypothesis of Theorem 19 are verified, then*

$$f'_{n,u}(\lambda) \leq -\frac{1}{n}\sum_{i=1}^{n} \psi'(\|X_i - T(P) - \lambda u\|). \tag{3.7.4}$$

In particular, because ψ' is non-negative, the function $f_{n,u}$ is non-increasing.

The proof of Lemma 9 is given Section 3.7.2. Hence, for all $n \in \mathbb{N}^*$, $\lambda \in \mathbb{R}$ and $u \in S^{d-1}$

$$\langle T(P) - T(\widehat{P}_n), u\rangle \geq \lambda \Rightarrow f_{n,u}(\langle T(P) - T(\widehat{P}_n), u\rangle) = 0 \leq f_{n,u}(\lambda).$$

And then,

$$\begin{aligned}\mathbb{P}(\|T(P) - T(\widehat{P}_n)\| \geq \lambda) &= \mathbb{P}(\exists u \in S^{d-1},\, \langle T(P) - T(\widehat{P}_n), u\rangle \geq \lambda) \\ &\leq \mathbb{P}(\exists u \in S^{d-1},\, f_{n,u}(\lambda) \geq 0).\end{aligned} \tag{3.7.5}$$

Let us show that with high probability, for all $u \in S^{d-1}$, $f_{n,u}(\lambda) \leq 0$. We apply Taylor's inequality to the function $t \mapsto f_{n,u}(t)$. As $f_{n,u}$ is non-increasing, for all $\lambda > 0$,

$$f_{n,u}(\lambda) \leq f_{n,u}(0) - \lambda \inf_{t\in[0,\lambda]} \left|f'_{n,u}(t)\right|. \tag{3.7.6}$$

Then, from equation (3.7.4),

$$|f'_{n,u}(t))| \geq \frac{1}{n}\sum_{i=1}^{n} \psi'(\|X_i - T(P) - tu\|). \tag{3.7.7}$$

The right-hand side of equation (3.7.6) is the minimum of the mean of n i.i.d random variables in $[0,1]$. Hence, the function

$$(X_1, \ldots, X_n) \mapsto \sup_{\substack{u\in S^{d-1}\\ t\in[0,\lambda]}} -|f'_{n,u}(t)|$$

satisfies, by sub-linearity of the supremum operator and triangular inequality, the bounded difference property, with differences bounded by $1/n$. By the bounded difference inequality, for all $\varepsilon > 0$, with probability larger than $1 - e^{-2n\varepsilon^2}$, we have

$$\sup_{\substack{u\in S^{d-1}\\ t\in[0,\lambda]}} -|f'_{n,u}(\lambda)| \leq \mathbb{E}\left[\sup_{\substack{u\in S^{d-1}\\ t\in[0,\lambda]}} -\frac{1}{n}\sum_{i=1}^{n} \psi'(\|X_i - T(P) - tu\|)\right] + \varepsilon.$$

This implies that, with the same probability,

$$\inf_{\substack{u\in S^{d-1}\\ t\in[0,\lambda]}} |f'_{n,u}(\lambda)| \geq \mathbb{E}\left[\inf_{\substack{u\in S^{d-1}\\ t\in[0,\lambda]}} \frac{1}{n}\sum_{i=1}^{n} \psi'(\|X_i - T(P) - tu\|)\right] - \varepsilon = m - \varepsilon,$$

where $m = \mathbb{E}\left[\inf_{\substack{u\in S^{d-1}\\ t\in[0,\lambda]}} \frac{1}{n}\sum_{i=1}^n \psi'(\|X_i - T(P) - tu\|)\right]$. For $\varepsilon = m/2$, we get that, with probability larger than $1 - e^{-nm^2/2}$, for all $u \in S^{d-1}$, $|f'_{n,u}(\lambda))| \geq m/2$. Use this equation and equation (3.7.5), equation (3.7.6), equation (3.7.7),

$$\begin{aligned}
\mathbb{P}(\|T(P) - T(\widehat{P}_n)\| \geq \lambda) &\leq \mathbb{P}(\exists u \in S^{d-1},\, f_{n,u}(\lambda) \geq 0) \\
&\leq 1 - \mathbb{P}\big(\forall u \in S^{d-1},\, f_{n,u}(0) - \lambda\big|f'_{n,u}(\lambda)\big| \leq 0\big) \\
&\leq 1 - \mathbb{P}\Big(\forall u \in S^{d-1},\, f_{n,u}(0) \leq \lambda\frac{m}{2}\Big) + e^{-nm^2/2} \\
&= \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \frac{X_i - T(P)}{\|X_i - T(P)\|}\psi(\|X_i - T(P)\|)\right\| \geq \lambda\frac{m}{2}\right) + e^{-nm^2/2} \\
&\leq t_{\mathrm{IF}}(\lambda m/2) + e^{-nm^2/2}. \qquad (3.7.8)
\end{aligned}$$

Finally, we bound m using Assumptions (1). For all $\lambda < \beta/2$,

$$\begin{aligned}
\mathbb{E}\left[\inf_{\substack{u\in S^{d-1}\\ t\in[0,\lambda]}} \frac{1}{n}\sum_{i=1}^n \psi'(\|X_i - T(P) - tu\|)\right] &\geq \gamma\mathbb{E}\left[\inf_{\substack{u\in S^{d-1}\\ t\in[0,\lambda]}} \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\|X_i - T(P) - tu\| \leq \beta\}\right] \\
&\geq \gamma\mathbb{E}\left[\inf_{u\in S^{d-1}} \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\|X_i - T(P)\| \leq \beta - \lambda\}\right] \\
&\geq \gamma\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\|X_i - T(P)\| \leq \beta/2\}\right] \\
&= \gamma\mathbb{P}(\|X_i - T(P)\| \leq \beta/2) \\
&= \gamma\mathbb{P}(\psi(\|X_i - T(P)\|) \leq \psi(\beta/2)).
\end{aligned}$$

The last inequality holds because ψ is increasing on $[0, \beta]$. As $V = \mathbb{E}[\psi(\|X - T(P)\|)^2] < \infty$, by Markov's inequality, it follows that

$$\mathbb{E}\left[\inf_{\substack{u\in S^{d-1}\\ t\in[0,\lambda]}} \frac{1}{n}\sum_{i=1}^n \psi'(\|X_i - T(P) - tu\|)\right] \geq \gamma\left(1 - \frac{V}{\psi(\beta/2)^2}\right). \qquad (3.7.9)$$

As $V \leq \psi(\beta/2)^2/2$, we get $m \geq \gamma/2$. Plugging this bound in equation (3.7.8) concludes the proof.

Proof of Theorem 21

Let us show that for all $H(t)$ such that $t\mathbb{E}_{H(t)}[\psi(\|X\|)] \xrightarrow[t\to 0]{} 0$, we have that W_ψ is $(H, \mathcal{P}_\psi)$-stable. For all $t \in [0,1]$, we have

$$\begin{aligned} W_\psi((1-t)P + tH(t), P) &= \sup_{h \preceq \psi} \left(\int h(x) d((1-t)P + tH(t))(x) - \int h(x) dP(x) \right) \\ &= \sup_{h \preceq \psi} \left((1-t) \int h(x) dP(x) + t \int h(x) dH(t)(x) - \int h(x) dP(x) \right) \\ &= t \sup_{h \preceq \psi} \left(\int h(x) dH(t)(x) - \int h(x) dP(x) \right) \\ &\leq t \mathbb{E}_{\substack{X \sim H(t) \\ Y \sim P}} [\psi(\|X - Y\|)], \end{aligned}$$

where $\mathbb{E}_{\substack{X \sim H(t) \\ Y \sim P}}$ is the expectation with X and Y independents, $X \sim H(t)$ and $Y \sim P$. The last line comes from $h \preceq \psi$. Since ψ is sub-additive and increasing, for all x, y, $\psi(\|x-y\|) \leq \psi(\|x\|)+\psi(\|y\|)$. Hence,

$$W_\lambda((1-t)P + tH(t), P) \leq t\big(\mathbb{E}_{H(t)}[\psi(\|X\|)] + \mathbb{E}_P[\psi(\|Y\|)]\big). \tag{3.7.10}$$

This tends to 0 as t tends to 0.

In particular, if $H(t) = \delta_{g(t)}$ is a Dirac distribution in $g(t)$, where g satisfies $t\psi(\|g(t)\|) \xrightarrow[t\to 0]{} l > 0$. Then,

$$\begin{aligned} W_\psi((1-t)P + tH(t), P) &= t \sup_{h \preceq \psi} \left\{ \mathbb{E}_{\substack{X \sim H(t) \\ Y \sim P}} [h(X) - h(Y)] \right\} \\ &= t \sup_{h \preceq \psi} \{h(g(t)) - \mathbb{E}_P[h(Y)]\}. \end{aligned}$$

Consider $h(x) = \psi(\|x\|)$. By sub-additivity of ψ, $h \preceq \psi$. Thus,

$$W_\psi((1-t)P + tH(t), P) \geq t(\psi(\|g(t)\|) - \mathbb{E}_P[\psi(\|Y\|)]).$$

The right hand side tends to $l > 0$ when $t \to 0$.

Proof of Theorem 22

Let $P \in \mathcal{P}_\psi$, we want to show that T is continuous in P. Let us denote

$$Z_P(\theta) = \mathbb{E}_P \left[\frac{X - \theta}{\|X - \theta\|} \psi(\|X - \theta\|) \right].$$

First, we show that $Z_P(T(Q))$ tends to 0. Denote, for $i \in \{1, \dots, d\}$,

$$Z_P^i(\theta) = \mathbb{E}_P \left[\frac{X_i - \theta_i}{\|X - \theta\|} \psi(\|X - \theta\|) \right].$$

By equation (3.2.2), $Z_P^i(T(P)) = 0$ and

$$Z_P^i(T(Q)) = \int \frac{x_i - T(Q)_i}{\|x - T(Q)\|} \psi(\|x - T(Q)\|) dP(x) - \int \frac{y_i - T(Q)_i}{\|y - T(Q)\|} \psi(\|y - T(Q)\|) dQ(y).$$

Then, we use the following lemma, whose proof is postponed to Section 3.7.2.

Lemma 10. *Let ψ satisfy Assumptions 1. Then, for all $\theta, x, y \in \mathbb{R}^d$,*

$$\left\| \frac{x-\theta}{\|x-\theta\|}\psi(\|x-\theta\|) - \frac{y-\theta}{\|y-\theta\|}\psi(\|y-\theta\|) \right\| \leq 7\psi(\|x-y\|).$$

hence, $\psi_i : x \mapsto \frac{x_i - T(Q)_i}{\|x-T(Q)\|}\psi(\|x-T(Q)\|)$ verifies that $\frac{1}{7}\psi_i \preceq \psi$. This implies that $Z_P^i(T(Q)) \leq 7W_\psi(P,Q)$ and thus

$$Z_P^i(T(Q)) \xrightarrow[W_\psi(P,Q)\to 0]{} 0. \tag{3.7.11}$$

Now, let Q_n denote a sequence such that $Z_P(T(Q_n)) \xrightarrow[n\to\infty]{} 0$. We want to show that $T(Q_n) \to 0$. First, we show that the sequence $T(Q_n)$ is bounded. By contradiction, suppose that there exists a subsequence $(k_n)_{n\in\mathbb{N}}$ such that $\|T(Q_{k_n}\| \xrightarrow[n\to\infty]{} \infty$. We show that $Z_P(T(Q_{k_n}))$ would not tend to zero. Let $u_n = \frac{-T(Q_{k_n})}{\|T(Q_{k_n})\|}$, we have

$$\begin{aligned}
\|Z_P(T(Q_{k_n}))\| =& \left\| \mathbb{E}\left[\frac{X - T(Q_{k_n})}{\|X - T(Q_{k_n})\|}\psi(\|X - T(Q_{k_n})\|) \right] \right\| \\
\geq& \mathbb{E}\left[\frac{\langle X - T(Q_{k_n}), u_n\rangle}{\|X - T(Q_{k_n})\|}\psi(\|X - T(Q_{k_n})\|) \right] \\
\geq& \psi(\|T(Q_{k_n})\|)\mathbb{E}\left[\frac{\langle X - T(Q_{k_n}), u_n\rangle}{\|X - T(Q_{k_n})\|} \right] - \mathbb{E}\left[\frac{|\langle X - T(Q_{k_n}), u_n\rangle|}{\|X - T(Q_{k_n})\|}\psi(\|X\|) \right].
\end{aligned} \tag{3.7.12}$$

Observe that we have the two inequalities

$$\|T(Q_{k_n})\| - \|X\| \leq \|X - T(Q_{k_n})\| \leq \|X\| + \|T(Q_{k_n})\|,$$

and because $\|T(Q_{k_n})\| \to \infty$ we have that

$$\frac{\|X - T(Q_{k_n})\|}{\|T(Q_{k_n})\|} \xrightarrow[n\to\infty]{a.s.} 1 \quad \text{and} \quad \frac{\langle X - T(Q_{k_n}), -T(Q_{k_n})\rangle}{\|T(Q_{k_n})\|^2} \xrightarrow[n\to\infty]{a.s.} 1.$$

This implies, by dominated convergence theorem (dominated by 1),

$$\mathbb{E}\left[\frac{\langle X - T(Q_{k_n}), -T(Q_{k_n})\rangle}{\|X - T(Q_{k_n})\|\|T(Q_{k_n})\|} \right] \xrightarrow[n\to\infty]{} 1$$

and again by dominated convergence theorem (dominated by $\psi(\|X\|)$ integrable),

$$\mathbb{E}\left[\frac{\langle X - T(Q_{k_n}), -T(Q_{k_n})\rangle}{\|X - T(Q_{k_n})\|\|T(Q_{k_n})\|}\psi(\|X\|) \right] \xrightarrow[n\to\infty]{} \mathbb{E}[\psi(\|X\|)].$$

From these two limits and equation (3.7.13), we get as n tends to infinity

$$\liminf_{n\to\infty} \|Z_P(T(Q_{k_n}))\| \geq \psi(\infty) - \mathbb{E}[\psi(\|X\|)] \tag{3.7.13}$$

the right-hand side is strictly positive by hypothesis and this contradicts $\|Z_P(T(Q_{k_n}))\| \xrightarrow[n\to\infty]{} 0$.

Hence, $\|T(Q_n)\|$ is a bounded sequence. Then, by Bolzano–Weierstrass theorem, there exists a subsequence $(k_n)_{n\in\mathbb{N}}$ with k_n that tends to infinity as n tends to infinity and with $T(Q_{k_n})$ that converges to a limit $l \in \mathbb{R}^d$. By continuity of Z_P, we have that

$$Z_P(T(Q_{k_n})) \xrightarrow[n\to\infty]{} Z_P(l)$$

and by (3.7.11), we also have $Z_P(T(Q_{k_n})) \xrightarrow[n\to\infty]{} 0$, hence by unicity of the limit $Z_P(l) = 0$ and by definition of $T(P)$, we have that $l = T(P)$ because we assumed that $T(P)$ was unique (see Section 3.2). All subsequences of $T(Q_n)$ converge to $T(P)$, this means $T(Q_n)$ converges to $T(P)$.

3.7.2 Proof of auxiliary results.

Proof of Lemma 3

Define $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ by

$$\rho(x) = \int_0^x \psi(t)dt.$$

ρ is C^2, even and increasing on $\mathbb{R}_+$. Consider the problem equivalent to (3.2.2) of finding

$$T(P) \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \mathbb{E}[\rho(\|X-\theta\|)]. \tag{3.7.14}$$

Let $J(\theta) = \mathbb{E}[\rho(\|X-\theta\|)]$, we prove that finding $T(P)$ is a convex problem. J is differentiable and its gradient is

$$\nabla J(\theta) = -\mathbb{E}\left[\frac{X-\theta}{\|X-\theta\|}\psi(\|X-\theta\|)\right].$$

Let us compute the Hessian. Let I_d be the identity matrix in $\mathbb{R}^{d\times d}$, we have

$$\mathrm{Hess}(J)(\theta) = \mathbb{E}\left[\left(\frac{I_d}{\|X-\theta\|} - \frac{(X-\theta)(X-\theta)^T}{\|X-\theta\|^3}\right)\psi(\|X-\theta\|) + \frac{(X-\theta)(X-\theta)^T}{\|X-\theta\|^2}\psi'(\|X-\theta\|)\right].$$

Then, for all $u \in \mathbb{R}^d$, $u \neq 0$

$$u^T\mathrm{Hess}(J)(\theta)u = \mathbb{E}\left[\left(\frac{\|u\|^2}{\|X-\theta\|} - \frac{\langle u, X-\theta\rangle^2}{\|X-\theta\|^3}\right)\psi(\|X-\theta\|) + \frac{\langle u, X-\theta\rangle^2}{\|X-\theta\|^2}\psi'(\|X-\theta\|)\right]. \tag{3.7.15}$$

Now, use Cauchy-Schwarz inequality to get

$$\frac{\|u\|^2}{\|X-\theta\|} - \frac{\langle u, X-\theta\rangle^2}{\|X-\theta\|^3} \geq 0 \tag{3.7.16}$$

and because ψ is concave and $\psi(0) = 0$, we have that for all $x > 0$, $\psi(x) \geq x\psi'(x)$. Hence, from equations (3.7.15) and (3.7.16),

$$\begin{aligned} u^T\mathrm{Hess}(J)(\theta)u &\geq \mathbb{E}\left[\left(\|u\|^2 - \frac{\langle u, X-\theta\rangle^2}{\|X-\theta\|^2}\right)\psi'(\|X-\theta\|) + \frac{\langle u, X-\theta\rangle^2}{\|X-\theta\|^2}\psi'(\|X-\theta\|)\right] \\ &= \mathbb{E}\left[\|u\|^2\psi'(\|X-\theta\|)\right] \geq 0 \end{aligned} \tag{3.7.17}$$

Hence, J is convex. Moreover, we have from Assumptions 1,

$$\mathbb{E}[\psi'(\|X - T(P)\|)] \geq \gamma \mathbb{P}(\|X - T(P)\| \leq \beta) = \gamma \mathbb{P}(\rho(\|X - T(P)\|) \leq \rho(\beta))$$

and from Markov inequality

$$\mathbb{E}[\psi'(\|X - T(P)\|)] \geq \gamma \left(1 - \frac{\mathbb{E}[\rho(\|X - T(P)\|)]}{\rho(\beta)}\right)$$

now, because of equation (3.7.14),

$$\mathbb{E}[\psi'(\|X - T(P)\|)] \geq \gamma \left(1 - \frac{\mathbb{E}[\rho(\|X - \mathbb{E}[X])\|)]}{\rho(\beta)}\right) > 0.$$

Hence, from Equation (3.7.17), for all $u \in \mathbb{R}^d$ $u \neq 0$,

$$u^T \mathrm{Hess}(J)(T(P))u > 0$$

The Hessian is definite positive at $T(P)$, hence $T(P)$ is unique.

Proof of Lemma 4

Huber's score function: The equality for the Huber's score function is immediate by derivation of ψ_H.

Catoni's score function: ψ_C is differentiable, and we have for all $x \geq 0$,

$$\psi_C'(x) = \frac{1 + \frac{x}{\beta}}{1 + \frac{x}{\beta} + \frac{x^2}{2\beta^2}}.$$

This function is decreasing on $\mathbb{R}_+$, positive and even, hence

$$\psi_C'(x) \geq \psi_C'(\beta) \mathbb{1}\{x \leq \beta\} = \frac{4}{5} \mathbb{1}\{x \leq \beta\}.$$

Polynomial score function: ψ_P is differentiable, and we have for all $x \geq 0$

$$\psi_P(x)' = \frac{1 + \frac{1}{p}\left(\frac{x}{\beta}\right)^{1-1/p}}{\left(1 + \left(\frac{x}{\beta}\right)^{1-1/p}\right)^2}.$$

As in the case of Catoni's score function, this function is decreasing over $\mathbb{R}_+$, positive and even. Then, we get

$$\psi_P'(x) \geq \psi_P'(\beta) \mathbb{1}\{x \leq \beta\} = \frac{1}{4}\left(1 + \frac{1}{p}\right) \mathbb{1}\{x \leq \beta\}.$$

Proof of Lemma 5

First, we can show that for all $q \in \mathbb{N}^*$, we have that

$$g_q : x \mapsto \begin{cases} q^q x/(e^q - 1) & if\ x \in [0, e^q - 1] \\ \log(1+x)^q & if\ x > e^q - 1 \end{cases}$$

is a concave function over $\mathbb{R}_+$.

g_q is continuous at $e^q - 1$, the left and right limits are equal to q^q. g_q is derivable on $[0, e^q - 1)$ and $(e^q - 1, \infty)$. This derivative is non-increasing on both intervals. At $e^q - 1$, the left derivative is $q^q(e^q - 1)^{-1}$ while the derivative on the right is $qq^{q-1}e^{-q} = q^q e^{-q}$. Thus, the left derivative at $e^q - 1$ is larger than the right derivative. Hence the derivative is non-increasing on $\mathbb{R}_+$, g_q is concave on $\mathbb{R}_+$.

By concavity, $\log(1+x)^q$ is smaller than its tangent in $e^{q-1} - 1$. This tangent is given by the function

$$x \mapsto (q-1)^q + \frac{q^q}{e^{q-1} - 1}\big(x - (e^{q-1} - 1)\big).$$

This last function is clearly smaller than the function $x \mapsto q^q x/(e^q - 1)$. Hence, $x \mapsto \log(1+x)^q$ is smaller than g_q, we found a concave upper bound of $x \mapsto \log(1+x)^q$.

Since g_q is concave, by Jensen's inequality, for any positive random variable Z such that $\mathbb{E}[Z] < \infty$, we have

$$\mathbb{E}[\log(1+Z)^q] \leq \mathbb{E}[g_q(Z)] \leq g_q(\mathbb{E}[Z]).$$

Then, for all x, we have $g_q(x) \leq \max(q^q, \log(1+x)^q)$, hence,

$$\mathbb{E}[\log(1+Z)^q] \leq \max(q^q, \log(1+\mathbb{E}[Z])^q).$$

Finally, use that $q^q \leq q! e^q$ to get

$$\mathbb{E}[\log(1+Z)^q] \leq q! \max(e, \log(1+\mathbb{E}[Z]))^q. \tag{3.7.18}$$

Denote

$$s = \max\left(e, \log\left(1 + \frac{\mathbb{E}[\|X - T_C(P)\|])}{\beta} + \frac{\mathbb{E}[\|X - T_C(P)\|^2]}{2\beta^2}\right)\right),$$

and apply equation (3.7.18) to $Z = X/\beta + X^2/(2\beta^2)$ to get

$$\mathbb{E}[\psi_C(\|X - T_C(P)\|)^q] \leq q!(\beta s)^q.$$

Proof of Lemma 6

By Markov's inequality, we have

$$t_{\mathrm{IF}}(\lambda) \leq \frac{\mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^n \frac{X_i - T_P(P)}{\|X_i - T_P(P)\|}\psi_P(\|X_i - T_P(P)\|)\right\|^{qp}\right]}{\lambda^{qp}}. \tag{3.7.19}$$

Let $Y_i = \frac{1}{n}\sum_{i=1}^n \frac{X_i - T_P(P)}{\|X_i - T_P(P)\|}\psi_P(\|X_i - T_P(P)\|)$ for $1 \le i \le d$, from [DG12, Theorem 1.2.5], there exists an absolute constant $K > 0$ such that

$$\mathbb{E}\left[\left\|\sum_{i=1}^n Y_i\right\|^{pq}\right]^{1/(pq)} \le Kpq\left(\mathbb{E}\left[\left\|\sum_{i=1}^n Y_i\right\|^2\right]^{1/2} + \mathbb{E}\left[\max_{1\le i\le n}\|Y_i\|^{pq}\right]^{1/(pq)}\right). \tag{3.7.20}$$

Let $\varepsilon_1, \dots, \varepsilon_n$ denote i.i.d Rademacher random variable independents from $Y_1, \dots, Y_n$. By the symmetrization lemma (see [DG12, Lemma 1.2.6]),

$$\mathbb{E}\left[\left\|\sum_{i=1}^n Y_i\right\|^2\right] \le 4\mathbb{E}\left[\left\|\sum_{i=1}^n \varepsilon_i Y_i\right\|^2\right] = 4\mathbb{E}\left[\sum_{i=1}^n \|Y_i\|^2\right] = 4n\mathbb{E}[\|Y\|^2].$$

Thus, by Jensen's inequality,

$$\mathbb{E}\left[\left\|\sum_{i=1}^n Y_i\right\|^2\right] \le 4n\mathbb{E}[\|Y\|^{pq}]^{2/(pq)}.$$

For the second term, as the max of n non-negative real numbers is smaller than their sum, we have

$$\mathbb{E}\left[\max_{1\le i\le n}\|Y_i\|^{pq}\right] \le \mathbb{E}\left[\sum_{1\le i\le n}\|Y_i\|^{pq}\right] \le n\mathbb{E}[\|Y\|^{pq}] \le n^{pq/2}\mathbb{E}[\|Y\|^{pq}].$$

Putting these two equations together we obtain from equation (3.7.20),

$$\mathbb{E}\left[\left\|\sum_{i=1}^n Y_i\right\|^{pq}\right]^{1/(pq)} \le 3Kpq\sqrt{n}\mathbb{E}[\|Y\|^{pq}]^{1/(pq)}. \tag{3.7.21}$$

From equations (3.7.19) and (3.7.21) and if we reinject the definition of Y_i's, we get

$$t_{\mathrm{IF}}(\lambda) \le \mathbb{E}[\psi_P(\|X - T_P(P)\|)^{pq}]\left(\frac{Kpq}{\sqrt{n}\lambda}\right)^{qp}.$$

Then, use that $\psi_p(\|x\|) \le \|x\|^{1/p}\beta^{1-1/p}$ to get

$$t_{\mathrm{IF}}(\lambda) \le \frac{\mathbb{E}[\|X - T_P(P)\|^q]}{\beta^q}\left(\frac{Kpq\beta}{\sqrt{n}\lambda}\right)^{qp}.$$

Proof of Lemma 8

We have

$$\begin{aligned} W_\psi\left(\frac{1}{n}\sum_{i=1}^n \delta_{X_i}, \frac{1}{n}\sum_{i=1}^n \delta_{X_i'}\right) &= \sup_{h\preceq\psi}\left(\frac{1}{n}\sum_{i=1}^n h(X_i) - \frac{1}{n}\sum_{i=1}^n h(X_i')\right) \\ &= \sup_{h\preceq\psi}\left(\frac{1}{n}\sum_{i\in\mathcal{O}} h(X_i) - h(X_i')\right) \\ &\le \frac{1}{n}\sum_{i\in\mathcal{O}}\psi(\|X_i - X_i'\|). \end{aligned} \tag{3.7.22}$$

If $(X_i)_{i\in\mathcal{O}}$ are i.i.d, then by Markov inequality, we have that for all $\lambda > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=n-k_n}^{n}\psi(\|X_i - X_i'\|) > \lambda\right) \leq \frac{k_n\mathbb{E}[\psi(\|X_i - X_i'\|)]}{n\lambda} \leq \frac{k_n(\mathbb{E}[\psi(\|X_i\|)] + \mathbb{E}[\psi(\|X_i'\|)])}{n\lambda}.$$

By hypothesis this last upper bound converges to 0 as n tends to infinity. This proves the convergence in probability. Now, by the triangular inequality we have,

$$W_\psi\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}, P\right) \leq W_\psi\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}, \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i'}\right) + W_\psi\left(P, \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i'}\right).$$

We have proved that the first term converges to 0 The second term converges to 0 by Lemma 7.

Proof of Lemma 9

$f_{n,u}$ derivable and we have

$$\begin{aligned} f_{n,u}'(\lambda) =& \frac{1}{n}\sum_{i=1}^{n}\left(-1 + \frac{\langle X_i - T(P) - \lambda u, u\rangle^2}{\|X_i - T(P) - \lambda u\|^2}\right)\frac{\psi(\|X_i - T(P) - \lambda u\|)}{\|X_i - T(P) - \lambda u\|} \\ & - \frac{1}{n}\sum_{i=1}^{n}\frac{\langle X_i - T(P) - \lambda u, u\rangle^2}{\|X_i - T(P) - \lambda u\|^2}\psi'(\|X_i - T(P) - \lambda u\|). \end{aligned} \tag{3.7.23}$$

By Cauchy-Schwarz inequality,

$$\forall i \in \{1, \dots, n\}, \quad \frac{-1}{\|X_i - T(P) - \lambda u\|} + \frac{\langle X_i - T(P) - \lambda u, u\rangle^2}{\|X_i - T(P) - \lambda u\|^3} \leq 0.$$

As ψ is concave on $\mathbb{R}_+$, we also have that, for all $y > 0$, $y\psi'(y) \leq \psi(y)$. Combining these two inequalities, we get

$$\begin{aligned} f_{n,u}'(\lambda) \leq& \frac{1}{n}\sum_{i=1}^{n}\left(-1 + \frac{\langle X_i - T(P) - \lambda u, u\rangle^2}{\|X_i - T(P) - \lambda u\|^2}\right)\psi'(\|X_i - T(P) - \lambda u\|) \\ & - \frac{1}{n}\sum_{i=1}^{n}\frac{\langle X_i - T(P) - \lambda u, u\rangle^2}{\|X_i - T(P) - \lambda u\|^2}\psi'(\|X_i - T(P) - \lambda u\|) \\ =& -\frac{1}{n}\sum_{i=1}^{n}\psi'(\|X_i - T(P) - \lambda u\|). \end{aligned}$$

Proof of Lemma 10

Let $\delta > 0$, we want to bound, for all $h \in \mathbb{R}^d$ such that $\|h\| \leq \delta$, the quantity

$$J(x, h) = \left\|\frac{x+h}{\|x+h\|}\psi(\|x+h\|) - \frac{x}{\|x\|}\psi(\|x\|)\right\|.$$

We consider two cases

Case 1: $\|x\| \leq 2\delta$.
As $\|h\| \leq \delta$, both x and $x+h$ have a norm bounded from above by 3δ. Then, by triangular inequality we have

$$J(x,h) \leq \psi(\|x\|) + \psi(\|x+h\|) \leq 2\psi(3\delta).$$

and finally, by sub-additivity, we have $\psi(3\delta) \leq 3\psi(\delta)$ which implies that $J(x,h) \leq 6\psi(\delta)$.

Case 2: $\|x\| > 2\delta$.
Let $x_h = \|x+h\|\frac{x}{\|x\|}$. By the triangular inequality

$$\begin{aligned} J(x,h) &\leq \left\| \frac{x+h}{\|x+h\|}\psi(\|x+h\|) - \frac{x_h}{\|x_h\|}\psi(\|x_h\|) \right\| + \left\| \frac{x_h}{\|x_h\|}\psi(\|x_h\|) - \frac{x}{\|x\|}\psi(\|x\|) \right\| \\ &= \left\| \frac{x+h}{\|x+h\|}\psi(\|x+h\|) - \frac{x}{\|x\|}\psi(\|x+h\|) \right\| + \left\| \frac{x}{\|x\|}\psi(\|x+h\|) - \frac{x}{\|x\|}\psi(\|x\|) \right\|. \quad (3.7.24) \end{aligned}$$

We use the following lemma whose proof is given Section 3.7.2.

Lemma 11. *If ψ verifies Assumptions 1. Then for all $h \in \mathbb{R}^d$ such that $\|h\| \leq \delta$, for all $x \in \mathbb{R}^d$ with $\|x\| \geq 2\delta$,*

$$\left\| \frac{x+h}{\|x+h\|}\psi(\|x+h\|) - \frac{x}{\|x\|}\psi(\|x+h\|) \right\| \leq 6\psi(\delta).$$

In the second term of the right hand side of equation (3.7.24), we are reduced to the 1-dimensional case,

$$\left\| \frac{x}{\|x\|}\psi(\|x+h\|) - \frac{x}{\|x\|}\psi(\|x\|) \right\| = |\psi(\|x+h\|) - \psi(\|x\|)|.$$

Then, by sub-additivity of ψ and triangular inequality,

$$\psi(\|x\|) - \psi(\|h\|) \leq \psi(\|x+h\|) \leq \psi(\|h\|) + \psi(\|x\|)$$

which shows that

$$\left\| \frac{x}{\|x\|}\psi(\|x+h\|) - \frac{x}{\|x\|}\psi(\|x\|) \right\| \leq \psi(\|h\|) \leq \psi(\delta).$$

Inject this and Lemma 11 in equation (3.7.24) to get for all $x, h \in \mathbb{R}^d$ with $\|x\| \geq 2\delta$ and $\|h\| \leq \delta$

$$J(x,h) \leq 7\psi(\delta).$$

Proof of Lemma 11

We have

$$\left\| \frac{x+h}{\|x+h\|}\psi(\|x+h\|) - \frac{x}{\|x\|}\psi(\|x+h\|) \right\| = \left\| \frac{x+h}{\|x+h\|} - \frac{x}{\|x\|} \right\| \psi(\|x+h\|). \quad (3.7.25)$$

Set $f(x) = \frac{x}{\|x\|}$. The function f is derivable away from 0 and as $\|x\| \geq 2\delta$, both x and $x+h$ are bounded away from 0. By Taylor inequality, we have

$$\|f(x+h) - f(x)\| \leq \|h\| \sup_{t\in[0,1]} \|df_{x+th}\|_{\text{op}}, \quad (3.7.26)$$

where df is the differential of f and $\|\cdot\|_{\mathrm{op}}$ is the operator norm with respect to the Euclidean norm. We have, for all $y \in S^{d-1}$,

$$df_{x+th}(y) = \frac{y}{\|x+th\|} - \frac{\langle y, x+th\rangle (x+th)}{\|x+th\|^3}.$$

Suppose that $\|x\| \geq 2\delta$, by triangular inequality and Cauchy-Schwarz inequality, we have

$$\|df_{x+th}\|_{\mathrm{op}} \leq \frac{\|y\|}{\|x+th\|} + \frac{|\langle y, x+th\rangle| \|x+th\|}{\|x+th\|^3} \leq \frac{2}{\|x+th\|} \leq \frac{2}{\|x\| - t\|h\|} \leq \frac{2}{\|x\| - \|h\|},$$

inject this in equation (3.7.26) and because $\|h\| \leq \delta$,

$$\|f(x+h) - f(x)\| \leq \|h\| \frac{2}{\|x\| - \|h\|} \leq \frac{2\delta}{\|x\| - \delta}.$$

From equation (3.7.25)

$$\left\| \frac{x+h}{\|x+h\|} \psi(\|x+h\|) - \frac{x}{\|x\|} \psi(\|x+h\|) \right\| \leq \frac{2\delta}{\|x\| - \delta} \psi(\|x+h\|).$$

By sub-linearity of ψ, this implies

$$\begin{aligned} \left\| \frac{x+h}{\|x+h\|} \psi(\|x+h\|) - \frac{x}{\|x\|} \psi(\|x+h\|) \right\| &\leq \frac{2\delta}{\|x\| - \delta} \psi(\|x\| + \delta) \\ &\leq \frac{2\delta}{\|x\| - \delta} (\psi(\|x\| - \delta) + \psi(2\delta)) \end{aligned} \tag{3.7.27}$$

then, because ψ is concave and $\psi(0) = 0$, we have that $\lambda \mapsto \psi(\lambda)/\lambda$ is non-increasing over $[0, \infty)$ and because $\|x\| - \delta \geq \delta$, we have

$$\frac{2\delta}{\|x\| - \delta} \psi(\|x\| - \delta) \leq \frac{2\delta}{\delta} \psi(\delta) = 2\psi(\delta).$$

The second term in equation (3.7.27) can be directly bounded using $\|x\| \geq 2\delta$,

$$\frac{2\delta}{\|x\| - \delta} \psi(2\delta) \leq \frac{2\delta}{\delta} \psi(2\delta) \leq 2\psi(2\delta)$$

and because ψ is sub-additive, $\psi(2\delta) \leq 2\psi(\delta)$. Inject this in equation (3.7.27) to get

$$\left\| \frac{x+h}{\|x+h\|} \psi(\|x+h\|) - \frac{x}{\|x\|} \psi(\|x+h\|) \right\| \leq 6\psi(\delta).$$

3.7.3 Technical tools

We remind the reader of Bernstein inequality, a classical concentration inequality, this form of Bernstein inequality is borrowed from [BLM13, Theorem 2.10].

Theorem 23. *Let $X_1, \ldots, X_n$ be independent real-valued random variables. Assume that there exist positive numbers v and c such that $\sum_{i=1}^n \mathbb{E}[X_i^2] \leq v$ and*

$$\sum_{i=1}^n \mathbb{E}[(X_i)_+^q v c^{q-2} \quad \text{for all integers } q \geq 3,$$

where $x_+ = \max(0, x)$. *Then for all* $t > 0$

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq \sqrt{2vt} + ct\right) \leq e^{-t}.$$

The following theorem is borrowed from [A+08, Theorem 4], it is a concentration inequality for suprema of sums of independent random variables.

Theorem 24. *Let* $X_1, \ldots, X_n$ *be independent random variables with values in a measurable space* $(\mathcal{S}, \mathcal{B})$ *and let* $\mathcal{F}$ *be a countable class of measurable functions* $f : \mathcal{S} \to \mathbb{R}$. *Assume that for every* $f \in \mathcal{F}$ *and every* i, $\mathbb{E}[f(X_i)] = 0$ *and for any* $\alpha \in (0, 1]$ *and all* i, $\|\sup_f |f(X_i)|\|_{\psi_\alpha} < \infty$. *Let*

$$Z = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} f(X_i) \right|.$$

Define moreover

$$\sigma^2 = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathbb{E}[f(X_i)^2].$$

Then, for all $0 < \eta < 1$ *and* $\delta > 0$, *there exists a constant* $C = C(\alpha, \eta, \delta) > 0$ *such that for all* $t \geq 0$,

$$\mathbb{P}(Z \geq (1+\eta)\mathbb{E}[Z] + t) \leq \exp\left(-\frac{t^2}{2(1+\delta)\sigma^2}\right) + 3\exp\left(-\left(\frac{t}{C\|\max_i \sup_{f \in \mathcal{F}} |f(X_i)|\|_{\psi_\alpha}}\right)^\alpha\right), \quad (3.7.28)$$

and

$$\mathbb{P}(Z \leq (1-\eta)\mathbb{E}[Z] - t) \leq \exp\left(-\frac{t^2}{2(1+\delta)\sigma^2}\right) + 3\exp\left(-\left(\frac{t}{C\|\max_i \sup_{f \in \mathcal{F}} |f(X_i)|\|_{\psi_\alpha}}\right)^\alpha\right). \quad (3.7.29)$$

The following theorem is found as Theorem 3.1 in [DLLO16] and show that the rate of convergence of an estimator towards the mean when $\mathbb{E}[X^2] = \infty$ cannot be of order $O(1/\sqrt{n})$.

Theorem 25. *For any* $M > 0$ *and* $\alpha \in (0, 1)$, *let* $\mathcal{P}_{1+\alpha}^M$ *be the set of all distributions on* $\mathbb{R}$ *such that* $\mathbb{E}[|X - \mathbb{E}[X]|^\alpha] = M$. *Let* $n > 5$ *be a positive integer and* $\delta \in (2e^{-n/4}, 1/2)$. *Then, for any estimator of the mean* $\widehat{E}_n$,

$$\sup_{P \in \mathcal{P}_{1+\alpha}^M} \mathbb{P}\left(\left|\widehat{E}_n(X_1^n, \delta) - \mathbb{E}[X]\right| > \left(\frac{M^{1/\alpha}\log(1/\delta)}{n}\right)^{\alpha/(\alpha+1)}\right) \geq \delta.$$

Chapter 4

Tractable robust mean estimation using M-Estimators.

Abstract

We present a new analysis of M-estimators of locations in $\mathbb{R}^d$ using results from [Mat20b]. In particular, we use concentration inequality on M-estimators from [Mat20b] to investigate the robust estimation of the mean in high dimension in a corrupted setting (Huber corruption neighborhood). The bounds we present are for $q > 1$ finite moments, for a sample of size n and covariance matrix Σ, we attain the minimax speed $\sqrt{Tr(\Sigma)/n} + \sqrt{\|\Sigma\|_{op}/n}$ in a heavy-tailed setting. In a corrupted setting, we attain speed $\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}\varepsilon^{1-1/q}$ when X has q finite moments. One of the major advantages of our approach compared to others recently proposed is that our estimator is tractable and fast to compute even in very high dimension, we present simulation results that show such computation using the iterative reweighting algorithm.

4.1 Introduction

One of the first tasks considered in robustness theory has been to compute so-called locations estimators meant to exhibit a central tendency of the data. Let $X \sim P$ for some P probability measure on $\mathbb{R}^d$, let ρ be an increasing function from $\mathbb{R}_+$ to $\mathbb{R}_+$, $\beta > 0$, we are interested in estimating the location parameter $T(P)$ defined by

$$T(P) \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \mathbb{E}\left[\rho\left(\frac{\|X - \theta\|}{\beta}\right)\right], \tag{4.1.1}$$

where $\|\cdot\|$ is the euclidean norm. Alternatively, if ρ is smooth enough (which will be the case in this article), we define $T(P)$ by

$$\mathbb{E}\left[\frac{X - T(P)}{\|X - T(P)\|}\psi\left(\frac{\|X - T(P)\|}{\beta}\right)\right] = 0, \tag{4.1.2}$$

where $\psi = \rho'$ is called the score function. The existence and unicity of $T(P)$ is assured under some hypothesis on ρ and P, see for instance Lemma 1 in [Mat20b]. To avoid cluttered notation, if ψ and β do not change, the dependency of T on β and ψ will be assumed without it being shown, otherwise it will be indicated in subscript.

The empirical estimator obtained by plugging the empirical density $\widehat{P}_n$ in equation (4.1.2) is called M-estimator (or Z-estimator) associated with ψ, it is denoted $T(\widehat{P}_n)$ and computed from an i.i.d sample $X_1, \dots, X_n$ using the following equation:

$$\sum_{i=1}^{n} \frac{X_i - T(\widehat{P}_n)}{\|X_i - T(\widehat{P}_n)\|} \psi\left(\frac{\|X_i - T(\widehat{P}_n)\|}{\beta}\right) = 0. \tag{4.1.3}$$

$T(\widehat{P}_n)$ always exists if ψ is non-decreasing for instance but it is not necessarily unique, if there are several possible choices we choose one arbitrarily. An estimator of this type has already been studied in [CG17]. This way of estimating $T(P)$ is taken from empirical risk minimization theory and a particular case of $T(P)$ is obtained when choosing $\psi(x) = x$ in which case $T(P) = \mathbb{E}[X]$ and $T(\widehat{P}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$, however it is well known that the empirical mean is not robust to both outliers and heavy-tailed data [Cat12] for instance in a corruption setting. A careful choice of the function ψ yields estimators that are more robust to corrupted and heavy-tailed data as it is shown in [Mat20b].

We consider $T(\widehat{P}_n)$ as an estimator of the mean and when we talk of the bias in this article, we mean the quantity $\|\mathbb{E}[X] - T(P)\|$. We are interested in the robustness of the empirical estimator $T(\widehat{P}_n)$ through the lense of the tools introduced in the article by the same author in [Mat20b]. Informally, in [Mat20b], we showed that the deviations of $T(\widehat{P}_n)$ around $T(P)$ could be controlled using the deviations of a sum of i.i.d random variables:

$$\|T(\widehat{P}_n) - T(P)\| \simeq \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{X_i - T(P)}{\|X_i - T(P)\|} \psi\left(\frac{\|X_i - T(P)\|}{\beta}\right) \right\|. \tag{4.1.4}$$

Then, we only need to bound the bias $\|T(P) - \mathbb{E}[X]\|$ in order to have a control of the deviation of $T(\widehat{P}_n)$ around the mean $\mathbb{E}[X]$ because the term on the right hand side of equation (4.1.4) is easily controlled using standard concentration inequalities (see [BLM13] for general concentration inequalities and [Mat20b] for their application to our problem). One of the consequences of this result is that if ψ is bounded, we can get a control on $\|T(\widehat{P}_n) - T(P)\|$ even when X does not have a finite second moment, see Corollary 1 in [Mat20b]. The moment condition comes from the need to control the bias $\|T(P) - \mathbb{E}[X]\|$. When P is symmetric, there is no need to control the bias and we obtain very fast concentration of $T(\widehat{P}_n)$ around $\mathbb{E}[X]$ even when the second moment is infinite, see Corollary 10 below.

In Section 4.3, we provides bounds on the bias $\|T(P) - \mathbb{E}[X]\|$ and on the variance terms in the concentration inequality from [Mat20b]. Bounding the bias has often been a problem ignored in robust statistics by saying that the methods work on symmetric distributions where we know that all the locations estimators are equal to $\mathbb{E}[X]$ and if the distribution is skewed we only say that we estimate a quantity meant to quantify a central tendency of P but not $\mathbb{E}[X]$ directly. However in statistical learning for example, estimating the mean is not just an arbitrary choice and we don't want to estimate a central tendency of the dataset, we want to estimate its mean. In this article, we give explicit bounds on the bias and we use those bounds (in Section 4.4.2) to give concentration results on $T(\widehat{P}_n)$ around $\mathbb{E}[X]$ in the context of heavy-tailed distribution even

beyond the L^2 case. We attain minimax rates as soon as X has strictly more than 2 moments provided a careful choice of the parameter of the score function is made. The choice of the parameter is a trade-off between bias (controlled in Section 4.3) and variance (controlled in Section 4.4.1 and through concentration inequalities from [Mat20b]) and we explicit this trade-off. This distinction between bias term and variance term in the bounds was not present in other works on robust estimation of the mean vector, to our knowledge.

In the literature, there are estimators that have strong theoretical guarantees but that are intractable, for example one can see estimators based on the aggregation of one-dimensional estimators (same idea as projection pursuits), see [Ler19, Theorem 44] and reference therein, see also [LM+19d] and there has also been estimators based on depth, for example Tukey's median [CGR+18]. On the other hand, there are tractable algorithms but whose theoretical guarantees are lacking for example the coordinate-wise median or the geometrical median [Min15, CG17], our work belong to this type of methods, our estimator is easily computable and even though the obtained error bounds are much better than for the coordinate-wise median, at least in corrupted setting it is not minimax. Recently there have been several propositions of algorithms whose goal was to be at the same time tractable and minimax, see [DKP20, DL19, Hop20] however these algorithms are often hard to implement and in practice the complexity makes them intractable for high-dimensional problems.

In this context, in Section 4.4.3, we show that $T(\widehat{P}_n)$ is suitable to estimate the mean in high dimension in a heavy-tailed and corrupted setting. In a corrupted setting we show a sub-optimal error bound, our estimator is not minimax when the data comes from Huber contamination distributions (except in dimension 1). Our result takes the following form informally: if $X_1, \dots, X_n$ are i.i.d from a mixture distribution $(1-\varepsilon)P + \varepsilon Q$ with P having $q \geq 2$ finite moments and Σ the covariance matrix of P. Then, under some assumptions on ψ and n, for all $0 < \lambda \lesssim n$, with probability larger than $1 - 8\exp(-\lambda/8)$,

$$\left\|\mathbb{E}[X] - T(\widehat{P}_n)\right\| \lesssim \frac{\sqrt{Tr(\Sigma)} + \sqrt{\|\Sigma\|_{op}\lambda}}{\sqrt{n}} \bigvee \mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q} \varepsilon^{1-1/q}.$$

See Theorem 27 below for the formal and more precise statement.

This type of bound is not really surprising and because of the factor in front of the ε term, this is not minimax. The dependency in the number of finite moments links the two common settings: when P has two finite moments, the bound will be in $\sqrt{\varepsilon}$ as in [DL19, DKP20] while if P is Gaussian, it is known for example in [CGR+18] that the dependency can be of order ε. Our result extrapolate between these two results. A maybe more surprising consequence of the fact that we separate the effect of the bias and the effect of the variance is that we show that in fact to achieve rates in ε, we don't need the inliers to be Gaussian, we only need them to be symmetric, see Corollary 10. Moreover, when the inliers are symmetric we have a bound that greatly improve on the asymmetric case because we replace the terms in $Tr(\Sigma)$ and $\|\Sigma\|_{op}$ can be replaced by smaller terms that are assured to be finite even if X does not have a finite second moment.

In Section 4.4.4 we study the case where before feeding the samples to the estimator $T(\widehat{P}_n)$, we begin by grouping them and computing the mean on each group using the same principle as median of means estimator. This section gives advances in answering the question: is it interesting to make groups and take the route of median of means estimators or is it better to just use M-estimators out of the box? Our answer is that when the density is very skewed and heavy-tailed, it is interesting to make blocks but if the density is symmetric, using a Huber estimator is advised. This observation is later illustrated in the numerical experiments of Section 4.6.2. In practice

we gave indication on how to choose the parameter β of a M-estimator but in doing so, we used characteristics of the underlying distribution such as its variance which is not available to the statistician in most cases. To resolve this problem, in Section 4.5, we show how to tune the hyperparameter β adaptively in dimension 1 through Lepski's method.

Finally, Section 4.6 presents an algorithm to compute $T(\widehat{P}_n)$ and a proof of convergence of this algorithm. The algorithm used is the well-known iterative-reweighting algorithm used to compute M-estimators for example in the context of regression (see [HR09, Section 7.8] or [BT74, HI17a]), this algorithm has a complexity $O(Tnd)$ where T is the number of iterations, n is the sample size and d the dimension, it is very fast in practice. One of the advantages of an iterative-reweighting algorithm is that the computation is rather fast, the complexity is linear in both the sample size and the dimension and it can be used in a high-dimensional setting. Using this algorithm, we illustrate the techniques highlighted in this article with a study of Lepski's method in dimension 1 and a comparison of diverse robust mean estimators in high dimension. The code for this last section can be found on github at https://github.com/TimotheeMathieu/RobustMeanEstimator/.

4.2 Setting and Notations

4.2.1 Setting

Throughout the article, we use the following assumptions on ψ.

Assumptions 3. *Suppose that X is a continuous random variable, ψ is a continuous, non-decreasing, concave function on $\mathbb{R}_+$, differentiable almost everywhere, $\psi(0) = 0$ and we have for almost every $x > 0$,*

$$\gamma \mathbb{1}\{x \leq 1\} \leq \psi'(x) \leq 1$$

for some $\gamma > 0$. Let $\beta > 0$ and let $\rho(x) = \int_0^x \psi(t)dt$, we assume that

$$\rho(1/3) \geq \mathbb{E}[\rho(\|X - \mathbb{E}[X]\|/\beta)]. \tag{4.2.1}$$

Remark that there exists always a $\beta > 0$ that satisfies equation (4.2.1) as long as the right hand side of (4.2.1) is finite, just take β large enough. Equation (4.2.1) makes us choose β large enough for us to be able to do our analysis, this is a technical assumption that could be weakened at the cost of simplicity. Similarly, we use continuous random variables by simplicity, this could also be weakened. If Assumptions 3 are satisfied, we have in particular that our problem is well defined in the sense that $T(P)$ exists and is unique. See [Mat20b, Lemma 1] for a proof of this fact.

Of special interest for us will be Huber's estimator whose score function ψ_H is defined as follows. For all $x \geq 0$, let

$$\psi_H(x) = x\, \mathbb{1}\{x \leq 1\} + \mathbb{1}\{x > 1\}. \tag{4.2.2}$$

In dimension 1, the M-estimator constructed from this score function is called the Huber's estimator [Hub64]. ψ_H is differentiable except in 1 and we have $\psi'(x) = \mathbb{1}\{x \leq 1\}$ for $x \neq 1$ which verifies Assumptions 3 with $\gamma = 1$. The corresponding ρ function is for $x > 0$,

$$\rho_H(x) = \frac{x^2}{2}\mathbb{1}\{x \leq 1\} + \left(x - \frac{1}{2}\right)\mathbb{1}\{x > 1\}.$$

Assumptions 3 are verified for β large enough if X has a finite first moment.

Asymptotically, we know that in a Gaussian contamination neighborhood, ψ_H is optimal for the variance (see [HR09]). However it is not clear which ψ is optimal in the context of heavy-tailed distribution to estimate the mean or if unbounded ψ function can be sometimes interesting and even though [Mat20b] tries to expound on this point there is no definite answer given.

4.2.2 Notations

Let $\mathcal{P}$ denote the set of probability distributions on $\mathbb{R}^d$, $S^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ where $\|\cdot\|$ is the Euclidean norm. For any two reals a, b, denote $a \lesssim b$ is $a \leq Cb$ for some universal constant $C > 0$.

Let $X, X_1, \dots, X_n$ denote i.i.d random variables such that $X \sim P \in \mathcal{P}$. Let $\widehat{P}_n$ denotes the empirical distribution given by $\widehat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ where, for any $x \in \mathbb{R}^d$, δ_x is the Dirac distribution in x. If it exists, we will denote Σ the covariance matrix of X.

Let T_H, denote the functional such that, for Huber's score function ψ_H defined in (4.2.2)

$$\mathbb{E}\left[\frac{X - T_H(P)}{\|X - T_H(P)\|}\psi_H\left(\frac{\|X - T_H(P)\|}{\beta}\right)\right] = 0.$$

Define the following variance terms appearing in [Mat20b],

$$V = \mathbb{E}\left[\beta^2\psi\left(\frac{\|X - T(P)\|}{\beta}\right)^2\right], \quad v = \left\|\mathbb{E}\left[\frac{(X - T(P))(X - T(P))^T}{\|X - T(P)\|^2}\psi\left(\frac{\|X - T(P)\|}{\beta}\right)^2\right]\right\|_{op} \tag{4.2.3}$$

where $\|\cdot\|_{op}$ denotes the operator norm associated with $\|\cdot\|$. In the special case of Huber's estimator, denote

$$V_H = \mathbb{E}\left[\psi_H\left(\frac{\|X - T_H(P)\|}{\beta}\right)^2\right], \quad v_H = \left\|\mathbb{E}\left[\frac{(X - T_H(P))(X - T_H(P))^T}{\|X - T_H(P)\|^2}\psi_H\left(\frac{\|X - T_H(P)\|}{\beta}\right)^2\right]\right\|_{op}.$$

In the Gaussian setting $X \sim \mathcal{N}(\mu, \Sigma)$ for $\psi(x) = x$, we have $V = Tr(\Sigma)$ and $v = \|\Sigma\|_{op}$ and, from Hanson-Wright inequality (see equation (4.4.2)), we have that $Tr(\Sigma)$ and $\|\Sigma\|_{op}$ describe the spread of the empirical mean in high dimension. Here we are not in a Gaussian setting and for example in the case of Huber's estimator, V_H and v_H will describe the variability of the influence function of Huber's estimator, notice that V_H and v_H are always finite, for any distribution P.

4.3 Bias and variance of M-estimators when considered as estimators of the mean

The problem of bounding the bias $\|T(P) - \mathbb{E}[X]\|$ has been avoided in a number of articles on robust statistics by saying that if P is symmetric then $T(P) = \mathbb{E}[X]$, which is true but unfortunately in the case of skewed distribution this bias can be very large and the choice of β

will determine how large the bias is. In this section, we aim at finding how the bias behaves as β grows.

We introduce the following function

$$Z_\beta : \theta \mapsto \mathbb{E}\left[\beta \frac{(X-\theta)}{\|X-\theta\|}\psi\left(\frac{\|X-\theta\|}{\beta}\right)\right].$$

Z_β is linked to the influence function of T which can be defined as a Gâteaux differential of the functional T and which informally measures the influence that a point placed on θ has on the value of T, see [HRRS86] for a formal definition. The following lemma is the cornerstone of this article as it links Z_β with the distance between $T(P)$ and $\mathbb{E}[X]$.

Theorem 26. *Let X be a random vector in $\mathbb{R}^d$, $X \sim P$ with finite expectation and suppose Assumptions 3, then*

$$\frac{1}{3}\|Z_\beta(\mathbb{E}[X])\| \leq \|\mathbb{E}[X] - T(P)\| \leq \frac{2}{\gamma}\|Z_\beta(\mathbb{E}[X])\|$$

We postpone the proof to Section 4.7.1. From Theorem 26, it is sufficient to upper bound $\|Z_\beta(\mathbb{E}[X])\|$ to get a bound on the bias.

The choice of β is a very important problem when estimating $\mathbb{E}[X]$ using $T(\widehat{P}_n)$ and in particular we will need to choose β carefully as a function of n in order to have $T(\widehat{P}_n)$ that converges to $\mathbb{E}[X]$. The choice of β will entail a sort of bias-variance tradeoff. Remark that we do not need a finite second moment for our analysis to work, we only need $\mathbb{E}[\rho(\|X - T(P)\|/\beta)] < \infty$ which is linked to the first moment of X in the case of $\psi = \psi_H$.

4.3.1 Bias of Huber's estimator

We begin with the bias of the Huber estimator obtained from equation (4.1.3) with $\psi = \psi_H$. In the case of Huber's estimator, we can do some explicit computations for example with the Pareto distribution and this will give us some baseline.

Lemma 12. *Let T_H be Huber functional defined in equation (4.1.2), if X follows a Pareto distribution with shape parameter α (i.e. X has density $f(x) = \alpha \mathbb{1}\{x \geq 1\}/x^{\alpha+1}$), then when $\beta \to \infty$,*

$$|\mathbb{E}[X] - T_H(P)| = \Theta\left(\frac{1}{\beta^{\alpha-1}}\right).$$

Where Θ is the Landau notation that corresponds to being lower bounded and upper bounded by constant times the function under the parenthesis. The proof is in Section 4.8.1. Lemma 12 shows that the distance to the mean depends strongly on the tail of the distribution of X and this Lemma gives some lower bound on the attainable bias when we have only a few finite moments. Then, we can show the following lemma that gives a bound on the bias of the Huber estimator for a distribution with a finite number of moments.

Lemma 13. *Let X be a random variable with $\mathbb{E}[\|X\|^q] < \infty$ for $q \in \mathbb{N}^*$ and suppose that Assumptions 3 hold. Then*

$$\|\mathbb{E}[X] - T_H(P)\| \leq \frac{2\mathbb{E}[\|X - \mathbb{E}[X]\|^q]}{(q-1)\beta^{q-1}}.$$

The proof is in Section 4.8.2. Lemma 13 is not exactly tight as can be seen on the example of the Pareto distribution for $d = 1$ from Lemma 12: when $\alpha = 2$, we have only one finite moment but we obtain nonetheless a rate of $1/\beta$. However, our result is almost tight as we see that in fact if $\alpha > 2$, then we have two finite moments and a rate of $1/\beta$.

In addition to Lemma 13 we can also show an exponential bound on the bias when the random variable X is sub-exponential however because the primary use of Huber estimator is with robust statistics, we only state the result for a finite number of finite moments as it is what will interest us. An interested reader can adapt the proof to lighter-tailed distributions and obtain bounds that are exponentially small with β.

4.3.2 Bias of smooth M-estimators

For a ψ function that is not Huber's score function, the bias also depends strongly on the behavior of ψ near 0.

Lemma 14. *Suppose that ψ is C^k with bounded k^{th} derivative and that Assumptions 3 hold, $\psi'(0) = 1$ and for $2 \leq j \leq k-1$, $\psi^{(j)}(0) = 0$. Let X be a random variable such that $\mathbb{E}[\|X\|^k] < \infty$, then,*

$$\|Z_B(\mathbb{E}[X])\| \leq \frac{\|\psi^{(k)}\|_\infty}{k!\beta^{k-1}}\mathbb{E}\big[\|X - T(P)\|^k\big] \tag{4.3.1}$$

And if X follows a Bernoulli distribution of parameter p, this bound is tight in its dependency in β, when $\beta \to \infty$, we have

$$Z_\beta(\mathbb{E}[X]) = \psi^{(k)}(0)\frac{p(1-p)^k - (1-p)p^k}{k!\beta^{k-1}} + o\left(\frac{1}{\beta^{k-1}}\right)$$

For example, we can show that for Catoni's score function $\psi(x) = \log(1 + x + x^2/2)$ whose second derivative is $\psi''(x) = -(x + x^2/2)/(1 + x + x^2/2)^2$, we have that $\psi(x) = x + x^3/6 + o(x^3)$ and then the bias of Catoni's estimator is in general of order $1/\beta^2$. Lemma 14 shows that the bias depends on the smoothness of the function near 0 and also the number of finite moments, when the distribution is light-tailed it can be interesting to choose an estimator that is equal to the identity near 0 in order to have a bound on the bias that is similar to Huber estimator, then we have to choose the behavior of ψ near infinity and this depends on the robustness/efficiency trade-off see [Mat20b] for more information on the choice of ψ at infinity.

4.4 Concentration inequalities of Huber's estimator and HOME

In this section, we investigate the concentration of Huber's estimator around the mean in diverse settings. The goal will be to recover deviations similar to the one we would have in a Gaussian setting, but when the data are not Gaussian. It can be that the data are heavy-tailed (sere Section 4.4.2) or corrupted by outliers (see Section 4.4.3). The gold standard in this context is the deviation of the empirical mean in a Gaussian setting (see [BLM13]). If $X_1, \ldots, X_n$ are i.i.d

from $\mathcal{N}(\mu, \sigma^2)$ for some $\mu \in \mathbb{R}$ and $\sigma > 0$, then for all $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu'\right| > \sigma\sqrt{\frac{t}{2n}}\right) \leq e^{-t}. \tag{4.4.1}$$

An equivalent of this in the multi-dimensional setting is Hanson-Wright inequality [HW71]: let $X \sim \mathcal{N}(\mu, \Sigma)$ for Σ a positive definite matrix, $\mu \in \mathbb{R}^d$. Then, for any $t > 0$,

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right\|^2 > \frac{2Tr(\Sigma)}{n} + \frac{9t\|\Sigma\|_{op}}{n}\right) \leq e^{-t}. \tag{4.4.2}$$

This form of Hanson-Wright inequality can be found for example in [Ler19]. These two deviation bounds will be the gold standard for our task, we obtain deviations similar to the ones in equations (4.4.1) and (4.4.2) but in a non-Gaussian setting.

4.4.1 Bound on the variance of M-estimators

First, we have to control the variability of $T(\widehat{P}_n)$ in order to control its deviations. The following lemma gives an upper bound on both V and v.

Lemma 15. *Suppose that Assumptions 3 are satisfied, we have that* $V \leq \mathbb{E}[\|X - \mathbb{E}[X]\|^2] = \mathrm{Tr}(\Sigma)$, *and* $v \leq \|\Sigma\|_{op} + \|\mathbb{E}[X] - T(P)\|^2$.

Lemma 15 (proven in Section 4.8.4) gives a control on V and v using the properties of X. Next we show that Lemma 15 is tight in the case of Huber's estimator as long as X is sufficiently concentrated using the following lemma whose proof is provided in Section 4.8.5.

Lemma 16. *Suppose that Assumptions 3 are satisfied and that* X *is such that* $\mathbb{E}[\|X\|^{2q}] < \infty$ *for some* $q > 1$, *then*

$$V_H \geq \mathbb{E}[\|X - \mathbb{E}[X]\|^2] - 4^q \frac{\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right]^{1-1/q}}{\left(\mathbb{E}[\|X - T_H(P)\|^{2q}] + \beta^{2q}\right)^{1-2/q}}.$$

and

$$v_H \geq \|\Sigma\|_{op} - 4^q \frac{\mathbb{E}\left[\|X - T_H(P)\|^{2q}\right]^{1-1/q}}{\left(\mathbb{E}[\|X - T_H(P)\|^{2q}] + \beta^{2q}\right)^{1-2/q}}$$

Lemma 15 and 16 imply that if X has enough moments, say with 4 finite moments, and if β is sufficiently large, then the behavior of the variance term is the same as the variance term for the empirical mean. On the other hand, if X is not very concentrated, Lemma 15 can be a very rough bound and in the case of Huber estimator if X has only a finite first moment but no finite variance, then V_H and v_H are finite even though $Tr(\Sigma) = \|\Sigma\|_{op} = \infty$.

4.4.2 Concentration of Huber's estimator

In this section we use the concentration inequalities of $T_H(\widehat{P}_n)$ around $T_H(P)$ proved in [Mat20b] and the bound on the bias obtained in Section 4.3 to get the concentration of $T_H(\widehat{P}_n)$ around

$\mathbb{E}[X]$. This allows us to get results with a similar flavour to the results from [Cat12] and [CG17]. Let X be a real random variable, in dimension 1, V_H reduces to $V_H = \mathbb{E}[\beta^2 \wedge |X - T(P)|^2]$. From [Mat20b], we have the following lemma.

Lemma 17. *Let X be a real random variable with law P. If $8V_H \leq \beta^2 < \infty$, then for all $t > 0$ such that $4\sqrt{2V_H t/n} + 4\beta t/n \leq \beta/2$, with probability greater than $1 - 2e^{-t} - e^{-n/8}$,*

$$\left|T_H(\widehat{P}_n) - T_H(P)\right| \leq 4\sqrt{\frac{2V_H t}{n}} + 4\frac{\beta t}{n}. \tag{4.4.3}$$

Then, using Lemma 13 we have a bound on what we call the **bias term** and using Lemma 17 we have a bound on what one could call the **variance term**. We use this terminology because in practice, the choice of β will imply a tradeoff which will be very similar to the bias/variance tradeoff in classic statistics. In the rest of the article, bias term will denote the bound on the bias from Lemma 13 and the variance term will always be from a concentration inequality around $T(P)$ similar to Lemma 17. We have the following lemma.

Lemma 18. *Let X be a real random variable with $\mathbb{E}[|X|^q] < \infty$ for some $q > 1$, suppose that Assumptions 3 are satisfied and $8V_H \leq \beta^2$. For all $t > 0$ such that $4\sqrt{2V_H t/n} + 4\beta t/n \leq \beta/2$, with probability greater than $1 - 2e^{-t} - e^{-n/8}$,*

$$\left|T_H(\widehat{P}_n) - \mathbb{E}[X]\right| \leq 4\sqrt{\frac{2V_H t}{n}} + 4\frac{\beta t}{n} + \frac{2\mathbb{E}[|X - \mathbb{E}[X]|^q]}{(q-1)\beta^{q-1}}. \tag{4.4.4}$$

To conclude on the concentration of Huber estimator, we have to choose β and we see that the rate of convergence of order $O(\sqrt{V_H t/n})$ is preserved if β satisfies

$$\left(\frac{\mathbb{E}[|X - \mathbb{E}[X]|^q]\sqrt{n}}{\sqrt{V_H t}}\right)^{1/(q-1)} \leq \beta \leq \sqrt{nV_H}.$$

Then, choosing such a β gives us sub-Gaussian rates similar to equation (4.4.1).

From Section 4.4.1 we have that as soon as $q \geq 4$ and $\beta \geq \mathbb{E}[|X - \mathbb{E}[X]|^q]^{1/q}$, even though V_H is in general smaller than σ^2 we have in fact that $V_H \simeq \sigma^2$, and then we may want to choose β that optimizes $4\frac{\beta t}{n} + \frac{2\mathbb{E}[|X-\mathbb{E}[X]|^q]}{(q-1)\beta^{q-1}}$ which corresponds to $\beta^q = \frac{n\mathbb{E}[|X-\mathbb{E}[X]|^q]}{2t}$, in this case the dominant term is the variance term $4\sqrt{2V_H t/n} \leq 4\sigma\sqrt{2t/n}$. On the other hand, if one does not care about the constants, we can deduce the following alternative result when X has a finite variance, this is another flavor of the concentration of Huber's estimator obtained by Catoni in [Cat12] but with worst constants. Take $\beta = \sigma\sqrt{n/(2t)}$, the condition on t and β can both be resumed to one condition which gives us the following lemma.

Lemma 19. *Let X be a real random variable with $\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] < \infty$. For all $t \in (0, n/16)$, with probability greater than $1 - 2e^{-t} - e^{-n/8}$,*

$$\left|T_H(\widehat{P}_n) - \mathbb{E}[X]\right| \leq 8\sigma\sqrt{\frac{2t}{n}}. \tag{4.4.5}$$

In higher dimension, we can also have similar results to the one-dimensional case. From [Mat20b], we have the following lemma.

Lemma 20. *Let $X_1, \dots, X_n$ be i.i.d random variables with law P on $\mathbb{R}^d$. If $\beta^2 \geq 8V_H$, then there exist an absolute constant $C > 0$ such that, for all $0 < \lambda \lesssim n$, with probability larger than $1 - 4\exp(-\lambda) - \exp(-n/8)$,*

$$\left\| T_H(P) - T_H(\widehat{P}_n) \right\| \leq 6\frac{V_H^{1/2}}{\sqrt{n}} + 8\sqrt{v_H \frac{\lambda}{n}} + \frac{C}{n}\lambda\beta. \tag{4.4.6}$$

Once again, we have a bias term and a variance term. Then, we inject the control on the bias obtained in Lemma 13 in Lemma 20 to get under the same hypothesis as in Lemma 20, for all $\lambda \leq \lambda_{max}$, with probability greater than $1 - 4e^{-\lambda} - e^{-n/8}$,

$$\left\| \mathbb{E}[X] - T_H(\widehat{P}_n) \right\| \leq 6\frac{V_H^{1/2}}{\sqrt{n}} + 8\sqrt{v_H \frac{\lambda}{n}} + \frac{C}{n}\lambda\beta + \frac{2\mathbb{E}[\|X - \mathbb{E}[X]\|^q]}{(q-1)\beta^{q-1}}. \tag{4.4.7}$$

Let us explicit the rate of convergence in equation (4.4.7). As soon as

$$\left(\frac{\mathbb{E}[\|X - \mathbb{E}[X]\|^q]\sqrt{n}}{\sqrt{V_H}}\right)^{1/(q-1)} \leq \beta \leq \sqrt{\frac{nv_H}{\lambda}},$$

the bias term is of a smaller order of magnitude than the variance term. Contrary to the $d = 1$ case, we don't recover Catoni's result from [CG17], if we choose β optimally in the case of 2 finite moments we get a bound of order $\sqrt{Tr(\Sigma)\lambda/n}$. However, as soon as we have strictly more than 2 moments the bias becomes negligible in front of the variance term and we recover Catoni's result in [CG17] with worst constants, specifying equation (4.4.7) with $\beta = \sqrt{\frac{\mathbb{E}[\|X-\mathbb{E}[X]\|^3]\sqrt{n}}{\sqrt{Tr(\Sigma)}}}$.

Lemma 21. *Let X be such that $\mathbb{E}[\|X\|^3] < \infty$, suppose that Assumptions 3 are satisfied. For all $n \geq 64$, there exists a universal constant $C > 0$ such that for all $0 < \lambda \lesssim n$, with probability larger than $1 - 4\exp(-\lambda) - \exp(-n/8)$,*

$$\left\| \mathbb{E}[X] - T_H(\widehat{P}_n) \right\| \lesssim \frac{Tr(\Sigma)^{1/2}}{\sqrt{n}} + \sqrt{\|\Sigma\|_{op}\frac{\lambda}{n}} + \frac{\lambda}{n^{3/4}}\frac{\mathbb{E}[\|X - \mathbb{E}[X]\|^3]^{1/2}}{Tr(\Sigma)^{1/4}}$$

We simplified the condition on λ_{max} thanks to the given value of β. The condition $n \geq 64$ can be weakened to $n \geq 64Tr(\Sigma)^{3/2}/\mathbb{E}[\|X - \mathbb{E}[X]\|^3]$ if one needs to use this inequality for small sample-size.

Remark 1 (Comparison of bias term and variance term)**.** *To compare the bias and variance terms, we have to understand the relation of $\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}$ with $Tr(\Sigma)$ and $\|\Sigma\|_{op}$. For instance, let $X^{(i)}$ be the i^{th} coordinate of X, by Jensen's inequality gives us for $q \geq 2$,*

$$\mathbb{E}[\|X - \mathbb{E}[X]\|^q] = \mathbb{E}\left[\left(\sum_{i=1}^d (X^{(i)} - \mathbb{E}[X^{(i)}])^2\right)^{q/2}\right] \leq d^{q/2-1}\sum_{i=1}^d \mathbb{E}\left[|X^{(i)} - \mathbb{E}[X^{(i)}]|^q\right]$$

Hence, the dependency of $\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}$ in the dimension is of order $\sqrt{d}$ if the distribution have the same marginal on each direction and this is the behavior recovered in the case of Gaussian

random variables with covariance matrix proportional to the identity matrix. Indeed, it can be shown ([BLM13]) that for $X \sim \mathcal{N}(0, \Sigma)$ for some positive definite matrix Σ,

$$\mathbb{E}[\|X\|^{2q}]^{1/2q} \leq 4(q!)^{1/2q}\sqrt{Tr(\Sigma)} + ((2q)!)^{1/2q}8\sqrt{\|\Sigma\|_{op}}.$$

The Gaussian behavior is recovered as long as the marginals of X have sub-Gaussian behavior of its i^{th} moment for $1 \leq i \leq q$. In this setting, as long as all the quantities mentioned are finite, in regard to the dimension, we can treat similarly $\sqrt{Tr(\Sigma)}$ and $\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}$.

Lemma 21 can be extended to smooth M-estimators using the bound on the bias found in Lemma 14 and concentration inequalities on the influence function using tools in [Mat20b], such a concentration inequality would depend on the properties of X and ψ.

4.4.3 Huber estimator in Huber corruption setting

Let $P_\varepsilon = (1-\varepsilon)P + \varepsilon Q$ for some probability distribution Q. We want to estimate the expectation of the distribution P while we have only access to a sample $X_1, \dots, X_n$ i.i.d from P_ε. The variance term that takes the place of V_H is

$$V_\varepsilon = (1-\varepsilon)V_H + \varepsilon \mathbb{E}_{X\sim Q}\left[\beta^2 \psi_H\left(\frac{\|X - T(P)\|}{\beta}\right)^2\right] \leq V_H + \varepsilon\beta^2 \tag{4.4.8}$$

the last inequality being a consequence of $\psi_H \leq 1$. Similarly,

$$v_\varepsilon \leq (1-\varepsilon)v_H + \varepsilon\left\|\mathbb{E}_{X\sim Q}\left[\beta^2 \frac{(X - T(P))(X - T(P))^T}{\|X - T(P)\|^2}\psi\left(\frac{\|X - T(P)\|}{\beta}\right)^2\right]\right\|_{op} \leq v_H + \varepsilon\beta^2. \tag{4.4.9}$$

This conclude the bound on the variance term, now let us control the bias term, we have

$$\|T_H(P_\varepsilon) - \mathbb{E}[X]\| \leq \|T_H(P_\varepsilon) - T_H(P)\| + \|T_H(P) - \mathbb{E}[X]\|. \tag{4.4.10}$$

The first term in the right hand side of equation (4.4.10) is controlled by the following lemma whose proof is in Section 4.8.6

Lemma 22. *If $\varepsilon \leq 1/8$ and $8V_H \leq \beta^2$, then*

$$\|T_H(P) - T_H(P_\varepsilon)\| \leq 4\varepsilon\beta$$

Bounding the difference $\|T_H(P_\varepsilon) - T_H(P)\|$, is a rather old problem and it has already been treated in numerous application with no explicit bound, see for example the computation of the breakdown point of M-estimators in [HR09] and the gross-error sensitivity in [HRRS86].

From Lemma 20, if $8V_\varepsilon \leq \beta^2 < \infty$, there exists an absolute constant $C > 0$ such that, for all $\lambda \in (0, \lambda_{max})$, with probability larger than $1 - 4\exp(-\lambda) - \exp(-n/8)$,

$$\left\|T_H(P) - T_H(\widehat{P}_n)\right\| \leq 6\frac{V_\varepsilon^{1/2}}{\sqrt{n}} + 8\sqrt{v_\varepsilon\frac{\lambda}{n}} + \frac{C}{n}\lambda\beta. \tag{4.4.11}$$

Where λ_{max} is such that

$$\frac{3V_\varepsilon^{1/2}}{2\sqrt{n}} + 2\sqrt{v_\varepsilon \frac{\lambda_{\max}}{n}} + \frac{C}{n}\lambda_{\max}\beta \leq \frac{\beta}{2}$$

which is verified by $\lambda_{max} \lesssim n$ when $\beta \geq 8V_\varepsilon$. A direct corollary of Lemma 22 and equation (4.4.11) in the case of symmetric P and choosing $\beta^2 \lesssim V_H$, we have the following.

Corollary 10. *Suppose P is a symmetric distribution in $\mathbb{R}^d$, let $X_1, \dots, X_n$ be i.i.d random variables with law P_ε. Suppose $V_H < \infty$. For all $0 < \lambda \lesssim n$, with probability larger than $1 - 4\exp(-\lambda) - \exp(-n/8)$,*

$$\left\| T_H(P) - T_H(\widehat{P}_n) \right\| \lesssim \frac{\sqrt{V_H}}{\sqrt{n}} + \sqrt{\frac{v_H \lambda}{n}} + \frac{\lambda\sqrt{V_H}}{n} + \varepsilon\sqrt{V_H}. \tag{4.4.12}$$

Then, from Lemma 22 and Lemma 13, we get

$$\|T_H(P_\varepsilon) - \mathbb{E}[X]\| \leq 4\beta\varepsilon + \frac{2\mathbb{E}[\|X - \mathbb{E}[X]\|^q]}{(q-1)\beta^{q-1}}. \tag{4.4.13}$$

From equation (4.4.11) simplified using equations (4.4.8) and (4.4.9) and using equation (4.4.13), we get the following lemma.

Lemma 23. *Suppose P is a distribution in $\mathbb{R}^d$ with covariance matrix Σ, let $X_1, \dots, X_n$ be i.i.d random variables with law P_ε. Suppose $8(V_H + \varepsilon\beta^2) \leq \beta^2 < \infty$, $\mathbb{E}[\|X\|^q] < \infty$ for some $q > 1$ and Assumptions 3 are satisfied, then there exists an absolute constant $C > 0$ such that, for all $0 \leq \lambda \lesssim n$, with probability larger than $1 - 4\exp(-\lambda) - \exp(-n/8)$,*

$$\left\| \mathbb{E}[X] - T_H(\widehat{P}_n) \right\| \leq 6\sqrt{\frac{V_H + \varepsilon\beta^2}{n}} + 8\sqrt{\frac{(v_H + \varepsilon\beta^2)\lambda}{n}} + \frac{C}{n}\lambda\beta + 4\beta\varepsilon + \frac{2\mathbb{E}[\|X - \mathbb{E}[X]\|^q]}{(q-1)\beta^{q-1}}. \tag{4.4.14}$$

For the sake of comparison with the state of the art, we can also simplify Lemma 23 as the following theorem proved in Section 4.7.2.

Theorem 27. *Suppose P is a distribution in $\mathbb{R}^d$ with covariance matrix Σ, let $X_1, \dots, X_n$ be i.i.d random variables with law P_ε. Suppose that $\varepsilon \leq 1/16$, $n \geq 16^{q-1}\frac{Tr(\Sigma)^q}{\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^2}$, $\mathbb{E}_P[\|X\|^3] < \infty$ and Assumptions 3 are satisfied, then for all $0 < \lambda \lesssim n$, with probability larger than $1 - 4\exp(-\lambda) - \exp(-n/8)$,*

$$\left\| \mathbb{E}[X] - T_H(\widehat{P}_n) \right\| \lesssim \left(\frac{\sqrt{Tr(\Sigma)} + \sqrt{\|\Sigma\|_{op}\lambda}}{\sqrt{n}} \bigvee \mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q} \varepsilon^{1-1/q} \right) g\left(\lambda, \frac{\varepsilon^{1/2-1/q}}{M}, \frac{1}{M^{\frac{q}{2}} n^{\frac{q-2}{4}}} \right).$$

Where $M = \frac{\sqrt{Tr(\Sigma)}}{\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}}$, and $g : \mathbb{R}^3 \to \mathbb{R}_+$ is such that $g(\lambda, x, y) = 1 + o(\lambda) + o(x) + o(y)$ for (λ, x, y) that tend to 0.

Notice that ε is multiplied by a quantity that increases with the dimension in general, this bound is not minimax at least in the case of Gaussian inliers, see [DL19] who achieve a sharper bound. However, we see that the bound is of order $\varepsilon^{1-1/q}$ which is what is found for example in [HL19], one highlight of this section however is that in Corollary 10, it was not Gaussian inliers that were needed to have a dependency in ε but only symmetric inliers, from Corollary 10.

4.4.4 Application to the concentration of HOME

Let $X_1, \dots, X_n$ be i.i.d random variables on $\mathbb{R}^d$, let $K \in \mathbb{N}$ and suppose that K divides n. Let $B_1, \dots, B_K$ be a partition of $\{1, \dots, n\}$ and $b \in \mathbb{N}^*$ be such that

$$\forall k \in \{1, \dots, K\}, \quad |B_k| = b, \quad \forall k \neq j, \quad B_k \cap B_j = \emptyset \quad \text{and} \quad \cup_{k=1}^K B_k = \{1, \dots, n\}$$

We define Huber of Means Estimator as HOME_K, solution of

$$\frac{1}{K}\sum_{k=1}^{K} \frac{\sum_{i\in B_k}(X_i - \text{HOME}_K(X_1^n))}{\left\|\sum_{i\in B_k}(X_i - \text{HOME}_K(X_1^n))\right\|} \psi\left(\frac{1}{b\beta}\left\|\sum_{i\in B_k}(X_i - \text{HOME}_K(X_1^n))\right\|\right) = 0$$

The theoretical counterpart of $\text{HOME}_K(X_1^n)$ will here be $T_H(P_B)$ where P_B is the law of the empirical mean $\frac{1}{b}\sum_{i=1}^b X_i$.

Lemma 24. *Suppose that X has a finite second moment, then there exists a constant C' that does not depend on q or b such that,*

$$\|\mathbb{E}[X] - T_H(P_B)\| \leq C' q \frac{\sqrt{bTr(\Sigma)} + b^{1/q}\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}}{(q-1)\beta^{q-1}}.$$

The proof is provided in Section 4.8.7. This takes care of the bias term Then we can control the variance term in the concentration inequality, we have by elementary equalities that $\mathbb{E}[\|\frac{1}{b}\sum_{i=1}^b (X_i - \mathbb{E}[X])\|^2] = Tr(\Sigma)/b$ and

$$\left\|\mathbb{E}\left[\left(\frac{1}{b}\sum_{i=1}^b (X_i - \mathbb{E}[X])\right)\left(\frac{1}{b}\sum_{i=1}^b (X_i - \mathbb{E}[X])\right)^T\right]\right\|_{op} = \frac{\|\Sigma\|_{op}}{b}.$$

Then, from Lemma 20, Lemma 15 and Lemma 24, we can obtain a lemma that gives the deviations of $\text{HOME}_K(X_1^n)$ as a function of β and like before we choose β accordingly to get the following lemma.

Lemma 25. *Let $X_1, \dots, X_n$ be i.i.d random variables with law P on $\mathbb{R}^d$ with $\mathbb{E}[\|X\|^q] < \infty$ for some $q \geq 2$. If $\beta^2 \geq 8Tr(\Sigma)/b$, then there exist absolute constants $C > 0$ such that, for all $0 < \lambda \lesssim K$, with probability larger than $1 - 4\exp(-\lambda) - \exp(-K/8)$,*

$$\begin{aligned}\|\text{HOME}_K(X_1^n) - \mathbb{E}[X]\| &\leq 12\frac{\sqrt{Tr(\Sigma)}}{\sqrt{n}} + 16\sqrt{\|\Sigma\|_{op}\frac{\lambda}{n}} \\ &+ \frac{C\lambda q^{\frac{q}{q-1}}}{\sqrt{n}}\left(\frac{\sqrt{Tr(\Sigma)}}{K^{\frac{q-2}{2q}}} + K^{\frac{1}{2q}}\frac{\mathbb{E}[\|X_i - \mathbb{E}[X]\|^q]^{1/q}}{n^{\frac{1}{2}-\frac{1}{q}}}\right)^{\frac{q}{q-1}}\left(\frac{1}{Tr(\Sigma)}\right)^{\frac{1}{2(q-1)}}. \quad (4.4.15)\end{aligned}$$

Lemma 25 shows that the parameter K will allow a trade-off between $\sqrt{Tr(\Sigma)}$ and $\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}$. When K increases, the bias term that depends on $\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}$ increases while the bias term that depends on $Tr(\Sigma)$ decreases in equation (4.4.15). The difference between $\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}$ and $\sqrt{Tr(\Sigma)}$ will strongly depend on the tail of the distribution, if the distribution is light tailed (for example Gaussian, see Remark 1) then $\sqrt{Tr(\Sigma)}$ is not very far from $\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}$ but if the distribution is heavy tailed $\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}$ can be much larger than $\sqrt{Tr(\Sigma)}$ and making blocks can be interesting.

4.4.5 Comparison HOME and Huber on Stable distributions

As said previously in a comment to Lemma 25, when the q^{th} central moment of X is much bigger than its second central moment, it can be interesting to make blocks. There is another situation when it can be interesting to make blocks and that is when the distribution is very skewed. It can be seen in the example of stable distributions in dimension 1.

Let $X_1, \dots, X_n$ be i.i.d sample from a stable distribution with the characteristic function of X_i defined for some $\alpha \in (1,2)$, $c > 0$ and $\gamma \in [-1,1]$,

$$\forall y \in \mathbb{R}, \quad \phi_X(y) = \exp\Big(-|cy|^\alpha \Big(1 - i\gamma \text{sign}(y) \tan\Big(\frac{\pi\alpha}{2}\Big)\Big)\Big)$$

it is known (see [Fel]) that with this choice of α, the mean of X_i is 0, the variance is infinite and if $\gamma \neq 1$, the distribution is skewed. This constitutes an example of a heavy-tailed distribution which does not have a finite second moment. We have the following property of stable distribution: if $x \mapsto f(x)$ is the density of X_i and $1 \leq b \leq n$ is an integer, then $\frac{1}{b}\sum_{i=1}^b X_i$ has the density $y \mapsto b^{\frac{\alpha-1}{\alpha}} f\Big(y b^{\frac{\alpha-1}{\alpha}}\Big)$.

Then, we can easily compare the bias of HOME with Huber's estimator (we compare the distance to the mean which is here 0). Let $\beta_1 > 0$ and $\theta \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\left[\beta_1 \text{sign}\left(\frac{1}{b}\sum_{i=1}^b X_i - \theta\right)\psi\left(\frac{|\frac{1}{b}\sum_{i=1}^b X_i - \theta|}{\beta_1}\right)\right] &= \mathbb{E}\left[\beta_1 b^{\frac{\alpha-1}{\alpha}} \text{sign}\Big(X b^{-\frac{\alpha-1}{\alpha}} - \theta\Big)\psi\left(\frac{|X b^{-\frac{\alpha-1}{\alpha}} - \theta|}{\beta_1}\right)\right] \\ &= \mathbb{E}\left[\beta_2 \text{sign}(X - \theta b^{\frac{\alpha-1}{\alpha}}\theta)\psi\left(\frac{|X - \theta b^{\frac{\alpha-1}{\alpha}}\theta|}{\beta_2}\right)\right],\end{aligned}$$

where we used $\beta_2 = \beta_1 b^{\frac{\alpha-1}{\alpha}}$.

Then, based on the definition of HOME and Huber's estimator, we have that

$$\text{HOME}_{K,\beta_1}(X) = b^{-\frac{\alpha-1}{\alpha}} T_{H,\beta_2}(P).$$

As expected the procedure of using blocks decreases the bias. Moreover, we have

$$\mathbb{E}\left[\beta_1^2 \psi\left(\frac{|\frac{1}{b}\sum_{i=1}^b X_i - \text{HOME}_{K,\beta_1}(X_1^n)|}{\beta_1}\right)^2\right] = b^{-\frac{\alpha-1}{\alpha}}\mathbb{E}\left[\beta_2^2\psi\left(\frac{|X - T_{H,\beta_2}(P)|}{\beta_2}\right)^2\right] = b^{-\frac{\alpha-1}{\alpha}} V_{H,\beta_2}.$$

Then, using the concentration inequality from Lemma 17, we get the following statistical guarantees for Huber's estimator and HOME:

Huber's estimator : suppose $8V_{H,\beta_2} \leq \beta_2^2$. For all $t > 0$ such that $4\sqrt{2V_{H,\beta_2}t/n} + 4\beta_2 t/n \leq \beta_2/2$, with probability greater than $1 - 2e^{-t} - e^{-n/8}$,

$$\Big|T_{H,\beta_2}(\widehat{P}_n) - \mathbb{E}[X]\Big| \leq 4\sqrt{\frac{2V_{H,\beta_2}t}{n}} + 4\frac{\beta_2 t}{n} + |T_{H,\beta_2}(P)|. \tag{4.4.16}$$

HOME : suppose $8V_{H,\beta_2} \leq \beta_2^2 b^{-(\alpha-1)/\alpha}$. For all $t > 0$ such that $4\sqrt{2V_{H,\beta_2}tb^{\frac{\alpha-1}{\alpha}}/K} +$

$4\beta_2 t/K \leq \beta_2/2$, with probability greater than $1 - 2e^{-t} - e^{-K/8}$,

$$|\mathrm{HOME}_{K,\beta_1}(X_1^n) - \mathbb{E}[X]| \leq 4\sqrt{\frac{2V_{H,\beta_2}t}{Kb^{\frac{\alpha-1}{\alpha}}}} + 4\frac{\beta_2 t}{Kb^{\frac{\alpha-1}{\alpha}}} + \frac{1}{b^{\frac{\alpha-1}{\alpha}}}|T_{H,\beta_2}(P)|. \tag{4.4.17}$$

By using blocks, we increase the variance term because $Kb^{(\alpha-1)/\alpha} = n^{1-1/\alpha}K^{1/\alpha} < n$, we also decrease the bias term and decrease the probability with which we make the decision. Hence, it can be interesting to make blocks when T_H is very biased and n large, which happens when the distribution is very skewed with γ close to 1 and increasing c increase the bias while we always have $V_{H,\beta} \leq \beta^2$.

Then, we conjecture that more generally, there is interest in making blocks of data when the distribution is at the same time heavy tailed (so that the bias term is not negligible compared to the variance term, see Section 4.4.1) and also very skewed (so that the bias is large) and the sample size is large. On the other hand, if the distribution is symmetric we advise to use Huber's estimator instead of HOME.

4.5 Algorithm and choice of β

4.5.1 Algorithm: iterative re-weighting

To compute $T(\widehat{P}_n)$, we use an iterative re-weighting algorithm. This algorithm is rather well known to compute M-estimators, see [HR09, Section 7] and it has already been extensively studied. The principle is to rewrite the definition of $T(\widehat{P}_n)$ from equation (4.1.3) as

$$T(\widehat{P}_n)\sum_{i=1}^n \frac{\psi\left(\frac{\|X_i - T(\widehat{P}_n)\|}{\beta}\right)}{\|X_i - T(\widehat{P}_n)\|} = \sum_{i=1}^n X_i \frac{\psi\left(\frac{\|X_i - T(\widehat{P}_n)\|}{\beta}\right)}{\|X_i - T(\widehat{P}_n)\|},$$

then, denote $w_i = \frac{\beta\psi\left(\frac{\|X_i - T(\widehat{P}_n)\|}{\beta}\right)}{\|X_i - T(\widehat{P}_n)\|}$, we get an expression of $T(\widehat{P}_n)$ as a weighted sum:

$$T(\widehat{P}_n) = \sum_{i=1}^n X_i \frac{w_i}{\sum_{i=1}^n w_i}.$$

The weights w_i depend on $T(\widehat{P}_n)$ and the principle of the algorithm is as follows. Initialize θ_0 with the coordinate-wise median and iterate the following

$$w_i^{(m)} = \frac{\beta\psi\left(\frac{\|X_i - \theta^{(m)}\|}{\beta}\right)}{\|X_i - \theta^{(m)}\|}$$

$$\theta^{(m+1)} = \sum_{i=1}^n X_i \frac{w_i^{(m)}}{\sum_{i=1}^n w_i^{(m)}}.$$

We show that this algorithm allows us to find a minimizer of

$$Z_n(\theta) = \frac{1}{n}\sum_{i=1}^n \rho\left(\frac{\|X_i - \theta\|}{\beta}\right).$$

Lemma 26. *Assume that ρ is convex, that $x \mapsto \psi(x)/x$ is bounded and decreasing, then unless $\theta^{(m)}$ is the minimizer of Z_n we have that $Z_n(\theta^{(m+1)}) < Z_n(\theta^{(m)})$.*

The proof is provided in Section 4.8.9. Then, from Lemma 26 and as Z_n is non-negative, we have that the sequence $Z_n(\theta^{(m)})$ converge to its minimum $Z_n(T(\widehat{P}_n))$, hence $\theta^{(m)}$ converges to $T(\widehat{P}_n)$. The Lemma does not provide the number of iterations needed to attain a certain precision; it only provides the convergence, in practice the algorithm seems to converge in a few iterations.

4.5.2 Choice of β: Lepski's method

In this section because we will often change the β parameter, we indice the quantities that depend on a parameter β by this parameter, T_H becomes $T_{H,\beta}$, ψ_H becomes $\psi_{H,\beta}$ and V_H becomes $V_{H,\beta}$. We define the following quantity

$$\widehat{V}_\beta = \frac{1}{n(n-1)} \sum_{i \neq j} \psi_{H,\beta}(X_i - X_j)^2,$$

Let $\mathcal{B}$ be a finite grid of $[0, \beta_{max}]$ where β_{max} is the solution of $\beta = \sqrt{2n\widehat{V}_\beta}$ (this upper bound is dictated by the theory so as not to lower the rate of convergence of $T_{H,\beta}(\widehat{P}_n)$, see equation (4.4.5)). Define

$$\widehat{I}_\beta(t) = \left[T_{H,\beta}(\widehat{P}_n) - 4\sqrt{\frac{\widehat{V}_\beta t}{n} + \frac{\sqrt{2}t^{3/2}\beta^2}{n^{3/2}}} - \frac{4\beta t}{n} - \frac{2\sigma^2}{\beta}, T_{H,\beta}(\widehat{P}_n) + 4\sqrt{\frac{\widehat{V}_\beta t}{n} + \frac{\sqrt{2}t^{3/2}\beta^2}{n^{3/2}}} + \frac{4\beta t}{n} + \frac{2\sigma^2}{\beta} \right].$$

So that

$$\widehat{I}_\beta(t/\beta^2) = \left[T_{H,\beta}(\widehat{P}_n) - 4\sqrt{\frac{\widehat{V}_\beta t}{n\beta^2} + \frac{\sqrt{2}t^{3/2}}{\beta n^{3/2}}} - \frac{4t}{\beta n} - \frac{2\sigma^2}{\beta}, T_{H,\beta}(\widehat{P}_n) + 4\sqrt{\frac{\widehat{V}_\beta t}{n\beta^2} + \frac{\sqrt{2}t^{3/2}}{\beta n^{3/2}}} + \frac{4t}{\beta n} + \frac{2\sigma^2}{\beta} \right]. \tag{4.5.1}$$

and finally define the estimator of β given by

$$\widehat{\beta}_t = \max\left\{ \beta \in \mathcal{B} : \cap_{\substack{b \in \mathcal{B} \\ b \leq \beta}} \widehat{I}_b(t/b^2) \neq \emptyset \right\} \tag{4.5.2}$$

We have the following lemma

Lemma 27. *Let $X_1, \ldots, X_n$ be i.i.d from a distribution P and suppose that Assumptions 3 hold, let $\widehat{\beta}_t$ be constructed from equation (4.5.2), then*

$$\mathbb{P}\left(|\widehat{T}_{H,\widehat{\beta}_t}(P) - \mathbb{E}[X]| > \inf_{\beta \in \mathcal{B}} 8\sqrt{\frac{\widehat{V}_\beta t}{\beta^2 n} + \frac{\sqrt{2}t^{3/2}}{\beta n^{3/2}}} + \frac{8t}{\beta n} + \frac{2\sigma^2}{\beta} \right) \leq \sum_{b \in \mathcal{B}} \left(4e^{-t/b^2} + e^{-n/8} \right)$$

From Lemma 27, we can use $\widehat{\beta}_t$ to choose β adaptively. There is still a parameter that has to be set and that is the value of t. The choice of the parameter t has some impact on the algorithm, theoretically, a large value of t gets results that are less accurate than for small values but with

a higher confidence. In numerical applications, we see that a choice of t between 1 and 10 was suitable for most distributions.

We may use Lepski's method in a multi-dimensional setting but it would add a lot of computational time, and the estimation of $\|\Sigma\|_{op}$ is not easy and not efficient in practice and moreover the added complexity would make the algorithm computationally intensive.

4.6 Numerical illustrations

The code for this section can be found on github: https://github.com/TimotheeMathieu/RobustMeanEstimator/.

4.6.1 Numerical illustration of Lepski's method in dimension 1

In this section, we apply Lepski's method to the problem of the estimation of the mean in one dimension. If we look carefully at the proof of Lemma 27 and Lemma 13, we see that in fact the term σ^2/β is not optimal as soon as the distribution has more than 2 moments and in fact in practice, we witnessed that it is much more efficient to use

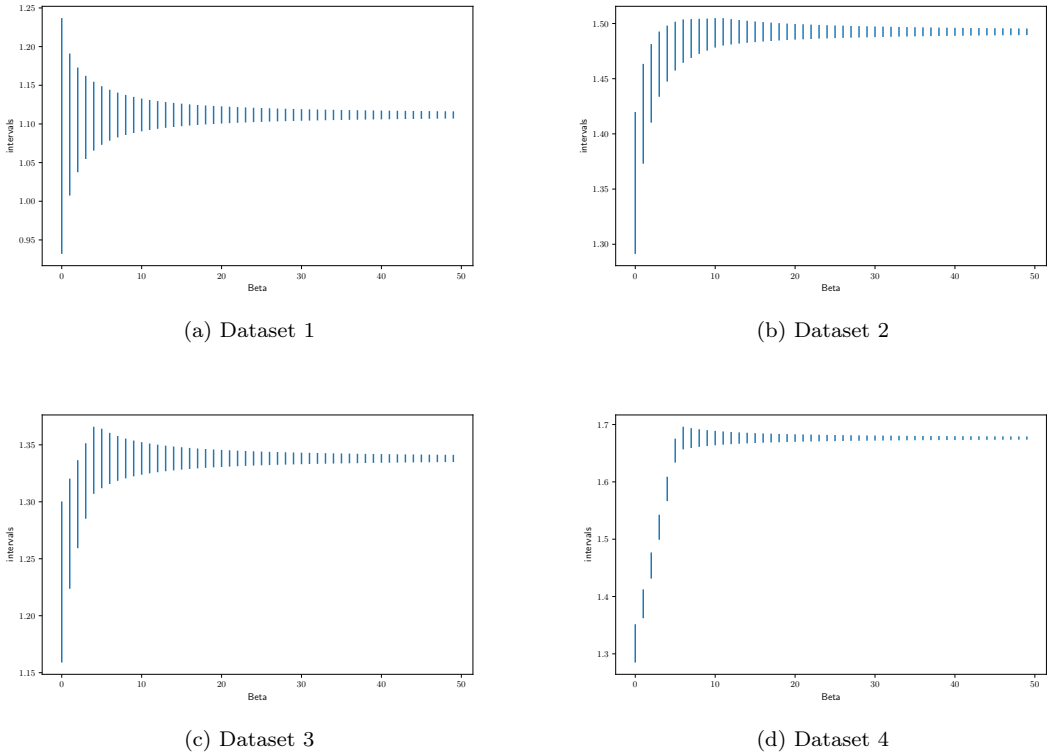
$$\tilde{I}_\beta(t/\beta^2) = \left[T_{H,\beta}(\widehat{P}_n) - 4\sqrt{\frac{\widehat{V}_\beta t}{n\beta^2} + \frac{\sqrt{2}t^{3/2}}{\beta n^{3/2}}} - \frac{4t}{\beta n}, T_{H,\beta}(\widehat{P}_n) + 4\sqrt{\frac{\widehat{V}_\beta t}{n\beta^2} + \frac{\sqrt{2}t^{3/2}}{\beta n^{3/2}}} + \frac{4t}{\beta n}\right],$$

instead of $\widehat{I}_\beta(t/\beta)$ defined in equation (4.5.1). Although Lemma 27 does not use these intervals, in fact we can see (using the results from Lemma 21) that the bias term σ^2/β is negligible compared to the other terms for a large choice of β and in practice this is much more efficient.

To compute the estimator $T(\widehat{P}_n)$ we use an iterated reweighting algorithm (see Section 4.5.1), we use a grid $\mathcal{B}$ of 50 points linearly spaced in $[0, \beta_{max}]$ (β_{max} is computed using newton algorithm). We use simulated examples so that we can sample at will from the initial distribution and use Monte-Carlo simulation. We consider 3 datasets:

- **Dataset 1:** $X_1, \dots, X_n$ are i.i.d from a Pareto distribution with shape parameter 10 and scale parameter 1.
- **Dataset 2:** $X_1, \dots, X_n$ are i.i.d from a Pareto distribution with shape parameter 3 and scale parameter 1.
- **Dataset 3:** $X_1, \dots, X_n$ are i.i.d from a mildly corrupted Pareto distribution with shape parameter 5 and scale parameter 1: $X_i \sim (1-\varepsilon)Pareto(5,1) + \varepsilon\delta_3$, $\varepsilon = 0.05$.
- **Dataset 4:** $X_1, \dots, X_n$ are i.i.d from a highly corrupted Pareto distribution with shape parameter 5 and scale parameter 1: $X_i \sim (1-\varepsilon)Pareto(5,1) + \varepsilon\delta_{10}$, $\varepsilon = 0.05$

The different datasets represent respectively a light-tailed dataset, a heavy-tailed dataset, a dataset with mild corruption and a dataset with heavy corruption. Figure 4.1 show the intervals



(a) Dataset 1

(b) Dataset 2

(c) Dataset 3

(d) Dataset 4

Figure 4.1: Plot of the intervals $I_\beta(t/\beta)$ for β in $\mathcal{B}$.

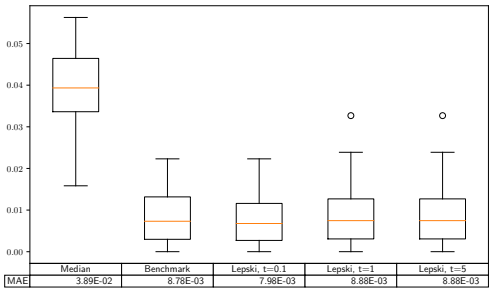
$I_\beta(t/\beta)$ for $\beta \in \mathcal{B}$ and $t = 5$. As expected the intervals will quickly have an empty intersection on corrupted datasets (Datasets 3 and 4) whereas on non-corrupted data sets the value of β chosen by the procedure will be much larger and it even attain the upper bound of β_{max} in Dataset 1.

Let $n = 100, M = 500$ and we sample $(X_{i,j})_{\substack{1\leq i\leq n \\ 1\leq j\leq M}}$ i.i.d from one of the four datasets and on each sample we compute the absolute deviation from the theoretical mean

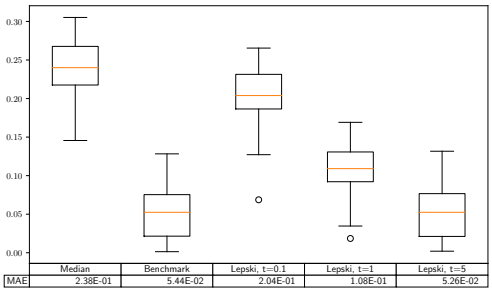
$$\left| T\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{X_{i,j}}\right) - \mu \right|,$$

where μ is the expectation if the dataset is not corrupted (Dataset 1 and 2) and μ is the expectation of the inliers when the dataset is corrupted (Dataset 3 and 4). Figure 4.2 summarizes these results using boxplots. The Benchmark estimator is Huber estimator where β is selected using a grid search (this is an oracle estimator normally not accessible to the statistician if we can't simulate according to the estimator and if we don't know the theoretical mean), lepski estimators are Huber estimators where β is tuned as described before. In Figure 4.2 we see that we can't use a value of t that is good for all the examples, if the corruption is very high a small value of t is preferable and if the dataset is not corrupted a high value of t may be more efficient. We may think that this procedure is not useful as we still have a parameter to choose but experience

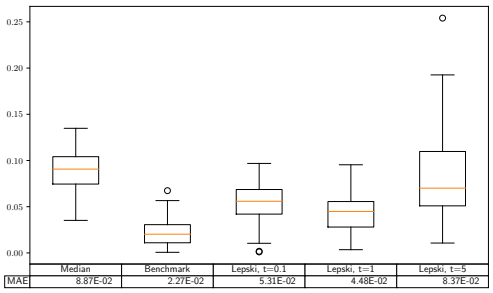
shows that t is much easier to tune than β.



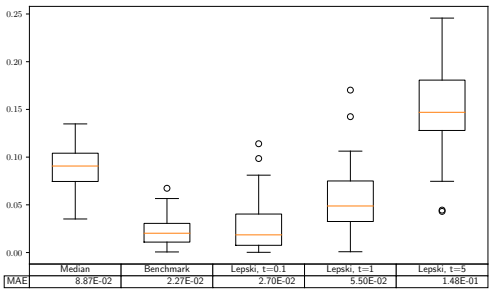
(a) Dataset 1



(b) Dataset 2



(c) Dataset 3



(d) Dataset 4

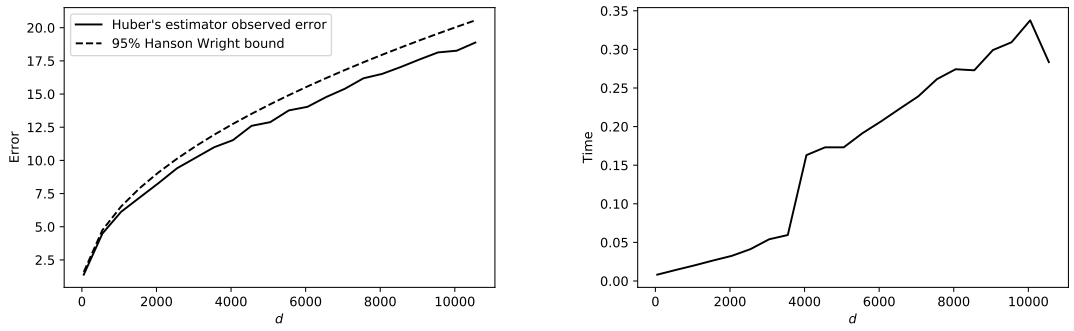
Figure 4.2: Boxplots of the MAE of the median and of Huber estimator with β chosen with grid search (benchmark) and Huber estimator with β chosen using Lepski's method (for $t \in \{0.1, 1, 5\}$). MAE stands for mean absolute error.

4.6.2 Numerical illustration in dimension $d > 1$

In this section, we apply our algorithm in a high dimensional setting up to $d = 10\,000$. We consider two datasets, first is a multivariate Gaussian Dataset with outliers and second is a skewed, heavy-tailed and corrupted dataset. In this section, all the data is simulated and the parameters β and K are tuned to minimize the error (we used additional simulations of the dataset to tune the parameters).

Dataset 5. In this dataset, we consider $X_1, \dots, X_{n-5}$ i.i.d from a standard normal and $X_{n-4}, \dots, X_n$ i.i.d from $\mathcal{N}(10 \cdot \mathbf{1}, I_d)$ with $n = 50$ and $\mathbf{1}$ the vector with all the coordinates set to 1. We plot the distance of $T(\widehat{P}_n)$ to 0 as a function of the dimension d. For comparison purposes, we also plot Hanson-Wright bound which is here $\sqrt{2d/n + 9t/n}$ where $t = \log(1/0.05)$ is chosen to get a 95% confidence into Hanson-Wright inequality. The result is provided in Figure 4.3.

Also included in Figure 4.3 is a plot of the computation time on a i9 CPU (Intel(R) Core(TM) i9-9980HK CPU @ 2.40GHz).



(a) Dataset 5

(b) Computation time on Dataset 5

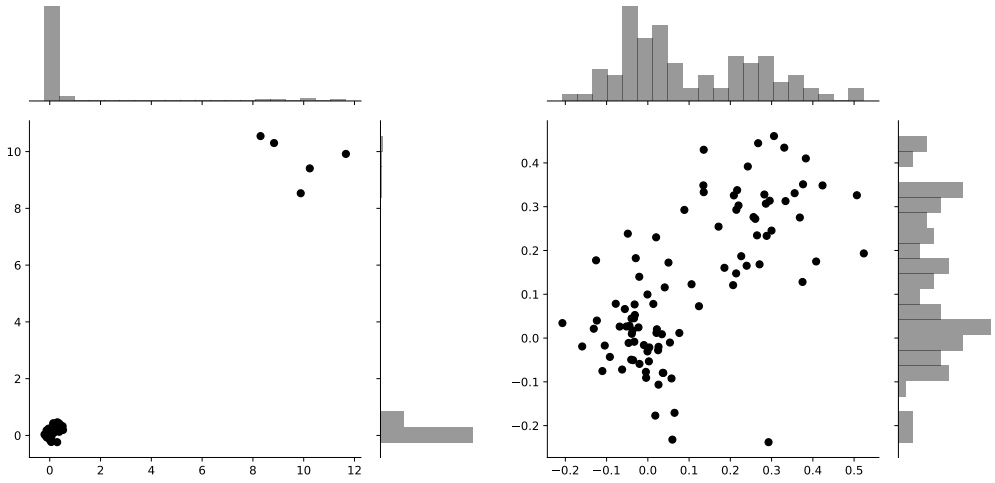
Figure 4.3: In Figure 4.3a: plot of the error $\|T(\widehat{P}_n) - \mathbb{E}[X]\|$ for one different sample for each point as a function of the dimension d and Hanson-Wright bound. Figure 4.3b: plot of the mean computation time in second on 100 runs on Dataset 5.

Figure 4.3a shows that the error evolves similarly to Hanson-Wright inequality which confirms that the estimator is minimax in this context. Figure 4.3b show the computation time of the algorithm which is rather fast, even in dimension $d = 10\,000$, and although the theory does not assure convergence in a few steps, the plot seems to indicate that the algorithm converges in a few steps in practice and that the complexity is linear in d.

Dataset 6. In this dataset, we consider $X_1, \dots, X_{n-5}$ i.i.d from a mixture of two multivariate student distributions. $X \sim 0.6t_6(0, \Sigma) + 0.4t_6(0.3 \cdot \mathbf{1}, \Sigma)$ where $t_\nu(\mu, \Sigma)$ is a multivariate t distribution with ν degree of freedom and parameters (μ, Σ). Σ has been sampled from an inverse Wishart distribution, its trace (on this example) is $Tr(\Sigma) \simeq 0.99$ and its operator norm is $\|\Sigma\|_{op} \simeq 0.03$. The mean of this distribution is $(0.4 \times 0.3) \cdot \mathbf{1}$. Finally, $X_{n-4}, \dots, X_n$ are outliers sampled from $\mathcal{N}(10 \cdot \mathbf{1}, I_d)$. The dimension $d = 200$ and the number of points is $n = 100$.

Figure 4.4 represents the first two coordinates of one sample dataset and a zoom on the inliers. To measure the performance of an estimator, as previously, we will look at the distance (in euclidean norm) to the mean of the inliers over $M = 30$ runs, the result is presented in Figure 4.5, the estimators considered are the empirical mean, the coordinate-wise median, Huber's estimator and HOME.

In Figure 4.5 we see that the empirical mean performs worst on this task, this is not surprising because it is the only non-robust estimator. Then, the coordinate-wise median comes second because of its large bias. Finally, HOME is a little bit better than Huber's estimator which is within expectation because the dataset is skewed and Heavy-tailed which favors HOME as said in Section 4.4.4. We can't really compare to Hanson-Wright inequality because we are not in a Gaussian setting, on the other hand, we can compute what is the error of the empirical mean computed only on the inliers because we know which points are the inliers. Doing this, we obtain an MAE $\simeq -2.2$ hence our estimator seems to attain what was announced which is to be as good



(a) Dataset 6 (b) Dataset 6 zoomed

Figure 4.4: Plot of the first two coordinates of one realization of Dataset 6.

as the empirical mean would be if the data were not corrupted.

4.7 Proof of Theorems

4.7.1 Proof of Theorem 26

The function $\theta \mapsto Z_\beta(\theta)$ is differentiable and by the mean value theorem, we have

$$\|Z_\beta(\mathbb{E}[X]) - Z_\beta(T(P))\| \geq \|\mathbb{E}[X] - T(P)\| \inf_{t\in[0,1]} \|Jac(Z_\beta)(t\mathbb{E}[X] + (1-t)T(P))\|_{op} \quad (4.7.1)$$

Where Jac denotes the Jacobian matrix that we control with the following lemma.

Lemma 28. *Let $u \in S^{d-1}$ and $\theta \in \mathbb{R}^d$,*

$$u^T Jac(Z_\beta)(\theta) u \leq -\mathbb{E}\left[\psi'\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right]$$

Proof.

$$\begin{aligned} Jac(Z_\beta)(\theta) = &- \mathbb{E}\left[\beta \frac{I_d}{\|X-\theta\|}\psi\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right] + \mathbb{E}\left[\beta\frac{(X-\theta)(X-\theta)^T}{\|X-\theta\|^3}\psi\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right] \\ &- \mathbb{E}\left[\frac{(X-\theta)(X-\theta)^T}{\|X-\theta\|^2}\psi'\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right] \end{aligned}$$

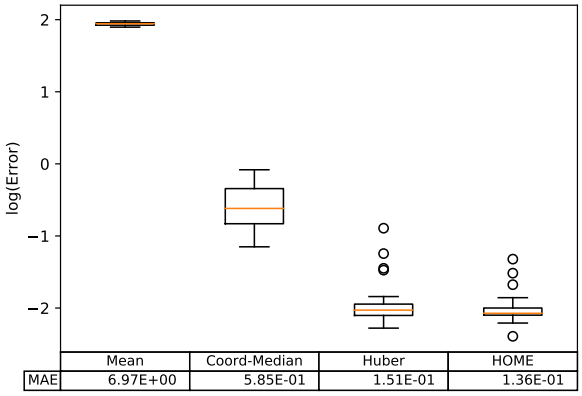


Figure 4.5: Boxplots of the performance of different estimators on Dataset 6. MAE stands for mean absolute error.

Then, for all $u \in S^{d-1}$,

$$\begin{aligned} u^T Jac(Z_\beta)(\theta)u =& -\mathbb{E}\left[\beta\frac{u^T u}{\|X-\theta\|}\psi\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right] + \mathbb{E}\left[\beta\frac{u^T(X-\theta)(X-\theta)^T u}{\|X-\theta\|^3}\psi\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right] \\ & - \mathbb{E}\left[\frac{u^T(X-\theta)(X-\theta)^T u}{\|X-\theta\|^2}\psi'\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right] \\ =& -\mathbb{E}\left[\beta\frac{1-\langle\frac{X-\theta}{\|X-\theta\|},u\rangle^2}{\|X-\theta\|}\psi\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right] - \mathbb{E}\left[\frac{\langle X-\theta,u\rangle^2}{\|X-\theta\|^2}\psi'\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right] \end{aligned}$$

then, because ψ is concave and $\psi(0)=0$, we have that $\forall y \geq 0$, $\psi(y) \geq y\psi'(y)$ and by Cauchy-Schwarz inequality we have that $1 - \left\langle \frac{X-\theta}{\|X-\theta\|}, u \right\rangle^2 \geq 0$. Hence,

$$\begin{aligned} u^T Jac(Z_\beta)(\theta)u \leq& -\mathbb{E}\left[\left(1-\left\langle\frac{X-\theta}{\|X-\theta\|},u\right\rangle^2\right)\psi'\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right] - \mathbb{E}\left[\frac{\langle X-\theta,u\rangle^2}{\|X-\theta\|^2}\psi'\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right] \\ =& -\mathbb{E}\left[\psi'\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right] \end{aligned}$$

■

Hence, for all $u \in S^{d-1}$ and $\theta \in \mathbb{R}^d$, because $\psi'(x) \geq 0$ for $x \geq 0$,

$$|u^T Jac(Z_\beta)(\theta)u| \geq \mathbb{E}\left[\psi'\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right],$$

which implies

$$\|Jac(Z_\beta)(\theta)\|_{op} \geq \mathbb{E}\left[\psi'\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right].$$

Then, by assumption on ψ,

$$\|Jac(Z_\beta)(\theta)\|_{op} \geq \gamma\mathbb{P}(\|X-\theta\| \leq \beta)$$

Hence, for all $t \in [0,1]$,

$$\begin{aligned}\|Jac(Z_\beta)(t\mathbb{E}[X] + (1-t)T(P))\|_{op} &\geq \gamma\mathbb{P}(\|X - t\mathbb{E}[X] - (1-t)T(P)\| \leq \beta) \\ &\geq \gamma\mathbb{P}(\|X - T(P)\| \leq \beta - \|\mathbb{E}[X] - T(P)\|)\end{aligned}$$

Then, use the following lemma proven in Section 4.8.11.

Lemma 29. *If* $\rho(1/3) \geq \mathbb{E}[\rho(\|X - \mathbb{E}[X]\|/\beta)]$, *then* $\|\mathbb{E}[X] - T(P)\| \leq \frac{\beta}{3}$.

We get,

$$\begin{aligned}\|Jac(Z_\beta)(t\mathbb{E}[X] + (1-t)T(P))\|_{op} &\geq \gamma\mathbb{P}(\|X - T(P)\| \leq 2\beta/3) \\ &= \gamma\mathbb{P}\left(\rho\left(\frac{\|X - T(P)\|}{\beta}\right) \leq \rho(2/3)\right) \\ &\geq \gamma\mathbb{P}\left(\rho\left(\frac{\|X - T(P)\|}{\beta}\right) \leq 2\rho(1/3)\right)\end{aligned}$$

because ρ is increasing and super-additive on $\mathbb{R}_+$ (ρ is increasing because $\psi(0) = 0$ and ψ is non-decreasing because $\psi' \geq 0$, hence $\psi = \rho' \geq 0$). Hence, by Markov's inequality and using the hypothesis,

$$\|Jac(Z_\beta)(t\mathbb{E}[X] + (1-t)T(P))\|_{op} \geq \gamma\left(1 - \frac{\mathbb{E}\left[\rho\left(\frac{\|X-T(P)\|}{\beta}\right)\right]}{2\rho(1/3)}\right) \geq \frac{\gamma}{2}$$

Then, from equation (4.7.1), we get the result

For the inequality in the other direction, write that by mean value theorem,

$$\|Z_\beta(\mathbb{E}[X]) - Z_\beta(T(P))\| \leq \|\mathbb{E}[X] - T(P)\| \sup_{t\in[0,1]} \|Jac(Z_\beta(t\mathbb{E}[X] + (1-t)T(P)))\|_{op} \tag{4.7.2}$$

In proof of Lemma 28 we showed that for all $u \in S_d$,

$$|u^T Jac(Z_\beta)(\theta)u| = \left|-\mathbb{E}\left[\beta\frac{1 - \langle\frac{X-\theta}{\|X-\theta\|}, u\rangle^2}{\|X-\theta\|}\psi\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right] - \mathbb{E}\left[\frac{\langle X-\theta, u\rangle^2}{\|X-\theta\|^2}\psi'\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right]\right|$$

hence, by triangular inequality, and Cauchy-Schwarz inequality,

$$|u^T Jac(Z_\beta)(\theta)u| \leq \mathbb{E}\left[\beta\frac{2}{\|X-\theta\|}\psi\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right] + \mathbb{E}\left[\psi'\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right]$$

and finally, using that $\psi' \leq 1$ and hence ψ is 1-Lipshitz and $\psi(0) = 0$, for all $u \in S^{d-1}$,

$$|u^T Jac(Z_\beta)(\theta)u| \leq 3$$

which prove the result by injecting this equation in equation (4.7.2).

4.7.2 Proof of Theorem 27

From Lemma 23, if $8(V_H + \varepsilon\beta^2) \leq \beta^2 < \infty$ and $\mathbb{E}[\|X - T_H(P)\|] \leq \beta/4$, $\mathbb{E}[\|X\|^q] < \infty$ and Assumptions 3 are satisfied, then there exists an absolute constant $C > 0$ such that, for all $\lambda \in (0, \lambda_{max})$, with probability larger than $1 - 4\exp(-\lambda) - \exp(-n/8)$,

$$\left\|\mathbb{E}[X] - T_H(\widehat{P}_n)\right\| \leq 6\sqrt{\frac{Tr(\Sigma)}{n}} + 8\sqrt{\frac{\|\Sigma\|_{op}\lambda}{n}} + \frac{\sqrt{\varepsilon}\beta}{\sqrt{n}}\left(1+\sqrt{\lambda}\right) + \frac{C}{n}\lambda\beta + 4\beta\varepsilon + \frac{\mathbb{E}[\|X - \mathbb{E}[X]\|^q]}{(q-1)\beta^{q-1}}. \tag{4.7.3}$$

using the sub-linearity of the square root and Lemma 15.

Denote $R_\lambda = \sqrt{\frac{Tr(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\|_{op}\lambda}{n}}$, $M = \frac{\sqrt{Tr(\Sigma)}}{\mathbb{E}[\|X-\mathbb{E}[X]\|^q]^{1/q}}$ and choose

$$\beta = \left(\frac{\mathbb{E}[\|X - \mathbb{E}[X]\|^q]\sqrt{n}}{\sqrt{Tr(\Sigma)}}\right)^{1/(q-1)} = \mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}\left(\frac{\sqrt{n}}{M}\right)^{\frac{1}{q-1}}.$$

Then

$$\left\|\mathbb{E}[X] - T_H(\widehat{P}_n)\right\| \lesssim R_\lambda\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{\varepsilon}\left(1+\sqrt{\lambda}\right)n^{\frac{1}{2(q-1)}}}{M^{1+\frac{1}{q-1}}\sqrt{n}} + \frac{\lambda n^{\frac{1}{2(q-1)}}}{M^{1+\frac{1}{q-1}}n} + \frac{n^{\frac{1}{2(q-1)}}\varepsilon}{M^{1+\frac{1}{q-1}}}\right). \tag{4.7.4}$$

Then, we proceed in two parts.
First part: if $\varepsilon^{1-1/q} \lesssim \frac{M}{\sqrt{n}}$.
From equation (4.7.4),

$$\left\|\mathbb{E}[X] - T_H(\widehat{P}_n)\right\| \lesssim R_\lambda\left(\frac{1}{\sqrt{n}} + \frac{M^{\frac{q}{2(q-1)}}\left(1+\sqrt{\lambda}\right)}{M^{\frac{q}{q-1}}n^{\frac{1}{2}-\frac{1}{2(q-1)}+\frac{q}{4(q-1)}}} + \frac{\lambda}{M^{\frac{q}{q-1}}n^{\frac{2q-3}{2q-2}}}\right). \tag{4.7.5}$$

Which simplifies in

$$\left\|\mathbb{E}[X] - T_H(\widehat{P}_n)\right\| \lesssim \frac{R_\lambda}{\sqrt{n}}\left(1 + \frac{\left(1+\sqrt{\lambda}\right)}{M^{\frac{q}{2(q-1)}}n^{\frac{q-2}{4(q-1)}}} + \frac{\lambda}{M^{\frac{q}{q-1}}n^{\frac{q-2}{2q-2}}}\right). \tag{4.7.6}$$

Second part: if $\frac{M}{\sqrt{n}} \lesssim \varepsilon^{1-1/q}$.

$$\left\|\mathbb{E}[X] - T_H(\widehat{P}_n)\right\| \lesssim R_\lambda\left(\frac{\varepsilon^{1-1/q}}{M} + \frac{\varepsilon^{\frac{1}{2}+\frac{q-1}{q}-\frac{1}{q}}\left(1+\sqrt{\lambda}\right)}{M^2} + \frac{\lambda\varepsilon^{2\frac{q-1}{q}}}{\varepsilon^{1/q}M^3}\right) \tag{4.7.7}$$

$$= \frac{R_\lambda}{M}\varepsilon^{1-1/q}\left(1 + \frac{\varepsilon^{\frac{1}{2}-\frac{1}{q}}\left(1+\sqrt{\lambda}\right)}{M} + \frac{\lambda\varepsilon^{1-\frac{2}{q}}}{M^2}\right) \tag{4.7.8}$$

$$\tag{4.7.9}$$

Now, we simplify the conditions. The first condition is $8(V_H + \varepsilon\beta^2) \leq \beta^2$, which is implied, with our choice of β and using Lemma 15, to

$$8Tr(\Sigma) \leq (1 - 8\varepsilon)\left(\frac{\mathbb{E}[\|X - \mathbb{E}[X]\|^q]\sqrt{n}}{\sqrt{Tr(\Sigma)}}\right)^{2/(q-1)}.$$

Then, use that $\varepsilon \leq 1/16$ and isolate n to get the form used in the lemma. The condition on λ_{max} is under the form

$$\sqrt{\frac{V_H}{n}} + \sqrt{\frac{v_H\lambda_{max}}{n}} + \frac{\sqrt{\varepsilon}}{\sqrt{n}}\Big(1 + \sqrt{\lambda_{max}}\Big)\beta + \frac{C}{n}\lambda_{max}\beta \lesssim \frac{\beta}{2}.$$

then, because $\beta \geq \mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q} n^{1/(2(q-1))}$, this condition can be simplified

$$\sqrt{\frac{V_H}{n}} + \sqrt{\frac{v_H\lambda_{max}}{n}} + \frac{\sqrt{\varepsilon}}{\sqrt{n^{\frac{q-2}{q-1}}}}\Big(1 + \sqrt{\lambda_{max}}\Big)\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q} + \frac{C}{n^{\frac{2q-3}{2q-2}}}\lambda_{max}\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q} \lesssim \frac{\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q} n^{1/(2(q-1))}}{2}. \quad (4.7.10)$$

which is verified as long as $\lambda_{max} \lesssim n$.

4.8 Proofs of the lemmas

4.8.1 Proof of Lemma 12

We compute Z_β. For all $\theta \in \mathbb{R}$,

$$\begin{aligned}
Z_\beta(\theta) &= \mathbb{E}[(X - \theta)\mathbb{1}\{|X - \theta| \leq \beta\}] + \beta\mathbb{P}(X - \theta > \beta) \\
&= \alpha\int_1^{\theta+\beta} \frac{x - \theta}{x^{\alpha+1}}dx + \frac{\beta}{(\theta + \beta)^\alpha} \\
&= \alpha\left(\left[\frac{1}{(\alpha - 1)x^{\alpha-1}} - \frac{\theta}{\alpha x^\alpha}\right]_1^{\theta+\beta}\right) + \frac{\beta}{(\theta + \beta)^\alpha} \\
&= \alpha\left(\frac{1}{(\alpha - 1)(\theta + \beta)^{\alpha-1}} - \frac{\theta}{\alpha(\theta + \beta)^\alpha} - \frac{1}{\alpha - 1} + \frac{\theta}{\alpha}\right) + \frac{\beta}{(\theta + \beta)^\alpha}
\end{aligned}$$

Then, $Z_\beta(\mathbb{E}[X]) = Z_\beta(\alpha/(\alpha - 1))$

$$\begin{aligned}
Z_\beta(\mathbb{E}[X]) &= \alpha\left(\frac{1}{(\alpha - 1)(\alpha/(\alpha - 1) + \beta)^{\alpha-1}} - \frac{1}{(\alpha - 1)(\alpha/(\alpha - 1) + \beta)^\alpha}\right) + \frac{\beta}{(\alpha/(\alpha - 1) + \beta)^\alpha} \\
&= \frac{\alpha}{(\alpha - 1)(\alpha/(\alpha - 1) + \beta)^{\alpha-1}}\left(1 - \frac{1}{(\alpha/(\alpha - 1) + \beta)}\right) + \frac{\beta}{(\alpha/(\alpha - 1) + \beta)^\alpha}
\end{aligned}$$

When β gets large, we have $Z_\beta(\mathbb{E}[X]) = O(1/\beta^{\alpha-1})$. Hence by Theorem 26 (because ψ_H satisfies Assumptions 3), the bias is $|\mathbb{E}[X] - T(P)| = O(1/\beta^{\alpha-1})$.

4.8.2 Proof of Lemma 13

From Theorem 26 we only need to control $Z_\beta(\mathbb{E}[X])$. We have,

$$\begin{aligned} Z_\beta(\mathbb{E}[X]) &= \mathbb{E}\left[\beta\frac{(X-\mathbb{E}[X])}{\|X-\mathbb{E}[X]\|}\psi\left(\left\|\frac{X-\mathbb{E}[X]}{\beta}\right\|\right)\right] \\ &= \mathbb{E}\left[\beta\frac{(X-\mathbb{E}[X])}{\|X-\mathbb{E}[X]\|}\psi\left(\left\|\frac{X-\mathbb{E}[X]}{\beta}\right\|\right)\right] - \mathbb{E}\left[\beta\frac{(X-\mathbb{E}[X])}{\|X-\mathbb{E}[X]\|}\left\|\frac{X-\mathbb{E}[X]}{\beta}\right\|\right] \end{aligned}$$

Hence, by triangular inequality,

$$\|Z_\beta(\mathbb{E}[X])\| \leq \beta\mathbb{E}\left[\left|\psi\left(\left\|\frac{X-\mathbb{E}[X]}{\beta}\right\|\right) - \left\|\frac{X-\mathbb{E}[X]}{\beta}\right\|\right|\right]$$

We denote $Y = \|X-\mathbb{E}[X]\|/\beta$, we have

$$\begin{aligned} \|Z_\beta(\mathbb{E}[X])\| \leq \beta\mathbb{E}[|\psi(Y)-Y|] &= \beta\int|\psi(y)-y|dF_Y(y) \\ &= \beta\int_0^\infty(y-\psi(y))dF_Y(y) \end{aligned}$$

because ψ is 1-Lipshitz and $\psi(0)=0$. Then, by integration by part,

$$\|Z_\beta(\mathbb{E}[X])\| \leq \beta\int_0^\infty(1-\psi'(y))(1-F_Y(y))dy$$

Until now, the proof was valid for any ψ, for the specific case of Huber score function (see (4.2.2)), we get that

$$\|\mathbb{E}[X]-T_H(P)\| \leq 2\beta\int_1^\infty\mathbb{P}(\|X-\mathbb{E}[X]\|\geq\beta y)dy$$

Then, use Markov's inequality,

$$\|\mathbb{E}[X]-T_H(P)\| \leq 2\beta\int_1^\infty\frac{\|X-\mathbb{E}[X]\|^q}{\beta^q y^q}dy = \frac{2\|X-\mathbb{E}[X]\|^q}{(q-1)\beta^{q-1}}.$$

4.8.3 Proof of Lemma 14

We have,

$$Z_\beta(\mathbb{E}[X]) = \mathbb{E}\left[\frac{X-\mathbb{E}[X]}{\|X-\mathbb{E}[X]\|}\beta\psi\left(\left\|\frac{X-\mathbb{E}[X]}{\beta}\right\|\right)\right]$$

Then, by Taylor expansion

$$\begin{aligned} \|Z_\beta(\mathbb{E}[X])\| \leq& \left\|\mathbb{E}\left[\frac{X-\mathbb{E}[X]}{\|X-\mathbb{E}[X]\|}\beta\left\|\frac{X-\mathbb{E}[X]}{\beta}\right\|\right]\right\| + \beta\mathbb{E}\left[\frac{\|\psi^{(k)}\|_\infty}{k!}\left\|\frac{X-\mathbb{E}[X]}{\beta}\right\|^k\right] \\ =& \mathbb{E}\left[\frac{\|\psi^{(k)}\|_\infty\mathbb{E}[\|X-\mathbb{E}[X]\|^k]}{k!\beta^{k-1}}\right] \end{aligned}$$

which proves the first part of the lemma. In the case of Bernoulli distribution, the result follows from a Taylor expansion:

$$\begin{aligned}
Z_\beta(\mathbb{E}[X]) &= \mathbb{E}\left[\operatorname{sign}(X-\mathbb{E}[X])\beta\psi\left(\left|\frac{X-\mathbb{E}[X]}{\beta}\right|\right)\right] \\
&= p\left(\beta\psi\left(\frac{1-p}{\beta}\right)\right) + (1-p)\left(\beta\psi\left(\frac{-p}{\beta}\right)\right) \\
&= p\beta\left(\frac{1-p}{\beta} + \frac{1}{k!}\psi^{(k)}(0)\frac{(1-p)^k}{\beta^k} + o\left(\frac{1}{\beta^k}\right)\right) - (1-p)\beta\left(\frac{p}{\beta} + \frac{1}{k!}\psi^{(k)}(0)\frac{p^k}{\beta^k} + o\left(\frac{1}{\beta^k}\right)\right)
\end{aligned}$$

4.8.4 Proof of Lemma 15

First, remark that we have for all $x \in \mathbb{R}_+$, $\psi^2(x) \leq 2\rho(x)$. Indeed, let $h(x) = \psi^2(x) - 2\rho(x)$, its derivative is $h'(x) = 2\psi(x)(\psi'(x) - 1)$ and because $\psi' \leq 1$ and $\psi(0) = 0$, we get that h is decreasing, the fact that $h(0) = 0$ implies that for all $x \in \mathbb{R}_+$, $\psi^2(x) \leq 2\rho(x)$. Then,

$$V \leq 2\beta^2\mathbb{E}\left[\rho\left(\frac{\|X - T(P)\|}{\beta}\right)\right]. \tag{4.8.1}$$

Define $J(\theta) = \mathbb{E}\left[\rho\left(\frac{\|X-\theta\|}{\beta}\right)\right]$ by definition 4.1.1, $T(P)$ is the minimum of J and by equation (4.8.1),

$$V \leq 2\beta^2\mathbb{E}\left[\rho\left(\frac{\|X - \mathbb{E}[X]\|}{\beta}\right)\right]$$

Then finally, using that by integration of $\psi' \leq 1$ we have $\rho(x) \leq x^2/2$, hence the result. Similarly, note that

$$\begin{aligned}
v^2 &= \beta^2 \sup_{u\in S^{d-1}} \mathbb{E}\left[\frac{u^T(X-T(P))(X-T(P))^T u}{\|X-T(P)\|^2}\psi\left(\frac{\|X-T(P)\|}{\beta}\right)^2\right] \\
&= \beta^2 \sup_{u\in S^{d-1}} \mathbb{E}\left[\frac{\langle u, X-T(P)\rangle^2}{\|X-T(P)\|^2}\psi\left(\frac{\|X-T(P)\|}{\beta}\right)^2\right] \\
&\leq \beta^2 \sup_{u\in S^{d-1}} \mathbb{E}\left[\frac{\langle u, X-T(P)\rangle^2}{\|X-T(P)\|^2}2\rho\left(\frac{\|X-T(P)\|}{\beta}\right)\right] \\
&\leq \beta^2 \sup_{u\in S^{d-1}} \mathbb{E}\left[\frac{\langle u, X-T(P)\rangle^2}{\|X-T(P)\|^2}\left(\frac{\|X-T(P)\|}{\beta}\right)^2\right] \\
&= \sup_{u\in S^{d-1}} \mathbb{E}\left[\langle u, X-T(P)\rangle^2\right] \\
&= \sup_{u\in S^{d-1}} \mathbb{E}\left[\left(\langle u, X-\mathbb{E}[X]\rangle + \langle u, \mathbb{E}[X]-T(P)\rangle\right)^2\right] \\
&= \sup_{u\in S^{d-1}} \mathbb{E}\left[\langle u, X-\mathbb{E}[X]\rangle^2 + \langle u, \mathbb{E}[X]-T(P)\rangle^2\right] = \|\Sigma\|_{op} + \|\mathbb{E}[X]-T(P)\|^2
\end{aligned}$$

4.8.5 Proof of Lemma 16

We have

$$V_H = \mathbb{E}\left[\beta^2 \psi_H\left(\frac{\|X - T_H(P)\|}{\beta}\right)^2\right] = \mathbb{E}[\beta^2 \wedge (\|X - T_H(P)\|)^2]$$
$$= \mathbb{E}[\|X - T_H(P)\|^2] - \mathbb{E}[\left(\|X - T_H(P)\|^2 + \beta^2\right)\mathbb{1}\{\|X - T_H(P)\| > \beta\}]$$

Then, by Hölder inequality,

$$V_H \geq \mathbb{E}[\|X - T_H(P)\|^2] - \mathbb{E}\Big[\left(\|X - T_H(P)\|^2 + \beta^2\right)^q\Big]^{1/q} \mathbb{P}(\|X - T_H(P)\| > \beta)^{1-1/q} \quad (4.8.2)$$

Then, use the following lemma

Lemma 30. *Let Y be a positive real random variable, $\mathbb{E}[Y^q] < \infty$. We have for all $\lambda > 0$,*

$$\mathbb{P}(Y \geq \lambda) \leq 2^{q-1} \frac{\mathbb{E}[Y^q]}{\lambda^q + \mathbb{E}[Y^q]}$$

See Section 4.8.12 for the proof. Then, for $Y = \|X - T_H(P)\|$, we get,

$$V_H \geq \mathbb{E}[\|X - T_H(P)\|^2] - \mathbb{E}\Big[\left(\|X - T_H(P)\|^2 + \beta^2\right)^q\Big]^{1/q} \left(2^{2q-1} \frac{\mathbb{E}\big[\|X - T_H(P)\|^{2q}\big]}{\mathbb{E}[\|X - T_H(P)\|^{2q}] + \beta^{2q}}\right)^{1-1/q}.$$

Use the fact that $(a+b)^q \leq 2^{q-1}(a^q + b^q)$,

$$V_H \geq \mathbb{E}[\|X - T_H(P)\|^2] - 2^{(q-1)/q + (2q-1)(1-1/q)} \frac{\mathbb{E}\big[\|X - T_H(P)\|^{2q}\big]^{1-1/q}}{\left(\mathbb{E}[\|X - T_H(P)\|^{2q}] + \beta^{2q}\right)^{1-2/q}}$$
$$\geq \mathbb{E}[\|X - T_H(P)\|^2] - 2^{2q} \frac{\mathbb{E}\big[\|X - T_H(P)\|^{2q}\big]^{1-1/q}}{\left(\mathbb{E}[\|X - T_H(P)\|^{2q}] + \beta^{2q}\right)^{1-2/q}}$$

And finally, because $\mathbb{E}[X]$ is the minimizer of the quadratic loss,

$$V_H \geq \mathbb{E}[\|X - \mathbb{E}[X]\|^2] - 4^q \frac{\mathbb{E}\big[\|X - T_H(P)\|^{2q}\big]^{1-1/q}}{\left(\mathbb{E}[\|X - T_H(P)\|^{2q}] + \beta^{2q}\right)^{1-2/q}}.$$

Then, we operate the same manner for the bound on v_H. We have,

$$v_H = \sup_{u \in S^{d-1}} \mathbb{E}\left[\frac{\langle u, X - T_H(P)\rangle^2}{\|X - T_H(P)\|^2}\left(\beta^2 \wedge (\|X - T_H(P)\|)^2\right)\right]$$
$$= \sup_{u \in S^{d-1}} \mathbb{E}\big[\langle u, X - T_H(P)\rangle^2\big] - \mathbb{E}\left[\frac{\langle u, X - T_H(P)\rangle^2}{\|X - T_H(P)\|^2}\left(\beta^2 - (\|X - T_H(P)\|)^2\right)\mathbb{1}\{\|X - T_H(P)\| \geq \beta\}\right]$$

Then, use Cauchy-Schwarz inequality,

$$v_H \geq \sup_{u \in S^{d-1}} \mathbb{E}\big[\langle u, X - T_H(P)\rangle^2\big] - \mathbb{E}\big[\left(\beta^2 - (\|X - T_H(P)\|)^2\right)\mathbb{1}\{\|X - T_H(P)\| \geq \beta\}\big]$$
$$\geq \sup_{u \in S^{d-1}} \mathbb{E}\big[\langle u, X - \mathbb{E}[X]\rangle^2\big] + \langle u, \mathbb{E}[X] - T_H(P)\rangle^2 - \mathbb{E}\big[\left(\beta^2 - (\|X - T_H(P)\|)^2\right)\mathbb{1}\{\|X - T_H(P)\| \geq \beta\}\big]$$
$$\geq \sup_{u \in S^{d-1}} \mathbb{E}\big[\langle u, X - \mathbb{E}[X]\rangle^2\big] - \mathbb{E}\big[\left(\beta^2 - (\|X - T_H(P)\|)^2\right)\mathbb{1}\{\|X - T_H(P)\| \geq \beta\}\big]$$

Then, use the same reasoning as for the bound on V_H to conclude that

$$v_H \geq \|\Sigma\|_{op} - 4^q \frac{\mathbb{E}\big[\|X - T_H(P)\|^{2q}\big]^{1-1/q}}{\left(\mathbb{E}[\|X - T_H(P)\|^{2q}] + \beta^{2q}\right)^{1-2/q}}$$

4.8.6 Proof of Lemma 22

Let

$$Z_P(\theta) = \mathbb{E}_P\left[\frac{X-\theta}{\|X-\theta\|}\psi_H\left(\frac{\|X-\theta\|}{\beta}\right)\right].$$

We have that, for all $u \in S^{d-1}$ and $C > 0$,

$$\langle Z_{P_\varepsilon}(T_H(P) + Cu), u\rangle = (1-\varepsilon)\langle Z_P(T_H(P) + Cu), u\rangle + \varepsilon\langle Z_Q(T_H(P) + Cu), u\rangle$$

then, by Cauchy-Schartz inequality,

$$\langle Z_{P_\varepsilon}(T_H(P) + Cu), u\rangle \leq (1-\varepsilon)\langle Z_P(T_H(P) + Cu), u\rangle + \varepsilon\beta. \tag{4.8.3}$$

Use Taylor inequality on the function $f : t \mapsto \langle Z_P(T_H(P) + tCu), u\rangle$, we have

$$\begin{aligned}\langle Z_P(T_H(P) + Cu), u\rangle &\leq C \sup_{t\in[0,1]} u^T Jac(Z_P)(tT_H(P) + (1-t)(T_H(P) + Cu))u \\ &= C \sup_{t\in[0,1]} u^T Jac(Z_P)(T_H(P) + (1-t)Cu)u,\end{aligned}$$

where we used that $Z_P(T_H(P)) = 0$. But, from the lemma 28,

$$u^T Jac(Z_P)(\theta)u \leq -\mathbb{E}\left[\psi_H'\left(\left\|\frac{X-\theta}{\beta}\right\|\right)\right] = -\mathbb{P}(\|X-\theta\| \leq \beta)$$

Hence,

$$\begin{aligned}\langle Z_P(T_H(P) + Cu), u\rangle &\leq -C \inf_{t\in[0,1]} \mathbb{P}(\|X - T_H(P) - (1-t)Cu\| \leq \beta) \\ &\leq -C\mathbb{P}(\|X - T_H(P)\| \leq \beta - C) \\ &\leq -C\mathbb{P}(\psi_H(\|X - T_H(P)\|) \leq \psi_H(\beta - C)) \\ &\leq -C\left(1 - \frac{V_H}{(\beta - C)^2}\right)\end{aligned}$$

Take $C = 4\varepsilon\beta$,

$$\langle Z_P(T_H(P) + Cu), u\rangle \leq -4\varepsilon\beta\left(1 - \frac{V_H}{\beta^2(1-4\varepsilon)^2}\right)$$

and given the hypothesis that $\varepsilon \leq \varepsilon_{max} \leq 1/8$, we get

$$\langle Z_P(T_H(P) + Cu), u\rangle \leq -4\varepsilon\beta\left(1 - \frac{4V_H}{\beta^2}\right)$$

then, because $4V_H \leq \beta^2/2$, we get that $\langle Z_{P_\varepsilon}(T_H(P) + Cu), u\rangle \leq -2\varepsilon\beta$. Then, from equation (4.8.3),

$$\langle Z_{P_\varepsilon}(T_H(P) + Cu), u\rangle \leq -(1-\varepsilon)2\varepsilon\beta + \varepsilon\beta \leq -\left(1 - \frac{1}{8}\right)2\varepsilon\beta + \varepsilon\beta < 0. \tag{4.8.4}$$

Now, let $f_u : \lambda \mapsto \langle Z_{P_\varepsilon}(T_H(P) + \lambda u), u\rangle$. For all $\lambda \in \mathbb{R}$, we have

$$f_u'(\lambda) = (1-\varepsilon)u^T Jac(Z_P)(T_H(P) + \lambda u)u + \varepsilon u^T Jac(Z_H)(T_H(P) + \lambda u)u.$$

Then, using Lemma 28, we have that both jacobian matrices are non-positive and hence f_u' is non-positive. which proves that f_u is non-increasing. Then, as $f_u(\langle T_H(P) - T_H(P_\varepsilon), u\rangle) = 0$, equation (4.8.4) translates in $f_u(C) \leq f_u(\langle T_H(P) - T_H(P_\varepsilon), u\rangle)$, which implies because f_u is non-increasing that

$$C = 4\varepsilon\beta \geq \langle T_H(P) - T_H(P_\varepsilon), u\rangle.$$

This is valid for any $u \in S^{d-1}$, hence

$$\|T_H(P) - T_H(P_\varepsilon)\| \leq 4\varepsilon\beta$$

4.8.7 Proof of Lemma 24

Using Lemma 13, we have

$$\|\mathbb{E}[X] - T_H(P_B)\| \leq \frac{2\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^b X_i - \mathbb{E}[X]\right\|^q\right]}{(q-1)\beta^{q-1}}.$$

We only have to control the numerator. Then, use a bound on the moments of a sum of i.i.d random variables found [DG12, Theorem 1.2.5], which says that there exists an absolute constant $K > 0$ such that

$$\mathbb{E}\left[\left\|\sum_{i=1}^b X_i - \mathbb{E}[X]\right\|^q\right]^{1/q} \leq Kq\left(\mathbb{E}\left[\left\|\sum_{i=1}^b X_i - \mathbb{E}[X]\right\|^2\right]^{1/2} + \mathbb{E}\left[\max_{1\leq i\leq b}\|X_i - \mathbb{E}[X]\|^q\right]^{1/q}\right) \tag{4.8.5}$$

Let $\varepsilon_1, \ldots, \varepsilon_n$ denote i.i.d Rademacher random variable independents from $Y_1, \ldots, Y_n$. By the symmetrization lemma (see [DG12, Lemma 1.2.6]),

$$\mathbb{E}\left[\left\|\sum_{i=1}^b X_i - \mathbb{E}[X]\right\|^2\right] \leq 4\mathbb{E}\left[\left\|\sum_{i=1}^b \varepsilon_i X_i\right\|^2\right] = 4\mathbb{E}\left[\sum_{i=1}^b \|X_i\|^2\right] = 4b\mathbb{E}[\|X\|^2] = 4bTr(\Sigma).$$

Then, inject this equation in equation (4.8.6)

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{i=1}^b X_i - \mathbb{E}[X]\right\|^q\right]^{1/q} &\leq Kq\left(2\sqrt{bTr(\Sigma)} + \mathbb{E}\left[\max_{1\leq i\leq b}\|X_i - \mathbb{E}[X]\|^q\right]^{1/q}\right) && (4.8.6)\\
&\leq Kq\left(2\sqrt{bTr(\Sigma)} + \mathbb{E}\left[\sum_{i=1}^b \|X_i - \mathbb{E}[X]\|^q\right]^{1/q}\right) && (4.8.7)\\
&= Kq\left(2\sqrt{bTr(\Sigma)} + b^{1/q}\mathbb{E}[\|X - \mathbb{E}[X]\|^q]^{1/q}\right). && (4.8.8)
\end{aligned}$$

Hence the result.

4.8.8 Proof of Lemma 25

from Lemma 20, Lemma 15 and Lemma 24, if $\beta^2 \geq 8Tr(\Sigma)/b$, then there exist absolute constants $C, C' > 0$ such that, for all $\lambda \in (0, \lambda_{max})$, with probability larger than $1 - 4\exp(-\lambda) - \exp(-K/8)$,

$$\|\text{HOME}_K(X_1^n) - \mathbb{E}[X]\| \leq 6\frac{\sqrt{Tr(\Sigma)}}{\sqrt{n}} + 8\sqrt{\|\Sigma\|_{op}\frac{\lambda}{n}} + \frac{C}{K}\lambda\beta + C'^q q^q \frac{\left(\sqrt{bTr(\Sigma)} + b^{1/q}\mathbb{E}[\|X_i - \mathbb{E}[X]\|^q]^{1/q}\right)^q}{(q-1)b^q\beta^{q-1}}. \tag{4.8.9}$$

where λ_{max} is such that

$$\frac{3\sqrt{Tr(\Sigma)}}{2\sqrt{n}} + 2\sqrt{\|\Sigma\|_{op}\frac{\lambda_{\max}}{n}} + \frac{C}{K}\lambda_{\max}\beta \leq \frac{\beta}{2}.$$

Using the same reasoning used to simplify this condition under Lemma 21, the condition on λ_{max} is implied by

$$\frac{3}{2\sqrt{n}} + 2\sqrt{\frac{\lambda_{\max}}{n}} + \frac{C}{K}\lambda_{\max}\frac{\beta}{\sqrt{Tr(\Sigma)}} \leq \frac{\beta}{2\sqrt{Tr(\Sigma)}}.$$

Choose $\beta = C'^{\frac{q}{q-1}} q^{\frac{q}{q-1}} b^{-\frac{q}{2(q-1)}} \left(\sqrt{Tr(\Sigma)} + b^{\frac{1}{q}-\frac{1}{2}}\mathbb{E}[\|X_i - \mathbb{E}[X]\|^q]^{1/q}\right)^{\frac{q}{q-1}} \left(\frac{n}{Tr(\Sigma)}\right)^{\frac{1}{2(q-1)}}$. Then, inject this in equation (4.8.9), there exists a constant $C > 0$ such that

$$\|\text{HOME}_K(X_1^n) - \mathbb{E}[X]\| \leq 12\frac{\sqrt{Tr(\Sigma)}}{\sqrt{n}} + 16\sqrt{\|\Sigma\|_{op}\frac{\lambda}{n}} + \frac{C\lambda q^{\frac{q}{q-1}}}{K^{1-\frac{q}{2q-2}}\sqrt{n}}\left(\sqrt{Tr(\Sigma)} + K^{\frac{1}{2}-\frac{1}{q}}\frac{\mathbb{E}[\|X_i - \mathbb{E}[X]\|^q]^{1/q}}{n^{\frac{1}{2}-\frac{1}{q}}}\right)^{\frac{q}{q-1}}\left(\frac{1}{Tr(\Sigma)}\right)^{\frac{1}{2(q-1)}}. \tag{4.8.10}$$

which implies

$$\|\text{HOME}_K(X_1^n) - \mathbb{E}[X]\| \leq 12\frac{\sqrt{Tr(\Sigma)}}{\sqrt{n}} + 16\sqrt{\|\Sigma\|_{op}\frac{\lambda}{n}} + \frac{C\lambda q^{\frac{q}{q-1}}}{\sqrt{n}}\left(\frac{\sqrt{Tr(\Sigma)}}{K^{\frac{q-2}{2q}}} + K^{\frac{1}{2q}}\frac{\mathbb{E}[\|X_i - \mathbb{E}[X]\|^q]^{1/q}}{n^{\frac{1}{2}-\frac{1}{q}}}\right)^{\frac{q}{q-1}}\left(\frac{1}{Tr(\Sigma)}\right)^{\frac{1}{2(q-1)}}. \tag{4.8.11}$$

Finally let us simplify the condition on λ_{max}. We have

$$\frac{\beta}{\sqrt{Tr(\Sigma)}} = C'^{\frac{q}{q-1}} q^{\frac{q}{q-1}} \frac{\sqrt{n}}{K^{1-\frac{1}{2(q-1)}}}\left(1 + \frac{\mathbb{E}[\|X_i - \mathbb{E}[X]\|^q]^{1/q}}{b^{\frac{1}{2}-\frac{1}{q}}\sqrt{Tr(\Sigma)}}\right)^{\frac{q}{q-1}} \geq C' q \frac{\sqrt{n}}{K^{1-\frac{1}{2(q-1)}}}.$$

Then, the condition

$$\frac{3}{2\sqrt{n}} + 2\sqrt{\frac{\lambda_{\max}}{n}} + \frac{C}{K}\lambda_{\max}\frac{\beta}{\sqrt{Tr(\Sigma)}} \leq \frac{\beta}{2\sqrt{Tr(\Sigma)}},$$

is implied by

$$\frac{3}{2\sqrt{n}} + 2\sqrt{\frac{\lambda_{\max}}{n}} + \frac{\lambda_{max}}{K}\frac{C''\sqrt{n}}{K^{1-\frac{1}{2(q-1)}}} \leq C' q \frac{\sqrt{n}}{K^{1-\frac{1}{2(q-1)}}},$$

for some absolute constant $C'' > 0$.

4.8.9 Proof of Lemma 26

The proof is derived from the proof of iterative reweighting algorithm for regression found in [HR09, Section 7.8] and surprisingly we don't need to change anything in the proof. We only need to check that it works also for our setting. We remind here the proof for completeness purpose.

Define

$$\mathbf{U}(\theta) = \frac{1}{n}\sum_{i=1}^{n} U_i\left(\frac{\|X_i - \theta\|}{\beta}\right), \tag{4.8.12}$$

where $U_i(x) = a_i + \frac{1}{2}b_i x^2$ with $a_i, b_i \in \mathbb{R}$ such that, for all $1 \leq i \leq n$,

$$U_i(r_i) \geq \rho(r_i) \quad \text{and} \quad U_i(r_i) = \rho(r_i) \tag{4.8.13}$$

with $r_i = \|X_i - \theta^{(m)}\|/\beta$, see Figure 4.6. Equation (4.8.13) implies that U_i and ρ have the same

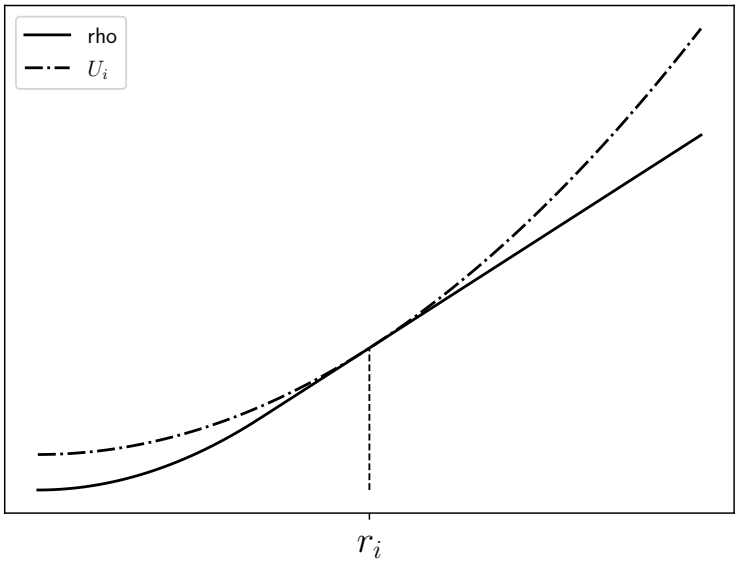


Figure 4.6: Comparison function for proof of convergence

tangent at r_i:

$$U_i'(r_i) = b_i r_i = \psi(r_i). \tag{4.8.14}$$

Hence,

$$w_i = \frac{\psi(r_i)}{r_i} = b_i,$$

and

$$a_i = \rho(r_i) - \frac{1}{2}r_i\psi(r_i).$$

Let us check that equation (4.8.13) holds. Let for $x \geq 0$,

$$z_i(x) = U_i(x) - \rho(x) = \rho(r_i) - \frac{1}{2} r_i \psi(r_i) + \frac{\psi(r_i)}{2r_i} x^2 - \rho(x).$$

We have that z satisfies

$$z_i(r_i) = 0 \quad \text{and} \quad z_i'(r_i) = 0$$

and

$$z_i'(x) = \frac{\psi(r_i)}{r_i} x - \psi(x).$$

Then, since $x \mapsto \psi(x)/x$ is decreasing for $x > 0$, this implies for $0 \leq x \leq r_i$, $z_i'(x) \leq 0$ and for $x \geq r_i$, $z_i'(x) \geq 0$. Hence, $z_i(x) \geq z_i(r_i) = 0$ which proves that $\mathbf{U}$ verifies equations (4.8.13).

Now, rewriting equation (4.8.12) using equation (4.8.14),

$$\mathbf{U}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{w_i}{2} \left(\frac{\|X_i - \theta\|}{\beta} \right)^2 + \rho(r_i) - \frac{1}{2} r_i \psi(r_i) \right).$$

$\mathbf{U}$ is a strictly convex function and its minimum is found in $\theta = \theta^{(m+1)}$. This proves that if $\theta^{(m)}$ does not realize the minimum of $\mathbf{U}$, then

$$\mathbf{U}(\theta^{(m+1)}) < \mathbf{U}(\theta^{(m)}) = \frac{1}{n} \sum_{i=1}^{n} U_i(r_i).$$

Hence, $Z_n(\theta^{(m+1)}) \leq \mathbf{U}(\theta^{(m+1)}) < \sum_{i=1}^{n} \rho(r_i) = Z_n(\theta^{(m)})$ except if $\theta^{(m)}$ realizes the minimum of Z_n in which case $\theta^{(m)} = \theta^{(m+1)}$.

4.8.10 Proof of Lemma 27

To prove the lemma, the goal is to use Theorem 18 in Section 2.5 of [Ler19] that we recall here.

Theorem 28. *Let $\mu \in \mathbb{R}$, assume that, for any β in a finite set $\mathcal{B} \subset \mathbb{R}$, there exists a confidence interval $\widehat{I}_\beta$ such that*

- *for any $\beta, \beta' \in \mathcal{B}$ such that $\beta \leq \beta'$, $|\widehat{I}_\beta| \geq |\widehat{I}_{\beta'}|$,*
- $\mathbb{P}\Big(\mu \in \widehat{I}_\beta\Big) \geq 1 - \alpha_\beta$

Then, if one defines

$$\widehat{\beta} = \min\{\beta \in \mathcal{B} : \cap_{\substack{b \in \mathcal{B} \\ b \leq \beta}} \widehat{I}_b \neq \emptyset\}, \quad \text{and} \quad \widehat{\mu} \in \widehat{I}_{\widehat{\beta}}$$

we have,

$$\forall \beta \in \mathcal{B}, \quad \mathbb{P}\Big(|\widehat{\mu} - \mu| > 2|\widehat{I}_\beta|\Big) \leq \sum_{\substack{b \in \mathcal{B} \\ b \leq \beta}} \alpha_b$$

$\psi_{H,\beta}$ is sub-additive, thus we have,

$$V_{H,\beta} = \mathbb{E}\Big[\big(\psi_{H,\beta}(X - T(P)) - \mathbb{E}[\psi_{H,\beta}(X' - T(P))]\big)^2\Big] \leq \mathbb{E}\Big[\big(\mathbb{E}[\psi_{H,\beta}(X - X')|X]\big)^2\Big],$$

where X and X' are independent and with the same law. Then, by Jensen's inequality $V_{H,\beta} \leq \mathbb{E}[\psi_{H,\beta}(X - X')^2]$ which we estimate using the U-Statistic

$$\widehat{V}_\beta = \frac{1}{n(n-1)} \sum_{i \neq j} \psi_{H,\beta}(X_i - X_j)^2.$$

Then, we verify that for any β, β' in $\mathcal{B}$, $\beta \geq \beta'$, we have that

$$|\tilde{I}_\beta(t/\beta)| \leq |\tilde{I}_{\beta'}(t/\beta')|$$

because $\widehat{V}_\beta/\beta^2 = \frac{1}{n(n-1)} \sum_{i \neq j} 1 \wedge \frac{(X_i - X_j)^2}{\beta^2}$ is a non-decreasing function of β, this is the first hypothesis of Theorem 28 with $\widehat{I}_\beta = \tilde{I}_\beta(t/\beta^2)$. Now we have to bound $\mathbb{P}(\mu \notin \tilde{I}_\beta(t))$. First, by an equivalent of Hoeffding's inequality for U-Statistics (see Theorem 8.1.1 in [KB13]), because ψ_H is bounded by β, we have

$$\mathbb{P}\left(|\widehat{V}_\beta - \mathbb{E}[\psi_H(X - X')^2]| > \frac{\sqrt{t}\beta^2}{\sqrt{2\lfloor n/2 \rfloor}}\right) \leq 2e^{-t}$$

which implies

$$\mathbb{P}\left(|\widehat{V}_\beta - \mathbb{E}[\psi_H(X - X')^2]| > \frac{\sqrt{t}\beta^2}{\sqrt{n/2}}\right) \leq 2e^{-t}$$

Then, with probability greater than $1 - 4e^{-t} - e^{-n/8}$, we have

$$|T_H(\widehat{P}_n) - T_H(P)| \leq 4\sqrt{\frac{\widehat{V}_\beta t}{n} + \frac{\sqrt{2}t^{3/2}\beta^2}{n^{3/2}}} + \frac{4\beta t}{n}$$

and taking the bound on the bias into account, this translates in $\mathbb{P}(\mathbb{E}[X] \in \tilde{I}_\beta(t)) \geq 1 - 4e^{-t} - e^{-n/8}$, hence,

$$\mathbb{P}\Big(\mathbb{E}[X] \in \tilde{I}_\beta\big(t/\beta^2\big)\Big) \geq 1 - 4e^{-t/\beta^2} - e^{-n/8}.$$

Then, by Theorem 28, we have for all $\beta \in \mathcal{B}$,

$$\mathbb{P}\left(|\widehat{T}_{H,\tilde{\beta}}(P) - \mathbb{E}[X]| > 8\sqrt{\frac{\widehat{V}_\beta t}{\beta^2 n} + \frac{\sqrt{2}t^{3/2}}{\beta n^{3/2}}} + \frac{8t}{\beta n} + \frac{2\sigma^2}{\beta}\right) \leq \sum_{\substack{\beta \in \mathcal{B} \\ b \leq \beta}} \Big(4e^{-t/\beta^2} + e^{-n/8}\Big)$$

4.8.11 Proof of Lemma 29

ρ is convex because $\rho'' = \psi' \geq 0$ and it is increasing because $\psi = \rho' \geq 0$ ($\psi(0) = 0$ and ψ increasing). Then, from triangular inequality and Jensen's inequality, we have

$$\rho\left(\frac{\|\mathbb{E}[X] - T(P)\|}{\beta}\right) \leq \rho\left(\frac{\mathbb{E}[\|X - T(P)\|]}{\beta}\right) \leq \mathbb{E}\left[\rho\left(\frac{\|X - T(P)\|}{\beta}\right)\right].$$

By definition of $T(P)$, it is a minimizer of $\theta \mapsto \mathbb{E}\Big[\rho\Big(\frac{\|X-\theta\|}{\beta}\Big)\Big]$, hence,

$$\rho\bigg(\frac{\|\mathbb{E}[X]-T(P)\|}{\beta}\bigg) \leq \mathbb{E}\bigg[\rho\bigg(\frac{\|X-\mathbb{E}[X]\|}{\beta}\bigg)\bigg]$$

then, use the hypothesis to upper bound the right-hand side by $\rho(1/3)$, we get

$$\rho\bigg(\frac{\|\mathbb{E}[X]-T(P)\|}{\beta}\bigg) \leq \rho(1/3).$$

Finally, because ρ is non-decreasing on $\mathbb{R}_+$ (its derivative is non-negative), we get the result.

4.8.12 Proof of Lemma 30

We have for all $u, \lambda > 0$,

$$\begin{aligned}\mathbb{P}(Y \geq \lambda) &= \mathbb{P}((Y+u)^q \geq (\lambda+u)^q) \\ &\leq \frac{\mathbb{E}[(Y+u)^q]}{(\lambda+u)^q} \leq \frac{\mathbb{E}[(Y+u)^q]}{(\lambda^{q/2}+u^{q/2})^2}\end{aligned}$$

Then, use that by convexity of the q^{th}-power function, $(a+b)^q \leq 2^{q-1}(a^q+b^q)$ and also $(a+b)^q \geq a^q + b^q$,

$$\mathbb{P}(Y \geq \lambda) \leq 2^{q-1}\frac{\mathbb{E}[Y^q+u^q]}{(\lambda^{q/2}+u^{q/2})^2}$$

Take $u = \mathbb{E}[Y^q]^{2/q}/\lambda$,

$$\mathbb{P}(Y \geq \lambda) \leq 2^{q-1}\frac{\mathbb{E}[Y^q] + \frac{\mathbb{E}[Y^q]^2}{\lambda^q}}{\lambda^q(1+\frac{\mathbb{E}[Y^q]}{\lambda^q})^2} = 2^{q-1}\frac{\mathbb{E}[Y^q]}{\lambda^q(1+\frac{\mathbb{E}[Y^q]}{\lambda^q})} = 2^{q-1}\frac{\mathbb{E}[Y^q]}{\lambda^q+\mathbb{E}[Y^q]}$$

4.9 Addendum: towards a faster estimator

As said in Section 4.4.3, $T_H(\widehat{P}_n)$ is not minimax because its deviations are of the type

$$\Big\|\mathbb{E}[X]-T_H(\widehat{P}_n)\Big\| \lesssim \frac{\sqrt{Tr(\Sigma)}+\sqrt{\|\Sigma\|_{op}\lambda}}{\sqrt{n}} \bigvee \mathbb{E}[\|X-\mathbb{E}[X]\|^q]^{1/q}\varepsilon^{1-1/q}, \tag{4.9.1}$$

and we know that there are estimators that achieve on a Gaussian dataset

$$\Big\|\mathbb{E}[X]-T(\widehat{P}_n)\Big\| \lesssim \frac{\sqrt{Tr(\Sigma)}+\sqrt{\|\Sigma\|_{op}\lambda}}{\sqrt{n}} \bigvee \varepsilon\sqrt{\|\Sigma\|_{op}}, \tag{4.9.2}$$

which can be a lot tighter (see [CGR+18]). In practice, we can see that indeed, our estimator achieves (4.9.1) and that this bound is tight. In this section, we propose an algorithm and we conjecture that this estimator follows a bound of the type (4.9.2), we only test empirically and we don't give any theoretical proof of this fact.

Estimator	$\varepsilon = 0.1$	$\varepsilon = 0.2$	slope between $\varepsilon = 0.2$ and $\varepsilon = 0.1$
Same β for all coordinates	0.10	0.24	1.4
Different β for each coordinates	0.28	0.56	2.6

Figure 4.7: Results of the experience. Given equations (4.9.2) and (4.9.1), we can expect an error on these numbers of at most $\sqrt{Tr(\Sigma)/n} \simeq 0.02$ (when doing several times the experiment we indeed observe this error) which may change the results a bit but will not change the conclusion.

The estimator: we propose to use the estimator T_n which corresponds to $T_B(X_1^n) = T(X_1B^{-1}, \dots, X_nB^{-1})B$ where B is an invertible parameter matrix of size $d \times d$. For now, we restrict ourselves to diagonal B matrices. This is an extension of what has been done in this article where B was a constant times the identity matrix.

To choose B, we use a gradient descent on a bootstrap estimation of the variance: for M bootstrap samples $(X^*_{i,j})_{1\leq i\leq n, 1\leq j\leq M}$

$$V_{boot} = \frac{1}{M}\sum_{j=1}^{M}\left\| T_B((X^*_{i,j})_{1\leq i\leq n}) - \frac{1}{M}\sum_{j=1}^{M} T_B((X^*_{i,j})_{1\leq i\leq n})\right\|^2.$$

The bootstrap estimation of the variance seems to be a good measure of performance in our case. Intuitively this can be understood that if there are corrupted sample in our dataset, only a portion of the bootstrap samples will be corrupted and V_{boot} will be a sort of distance between the estimation on corrupted and the estimation on non-corrupted.

Due to the bootstrap estimation of the variance at each step, this algorithm is very heavy and could use more work.

The simulated dataset: To assess whether an estimator has a bound of the type (4.9.1) or (4.9.2), we use a dataset in which $X_1, \dots, X_n$ are i.i.d from a corrupted Gaussian $(1-\varepsilon)\mathcal{N}(0, \Sigma) + \varepsilon\mathcal{N}(1000 \cdot \mathbf{1}, I_d)$, with $d = 6$, $\Sigma = \mathrm{diag}(1, 2, 1, 0.1)$, $n = 10000$ and $\varepsilon \in \{0.1, 0.2\}$. The high value of n makes it so that the variance of the estimator is of an order smaller than the bias of the estimator and all the error will come from the bias: the variance is of order $\sqrt{Tr(\Sigma)/n} \simeq 0.020$ and the bias is (hopefully) of order $\varepsilon\sqrt{\|\Sigma\|_{op}} \geq 0.14$. The goal is to see whether the coefficient of proportionality between the the error for a corruption $\varepsilon = 0.2$ and the error for a corruption $\varepsilon = 0.1$ is $0.1 \times \sqrt{\|\Sigma\|_{op}} \simeq 0.14$ or if it is closer to the non-optimal $0.1 \times \sqrt{Tr(\Sigma)} \simeq 0.20$. We realize only one run because the randomness doesn't play a prevalent role in this study as n is very big (and also because the algorithm is rather slow and making several runs would take time).

The results are represented in Figure 4.7 which indicates that indeed our estimator seems to be minimax while the estimator where all the coordinates are treated in the same way is not minimax (the same algorithm is used to optimize the two estimators), this is provided that our algorithm did indeed attain a global minimum (which is in no way certain because of the lack of theoretical result on this). There is still a lot of work to be done in this line of thought. The algorithm is very simplistic and a better optimization and/or objective function would be beneficial. There are for now no theoretical results, but nonetheless this is encouraging as this may be a way to solve the problem of tractable minimax robust mean estimation is high dimension.

Chapter 5

Robust classification via MOM minimization

Abstract

We present an extension of Chervonenkis and Vapnik's classical empirical risk minimization (ERM) where the empirical risk is replaced by a median-of-means (MOM) estimator of the risk. The resulting new estimators are called MOM minimizers. While ERM is sensitive to corruption of the dataset for many classical loss functions used in classification, we show that MOM minimizers behave well in theory, in the sense that it achieves Vapnik's (slow) rates of convergence under weak assumptions: the function in the hypothesis class are only required to have a finite second moment and some outliers may also have corrupted the dataset.

We propose algorithms, inspired by MOM minimizers, which may be interpreted as MOM version of Block Stochastic Gradient Descent (BSGD). The key point of these algorithms is that the block of data onto which a descent step is performed is chosen according to its " centrality" among the other blocks. This choice of " descent block" make these algorithms robust to outliers also this is the only extra step added to classical BSGD algorithms. As a consequence, classical BSGD algorithms can be easily turn into robust MOM versions. Moreover, MOM algorithms perform a smart subsampling which may help to reduce substantially time computations and memory resources when applied to non linear algorithms. These empirical performances are illustrated on both simulated and real datasets.

5.1 Introduction

The article presents a class of robust (to outliers and heavy-tailed data) estimators and algorithms for the classification problem. Consider the classical binary classification problem, let $\mathcal{F}$ denote a class of functions from $\mathcal{X}$ to $\{\pm 1\}$, the empirical risk minimizer (ERM) is defined by

$$\widehat{f}_{\mathrm{ERM}} \in \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} I\{Y_i \neq f(X_i)\} \tag{5.1.1}$$

where $I\{Y_i \neq f(X_i)\} = 1$ if $Y_i \neq f(X_i)$ and 0 otherwise. In this paper, we are interested in the case where the random variables $f(X_i)$ only satisfy a second moment assumption and where the dataset $\{(X_i, Y_i)_{i\in\{1,\dots,N\}}\}$ may contain outliers. The ERM behaves well under these assumptions (see Theorem 29 below). The reason is that the $0-1$ loss $\ell_f^{0-1}(x,y) = I_{\{y\neq f(x)\}}$ is bounded, which grants concentration no matter the distribution of X and a small number of data cannot really impact the empirical mean performance. However, it is well known that ERM is a theoretical estimators that can only be approximated in most situations by efficient algorithms. Indeed, the minimization problem (5.1.1) is NP-hard even for classes $\mathcal{F}$ of half-spaces indicators [GR09, FGRW12]. One of the most classical way to approximate ERM is to choose a convex relaxation of the problem (5.1.1) and design an algorithm solving the associated convex problem. The problem of these approaches in the setting of this paper is that the relaxed criteria are unbounded and therefore way more sensitive to outliers or heavy tailed inputs. This results into poor performance of the algorithms on corrupted and/or heavy-tailed data. Figure 5.1 illustrates this problem on a toy example where most data would be well separated by a linear classifier like Perceptron [Ros58] or logistic classifier, but some anomalies flaw these algorithms.

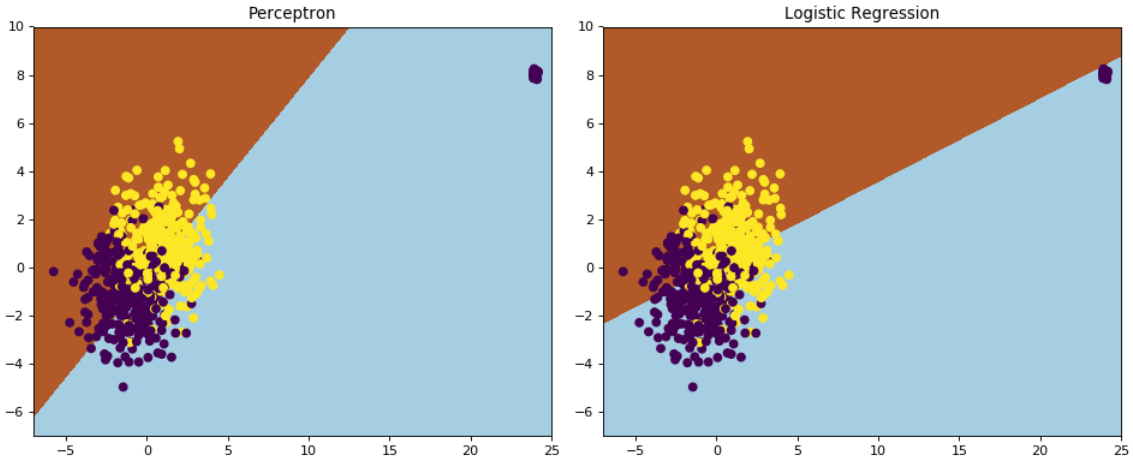


Figure 5.1: Scatter plot of the toy dataset, the color of the points gives their class. The background color gives the linear separation provided by the perceptron (left) and the logistic regression (right) trained on this corrupted dataset.

The example in Figure 5.1 is representative of a general problem that this paper intends to study. Robust learning has received particular attention in recent years by practitioners working on large datasets which are particularly sensitive to data corruption. Challenges recently posted on "kaggle", the most popular data science competition platform, have put forward this topic (see, the 1.5 million dollars problem "Passenger Screening Algorithm Challenge" involves the discovery of terrorist activity from 3D images or the challenge named "NIPS 2017: Defense Against Adversarial Attack" consists in building algorithms robust to adversarial data). Robust algorithms have also been studied theoretically both in statistical and computer science communities. In statistics, robust results usually deal with issues arising when data have heavy-tailed distribution [LM19c, Min15, CGR$^+$18, FK18]. In computer science, most works deal with corrupted datasets, in particular when this corruption arise from adversarial outliers [DKK$^+$19a, CDG19, DKK$^+$17]. Only few papers consider both problems simultaneously [LL18, LL20].

In learning theory, most alternatives to ERM manage the problem of outliers and heavy tail distributions for outputs only. These solutions are based on the pioneering work of John Tukey [Tuk60, Tuk62], Peter Huber [Hub64, Hub67] and Frank Hampel [Ham71, Ham74], replacing the square loss by a robust alternative like Huber loss or Tukey's biweight loss. These methods do not allow to treat the case where the inputs are with heavy tails or corrupted, which is a classical problem in robust statistics also known as the "*leverage point problem*", see [HR09].

In this article, we address this question by considering an alternative to M-estimators, called median-of-means (MOM) minimizers. Several estimators based on MOM have recently been proposed in the literature [Min15, LL18, LM19a, LM$^+$19d, LM19c, Men, LL20]. To our knowledge, these articles use the small ball hypothesis [KM15, Men15] to treat problems of least squares regression or Lipschitzian loss regression. This assumption is restrictive in some classic functional regression frameworks [Sau18, HW17] or for problems such as the construction of recommendation system where inputs are sampled in the canonical basis and therefore do not satisfy a small ball condition.

We construct a natural estimator based on the MOM principle, which is called MOM minimizer. This estimator is studied here without the small ball hypothesis. Instead, we assume an a priori bound on the L^2-norm of learning functions. We can identify mainly two streams of hypothesis in Learning theory: 1) boundedness with respect to some norm of the class F of functions and the output Y, the typical example is the boundedness in L^∞ assumption or 2) norm equivalence assumption over the class F (or, more precisely, on the shifted class $F - f^* = \{f - f^* : f \in F\}$ where f^* is the oracle in F, i.e. the minimizer of the theoretical risk among the functions in F) and Y, the typical example being the subgaussian assumption, i.e. $\|f - f^*\|_{\psi_2} \leq L\|f - f^*\|_{L^2}, \forall f \in F$ where for $g \in F$ $\|g\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}[\exp(X^2/t^2)] \leq 2\}$. The small ball assumption is a norm equivalence assumption between the L^1 and L^2 norms and is concerned with the second type of assumptions. Our approach here deals with the first type of assumption. As we only assume boundedness in L^2-norm, this can be seen as a significant relaxation upon the L^∞ boundedness assumption. It turns out that, in this relaxed setting, MOM minimizers achieve minimax rates of convergence [DGL96] in the absence of a margin condition [MT99] even under a L^∞ assumption.

The estimation of the expectation of a univariate variable by median-of-means (MOM) [AMS99, JGV86, NY83] is done as follows: given a partition of the dataset into blocks of the same size, an empirical mean is constructed on each block and the MOM estimator is obtained by taking the median of these empirical means (see Section 5.2.2 for details). These estimators are naturally resistant to the presence of a few outliers in the dataset: if the number of these outliers does not exceed half the number of blocks, more than half of these blocks are made of "clean" data and the median is a reliable estimator.

On the practical side, we introduce algorithms inspired by the MOM minimizers. In these algorithms, the MOM principle is used within algorithms originally intended for the evaluation of ERM estimators associated to convex loss functions. In Section 5.4, we present a " MOM version" of gradient descent algorithms following this approach. The general principle of this iterative algorithm is as follows: at iteration t, a dataset equipartition $B_1, \ldots, B_K$ is selected uniformly at random and the most central block B_{med} is determined according to the following formula

$$\sum_{i \in B_{\text{med}}} \ell_{f_t}(X_i, Y_i) = \text{median}\left(\sum_{i \in B_k} \ell_{f_t}(X_i, Y_i) : \quad k = 1, \ldots, K \right) = \text{MOM}_K\left(\ell_{f_t}\right) \tag{5.1.2}$$

where $\ell_{f_t}(X_i, Y_i) = \ell(f_t(X_i), Y_i)$ is the loss of the prediction $f_t(X_i)$ of the label Y_i. Next iteration

f_{t+1} is then produced by taking from f_t a step down in the direction opposite to the gradient of $f \to \sum_{i \in B_{\text{med}}} \ell_f(X_i, Y_i)$ at f_t, cf. Algorithm 1. The underlying heuristic is that the data in the selected block B_{med} are safe for estimating the risk of f_t, in the sense that empirical risk $|B_{\text{med}}|^{-1} \sum_{i \in B_{\text{med}}} \ell_{f_t}(X_i, Y_i)$ is a subgaussian estimator of $\mathbb{E}\ell_f(X_i, Y_i)$, cf. [DLLO16] and that data indexed by B_{med} should not be outliers. The differentiation properties of $f \to \text{MOM}_K(\ell_f)$ are studied in Section 5.4.2. One additional advantage of our algorithm is that it is based on a simple idea: select a "good" block of data in such as way that it does not contain outliers and it is a subgaussian estimator of the risk. As a result, it requires only little modifications on existing Gradient descent based algorithms to make them robust to outliers and heavy-tailed data. As a proof of concept, in this article, we perform this "MOM modification" to the Logistic Regression, Perceptron and SVM-like algorithm.

In Section 5.5, the practical performances of thess algorithms are illustrated on several simulations, involving in particular different loss functions. These simulations illustrate not surprisingly the gain of robustness that there is to use these algorithms in their MOM version rather than in their traditional version, as can for example be appreciated on the toy-example of Figure 5.1 (see also Figure 5.4 below). MOM estimators are compared to different learning algorithms on real datasets that can be modeled by heavy tailed data, obtaining in each case performances comparable to the best of these benchmarks.

Another advantage of our procedure is that it works on blocks of data. This can improve speed of execution and reduce memory requirements, which can be decisive on massive datasets and/or when one wishes to use non-linear algorithms as in Section 5.4.3. This principle of dividing the dataset to calculate estimators more quickly and then aggregating them is a powerful tool in statistics and machine learning [Jor13]. Among others, one can mention bagging methods [Bre96] or subagging —a variant of bagging where the bootstrap is replaced by subsampling— [BY02]. These methods are considered difficult to study theoretically in general and their analysis is often limited to obtaining asymptotic guarantees. By contrast, the theoretical tools for non-asymptotic risk analysis of MOM minimizers have already essentially been developed. Finally, subsampling by the central block B_{med} ensures robustness properties that cannot be guaranteed by traditional alternatives.

Moreover, the algorithm provides an empirical notion of data depth: data providing good risk estimates of $f \to \mathbb{E}\ell_f(X, Y)$ are likely to be selected many times in the central block B_{med} along the descent, while outliers will be systematically ignored. This notion of depth, based on the risk function, is very natural for prediction problems. It is complemented by an outliers detection procedure: data that are selected a number of times below a predetermined threshold are classified outliers. This procedure is evaluated on our toy example of Figure 5.1 – for this example, data represented by the dots in the top right corner (the outliers) all end with a null score (see Figure 5.8 below). The procedure is then tested on a real dataset on which the conclusions are more interesting. On this experiment, according to the theoretical upper bounds in Theorem 30, MOM minimizer's prediction qualities are deteriorated with large values of K, and this result is verified in some practical cases cf. Figure 5.11. On the other hand, when there are enough data and when the data are not too heavy tailed (finite third moment of the $f(X_i)$), the article [MS17] decouples K and N in the risk bound and find an optimal scaling of $K \asymp \sqrt{N}$, and one might think that this decoupling ought to be possible also in our context. On the other hand, outlier detection is best when the number of blocks is large, cf. Figure 5.9. Outlier resistance and anomaly detection tasks can therefore both be handled using the MOM principle, but the main hyper-parameter K – the number of blocks of data – for setting this method must be chosen

carefully according to the objective. A number of blocks as small as possible (about twice the number of outliers) will give the best predictions, while large values of this number of blocks will accelerate the detection of anomalies. Note that it is essential for outliers detection to use different (for instance, random) partitions at each step of the descent to avoid giving the same score to an outlier and to all the data in the same block containing it.

Detecting outliers is usually performed in machine learning via some unsupervised preprocessing algorithm that detects outliers outside a bulk of data, see for example [HD04, HF00, CD01b, Nec15] or other algorithms like DBSCAN [BK07] or isolation forest [LTZ08]. These algorithms assume elliptical symmetry of the data, a solution for skewed data can also be found in [HV10]. These unsupervised preprocessing removes outliers in advance, i.e. before starting any learning task. As expected, these strategies work well in the toy example from Figure 5.1. There are several cases where it will fail though. First, as explained in [HR09], this strategy classifies data independently of the risk, it is likely to remove from the dataset outlier coming from heavy-tailed distribution, yielding biased estimators. Moreover, a small group of misclassified data inside a bulk won't be detected. Our notion of depth, based on the risk, seems more adapted to the learning task than any preprocessing procedure blind to the risk.

The paper is organized as follows. Section 5.2 presents the classification problem, the ERM and its MOM versions and gathers the assumptions granted for the main results. Section 5.3 presents theoretical risk bounds for the ERM estimator and MOM minimizers on corrupted datasets. Section 5.4 deals with theoretical results on the algorithm computing MOM minimizers. We present the algorithm, study the differentiation property of the objective function $f \to \mathrm{MOM}_K\big(\ell_f\big)$ and provide theoretical bounds on its complexity. Section 5.5 shows empirical performance of our estimators in both simulated and real datasets. Proofs of the main results are postponed to Section 5.6 where we also added heuristics on the practical choice of the hyper-parameters.

5.2 Setting

5.2.1 Empirical risk minimization for binary classification

Consider the supervised binary classification problem, where one observes a sample $(X_1, Y_1), \ldots, (X_N, Y_N)$ taking values in $\mathcal{X} \times \mathcal{Y}$. The set $\mathcal{X}$ is a measurable space and $\mathcal{Y} = \{-1, 1\}$. The goal is to build a classifier —that is, a measurable map $f : \mathcal{X} \to \mathcal{Y}$— such that, for any new observation (X, Y), $f(X)$ is a good prediction for Y. For any classifier f, let

$$\ell_f^{0-1}(x, y) = I\{y \neq f(x)\}, \qquad R^{0-1}(f) = P\ell_f^{0-1} = \mathbb{P}_{(X,Y)\sim P}\big(Y \neq f(X)\big) \ .$$

The $0-1$ risk $R^{0-1}(\cdot)$ is a standard measure of the quality of a classifier. Following Chervonenkis and Vapnik [Vap00], a popular way to build estimators is to replace the unknown measure P in the definition of the risk by the empirical measure P_N defined for any real valued function g by $P_N g = N^{-1}\sum_{i=1}^N g(X_i, Y_i)$ and minimize the empirical risk. The *empirical risk minimizer* for the $0-1$ loss on a class $\mathcal{F}$ of classifiers is $\widehat{f}_{\mathrm{ERM}}^{0-1} \in \mathrm{argmin}_{f\in\mathcal{F}}\{P_N\ell_f^{0-1}\}$.

The main issue with $\widehat{f}_{\mathrm{ERM}}^{0-1}$ is that it cannot be computed efficiently in general. One source of computational complexity is that both $\mathcal{F}$ and the $0-1$ loss function are non-convex. This is why various convex relaxations of the $0-1$ loss have been introduced in statistical learning

theory. These proceed in two steps. First, $\mathcal{F}$ should be replaced by a convex set F of functions taking values in $\mathbb{R}$. Then one builds an alternative loss function ℓ for ℓ^{0-1} defined for all $f \in F$. The new function ℓ should be convex and put less weight on those $f \in F$ such that $f(X_i)Y_i > 0$, these loss functions are commonly called "classification-calibrated losses" in the literature. Classical examples include the *hinge loss* $\ell_f^{\text{hinge}}(x,y) = (1 - yf(x))_+$, or the *logistic loss* $\ell_f^{\text{logistic}}(x,y) = \log(1 + e^{-yf(x)})$. A couple (F,ℓ) such that F is a convex set of real valued functions and ℓ is a convex function (i.e. for all $y \in \{-1,1\}$ and $x \in \mathcal{X}$, $f \in F \to \ell_f(x,y)$ is convex) such that $\ell_f(x,y) < \ell_f(x,-y)$ whenever $yf(x) > 0$ will be called a convex relaxation of $(\mathcal{F}, \ell^{0-1})$. Given a convex relaxation (F,ℓ) of $(\mathcal{F}, \ell^{0-1})$, one can define the associated empirical risk minimizer by

$$\widehat{f}_{\text{ERM}} \in \operatorname*{argmin}_{f \in F} P_N \ell_f \quad . \tag{5.2.1}$$

Note that $\widehat{f}_{\text{ERM}}$ does not build a classifier. To deduce, a classification rule from $\widehat{f}_{\text{ERM}}$ one can simply consider its sign function defined for all $x \in \mathcal{X}$ by $\text{sign}(\widehat{f}_{\text{ERM}}(x)) = 2(I\{\widehat{f}_{\text{ERM}}(x) \geq 0\} - 1/2)$. The procedure $\widehat{f}_{\text{ERM}}$ is solution of a convex optimization problem that can therefore be approximated using a descent algorithm. We refer for example to [Bub15] for a recent overview of this topic and Section 5.4 for more examples.

5.2.2 Corrupted datasets

In this paper, we consider a framework where the dataset may have been corrupted by *outliers* (or anomalies). There are several definitions of outliers in the literature, here, we assume that the dataset is divided into two parts. The first part is the set of inliers, indexed by $\mathcal{I}$, data $(X_i, Y_i)_{i \in \mathcal{I}}$ are hereafter always assumed to be independent and identically distributed (i.i.d.) with common distribution P. The second one is the set of outliers, indexed by $\mathcal{O} \subset [N]$ which has cardinality $|\mathcal{O}|$. Nothing is assumed on these data which may not be independent, have distributions P_i totally different from P, satisfying $P_i|f|^\alpha = \infty$ for any $\alpha > 0$, etc... Doing no hypothesis on the outliers is commonly done in Machine Learning with adversarial examples, see [CGR$^+$18, DKK$^+$17] for examples of such application. In particular, this framework is sufficiently general to cover the case where outliers are i.i.d. with distribution $Q \neq P$ as in the ϵ-contamination model [HR09, CGR$^+$18, Gao17, DM15].

Our first result shows that the rate of convergence of $\widehat{f}^{0-1}_{\text{ERM}}$ is not affected by this corruption as long as $|\mathcal{O}|$ does not exceed $N \times$ (rate of convergence) see Theorem 29 and the remark afterward. However, it is easy to remark that, when the number N of data is finite as it is always the case in practice, even one aggressive outliers may yield disastrous breakdown of the empirical mean's statistical performance. Consequently, even if $\widehat{f}^{0-1}_{\text{ERM}}$ behaves correctly, its proxy $\widehat{f}_{\text{ERM}}$ defined in (5.2.1) for a convex relaxation (F,ℓ) can have disastrous statistical performances, particularly when F and ℓ are unbounded, cf. Figure 5.4 for an illustration.

To bypass this problem, we consider in this paper an alternative to the empirical mean called *median-of-means* [AMS99, JGV86, NY83]. Let $K \leq N$ denote an integer and let $B_1, \dots, B_K$ denote a partition of $\{1, \dots, N\}$ into bins B_k of equal size $|B_k| = N/K$. If K doesn't divide N, one can always drop a few data. For any function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and any non-empty subset $B \subset \{1, \dots, N\}$, define the empirical mean on B by $P_B f = |B|^{-1} \sum_{i \in B} f(X_i, Y_i)$. The median-of-means (MOM) estimator of Pf is defined as the empirical median of the empirical

means on the blocks B_k

$$\mathrm{MOM}_K(f) = \mathrm{median}\{P_{B_k} f : \ k = 1, \dots, K\} \ .$$

As the classical Huber's estimator [Hub64], MOM estimators interpolate between the unbiased but non robust empirical mean (obtained for $K = 1$) and the robust but biased median (obtained for $K = N$). In particular, when applied to loss functions, these new estimators of the risk $P\ell_f, f \in F$ suggest to define the following alternative to Chervonenkis and Vapnik's ERM estimator, called MOM minimizers

$$\widehat{f}_{\mathrm{MOM},K} \in \operatorname*{argmin}_{f \in F} \mathrm{MOM}_K(\ell_f) \ . \tag{5.2.2}$$

From a theoretical point of view, we will prove that, when the number $|\mathcal{O}|$ of outliers is smaller than $N \times$ (rate of convergence), $\widehat{f}_{\mathrm{MOM},K}$ performs well under a second moment assumptions on F and ℓ. To illustrate our main assumptions and theoretical results, we will regularly use the following classical example.

Example 1 (linear classification.)**.** *Let $\mathcal{X} = \mathbb{R}^p$ and let $\|\cdot\|_2$ denote the classical Euclidean norm on $\mathbb{R}^p$. Let F denote a set of linear functions*

$$F = \{f_t : x \mapsto \langle x, t\rangle : \|t\|_2 \leq \Gamma\} \ .$$

Let ℓ denote either the hinge loss or the logistic loss defined respectively for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $f \in F$ by

$$\ell_f^{hinge}(x, y) = (1 - yf(x))_+, \qquad \ell_f^{logistic}(x, y) = \log(1 + e^{-yf(x)}) \ .$$

Remark that the case with an intercept is included in this linear case by adding an artificial $(p+1)^{th}$ dimension: we consider $x' = (x_1, \dots, x_p, 1)$ where $x_1, \dots, x_p$ are the coordinated of x, and then $\langle x', (t_1, \dots, t_p, t_{p+1}\rangle = \langle x, (t_1, \dots, t_p)\rangle + t_{p+1}$. In practice this correspond to adding a column of 1 at the end of the design matrix.

5.2.3 Main assumptions

As already mentioned, data are divided into two groups, a subset $\{(X_i, Y_i) : i \in \mathcal{O}\}$ made of outliers (on which we will make no assumption) and the remaining data $\{(X_i, Y_i) : i \in \mathcal{I}\}$ contains all data that bring information on the target/oracle

$$f^* \in \operatorname*{argmin}_{f \in F} P\ell_f \ .$$

Data indexed by $\mathcal{I}$ are therefore called *inliers* or *informative data*. To keep the presentation as simple as possible, inliers are assumed to be i.i.d. distributed according to P although this assumption could be relaxed as in [LL18, LL20]. Finally, note that the $\mathcal{O} \cup \mathcal{I} = \{1, \dots, N\}$ partition of the dataset is of course unknown from the statistician. Moreover, since no assumption is granted on the set of data indexed by $\mathcal{O}$, this setup covers the framework of adversarial attack where one may imagine that the data indexed by $\mathcal{O}$ have been changed in the worst possible way by some malicious adverser.

Let us now turn to the set of assumptions we will use to study MOM minimizers procedures. For any measure Q and any function f for which it makes sense, denote by $Qf = \int f dQ$. Denote

also, for all $q \geq 1$, by L^q the set of real valued functions f such that $\int |f|^q dP < \infty$ and, for any $f \in L^q$, by

$$\|f\|_{L^q} = \left(\int |f|^q dP \right)^{1/q} .$$

Our first assumption is an L^2-assumption on the functions in F.

Assumption 1. *For all $f \in F$, we have $\|f\|_{L^2} \leq \theta_2$.*

Of course, Assumption 1 is granted if F is a set of classifiers. It also holds for the linear class of functions from Example 1 as long as $P\|X\|_2^2 < \infty$ with $\theta_2 = \Gamma(P\|X\|_2^2)^{1/2}$. As announced in the introduction, it is a boundedness assumption (w.r.t. the L_2-norm) and not a norm equivalence assumption. For instance, it covers cases that cannot be handled via norm equivalence. A typical example is for matrix completion problems where X is uniformly distributed over the canonical basis $(E_{pq} : p \in [m], q \in [T])$ of the linear space $\mathbb{R}^{m \times T}$ of $m \times T$ matrices. One has for $f(\cdot) = \langle \cdot, E_{11} \rangle$ and any $r \geq 1$, $\|f\|_{L_r} = (\mathbb{E}|f(X)|^r)^{1/r} = (1/(mT))^{1/r}$. Hence, any norm equivalence assumption on the class $F = \{f_A = \langle \cdot, A \rangle : \|f_A\|_{L^2} \leq \theta_2\}$ will depend on the dimension mT of the problem resulting either in wrong rates of convergence or in assumption on the number of data. Our approach does not use any norm equivalence assumption so that our rates of convergence do not depend on dimension dependent ratio. Rates depend only on the L_2 radius θ_2 of F from Assumption 1.

The second assumption deals with the complexity of the class F. This complexity appears in the upper bound of the risk. It is defined using only informative data. Let

$$\mathcal{K} = \{k \in \{1, \dots, K\} : B_k \cap \mathcal{O} = \emptyset\} \quad \text{and} \quad \mathcal{J} = \cup_{k \in \mathcal{K}} B_k .$$

Definition 3. *Let $\mathcal{G}$ denote a class of functions $f : \mathcal{X} \to \mathbb{R}$ and let $(\epsilon_i)_{i \in \mathcal{I}}$ denote i.i.d. Rademacher random variables independent from $(X_i, Y_i)_{i \in \mathcal{I}}$. The* Rademacher complexity *of $\mathcal{G}$ is defined by*

$$\mathcal{R}(\mathcal{G}) = \max_{A \in \{\mathcal{I}, \mathcal{J}\}} \mathbb{E}\left[\sup_{f \in \mathcal{G}} \sum_{i \in A} \epsilon_i f(X_i) \right] .$$

The Rademacher complexity is a standard measure of complexity in classification problems [BM02]. It can be upper bounded by $\text{comp}/\sqrt{N}$ where comp is a measure of complexity such as the square root of the VC dimension or the Dudley's entropy integral or the Gaussian mean width of the class F see for example [BBL05, Kol11, BM02, BLM13, DGL96] for a presentation of these classical bounds. Our second assumption is simply that the Rademacher complexity of the class F is finite.

Assumption 2. *The Rademacher complexity of F is finite, $\mathcal{R}(F) < \infty$.*

Assumption 2 holds in the linear classification example under Assumption 1 since it follows from Cauchy-Schwarz inequality that $\mathcal{R}(F) \leq \theta_2 \sqrt{|\mathcal{I}|p}$. Finally, our last assumption is that the loss function ℓ considered is Lipschitz in the following sense.

Assumption 3. *The loss function ℓ satisfies for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and all $f, f' \in F$,*

$$|\ell_f(x, y) - \ell_{f'}(x, y)| \leq L|f(x) - f'(x)| .$$

Assumption 3 holds for classical convex relaxation of the $0-1$ loss such as hinge loss ℓ^{hinge} or logistic loss ℓ^{logistic} as in Example 1. In these examples, the constant L can be chosen equal to 1. It also covers non-convex loss functions such as the one in [BBS17, Cat12, AC11] or sigmoid loss functions such as the one used in Deep Learning. In particular, our results do not follow from other work on MOM estimators using convex loss functions such as in [CLL19b].

5.3 Theoretical guarantees

Our first result follows Vapnik-Chervonenkis's original risk bound for the ERM and shows that $\widehat{f}^{0-1}_{\text{ERM}}$ is insensitive to the presence of outliers in the dataset. Moreover, it quantifies this robustness property since Vapnik-Chervonenkis's rate of convergence is still achieved by $\widehat{f}^{0-1}_{\text{ERM}}$ when there are less than (number of observations) times (Vapnik's rate of convergence) outliers.

Theorem 29. *Let $\mathcal{F}$ denote a collection of classifiers. Let $\mathcal{L}^{0-1}_{\mathcal{F}} = \{\ell^{0-1}_f - \ell^{0-1}_{f^*} : f \in \mathcal{F}\}$ be the family of excess loss functions indexed by $\mathcal{F}$ where $f^* \in \operatorname{argmin}_{f\in\mathcal{F}} R^{0-1}(f)$. For all $K > 0$, with probability at least $1 - e^{-K}$, we have*

$$R^{0-1}(\widehat{f}^{0-1}_{ERM}) - \inf_{f\in\mathcal{F}} R^{0-1}(f) \leq \frac{2\mathcal{R}(\mathcal{L}^{0-1}_{\mathcal{F}})}{N} + \sqrt{\frac{K}{2|\mathcal{I}|}} + \frac{2|\mathcal{O}|}{N}.$$

Theorem 29 is proved in Section 5.6.1. It is an adaptation of Vapnik-Chervonenkis's proof of the excess risk bounds satisfied by $\widehat{f}^{0-1}_{\text{ERM}}$ in the presence of outliers.

Remark 2. *In the last result, one can easily bound the excess risk using $\mathcal{R}(\mathcal{F})$ instead of $\mathcal{R}(\mathcal{L}^{0-1}_{\mathcal{F}})$ since*

$$\mathcal{R}(\mathcal{L}^{0-1}_{\mathcal{F}}) = \max_{A\in\{\mathcal{I},\mathcal{J}\}} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{i\in A} \epsilon_i(f(X_i) - f^*(X_i))\right] = \mathcal{R}(\mathcal{F}) \ .$$

The final bound is of similar flavor: for all $K > 0$, with probability at least $1 - e^{-K}$, we have

$$R^{0-1}(\hat{f}^{0-1}_{ERM}) - \inf_{f\in\mathcal{F}} R^{0-1}(f) \lesssim \max\left(\frac{\mathcal{R}(\mathcal{F})}{N}, \sqrt{\frac{K}{|\mathcal{I}|}}, \frac{|\mathcal{O}|}{N}\right) \ . \tag{5.3.1}$$

Remark 3. *When $\mathcal{F}$ is the class of all linear classifiers, that is when $\mathcal{F} = \{\text{sgn}(\langle t, \cdot\rangle) : t \in \mathbb{R}^p\}$, one has $\mathcal{R}(\mathcal{F}) \leq \sqrt{|\mathcal{I}|p}$ (see Theorem 3.4 in [BBL05]). Therefore, when $|\mathcal{I}| \geq N/2$, Theorem 29 implies that for all $1 \leq K \leq p$, with probability at least $1 - \exp(-K)$,*

$$R^{0-1}(\hat{f}^{0-1}_{ERM}) - \inf_{f\in\mathcal{F}} R^{0-1}(f) \lesssim \max(\sqrt{p/N}, |\mathcal{O}|/N) \ .$$

As a consequence, when the number of outliers is such that $|\mathcal{O}| \lesssim N\times\sqrt{p/N}$, Vapnik-Chervonenkis's classical "slow" rate of convergence $\sqrt{p/N}$ is still achieved by the ERM estimator even if $|\mathcal{O}|$ outliers have polluted the dataset. The interested reader can also check that "fast rates" p/N could also be achieved by the ERM estimator in the presence of outliers if $|\mathcal{O}| \lesssim p$ and when the so-called strong margin assumption holds (see, [BBL05]). Note also that the previous remark also holds if F is a class with VC dimension p beyond the case of indicators of half spaces.

The conclusion of Theorem 29 can be misleading in practice. Indeed, theoretical performance of the ERM estimator for the $0-1$ loss function are not downgraded by outliers, but its proxies based on convex relaxation (F, ℓ) of $(\mathcal{F}, \ell^{0-1})$ are. This can be seen on the toy example in Figure 5.1 and in Figure 5.4 from Section 5.5. In this work, we propose a robust surrogate, based on MOM estimators of the risk and defined in (5.1.2), to the natural empirical risk estimation of the risk which works for unbounded loss functions. In the next result, we prove that the MOM minimizer $\hat{f}_{\mathrm{MOM},K}$ defined as

$$\hat{f}_{\mathrm{MOM},K} \in \operatorname*{argmin}_{f \in F} \mathrm{MOM}_K(\ell_f) \tag{5.3.2}$$

satisfies an excess risk bound under weak assumptions introduced in Section 5.2.

Theorem 30. *Grant Assumptions 1, 2 and 3. Assume that $N > K > 4|\mathcal{O}|$ and let $\Delta = 1/4 - |\mathcal{O}|/K$. Then, with probability larger than $1 - 2\exp(-2\Delta^2 K)$, we have*

$$R(\hat{f}_{MOM,K}) \leq \inf_{f \in F} R(f) + 4L \max\left(\frac{4\mathcal{R}(F)}{N}, 2\theta_2\sqrt{\frac{K}{N}}\right) .$$

Theorem 30 is proved in Section 5.6.2. Compared to Theorem 29, $\hat{f}_{\mathrm{MOM},K}$ achieves the same rate $(\mathcal{R}(F)/N) \vee (\sqrt{K/N})$ under the same conditions on the number of outliers with the same exponential control of the probability as for the ERM estimator f^{0-1}_{ERM}. The main difference is that the loss function may be unbounded, which is often the case in practice. Moreover, unlike classical analysis of ERM obtained by minimizing an empirical risk associated with a convex surrogate loss function, we only need a second moment assumption on the class F.

These theoretical improvements have already been noticed in previous works [Min15, DLLO16, LL18, LM19b, LM+19d, LM19c, Men, LL20]. Contrary to tournaments of [LM19c], Le Cam MOM estimators of [LL18] or minmax MOM estimators [LL20], Theorem 30 does not require the small ball assumption on F but only shows "slow rates" of convergence. These slow rates are minimax optimal in the absence of a margin or Bernstein assumption [BM06, MT99]. Removing the small ball assumption may be useful in some examples. As an illustration, consider the toy example where the design

$$X = \begin{bmatrix} \mathbf{1}_{W \in I_1} \\ \vdots \\ \mathbf{1}_{W \in I_d} \end{bmatrix}$$

where $I_1, \ldots, I_d$ is a partition of a measurable set $\mathbb{W}$ into subsets such that $\mathbb{P}(W \in I_i) = 1/d$ for each $i \in \{1, \ldots, d\}$. Then $\mathbb{X} = [0,1]^d$ and one can consider the set F of linear functions $f(X) = \langle t, X \rangle$, where the Euclidean norm of t satisfies $\|t\| \leqslant B\sqrt{d}$. Then, as $\|\langle t, \cdot \rangle\|^2_{L^2} = \sum_{i=1}^d t_i^2 \mathbb{P}(W \in I_i) = \|t\|^2/d$, Assumption 1 holds with $\theta_2 = B$. In this example, Assumption 2 holds with $\mathcal{R}(F) \leq \theta_2\sqrt{|\mathcal{I}|d} \leqslant B\sqrt{Nd}$. It follows from Theorem 30 that the remainder term in this example is bounded from above by

$$4LB \max\left(4\sqrt{\frac{d}{N}}, 2\theta_2\sqrt{\frac{K}{N}}\right) .$$

In particular, it converges to 0 if $d \vee K \ll N$. By comparison, in the same example, it is shown in [CLL19b] that the remainder term converges to 0 only if $d \lesssim \sqrt{N}$.

Proof of Theorem 30 does not enable fast rates to be obtained. Indeed, the non-linearity of the median excludes the possibility of using localization techniques leading to these fast rates. However, we show in the simulation study (cf. left side picture of Figure 5.13) that fast rates seem to be reached by the MOM minimizer.

Remark 4. *The MOM principle has been used together with Lipschitz loss functions recently in [CLL19b]. In this paper, a minmax MOM estimator is constructed which can achieve fast rates of convergence under a margin condition. The argument from [CLL19b] relies heavily on the convexity of the loss – an assumption we do not have here. The reason why the convexity of the loss is so important in [CLL19b] is that it allows to exclude (as potential minmax MOM estimator) all the functions in F outside a L_2-ball centered in f^* with radius r if all the functions in F in the sphere $f^* + r\mathcal{S}_2$ are excluded. Therefore, thanks to convexity, the latter " homogeneity argument" reduces the problem to the study of the sub-model $F \cap (f^* + r\mathcal{S}_2)$ (which is bounded in L_2 with the right radius r). Here, no such homogeneity argument can be used because we did not assume the loss to be convex. Nevertheless, if we assume that the loss is convex then we may still apply Theorem 4 in [CLL19b] and replace all the localized sets by the entire set F and the variance term by the L_2 uniform bound θ_2 coming from Assumption 1 to obtain a similar result as Theorem 30 for a minmax MOM estimator. These stronger results require the convexity of the loss and a Bernstein assumption that may be satisfied only under strong assumptions as discussed in the toy example.*

Finally, the main advantage of our approach is its simplicity, we just have to replace empirical means by their MOM alternative in the definition of the ERM estimator. Moreover, as expected, this simple alternative to ERM estimators yields a systematically way to modify algorithms designed for approximating the ERM estimator. The resulting "MOM versions" of these algorithms are both faster and more robust than their original "ERM version". Before illustrating these facts on simulations, let us describe algorithms approximating MOM minimizers.

5.4 Computation of MOM minimizers

In this section, we present a generic algorithm to provide a MOM version of descent algorithms. We study the differentiation property of the objective function $f \to \mathrm{MOM}_K(\ell_f)$. Then we check on simulated and real databases the robustness and outlier detection property of these MOM algorithms.

5.4.1 MOM algorithms

The general idea is that any descent algorithms such as gradient descent, Newton method, alternate gradient descent, etc. (cf. [MB11, Bub15, BV04, BJMO12]) can easily be turned into a robust MOM-version. To illustrate this idea, a basic gradient descent is analyzed in the sequel. We start with a block splitting policy of the database.

The choice of blocks greatly influences the practical performance of the algorithm. In particular, a recurring flaw is that iterations tend to get stuck in local minima, which greatly slows the convergence of the alogorithm. To overcome this default and improve the stability of the procedure,

a new partition is constructed at each iteration by drawing it uniformly at random, cf. step **2** of Algorithm 1.

Let $\mathcal{S}_N$ denote the set of permutations of $\{1,\dots,N\}$. For each $\sigma \in \mathcal{S}_N$, let $B_0(\sigma) \cup \cdots \cup B_{K-1}(\sigma) = \{1,\dots,N\}$ denote an equipartition of $\{1,\dots,N\}$ defined for all $j \in [\![0, K-1]\!]$ by

$$B_j(\sigma) = \{\sigma(Kj+1),\dots,\sigma(K(j+1))\} = \sigma(\{Kj+1,\dots,K(j+1)\}) \ .$$

To simplify the presentation, let us assume the class F to be parametrized $F = \{f_u : u \in \mathbb{R}^p\}$, for some $p \in \mathbb{N}^*$. Let's assume that the function $u \mapsto f_u$ is as regular as needed and convex (a typical example is $f_u(x) = \langle u, x \rangle$ for all $x \in \mathbb{R}^p$). Denote by $\nabla_u \ell_{f_u}$ the gradient or a subgradient of $u \mapsto \ell_{f_u}$ in $u \in \mathbb{R}^p$. The step-sizes sequence is denoted by $(\eta_t)_{t\geq 0}$ and satisfies the classical conditions: $\sum_{t=1}^{\infty} \eta_t = \infty$ and $\sum_{t=1}^{\infty} \eta_t^2 < \infty$. Iterations will go on until a stopping time $T \in \mathbb{N}^*$ has been achieved. With these notations, a generic MOM version of a gradient descent algorithm (with random choice of blocks) is detailed in Algorithm 1 below.

Algorithm 1: MOM gradient descent algorithm.

input : $u_0 \in \mathbb{R}^p$, $K \in [\![3, N/2]\!]$, $T \in \mathbb{N}^*$ and $(\eta_t)_{t\in\{0,\dots,T-1\}} \in \mathbb{R}_+^T$
output : a MOM version of BSGD

for $t = 0, \cdots, T-1$ **do**
- choose a permutation at random: $\sigma_t \sim \mathrm{Unif}(\mathcal{S}_N)$,
- build a partition of the dataset: $B_0(\sigma_t), \dots, B_{K-1}(\sigma_t)$,
- find a median block: $k_{med}(t)$ s.t. $\mathrm{MOM}_K\big(\ell_{f_{u_t}}\big) = P_{B_{k_{med}(t)}(\sigma_t)}\big(\ell_{f_{u_t}}\big)$,
- do a descent step on the median block

$$u_{t+1} = u_t - \eta_t \nabla_t \text{ where } \nabla_t = \sum_{i \in B_{k_{med}(t)}(\sigma_t)} \nabla_{u_t} \ell_{f_{u_t}}(X_i, Y_i).$$

end
Return u_T

Remark 5 (MOM gradient descent algorithm and stochastic block gradient descent)**.** *Algorithm 1 can be seen as a stochastic block gradient descent (SBGD) algorithm minimizing the function $t \to \mathbb{E}\ell_t(X,Y)$ using a given dataset. The main difference with the classical SBGD is that the choice of the block along which the gradient direction is performed is chosen according to a centrality measure computed thanks to the median operator in step **4** of Algorithm 1.*

In Section 5.5, we use the MOM principle (as in the generic Algorithm 1) to construct MOM versions for various classical algorithms such as Perceptron, Logistic Regression, Kernel Logistic Regression, SGD Classifiers or Multi-layer Perceptron.

5.4.2 Differentiation properties of $f \to \mathrm{MOM}_K(\ell_f)$, random partition and local minima

Let us try to explain the choice of the descent direction ∇_t in step **5** of Algorithm 1. In the previous sections, we introduced and studied MOM minimization procedures which are minimizers of $f \to \mathrm{MOM}_K(\ell_f)$ over F. The optimization problem that needs to be solved to construct a MOM minimizer is not convex, in general. It therefore raises difficulties since classical tools and algorithms from the convex optimization toolbox cannot be used a priori. Nevertheless, one may still try to do a gradient descent algorithm for this (non-convex) optimization problem with objective function given by $f \to \mathrm{MOM}_K(\ell_f)$. To do so, we first need to check the differentiation properties of $f \to \mathrm{MOM}_K(\ell_f)$ over F.

First observe that the descent direction ∇_t is the gradient of the empirical risk constructed on the median block of data $B_{k_{med}(t)}(\sigma_t)$ at f_{u_t} (we recall that F is parametrized like $\{f_u : u \in \mathbb{R}^p\}$). A classical Gradient Descent algorithm on $f \to \mathrm{MOM}_K(\ell_f)$ starting from f_{u_t} would use a gradient at f_{u_t} of the objective function. Let us first identify situations where this is indeed the case i.e. when ∇_t is the gradient of $f \to \mathrm{MOM}_K(\ell_f)$ in f_{u_t}.

Assumption 4. *For almost all datasets $\mathcal{D}_N = \{(X_i, Y_i) : i = 1, \dots, N\}$ and Lebesgue almost all $u \in \mathbb{R}^p$, there exists an open convex set B containing u such that for any equipartition of $\{1, \dots, N\}$ into K blocks $B_1, \dots, B_K$ there exists $k_{med} \in \{1, \cdots, K\}$ such that for all $v \in B$, $P_{B_{k_{med}}}(\ell_{f_v}) \in MOM_K(\ell_{f_v})$.*

In other word, under Assumption 4, for almost all $u_0 \in \mathbb{R}^p$, the median block $B_{k_{med}}$ achieving $\mathrm{MOM}_K(\ell_{f_{u_0}})$ is the same as the one achieving $\mathrm{MOM}_K(\ell_{f_u})$ for all u in an open and convex neighborhood B of u_0. It means that the objective function $u \to \mathrm{MOM}_K(\ell_{f_u})$ is equal to the empirical risk function over the same block of data $B_{k_{med}}$: $u \to P_{B_{k_{med}}} \ell f_u$, on B. Since B is an open set and that $u \to P_{B_{k_{med}}} \ell_{f_u}$ is differentiable in u_0 then the objective function $u \to \mathrm{MOM}_K(\ell_{f_u})$ is also differentiable in u_0 and the two gradients coincide:

$$\nabla\big(u \to \mathrm{MOM}_K(\ell_{f_u})\big)_{|u_0} = \nabla\Big(u \to P_{B_{k_{med}}} \ell_{f_u}\Big)_{|u_0}. \tag{5.4.1}$$

Under Assumption 4, Algorithm 1 is indeed a gradient descent algorithm performed on the objective function $u \in \mathbb{R}^p \to \mathrm{MOM}_K(\ell_{f_u})$.

Let us give an example where Assumption 4 is satisfied. Let $B_1 \cup \cdots \cup B_K = \{1, \dots, N\}$ be an equipartition and let ψ be defined for all $x = (x_i)_{i=1}^N \in \mathbb{R}^N$ and $u \in \mathbb{R}^p$ by,

$$\psi_u(x) = \mathrm{MOM}_K\big(f_u(x)\big) = \mathrm{median}\left(\frac{K}{N}\sum_{i \in B_k} f_u(x_i), 1 \le k \le K\right) = P_{B(K/2)(u)}(f_u),$$

where for all blocks $B \subset \{1, \dots, N\}$, $P_B f_u = |B|^{-1} \sum_{i \in B} f_u(x_i)$ and the blocks $B(k)(u), k = 1, \dots, K$ are rearranged blocks defined such that $P_{B(1)(u)}(f_u) \ge \cdots \ge P_{B(K)(u)}(f_u)$. Proposition 1 below shows that Assumption 4 is satisfied in several situations. Its proof can be found in Section 5.6.

Proposition 1. *Let $X_1, \dots, X_N$ be N real-valued random variables, suppose K is odd and N is a multiple of K. Let $(f_u)_{u \in \mathbb{R}^d}$ be a family of functions with values in $\mathbb{R}$. Assume that for all*

$x \in \mathbb{R}$, the function $u \mapsto f_u(x)$ is Lipschitz and the probability distribution of $f_u(X_1)$ has a law absolutely continuous with respect to Lebesgue measure. Then, with probability 1, Assumption 4 is satisfied, in particular, the partial derivative of $u \mapsto \psi_{f_u}((X_i)_{i=1}^N) = MOM_K(f_u((X_i)_{i=1}^N))$ with respect to the j^{th} coordinate is given for almost all $X_1, \dots, X_N$ by

$$\partial_j \psi_{f_u}((X_i)_{i=1}^N) = \frac{K}{N} \sum_{i \in B(\lceil K/2 \rceil)(u)} \partial_j f_u(X_i)$$

where ∂_j denote the derivative with respect to the j^{th} coordinate of u.

Under Assumption 4, the picture of the MOM gradient descent algorithm is pretty simple and depicted in Figure 5.2. At every step t, the median operator makes a partition of $\mathbb{R}^p$ into K cells $\mathcal{C}_k(t) = \{u \in \mathbb{R}^p : \mathrm{MOM}_K(\ell_{f_u}) = P_{B_k}\ell_{f_u}\}$ for $k = 1, \dots, K$ – this partition changes at every step because the blocks $B_1, \dots, B_K$ are chosen randomly at the beginning of every step according to the random partition σ_t. We want every iteration u_t of the MOM algorithm to be in the interior of a cell and not on a frontier in order to differentiate the objective function $u \to \mathrm{MOM}_K(\ell_{f_u})$ at u_t. This is indeed the case under Assumption 4, given that in that case, there is an open neighbor B of u_t such that for all $v \in B$, $\mathrm{MOM}_K(\ell_{f_v}) = P_{B_k}\ell_{f_v}$ where the index $k = k_{med}$ of the block is common to every $v \in B$. Therefore, to differentiate the objective function $u \to \mathrm{MOM}_K(\ell_{f_u})$ at u_t one just needs to differentiate $u \to P_{B_k}\ell_{f_u}$ at u_t. The objective function to minimize is differentiable almost everywhere under Assumption 4 and a gradient of the objective function is given by $\nabla(u \to P_{B_k}\ell_{f_u})_{|u=u_t}$, that is ∇_t from step **5** of Algorithm 1.

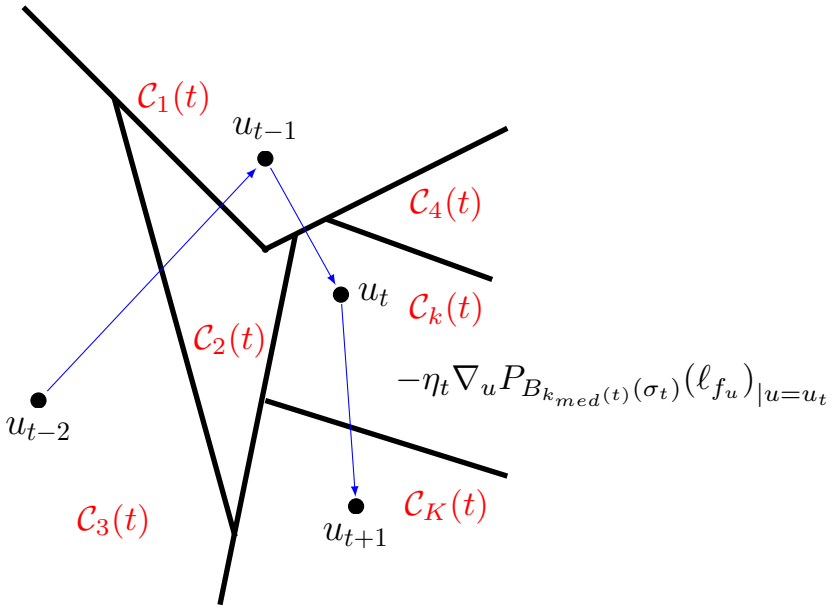


Figure 5.2: Partition of $\mathbb{R}^p$ at step t by the median operator and iteration number $t-2, t-1, t$ and $t+1$ of the MOM gradient descent algorithm. Under Assumption 4, there is a natural descent direction given at step t by $-\nabla_u(u \to P_{B_{k_{med}(t)}(\sigma_t)}(\ell_{f_u})))_{|u=u_t}$.

Under Assumption 4, the importance of partitioning the dataset at each new iteration is more transparent. Indeed, if we were to perform the MOM gradient descent such as in

Algorithm 1 but without a new partition at each step then local minima of the K empirical risks $u \to P_{B_k}\ell_{f_u}, k \in [K]$ may mislead the descent algorithm. Indeed, if a minimum of $u \to P_{B_k}\ell_{f_u}$ for some $k \in [K]$ is in the cell $\mathcal{C}_k$ then the algorithm will reach this minimum without noticing that a " better" minimum is in another cell. That is why re-partitioning the dataset of every iteration avoid this effect and speed up the convergence (see [LL20] for experiments).

5.4.3 Complexity of MOM risk minimization algorithms

In this section, we compute the computational cost of several MOM versions of some classical algorithms. Let $C(m)$ be the computational complexity of a single standard gradient descent update step on a dataset of size m and let $L(m)$ be the computational complexity of the evaluation of the empirical risk $(1/m)\sum_{i\in B}\ell_f(X_i, Y_i)$ of some $f \in F$ on a dataset B containing m data. Here the computational complexity is simply the number of basic operations needed to perform a task [AB09b].

For each epoch, we begin by computing the "MOM empirical risk". We perform K times N/K evaluations of the loss function, then we sort the K means of these blocks of loss to finally get the median. The complexity of this step is then $O(KL(N/K) + K\ln(K))$, assuming that the sort algorithm is in $O(K\ln(K))$ (like *quick sort* [Hoa62]). Then we do the gradient step on a sample of size N/K. Hence, the time complexity of this algorithm is

$$O(T(KL(N/K) + K\ln(K) + C(N/K))).$$

Example 2 (Linear complexity " ERM version" algorithms)**.** *For example, if the standard gradient step and the loss function evaluation have linear complexity – like Perceptron or Logistic Regression – the complexity of the MOM algorithm is $O(T(N + K\log(K)))$ against $O(TN)$ for the ERM algorithm. Therefore, the two complexities are of the same order and the only advantage of MOM algorithms lies in their robustness to outliers and heavy-tailed properties.*

Example 3 (Super-linear complexity " ERM version" algorithms)**.** *If, on the other hand, the complexity is more than linear as for Kernel Logistic Regression (KLR), taking into account the matrix multiplications whose complexity can be found in [Gal14], the complexity of the MOM version of KLR, due to the additional need of the computation of the kernel matrix, is $O(N^2 + T(N^2/K + K\log(K) + (N/K)^{2.373}))$ against $O(TN^{2.373})$ for the standard " ERM version". MOM versions of KLR are therefore faster than the classical version of KLR on top of being more robust. This advantage comes from the fact that MOM algorithms work on blocks of data instead on the entire dataset at every step. More informations about Kernel Logistic Regression can be found in [Rot01] for example.*

In this last example, the complexity comes in part from the evaluation of the kernel matrix that can be computationally expensive. Following the idea that MOM algorithms are performing ERM algorithm restricted to a wisely chosen block of data, then one can modify our generic strategy in this particular example to reduce drastically its complexity. The idea here is that we only need to construct the kernel matrix on the median block. The resulting algorithm, called Fast KLR MOM is described in Figure 2.

In Figure 2, we compute only the block kernel matrices, denoted by $N^1, \ldots, N^k$ and constructed from the samples in the block B_k. We also denote by N_i^k the i^{th} row in N^k.

Algorithm 2: Description of Fast KLR MOM algorithm.

input : $\alpha_0 \in \mathbb{R}^p$, $K \in [\![3, N/2]\!]$, $T \in \mathbb{N}^*$, $(\eta_t)_{t\in\{0,\dots,T-1\}} \in \mathbb{R}_+^T$, $\beta \in \mathbb{R}_+^*$, $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a positive definite kernel and a bloc decomposition $B_1, \dots, B_K$ of $\{1, \dots, N\}$.

output : a MOM version of KLR classifiers

Construct the bloc Kernel matrices $N^k = (\kappa(X_i, X_j))_{i,j\in B_k}$ for $1 \le k \le K$,

for $t = 0, \cdots, T-1$ **do**

find a median block: $k_{med}(t)$ s.t. $\mathrm{MOM}_K\big(\ell_{f_{\alpha_t}}\big) = P_{B_{k_{med}(t)}}\big(\ell_{f_{\alpha_t}}\big)$ with

$$P_{B_k}(\ell_{f_{\alpha_t}}) = \frac{1}{|B_k|}\sum_{i\in B_k} \ln(1 + e^{-N_i^k \alpha_t^k Y_i}) + \beta \sum_{k=1}^{K} (\alpha_t^k)^T N^k \alpha_t^k,$$

where α_t^k is the vector in $\mathbb{R}^{|B_k|}$ made of the coordinates of α_t with indices in B_k.

Do an IRLS descent step for KLR with weight matrice $W_{k_{med}(t)}$, design matrice $X_{k_{med}(t)}$ and labels $y_{k_{med}(t)}$ on $B_{k_{med}(t)}$

$$\alpha_{t+1}^{k_{med}(t)} = \alpha_t^{k_{med}(t)}(1-\eta_t) + \eta_t (X_{k_{med}(t)}^T W_{k_{med}(t)} X_{k_{med}(t)})^{-1} X_{k_{med}(t)}^T W_{k_{med}(t)} y_{k_{med}(t)}.$$

$$\alpha_{t+1}^{k} = \alpha_t^{k}(1-\eta_t), \quad \forall k \neq k_{med}(t).$$

end

Return $\alpha_T, N_{k_{med}(T)}$

There are several drawbacks in the approach of Algorithm 2. First, the blocks are fixed at the beginning of the algorithm; therefore the algorithm needs a bigger dataset to work well and it may converge to a local minimum. Nonetheless, from the complexity point of view, this algorithm will be much faster than both the classical KLR and MOM KLR (see below for a computation of its complexity) which is important given the growing use of kernel methods on very large databases for example in biology. The choice of K should ultimately realize a trade-off between complexity and performance (in term of accuracy for example) when dealing with big databases containing few outliers.

Example 4 (Complexity of Fast KLR-MOM algorithm)**.** *The complexity of Fast KLR-MOM is* $O(N^2/K + T(N^2/K + K\log(K) + (N/K)^{2.373}))$ *against* $O(TN^{2.373})$ *for the ERM version.*

5.5 Implementation and Simulations

5.5.1 Basic results on a toy dataset

The toy model we consider models outliers due to human or machine errors we would like to ignore in our learning process. It is also a dataset corrupted to make linear classifiers fail. The dataset is a 2D dataset constituted of three "labeled Gaussian distribution". Two informative Gaussians $\mathcal{N}((-1,-1), 1.4I_2)$ and $\mathcal{N}((1,1), 1.4I_2)$ with label respectively 1 and -1 and one outliers Gaussian $\mathcal{N}((24,8), 0.1I_2)$ with label 1. In other words, the distribution of informative data is given by $\mathcal{L}(X|Y=1) = \mathcal{N}((-1,-1), 1.4I_2)$, $\mathcal{L}(X|Y=-1) = \mathcal{N}((1,1), 1.4I_2)$ and

$\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1) = 1/2$. Outliers data have distribution given by $Y = 1$ a.s. and $X \sim \mathcal{N}((24, 8), 0.1I_2)$.

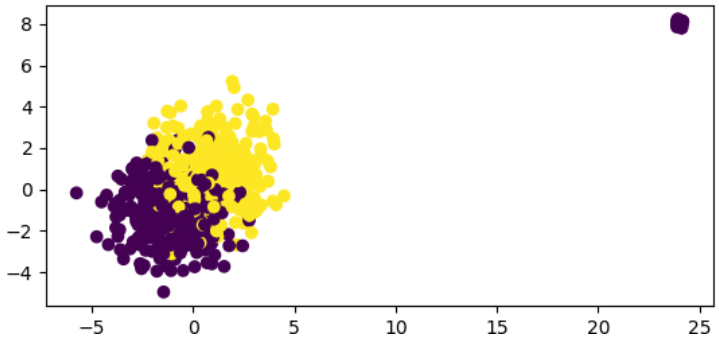


Figure 5.3: Scatter plot of 630 samples from the training dataset (600 informative data, 30 outliers), the color of the points correspond to their labels.

The algorithms we study are the MOM adaptations of Perceptron, Logistic Regression and Kernel Logistic Regression.

Based on our theoretical results, we know that the number of blocks K has to be larger than 4 times the number of outliers for our procedure to be on the safe side. The value $K = 120$ is therefore used in all subsequent applications of MOM algorithms on the toy dataset except when told otherwise. To quantify performance, we compute the miss-classification error on a clean dataset made of data distributed like the informative data.

For Kernel Logistic Regression, we study here a linear kernel because outliers in this dataset are clearly adversarial when dealing with linear classifiers. The algorithm can also use more sophisticated kernels, a comparison of the MOM algorithms with similar ERM algorithms is represented in figure 5.4, the ERM algorithms are taken from the python library scikit-learn [PVG+11] with their default parameters.

Figure 5.4 illustrates resistance to outliers of MOM's algorithms compared to their classical version.

These first results are completed in Figure 5.5 where we computed accuracy on several run of the algorithms. These results confirm the visual impression of our first experiment.

Finally, we illustrate our results regarding complexities of the algorithms on a simulated example. MOM algorithms have been computed together with state-of-the art algorithms from scikit-learn [PVG+11] (we use Random forest, SVM classifier as well as SGD classifier optimizing Huber loss which entail a robustness in Y but not in X, see [HR09, Chapter 7]) on a simulated dataset composed of two Gaussian blobs $\mathcal{N}((-1, -1), 1.4I_2)$ and $\mathcal{N}((1, 1), 1.4I_2)$ with label respectively 1 and -1. We sample 20000 points for the training dataset and 20000 for the test dataset. The parameters used in the algorithms are those for which we obtained the optimal accuracy, (this accuracy is illustrated in the next section). Time of training plus time of evaluation on the test dataset are gathered in Figure 5.6.

Not surprisingly, very efficient versions of linear algorithms from Python's library are extremely

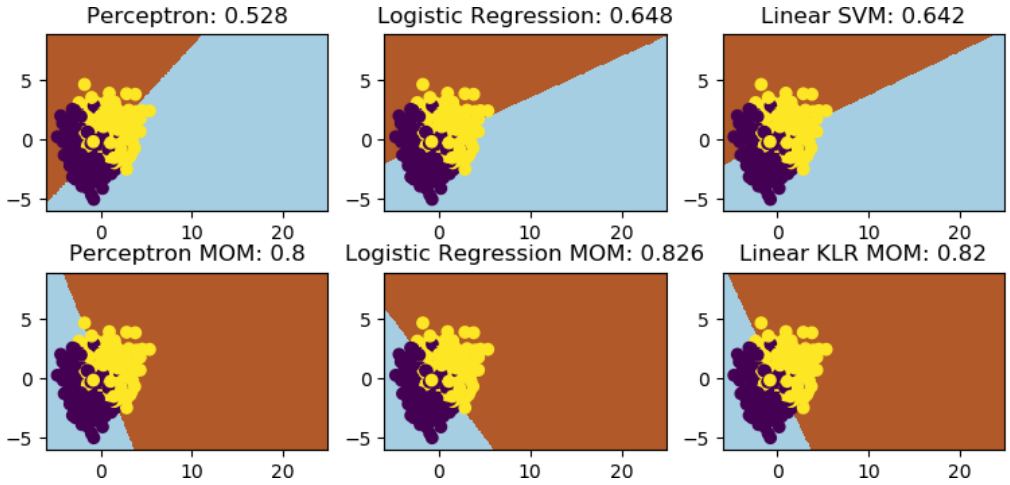


Figure 5.4: Scatter plot of 500 samples from the test dataset (500 informative data), the color of the points correspond to their labels and the background color correspond to the prediction. The score in the title of each subfigure is the accuracy of the algorithm.

fast (results are sometimes provided before we even charged the dataset in some experiments). The performance of our algorithm are nevertheless acceptable in general (around 5 times longer than random forest for example). The important fact here is that non linear algorithms such as SVM take much more time to provide a result. FAST KLR MOM is able to reduce substantially the execution time of SVM with comparable predictive performance.

5.5.2 Applications on real datasets

We used the HTRU2 dataset, also studied in [LSCB15], that is provided by the UCI Machine Learning Repository. The goal is to detect pulsars (a rare type of Neutron star) based on radio emission detectable on earth from which features are extracted to gives us this dataset. The problem is that most of the signal comes from noise and not pulsar, the goal is then to classify pulsar against noise, using the 17 898 points in the dataset.

The accuracy of different algorithms is obtained using on several runs of the algorithms each using 4/5 of the datasets for training and 1/5 for testing algorithms. Boxplots presenting performance of various algorithms are displayed in Figure 5.7. To improve performance, RBF kernel was used both for KLR MOM and Fast KLR MOM.

5.5.3 Outlier detection with MOM algorithms

When we run MOM version of a descent algorithm, we select at each step a block of data points realizing the median of a set of "local/block empirical risk" at the current iteration of the algorithm. The number of times a point is selected by the algorithm can be used as a depth function measuring reliability of the data. Note that this definition of depth of a data point has

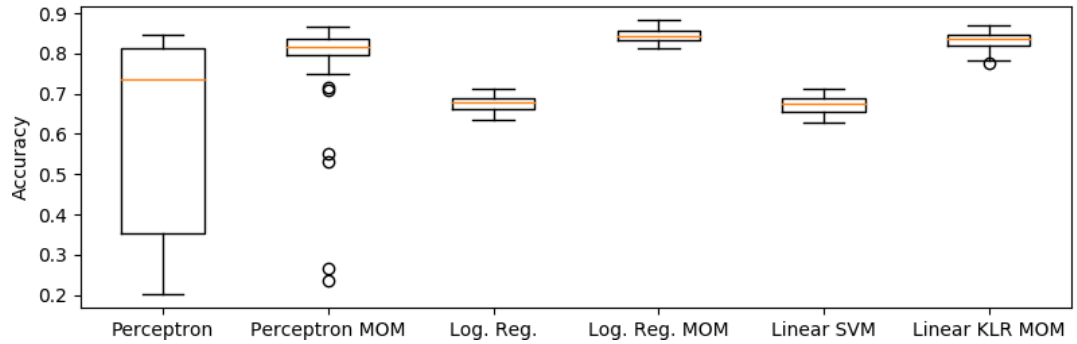


Figure 5.5: Comparison of the MOM algorithms and their counterpart with the boxplots of the accuracy on the test dataset from 50 runs of the algorithms on 50 sample of the training/test toy dataset (one run for each dataset sampled).

Algorithm	Perceptron MOM	Log. Reg. MOM	KLR MOM	Fast KLR MOM
Time (s)	1.06	1.05	13.6	1.2
Algorithm	Rand. Forest	SVM	SGD Hub. loss	
Time (s)	0.21	9.0	0.0078	

Figure 5.6: Time of different algorithms on a simulated dataset .

the advantage of taking into account the learning task we want to solve, that is the loss ℓ and the class F. It means that outliers are considered w.r.t. the problem we want to solve and not w.r.t. some a priori notion of centrality of points in $\mathbb{R}^d$ unrelated with the problem considered at the beginning.

We apply this idea on the toy dataset with the Logistic Regression MOM algorithm. Results are gathered in a sorted histogram given in Figure 5.8. Red bars represent outliers in the original datasets.

Quite remarkably, outliers are in fact those data that have been used the smallest number of times. The method targets a very specific type of outliers, those disturbing the classification task at hand. If there was a point very far away from the bulk of data but in the half-space of its label, it wouldn't be detected.

This detection algorithm doesn't scale well when the dataset gets bigger as a large number of iterations is necessary to choose each point a fair number of times. For bigger datasets, we suggest to adapt usual outlier detection algorithms [Agg13]. We emphasize that clustering techniques and K-Means are rather easy to adapt in a MOM algorithm and detect points far from the bulk of data. This technique might greatly improve usual K-Means as MOM K-Means is more robust.

Let us now analyze the effect of K on the outlier detection task. The histogram of the 1000 smaller counts of points of HTRU2 dataset as K gets bigger is plotted in Figure 5.9.

It appears from Figure 5.9 that K measures the sensitivity of the algorithm. Severe outliers (as in the toy example) are detected for small K while mild outliers are only discovered as K gets

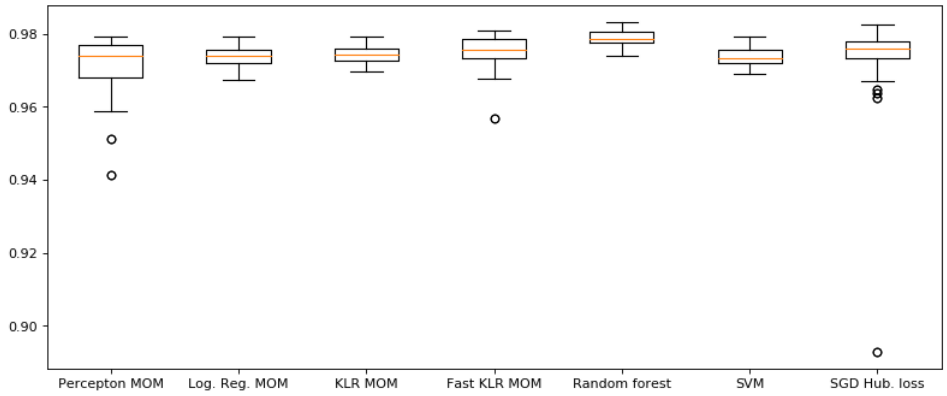


Figure 5.7: Comparison of the MOM algorithms and common algorithms with the boxplots and the medians of the accuracy $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{f}(x_i)=y_i\}$ on the test dataset from 50 runs of the algorithms on 100 sample of a 4/5 cut of the dataset HTRU2 (one run is trained on a sample of 4/5 of the dataset and tested on the remaining 1/5)

bigger.

It seems therefore that the optimal choice of K in MOM depends on the task one is interested in. For classification, K should be as small as possible to get better risk bounds (but it still should be larger than the number of outliers) whereas for detecting outliers we may want to choose K much larger to even detect an outlier, (but it should also be small enough for the underlying classification to perform correctly). As a proof of concept, for Pulsar database, we got optimal results choosing $K = 10$ for classification whereas we only detect a significant amount of outliers when K is around 1000.

5.6 Proofs

5.6.1 Proof of Theorem 29

We adapt Vapnik-Chervonenkis's classical analysis [Vap98] of excess risk bound of ERM to a dataset corrupted by outliers. We first recall that $f^* \in \operatorname{argmin}_{f\in\mathcal{F}} R^{0-1}(f)$ and for all $f \in \mathcal{F}$, the excess loss function of f is $\mathcal{L}_f^{0-1} = \ell_f^{0-1} - \ell_{f^*}^{0-1}$. For simplicity we denote $\hat{f} = \widehat{f}_{\mathrm{ERM}}^{0-1}$ and for all $f \in \mathcal{F}$, $\mathcal{L}_f^{0-1} = \mathcal{L}_f$ and $R(f) = R^{0-1}(f)$.

It follows from the definition of the ERM estimator that $P_N\mathcal{L}_{\hat{f}} \leq 0$. Therefore, if we denote by $P_{\mathcal{I}}$ (resp. $P_{\mathcal{O}}$) the empirical measure supported on $\{(X_i, Y_i) : i \in \mathcal{I}\}$ (resp. $\{(X_i, Y_i) : i \in \mathcal{O}\}$),

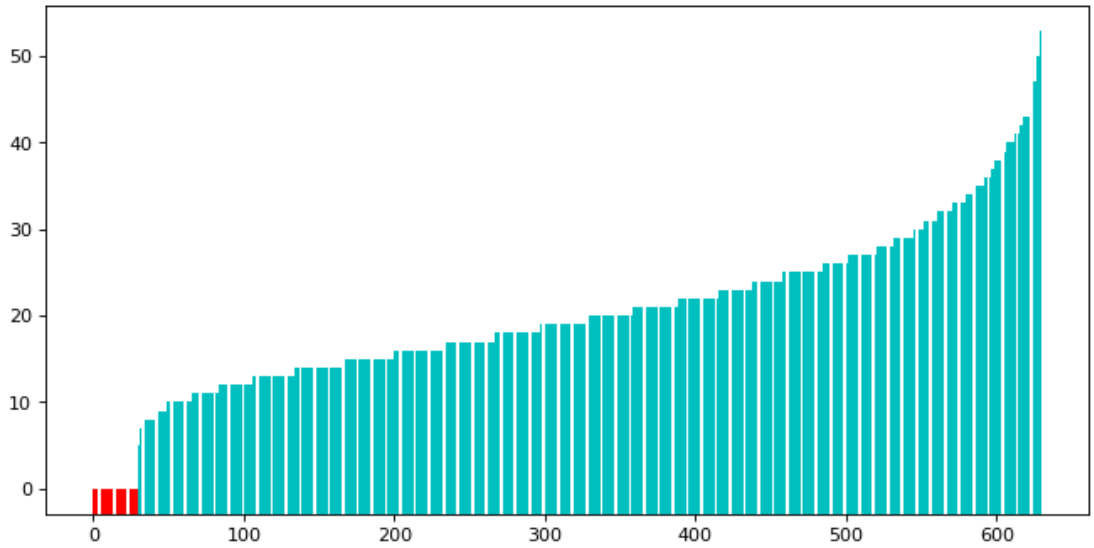


Figure 5.8: Sorted Histogram of the score (number of times a data belongs to the selected median block) of each points in a Logistic Regression MOM algorithm on a toy dataset. Red is an outlier and blue is an informative sample. $K = 120$ and $T = 2000$.

we have

$$R(\hat{f}) - R(f^*) = (P - P_N)\mathcal{L}_{\hat{f}} + P_N\mathcal{L}_{\hat{f}} \leq (P - P_N)\mathcal{L}_{\hat{f}} = \frac{|\mathcal{I}|}{N}(P - P_{\mathcal{I}})\mathcal{L}_{\hat{f}} + \frac{|\mathcal{O}|}{N}(P - P_{\mathcal{O}})\mathcal{L}_{\hat{f}}$$
$$\leq \frac{|\mathcal{I}|}{N}\sup_{f\in\mathcal{F}}(P - P_{\mathcal{I}})\mathcal{L}_f + \frac{2|\mathcal{O}|}{N}$$

because $|\mathcal{L}_{\hat{f}}| \leq 1$ a.s.. Then, by the bounded difference inequality [BLM13, Theorem 6.2], since all $f \in \mathcal{F}$ satisfies $-1 \leq \mathcal{L}_f \leq 1$, one has, for any $x > 0$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}(P - P_{\mathcal{I}})\mathcal{L}_f \geq \mathbb{E}[\sup_{f\in\mathcal{F}}(P - P_{\mathcal{I}})\mathcal{L}_f] + x\right) \leq e^{-2|\mathcal{I}|x^2} \quad .$$

Furthermore, by the symmetrization argument (cf. Chapter 4 in [LT91]),

$$\mathbb{E}[\sup_{f\in\mathcal{F}}(P - P_{\mathcal{I}})\mathcal{L}_f] \leq 2\frac{\mathcal{R}(\mathcal{L}_{\mathcal{F}})}{|\mathcal{I}|} \quad .$$

Therefore, for any $x > 0$, with probability larger than $1 - e^{-2|\mathcal{I}|x^2}$,

$$R(\hat{f}) - R(f^*) \leq \frac{2\mathcal{R}(\mathcal{L}_{\mathcal{F}})}{N} + x + \frac{2|\mathcal{O}|}{N} \quad .$$

The proof is completed by choosing $x = \sqrt{K/(2|\mathcal{I}|)}$.

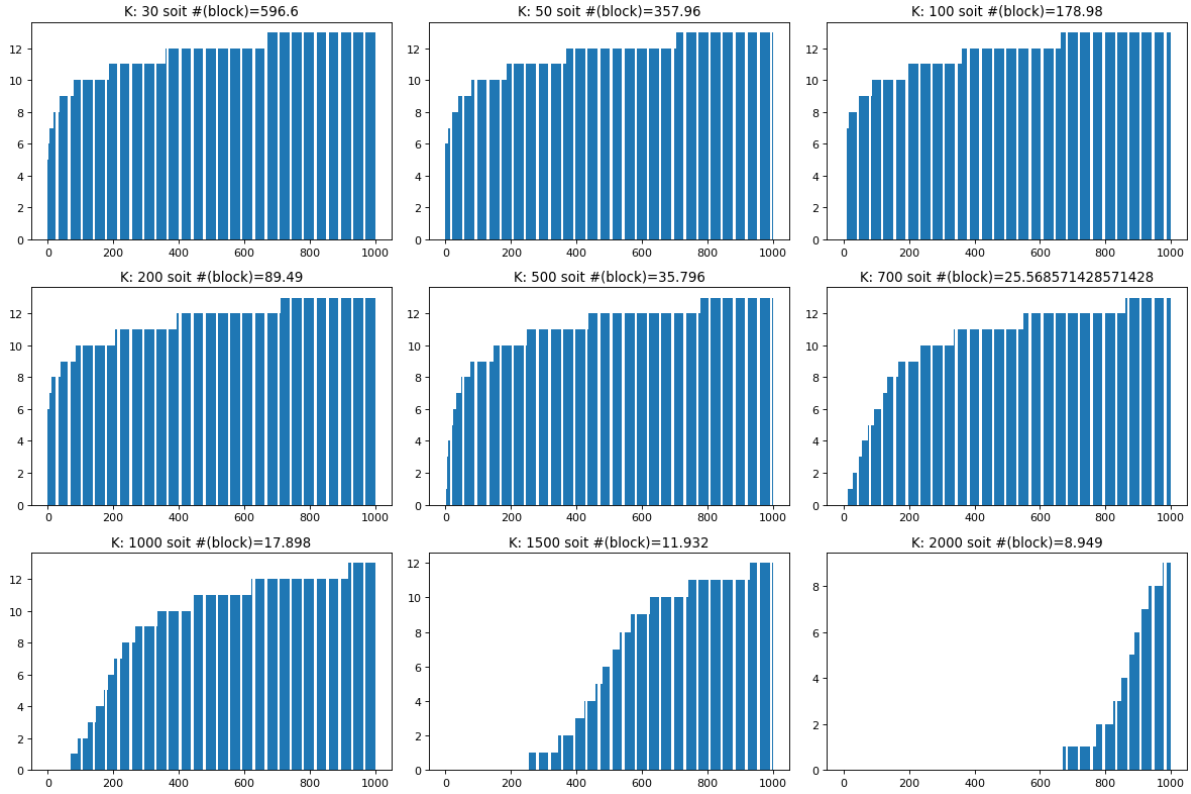


Figure 5.9: Sorted Histogram on the score (number of times a data is selected in a median block) of each points in a Logistic Regression MOM algorithm on the pulsar dataset for various values of K and $T = 20 \times K$ (only the 1000 smaller counts among the 17898 sample of the pulsar dataset are represented).

5.6.2 Proof of Theorem 30

Let $f^* \in \operatorname{argmin}_{f \in F} P\ell_f$. By definition, one has $\mathrm{MOM}_K\big(\ell_{\widehat{f}_{\mathrm{MOM},K}}\big) \leq \mathrm{MOM}_K\big(\ell_{f^*}\big)$, therefore,

$$R(\widehat{f}_{\mathrm{MOM},K}) - R(f^*) \leq P\ell_{\widehat{f}_{\mathrm{MOM},K}} - \mathrm{MOM}_K\big(\ell_{\widehat{f}_{\mathrm{MOM},K}}\big) - \big(P\ell_{f^*} - \mathrm{MOM}_K\big(\ell_{f^*}\big)\big) \ . \qquad (5.6.1)$$

Let us now control the two expressions in the right-hand side of (5.6.1). Let $x > 0$. We have

$$\mathbb{P}\big[P\ell_{f^*} - \mathrm{MOM}_K\big(\ell_{f^*}\big) > x\big] = \mathbb{P}\left[\sum_{k=1}^{K} I(P\ell_{f^*} - P_{B_k}\ell_{f^*} > x) \geq \frac{K}{2}\right] = \sum_{k=K/2}^{K} \binom{K}{k} p^k (1-p)^{K-k} \leq p^{K/2} 2^K$$

where $p = \mathbb{P}[P\ell_{f^*} - P_{B_k}\ell_{f^*} > x]$. Using Markov inequality together with $\mathrm{var}(\ell_{f^*}) \leq 2L^2\mathbb{E}(f^*(X))^2 \leq 2L^2\theta^2$, we obtain

$$\mathbb{P}\big[P\ell_{f^*} - \mathrm{MOM}_K\big(\ell_{f^*}\big) > x\big] \leq \left(\frac{4\mathrm{var}(\ell_{f^*})K}{Nx^2}\right)^{K/2} \leq \left(\frac{8L^2\theta^2 K}{Nx^2}\right)^{K/2} = \exp(-K/2)$$

when $x = 2L\theta\sqrt{2eK/N}$.

Now, for any $x > 0$, one has $\sup_{f\in F} \mathrm{MOM}_K\big(P\ell_f - \ell_f\big) > x$ iff

$$\sup_{f\in F} \sum_{k=1}^{K} I\{(P - P_{B_k})\ell_f > x\} \geq \frac{K}{2} \ . \tag{5.6.2}$$

Let us now control the probability that (5.6.2) holds via an adaptation of the small ball method [KM15, Men15]. Let $x > 0$ and let $\phi(t) = (t-1)I\{1 \leq t \leq 2\} + I\{t \geq 2\}$ be defined for all $t \in \mathbb{R}$. As $\phi(t) \geq I\{t \geq 2\}$, one has

$$\begin{aligned}
&\sup_{f\in F} \sum_{k=1}^{K} I\{(P - P_{B_k})\ell_f > x\} \\
&\leq \sup_{f\in F} \sum_{k\in\mathcal{K}} \mathbb{E}[\phi(2(P - P_{B_k})\ell_f/x)] + |\mathcal{O}| + \sup_{f\in F} \sum_{k\in\mathcal{K}} (\phi(2(P - P_{B_k})\ell_f/x) - \mathbb{E}[\phi(2(P - P_{B_k})\ell_f/x)])
\end{aligned}$$

where we recall that $\mathcal{K} = \{k \in \{1, \cdots, K\} : B_k \cap \mathcal{O} = \emptyset\}\}$.

Since, $\phi(t) \leq I\{t \geq 1\}$ and for all $f \in F$, $\mathrm{Var}(\ell_f) \leq 2L^2\mathbb{E}f(X)^2 \leq 2L^2\theta_2^2$, we have for all $f \in F$ and $k \in \mathcal{K}$,

$$\mathbb{E}[\phi(2(P - P_{B_k})\ell_f/x)] \leq \mathbb{P}\Big((P - P_{B_k})\ell_f \geq \frac{x}{2}\Big) \leq \frac{4\mathrm{Var}(\ell_f)}{x^2|B_k|} \leq \frac{8L^2\theta_2^2 K}{x^2 N} \ .$$

One has therefore

$$\begin{aligned}
&\sup_{f\in F} \sum_{k=1}^{K} I\{(P - P_{B_k})\ell_f > x\} \\
&\qquad \leq K\left(\frac{8L^2\theta_2^2 K}{x^2 N} + \frac{|\mathcal{O}|}{K} + \sup_{f\in F} \frac{1}{K} \sum_{k\in\mathcal{K}} \left(\phi\left(\frac{2(P - P_{B_k})\ell_f}{x}\right) - \mathbb{E}\left[\phi\left(\frac{2(P - P_{B_k})\ell_f}{x}\right)\right]\right)\right) \ .
\end{aligned}$$

As $0 \leq \phi(\cdot) \leq 1$, by the bounded-difference inequality, for any $y > 0$, with probability larger than $1 - e^{-2y^2K}$,

$$\begin{aligned}
&\sup_{f\in F} \frac{1}{K} \sum_{k\in\mathcal{K}} \left(\phi\left(\frac{2(P - P_{B_k})\ell_f}{x}\right) - \mathbb{E}\left[\phi\left(\frac{2(P - P_{B_k})\ell_f}{x}\right)\right]\right) \\
&\qquad\qquad \leq \mathbb{E}\left[\sup_{f\in F} \frac{1}{K} \sum_{k\in\mathcal{K}} \left(\phi\left(\frac{2(P - P_{B_k})\ell_f}{x}\right) - \mathbb{E}\left[\phi\left(\frac{2(P - P_{B_k})\ell_f}{x}\right)\right]\right)\right] + y \ .
\end{aligned}$$

Now, by the symmetrization inequality,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in F} \frac{1}{K} \sum_{k\in\mathcal{K}} \left(\phi\left(\frac{2(P - P_{B_k})\ell_f}{x}\right) - \mathbb{E}\left[\phi\left(\frac{2(P - P_{B_k})\ell_f}{x}\right)\right]\right)\right] \\
&\qquad\qquad \leq 2\mathbb{E}\left[\sup_{f\in F} \frac{1}{K} \sum_{k\in\mathcal{K}} \epsilon_k \phi\left(\frac{2(P - P_{B_k})\ell_f}{x}\right)\right] \ .
\end{aligned}$$

Since ϕ is 1-Lipschitz and $\phi(0) = 0$, by the contraction principle (see [LT91, Chapter 4] or more precisely equation (2.1) in [Kol11]),

$$\mathbb{E}\left[\sup_{f\in F}\frac{1}{K}\sum_{k\in\mathcal{K}}\epsilon_k\phi\left(\frac{(P-P_{B_k})\ell_f}{x}\right)\right] \leq \mathbb{E}\left[\sup_{f\in F}\frac{1}{xK}\sum_{k\in\mathcal{K}}\epsilon_k(P-P_{B_k})\ell_f\right] .$$

By the symmetrization principle,

$$\mathbb{E}\left[\sup_{f\in F}\frac{2}{xK}\sum_{k\in\mathcal{K}}\epsilon_k(P-P_{B_k})\ell_f\right] \leq \frac{2}{xN}\mathbb{E}\left[\sup_{f\in F}\sum_{i\in\mathcal{J}}\epsilon_i\ell_f(X_i,Y_i)\right] .$$

Finally, since ℓ is L-Lipschitz, by the contraction principle (see equation (2.1) in [Kol11]),

$$\mathbb{E}\left[\sup_{f\in F}\sum_{i\in\mathcal{J}}\epsilon_i\ell_f(X_i,Y_i)\right] \leq 2L\mathcal{R}(F) .$$

Thus, for any $y > 0$, with probability larger than $1 - \exp(-2y^2K)$,

$$\sup_{f\in F}\sum_{k=1}^{K} I\{(P-P_{B_k})\ell_f > x\} \leq K\left(\frac{8L^2\theta_2^2K}{x^2N} + \frac{|\mathcal{O}|}{K} + y + \frac{4L\mathcal{R}(F)}{xN}\right) .$$

Let $\Delta = 1/4 - |\mathcal{O}|/K$ and let $y = \Delta$ and $x = 8L\max\left(\theta_2\sqrt{K/N}, 4\mathcal{R}(F)/N\right)$ so

$$\mathbb{P}\left(\sup_{f\in F}\sum_{k=1}^{K} I\{(P-P_{B_k})\ell_f > x\} < \frac{K}{2}\right) \geq 1 - e^{-\Delta^2K/8} .$$

Going back to (5.6.2), this means that

$$\mathbb{P}\left(\sup_{f\in F}\mathrm{MOM}_K\big(\ell_f - P\ell_f\big) \leq 4L\max\left(\theta_2\sqrt{\frac{K}{N}}, \frac{4\mathcal{R}(F)}{N}\right)\right) \geq 1 - \exp(-2\Delta^2K) . \tag{5.6.3}$$

Plugging this result in (5.6.1) concludes the proof of the theorem.

5.6.3 Proof of Proposition 1

We denote by $B(1)(u), \ldots, B(K)(u)$ the blocks such that the corresponding empirical means $P_{B(k)(u)}(f_u(X_1^N)), k = 1, \ldots, K$ are sorted: $P_{B(1)(u)}(f_u(X_1^N)) \geq \cdots \geq P_{B(K)(u)}(f_u(X_1^N))$. Denote $J \in \mathbb{N}$ such that $K = 2J + 1$.

The goal is to show that $u \mapsto \psi_{f_u}((X_i)_{i=1}^N) = MOM_K(f_u((X_i)_{i=1}^N))$ is differentiable and to compute its partical derivatives. To that end, it suffices to show that for all ε with $\|\varepsilon\|_2$ sufficiently small, we have $B(J)(u) = B(J)(u + t\varepsilon)$ for all $t \in [0, 1]$ and for that it is sufficient to check that the same order of the K empirical means is preserved for all $f_{u+t\varepsilon}$:

$$\forall 1 \leq k \leq K-1, \forall t \in [0,1], \quad P_{B(k)(u)}(f_{u+t\varepsilon}) - P_{B(k+1)(u)}(f_{u+t\varepsilon}) > 0. \tag{5.6.4}$$

We decompose this difference in three parts,

$$\begin{aligned} P_{B(k)(u)}(f_{u+t\varepsilon}) - P_{B(k+1)(u)}(f_{u+t\varepsilon}) \geq & P_{B(k)(u)}(f_u) - P_{B(k+1)(u)}(f_u) \\ & - \left|P_{B(k)(u)}(f_u) - P_{B(k)(u)}(f_{(u+t\varepsilon)}\right| \\ & - \left|P_{B(k+1)(u)}(f_{u+t\varepsilon}) - P_{B(k+1)(u)}(f_u)\right| \end{aligned}$$

The two last terms are controlled by the Lipshitz property of $u \mapsto f_u$,

$$\forall t \in [0,1], \quad P_{B(k)(u)}(f_{u+t\varepsilon}) - P_{B(k+1)(u)}(f_{u+t\varepsilon}) \geq P_{B(k)(u)}(f_u) - P_{B(k+1)(u)}(f_u) - 2tL\|\varepsilon\|_2.$$

We denote by

$$h_k(\|\varepsilon\|_2) = \mathbb{P}\big(\forall t \in [0,1], \quad P_{B(k)(u)}(f_u) - P_{B(k+1)(u)}(f_u) - 2tL\|\varepsilon\|_2 \geq 0\big)$$

for all $1 \leq k \leq K-1$, h_k is an non-decreasing function. Because for all $1 \leq k \leq K$, $P_{B(k)(u)}(f_u)$ has a uniformly continuous law with respect to the Lebesgue measure (because its density is a convolution of several copies of the density of $f_u(X)$), there is no jump in the c.d.f and then h_k verifies that

$$h_k(\|\varepsilon\|_2) \xrightarrow[\|\varepsilon\|_2 \to 0]{} 1.$$

And again because for all $1 \leq k \leq K$, $P_{B(k)(u)}(f_u)$ has a uniformly continuous law with respect to the Lebesgue measure, we also have that

$$h_k(\|\varepsilon\|_2) = \mathbb{P}\big(\forall t \in [0,1], \quad P_{B(k)(u)}(f_u) - P_{B(k+1)(u)}(f_u) - 2tL\|\varepsilon\|_2 > 0\big).$$

Then, taking the union bound for $1 \leq k \leq K-1$,

$$\begin{aligned} h(\|\varepsilon\|_2) &:= \mathbb{P}\big(\forall 1 \leq k \leq K-1, \forall t \in [0,1], \quad P_{B(k)(u)}(f_u) - P_{B(k+1)(u)}(f_u) - 2tL\|\varepsilon\|_2 > 0\big) \\ &\geq 1 - \sum_{k=1}^{K-1}(1 - h_k(\|\varepsilon\|_2)). \end{aligned}$$

Moreover, h can be rewritten as a probability that the blocks don't change using the reasoning leading to equation (5.6.4), hence

$$h(\|\varepsilon\|_2) = \mathbb{P}(\forall 1 \leq k \leq K-1, \quad B(k)(u) = B(k)(u+t\varepsilon)) \leq \mathbb{P}(\forall t \in [0,1], \quad B(J)(u) = B(J)(u+t\varepsilon)).$$

We now compute the partial derivatives of the median of means ψ_{f_u}. Let $e_1, \ldots, e_p \in \mathbb{R}^p$ be the canonical basis of $\mathbb{R}^p$. For all $m \in \mathbb{N}$, we define $\varepsilon_m^j = \delta_m e_j$ with $(\delta_m)_m$ a decreasing sequence of $\mathbb{R}_+^*$ such that for all $1 \leq k \leq K-1$ we have $h_k(\delta_m) \geq 1 - 2^{-m}$, δ_m exists because $h_k(\delta) \to 1$ when $\delta \to 0$. Then,

$$h(\|\varepsilon_m^j\|_2) \geq 1 - K2^{-m}. \tag{5.6.5}$$

We denote by A_m^j the event $A_m^j := \left\{\forall t \in [0,1], \quad B_{(J)}(u) = B_{(J)}(u + t\varepsilon_{m'}^j)\right\}$ and we study the limiting event $\Omega^j = \overline{\lim}_{m\to\infty} A_m^j$.

First, let us note that for all $1 \leq j \leq p$, the sequence of set $(A_m^j)_n$ is non-increasing, hence

$$\Omega^j = \overline{\lim}_{m\to\infty} A_m^j = \underline{\lim}_{m\to\infty} A_m^j = (\overline{\lim}_{m\to\infty}(A_m^j)^c)^c,$$

then, for all $1 \leq j \leq d$, we can study the $\overline{\lim}_{m\to\infty}(A_m^j)^c$ with Borel-Cantelli Lemma. Indeed, we have from equation (5.6.5), $\mathbb{P}((A_m^j)^c) \leq K2^{-m}$. Hence, the series $\sum_m \mathbb{P}((A_m^j)^c)$ converges and by Borel Cantelli Lemma, $\mathbb{P}(\overline{\lim}_{m\to\infty}(A_m^j)^c) = 0$, then for all $1 \leq i \leq p$, $\mathbb{P}(\Omega^j) = 1$. In other words, we have that for all $\omega \in \Omega^j$, there exists $m \geq 1$ such that $\omega \in A_m^j$. Hence, there exists $m \geq 1$ such that for all $t \in [0,1]$, $\quad B(J)(u) = B(J)(u + t\varepsilon_m^j)$, which implies that for all $1 \leq j \leq p$,

$$\begin{aligned}\partial_j \psi_{f_u}(X) &= \lim_{t\to 0} \frac{\psi_{f_{u+t\varepsilon_m^j}}(X) - \psi_{f_u}(X)}{t} = \lim_{t\to 0} \frac{P_{B(J)(u)}(f_{u+t\varepsilon_m^j}) - P_{B(J)(u)}(f_u)}{t} \\ &= \frac{1}{N/K} \lim_{t\to 0} \sum_{i\in B(J)(u)} \frac{f_{u+t\varepsilon_m^j}(X_i) - f_u(X_i)}{t} = \frac{1}{N/K} \sum_{i\in B(J)(u)} \partial_j f_u(X_i).\end{aligned}$$

5.7 Annex

5.7.1 Choice of the number of blocks

Let us study the behaviour of our algorithms when the number of blocks changes. We plot the accuracy as a function of K averaged on 50 runs to have a good idea of the evolution of the performance with respect to K, the result is represented in figure 5.10.

There is a clear separation around $2|\mathcal{O}| = 60$ that is consistent with the theory. On the other hand the accuracy doesn't decrease when K gets bigger one would expect. This may be due to the symmetry of the dataset. If we run the same experiment on the real dataset, we get a much more regular plot, see Figure 5.11.

Figure 5.11 confirms our predictions on clean datasets, the accuracy getting better as K gets smaller (the MOM minimizer is the ERM estimator when $K = 1$ and ERM is optimal in the i.i.d. setup, [LM13]). This may be due to the small number of outliers in this dataset.

5.7.2 Illustration of convergence rate

In this section, we estimate the rate of convergence of the MOM risk minimization algorithm Logistic Regression on two databases (see figure 5.12). The first dataset is composed of points located on two interlaced half-circle with a Gaussian noise of standard deviation 0.3, the two "moons" are each of a different class. We assume that these moons don't satisfy the margin property (we checked that the rate was slow for ERM algorithms, using the vanilla logistic regression). The second dataset is composed of two Gaussians $\mathcal{N}((-1,-1), 1.4^2 I_2)$ and $\mathcal{N}((1,1), 1.4^2 I_2)$ with respective label 1 and 0, we can prove that this dataset verifies the margin property needed to obtain fast rate in ERM

There are no outliers in the datasets because we only want to test the rate of convergence. To illustrate the rates of convergence of our algorithms, we plot the curve $\log\left(\left|\hat{R}^{0-1}(\hat{f}_K)) - \hat{R}^{0-1}(f^*)\right|\right)$ as a function of $\log(n)$ where the risk is estimated by Monte-Carlo. The figure obtained for Logistic Regression MOM is represented in figure 5.13. It seems that MOM minimizers can achieve fast rates of convergence even if we did not prove them.

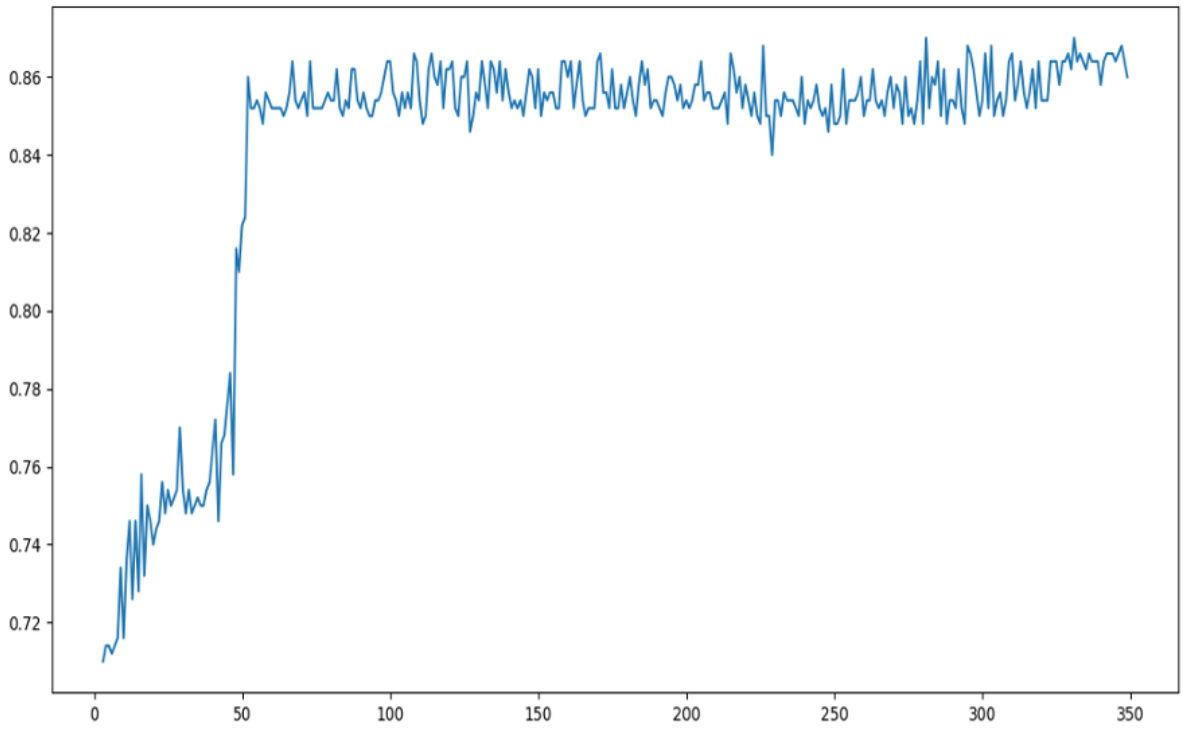


Figure 5.10: Plot of the accuracy on the toy dataset of Logistic Regression MOM as a function of K.

Remark 6. *We used random blocks sampled at each iteration for this application because it is the algorithm that we described earlier but even if we use one partition of blocks for the whole algorithm (as in the theory we developed) we obtain nonetheless fast rate for the Gaussians dataset.*

5.7.3 Comparison with robust algorithms based on M-estimators.

In this section we compare the algorithm Logistic Regression MOM with two other algorithms based on M-estimators, these algorithms are studied on the toy dataset presented in Section 5.5.

One algorithm is a gradient on the Huber estimation of the loss function, it follows the same reasoning as MOM risk minimization and minimizes $\mathbb{E}[\ell_f(X,Y)]$ using as a proxy the Huber estimator for this quantity. The Huber estimator is then defined as a M-estimator, denoted here $\hat{\mu}_f$, solution of

$$\sum_{i=1}^{n} \psi_c(\hat{\mu}_f - \ell(f(X_i), Y_i)) = 0$$

where $\psi_c = \max(-c, \min(c, x))$ is the Huber function, $c > 0$. Using this definition of $\hat{\mu}_f$, it is then easy to compute the gradient $\nabla\hat{\mu}_f$ and then use a gradient descent algorithm. The theory behind this algorithm is studied further in [BJL15].

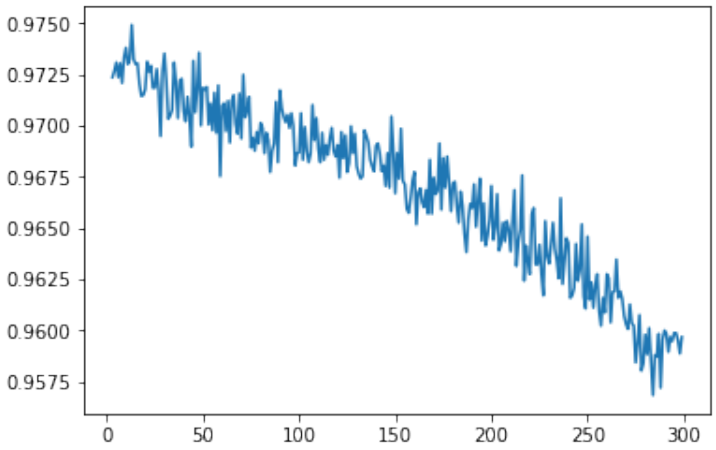
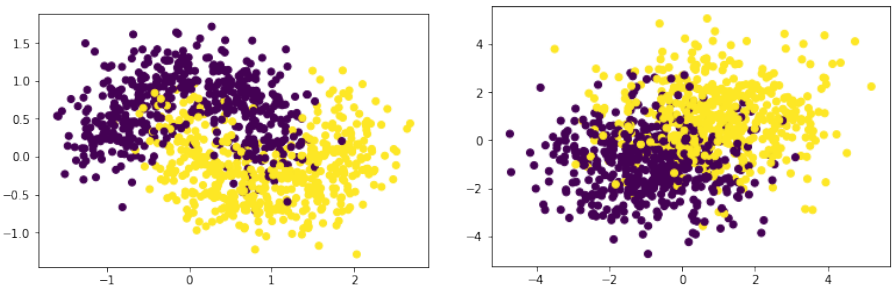


Figure 5.11: Plot of the accuracy on HTRU2 dataset of Logistic Regression MOM as a function of K.

The second algorithm uses a “redescending" loss function, in short we do ERM with a bounded loss function. Here we use Tukey biweight loss function rescaled by MADN scale estimator and IRLS algorithm to optimize the empirical risk.

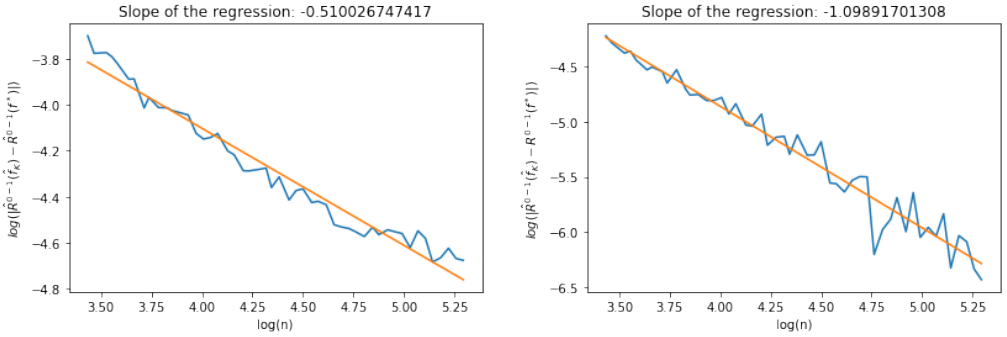
Figure 7.1 shows that all algorithms perform similarly on this easy, low dimensional dataset. The situation is quite different in higher dimension. In Figure 5.15 we used a 200 dimensional dataset and the algorithm using a redescending loss function does not perform well. This may be due to local minima in which the algorithm gets stuck, as local minima are multiplied when the dimension gets higher. The other algorithms don’t suffer this drawback since they use a “projection by the loss function" that makes the problem one dimensional.

The algorithm using redescending loss functions is a simple gradient descent that has linear complexity. The Huber gradient algorithm estimates at each iteration a Huber estimator of location. The complexity of this estimator depends on the algorithm used but for most M-estimators a commonly used algorithm is an iteratively reweighted algorithm whose complexity is linear in the sample size. In practice we can nonetheless notice a great complexity of the Huber estimator in some cases where data are not well spread. In most cases, Logistic Regression MOM is the fastest among these three algorithms and the gradient Huber is the slowest, even though logistic regression may need a lot more iterations than the other algorithms.



(a) Scatter plot of the Moons dataset (b) Scatter plot of the Gaussians dataset

Figure 5.12: Scatter plot of the two dataset used in this section, the color represent the class of the points.



(a) Convergence rate for the Moons dataset (b) Convergence rate for the Gaussians dataset

Figure 5.13: Plot of the logarithm of the excess risk as a function of $\log(n)$ in two cases: (a) where the margin assumption does not hold and (b) where the margin assumption holds. A linear regression is fitted on the curve, its slope is printed at the top of each figure revealing a slow $n^{-0.51}$ rate of convergence in case (a) and a fast $n^{-1.1}$ in case (b).

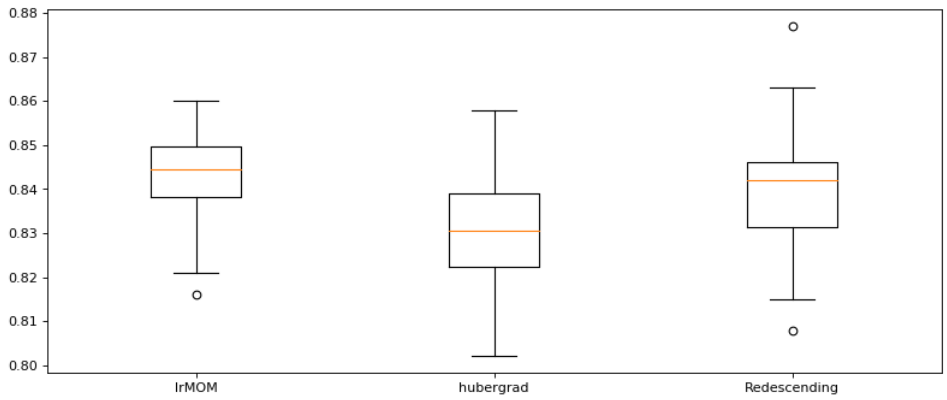


Figure 5.14: Boxplot of the accuracy obtained on 50 training/test run (1000 training sample, 2% corruption) of each algorithms on a 2-dimensional toy dataset.

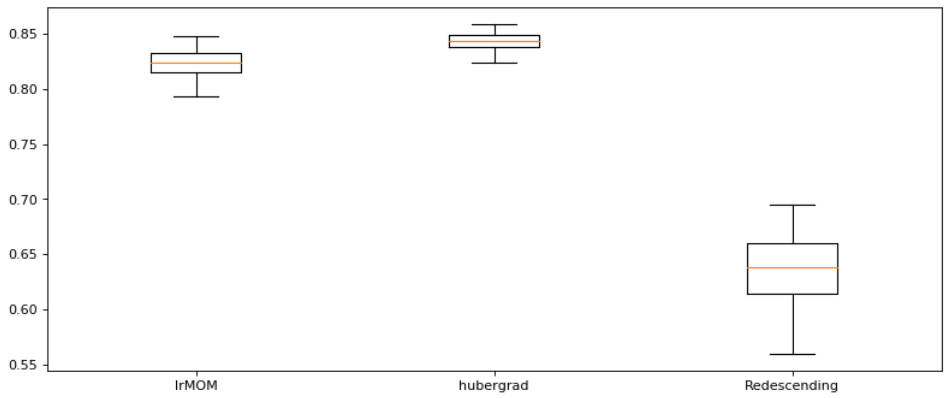


Figure 5.15: Boxplot of the accuracy obtained on 50 training/test run (2000 training sample, 2% corruption) of each algorithms on a 200-dimensional toy dataset.

Chapter 6

Excess risk bounds in robust empirical risk minimization

Abstract

This paper investigates robust versions of the general empirical risk minimization algorithm, one of the core techniques underlying modern statistical methods. Success of the empirical risk minimization is based on the fact that for a "well-behaved" stochastic process $\{f(X),\ f \in \mathcal{F}\}$ indexed by a class of functions $f \in \mathcal{F}$, averages $\frac{1}{N}\sum_{j=1}^{N} f(X_j)$ evaluated over a sample $X_1, \dots, X_N$ of i.i.d. copies of X provide good approximation to the expectations $\mathbb{E}f(X)$, uniformly over large classes $f \in \mathcal{F}$. However, this might no longer be true if the marginal distributions of the process are heavy-tailed or if the sample contains outliers. We propose a version of empirical risk minimization based on the idea of replacing sample averages by robust proxies of the expectations, and obtain high-confidence bounds for the excess risk of resulting estimators. In particular, we show that the excess risk of robust estimators can converge to 0 at fast rates with respect to the sample size N, referring to the rates faster than $N^{-1/2}$. We discuss implications of the main results to the linear and logistic regression problems, and evaluate the numerical performance of proposed methods on simulated and real data. Keywords: robust estimation, excess risk, median-of-means, regression, classification

6.1 Introduction

This work is devoted to robust algorithms in the framework of statistical learning. A recent Forbes article [Ron19] states that "Machine learning algorithms are very dependent on accurate, clean, and well-labeled training data to learn from so that they can produce accurate results" and "According to a recent report from AI research and advisory firm Cognilytica, over 80% of the time spent in AI projects are spent dealing with and wrangling data." While some abnormal elements of the sample, or outliers, can be detected and filtered during the preprocessing steps, others are more difficult to detect: for instance, a sophisticated adversary might try to "poison" data to force a desired outcome [MDC14]. Other seemingly abnormal observations could be inherent to the underlying data-generating process. An "ideal" learning method should not discard informative samples, while limiting the effect of individual observation on the output of the learning algorithm

at the same time. We are interested in robust methods that are model-free, and require minimal assumptions on the underlying distribution. We study two types of robustness: robustness to heavy tails expressed in terms of the moment requirements, as well as robustness to (a variant of) adversarial contamination. Heavy tails can be used to model variation and randomness naturally occurring in the sample, while adversarial contamination is a convenient way to model outliers of unknown nature.

The statistical framework used throughout the paper is defined as follows. Let $(S, \mathcal{S})$ be a measurable space, and let $X \in S$ be a random variable with distribution P. Suppose that $X_1, \dots, X_N$ are i.i.d. copies of X. Moreover, assume that $\mathcal{F}$ is a class of measurable functions from S to $\mathbb{R}$ and $\ell : \mathbb{R} \to \mathbb{R}_+$, where $\mathbb{R}_+$ is a set of non-negative integers, is a loss function. Many problems in statistical learning theory can be formulated as risk minimization of the form

$$\mathbb{E}\,\ell(f(X)) \to \min_{f \in \mathcal{F}}.$$

We will frequently write $P\ell(f)$ or simply $\mathcal{L}(f)$ in place of the expected loss $\mathbb{E}\ell(f(X))$. Throughout the paper, we will also assume that the minimum above is attained for some (unique) $f_* \in \mathcal{F}$ (however, f_* does not necessarily coincide with the global minimizer of $\mathcal{L}(f)$ over all measurable functions that might not belong to $\mathcal{F}$). For example, in the context of regression, $X = (Z, Y) \in \mathbb{R}^d \times \mathbb{R}$, $f(Z, Y) = Y - g(Z)$ for some g in a class $\mathcal{G}$ (such as the class of linear functions), $\ell(x) = x^2$, and $f_*(z, y) = y - g_*(z)$, where $g_* = \operatorname{argmin}_{g \in \mathcal{G}} \mathbb{E}(Y - g(Z))^2$. As the true distribution P is usually unknown, a proxy of f_* is obtained via *empirical risk minimization* (ERM), namely

$$\tilde{f}_N := \operatorname*{argmin}_{f \in \mathcal{F}} \mathcal{L}_N(f), \tag{6.1.1}$$

where P_N is the empirical distribution based on the sample $X_1, \dots, X_N$ and

$$\mathcal{L}_N(f) := P_N\,\ell(f) = \frac{1}{N}\sum_{j=1}^{N} \ell(f(X_j)).$$

Performance of any $f \in \mathcal{F}$ (in particular, $\tilde{f}_N$) is measured via the excess risk $\mathcal{E}(f) := P\ell(f) - P\ell(f_*)$. The excess risk of $\tilde{f}_N$ is a random variable defined as

$$\mathcal{E}(\tilde{f}_N) := P\,\ell\big(\tilde{f}_N\big) - P\ell(f_*) = \mathbb{E}\Big[\ell\big(\tilde{f}(X)\big)|X_1, \dots, X_N\Big] - \mathbb{E}\ell(f_*(X)).$$

General bounds for the excess risk have been extensively studied; a small subsample of the relevant works includes the papers [VvdG00, vdVW96, Kol11, AB09a, BBM05, Tsy04] and references therein. However, until recently sharp estimates were known only in the situation when the functions in the class $\ell(\mathcal{F}) := \{\ell(f),\ f \in \mathcal{F}\}$ are uniformly bounded, or when the envelope $F_\ell(x) := \sup_{f \in \mathcal{F}} |\ell(f(x))|$ of the class $\ell(\mathcal{F})$ possesses finite exponential moments. Our focus is on the situation when marginal distributions of the process $\{\ell(f(X)),\ f \in \mathcal{F}\}$ indexed by $\mathcal{F}$ are allowed to be heavy-tailed, meaning that they possess finite moments of low order only (in this paper, "low order" usually means between 2 to 4). In such cases, the tail probabilities of the random variables $\Big\{\frac{1}{\sqrt{N}}\sum_{j=1}^{N}(\ell(f(X_j)) - \mathbb{E}\ell(f(X))),\ f \in \mathcal{F}\Big\}$ decay polynomially, thus rendering many existing techniques ineffective. Moreover, we consider a challenging framework of *adversarial contamination* where the initial dataset of cardinality N is merged with a set of $\mathcal{O} < N$ outliers which are generated by an adversary who has an opportunity to inspect the data, and the combined dataset of cardinality $N^\circ = N + \mathcal{O}$ is presented to an algorithm; in this paper, we assume that the proportion of contamination $\frac{\mathcal{O}}{N}$ (or its upper bound) is known.

The approach that we propose is based on replacing the sample mean at the core of ERM by a more "robust" estimator of $\mathbb{E}\,\ell(f(X))$ that exhibits tight concentration under minimal moment assumptions. Well known examples of such estimators include the median-of-means estimator [NY83, AMS99, LO11] and Catoni's estimator [Cat12]. Both the median-of-means and Catoni's estimators gain robustness at the cost of being biased. The ways that the bias of these estimators is controlled is based on different principles however. Informally speaking, Catoni's estimator relies on delicate "truncation" of the data, while the median-of-means (MOM) estimator exploits the fact that the median and the mean of a symmetric distribution both coincide with its center of symmetry. In this paper, we will use "hybrid" estimators that take advantage of both symmetry and truncation. This family of estimators has been introduced and studied in [MS17, Min18], and we review the construction below.

6.1.1 Organization of the paper.

The main ideas behind the proposed estimators are explained in Section 6.1.3, followed by the high-level overview of the main theoretical results and comparison to existing literature in Section 6.1.4. The complete statements of the key results are given in Section 6.2, and in Section 6.3 we deduce the corollaries of these results for specific examples. Finally, the main ideas and key inequalities necessary for the proofs is explained in Section 6.4. The remaining technical arguments are contained in the supplementary material. Finally, in Section 6.7 of the supplement we discuss practical implementation and numerical performance of our methods on synthetic and real data.

6.1.2 Notation.

For two sequences $\{a_j\}_{j\geq 1} \subset \mathbb{R}$ and $\{b_j\}_{j\geq 1} \subset \mathbb{R}$ for $j \in \mathbb{N}$, the expression $a_j \lesssim b_j$ means that there exists a constant $c > 0$ such that $a_j \leq cb_j$ for all $j \in \mathbb{N}$; $a_j \asymp b_j$ means that $a_j \lesssim b_j$ and $b_j \lesssim a_j$. Absolute constants will be denoted c, c_1, C, C', etc, and may take different values in different parts of the paper. For a function $h : \mathbb{R}^d \mapsto \mathbb{R}$, we define

$$\operatorname*{argmin}_{y\in\mathbb{R}^d} h(y) = \{y \in \mathbb{R}^d : h(y) \leq h(x) \text{ for all } x \in \mathbb{R}^d\},$$

and $\|h\|_\infty := \operatorname{ess\,sup}\{|h(y)| : y \in \mathbb{R}^d\}$. Moreover, $L(h)$ will stand for a Lipschitz constant of h. For $f \in \mathcal{F}$, let $\sigma^2(\ell, f) = \operatorname{Var}(\ell(f(X)))$ and for any subset $\mathcal{F}' \subseteq \mathcal{F}$, denote $\sigma^2(\ell, \mathcal{F}') = \sup_{f\in\mathcal{F}'} \sigma^2(\ell, f)$. Additional notation and auxiliary results are introduced on demand.

6.1.3 Robust mean estimators.

Let $k \leq N$ be an integer, and assume that $G_1, \ldots, G_k$ are disjoint subsets of the index set $\{1, \ldots, N\}$ of cardinality $|G_j| = n \geq \lfloor N/k \rfloor$ each. Given $f \in \mathcal{F}$, let

$$\overline{\mathcal{L}}_j(f) := \frac{1}{n} \sum_{i\in G_j} \ell(f(X_i))$$

be the empirical mean evaluated over the subsample indexed by G_j. Given a convex, even function $\rho : \mathbb{R} \mapsto \mathbb{R}_+$ and $\Delta > 0$, set

$$\widehat{\mathcal{L}}^{(k)}(f) := \operatorname*{argmin}_{y\in\mathbb{R}} \sum_{j=1}^{k} \rho\left(\sqrt{n}\,\frac{\overline{\mathcal{L}}_j(f) - y}{\Delta}\right). \tag{6.1.2}$$

Clearly, if $\rho(x) = x^2$, $\widehat{\mathcal{L}}^{(k)}(f)$ is equal to the sample mean. If $\rho(x) = |x|$, then $\widehat{\mathcal{L}}^{(k)}(f)$ is the median-of-means estimator [NY83, AMS99, DLLO16]. We will be interested in the situation when ρ is smooth and "shaped" like Huber's loss, in particular, that ρ' is bounded and Lipchitz continuous (exact conditions imposed on ρ are specified in Assumption 1 below). Note that (6.1.2) defines a whole family of estimators for different values of k and n. It is instructive to consider two cases: first, when $k = N$ (so that $n = 1$) and the scaling factor $\Delta \asymp \sqrt{\mathrm{Var}(\ell(f(X)))}\sqrt{N}$, $\widehat{\mathcal{L}}^{(k)}(f)$ is akin to Catoni's estimator [Cat12], and when n is large (e.g. $\sqrt{N} \ll n \ll N$ and $\Delta \asymp \sqrt{\mathrm{Var}(\ell(f(X)))}$), $\widehat{\mathcal{L}}^{(k)}(f)$ is the "median-of-means type" estimator. Let us elaborate on these two cases further. Generally speaking, the estimator $\widehat{\mathcal{L}}^{(k)}(f)$ is biased, and, as we already mentioned in the introduction, one way to understand the difference between Catoni's and median-of-means type estimators is via the difference in mechanisms used to control the bias. In the case of Catoni's estimator, this mechanism is based on truncating each observation at the level of order $\sqrt{N}$ [1] encoded in the choice of $\Delta \asymp \sqrt{\mathrm{Var}(\ell(f(X)))}\sqrt{N}$, while in the case of the median-of-means estimator it relies on the approximate symmetry, implied by the Central Limit Theorem, of the distribution of the empirical averages $\overline{\mathcal{L}}_j(f)$, and in particular the fact that any reasonable estimator of location for this distribution will be close to the mean $\mathcal{L}(f)$ when $|G_j|$ is large. In Section 6.2.1, we formally introduce the key quantities that allow us to control the bias under various moment assumptions on the underlying classes.

We also construct a permutation-invariant version of the estimator $\widehat{\mathcal{L}}^{(k)}(f)$ that does not depend on the specific choice of the subgroups $G_1, \dots, G_k$. We conjecture that this estimator is more efficient than $\widehat{\mathcal{L}}^{(k)}(f)$; see remark 6.1.3 below for more details. Next, let

$$\mathcal{A}_N^{(n)} := \{J : \ J \subseteq \{1, \dots, N\}, \ |J| = n\}.$$

Let h be a measurable, permutation-invariant function of n variables. Recall that a U-statistic of order n with kernel h based on an i.i.d. sample $X_1, \dots, X_N$ is defined as [Hoe48]

$$U_{N,n} = \frac{1}{\binom{N}{n}} \sum_{J\in\mathcal{A}_N^{(n)}} h(\{X_j\}_{j\in J}). \tag{6.1.3}$$

Given $J \in \mathcal{A}_N^{(n)}$, let $\overline{\mathcal{L}}(f; J) := \frac{1}{n}\sum_{i\in J} f(X_i)$. Consider U-statistics of the form

$$U_{N,n}(z; f) = \sum_{J\in\mathcal{A}_N^{(n)}} \rho\left(\sqrt{n}\,\frac{\overline{\mathcal{L}}(f; J) - z}{\Delta}\right).$$

Then the permutation-invariant version of $\widehat{\mathcal{L}}^{(k)}(f)$ is defined as

$$\widehat{\mathcal{L}}_U^{(k)}(f) := \operatorname*{argmin}_{z\in\mathbb{R}} U_{N,n}(z; f).$$

[1] Reference to truncation can be made explicit by setting $\rho(x) = \min\big(x^2/2, |x| - 1/2\big)$ to be Huber's loss and considering the gradient descent iteration for the optimization problem (6.1.2).

Finally, assuming that $\widehat{\mathcal{L}}^{(k)}(f)$ provides good approximation of the expected loss $\mathcal{L}(f)$ of each individual $f \in \mathcal{F}$, it is natural to consider

$$\widehat{f}_N := \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{\mathcal{L}}^{(k)}(f), \tag{6.1.4}$$

as well as its permutation-invariant analogue

$$\widehat{f}_N^U := \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{\mathcal{L}}_U^{(k)}(f) \tag{6.1.5}$$

as an alternative to standard empirical risk minimization (6.1.1). The main goal of this paper is to obtain general bounds for the excess risk of the estimators $\widehat{f}_N$ and $\widehat{f}_N^U$ under minimal assumptions on the stochastic process $\{\ell(f(X)),\ f \in \mathcal{F}\}$. More specifically, we are interested in scenarios when the excess risk converges to 0 at fast, or "optimistic" rates, referring to the rates faster than $N^{-1/2}$. Rate of order $N^{-1/2}$ ("slow rates") are easier to establish: in particular, results of this type follow from bounds on the uniform deviations $\sup_{f\in\mathcal{F}}\left|\widehat{\mathcal{L}}^{(k)}(f) - \mathcal{L}(f)\right|$ that have been investigated in [Min18]. Proving fast rates is a more technically challenging task: to achieve the goal, we develop Bahadur-type representations [Bah66] of the estimators $\widehat{\mathcal{L}}^{(k)}(f)$ and $\widehat{\mathcal{L}}_U^{(k)}(f)$ that provide linear, in $\ell(f)$, approximations of these nonlinear statistics that are easier to study, and carefully analyze the remainder terms. Introduction of such representations in the framework of median-of-means estimation is one of the main technical novelties of the paper; the tools we develop could prove useful in other related problems, such as study of the asymptotic distributions of the robust estimators $\widehat{f}_N$ and $\widehat{f}_N^U$.

Remark. *The main reason we introduce the permutation-invariant estimator $\widehat{f}_N^U$ is our conjecture that it has superior, compared to $\widehat{f}_N$, performance. We were able to confirm this fact numerically in our experiments; however, complete theoretical confirmation is not yet available, and requires new technical tools beyond those developed in the present work. Specifically, we conjecture that $\widehat{f}_N^U$ is more* efficient *than $\widehat{f}_N$: when $\mathcal{F}$ is finite dimensional, this means, informally, that the asymptotic distribution of $\sqrt{N}(\widehat{f}_N^U - f_*)$ has smaller variance than the asymptotic distribution of $\sqrt{N}(\widehat{f}_N - f_*)$. In other words, the conjectured difference in performance is about the constant factors rather than the rates. Such improvements are too subtle to be captured by the non-asymptotic bounds for the excess risk that are being pursued in this work, nevertheless they are clearly noticeable in the simulations.*

It should also be acknowledged that exact evaluation of the U-statistics-based estimators $\widehat{\mathcal{L}}_U^{(k)}(f)$ and $\widehat{f}_N^U$ is not feasible due to the number of summands $\binom{N}{n}$ being very large even for small values of n. However, exact computation is typically not required, and throughout our detailed simulation studies, gradient descent methods proved to be very efficient for the problem (6.1.5) in scenarios like least-squares and logistic regression. These points, as well as comparison of the numerical performance of the estimators $\widehat{f}_N^U$ and $\widehat{f}_N$, are further discussed in Section 6.7 of the supplementary material.

6.1.4 Overview of the main results and comparison to existing bounds.

Our main contribution is the proof of high-confidence bounds for the excess risk of the estimators $\widehat{f}_N$ and $\widehat{f}_N^U$. First, we show (see Theorem 33 and (6.2.4)) that the excess risk is bounded from

above by the quantity of order $N^{-1/2}$ (referred to as "slow rates") with exponentially high probability if

$$\sigma^2(\ell, \mathcal{F}) = \sup_{f\in\mathcal{F}} \sigma^2(\ell, f) < \infty \text{ and } \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{\sqrt{N}} \sum_{j=1}^{N} (\ell(f(X_j)) - \mathbb{E}\ell(f(X))) < \infty.$$

The latter is true if the class $\{\ell(f),\ f \in \mathcal{F}\}$ is P-Donsker [Dud14], in other words, if the empirical process $f \mapsto \frac{1}{\sqrt{N}} \sum_{j=1}^{N} (\ell(f(X_j)) - \mathbb{E}\ell(f(X)))$ converges weakly to a Gaussian limit. This result is analogous to its counterpart in the standard empirical risk minimization framework. Moreover, it is known [Men08, LM10] that in general, the $N^{-1/2}$ rate for the excess risk of the empirical risk minimizers can not be improved. Our main contribution is the proof of the fact that that under additional assumption requiring that any $f \in \mathcal{F}$ with small excess risk is itself close to f_* (that minimizes the expected loss), $\widehat{f}_N$ and $\widehat{f}_N^U$ attain *fast* rates. This fact is well-known in the usual empirical risk minimization framework [BM06, Kol11] but is new for the type of robust estimators considered here. We state the bounds below only for $\widehat{f}_N$ while the results for the U-statistics based $\widehat{f}_N^U$ are similar, up to the change in absolute constants. In order to avoid excessive technical details at this stage, we will first illustrate our general results by stating corollaries for the popular frameworks of logistic regression and regression with quadratic loss, while the most general versions of the theorems and additional examples will be stated afterwards.

Binary classification and logistic regression. Assume that $(Z, Y) \in S \times \{\pm 1\}$ is a random couple where Z is an instance and Y is a binary label, and let $g_*(z) := \mathbb{E}[Y|Z = z]$ be the regression function. It is well-known that the binary classifier $b_*(z) := \text{sign}(g_*(z))$ achieves smallest possible misclassification error defined as $P(Y \neq g(Z))$. Let $\mathcal{F}$ be a given convex class of functions mapping S to $\mathbb{R}$, $\ell : \mathbb{R} \mapsto \mathbb{R}_+$ a convex, nondecreasing, Lipschitz loss function, and let

$$h_* = \underset{\text{all measurable } f}{\text{argmin}} \ \mathbb{E}\ell(Yf(Z)).$$

The loss ℓ is classification-calibrated if $\text{sign}(h_*(z)) = b_*(z)$ P-almost surely; we refer the reader to [BJM06] for a detailed exposition. In the case of logistic regression considered below, $S = \mathbb{R}^d$,

$$\ell(y, f(z)) = \ell(yf(z)) := \log\left(1 + e^{-yf(z)}\right)$$

is the classification-calibrated loss and $\mathcal{F} = \left\{f_\beta(\cdot) = \langle \cdot, v\rangle,\ v \in \mathbb{R}^d,\ \|v\|_2 \leq R\right\}$. Note that results stated below hold without assuming that $h_* \in \mathcal{F}$.

Regression with quadratic loss. Let $(Z, Y) \in S \times \mathbb{R}$ be a random couple satisfying $Y = f_*(Z) + \eta$ where the noise variable η is independent of Z and $f_*(z) = \mathbb{E}[Y|Z = z]$ is the regression function. Linear regression with quadratic loss corresponds to $S = \mathbb{R}^d$,

$$\ell(y, f(z)) = \ell(y - f(z)) := (y - f(z))^2$$

and $\mathcal{F} = \left\{f_\beta(\cdot) = \langle \cdot, v\rangle,\ \ v \in \mathbb{R}^d, \|v\|_2 \leq R\right\}$. In this case, we will assume that $f_* \in \mathcal{F}$; it is possible to avoid this assumption at the cost of additional technicalities and taking advantage of the deep results of S. Mendelson [Men16] on the multiplier inequalities.

In the statements below, we will assume that we are given an i.i.d. sample $(Z_1, Y_1), \ldots, (Z_N, Y_N)$ having the same distribution as (Z, Y) where the marginal distribution of Z is supported on a compact set. Moreover, suppose that $\mathbb{E}|\eta|^8 < \infty$ in the case of regression with quadratic loss; Section 6.3 contains other examples covering a wider class of distributions and classes $\mathcal{F}$.

Theorem 31 (Informal)**.** *Assume the framework of either logistic regression or linear regression with quadratic loss. Then, for appropriately chosen k and Δ,*

$$\mathcal{E}(\widehat{f}_N) \leq C(R,P,\rho)\left(\frac{d}{N} + \frac{s}{N^{3/4}} + \left(\frac{\mathcal{O}}{N}\right)^{3/4}\right)$$

with probability at least $1-e^{-s}$ for all $s \lesssim k$.

Moreover, we construct a two-step estimator $\widehat{f}''_N$ based on $\widehat{f}_N$ that is capable of achieving further improved rates.

Theorem 32 (Informal)**.** *Assume the framework of either logistic regression or linear regression with quadratic loss. There exists an estimator $\widehat{f}''_N$, defined later in the paper, such that*

$$\mathcal{E}\left(\widehat{f}''_N\right) \leq C(R,P,\rho)\left(\frac{d}{N} + \frac{s}{N} + \frac{\mathcal{O}}{N}\right)$$

with probability at least $1-e^{-s}$ for all $1 \leq s \leq s_{\max}$ where $s_{\max} := s_{\max}(N) \to \infty$ as $N \to \infty$.

The estimator $\widehat{f}''_N$ mentioned in Theorem 32 is based on a two-step procedure, where $\widehat{f}_N$ serves as an initial approximation that is refined on the second step via risk minimization restricted to a "small neighborhood" of $\widehat{f}_N$. All of the bounds in this paper have the form $\mathcal{E}(\widehat{f}_N) \leq \overline{\delta} + C(\mathcal{F},P)\left(\frac{s}{N^\gamma} + \left(\frac{\mathcal{O}}{N}\right)^\gamma\right)$, where $\frac{1}{2} \leq \gamma \leq 1$ and $\overline{\delta}$ is the quantity (formally defined in (6.2.5)) that often coincides, up to log-factors, with the optimal rate for the excess risk [ACL19, LM19c] – for instance, $\overline{\delta} \asymp \frac{d}{N}$ in the examples above. In the standard empirical risk minimization, the excess risk bounds in the linear and logistic regression admit the bounds of order $\frac{d}{N} + \frac{s}{N}$, albeit under more restrictive assumptions and in the corruption-free framework. Therefore, the bound of Theorem 31 is suboptimal in these cases due to the "remainder terms" being of order $N^{-3/4}$, and the improvement achieved by the two-step estimator $\widehat{f}''_N$, as described in Theorem 32, becomes important.

Next, we provide a brief overview of the literature on the topic and compare our results to the state of the art. Robustness of statistical learning algorithms has been studied extensively in recent years. Existing research has mainly focused on addressing robustness to heavy tails as well as adversarial contamination. One line of work investigated robust versions of the gradient descent method for the optimization problem (6.1.1) based on variants of the multivariate median-of-means technique [PSBR20, CSX17, YCKB18, AAZL18], as well as Catoni's estimator [HI17a]. The line works initiated in the theoretical computer science community [LRV16, DKK+19a, DKK+17, also see the survey paper [DK19]] tackled the problem of optimal mean estimation in the adversarial contamination framework by establishing deep connections between the mean and covariance estimation problems that culminated in the family of powerful filtering algorithms; these algorithms can also be used as subroutines in robust gradient descent-type methods [DKK+19b, CHK+20]. While these algorithms admit strong theoretical guarantees, they require robustly estimating the gradient vector at every step (with the exception of [DKK+19b] that offers a more efficient approach) hence are computationally demanding; moreover, results are weaker for losses that are not strongly convex (for instance, the hinge loss). The line of research that is closest in spirit to the approach of this paper includes the works that employ robust risk estimators based on Catoni's idea [AC11, BJL15, HI17b] and the median-of-means technique, such as "tournaments" and the

"min-max median-of-means" [LM19c, LM19b, LL20, LLM20, CLL19b], also see [CHK+20, Hop20] for the computationally efficient algorithms related to the tournament-type procedures. As it was mentioned in the introduction, the core of our methods can be viewed as a "hybrid" between Catoni's and the median-of-means estimators. We provide a more detailed comparison to the results of the aforementioned papers:

1. We show that risk minimization based on a version of Catoni's estimator is capable of achieving fast rates, thus improving the results and weakening the assumptions stated in [BJL15] that only allowed the slow rates to be established;
2. We develop new tools and techniques to analyze proposed estimators. In particular, we do not rely on the "small ball" method [KM15, Men15] and the standard "majority vote-based" analysis [LL20, LM19c] of the median-of-means estimators. Instead, we provide accurate bounds for the bias and investigate the remainder terms for the Bahadur-type linear approximations of the estimators defined in (6.1.2). In particular, we demonstrate that the order of typical deviations of the estimator $\widehat{\mathcal{L}}^{(k)}(f)$ around $\mathcal{L}(f)$ are significantly smaller than the deviations of the subsample averages $\overline{\mathcal{L}}_j(f)$, which is not easy to do using the majority vote-based proof techniques; consequently, this fact allows us to "decouple" the confidence parameter s that controls the deviation probabilities from parameters k and $\mathcal{O}$ responsible for the number of subsamples and the degree of contamination respectively. Unlike the tournaments-based estimators, in some regimes our algorithms admit a "universal" choice of k that is independent of the parameter $\overline{\delta}$ controlling the optimal rate. In the previous works, parameter k was often overloaded as it controlled the deviation probabilities while depending on $\overline{\delta}$ (or a closely related quantity) at the same time. Finally, our techniques allow us to establish bounds that are uniform over a certain range of confidence parameter s while the previously existing deviation results were only available for $s \asymp k$.
3. We are able to simultaneously treat the case of Lipschitz as well as non-Lipschitz (e.g., quadratic) loss functions ℓ. At the same time, in some situations (e.g. linear regression with quadratic loss), the required assumptions are stronger compared to the best results in the literature tailored specifically to the task, e.g. [LL20, LM19c] that treat the case of regression with quadratic loss.
4. Existing approaches based on the median-of-means estimators are either computationally intractable [LM19c], or outputs of practically efficient algorithms do not admit strong theoretical guarantees [LL20, LLM20, CLL19b]. We design numerical algorithms specifically for the estimators $\widehat{f}_N$ and $\widehat{f}_N^U$ defined via (6.1.4) and (6.1.5), and show that they enjoy good performance in numerical experiments as well as strong theoretical guarantees.

6.2 Theoretical guarantees for the excess risk.

In this section, we give complete statements of the main results and explain the high-level ideas behind their proofs.

6.2.1 Preliminaries.

We start by introducing the main quantities that appear in our results, and state the key assumptions. Recall that $\sigma^2(\ell, \mathcal{F}')$ stands for $\sup_{f\in\mathcal{F}'} \sigma^2(\ell, f)$, where $\mathcal{F}' \subseteq \mathcal{F}$. The loss functions

ρ that will be of interest to us satisfy the following assumption.

Assumption 1. *Suppose that the function $\rho : \mathbb{R} \mapsto \mathbb{R}$ is convex, even, 5 times continuously differentiable and such that*

(i) $\rho'(z) = z$ for $|z| \leq 1$ and $\rho'(z) = \text{const}$ for $z \geq 2$,
(ii) $z - \rho'(z)$ is nondecreasing.

An example of a function ρ satisfying required assumptions is given by "smoothed" Huber's loss defined as follows. Let

$$H(y) = \frac{y^2}{2} I\{|y| \leq 3/2\} + \frac{3}{2}\left(|y| - \frac{3}{4}\right) I\{|y| > 3/2\}$$

be the usual Huber's loss. Moreover, let ϕ be the "bump function" $\phi(x) = C \exp\left(-\frac{4}{1-4x^2}\right) I\left\{|x| \leq \frac{1}{2}\right\}$ where C is chosen so that $\int_{\mathbb{R}} \phi(x)dx = 1$. Then ρ given by the convolution $\rho(x) = (h * \phi)(x)$ satisfies Assumption 1.

Remark. *(a) The requirements that ρ is 5 times continuously differentiable is of the technical nature and is likely not necessary. It appears due to the fact that we need to control higher order terms in the Bahadur-Kiefer type representations of the estimator $\widehat{\mathcal{L}}^{(k)}(f)$, as well as rely on the Lindeberg replacement-type arguments in our proofs.*

(b) The derivative ρ' has a natural interpretation of being a smooth version of the truncation function. Moreover, observe that $\rho'(2) - 2 \leq \rho'(1) - 1 = 0$ by (ii), hence $\|\rho'\|_\infty \leq 2$. It is also easy to see that for any $x > y$, $\rho'(x) - \rho'(y) = y - \rho'(y) - (x - \rho'(x)) + x - y \leq x - y$, hence ρ' is Lipschitz continuous with Lipschitz constant $L(\rho') = 1$.

In section 6.1.3, we have briefly discussed the bias of robust mean estimators and various ways that it can be controlled. Now we will introduce the key quantities necessary to make the bounds precise. Everywhere below, $\Phi(\cdot)$ stands for the cumulative distribution function of the standard normal random variable and $W(f)$ denotes a random variable with distribution $N\big(0, \sigma^2(f)\big)$. For $f \in \mathcal{F}$ such that $\sigma(f) > 0$, $n \in \mathbb{N}$ and $t > 0$, define

$$\mathcal{M}_f(t, n) := \left| \Pr\left(\frac{\sum_{j=1}^n (f(X_j) - Pf)}{\sigma(f)\sqrt{n}} \leq t \right) - \Phi(t) \right|,$$

where $Pf := \mathbb{E}f(X)$. In other words, $\mathcal{M}_f(t, n)$ controls the rate of convergence in the central limit theorem. It follows from the results of L. Chen and Q.-M. Shao (Theorem 2.2 in in [CS01]) that

$$\mathcal{M}_f(t, n) \leq g_f(t, n) := C\left(\frac{\mathbb{E}(f(X) - \mathbb{E}f(X))^2 \, I\left\{ \frac{|f(X) - \mathbb{E}f(X)|}{\sigma(f)\sqrt{n}} > 1 + \left|\frac{t}{\sigma(f)}\right| \right\}}{\sigma^2(f)\left(1 + \left|\frac{t}{\sigma(f)}\right|\right)^2} \right.$$
$$\left. + \frac{1}{\sqrt{n}} \frac{\mathbb{E}|f(X) - \mathbb{E}f(X)|^3 \, I\left\{ \frac{|f(X) - \mathbb{E}f(X)|}{\sigma(f)\sqrt{n}} \leq 1 + \left|\frac{t}{\sigma(f)}\right| \right\}}{\sigma^3(f)\left(1 + \left|\frac{t}{\sigma(f)}\right|\right)^3} \right)$$

given that the absolute constant C is large enough. Note that, crucially, the control of the rate in terms of $g_f(t, n)$ is non-uniform, since $g_f(t, n)$ is a decreasing function of t. Moreover, let

$$G_f(n, \Delta) := \int_0^\infty g_f\left(\Delta\left(\frac{1}{2} + t\right), n\right) dt.$$

The quantity $\frac{G_f(n,\Delta)}{\sqrt{n}}$ plays the key role in controlling the bias of the estimator $\widehat{\mathcal{L}}^{(k)}(f)$: it decreases both as Δ get large and as the subsample size n increases, referring to different bias-controlling mechanisms of Catoni's and the median-of-means type estimators, see discussion after (6.1.2). The following statement provides simple upper bounds for $g_f(t,n)$ and $G_f(n,\Delta)$ that depend on the tail properties of $f(X)$; its proof can be found in [Min18, Section 4.4].

Lemma 31. *Let $X_1,\dots,X_n$ be i.i.d. copies of X, and assume that $Var(f(X)) < \infty$. Then $g_f(t,n) \to 0$ as $|t| \to \infty$ and $g_f(t,n) \to 0$ as $n \to \infty$, with convergence being monotone. Moreover, if $\mathbb{E}|f(X) - \mathbb{E}f(X)|^{2+\delta} < \infty$ for some $\delta \in [0,1]$, then for all $t > 0$*

$$g_f(t,n) \le C' \frac{\mathbb{E}\big|f(X) - \mathbb{E}f(X)\big|^{2+\delta}}{n^{\delta/2}(\sigma(f) + |t|)^{2+\delta}} \le C' \frac{\mathbb{E}\big|f(X) - \mathbb{E}f(X)\big|^{2+\delta}}{n^{\delta/2}|t|^{2+\delta}}, \tag{6.2.1}$$

$$G_f(n,\Delta) \le C'' \frac{\mathbb{E}\big|f(X) - \mathbb{E}f(X)\big|^{2+\delta}}{\Delta^{2+\delta} n^{\delta/2}},$$

where $C', C'' > 0$ are absolute constants.

We can rewrite the bound for $\sup_{f\in\mathcal{F}} G_f(n,\Delta)$ as $\sup_{f\in\mathcal{F}} G_f(n,\Delta) \le C'' \frac{\sup_{f\in\mathcal{F}} \mathbb{E}(|f(X)-\mathbb{E}f(X)|/\sigma(\ell,\mathcal{F}))^{2+\delta}}{(\Delta/\sigma(\ell,\mathcal{F}))^{2+\delta} n^{\delta/2}}$, where the numerator $\sup_{f\in\mathcal{F}} \mathbb{E}(|f(X) - \mathbb{E}f(X)|/\sigma(\ell,\mathcal{F}))^{2+\delta}$ is the quantity akin the kurtosis while the ratio $M_\Delta := \frac{\Delta}{\sigma(\ell,\mathcal{F})}$ appearing in the denominator can be interpreted as a truncation level expressed in the "units" of $\sigma(\ell,\mathcal{F})$. This "truncation level," along with the subgroup size n, are the two main quantities controlling the bias of the estimators $\widehat{\mathcal{L}}^{(k)}(f),\ f \in \mathcal{F}$.

6.2.2 Slow rates for the excess risk.

Let

$$\widehat{\delta}_N := \mathcal{E}(\widehat{f}_N) = \mathcal{L}\big(\widehat{f}_N\big) - \mathcal{L}(f_*),$$
$$\widehat{\delta}^U_N := \mathcal{E}(\widehat{f}^U_N) = \mathcal{L}\big(\widehat{f}^U_N\big) - \mathcal{L}(f_*)$$

be the excess risk of $\widehat{f}_N$ and its permutation-invariant analogue $\widehat{f}^U_N$ which are the main objects of our interest. The following bound for the excess risk is well known in the empirical risk minimization literature [Kol11], and it easily leads to control of the excess risk in terms of the uniform deviations of robust mean estimators.

$$\begin{aligned}
\mathcal{E}\big(\widehat{f}_N\big) &= \mathcal{L}\big(\widehat{f}_N\big) - \mathcal{L}(f_*) \\
&= \mathcal{L}\big(\widehat{f}_N\big) + \widehat{\mathcal{L}}^{(k)}(\widehat{f}_N) - \widehat{\mathcal{L}}^{(k)}(\widehat{f}_N) + \widehat{\mathcal{L}}^{(k)}(f_*) - \widehat{\mathcal{L}}^{(k)}(f_*) - \mathcal{L}(f_*) \\
&= \Big(\mathcal{L}\big(\widehat{f}_N\big) - \widehat{\mathcal{L}}^{(k)}(\widehat{f}_N)\Big) - \Big(\mathcal{L}(f_*) - \widehat{\mathcal{L}}^{(k)}(f_*)\Big) + \underbrace{\widehat{\mathcal{L}}^{(k)}(\widehat{f}_N) - \widehat{\mathcal{L}}^{(k)}(f_*)}_{\le 0} \\
&\le 2 \sup_{f\in\mathcal{F}} \Big|\widehat{\mathcal{L}}^{(k)}(f) - \mathcal{L}(f)\Big|.
\end{aligned} \tag{6.2.2}$$

The first result, Theorem 33 below, together with the inequality (6.2.2) immediately implies the "slow rate bound" (meaning rate not faster than $N^{-1/2}$) for the excess risk. This result has been previously established in [Min18]. Define

$$\widetilde{\Delta} := \max(\Delta, \sigma(\ell,\mathcal{F})).$$

Theorem 33. *There exist absolute constants $c, C > 0$ such that for all $s > 0$, n and k satisfying*

$$\frac{1}{\Delta}\left(\frac{1}{\sqrt{k}}\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{\sqrt{N}}\sum_{j=1}^{N}(\ell(f(X_j))-P\ell(f))+\sigma(\ell,\mathcal{F})\sqrt{\frac{s}{k}}\right)+\sup_{f\in\mathcal{F}}G_f(n,\Delta)+\frac{s}{k}+\frac{\mathcal{O}}{k}\le c, \quad (6.2.3)$$

the following inequality holds with probability at least $1-2e^{-s}$:

$$\sup_{f\in\mathcal{F}}\left|\widehat{\mathcal{L}}^{(k)}(f)-\mathcal{L}(f)\right| \le C\left[\frac{\widetilde{\Delta}}{\Delta}\left(\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{j=1}^{N}(\ell(f(X_j))-P\ell(f))+\sigma(\ell,\mathcal{F})\sqrt{\frac{s}{N}}\right)\right.$$
$$\left.+\widetilde{\Delta}\left(\sqrt{n}\frac{s}{N}+\frac{\sup_{f\in\mathcal{F}}G_f(n,\Delta)}{\sqrt{n}}+\frac{\mathcal{O}}{k\sqrt{n}}\right)\right].$$

Moreover, same bounds hold for the permutation-invariant estimators $\widehat{\mathcal{L}}_U^{(k)}(f)$, up to the change in absolute constants.

An immediate corollary is the bound for the excess risk

$$\mathcal{E}(\widehat{f}_N) \le C\left[\frac{\widetilde{\Delta}}{\Delta}\left(\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{j=1}^{N}(\ell(f(X_j))-P\ell(f))+\sigma(\ell,\mathcal{F})\sqrt{\frac{s}{N}}\right)\right.$$
$$\left.+\widetilde{\Delta}\sqrt{n}\left(\frac{s}{N}+\frac{\sup_{f\in\mathcal{F}}G_f(n,\Delta)}{n}+\frac{\mathcal{O}}{N}\right)\right] \quad (6.2.4)$$

that holds under the assumptions of Theorem 33 with probability at least $1-2e^{-s}$. When the class $\{\ell(f),\ f\in\mathcal{F}\}$ is P-Donsker [Dud14], $\limsup_{N\to\infty}\left|\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{\sqrt{N}}\sum_{j=1}^{N}(\ell(f(X_j))-P\ell(f))\right|$ is bounded, hence condition (6.2.3) holds for N large enough whenever s is not too big and Δ and k are not too small, namely, $s\le c'k$ and $\Delta\sqrt{k}\ge c''\sigma(\mathcal{F})$. The bound of Theorem 33 also suggests that the natural "unit" to measure the magnitude of the parameter Δ is $\sigma(\ell,\mathcal{F})$.

To put these results in perspective, let us consider two examples. First, assume that $n=1$, $k=N$ and set $\Delta=\Delta(s):=\sigma(\mathcal{F})\sqrt{\frac{N}{s}}$ for $s\le c'N$. Using Lemma 31 with $\delta=0$ to estimate $G_f(n,\Delta)$, we deduce that

$$\mathcal{E}(\widehat{f}_N) \le C\left[\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{j=1}^{N}(\ell(f(X_j))-P\ell(f))+\sigma(\ell,\mathcal{F})\left(\sqrt{\frac{s}{N}}+\frac{\mathcal{O}}{\sqrt{N}}\right)\right]$$

with probability at least $1-2e^{-s}$. This inequality improves upon excess risk bounds obtained for Catoni-type estimators in [BJL15], as it does not require functions in $\mathcal{F}$ to be uniformly bounded.

The second case we consider is when $N\gg n\ge 2$. For the choice of $\Delta\asymp\sigma(\ell,\mathcal{F})$, the estimator $\widehat{\mathcal{L}}^{(k)}(f)$ most closely resembles the median-of-means estimator, as we have explained in Section 6.1.3. In this case, Theorem 33 yields the excess risk bound of the form

$$\mathcal{E}(\widehat{f}_N) \le C\left[\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{j=1}^{N}(\ell(f(X_j))-P\ell(f))+\sigma(\ell,\mathcal{F})\left(\sqrt{\frac{s}{N}}+\sqrt{\frac{k}{N}}\sup_{f\in\mathcal{F}}G_f(n,\sigma(\mathcal{F}))+\frac{\mathcal{O}}{k}\sqrt{\frac{k}{N}}\right)\right]$$

that holds with probability $\geq 1-2e^{-s}$ for all $s \leq c'k$. As $\sup_{f\in\mathcal{F}} G_f(n,\Delta)$ is small for large n and $\frac{\mathcal{O}}{k}\sqrt{\frac{k}{N}} \leq \sqrt{\frac{\mathcal{O}}{N}}$ whenever $\mathcal{O} \leq k$, this bound improves upon Theorem 2 in [LLM20] that provides bounds for the excess risk for robust classifiers based on the the median-of-means estimators.

6.2.3 Towards fast rates for the excess risk.

It is well known that in regression and binary classification problems, excess risk often converges to 0 at a rate faster than $N^{-1/2}$, and could be as fast as N^{-1}. Such rates are often referred to as "fast" or "optimistic" rates. In particular, this is the case when there exists a "link" between the excess risk and the variance of the loss class, namely, if for some convex nondecreasing and nonnegative function ϕ such that $\phi(0) = 0$,

$$\mathcal{E}(f) = P\ell(f) - P\ell(f_*) \geq \phi\Big(\sqrt{\mathrm{Var}(\ell(f(X)) - \ell(f_*(X)))}\Big).$$

It is thus natural to ask if fast rates can be attained by estimators produced by the robust algorithms proposed above. Results presented in this section give an affirmative answer to this question. Let us introduce the main quantities that commonly appear in the excess risk bounds [Kol11, LM19c]. For $\delta > 0$, let

$$\begin{aligned}
\mathcal{F}(\delta) &:= \{\ell(f) : \ f \in \mathcal{F},\ \mathcal{E}(f) \leq \delta\},\\
\nu(\delta) &:= \sup_{\ell(f)\in\mathcal{F}(\delta)} \sqrt{\mathrm{Var}(\ell(f(X)) - \ell(f_*(X)))},\\
\omega(\delta) &:= \mathbb{E}\sup_{\ell(f)\in\mathcal{F}(\delta)} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \Big((\ell(f) - \ell(f_*))(X_j) - P(\ell(f) - \ell(f_*)) \Big) \right|.
\end{aligned}$$

Moreover, define

$$\mathfrak{B}(\ell,\mathcal{F}) := \frac{\sup_{f\in\mathcal{F}} \mathbb{E}^{1/4}(\ell(f(X)) - \mathbb{E}\ell(f(X)))^4}{\sigma(\ell,\mathcal{F})}.$$

The following condition, known as *Bernstein's condition* following [BM06], plays the crucial role in the analysis of excess risk bounds.

Assumption 2. *There exist constants $D > 0$, $\delta_B > 0$ such that*

$$Var(\ell(f(X)) - \ell(f_*(X))) \leq D^2\,\mathcal{E}(f)$$

whenever $\mathcal{E}(f) \leq \delta_B$.

Informally speaking, Assumption 2 postulates that any $f \in \mathcal{F}$ (more precisely, the loss $\ell(f)$ induced by it) with small excess risk is itself close to f_*. If this is true, it turns out that one can avoid global bounds for on the expected supremum of the empirical process used to obtain "slow" rates, and instead rely on the modulus of continuity $\omega(\delta)$ of the empirical process locally in the neighborhood of $\ell(f_*)$ in order to get better upper bounds on the excess risk. The basics of this approach in the classical empirical risk minimization frameworks are clearly explained in [Kol11, Chapter 1.2], and we rely on similar ideas below.

Assumption 2 is known to hold in many concrete cases of prediction and classification tasks, and we provide examples and references in Section 6.3 below. More general versions of the Bernstein's condition are often considered in the literature: for instance, it can be replaced by assumption requiring that $\mathrm{Var}(\ell(f(X)) - \ell(f_*(X))) \le D^2 \left(\mathcal{E}(f)\right)^\tau$ for some $\tau \in (0,1]$, as was done in [BM06]; clearly, our assumption corresponds to $\tau = 1$. Results of this paper admit straightforward extensions to the slightly less restrictive scenario when $\tau < 1$; we omit the details to reduce the level of technical burden on the statements of our results.

Following [Kol11, Chapter 4], we will say that the function $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is of concave type if it is nondecreasing and $x \mapsto \frac{\psi(x)}{x}$ is decreasing. Moreover, if for some $\gamma \in (0,1)$ $x \mapsto \frac{\psi(x)}{x^\gamma}$ is decreasing, we will say that ψ is of strictly concave type with exponent γ. We will assume that $\omega(\delta)$ admits an upper bound $\widetilde{\omega}(\delta)$ of strictly concave type (with some exponent γ), and that $\nu(\delta)$ admits an upper bound $\widetilde{\nu}(\delta)$ of concave type. For instance, when Assumption 2 holds, $\nu(\delta) \le D\sqrt{\delta}$ for $\delta \le \delta_B$, implying that $\widetilde{\nu}(\delta) = D\sqrt{\delta}$ is an upper bound for $\nu(\delta)$ of strictly concave type with $\gamma = \frac{1}{2}$. [2] Moreover, the function $\omega(\delta)$ often admits an upper bound of the form $\widetilde{\omega}(\delta) = R_1 + \sqrt{\delta} R_2$ where R_1 and R_2 do not depend on δ; such an upper bound is also of concave type. Next, set

$$\overline{\delta} := \min\left\{\delta > 0 : \ C_1(\rho)\frac{1}{\sqrt{N}}\frac{\widetilde{\Delta}}{\Delta}\frac{\widetilde{\omega}(\delta)}{\delta} \le \frac{1}{7}\right\}, \tag{6.2.5}$$

where $C_1(\rho)$ is a sufficiently large positive constant that depends only on ρ. The quantity $\overline{\delta}$ often coincides with the optimal rates for the excess risk in the classical empirical risk minimization framework: for example, it is of order $\frac{d}{N}$ up to logarithmic factors in linear regression with quadratic loss and in logistic regression when Bernstein's condition is satisfied; in general, the order of $\overline{\delta}$ ranges between the pessimistic $N^{-1/2}$ in "hard" problems and "optimistic" N^{-1} where the rates between correspond to weaker versions of Assumptions 2, for instance, see [BJM06]. Theorems below provide estimates for the excess risk of robust risk minimizers under various conditions on the tails of the random variables $\{f(X),\ f \in \mathcal{F}\}$. All these bounds have the same structure that includes the term $\overline{\delta}$ as well as the "remainder terms" that account for the bias of the robust risk estimators $\widehat{\mathcal{L}}^{(k)}(f)$ as well as the outlier contamination proportion $\frac{\mathcal{O}}{N}$; naturally, stricter moment conditions result in better remainder terms.

Theorem 34. *Assume that conditions of Theorem 33 hold. Additionally, suppose that* $M_\Delta := \frac{\Delta}{\sigma(\ell,\mathcal{F})} \ge 1$. *Then*

$$\widehat{\delta}_N \le \overline{\delta} + C(\rho)\left(D^2\left(\frac{1}{M_\Delta^2 n} + \frac{s+\mathcal{O}}{N}\right) + \sigma(\ell,\mathcal{F})\sqrt{n}M_\Delta\left(\frac{1}{M_\Delta^4 n} + \frac{s+\mathcal{O}}{N}\right)\right).$$

with probability at least $1 - 10e^{-s}$, *where the constant* $C(\rho)$ *depends on* ρ *only and* D *is a constant appearing in Assumption 2.*

Under stronger moment assumptions, the excess risk bound can be strengthened and take the following form.

Theorem 35. *Assume that conditions of Theorem 33 hold. Additionally, suppose that*

$$\sup_{f\in\mathcal{F}} \mathbb{E}^{1/4}(\ell(f(X)) - \mathbb{E}\ell(f(X)))^4 < \infty$$

[2] This is only true in some neighborhood of 0, but is sufficient for our purposes.

and that $M_\Delta := \frac{\Delta}{\sigma(\ell,\mathcal{F})} \geq 1$. *Then*

$$\widehat{\delta}_N \leq \overline{\delta} + C(\rho)\big(D^2 + \sigma(\ell,\mathcal{F})\sqrt{n}M_\Delta\big)\left(\frac{\mathfrak{B}^6(\ell,\mathcal{F})}{M_\Delta^4 n^2} + \frac{s+\mathcal{O}}{N}\right).$$

with probability at least $1-10e^{-s}$, *where the constant* $C(\rho)$ *depends on* ρ *only and* D *is a constant appearing in Assumption 2.*

The main ideas behind the proofs of Theorems 34 and 35 are explained in the beginning of Section 6.4.

Remark.

1. *The bounds of Theorems 34 and 35 hold for the excess risk* $\widehat{\delta}_N^U$ *of the permutation-invariant estimator* $\widehat{f}_N^U$, *up to a change in absolute constants.*

2. *It is evident that whenever* $\mathcal{O}=0$, *the best possible rates implied by Theorem 34 are of order* $N^{-2/3}$ *(indeed, this is the case whenever* $M_\Delta\sqrt{n} \asymp N^{1/3}$ *and* $\overline{\delta} \lesssim N^{-2/3}$*), while the best possible rates attained by Theorem 35 are of order* $N^{-3/4}$ *(when* $M_\Delta\sqrt{n} \asymp N^{1/4}$ *and* $\overline{\delta} \lesssim N^{-3/4}$*); in particular, in this case the choice of* M_Δ *and* n *is independent of* $\overline{\delta}$. *In general, if* $\mathcal{O}=\epsilon N$ *for* $\epsilon>0$, *the best rates implied by Theorems 34 and 35 are* $\overline{\delta}+C(\mathcal{F},\rho,P)\epsilon^{2/3}$ *and* $\overline{\delta}+C(\mathcal{F},\rho,P)\epsilon^{3/4}$ *respectively.*

3. *Assumption requiring that* $M_\Delta \geq 1$ *is introduced for convenience: without it, extra powers of the ratio* $\frac{\max(\Delta,\sigma(\ell,\mathcal{F}))}{\Delta}$ *appear in the bounds.*

Our next goal is to describe an estimator that is capable of achieving excess risk rates up to N^{-1}. The approach that we follow is similar in spirit to the "minmax" estimators studied in [AC11, LO11, LL20], among others, as well as the "median-of-means tournaments" introduced in [LM19c]; all these methods focus on estimating the differences $\mathcal{L}(f_1)-\mathcal{L}(f_2)$ for all $f_1,f_2\in\mathcal{F}$. Recall that $f_* = \operatorname{argmin}_{f\in\mathcal{F}} P\ell(f)$, and observe that for any fixed $f'\in\mathcal{F}$, f_* can be equivalently defined via

$$f_* = \operatorname*{argmin}_{f\in\mathcal{F}} P(\ell(f)-\ell(f')).$$

A version of the robust empirical risk minimizer (6.1.4) corresponding to this problem can be defined as

$$\widehat{\mathcal{L}}^{(k)}(f-f') := \operatorname*{argmin}_{y\in\mathbb{R}} \frac{1}{\sqrt{N}}\sum_{j=1}^k \rho\left(\sqrt{n}\,\frac{\big(\overline{\mathcal{L}}_j(f)-\overline{\mathcal{L}}_j(f')\big)-y}{\Delta}\right)$$

for appropriately chose $\Delta>0$, and

$$\widehat{f}'_N := \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{\mathcal{L}}^{(k)}(f-f').$$

Moreover, if $f'\in\mathcal{F}$ is a priori known to be "close" to f_*, then it suffices to search for the minimizer in a neighborhood $\mathcal{F}'$ of f' that contains f_* instead of all $f\in\mathcal{F}$:

$$\widehat{f}''_N := \operatorname*{argmin}_{f\in\mathcal{F}'} \widehat{\mathcal{L}}^{(k)}(f-f').$$

The advantage gained by this procedure is expressed by the fact that $\sup_{f\in\mathcal{F}'}\operatorname{Var}(\ell(f(X))-\ell(f'(X)))$ can be much smaller than $\sigma(\ell,\mathcal{F})$.

We will now formalize this argument and provide performance guarantees; we use the framework of Theorem 35 which leads to the bounds that are easier to state and interpret. However, similar reasoning applies to the setting of Theorem 34 as well. The presented algorithms also admit straightforward permutation-invariant modifications that we omit. Let

$$\widehat{\mathcal{E}}_N(f) := \widehat{\mathcal{L}}^{(k)}(f) - \widehat{\mathcal{L}}^{(k)}(\widehat{f}_N)$$

be the "empirical excess risk" of f. Indeed, this is a meaningful notion as $\widehat{f}_N$ is the minimizer of $\widehat{\mathcal{L}}^{(k)}(f)$ over $f \in \mathcal{F}$. Assume that the initial sample of size N is split into two disjoint parts S_1 and S_2 of cardinalities that differ at most by 1: $(X_1, Y_1), \dots, (X_N, Y_N) = S_1 \cup S_2$. The algorithm proceeds in the following way:

1. Let $\widehat{f}_{|S_1|}$ be the estimator (6.1.4) evaluated over subsample S_1 of cardinality $|S_1| \geq \lfloor N/2 \rfloor$, with the scale parameter Δ_1 and the partition parameter k_1 corresponding the group size $n_1 = \lfloor |S_1|/k_1 \rfloor$;

2. Let $\delta' = \overline{\delta} + C(\rho)\big(D^2 + \sigma(\ell, \mathcal{F})\sqrt{n} M_{\Delta_1}\big)\left(\frac{\mathfrak{B}^6(\ell,\mathcal{F})}{M_{\Delta_1}^4 n_1^2} + \frac{s+\mathcal{O}}{N}\right)$ be a known upper bound on the excess risk in Theorem 35 (while this condition is restrictive, it is similar to the requirements of existing approaches [BJL15, LM19c]; discussion of adaptation issues is beyond the scope of this paper and will be addressed elsewhere). Set

$$\widehat{\mathcal{F}}(\delta') := \Big\{ f \in \mathcal{F} : \ \widehat{\mathcal{E}}_N(f) \leq \delta' \Big\}.$$

3. Define $\widehat{f}''_N := \operatorname{argmin}_{f \in \widehat{\mathcal{F}}(\delta')} \widehat{\mathcal{L}}^{(k)}(f - \widehat{f}_{|S_1|})$ where

$$\widehat{\mathcal{L}}^{(k)}\Big(f - \widehat{f}_{|S_1|}\Big) = \operatorname*{argmin}_{y \in \mathbb{R}} \sum_{j=1}^{k_2} \rho\left(\sqrt{n}\, \frac{\Big(\overline{\mathcal{L}}_j(f) - \overline{\mathcal{L}}_j(\widehat{f}_{|S_1|})\Big) - y}{\Delta_2} \right)$$

is based on the subsample S_2 of cardinality $|S_2| \geq \lfloor N/2 \rfloor$, a scale parameter Δ_2 and the partition parameter k_2 corresponding the group size $n_2 = \lfloor |S_2|/k_2 \rfloor$.

It will be demonstrated in the course of the proofs that on event of high probability, $\widehat{\mathcal{F}}(\delta') \subseteq \mathcal{F}(c\delta')$ for an absolute constant $c \leq 7$. Hence, on this event $\sup_{f \in \widehat{\mathcal{F}}(\delta')} \operatorname{Var}(\ell(f(X)) - \ell(f_*(X))) \leq \nu^2(c\delta') \leq cD^2\delta'$ by the definition of $\nu(\delta)$ and Assumption 2, thus $\Delta_2 = D\, M_{\Delta_2} \sqrt{c\delta'}$ with $M_{\Delta_2} \geq 1$ often leads to an estimator with improved performance.

Theorem 36. *Suppose that*

$$\sup_{f \in \mathcal{F}} \mathbb{E}^{1/4}(\ell(f(X)) - \mathbb{E}\ell(f(X)))^4 < \infty$$

and that Δ_1, Δ_2 satisfy $M_{\Delta_1} := \frac{\Delta_1}{\sigma(\ell,\mathcal{F})} \geq 1$ and $M_{\Delta_2} := \frac{\Delta_2}{D\sqrt{7\delta'}} \geq 1$. Moreover, assume that for a sufficiently small absolute constant $c' > 0$, $\sup_{f \in \mathcal{F}} \max(G_f(n_1, \Delta_1), G_f(n_2, \Delta_2)) \leq c'$ and $\frac{s+\mathcal{O}}{\min(k_1,k_2)} \leq c'$. Finally, we require that

$$\sqrt{k_1} M_{\Delta_1} \geq \frac{c'}{\sigma(\ell,\mathcal{F})}\, \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{\sqrt{|S_1|}} \sum_{j=1}^{|S_1|} (\ell(f(X_j)) - P\,\ell(f)) \ \text{and} \tag{6.2.6}$$

$$\sqrt{k_2} M_{\Delta_2} \geq c' \frac{\sqrt{N\delta'}}{D}.$$

Then

$$\mathcal{E}\left(\widehat{f}_N''\right) \le \overline{\delta} + C(\rho)\Big(D^2 + D\sqrt{\delta'}\sqrt{n}M_{\Delta_2}\Big)\left(\frac{\mathfrak{B}^6(\ell,\mathcal{F})}{M_{\Delta_2}^4 n^2} + \frac{s+\mathcal{O}}{N}\right)$$

with probability at least $1-20e^{-s}$*, where* $C(\rho)$ *depends on* ρ *only and* D *is the constant appearing in Assumption 2.*

The statement of Theorem 36 is technical, so let us try to distill the main ideas. The key difference between Theorem 35 and Theorem 36 is that the "remainder term"

$$\sigma(\ell,\mathcal{F})\sqrt{n}M_\Delta\left(\frac{\mathfrak{B}^6(\ell,\mathcal{F})}{M_\Delta^4 n^2} + \frac{s+\mathcal{O}}{N}\right)$$

is replaced by a potentially much smaller quantity $\sqrt{\delta'}\sqrt{n}M_\Delta\left(\frac{\mathfrak{B}^6(\ell,\mathcal{F})}{M_\Delta^4 n^2} + \frac{s+\mathcal{O}}{N}\right)$. In particular, if $\delta' \ll \left(nM_\Delta^2\right)^{-1}$, this term often becomes negligible. To be more specific, assume that $\overline{\delta} = \frac{C(\mathcal{F})}{\sqrt{N}}\cdot h(N)$ where $h(N)\to 0$ as $N\to\infty$ (meaning that fast rates are achievable) and that $\mathcal{O} = \epsilon N$ for $\epsilon \ge \frac{1}{N}$. Moreover, suppose that $\mathfrak{B}(\ell,\mathcal{F})$ is bounded above by a constant. If Δ_1 is chosen such that $\Delta_1 \asymp \sigma(\ell,\mathcal{F})$, then $\delta' = C\Big(\overline{\delta} + \sigma(\ell,\mathcal{F})\Big(\left(\frac{k}{N}\right)^{3/2} + \frac{s+\mathcal{O}}{\sqrt{kN}}\Big)\Big)$. Hence, if $\max\Big(h(N)\sqrt{N}, N\epsilon^{2/3}\Big) \ll k_j \le CN\sqrt{\epsilon}$ for $j=1,2$ and $\Delta_2 \asymp \sqrt{\delta'}$, then

$$\delta'\cdot nM_{\Delta_2}^2 = O(1),$$

and the excess risk of $\widehat{f}_N''$ admits the bound

$$\mathcal{E}\left(\widehat{f}_N''\right) \le \overline{\delta} + C(\rho,D)\Big(\epsilon + \frac{s}{N}\Big)$$

that holds with probability at least $1-Ce^{-s}$. A possible choice satisfying all the required conditions is $k_j \asymp N\sqrt{\epsilon}$, $j=1,2$ (indeed, it this case it is straightforward to check that conditions (6.2.6) hold for sufficiently large N as $k_j \gtrsim \sqrt{N}$, $j=1,2$). Analysis of the case when $\mathcal{O}=0$ follows similar steps, with several simplifications.

6.3 Examples.

In this section, we consider two common prediction problems, regression and binary classification, and discuss the implications of our main results for these problems in detail.

6.3.1 Binary classification with convex surrogate loss.

The key elements of the binary classification framework were outlined in Section 6.1.4. Here, we recall few popular examples of classification-calibrated losses and present conditions that are sufficient for the Assumption 2 to hold.

Logistic loss $\ell(yf(z)) = \log\left(1+e^{-yf(z)}\right)$. Consider two scenarios:

1. Uniformly bounded classes, meaning that for all $f \in \mathcal{F}$, $\sup_{z\in S} |f(z)| \le B$. In this case, Assumption 2 holds with $D = 2e^B$ for all $f \in \mathcal{F}$. See [BJM04] and Proposition 6.1 in [ACL19].
2. Linear separators and Gaussian design: in this case, we assume that $S = \mathbb{R}^d$, $Z \sim N(0, I)$ is Gaussian, and $\mathcal{F} = \{\langle \cdot, v\rangle : \|v\|_2 \le R\}$ is a class of linear functions. In this case, according to the Proposition 6.2 in [ACL19], Bernstein's assumption is satisfied with $D = cR^{3/2}$ for some absolute constant $c > 0$.

Hinge loss $\ell(yf(z)) = \max(0, 1 - yf(z))$. In this case, sufficient condition for Assumption 2 to hold is the following: there exists $\tau > 0$ such that $|g_*(Z)| \ge \tau$ almost surely, where $g_*(z) = \mathbb{E}[Y|Z = z]$. It follows from Theorem 7 in [BJM04] (see also [Tsy04]) that Assumption 2 holds with $D = \frac{1}{\sqrt{2\tau}}$ in this case.

Bound for $\overline{\delta}$. Let Π stand for the marginal distribution of Z and recall that

$$\omega(\delta) := \mathbb{E} \sup_{\ell(f)\in\mathcal{F}(\delta)} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \Big((\ell(Y_j f(Z_j)) - \ell(Y_j f_*(Z_j))) - \mathbb{E}(\ell(Y f(Z)) - \ell(Y f_*(Z))) \Big) \right|.$$

Since ℓ is Lipchitz continuous by assumption (with Lipschitz constant denoted $L(\ell)$), consequent application of symmetrization and Talagrand's contraction inequalities [LT91, Van16] yields that

$$\omega(\delta) \le 4L(\ell)\, \mathbb{E} \sup_{\|f-f_*\|_{L_2(\Pi)}\le D\sqrt{\delta}} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \epsilon_j (f - f_*)(Z_j) \right|$$

where $\epsilon_1, \dots, \epsilon_N$ are i.i.d. random signs independent from Y_j's and Z_j's. The latter quantity is the modulus of continuity of a Rademacher process, and various upper bounds for it are well known. For instance, if $\mathcal{F}$ is a subset of a linear space of dimension d, then, according to Proposition 3.2 in [Kol11], $\mathbb{E}\sup_{\|f-f_*\|_{L_2(\Pi)}\le D\sqrt{\delta}} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \epsilon_j (f - f_*)(Z_j) \right| \le D\sqrt{\delta}\sqrt{d}$, whence $\widetilde{\omega}(\delta) := 4D\, L(\ell)\sqrt{\delta d}$ is an upper bound for $\omega(\delta)$ and is of concave type, implying that

$$\overline{\delta} \le C(\rho, \ell) D^2 \, \frac{d}{N}.$$

More generally, assume that the class $\mathcal{F}$ has a measurable envelope $F(z) := \sup_{f\in\mathcal{F}} |f(z)|$ that satisfies $\|F(Z)\|_{\psi_2} < \infty$, where $\|x\|_{\psi_2} := \inf\{C > 0 : \mathbb{E}\exp(|x/C|^2) \le 2\}$ is the ψ_2 (Orlicz) norm. Moreover, suppose that the covering numbers $N(\mathcal{F}, Q, \epsilon)$ of the class $\mathcal{F}$ with respect to the norm $L_2(Q)$ [3] satisfy the bound

$$N(\mathcal{F}, Q, \epsilon) \le \left(\frac{A\|F\|_{L_2(Q)}}{\epsilon} \right)^V \tag{6.3.1}$$

for some constants $A \ge 1$, $V \ge 1$, all $0 < \epsilon \le 2\|F\|_{L_2(Q)}$ and all probability measures Q. For instance, VC-subgraph classes are known to satisfy this bound with V being the VC dimension of $\mathcal{F}$ [vdVW96, Kol11]. In this case, it is not difficult to show (see for example the proof of Lemma 32 in the supplementary material) that

$$\mathbb{E} \sup_{\|f-f_*\|_{L_2(\Pi)}\le D\sqrt{\delta}} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \epsilon_j (f - f_*)(Z_j) \right| \le \widetilde{\omega}(\delta) := C\sqrt{V \log(e^2 A^2 N)} \left(\sqrt{\delta} + \sqrt{\frac{V}{N}} \log(A^2 N) \|F\|_{\psi_2} \right),$$

[3] Definition: the covering number $N(\mathcal{F}, Q, \epsilon)$ is the smallest integer $k \ge 1$ such that there exist $f_1, \dots, f_k \in L_2(Q)$ satisfying $\bigcup_{j=1}^{k} B(f_j, \epsilon) \supseteq \mathcal{F}$, where $B(f_j, \epsilon)$ is the $L_2(Q)$ ball of radius ϵ centered at f_j.

hence it is easy to check that in this case

$$\bar{\delta} \le C(\rho)\frac{V\log^{3/2}(e^2A^2N)\|F\|_{\psi_2}}{N}.$$

It immediately follows from the discussion following Theorem 36 that the excess risk of the estimator $\widehat{f}_N''$ satisfies

$$\mathcal{E}\left(\widehat{f}_N''\right) \le C(\rho, D)\left(\frac{\mathcal{O}}{N} + \frac{V\log^{3/2}(e^2A^2N)\|F\|_{\psi_2} + s}{N}\right)$$

with probability at least $1 - 20e^{-s}$. Note that we did not need to assume that the oracle $h_* := \underset{\text{all measurable } f}{\operatorname{argmin}} \mathbb{E}\ell(Yf(Z))$ belongs to $\mathcal{F}$. Similar results hold for regression problems with Lipschitz losses, such as Huber's loss or quantile loss [ACL19].

6.3.2 Regression with quadratic loss.

Let $X = (Z, Y) \in S \times \mathbb{R}$ be a random couple with distribution P satisfying $Y = f_*(Z) + \eta$ where the noise variable η is independent of Z and $f_*(z) = \mathbb{E}[Y|Z = z]$ is the regression function. Let $\|\eta\|_{2,1} := \int_0^\infty \sqrt{\Pr(|\eta| > t)}dt$, and observe that $\|\eta\|_{2,1} < \infty$ as $\sup_{f\in\mathcal{F}} \mathbb{E}(Y - f(Z))^4 < \infty$ by assumption. As before, Π will stand for the marginal distribution of Z. Let $\mathcal{F}$ be a given convex class of functions mapping S to $\mathbb{R}$ and such that the regression function f_* belongs to $\mathcal{F}$, so that

$$f_* = \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}(Y - f(Z))^2.$$

In this case, the natural choice for the loss function is the quadratic loss $\ell(x) = x^2$ which is not Lipschitz continuous on unbounded domains. Assume that the class $\mathcal{F}$ has a measurable envelope $F(z) := \sup_{f\in\mathcal{F}} |f(z)|$ that satisfies $\|F(Z)\|_{\psi_2} < \infty$. Moreover, suppose that the covering numbers $N(\mathcal{F}, Q, \epsilon)$ of the class $\mathcal{F}$ with respect to the norm $L_2(Q)$ satisfy the bound

$$N(\mathcal{F}, Q, \epsilon) \le \left(\frac{A\|F\|_{L_2(Q)}}{\epsilon}\right)^V \tag{6.3.2}$$

for some constants $A \ge 1$, $V \ge 1$, all $0 < \epsilon \le 2\|F\|_{L_2(Q)}$, and all probability measures Q. For instance, VC-subgraph classes are known to satisfy this bound with V being the VC dimension of $\mathcal{F}$ [vdVW96, Kol11].

Verification of Bernstein's assumption. It follows from Lemma 5.1 in [Kol11] that

$$\mathcal{F}(\delta) \subseteq \left\{(y - f(z))^2 : \ f \in \mathcal{F}, \ \mathbb{E}(f(Z) - f_*(Z))^2 \le 2\delta\right\},$$

hence $\nu(\delta) \le \sqrt{2\delta}$ so D can be taken to be $\sqrt{2}$ in Assumption 2.

Bound for $\bar{\delta}$. Required estimates follow from the following lemma:

Lemma 32. *Under the assumptions made in this section and for $\Delta \ge \sigma(\ell, \mathcal{F})$,*

$$\bar{\delta} \le C(\rho)\frac{V\log^2(A^2N)(\|F\|_{\psi_2}^2 + \|\eta\|_{2,1}^2)}{N}.$$

Moreover, if the functions if $\mathcal{F}$ are uniformly bounded, the $\log^2(A^2N)$ can be removed.

The proof is given in Section 6.5.9 of the supplementary material. An immediate corollary of the lemma, according to the discussion following Theorem 36, is that the excess risk of the estimator $\widehat{f}''_N$ satisfies the inequality

$$\mathcal{E}\Big(\widehat{f}''_N\Big) \le C(\rho)\left(\frac{\mathcal{O}}{N} + \frac{V\log^2(A^2N)(\|F\|^2_{\psi_2} + \|\eta\|^2_{2,1}) + s}{N}\right)$$

with probability at least $1 - 20e^{-s}$, for $0 < s \le cN^{1/4}$.

6.4 Proofs of the main results.

In the proofs of the main results, we will rely on the following convenient change of variables. Denote

$$\widehat{G}_k(z;f) = \frac{1}{\sqrt{k}}\sum_{j=1}^{k}\rho'\left(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f) - \mathcal{L}(f)) - z}{\Delta}\right),$$
$$G_k(z;f) = \sqrt{k}\,\mathbb{E}\rho'\left(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_1(f) - \mathcal{L}(f)) - z}{\Delta}\right).$$

In particular, when $\mathcal{O} = 0$, $G_k(z;f) = \mathbb{E}\widehat{G}_k(z;f)$. Let $\widehat{e}^{(k)}(f)$ and $e^{(k)}(f)$ be defined by the equations

$$\widehat{G}_k\Big(\widehat{e}^{(k)}(f); f\Big) = 0,$$
$$G_k\Big(e^{(k)}(f); f\Big) = 0.$$

Comparing this to the definition of $\widehat{\mathcal{L}}^{(k)}(f)$ (6.1.2), it is easy to see that $\widehat{e}^{(k)}(f) = \widehat{\mathcal{L}}^{(k)}(f) - \mathcal{L}(f)$.

Let us explain the main high-level ideas behind the proof. In the classical empirical risk minimization framework, $\widehat{\mathcal{L}}^{(k)}(f)$ is replaced by the empirical mean $P_N\ell(f) = \frac{1}{N}\sum_{j=1}^{N}\ell(f(X_j))$; in particular, it is linear in $\ell(f)$, meaning that $P_N(\ell(f_1) - \ell(f_2)) = P_N\ell(f_1) - P_N\ell(f_2)$, while $\widehat{\mathcal{L}}^{(k)}(f)$ lacks this property. Imagine that $\widehat{\mathcal{L}}^{(k)}(f)$ was linear in $\ell(f)$. Then, setting $\widehat{\delta}_N = \mathcal{L}(\widehat{f}_N) - \mathcal{L}(f_*)$, we would be able to write that

$$\widehat{\delta}_N = \mathcal{L}(\widehat{f}_N) - \mathcal{L}(f_*) = (\mathcal{L}(\widehat{f}_N) - \widehat{\mathcal{L}}^{(k)}(\widehat{f}_N)) - (\mathcal{L}(f_*) - \widehat{\mathcal{L}}^{(k)}(f_*)) + \underbrace{\widehat{\mathcal{L}}^{(k)}(\widehat{f}_N) - \widehat{\mathcal{L}}^{(k)}(f_*)}_{\le 0}$$
$$\le \sup_{f:\mathcal{E}(f)\le\widehat{\delta}_N}\Big|\widehat{\mathcal{L}}^{(k)}(f - f_*) - \mathcal{L}(f - f_*)\Big|. \quad (6.4.1)$$

It would then suffice to find a good upper bound for the supremum on the right side of (6.4.1) and solve the resulting inequality to get an upper bound for $\widehat{\delta}_N$. However, this argument does not work directly due to the lack of linearity. Instead, we use Bahadur-type representation of the $\widehat{e}^{(k)}(f)$ to introduce linearity into the problem. Specifically, we will show that $\widehat{e}^{(k)}(f) = -\frac{\widehat{G}_k(0;f)}{\partial_z G_k(0;f)} + r_N(f)$ where $r_N(f)$ is a small remainder term. The process $\widehat{G}_k(0;f)$ is "almost" linear in $\ell(f)$, the only

obstacle being the nonlinearity due to ρ'. Mimicking (6.4.1), we can write that

$$\widehat{\delta}_N = \widehat{e}^{(k)}(\widehat{f}_N) - \widehat{e}^{(k)}(f_*) + \underbrace{\widehat{\mathcal{L}}^{(k)}(\widehat{f}_N) - \widehat{\mathcal{L}}^{(k)}(f_*)}_{\leq 0} \leq \widehat{e}^{(k)}(\widehat{f}_N) - \widehat{e}^{(k)}(f_*)$$

$$= \left| \frac{\widehat{G}_k\Big(0; \widehat{f}_N\Big)}{\partial_z G\Big(0; \widehat{f}_N\Big)} - \frac{\widehat{G}_k(0; f_*)}{\partial_z G(0; f_*)} \right| + r'_N(\widehat{f}_N, f_*) \leq \sup_{f: \mathcal{E}(f) \leq \widehat{\delta}_N} \left(\left| \frac{\widehat{G}_k(0; f)}{\partial_z G(0; f)} - \frac{\widehat{G}_k(0; f_*)}{\partial_z G(0; f_*)} \right| + r'_N(f, f_*) \right)$$

for appropriately defined $r'_N(\cdot, \cdot)$. The difference $\frac{\widehat{G}_k(0;f)}{\partial_z G(0;f)} - \frac{\widehat{G}_k(0;f_*)}{\partial_z G(0;f_*)}$ can be tackled with the techniques commonly used to estimate suprema of the empirical processes; in particular, symmetrization and contraction inequalities for Rademacher sums [LT91] are used to remove the additional nonlinearity in the definition of $\widehat{G}_k(z, f)$ introduced by ρ'. At that point, one only needs to carefully estimate the remainder term r'_N.

6.4.1 Technical tools.

We summarize the key results that our proofs rely on.

Lemma 33. *Let ρ satisfy Assumption 1. Then for any random variable Y with $\mathbb{E}Y^2 < \infty$,*

$$Var(\rho'(Y)) \leq Var(Y).$$

Proof. See Lemma 5.3 in [Min18]. ∎

Lemma 34. *For any function h of with bounded third derivative and a sequence of i.i.d. random variables $x_1, \dots, x_n$ such that $\mathbb{E}x_1 = 0$ and $\mathbb{E}|x_1|^3 < \infty$,*

$$\left| \mathbb{E}h\left(\sum_{j=1}^n x_j \right) - \mathbb{E}h\left(\sum_{j=1}^n Z_j \right) \right| \leq Cn \, \|h'''\|_\infty \, \mathbb{E}|x_1|^3,$$

where $C > 0$ is an absolute constant and $Z_1, \dots, Z_n$ are i.i.d. centered normal random variables such that $Var(Z_1) = Var(x_1)$.

Proof. This bound follows from a standard application of Lindeberg's replacement method; see chapter 11 in [O'D14]. ∎

Lemma 35. *Assume that $\mathbb{E}|f(X) - \mathbb{E}f(X)|^2 < \infty$ for all $f \in \mathcal{F}$ and that ρ satisfies Assumption 1. Then for all $f \in \mathcal{F}$ and $z \in \mathbb{R}$ satisfying $|z| \leq \frac{\Delta}{\sqrt{n}} \frac{1}{2}$,*

$$\left| \mathbb{E}\rho'\left(\sqrt{n} \frac{(\bar{\theta}_j(f) - Pf) - z}{\Delta} \right) - \mathbb{E}\rho'\left(\frac{W(f) - \sqrt{n}z}{\Delta} \right) \right| \leq 2 \, G_f(n, \Delta).$$

Proof. See Lemma 4.2 in [Min18]. ∎

Given N i.i.d. random variables $X_1, \dots, X_N \in \mathcal{S}$, let $\|f-g\|_{L_\infty(\Pi_N)} := \max_{1\le j\le N} |f(X_j)-g(X_j)|$. Moreover, define

$$\Gamma_{n,\infty}(\mathcal{F}) := \mathbb{E}\gamma_2^2(\mathcal{F}; L_\infty(\Pi_N)),$$

where $\gamma_2(\mathcal{F}, L_\infty(\Pi_N))$ is Talagrand's generic chaining complexity [Tal14].

Lemma 36. *Let $\sigma^2 := \sup_{f\in\mathcal{G}} \mathbb{E}f^2(X)$. Then there exists a universal constant $C>0$ such that*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{j=1}^N f^2(X_j) - \mathbb{E}f^2(X)\right| \le C\left(\sigma\sqrt{\frac{\Gamma_{N,\infty}(\mathcal{F})}{N}} \bigvee \frac{\Gamma_{N,\infty}(\mathcal{F})}{N}\right).$$

Proof. See Theorem 3.16 in [Kol11]. ∎

The following form of Talagrand's concentration inequality is due to Klein and Rio (see Section 12.5 in [BLM13]).

Lemma 37. *Let $\{Z_j(f),\ f\in\mathcal{F}\}$, $j=1,\dots,N$ be independent (not necessarily identically distributed) separable stochastic processes indexed by class $\mathcal{F}$ and such that $|Z_j(f)-\mathbb{E}Z_j(f)|\le M$ a.s. for all $1\le j\le N$ and $f\in\mathcal{F}$. Then the following inequality holds with probability at least $1-e^{-s}$:*

$$\sup_{f\in\mathcal{F}}\left(\sum_{j=1}^N (Z_j(f)-\mathbb{E}Z_j(f))\right) \le 2\mathbb{E}\sup_{f\in\mathcal{F}}\left(\sum_{j=1}^N (Z_j(f)-\mathbb{E}Z_j(f))\right) + V(\mathcal{F})\sqrt{2s} + \frac{4Ms}{3} \quad (6.4.2)$$

where $V^2(\mathcal{F}) = \sup_{f\in\mathcal{F}}\sum_{j=1}^N Var(Z_j(f))$.

It is easy to see, applying (6.4.2) to processes $\{-Z_j(f),\ f\in\mathcal{F}\}$, that

$$\inf_{f\in\mathcal{F}}\left(\sum_{j=1}^N (Z_j(f)-\mathbb{E}Z_j(f))\right) \ge -2\mathbb{E}\sup_{f\in\mathcal{F}}\left(\sum_{j=1}^N (\mathbb{E}Z_j(f)-Z_j(f))\right) - V(\mathcal{F})\sqrt{2s} - \frac{4Ms}{3}$$

with probability at least $1-e^{-s}$.

6.4.2 Proof of Theorems 34 and 35.

We will provide detailed proofs for the estimator $\widehat{f}_N$ that is based on disjoint subsamples indexed by $G_1,\dots,G_k$. The bounds for its permutation-invariant version $\widehat{f}_N^U$ follow exactly the same steps where all applications of the Talagrand's concentration inequality (Lemma 37) should be replaced by its version (6.6.3) for nondegenerate U-statistics stated in Section 6.6 of the supplementary material.

Let $J\subset\{1,\dots,k\}$ of cardinality $|J|\ge k-\mathcal{O}$ be the set containing all j such that the subsample $\{X_i,\ i\in G_j\}$ does not include outliers. Clearly, $\{X_i:\ i\in G_j,\ j\in J\}$ are still conditionally i.i.d. as the partitioning scheme is independent of the data. Moreover, set $N_J := \sum_{j\in J}|G_j|$, and note that, since $\mathcal{O} < k/2$,

$$N_J \ge n|J| \ge \frac{N}{2}.$$

Consider stochastic process $R_N(f)$ defined as

$$R_N(f) = \widehat{G}_k(0;f) + \partial_z G_k(0;f)\cdot \widehat{e}^{(k)}(f), \tag{6.4.3}$$

where $\partial_z G_k(0;f) := \partial_z G_k(z;f)_{|_{z=0}}$. Whenever $\partial_z G_k(0;f)\neq 0$ (this assumption will be justified by Lemma 38 below), we can solve (6.4.3) for $\widehat{e}^{(k)}(f)$ to obtain

$$\widehat{e}^{(k)}(f) = -\frac{\widehat{G}_k(0;f)}{\partial_z G_k(0;f)} + \frac{R_N(f)}{\partial_z G_k(0;f)}, \tag{6.4.4}$$

which can be viewed as a Bahadur-type representation of $\widehat{e}^{(k)}(f)$. Setting $f := \widehat{f}_N$ and recalling that $\widehat{e}^{(k)}(f) = \widehat{\mathcal{L}}^{(k)}(f) - \mathcal{L}(f)$, we deduce that

$$\widehat{\mathcal{L}}^{(k)}(\widehat{f}_N) = \mathcal{L}(\widehat{f}_N) - \frac{\widehat{G}_k\Big(0;\widehat{f}_N\Big)}{\partial_z G_k\Big(0;\widehat{f}_N\Big)} + \frac{R_N(\widehat{f}_N)}{\partial_z G_k\Big(0;\widehat{f}_N\Big)}.$$

By the definition (6.1.4) of $\widehat{f}_N$, $\widehat{\mathcal{L}}^{(k)}(\widehat{f}_N)\le \widehat{\mathcal{L}}^{(k)}(f_*)$, hence

$$\mathcal{L}(\widehat{f}_N) - \frac{\widehat{G}_k\Big(0;\widehat{f}_N\Big)}{\partial_z G_k\Big(0;\widehat{f}_N\Big)} + \frac{R_N(\widehat{f}_N)}{\partial_z G_k\Big(0;\widehat{f}_N\Big)} \le \mathcal{L}(f_*) - \frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)} + \frac{R_N(f_*)}{\partial_z G_k(0;f_*)}.$$

Rearranging the terms, it is easy to see that

$$\widehat{\delta}_N = \mathcal{L}(\widehat{f}_N) - \mathcal{L}(f_*) \le \left|\frac{\widehat{G}_k\Big(0;\widehat{f}_N\Big)}{\partial_z G\Big(0;\widehat{f}_N\Big)} - \frac{\widehat{G}_k(0;f_*)}{\partial_z G(0;f_*)}\right| + 2\sup_{f\in\mathcal{F}(\widehat{\delta}_N)}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right|. \tag{6.4.5}$$

Remark. *Similar argument also implies, in view of the inequality $\mathcal{L}(f_*)\le \mathcal{L}(\widehat{f}_N)$, that*

$$\widehat{\mathcal{L}}^{(k)}(f_*) + \frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)} - \frac{R_N(f_*)}{\partial_z G_k(0;f_*)} \le \widehat{\mathcal{L}}^{(k)}(\widehat{f}_N) + \frac{\widehat{G}_k\Big(0;\widehat{f}_N\Big)}{\partial_z G_k\Big(0;\widehat{f}_N\Big)} - \frac{R_N(\widehat{f}_N)}{\partial_z G_k\Big(0;\widehat{f}_N\Big)},$$

hence

$$\widehat{\mathcal{L}}^{(k)}(f_*) - \widehat{\mathcal{L}}^{(k)}(\widehat{f}_N) \le \left|\frac{\widehat{G}_k\Big(0;\widehat{f}_N\Big)}{\partial_z G\Big(0;\widehat{f}_N\Big)} - \frac{\widehat{G}_k(0;f_*)}{\partial_z G(0;f_*)}\right| + 2\sup_{f\in\mathcal{F}(\widehat{\delta}_N)}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right|.$$

It follows from (6.4.5) that in order to estimate the excess risk of $\widehat{f}_N$, it suffices to obtain the upper bounds for

$$A_1 := \left|\frac{\widehat{G}_k\Big(0;\widehat{f}_N\Big)}{\partial_z G_k\Big(0;\widehat{f}_N\Big)} - \frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)}\right| \text{ and } A_2 := \sup_{f\in\mathcal{F}(\widehat{\delta}_N)}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right|.$$

Observe that

$$\frac{\widehat{G}_k\Big(0;\widehat{f}_N\Big)}{\partial_z G_k\Big(0;\widehat{f}_N\Big)} - \frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)}$$
$$= \frac{\widehat{G}_k\Big(0;\widehat{f}_N\Big) - \widehat{G}_k(0;f_*)}{\partial_z G_k\Big(0;\widehat{f}_N\Big)} + \frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\partial_z G_k\Big(0;\widehat{f}_N\Big)}\Big(\partial_z G_k(0;f_*) - \partial_z G_k\Big(0;\widehat{f}_N\Big)\Big).$$

Since ρ'' is Lipschitz continuous by assumption,

$$\left|\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\partial_z G_k\Big(0;\widehat{f}_N\Big)}\Big(\partial_z G_k(0;f_*) - \partial_z G_k\Big(0;\widehat{f}_N\Big)\Big)\right|$$
$$= \left|\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\partial_z G_k\Big(0;\widehat{f}_N\Big)}\frac{\sqrt{nk}}{\Delta}\mathbb{E}\left(\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f_*) - \mathcal{L}(f_*)}{\Delta}\right) - \rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(\widehat{f}_N) - \mathcal{L}(\widehat{f}_N)}{\Delta}\right)\right)\right|$$
$$\le L(\rho'')\left|\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\partial_z G_k\Big(0;\widehat{f}_N\Big)}\frac{\sqrt{nk}}{\Delta^2}\mathrm{Var}^{1/2}\Big(\ell(\widehat{f}_N(X)) - \ell(f_*(X))\Big)\right|$$
$$= C(\rho)\left|\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\partial_z G_k\Big(0;\widehat{f}_N\Big)}\right|\frac{\sqrt{nk}}{\Delta^2}\nu(\widehat{\delta}_N). \quad (6.4.6)$$

The following two lemmas are required to proceed.

Lemma 38. *There exist $C(\rho) > 0$ such that for any $f \in \mathcal{F}$,*

$$|\partial_z G_k(0;f)| \ge \frac{\sqrt{kn}}{2\sqrt{2\pi}\Delta}\left(\min\left(\frac{\Delta}{\sqrt{Var(\ell(f(X)))}}, 2\sqrt{\log 2}\right) - \frac{C(\rho)}{\sqrt{n}}\mathbb{E}\left|\frac{\ell(f(X)) - P\ell(f)}{\Delta}\right|^3\right).$$

Proof. See Section 6.5.1. ∎

In particular, the bound of Lemma 38 implies that for n large enough,

$$\inf_{f\in\mathcal{F}}|\partial_z G_k(0;f)| \ge \frac{1}{4\sqrt{2\pi}}\frac{\sqrt{kn}}{\max(\Delta,\sigma(\ell,\mathcal{F}))} = \frac{1}{4\sqrt{2\pi}}\frac{\sqrt{kn}}{\widetilde{\Delta}}. \quad (6.4.7)$$

It is also easy to deduce from the proof of Lemma 38 that for small n and $\Delta > \sigma(\ell,\mathcal{F})$, $\inf_{f\in\mathcal{F}}|\partial_z G_k(0;f)| \ge c(\rho)\frac{\sqrt{kn}}{\Delta}$ for some positive $c(\rho)$.

Lemma 39. *For any $f \in \mathcal{F}$,*

$$\widehat{G}_k(0;f) \le 2\left(\sqrt{k}\,G_f(n,\Delta) + \frac{\sigma(\ell,f)}{\Delta}\sqrt{s} + \frac{2s}{\sqrt{k}} + \frac{\mathcal{O}}{\sqrt{k}}\right)$$

with probability at least $1 - 2e^{-s}$, where $C > 0$ is an absolute constant.

Proof. See Section 6.5.2. ∎

Lemma 39 and (6.4.7) imply, together with (6.4.6), that

$$\left|\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\partial_z G_k\Big(0;\widehat{f}_N\Big)}\Big(\partial_z G_k(0;f_*)-\partial_z G_k\Big(0;\widehat{f}_N\Big)\Big)\right| \\ \le C(\rho)\frac{\widetilde{\Delta}^2}{\Delta^2}\left(\frac{\sigma(\ell,f_*)}{\Delta}\sqrt{\frac{s}{N}}+\frac{G_{f_*}(n,\Delta)}{\sqrt{n}}+\sqrt{n}\frac{s}{N}+\sqrt{n}\frac{\mathcal{O}}{N}\right)\nu(\widehat{\delta}_N)$$

on event Θ_1 of probability at least $1-2e^{-s}$. As $\widetilde{\Delta}\ge\sigma(\ell,\mathcal{F})$ by assumption, we deduce that

$$\left|\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\partial_z G_k\Big(0;\widehat{f}_N\Big)}\Big(\partial_z G_k(0;f_*)-\partial_z G_k\Big(0;\widehat{f}_N\Big)\Big)\right| \\ \le C(\rho)\nu(\widehat{\delta}_N)\left(\sqrt{\frac{s}{N}}+\frac{G_{f_*}(n,\Delta)}{\sqrt{n}}+\sqrt{n}\frac{s}{N}+\sqrt{n}\frac{\mathcal{O}}{N}\right).$$

Define

$$\overline{\delta}_1 := \min\left\{\delta>0:\ C_1(\rho)\left(\sqrt{\frac{s}{N}}+\frac{G_{f_*}(n,\Delta)}{\sqrt{n}}+\sqrt{n}\frac{s+\mathcal{O}}{N}\right)\frac{\widetilde{\nu}(\delta)}{\delta}\le\frac{1}{7}\right\} \tag{6.4.8}$$

where $C_1(\rho)$ is sufficiently large. It is easy to see that on event $\Theta_1\cap\{\widehat{\delta}_N>\overline{\delta}_1\}$,

$$\left|\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)\partial_z G_k\Big(0;\widehat{f}_N\Big)}\Big(\partial_z G_k(0;f_*)-\partial_z G_k\Big(0;\widehat{f}_N\Big)\Big)\right|\le\frac{\widehat{\delta}_N}{7}, \tag{6.4.9}$$

for appropriately chosen $C_1(\rho)$.

Our next goal is to obtain an upper bound for $\left|\frac{\widehat{G}_k\left(0;\widehat{f}_N\right)-\widehat{G}_k(0;f_*)}{\partial_z G_k\left(0;\widehat{f}_N\right)}\right|$. To this end, we will need to control the local oscillations of the process $\widehat{G}_k(0;f)$. Specifically, we are interested in the bounds on the random variable $\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_k(0;f)-\widehat{G}_k(0;f_*)\right|$. The following technical lemma is important for the analysis.

Lemma 40. *Let $(x_1,\eta_1),\dots,(x_n,\eta_n)$ be a sequence of independent identically distributed random couples such that $\mathbb{E}x_1=0$, $\mathbb{E}\eta_1=0$, and $\mathbb{E}|x_1|^2+\mathbb{E}|\eta_1|^2<\infty$. Let F be an odd, smooth function with bounded derivatives up to fourth order. Then*

$$\left|\mathbb{E}F\left(\sum_{j=1}^n x_j\right)-\mathbb{E}F\left(\sum_{j=1}^n\eta_j\right)\right|\le\max_{\alpha\in[0,1]}\sqrt{n}\ Var^{1/2}(x_1-\eta_1)\Big(\mathbb{E}|F'(S_n^\eta+\alpha(S_n^x-S_n^\eta))|^2\Big)^{1/2}.$$

Moreover, if $\mathbb{E}|x_1|^4+\mathbb{E}|\eta_1|^4<\infty$, then

$$\left|\mathbb{E}F\left(\sum_{j=1}^n x_j\right)-\mathbb{E}F\left(\sum_{j=1}^n\eta_j\right)\right|\le C(F)\cdot n\Big(\ Var^{1/2}(x_1-\eta_1)\big(R_4^2+\sqrt{n-1}R_4^3\big) \\ +\big(\mathbb{E}|x_1-\eta_1|^4\big)^{1/4}R_4^3\Big),$$

where $R_4=\big(\max\big(\mathbb{E}|x_1|^4,\mathbb{E}|\eta_1|^4\big)\big)^{1/4}$ and $C(F)>0$ is a constant that depends only on F.

Proof. See Section 6.5.3. ∎

Now we are ready to state the bound for the local oscillations of the process $\widehat{G}_k(0;f)$. Let

$$U(\delta,s) := \frac{2}{\Delta}\left(8\sqrt{2}\omega(\delta) + \nu(\delta)\sqrt{\frac{s}{2}}\right) + \frac{32s}{3\sqrt{k}}.$$

Moreover, if $\widetilde{\omega}(\delta)$ and $\widetilde{\nu}(\delta)$ are upper bounds for $\omega(\delta)$ and $\nu(\delta)$ and are of concave type, then

$$\widetilde{U}(\delta,s) := \frac{2}{\Delta}\left(c(\gamma)\,\widetilde{\omega}(\delta) + \widetilde{\nu}(\delta)\sqrt{\frac{s}{2}}\right) + \frac{32s}{\sqrt{k}}, \tag{6.4.10}$$

where $c(\gamma) > 0$ depends only on γ, is also an upper bound for $U(\delta,s)$ of strictly concave type. Moreover, define

$$\begin{aligned}
R_4(\ell,\mathcal{F}) &:= \sup_{f\in\mathcal{F}} \mathbb{E}^{1/4}\Big(\ell(f(X)) - \mathbb{E}\ell(f(X))\Big)^4,\\
\nu_4(\delta) &:= \sup_{f\in\mathcal{F}(\delta)} \mathbb{E}^{1/4}\Big(\ell(f(X)) - \ell(f_*(X)) - \mathbb{E}(\ell(f(X)) - \ell(f_*(X)))\Big)^4,\\
\mathfrak{B}(\ell,\mathcal{F}) &:= \frac{R_4(\ell,\mathcal{F})}{\sigma(\ell,\mathcal{F})},\\
\widetilde{B}(\delta) &:= \begin{cases} \frac{\widetilde{\nu}(\delta)}{\Delta}\frac{1}{M_\Delta}, & R_4(\ell,\mathcal{F}) = \infty,\\ \frac{\mathfrak{B}^3(\ell,\mathcal{F})}{\sqrt{n}}\left(\frac{\widetilde{\nu}(\delta)}{\Delta}\frac{1}{M_\Delta^2} + \frac{\widetilde{\nu}_4(\delta)}{\Delta}\frac{1}{M_\Delta^3\sqrt{n}}\right), & R_4(\ell,\mathcal{F}) < \infty, \end{cases}
\end{aligned}$$

where $\widetilde{\nu}_4(\delta)$ upper bounds $\nu_4(\delta)$ and is of concave type. Below, we will use a crude bound $\nu_4(\delta) \le 2R_4(\ell,\mathcal{F})$, but additional improvements are possible if better estimates of $\nu_4(\delta)$ are available.

Lemma 41. *With probability at least* $1 - e^{-2s}$,

$$\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_k(0;f) - \widehat{G}_k(0;f_*)\right| \le U(\delta,s) + C(\rho)\sqrt{k}\,\widetilde{B}(\delta) + 4\frac{\mathcal{O}}{\sqrt{k}},$$

where $C(\rho) > 0$ *is constant that depends only on* ρ.

Proof. See Section 6.5.4. ∎

Next, we state the "uniform version" of Lemma 41.

Lemma 42. *With probability at least* $1 - e^{-s}$, *for all* $\delta \ge \delta_{\min}$ *simultaneously,*

$$\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_k(0;f) - \widehat{G}_k(0;f_*)\right| \le C(\rho)\delta\left(\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}} + \sqrt{k}\frac{\widetilde{B}(\delta_{\min})}{\delta_{\min}}\right) + 4\frac{\mathcal{O}}{\sqrt{k}},$$

where $C(\rho) > 0$ *is constant that depends only on* ρ.

Proof. See Section 6.5.5. ∎

It follows from Lemma 42 and inequality (6.4.7) that on event Θ_2 of probability at least $1-e^{-s}$, for all $\delta \geq \delta_{\min}$ simultaneously,

$$\sup_{f\in\mathcal{F}(\delta)}\left|\frac{\widehat{G}_k(0;f)-\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f)}\right| \leq C(\rho)\delta\left(\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}}+\frac{\widetilde{\Delta}}{\sqrt{n}}\frac{\widetilde{B}(\delta_{\min})}{\delta_{\min}}\right)+4\widetilde{\Delta}\sqrt{n}\frac{\mathcal{O}}{N}.$$

Define

$$\overline{\delta}_2 := \min\left\{\delta>0:\ C_2(\rho)\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\delta,s)}{\delta}\leq\frac{1}{7}\right\},$$
$$\overline{\delta}_3 := \min\left\{\delta>0:\ C_3(\rho)\frac{\widetilde{\Delta}}{\sqrt{n}}\frac{\widetilde{B}(\delta)}{\delta}\leq\frac{1}{7}\right\},$$

where $C_2(\rho),\ C_3(\rho)$ are sufficiently large constants. Then, on event $\Theta_2\cap\left\{\widehat{\delta}_N>\max(\overline{\delta}_2,\overline{\delta}_3)\right\}$,

$$\sup_{f\in\mathcal{F}(\widehat{\delta}_N)}\left|\frac{\widehat{G}_k(0;f)-\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f)}\right| \leq \frac{2\,\widehat{\delta}_N}{7}+4\widetilde{\Delta}\sqrt{n}\frac{\mathcal{O}}{N} \tag{6.4.11}$$

for appropriately chosen $C_2(\rho), C_3(\rho)$.

Finally, we provide an upper bound for the process $R_N(f)$ defined via

$$R_N(f)=\widehat{G}_k(0;f)+\partial_z G_k(0;f)\cdot\widehat{e}^{(k)}(f).$$

Lemma 43. *Assume that conditions of Theorem 33 hold, and let $\delta_{\min}>0$ be fixed. Then for all $s>0$, $\delta\geq\delta_{\min}$, positive integers n and k such that*

$$\delta\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}\sqrt{k}}+\sup_{f\in\mathcal{F}}G_f(n,\Delta)+\frac{s+\mathcal{O}}{k}\leq c(\rho), \tag{6.4.12}$$

the following inequality holds with probability at least $1-7e^{-s}$, uniformly over all δ satisfying (6.4.12):

$$\sup_{f\in\mathcal{F}(\delta)}|R_N(f)|\leq C(\rho)\sqrt{N}\frac{\widetilde{\Delta}^2}{\Delta^2}\left(n^{1/2}\delta^2\left(\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}\sqrt{N}}\right)^2\bigvee\frac{\sigma^2(\ell,f_*)}{\Delta^2}\frac{n^{1/2}\,s}{N}\right.$$
$$\left.\bigvee n^{1/2}\left(\sup_{f\in\mathcal{F}}\frac{G_f(n,\Delta)}{\sqrt{n}}\right)^2\bigvee n^{3/2}\frac{s^2}{N^2}\bigvee n^{3/2}\frac{\mathcal{O}^2}{N^2}\right).$$

Moreover, the bound of Theorem 33 holds on the same event.

Proof. See Section 6.5.6. ∎

Recall that

$$\overline{\delta}_2=\min\left\{\delta>0:\ C_2(\rho)\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\delta,s)}{\delta}\leq\frac{1}{7}\right\},$$

where $C_2(\rho)$ is a large enough constant. Let Θ_3 be the event of probability at least $1-7e^{-s}$ on which Lemma 43 holds with $\delta_{\min}=\overline{\delta}_2$, and consider the event $\Theta_3\cap\{\widehat{\delta}_N>\overline{\delta}_2\}$. We will now show that on this event, Lemma 43 applies with $\delta=\widehat{\delta}_N$. Indeed, the bound of Theorem 33 is valid on Θ_3, hence the inequality (6.2.4) implies that on Θ_3, $\widehat{\delta}_N\le C(\rho)\frac{\widetilde{\Delta}}{\sqrt{n}}$, and it is straightforward to check that condition (6.4.12) of Lemma 43 holds with $\delta_{\min}=\overline{\delta}_2$ and $\delta=\widehat{\delta}_N$. It follows from inequality (6.4.7) that on event $\Theta_3\cap\{\widehat{\delta}_N\ge\overline{\delta}_2\}$,

$$\sup_{f\in\mathcal{F}(\widehat{\delta}_N)}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right|\le C(\rho)\frac{\widetilde{\Delta}^2}{\Delta^2}\left(\frac{n^{1/2}}{\widetilde{\Delta}}\widehat{\delta}_N^2\left(\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\delta_2,s)}{\delta_2}\right)^2\bigvee\widetilde{\Delta}\frac{\sigma^2(\ell,f_*)}{\Delta^2}\frac{n^{1/2}\,s}{N}\right.$$
$$\left.\bigvee n^{1/2}\widetilde{\Delta}\left(\sup_{f\in\mathcal{F}}\frac{G_f(n,\Delta)}{\sqrt{n}}\right)^2\bigvee n^{3/2}\widetilde{\Delta}\frac{s^2+\mathcal{O}^2}{N^2}\right).$$

Consider the expression

$$C(\rho)\frac{\widetilde{\Delta}^2}{\Delta^2}\frac{n^{1/2}}{\widetilde{\Delta}}\widehat{\delta}_N^2\left(\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\delta_2,s)}{\delta_2}\right)^2=C(\rho)\frac{\widetilde{\Delta}^2}{\Delta^2}\left(\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\delta_2,s)}{\delta_2}\right)^2\widehat{\delta}_N\cdot\frac{n^{1/2}\widehat{\delta}_N}{\widetilde{\Delta}},$$

and observe that whenever Theorem 33 holds, $\frac{n^{1/2}\widehat{\delta}_N}{\widetilde{\Delta}}\le c(\rho)$, hence the latter is bounded from above by

$$\widehat{\delta}_N\cdot C(\rho)\frac{\widetilde{\Delta}^2}{\Delta^2}\left(\frac{\widetilde{\Delta}}{\sqrt{N}}\frac{\widetilde{U}(\overline{\delta}_2,s)}{\overline{\delta}_2}\right)^2\le\frac{\widehat{\delta}_N}{7}$$

whenever $\Delta\ge\sigma(\ell,\mathcal{F})$ (so that $\widetilde{\Delta}=\Delta$) and $C_2(\rho)$ in the definition of $\overline{\delta}_2$ is large enough. Moreover,

$$C(\rho)\frac{\widetilde{\Delta}^3}{\Delta^3}\frac{\sigma^2(\ell,f_*)}{\Delta}\frac{n^{1/2}\,s}{N}\le C'(\rho)\cdot\sigma(\ell,f_*)\sqrt{n}\frac{s}{N}\le C'(\rho)\widetilde{\Delta}\sqrt{n}\frac{s}{N}$$

if $\widetilde{\Delta}\ge\sigma(\ell,f_*)$. As $\frac{s+\mathcal{O}}{k}\le c$ under the conditions of Theorem 33, $n^{3/2}\widetilde{\Delta}\frac{s^2+\mathcal{O}^2}{N^2}\le C\widetilde{\Delta}\sqrt{n}\frac{s+\mathcal{O}}{N}$. Combining the inequalities obtained above, we deduce on event $\Theta_3\cap\{\widehat{\delta}_N\ge\overline{\delta}_2\}$,

$$2\sup_{f\in\mathcal{F}(\widehat{\delta}_N)}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right|\le\frac{2\widehat{\delta}_N}{7}+C(\rho)\widetilde{\Delta}\left(\sqrt{n}\frac{s+\mathcal{O}}{N}\bigvee\frac{\sup_{f\in\mathcal{F}}\big(G_f(n,\Delta)\big)^2}{\sqrt{n}}\right)$$

whenever $\widetilde{\Delta}\ge\sigma(\ell,\mathcal{F})$. Finally, define

$$\overline{\delta}_4:=C_4(\rho)\widetilde{\Delta}\left(\sqrt{n}\frac{s+\mathcal{O}}{N}\bigvee\frac{\sup_{f\in\mathcal{F}}\big(G_f(n,\Delta)\big)^2}{\sqrt{n}}\right),$$

where $C_4(\rho)$ is sufficiently large. Then on event $\Theta_3\cap\Big\{\widehat{\delta}_N\ge\max\big(\overline{\delta}_2,7\overline{\delta}_4\big)\Big\}$,

$$2\sup_{f\in\mathcal{F}(\widehat{\delta}_N)}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right|+4\widetilde{\Delta}\sqrt{n}\frac{\mathcal{O}}{N}\le\frac{2\widehat{\delta}_N}{7}+\frac{\widehat{\delta}_N}{7}=\frac{3\widehat{\delta}_N}{7}.\tag{6.4.13}$$

Note that the expression above takes care of the term $4\widetilde{\Delta}\sqrt{n}\frac{\mathcal{O}}{N}$ that appeared in (6.4.11). Combining (6.4.9), (6.4.11), (6.4.13), we deduce that on event $\Theta_1\cap\Theta_2\cap\Theta_3\cap\Big\{\widehat{\delta}_N\ge\max\big(\overline{\delta}_1,\overline{\delta}_2,\overline{\delta}_3,7\,\overline{\delta}_4\big)\Big\}$,

$$\widehat{\delta}_N\le\frac{6}{7}\widehat{\delta}_N,$$

leading to a contradiction, hence on event $\Theta_1 \cap \Theta_2 \cap \Theta_3$ of probability at least $1 - 10e^{-s}$,

$$\widehat{\delta}_N \le \max\bigl(\overline{\delta}_1, \overline{\delta}_2, \overline{\delta}_3, 7\overline{\delta}_4\bigr). \tag{6.4.14}$$

Recall the definition (6.4.8) of $\overline{\delta}_1$. If condition 2 ("Bernstein condition") holds, then $\widetilde{\nu}(\delta) \le D\sqrt{\delta}$ for small enough δ, in which case

$$\overline{\delta}_1 \le C(\rho)D^2\left(\frac{s+\mathcal{O}}{N} + \frac{G^2_{f_*}(n,\Delta)}{n}\right),$$

where we used the fact that $\frac{s}{k} \le c$ by assumption. Together with the bound (6.2.1) for $G_{f_*}(n,\Delta)$, we deduce that, under the assumption that $R_4(\ell,\mathcal{F}) < \infty$,

$$\overline{\delta}_1 \le C(\rho)D^2\left(\frac{s+\mathcal{O}}{N} + \frac{\Bigl(\mathbb{E}\bigl|f_*(X) - \mathbb{E}f_*(X)\bigr|^3\Bigr)^2}{\Delta^6\, n^2}\right).$$

Since $\Delta = \sigma(\ell,\mathcal{F})M_\Delta$, $\frac{\mathbb{E}\left|f_*(X)-\mathbb{E}f_*(X)\right|^3}{\Delta^3} \le \frac{\sup_{f\in\mathcal{F}}\mathbb{E}\left|f(X)-\mathbb{E}f(X)\right|^3}{\sigma^3(\ell,\mathcal{F})M^3_\Delta} \le \frac{\mathfrak{B}^3(\ell,\mathcal{F})}{M^3_\Delta}$, where

$$\mathfrak{B}(\ell,\mathcal{F}) = \frac{\sup_{f\in\mathcal{F}}\mathbb{E}^{1/4}(\ell(f(X)) - \mathbb{E}\ell(f(X)))^4}{\sigma(\ell,\mathcal{F})},$$

hence

$$\overline{\delta}_1 \le C(\rho)D^2\left(\frac{s+\mathcal{O}}{N} + \frac{\mathfrak{B}^6(\ell,\mathcal{F})}{n^2 M^6_\Delta}\right). \tag{6.4.15}$$

At the same time, if only $\sigma(\ell,\mathcal{F}) < \infty$, we similarly obtain that

$$\overline{\delta}_1 \le C(\rho)D^2\left(\frac{s+\mathcal{O}}{N} + \frac{1}{M^4_\Delta\, n}\right). \tag{6.4.16}$$

Next we will estimate $\overline{\delta}_3$. Recall that, when $R_4(\ell,\mathcal{F}) < \infty$,

$$\widetilde{B}(\delta) = \frac{\mathfrak{B}^3(\ell,\mathcal{F})}{\sqrt{n}}\left(\frac{\widetilde{\nu}(\delta)}{\Delta}\frac{1}{M^2_\Delta} + \frac{\widetilde{\nu}_4(\delta)}{\Delta}\frac{1}{M^3_\Delta\sqrt{n}}\right).$$

For sufficiently small δ (namely, for which condition 2 holds) and $\Delta \ge \sigma(\ell,\mathcal{F})$,

$$\frac{\widetilde{\Delta}}{\sqrt{n}}\widetilde{B}(\delta) \le \frac{\mathfrak{B}^3(\ell,\mathcal{F})}{n}\left(\frac{\widetilde{\nu}(\delta)}{M^2_\Delta} + \frac{R_4(\ell,\mathcal{F})}{M^3_\Delta\sqrt{n}}\right) \le \frac{\mathfrak{B}^3(\ell,\mathcal{F})}{n}\left(D\frac{\sqrt{\delta}}{M^2_\Delta} + \sigma(\ell,\mathcal{F})\frac{\mathfrak{B}(\ell,\mathcal{F})}{M^3_\Delta\sqrt{n}}\right)$$

and

$$\overline{\delta}_3 \le C(\rho)\left(D^2\frac{\mathfrak{B}^6(\ell,\mathcal{F})}{n^2\, M^4_\Delta} + \sigma(\ell,\mathcal{F})\frac{\mathfrak{B}^4(\ell,\mathcal{F})}{n^{3/2}M^3_\Delta}\right). \tag{6.4.17}$$

At the same time, if only the second moments are finite, $\widetilde{B}(\delta) = \frac{\widetilde{\nu}(\delta)}{\Delta}\frac{1}{M_\Delta}$, and it is easy to deduce that in this case,

$$\overline{\delta}_3 \le C(\rho)\frac{D^2}{M^2_\Delta\, n}. \tag{6.4.18}$$

Next, we obtain a simpler bound for $\overline{\delta}_4$: as $\Delta \geq \sigma(\ell, \mathcal{F})$ by assumption, $\widetilde{\Delta} = \Delta = \sigma(\ell, \mathcal{F})\, M_\Delta$, and the estimate (6.2.1) for $G_{f_*}(n, \Delta)$ implies (if $R_4(\ell, \mathcal{F}) < \infty$) that

$$\overline{\delta}_4 \leq C(\rho)\, \sigma(\ell, \mathcal{F}) \left(\sqrt{n} M_\Delta \, \frac{s + \mathcal{O}}{N} + \frac{\mathfrak{B}^6(\ell, \mathcal{F})}{M_\Delta^5 n^{3/2}} \right). \tag{6.4.19}$$

If only $\sigma(\ell, \mathcal{F}) < \infty$, we similarly deduce from (6.2.1) that

$$\overline{\delta}_4 \leq C(\rho)\, \sigma(\ell, \mathcal{F}) \left(\sqrt{n} M_\Delta \cdot \frac{s + \mathcal{O}}{N} + \frac{1}{M_\Delta^3 \sqrt{n}} \right). \tag{6.4.20}$$

Finally, recall that $\widetilde{U}(\delta, s) = \frac{2}{\Delta}\big(c(\gamma)\, \widetilde{\omega}(\delta) + \widetilde{\nu}(\delta)\sqrt{\frac{s}{2}}\big) + \frac{32 s}{\sqrt{k}}$ and $\overline{\delta}_2 = \min\Big\{ \delta > 0 : \; C_2(\rho) \frac{\widetilde{\Delta}}{\sqrt{N}} \frac{\widetilde{U}(\delta, s)}{\delta} \leq \frac{1}{7} \Big\}$, hence

$$\overline{\delta}_2 \leq \overline{\delta} \bigvee C(\rho) D^2 \frac{s}{N} \bigvee C(\rho) \sigma(\ell, \mathcal{F}) \frac{s \sqrt{n} M_\Delta}{N}, \tag{6.4.21}$$

where $\overline{\delta}$ was defined in (6.2.5). Combining inequalities (6.4.15), (6.4.21), (6.4.17), (6.4.19) and (6.4.14), we obtain the final form of the bound under the stronger assumption $R_4(\ell, \mathcal{F}) < \infty$. Similarly, the combination of (6.4.16), (6.4.21), (6.4.18), (6.4.20) and (6.4.14) yields the bound under the weaker assumption $\sigma(\ell, \mathcal{F}) < \infty$.

6.4.3 Proof of Theorem 36.

Recall that $\widehat{\mathcal{E}}_N(f_*) := \widehat{\mathcal{L}}^{(k)}(f_*) - \widehat{\mathcal{L}}^{(k)}(\widehat{f}'_N)$ is the "empirical excess risk" of f_*, and let $\widehat{\delta}_N := \mathcal{E}(\widehat{f}'_N)$. It follows from Remark 6.4.2 that (using the notation used in the proof of Theorems 34 and 35)

$$\widehat{\mathcal{E}}_N(f_*) \leq \left| \frac{\widehat{G}_k\Big(0; \widehat{f}'_N\Big)}{\partial_z G\Big(0; \widehat{f}'_N\Big)} - \frac{\widehat{G}_k(0; f_*)}{\partial_z G(0; f_*)} \right| + 2 \sup_{f \in \mathcal{F}(\widehat{\delta}_N)} \left| \frac{R_N(f)}{\partial_z G_k(0; f)} \right|.$$

On the event of Theorem 35 of probability at least $1 - 10e^{-s}$,

$$\mathcal{E}(\widehat{f}'_N) \leq \delta' := \overline{\delta} + C(\rho)\big(D^2 \sigma(\ell, \mathcal{F}) \sqrt{n} M_\Delta\big) \left(\frac{\mathfrak{B}^6(\ell, \mathcal{F})}{M_\Delta^4 n^2} + \frac{s + \mathcal{O}}{N} \right),$$

hence on this event

$$\widehat{\mathcal{E}}_N(f_*) \leq \sup_{f \in \mathcal{F}(\delta')} \left| \frac{\widehat{G}_k(0; f)}{\partial_z G(0; f)} - \frac{\widehat{G}_k(0; f_*)}{\partial_z G(0; f_*)} \right| + 2 \sup_{f \in \mathcal{F}(\delta')} \left| \frac{R_N(f)}{\partial_z G_k(0; f)} \right| \leq \frac{6}{7} \delta'$$

where the last inequality again follows from main steps in the proof of Theorem 35; note that similar result holds if δ' is replaced by its analogue from Theorem 35. Consider the set $\widehat{\mathcal{F}}(\delta') = \Big\{ f \in \mathcal{F} : \; \widehat{\mathcal{E}}_N(f) \leq \delta' \Big\}$. First, observe that on the event $\mathcal{E}_1$ of Theorem 35, $f_* \in \widehat{\mathcal{F}}(\delta')$ as implied by the previous display. We will next show that $\widehat{\mathcal{F}}(\delta') \subseteq \mathcal{F}(7\delta')$ on the event $\mathcal{E}_1$ of Theorem 35, meaning that for any $f \in \widehat{\mathcal{F}}(\delta')$, $\mathcal{E}(f) \leq 7\delta'$. Indeed, let $f \in \widehat{\mathcal{F}}(\delta')$ be such that $\mathcal{E}(f) = \sigma$. Then (6.4.4) implies that

$$\begin{aligned} \mathcal{L}(f) - \mathcal{L}(f_*) &\leq \widehat{\mathcal{L}}^{(k)}(f) - \widehat{\mathcal{L}}^{(k)}(f_*) + \left| \frac{\widehat{G}_k(0; f)}{\partial_z G_k(0; f)} - \frac{\widehat{G}_k(0; f_*)}{\partial_z G_k(0; f_*)} \right| + \left| \frac{R_N(f)}{\partial_z G_k(0; f)} + \frac{R_N(f_*)}{\partial_z G_k(0; f_*)} \right| \\ &\leq \widehat{\mathcal{E}}_N(f) + \sup_{f \in \mathcal{F}(\sigma)} \left| \frac{\widehat{G}_k(0; f)}{\partial_z G_k(0; f)} - \frac{\widehat{G}_k(0; f_*)}{\partial_z G_k(0; f_*)} \right| + 2 \sup_{f \in \mathcal{F}(\sigma)} \left| \frac{R_N(f)}{\partial_z G_k(0; f)} \right|. \end{aligned}$$

Again, it follows from the arguments used in proof of Theorem 35 that on event $\mathcal{E}_1$ of probability at least $1-10e^{-s}$,

$$\sup_{f\in\mathcal{F}(\sigma)}\left|\frac{\widehat{G}_k(0;f)}{\partial_z G_k(0;f)}-\frac{\widehat{G}_k(0;f_*)}{\partial_z G_k(0;f_*)}\right|+2\sup_{f\in\mathcal{F}(\sigma)}\left|\frac{R_N(f)}{\partial_z G_k(0;f)}\right|\le\frac{6}{7}\max(\delta',\sigma).$$

Consequently, $\sigma\le\delta'+\frac{6}{7}\max(\delta',\sigma)$ on this event, implying that $\sigma\le 7\delta'$. Next, Assumption 2 yields that

$$\sup_{f\in\widehat{\mathcal{F}}(\delta')}\mathrm{Var}\Big(\ell(f(X))-\ell(\widehat{f}'_N)\Big)$$
$$\le 2\left(\sup_{f\in\widehat{\mathcal{F}}(\delta')}\mathrm{Var}(\ell(f(X))-\ell(f_*(X)))+\mathrm{Var}\Big(\ell(\widehat{f}'_N(X))-\ell(f_*(X))\Big)\right)\le 2D(\sqrt{7}+1)\delta'$$

on $\mathcal{E}_1$. It remains to apply Theorem 35, conditionally on $\mathcal{E}_1$, to the class

$$\widehat{\mathcal{F}}(\delta')-\widehat{f}'_N:=\Big\{f-\widehat{f}'_N,\ f\in\widehat{\mathcal{F}}(\delta')\Big\}.$$

To this end, we need to verify the assumption of Theorem 33 that translates into the requirement

$$c\Delta_2\ge\frac{1}{\sqrt{k_2}}\,\mathbb{E}\sup_{f\in\mathcal{F}(7\delta')}\frac{1}{\sqrt{|S_2|}}\sum_{j=1}^{|S_2|}(\ell(f(X_j))-\ell(f_*(X_j))-P(\ell(f)-\ell(f_*))).$$

As $\delta'>\overline{\delta}$ and $|S_2|\ge\lfloor N/2\rfloor$, we have the inequality

$$\mathbb{E}\sup_{f\in\mathcal{F}(7\delta')}\frac{1}{\sqrt{|S_2|}}\sum_{j=1}^{|S_2|}(\ell(f(X_j))-\ell(f_*(X_j))-P(\ell(f)-\ell(f_*)))\le C\delta'\sqrt{N},$$

hence it suffices to check that $\Delta_2=DM_{\Delta_2}\sqrt{7\delta'}\ge C\delta'\sqrt{\frac{N}{k_2}}$. The latter is equivalent to $\delta'\le CD^2M^2_{\Delta_2}\frac{k_2}{N}$ that holds by assumption. Result now follows easily as we assumed that the subsamples S_1 and S_2 used to construct $\widehat{f}'_N$ and $\widehat{f}''_N$ are disjoint.

Funding

Stanislav Minsker gratefully acknowledges support by the National Science Foundation [DMS-1712956 and CCF-1908905].

Supplementary material.

6.5 Remaining proofs.

6.5.1 Proof of Lemma 38.

As ρ is sufficiently smooth,

$$\partial_z G_k(0;f) = -\frac{\sqrt{kn}}{\Delta}\mathbb{E}\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right).$$

Let $W(\ell(f))$ denote a centered normal random variable variance equal to $\mathrm{Var}(\ell(f(X)))$. Lemma 34 implies that

$$\left|\mathbb{E}\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right) - \mathbb{E}\rho''\left(\frac{W(\ell(f))}{\Delta}\right)\right| \le C\frac{\|\rho^{(5)}\|_\infty}{\Delta^3\sqrt{n}}\mathbb{E}\left|\ell(f(X)) - P\ell(f)\right|^3.$$

Next, as $\rho''(x) \ge I\{|x|\le 1\}$ by assumption,

$$\mathbb{E}\rho''\left(\frac{W(\ell(f))}{\Delta}\right) \ge \Pr(|W(\ell(f))| \le \Delta).$$

Gaussian tail bound implies that

$$\Pr(|W(\ell(f))| \le \Delta) \ge 1 - 2\exp\left(-\frac{1}{2}\frac{\Delta^2}{\mathrm{Var}(\ell(f(X)))}\right) \ge \frac{1}{2}$$

whenever $\Delta^2 \ge 4\log(2)\mathrm{Var}(\ell(f(X)))$. On the other hand, if $x \sim N(0,1)$, then clearly $\Pr(Z \le |t|) \ge \frac{2|t|}{\sqrt{2\pi}}e^{-t^2/2}$, hence

$$\Pr(|W(\ell(f))| \le \Delta) \ge \frac{2\Delta}{\sqrt{2\pi\mathrm{Var}(\ell(f(X)))}}\exp\left(-\frac{1}{2}\frac{\Delta^2}{\mathrm{Var}(\ell(f(X)))}\right) \ge \frac{\Delta}{\sqrt{8\pi\mathrm{Var}(\ell(f(X)))}}$$

whenever $\Delta^2 < 4\log(2)\mathrm{Var}(\ell(f(X)))$. Combination of two bounds yields that

$$\Pr(|W(\ell(f))| \le \Delta) \ge \frac{1}{2\sqrt{2\pi}}\min\left(\frac{\Delta}{\sqrt{\mathrm{Var}(\ell(f(X)))}}, 2\sqrt{\log 2}\right).$$

6.5.2 Proof of Lemma 39.

Observe that

$$\begin{aligned}\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right) &= \frac{1}{\sqrt{k}}\sum_{j\in J}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right) + \frac{1}{\sqrt{k}}\sum_{j\notin J}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)\\ &\le \sqrt{\frac{|J|}{k}}\frac{1}{\sqrt{|J|}}\sum_{j\in J}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right) + 2\frac{\mathcal{O}}{\sqrt{k}},\end{aligned}$$

where we used the fact that $\|\rho'\|_\infty \le 2$. Bernstein's inequality implies that

$$\left|\frac{1}{\sqrt{|J|}}\left(\sum_{j\in J}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)-\mathbb{E}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)\right)\right| \le 2\left(\mathrm{Var}^{1/2}\left(\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)\right)\sqrt{s}+\frac{2s}{\sqrt{|J|}}\right)$$

with probability at least $1-2e^{-s}$, where we again used the fact that $\|\rho'\|_\infty \le 2$. Moreover,

$$\mathrm{Var}\left(\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)\right) \le \frac{\sigma^2(\ell,f)}{\Delta^2}$$

by Lemma 33, hence with the same probability

$$|\widehat{G}_k(0;f)| \le \sqrt{k}\left|\mathbb{E}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)\right| + 2\left(\frac{\sigma(\ell,f)}{\Delta}\sqrt{s}+\frac{2s}{\sqrt{k}}+\frac{\mathcal{O}}{\sqrt{k}}\right).$$

Lemma 6.2 in [Min18] implies that

$$\left|\mathbb{E}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)\right| \le \underbrace{\mathbb{E}\rho'\left(\frac{W(\ell(f))}{\Delta}\right)}_{=0} + 2G_f(n,\Delta),$$

hence the claim follows.

6.5.3 Proof of Lemma 40.

Since F is smooth, for any $x,y\in\mathbb{R}$, $F(y)-F(x)=\int_0^1 F'(x+\alpha(y-x))d\alpha\cdot(y-x)$. Let $S_n^x=\sum_{j=1}^n x_j$, $S_n^\eta=\sum_{j=1}^n \eta_j$. Then

$$F(S_n^x)-F(S_n^\eta)=(S_n^x-S_n^\eta)\int_0^1 F'(S_n^\eta+\alpha(S_n^x-S_n^\eta))d\alpha,$$

hence

$$\mathbb{E}(F(S_n^x)-F(S_n^\eta))=\int_0^1 \mathbb{E}[(S_n^x-S_n^\eta)F'(S_n^\eta+\alpha(S_n^x-S_n^\eta))]d\alpha.$$

Hö's inequality yields that

$$\begin{aligned}\left|\mathbb{E}(S_n^x-S_n^\eta)F'(S_n^\eta+\alpha(S_n^x-S_n^\eta))\right| &\le \left(\mathbb{E}|S_n^x-S_n^\eta|^2\right)^{1/2}\left(\mathbb{E}|F'(S_n^\eta+\alpha(S_n^x-S_n^\eta))|^2\right)^{1/2}\\ &\le \sqrt{n}\,\mathrm{Var}^{1/2}(x_1-\eta_1)\left(\mathbb{E}|F'(S_n^\eta+\alpha(S_n^x-S_n^\eta))|^2\right)^{1/2},\end{aligned}$$

implying the first inequality. The rest of the proof is devoted to the second inequality of the lemma. Let (W,Z) be a centered Gaussian vector with the same covariance as (x_1,η_1), and let

$(W_1, Z_1), \dots, (W_n, Z_n)$ be i.i.d. copies of (W, Z). We also set $S_n^W = \sum_{j=1}^n W_j$, $S_n^Z = \sum_{j=1}^n Z_j$. As $\mathbb{E}F\big(S_n^W\big) = \mathbb{E}F\big(S_n^Z\big) = 0$ for bounded odd F, it is easy to see that

$$\begin{aligned}|\mathbb{E}(F(S_n^x) - F(S_n^\eta))| &= \big|\mathbb{E}(F(S_n^x) - F(S_n^\eta)) - \mathbb{E}\big(F\big(S_n^W\big) - F\big(S_n^Z\big)\big)\big| \\ &= \Big|\int_0^1 \Big(\mathbb{E}(S_n^x - S_n^\eta)F'(S_n^\eta + \alpha(S_n^x - S_n^\eta)) - \mathbb{E}\big(S_n^W - S_n^Z\big)F'\big(S_n^Z + \alpha\big(S_n^W - S_n^Z\big)\big)\Big)d\alpha\Big| \\ &\le \int_0^1 \Big|\mathbb{E}(S_n^x - S_n^\eta)F'(S_n^\eta + \alpha(S_n^x - S_n^\eta)) - \mathbb{E}\big(S_n^W - S_n^Z\big)F'\big(S_n^Z + \alpha\big(S_n^W - S_n^Z\big)\big)\Big|d\alpha.\end{aligned}$$

Next we will estimate, for each $\alpha \in [0,1]$, the expression

$$\Big|\mathbb{E}(S_n^x - S_n^\eta)F'(S_n^\eta + \alpha(S_n^x - S_n^\eta)) - \mathbb{E}\big(S_n^W - S_n^Z\big)F'\big(S_n^Z + \alpha\big(S_n^W - S_n^Z\big)\big)\Big|. \tag{6.5.1}$$

To this end, we will use Lindeberg's replacement method. For $i = 0, \dots, n$, denote

$$T_i = (x_1 - \eta_1, \dots, x_i - \eta_i, W_{i+1} - Z_{i+1}, \dots, W_n - Z_n, \eta_1, \dots, \eta_i, Z_{i+1}, \dots, Z_n).$$

Then the expression in (6.5.1) is equal to $|\mathbb{E}G(T_n) - \mathbb{E}G(T_0)|$, where

$$G(T) = \left(\sum_{i=1}^n T^{(i)}\right) F'\left(\sum_{j=1}^n \Big(T^{(j+n)} + \alpha T^{(j)}\Big)\right)$$

and $T^{(j)}$ stands for the j-th coordinate of T. Clearly,

$$|\mathbb{E}G(T_n) - \mathbb{E}G(T_0)| \le \sum_{i=1}^n |\mathbb{E}G(T_i) - \mathbb{E}G(T_{i-1})|. \tag{6.5.2}$$

Fix i, and consider the Taylor expansions of $G(T_i)$ and $G(T_{i-1})$ at the point

$$T_i^0 = (x_1 - \eta_1, \dots, x_{i-1} - \eta_{i-1}, 0, W_{i+1} - Z_{i+1}, \dots, W_n - Z_n, \eta_1, \dots, \eta_{i-1}, 0, Z_{i+1}, \dots, Z_n)$$

(note that T_i^0 does not depend on x_i, η_i, W_i and Z_i). For $G(T_i)$ we get, setting $\delta_i = x_i - \eta_i$ and using $\partial^{(m)}_{i_1,\dots,i_m}$ to denote the m-th order partial derivative with respect to the $i_1, \dots, i_m$-th variables, that

$$\begin{aligned}G(T_i) &= G(T_i^0) + \partial_i G(T_i^0)\cdot\delta_i + \partial_{n+i}G(T_i^0)\cdot\eta_i \\ &\quad + \frac{1}{2}\big(\partial^2_{i,i}G(T_i^0)\cdot\delta_i^2 + 2\partial^2_{i,n+i}G(T_i^0)\cdot\delta_i\eta_i + \partial^2_{n+i,n+i}G(T_i^0)\cdot\eta_i^2\big) \\ &\quad + \frac{1}{6}\Big(\partial^3_{i,i,i}G(\tilde T_i^0)\cdot\delta_i^3 + \partial^3_{n+i,n+i,n+i}G(\tilde T_i^0)\cdot\eta_i^3 + \partial^3_{n+i,n+i,i}G(\tilde T_i^0)\cdot\eta_i^2\delta_i + \partial^3_{n+i,i,i}G(\tilde T_i^0)\cdot\eta_i\delta_i^2\Big),\end{aligned}$$

where $\tilde T_i^0$ is a point on a line segment between T_i^0 and T_i. Similarly, setting $\Delta_i = W_i - Z_i$,

$$\begin{aligned}G(T_{i-1}) &= G(T_i^0) + G(T_i^0) + \partial_i G(T_i^0)\cdot\Delta_i + \partial_{n+i}G(T_i^0)\cdot Z_i \\ &\quad + \frac{1}{2}\big(\partial^2_{i,i}G(T_i^0)\cdot\Delta_i^2 + \partial^2_{i,n+i}G(T_i^0)\cdot\Delta_i Z_i + \partial^2_{n+i,n+i}G(T_i^0)\cdot Z_i^2\big) \\ &\quad + \frac{1}{6}\Big(\partial^3_{i,i,i}G(\tilde T_i^0)\cdot\Delta_i^3 + \partial^3_{n+i,n+i,n+i}G(\tilde T_i^0)\cdot Z_i^3 + 3\partial^3_{n+i,n+i,i}G(\tilde T_i^0)\cdot Z_i^2\Delta_i + 3\partial^3_{n+i,i,i}G(\tilde T_i^0)\cdot Z_i\Delta_i^2\Big),\end{aligned} \tag{6.5.3}$$

where $\bar{T}_i^0$ is a point on a line segment between T_i^0 and T_{i-1}. Using independence of T_i^0 and (x_i, η_i, W_i, Z_i) and the fact that covariance structures of (x_i, η_i) and (W, Z) are the same, we deduce that

$$|\mathbb{E}G(T_i) - \mathbb{E}G(T_{i-1})| \le \frac{1}{6}\mathbb{E}\Big|\partial^3_{i,i,i}G(\tilde{T}_i^0)\cdot\delta_i^3 + \partial^3_{n+i,n+i,n+i}G(\tilde{T}_i^0)\cdot\eta_i^3 + 3\partial^3_{n+i,n+i,i}G(\tilde{T}_i^0)\cdot\eta_i^2\delta_i + 3\partial^3_{n+i,i,i}G(\tilde{T}_i^0)\cdot\eta_i\delta_i^2\Big|$$
$$+\frac{1}{6}\mathbb{E}\Big|\partial^3_{i,i,i}G(\tilde{T}_i^0)\cdot\Delta_i^3+\partial^3_{n+i,n+i,n+i}G(\tilde{T}_i^0)\cdot Z_i^3+3\partial^3_{n+i,n+i,i}G(\tilde{T}_i^0)\cdot Z_i^2\Delta_i+3\partial^3_{n+i,i,i}G(\tilde{T}_i^0)\cdot Z_i\Delta_i^2\Big|.$$

It remains to estimate each of the terms above. Assume that $\tau \in [0,1]$ is such that

$$\tilde{T}_i^0 = (x_1-\eta_1, \dots, x_{i-1}-\eta_{i-1}, \tau(x_i-\eta_i), W_{i+1}-Z_{i+1}, \dots, W_n-Z_n, \eta_1, \dots, \eta_{i-1}, \tau\eta_i, Z_{i+1}, \dots, Z_n).$$

1. Direct computation implies that

$$\partial^3_{i,i,i}G(\tilde{T}_i^0) = 3\alpha^2F'''\left(\sum_{j\neq i}\Big(\eta_j+\alpha\delta_j\Big)+\tau(\eta_i+\alpha\delta_i)\right) + \alpha^3F''''\left(\sum_{j\neq i}\Big(\eta_j+\alpha\delta_j\Big)+\tau(\eta_i+\alpha\delta_i)\right)\Big(\sum_{j\neq i}\delta_j+\tau\delta_i\Big),$$

hence

$$\mathbb{E}\Big|\partial^3_{i,i,i}G(\tilde{T}_i^0)\cdot\delta_i^3\Big| \le 3\alpha^2\|F'''\|_\infty\mathbb{E}|\delta_i^3| + \alpha^3\|F''''\|_\infty\left(\mathbb{E}\Big|\sum_{j\neq i}\delta_j\Big|\,\mathbb{E}|\delta_i|^3 + \mathbb{E}|\delta_i|^4\right)$$
$$\le 3\alpha^2\|F'''\|_\infty\big(\mathbb{E}\delta_i^2\big)^{1/2}\big(\mathbb{E}\delta_i^4\big)^{1/2} + \alpha^3\|F''''\|_\infty\left(\sqrt{\sum_{j\neq i}\mathbb{E}\delta_j^2}\,\big(\mathbb{E}\delta_i^2\big)^{1/2}\big(\mathbb{E}\delta_i^4\big)^{1/2} + \mathbb{E}|\delta_i|^4\right), \tag{6.5.4}$$

where we used Hölder's inequality in the last step.

2. Next,

$$\partial^3 G_{\eta_i,\eta_i,\eta_i}(\tilde{T}_i^0) = F''''\left(\sum_{j\neq i}\Big(\eta_j+\alpha\delta_j\Big)+\tau(\eta_i+\alpha\delta_i)\right)\Big(\sum_{j\neq i}\delta_j+\tau\delta_i\Big),$$

hence Hölder's inequality, together with the identity $\|F''''\|_\infty = M^{-3}\|H''''\|_\infty$, imply that

$$\mathbb{E}\Big|\partial^3_{n+i,n+i,n+i}G(\tilde{T}_i^0)\cdot\eta_i^3\Big| \le \|F''''\|_\infty\left(\mathbb{E}|\eta_i|^3\mathbb{E}\left|\sum_{j\neq i}\delta_j\right| + \mathbb{E}|\delta_i\eta_i^3|\right)$$
$$\le \|F''''\|_\infty\left(\mathbb{E}|\eta_i|^3\sqrt{\sum_{j\neq i}\mathbb{E}\delta_j^2} + \big(\mathbb{E}\delta_i^4\big)^{1/4}\big(\mathbb{E}\eta_i^4\big)^{3/4}\right). \tag{6.5.5}$$

3. Proceeding in a similar fashion, we deduce that

$$\partial^3 G_{n+i,n+i,i}(\tilde{T}_i^0) = F'''\left(\sum_{j\neq i}\Big(\eta_j+\alpha\delta_j\Big)+\tau(\eta_i+\alpha\delta_i)\right) + \alpha F''''\left(\sum_{j\neq i}\Big(\eta_j+\alpha\delta_j\Big)+\tau(\eta_i+\alpha\delta_i)\right)\Big(\sum_{j\neq i}\delta_j+\tau\delta_i\Big),$$

so that, applying Hölder's inequality, we obtain

$$\begin{aligned}\mathbb{E}\Big|\partial^3_{n+i,n+i,i}G(\tilde{T}_i^0)\cdot\eta_i^2\delta_i\Big| &\leq \|F'''\|_\infty\big(\mathbb{E}\eta_i^4\big)^{1/2}\big(\mathbb{E}\delta_i^2\big)^{1/2}+\alpha\|F''''\|_\infty\mathbb{E}\Big|\eta_i^2\delta_i\Big(\sum_{j\neq i}\delta_j+\tau\delta_i\Big)\Big| \\ &\leq \|F'''\|_\infty\big(\mathbb{E}\eta_i^4\big)^{1/2}\big(\mathbb{E}\delta_i^2\big)^{1/2}+\alpha\|F''''\|_\infty\left(\sqrt{\sum_{j\neq i}\mathbb{E}\delta_j^2}\big(\mathbb{E}\eta_i^4\big)^{1/2}\big(\mathbb{E}\delta_i^2\big)^{1/2}+\sqrt{\mathbb{E}\delta_i^4\,\mathbb{E}\eta_i^4}\right).\end{aligned} \tag{6.5.6}$$

4. Finally,

$$\partial^3 G_{n+i,i,i}(\tilde{T}_i^0) = 2\alpha F'''\left(\sum_{j\neq i}\Big(\eta_j+\alpha\delta_j\Big)+\tau(\eta_i+\alpha\delta_i)\right) + \alpha^2 F''''\left(\sum_{j\neq i}\Big(\eta_j+\alpha\delta_j\Big)+\tau(\eta_i+\alpha\delta_i)\right)\Big(\sum_{j\neq i}\delta_j+\tau\delta_i\Big).$$

Hölder's inequality implies that $\mathbb{E}|\eta_i\delta_i^2| = \mathbb{E}|\eta_i\delta_i\delta_i| \leq \big(\mathbb{E}\delta_i^2\big)^{1/2}\big(\mathbb{E}\delta_i^4\big)^{1/4}\big(\mathbb{E}\eta_i^4\big)^{1/4}$, hence

$$\begin{aligned}\Big|\mathbb{E}\partial^3_{n+i,i,i}G(\tilde{T}_i^0)\cdot\eta_i\delta_i^2\Big| &\leq 2\alpha\|F'''\|_\infty\big(\mathbb{E}\delta_i^2\big)^{1/2}\big(\mathbb{E}\delta_i^4\big)^{1/4}\big(\mathbb{E}\eta_i^4\big)^{1/4}+\alpha^2\|F''''\|_\infty\mathbb{E}\Big|\eta_i\delta_i^2\Big(\sum_{j\neq i}\delta_j+\tau\delta_i\Big)\Big| \\ &\leq 2\alpha\|F'''\|_\infty\big(\mathbb{E}\delta_i^2\big)^{1/2}\big(\mathbb{E}\delta_i^4\big)^{1/4}\big(\mathbb{E}\eta_i^4\big)^{1/4} \\ &\quad+\alpha^2\|F''''\|_\infty\left(\sqrt{\sum_{j\neq i}\mathbb{E}\delta_j^2}\big(\mathbb{E}\delta_i^2\big)^{1/2}\big(\mathbb{E}\delta_i^4\big)^{1/4}\big(\mathbb{E}\eta_i^4\big)^{1/4}+\big(\mathbb{E}\delta_i^4\big)^{3/4}\big(\mathbb{E}\eta_i^4\big)^{1/4}\right).\end{aligned} \tag{6.5.7}$$

Similar calculations yield an analogous bound for the terms in the expansion (6.5.3) of $G(T_{i-1})$. The equivalence of the moments of Gaussian random variables together with the fact that the covariance structure of (W,Z) matches that of (x_1,η_1) imply that the upper bounds (6.5.4), (6.5.5), (6.5.6), (6.5.7) remain valid for the terms in (6.5.3), up to an additional absolute multiplicative constant. Hence, combination of (6.5.2), (6.5.4), (6.5.5), (6.5.6), (6.5.7) and straightforward application of Hölder's inequality yields the result.

6.5.4 Proof of Lemma 41.

Define

$$D(\delta) := \sup_{\ell(f)\in\mathcal{F}(\delta)}\mathbb{E}^{1/2}\left(\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)-\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right)^2.$$

Recall that ρ' is Lipschitz continuous and $L(\rho') = 1$, hence

$$\left(\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f) - \mathcal{L}(f)}{\Delta}\right) - \rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f_*) - \mathcal{L}(f_*)}{\Delta}\right)\right)^2 \leq \left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f) - \overline{\mathcal{L}}_1(f_*) - (\mathcal{L}(f) - \mathcal{L}(f_*))}{\Delta}\right)^2, \quad (6.5.8)$$

which implies that

$$D(\delta) \leq \frac{\nu(\delta)}{\Delta}. \quad (6.5.9)$$

Next, observe that $\widehat{G}_k(0; f) = \frac{1}{\sqrt{k}}\sum_{j\in J}\rho'\Big(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\Big) + \frac{1}{\sqrt{k}}\sum_{j\notin J}\rho'\Big(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\Big)$, hence application of the triangle inequality yields that

$$\sup_{f\in\mathcal{F}(\delta)}\Big|\widehat{G}_k(0; f) - \widehat{G}_k(0; f_*)\Big| \leq \sup_{f\in\mathcal{F}(\delta)}|G_k(0; f) - G_k(0; f_*)| + \sqrt{\frac{|J|}{k}}\sup_{f\in\mathcal{F}(\delta)}\Big|\widehat{G}_{|J|}(0; f) - \widehat{G}_{|J|}(0; f_*) - \mathbb{E}\Big(\widehat{G}_{|J|}(0; f) - \widehat{G}_{|J|}(0; f_*)\Big)\Big| + 4\frac{\mathcal{O}}{\sqrt{k}}, \quad (6.5.10)$$

where $\widehat{G}_{|J|}(0; f) := \frac{1}{\sqrt{|J|}}\sum_{j\in J}\rho'\Big(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\Big)$. Talagrand's concentration inequality (specifically, the bound of Lemma 37) implies, together with the inequalities $\|\rho'\|_\infty \leq 2$ and $|J| > k/2$, that for any $s > 0$

$$\sup_{f\in\mathcal{F}(\delta)}\Big|\widehat{G}_{|J|}(0; f) - \widehat{G}_{|J|}(0; f_*) - \mathbb{E}\Big(\widehat{G}_{|J|}(0; f) - \widehat{G}_{|J|}(0; f_*)\Big)\Big| \leq 2\left[\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\Big|\widehat{G}_{|J|}(0; f) - \widehat{G}_{|J|}(0; f_*) - \mathbb{E}\Big(\widehat{G}_{|J|}(0; f) - \widehat{G}_{|J|}(0; f_*)\Big)\Big| + D(\delta)\sqrt{\frac{s}{2}} + \frac{32\sqrt{2}s}{3\sqrt{k}}\right]$$

with probability at least $1 - 2e^{-s}$. According to (6.5.9), $D(\delta) \leq \frac{L(\rho')}{\Delta}\nu(\delta)$. Hence, it remains to estimate the expected supremum. Sequential application of symmetrization, contraction and desymmetrization inequalities implies that

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\Big|\widehat{G}_{|J|}(0; f) - \widehat{G}_{|J|}(0; f_*) - \mathbb{E}\Big(\widehat{G}_{|J|}(0; f) - \widehat{G}_{|J|}(0; f_*)\Big)\Big| \\
&\quad\leq 2\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\frac{1}{\sqrt{|J|}}\sum_{j\in J}\epsilon_j\left(\rho'\left(\sqrt{n}\,\frac{\overline{\mathcal{L}}_j(f) - \mathcal{L}(f)}{\Delta}\right)\right) - \rho'\left(\sqrt{n}\,\frac{\overline{\mathcal{L}}_j(f_*) - \mathcal{L}(f_*)}{\Delta}\right)\right| \\
&\quad\leq \frac{4L(\rho')}{\Delta}\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\frac{\sqrt{n}}{\sqrt{|J|}}\sum_{j\in|J|}\epsilon_j\big((\overline{\mathcal{L}}_j(f) - \mathcal{L}(f))(X_j) - (\overline{\mathcal{L}}_j(f_*) - \mathcal{L}(f_*))(X_j)\big)\right| \\
&\leq \frac{8\sqrt{2}L(\rho')}{\Delta}\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^{N_J}\Big((\ell(f) - \ell(f_*))(X_j) - P(\ell(f) - \ell(f_*))\Big)\right| \leq \frac{8\sqrt{2}}{\Delta}\,\omega(\delta) \quad (6.5.11)
\end{aligned}$$

since $L(\rho') = 1$. To estimate $\sup_{f\in\mathcal{F}(\delta)}|G_k(0; f) - G_k(0; f_*)|$, we consider 2 cases: the first case when only 2 finite moments of $\ell(f(X))$, $f \in \mathcal{F}$ exist, and the second case when 4 moments are

finite. To obtain the bound in the first case, we observe that, since $\mathbb{E}\Big(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\Big)=0$ for any $f\in\mathcal{F}$,

$$\left|\mathbb{E}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)-\mathbb{E}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right| = \left|\mathbb{E}T\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)-\mathbb{E}T\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right|$$

where $T(x)=x-\rho'(x)$. Next, we apply Lemma 40 with $F=T$, $x_j=\frac{\ell(f(X_j)-\mathbb{E}\ell(f(X_j))}{\Delta\sqrt{n}}$ and $\eta_j=\frac{\ell(f_*(X_j)-\mathbb{E}\ell(f_*(X_j))}{\Delta\sqrt{n}}$. The first inequality of the lemma implies that

$$\left|\mathbb{E}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)-\mathbb{E}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right| \le \sqrt{\mathrm{Var}\left(\frac{\ell(f(X))-\ell(f_*(X))}{\Delta}\right)} \times \max_{\alpha\in[0,1]}\sqrt{\mathbb{E}\left(T'\left(\alpha\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}+(1-\alpha)\sqrt{n}\frac{\overline{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right)^2}.$$

Observe that $T'(x)=1-\rho''(x)\le I\{|x|\ge 1\}$ by Assumption 1. It implies that for any $\alpha\in[0,1]$,

$$\mathbb{E}\left(T'\left(\alpha\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}+(1-\alpha)\sqrt{n}\frac{\overline{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right)^2 \le \Pr\left(\left|\alpha\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}+(1-\alpha)\sqrt{n}\frac{\overline{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right|\ge 1\right) \le \sup_{f\in\mathcal{F}}\mathrm{Var}\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)=\sup_{f\in\mathcal{F}}\frac{\sigma^2(\ell,f)}{\Delta^2}.$$

by Chebyshev's inequality. Hence

$$\left|\mathbb{E}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)-\mathbb{E}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right| \le \mathrm{Var}^{1/2}(\ell(f(X))-\ell(f_*(X)))\frac{\sigma(\ell,\mathcal{F})}{\Delta^2}.$$

and, taking supremum over $f\in\mathcal{F}(\delta)$ and recalling that $\Delta=M_\Delta\cdot\sigma(\ell,\mathcal{F})$ for $M_\Delta\ge 1$, we obtain the inequality

$$\sup_{f\in\mathcal{F}(\delta)}|G_k(0;f)-G_k(0;f_*)|\le\sqrt{k}\frac{\nu(\delta)}{\Delta}\frac{1}{M_\Delta}\le\sqrt{k}\,\widetilde{B}(\delta).$$

On the other hand, under the assumption of existence of 4 moments, we get that

$$\left|\mathbb{E}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right)-\mathbb{E}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right| \le \frac{C(\rho)}{\sqrt{n}\Delta}\left(\mathrm{Var}^{1/2}(\ell(f(X))-\ell(f_*(X)))\left(\frac{R_4^2(\ell,\mathcal{F})}{\Delta^2}+\frac{R_4^3(\ell,\mathcal{F})}{\Delta^3}\right)+\frac{\mathbb{E}^{1/4}(\ell(f(X))-\ell(f_*(X)))^4}{\sqrt{n}}\frac{R_4^3(\ell,\mathcal{F})}{\Delta^3}\right),$$

Again, taking supremum over $f \in \mathcal{F}(\delta)$ and recalling that $\Delta = M_\Delta \cdot \sigma(\ell, \mathcal{F})$ for $M_\Delta \geq 1$, we deduce that

$$\sup_{f\in\mathcal{F}(\delta)} |G_k(0;f) - G_k(0;f_*)| \leq C(\rho)\sqrt{\frac{k}{n}}\left(\frac{\nu(\delta)}{\Delta}\left(\frac{\mathfrak{B}^3(\ell,\mathcal{F})}{M_\Delta^3} \vee \frac{\mathfrak{B}^2(\ell,\mathcal{F})}{M_\Delta^2}\right) + \frac{\nu_4(\delta)}{\Delta}\frac{\mathfrak{B}^3(\ell,\mathcal{F})}{M_\Delta^3\sqrt{n}}\right)$$
$$\leq C(\rho)\sqrt{\frac{k}{n}}\mathfrak{B}^3(\ell,\mathcal{F})\left(\frac{\nu(\delta)}{\Delta}\frac{1}{M_\Delta^2} + \frac{\nu_4(\delta)}{\Delta}\frac{1}{M_\Delta^3\sqrt{n}}\right) \leq C(\rho)\sqrt{k}\,\widetilde{B}(\delta), \quad (6.5.12)$$

implying the result.

6.5.5 Proof of Lemma 42.

Recall that $\widehat{G}_{|J|}(0;f) := \frac{1}{\sqrt{|J|}}\sum_{j\in J}\rho'\Big(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\Big)$. Given $\delta \geq \delta_{\min}$, define

$$\widehat{Q}_{|J|}(\delta) := \sup_{f\in\mathcal{F}(\delta)} \frac{\delta_{\min}}{\delta}\Big|\widehat{G}_{|J|}(0;f) - \widehat{G}_{|J|}(0;f_*)\Big|,$$
$$\widehat{T}_{|J|}(\delta_{\min}) := \sup_{\delta\geq\delta_{\min}} \widehat{Q}_{|J|}(\delta).$$

Observe that for any $\delta \geq \delta_{\min}$,

$$\sup_{f\in\mathcal{F}(\delta)}\Big|\widehat{G}_{|J|}(0;f) - \widehat{G}_{|J|}(0;f_*)\Big| \leq \frac{\delta}{\delta_{\min}}\widehat{T}_{|J|}(\delta_{\min}). \quad (6.5.13)$$

Hence, our goal will be to find an upper bound for $\widehat{T}_{|J|}(\delta_{\min})$. To this end, note that

$$\widehat{T}_{|J|}(\delta_{\min}) \leq \sup_{\delta\geq\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\Big|\widehat{G}_{|J|}(0;f) - \widehat{G}_{|J|}(0;f_*) - \mathbb{E}\Big(\widehat{G}_{|J|}(0;f) - \widehat{G}_{|J|}(0;f_*)\Big)\Big|$$
$$+ \sup_{\delta\geq\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}|G_k(0;f) - G_k(0;f_*)|. \quad (6.5.14)$$

It remains to estimate both terms in the inequality above. Inequality (6.5.8) implies the bound

$$\sup_{\delta\geq\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\mathrm{Var}^{1/2}\left(\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right) - \rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right)$$
$$\leq \frac{L(\rho')}{\Delta}\sup_{\delta\geq\delta_{\min}}\frac{\delta_{\min}}{\delta}\nu(\delta) \leq \frac{L(\rho')}{\Delta}\sup_{\delta\geq\delta_{\min}}\frac{\delta_{\min}}{\delta}\widetilde{\nu}(\delta) \leq \frac{1}{\Delta}\widetilde{\nu}(\delta_{\min})$$

since $\widetilde{\nu}$ is a function of concave type. Moreover, it is clear that for any $\delta \geq \delta_{\min}$,

$$\left|\frac{\delta_{\min}}{\delta}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f)-\mathcal{L}(f)}{\Delta}\right) - \frac{\delta_{\min}}{\delta}\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_1(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right| \leq 2\|\rho'\|_\infty \leq 4$$

almost surely. Now, Talagrand's concentration inequality implies that for any $s > 0$,

$$\sup_{\delta\ge\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\Big|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\Big(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\Big)\Big|$$
$$\le 2\Big[\mathbb{E}\sup_{\delta\ge\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\Big|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\Big(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\Big)\Big|$$
$$+\frac{L(\rho')}{\Delta}\widetilde{\nu}(\delta_{\min})\sqrt{\frac{s}{2}}+\frac{32\sqrt{2}s}{3\sqrt{k}}\Big] \quad (6.5.15)$$

with probability at least $1-e^{-s}$. To estimate the expectation, we proceed as follows: for $j\in\mathbb{Z}$, set $\delta_j := 2^{-j}$, and observe that

$$\mathbb{E}\sup_{\delta\ge\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\Big|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\Big(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\Big)\Big|$$
$$\le \mathbb{E}\sup_{j:\delta_j\ge\delta_{\min}}\sup_{\delta\in(\delta_{j+1},\delta_j]}\frac{\delta_{\min}}{\delta}\sup_{f\in\mathcal{F}(\delta)}\Big|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\Big(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\Big)\Big|$$
$$\le \sum_{j:\delta_j\ge\delta_{\min}}\frac{\delta_{\min}}{\delta_{j+1}}\mathbb{E}\sup_{\delta\in(\delta_{j+1},\delta_j]}\sup_{f\in\mathcal{F}(\delta)}\Big|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\Big(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\Big)\Big|$$
$$\le 2\sum_{j:\delta_j\ge\delta_{\min}}\frac{\delta_{\min}}{\delta_j}\mathbb{E}\sup_{f\in\mathcal{F}(\delta_j)}\Big|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\Big(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\Big)\Big|,$$

where the last inequality relied on the fact that $\mathcal{F}(\delta)\subseteq\mathcal{F}(\delta')$ for $\delta\le\delta'$. It follows from (6.5.11) that

$$\mathbb{E}\sup_{f\in\mathcal{F}(\delta_j)}\Big|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\Big(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\Big)\Big|\le\frac{8\sqrt{2}L(\rho')}{\Delta}\omega(\delta_j)\le\frac{8\sqrt{2}}{\Delta}\widetilde{\omega}(\delta_j),$$

where $\widetilde{\omega}(\cdot)$ is an upper bound on $\omega(\cdot)$ of strictly concave type (with exponent γ for some $\gamma\in(0,1)$). Hence, applying Proposition 4.2 in [Kol11], we deduce that

$$\mathbb{E}\sup_{\delta\ge\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\Big|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\Big(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\Big)\Big|$$
$$\le\frac{16}{\Delta}\delta_{\min}\sum_{j:\delta_j\ge\delta_{\min}}\frac{\widetilde{\omega}(\delta_j)}{\delta_j}\le\frac{c(\gamma)}{\Delta}\delta_{\min}\frac{\widetilde{\omega}(\delta_{\min})}{\delta_{\min}}=\frac{c(\gamma)}{\Delta}\widetilde{\omega}(\delta_{\min}),$$

and (6.5.15) yields the inequality

$$\sup_{\delta\ge\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}\Big|\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)-\mathbb{E}\Big(\widehat{G}_{|J|}(0;f)-\widehat{G}_{|J|}(0;f_*)\Big)\Big|\le\widetilde{U}(\delta_{\min},s), \quad (6.5.16)$$

where $\widetilde{U}(\delta,s)$ was defined in (6.4.10). For the second term in (6.5.14), inequality (6.5.12) implies that

$$\sup_{\delta\ge\delta_{\min}}\sup_{f\in\mathcal{F}(\delta)}\frac{\delta_{\min}}{\delta}|G_k(0;f)-G_k(0;f_*)|$$
$$\le C(\rho)\delta_{\min}\sqrt{\frac{k}{n}}\overline{R}^3(\ell,\mathcal{F},\Delta)\sup_{\delta\ge\delta_{\min}}\left(\frac{\nu(\delta)}{\delta\Delta}\frac{1}{M_\Delta^2}+\frac{\nu_4(\delta)}{\delta\,\Delta}\frac{1}{M_\Delta^3\sqrt{n}}\right)$$
$$\le C(\rho)\sqrt{k}\,\mathfrak{B}^3(\ell,\mathcal{F})\left(\frac{\widetilde{\nu}(\delta_{\min})}{\Delta}\frac{1}{M_\Delta^2}+\frac{\widetilde{\nu}_4(\delta_{\min})}{\Delta}\frac{1}{M_\Delta^3\sqrt{n}}\right)$$

since $\nu(\delta) \le \widetilde{\nu}(\delta)$, $\nu_4(\delta) \le \widetilde{\nu}_4(\delta)$ and $\widetilde{\nu}(\delta)$, $\widetilde{\nu}_4(\delta)$ are functions of concave type. Combining the bound above with (6.5.16), we deduce that

$$\widehat{T}_{|J|}(\delta_{\min}) \le \widetilde{U}(\delta_{\min}, s) + C(\rho)\sqrt{k}\widetilde{B}(\delta_{\min}),$$

hence (6.5.10) and (6.5.13) imply that for all $\delta \ge \delta_{\min}$ simultaneously,

$$\sup_{f\in\mathcal{F}(\delta)} \left|\widehat{G}_k(0;f) - \widehat{G}_k(0;f_*)\right| \le C(\rho)\delta\left(\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}} + \sqrt{k}\frac{\widetilde{B}(\delta_{\min})}{\delta_{\min}}\right) + 4\frac{\mathcal{O}}{\sqrt{k}}$$

with probability at least $1 - e^{-s}$.

6.5.6 Proof of Lemma 43.

The following identity is immediate:

$$R_N(f) = \underbrace{\widehat{G}_k\Big(\widehat{e}^{(k)}(f); f\Big)}_{=0} + \partial_z G_k(0;f)\cdot\widehat{e}^{(k)}(f) - \Big(\widehat{G}_k\Big(\widehat{e}^{(k)}(f);f\Big) - \widehat{G}_k(0;f)\Big).$$

Assumptions on ρ imply that for any $f \in \mathcal{F}$ and $j = 1,\dots,k$, there exists $\tau_j \in [0,1]$ such that

$$\rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f) - \mathcal{L}(f) - \widehat{e}^{(k)}(f)}{\Delta}\right) = \rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f) - \mathcal{L}(f)}{\Delta}\right) - \frac{\sqrt{n}}{\Delta}\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f) - \mathcal{L}(f)}{\Delta}\right)\cdot\widehat{e}^{(k)}(f)$$
$$+ \frac{n}{\Delta^2}\rho'''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f) - \mathcal{L}(f) - \tau_j\widehat{e}^{(k)}(f)}{\Delta}\right)\cdot\Big(\widehat{e}^{(k)}(f)\Big)^2,$$

hence

$$\widehat{G}_k\Big(\widehat{e}^{(k)}(f);f\Big) - \widehat{G}_k(0;f) = -\frac{\sqrt{n}}{\Delta}\frac{\widehat{e}^{(k)}(f)}{\sqrt{k}}\sum_{j=1}^k \rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f) - \mathcal{L}(f)}{\Delta}\right)$$
$$+ \frac{n}{\Delta^2}\frac{\big(\widehat{e}^{(k)}(f)\big)^2}{\sqrt{k}}\sum_{j=1}^k \rho'''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f) - \mathcal{L}(f) - \tau_j\widehat{e}^{(k)}(f)}{\Delta}\right),$$

and

$$R_N(f) = \frac{\sqrt{n}}{\Delta}\frac{\widehat{e}^{(k)}(f)}{\sqrt{k}}\sum_{j=1}^k\left(\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f) - \mathcal{L}(f)}{\Delta}\right) - \mathbb{E}\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f) - \mathcal{L}(f)}{\Delta}\right)\right)$$
$$- \frac{n}{\Delta^2}\frac{\big(\widehat{e}^{(k)}(f)\big)^2}{\sqrt{k}}\sum_{j=1}^k \rho'''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f) - \mathcal{L}(f) - \tau_j\widehat{e}^{(k)}(f)}{\Delta}\right). \quad (6.5.17)$$

We will need the following modification of Theorem 33 that is stated below and proved in Section 6.5.7.

Lemma 44. *Then there exist positive constants $c(\rho)$, $C(\rho)$ with the following properties. Fix $\delta_{\min} > 0$. Then for all $s > 0$, $\delta \geq \delta_{\min}$, positive integers n and k such that*

$$\delta \frac{\widetilde{U}(\delta_{\min}, s)}{\delta_{\min}\sqrt{k}} + \sup_{f\in\mathcal{F}} G_f(n,\Delta) + \frac{s+\mathcal{O}}{k} \leq c(\rho),$$

the following inequality holds with probability at least $1 - 2e^{-s}$:

$$\sup_{f\in\mathcal{F}(\delta)} \left|\widehat{e}^{(k)}(f)\right| \leq C(\rho)\widetilde{\Delta}\left[\frac{\delta}{\sqrt{N}}\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}} + \frac{\sigma(\ell,f_*)}{\Delta}\sqrt{\frac{s}{N}} + \frac{\sup_{f\in\mathcal{F}} G_f(n,\Delta)}{\sqrt{n}} + \frac{(s+\mathcal{O})\sqrt{n}}{N}\right]. \tag{6.5.18}$$

In the rest of the proof, we will assume that conditions of Lemma 44 and Theorem 33 hold, and let Θ' be an event of probability at least $1 - 4e^{-s}$ on which inequalities (6.5.18) and (6.2.4) are valid. On event Θ', the last term in (6.5.17) can thus be estimated as

$$\begin{aligned}
\sup_{f\in\mathcal{F}(\delta)} \left| \frac{n}{\Delta^2} \frac{\left(\widehat{e}^{(k)}(f)\right)^2}{\sqrt{k}} \sum_{j=1}^k \rho'''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f) - \mathcal{L}(f) - \tau_j \widehat{e}^{(k)}(f)}{\Delta}\right)\right| &\leq C_1(\rho)\frac{\sqrt{nN}}{\Delta^2} \sup_{f\in\mathcal{F}(\delta)} \left|\widehat{e}^{(k)}(f)\right|^2 \\
&\leq C_2(\rho)\sqrt{N}\frac{\widetilde{\Delta}^2}{\Delta^2}\left(\frac{n^{1/2}\delta^2}{N}\left(\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}}\right)^2 \bigvee \frac{\sigma^2(\ell,f_*)}{\Delta^2}\frac{n^{1/2}\, s}{N}\right. \\
&\qquad \left. \bigvee\, n^{1/2}\left(\sup_{f\in\mathcal{F}}\frac{G_f(n,\Delta)}{\sqrt{n}}\right)^2 \bigvee\, n^{3/2}\frac{s^2+\mathcal{O}^2}{N^2}\right), \qquad (6.5.19)
\end{aligned}$$

where we used the fact that $\|\rho'''\|_\infty < \infty$. It remains to estimate the first term in (6.5.17). The required bound will follow from the combination of Theorem 44 and the following lemma that is proved in Section 6.5.8.

Lemma 45. *Fix $\delta_{\min} > 0$. With probability at least $1 - 3e^{-s}$, for all $\delta \geq \delta_{\min}$ simultaneously,*

$$\begin{aligned}
&\sup_{f\in\mathcal{F}(\delta)} \left| \frac{1}{\sqrt{k}} \sum_{j=1}^k \left(\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f) - \mathcal{L}(f)}{\Delta}\right) - \mathbb{E}\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f) - \mathcal{L}(f)}{\Delta}\right)\right)\right| \\
&\qquad\qquad \leq C(\rho)\left(\delta\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}} + \frac{\sigma(\ell,f_*)}{\Delta}\sqrt{s} + \frac{s+\mathcal{O}}{\sqrt{k}}\right).
\end{aligned}$$

Let Θ'' be the event of probability at least $1 - 3e^{-2s}$ on which the inequality of Lemma 45 holds. Then simple algebra yields that on event $\Theta' \cap \Theta''$ of probability at least $1 - 7e^{-s}$,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}(\delta)} \left| \frac{\sqrt{n}}{\Delta}\frac{\widehat{e}^{(k)}(f)}{\sqrt{k}} \sum_{j=1}^k \left(\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f) - \mathcal{L}(f)}{\Delta}\right) - \mathbb{E}\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f) - \mathcal{L}(f)}{\Delta}\right)\right)\right| \\
&\qquad \leq C_3(\rho)\sqrt{N}\frac{\widetilde{\Delta}}{\Delta}\left(\frac{n^{1/2}\delta^2}{N}\left(\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}}\right)^2 \bigvee \frac{\sigma^2(\ell,f_*)}{\Delta^2}\frac{n^{1/2}\, s}{N}\right. \\
&\qquad\qquad \left. \bigvee\, n^{1/2}\left(\sup_{f\in\mathcal{F}}\frac{G_f(n,\Delta)}{\sqrt{n}}\right)^2 \bigvee\, n^{3/2}\frac{s^2+\mathcal{O}^2}{N^2}\right). \qquad (6.5.20)
\end{aligned}$$

Combination of inequalities (6.5.19) and (6.5.20) that hold with probability at leat $1-7e^{-s}$ yields the result.

6.5.7 Proof of Lemma 44.

In the situation when δ is fixed, the argument mimics the proof of Theorem 4.1 in [Min18], with minor modifications outlined below. Recall that

$$\widehat{G}_k(z;f) = \frac{1}{\sqrt{k}}\sum_{j=1}^{k}\rho'\biggl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\biggr).$$

Let z_1, z_2 be such that on an event of probability close to 1, $\widehat{G}_k(z_1;f)>0$ and $\widehat{G}_k(z_2;f)<0$ for all $f\in\mathcal{F}(\delta)$ simultaneously. Since $\widehat{G}_k$ is decreasing in z, it is easy to see that $\widehat{e}^{(k)}(f)\in(z_1,z_2)$ for all $f\in\mathcal{F}(\delta)$ on this event. Hence, our goal is to find z_1, z_2 satisfying conditions above and such that $|z_1|$, $|z_2|$ are as small as possible. Observe that

$$\widehat{G}_k(z;f) = \frac{1}{\sqrt{k}}\sum_{j\in J}\rho'\biggl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\biggr) + \frac{1}{\sqrt{k}}\sum_{j\notin J}\rho'\biggl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\biggr)$$

and $\Bigl|\frac{1}{\sqrt{k}}\sum_{j\notin J}\rho'\Bigl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\Bigr)\Bigr|\le 2\frac{\mathcal{O}}{\sqrt{k}}$. Moreover,

$$\begin{aligned}
&\frac{1}{\sqrt{k}}\sum_{j\in J}\rho'\biggl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\biggr)\\
&\quad= \frac{1}{\sqrt{k}}\sum_{j\in J}\biggl(\rho'\biggl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\biggr)-\rho'\biggl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\biggr)\\
&\qquad-\mathbb{E}\biggl[\rho'\biggl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\biggr)-\rho'\biggl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\biggr)\biggr]\biggr)\\
&\qquad+\frac{1}{\sqrt{k}}\sum_{j\in J}\biggl(\rho'\biggl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\biggr)-\mathbb{E}\rho'\biggl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\biggr)\biggr)\\
&\qquad+\frac{1}{\sqrt{k}}\sum_{j\in J}\biggl(\mathbb{E}\rho'\biggl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\biggr)-\mathbb{E}\rho'\biggl(\frac{W(\ell(f))-\sqrt{n}z}{\Delta}\biggr)\biggr)\\
&\qquad+\frac{1}{\sqrt{k}}\sum_{j\in J}\mathbb{E}\rho'\biggl(\frac{W(\ell(f))-\sqrt{n}z}{\Delta}\biggr).
\end{aligned}$$

We will proceed in 4 steps: first, we will find $\epsilon_1>0$ such that for any $z\in\mathbb{R}$ and all $f\in\mathcal{F}(\delta)$,

$$\begin{aligned}
&\frac{1}{\sqrt{k}}\sum_{j\in J}\biggl(\rho'\biggl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\biggr)-\rho'\biggl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\biggr)\\
&\qquad-\mathbb{E}\biggl[\rho'\biggl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\biggr)-\rho'\biggl(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\biggr)\biggr]\biggr)\le\epsilon_1
\end{aligned}$$

with high probability, then $\epsilon_2 > 0$ such that

$$\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\rho'\left(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\right)-\mathbb{E}\rho'\left(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\right)\right)\le\epsilon_2,$$

ϵ_3 satisfying

$$\sup_{f\in\mathcal{F}(\delta)}\left|\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\mathbb{E}\rho'\left(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-z}{\Delta}\right)-\mathbb{E}\rho'\left(\frac{W(\ell(f))-\sqrt{n}z}{\Delta}\right)\right)\right|\le\epsilon_3,$$

and finally we will choose $z_1 < 0$ such that for all $f \in \mathcal{F}(\delta)$,

$$\frac{1}{\sqrt{k}}\sum_{j\in J}\mathbb{E}\rho'\left(\frac{W(\ell(f))-\sqrt{n}z}{\Delta}\right)>\epsilon_1+\epsilon_2+\epsilon_3+2\frac{\mathcal{O}}{\sqrt{k}}. \tag{6.5.21}$$

Talagrand's concentration inequality (e.g. Corollary 16.1 in [Van16]), together with the bound $\|\rho'\|_\infty \le 2$, implies that for any $s > 0$,

$$\sqrt{\frac{|J|}{k}}\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_{|J|}(z;f)-\widehat{G}_{|J|}(z;f_*)-\mathbb{E}\Big(\widehat{G}_{|J|}(z;f)-\widehat{G}_{|J|}(z;f_*)\Big)\right|\le$$
$$2\left[\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_{|J|}(z;f)-\widehat{G}_{|J|}(z;f_*)-\mathbb{E}\Big(\widehat{G}_{|J|}(z;f)-\widehat{G}_{|J|}(z;f_*)\Big)\right|+D(\delta)\sqrt{\frac{s}{2}}+\frac{32}{3}\frac{s}{\sqrt{k}}\right]$$

with probability at least $1-2e^{-s}$. It has been observed in (6.5.9) that $D(\delta)\le\frac{\nu(\delta)}{\Delta}$. It remains to estimate the expected supremum. Sequential application of symmetrization, contraction and desymmetrization inequalities, together with the fact that $L(\rho')=1$, implies that

$$\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\widehat{G}_{|J|}(z;f)-\widehat{G}_{|J|}(z;f_*)-\mathbb{E}\Big(\widehat{G}_{|J|}(z;f)-\widehat{G}_{|J|}(z;f_*)\Big)\right|$$
$$\le 2\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\frac{1}{\sqrt{|J|}}\sum_{j\in J}\epsilon_j\left(\rho'\left(\sqrt{n}\,\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)-z}{\Delta}\right)\right)-\rho'\left(\sqrt{n}\,\frac{\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)-z}{\Delta}\right)\right|$$
$$\le\frac{4}{\Delta}\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\frac{\sqrt{n}}{\sqrt{|J|}}\sum_{j\in J}\epsilon_j\big((\overline{\mathcal{L}}_j(f)-\mathcal{L}(f))-(\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))\big)\right|$$
$$\le\frac{8\sqrt{2}}{\Delta}\mathbb{E}\sup_{f\in\mathcal{F}(\delta)}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^{N_J}\Big((\ell(f)-\ell(f_*))(X_j)-P(\ell(f)-\ell(f_*))\Big)\right|\le\frac{8\sqrt{2}}{\Delta}\,\omega(\delta).$$

Hence, it suffices to choose

$$\epsilon_1=\frac{8\sqrt{2}}{\Delta}\,\omega(\delta)+\frac{\nu(\delta)}{\Delta}\,\sqrt{s}+\frac{32}{3}\frac{s}{\sqrt{k}}.$$

Next, Bernstein's inequality and Lemma 33 together yield that with probability at least $1-2e^{-s}$,

$$\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\rho'\left(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\right)-\mathbb{E}\rho'\left(\sqrt{n}\,\frac{(\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*))-z}{\Delta}\right)\right)$$
$$\le 2\left(\frac{\sigma(\ell,f_*)}{\Delta}\sqrt{s}+\frac{3s}{\sqrt{k}}\right),$$

thus we can set $\epsilon_2 = 2\Big(\frac{\sigma(\ell,f_*)}{\Delta}\sqrt{s} + 3\frac{s}{\sqrt{k}}\Big)$. Lemma 35 implies that ϵ_3 can be chosen as

$$\epsilon_3 = \sqrt{k}\sup_{f\in\mathcal{F}(\delta)} G_f(n,\Delta).$$

Finally, we apply Lemma 6.3 of [Min18] with

$$\epsilon := \epsilon_1 + \epsilon_2 + \epsilon_3 + 2\frac{\mathcal{O}}{\sqrt{k}}$$

to deduce that

$$z_1 = -C\frac{\widetilde{\Delta}}{\sqrt{N}}\cdot\Big(\epsilon_1+\epsilon_2+\epsilon_3+2\frac{\mathcal{O}}{\sqrt{k}}\Big),$$

satisfies (6.5.21) under assumption that $\frac{\epsilon_1+\epsilon_2+\epsilon_3}{\sqrt{k}} + \frac{\mathcal{O}}{k} \le c$ for some absolute constants $c, C > 0$. Proceeding in a similar way, it is easy to see that setting $z_2 = -z_1$ guarantees that $\widehat{G}_k(z_2; f) < 0$ for all $f \in \mathcal{F}(\delta)$ with probability at least $1 - e^{-s}$, hence the claim follows.

It remains to make the bound uniform in $\delta \ge \delta_{\min}$. To this end, we need to repeat the "slicing argument" of Lemma 42 below (specifically, see (6.5.16)) to deduce that with probability at least $1 - 2e^{-s}$,

$$\sup_{f\in\mathcal{F}(\delta)}\Big|\widehat{G}_{|J|}(z;f) - \widehat{G}_{|J|}(z;f_*) - \mathbb{E}\Big(\widehat{G}_{|J|}(z;f) - \widehat{G}_{|J|}(z;f_*)\Big)\Big| \le \delta\,\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}}$$

uniformly for all $\delta \ge \delta_{\min}$, hence the value of ϵ_1 should be replaced by $\epsilon_1 = \delta\,\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}}$.

6.5.8 Proof of Lemma 45.

Observe that

$$\begin{aligned}
&\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\Big(\rho''\Big(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\Big) - \mathbb{E}\rho''\Big(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\Big)\Big)\\
&\qquad= \frac{1}{\sqrt{k}}\sum_{j\notin J}\Big(\rho''\Big(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\Big) - \mathbb{E}\rho''\Big(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\Big)\Big)\\
&\qquad+ \frac{1}{\sqrt{k}}\sum_{j\in J}\Big(\rho''\Big(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\Big) - \rho''\Big(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\Big)\\
&\qquad- \mathbb{E}\Big(\rho''\Big(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\Big) - \rho''\Big(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\Big)\Big)\Big)\\
&\qquad+ \frac{1}{\sqrt{k}}\sum_{j\in J}\Big(\rho''\Big(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\Big) - \mathbb{E}\rho''\Big(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\Big)\Big).
\end{aligned}$$

Clearly, as $\|\rho''\|_\infty \le 1$, $\Big|\frac{1}{\sqrt{k}}\sum_{j\notin J}\Big(\rho''\Big(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\Big) - \mathbb{E}\rho''\Big(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\Big)\Big)\Big| \le 2\frac{\mathcal{O}}{\sqrt{k}}$. Next, repeating the "slicing argument" of Lemma 42, it is not difficult to deduce that with probability

at least $1-2e^{-2s}$,

$$\sup_{f\in\mathcal{F}(\delta)}\left|\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)-\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right.$$
$$\left.-\mathbb{E}\left(\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)-\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right)\right|\le C(\rho)\,\delta\,\frac{\widetilde{U}(\delta_{\min},s)}{\delta_{\min}}$$

uniformly for all $\delta\ge\delta_{\min}$. Next, we will apply Bernstein's inequality to estimate the remaining term. Since ρ is convex, ρ'' is nonnegative, moreover, it follows from Assumption 1 that $\rho''(x)\neq 0$ for $|x|\le 2$, $\rho''(x)=1$ for $|x|\le 1$, and $\|\rho''\|_\infty=1$, hence $\left(\mathbb{E}\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)\right)^2\ge\left(\Pr\left(\left|\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right|\le 1\right)\right)^2$,

$$\mathbb{E}\left(\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)\right)^2\le\Pr\left(\left|\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right|\le 1\right)+\Pr\left(\left|\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right|\in[1,2]\right),$$

and

$$\mathrm{Var}\left(\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right)\right)\le\Pr\left(\left|\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right|\le 1\right)-\left(\Pr\left(\left|\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right|\le 1\right)\right)^2$$
$$+\Pr\left(\left|\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right|\ge 1\right)$$
$$\le 2\Pr\left(\left|\sqrt{n}\frac{\overline{\mathcal{L}}_j(f)-\mathcal{L}(f)}{\Delta}\right|\ge 1\right)\le 2\frac{\mathrm{Var}(\ell(f(X)))}{\Delta^2}.$$

Bernstein's inequality implies that with probability at least $1-e^{-s}$,

$$\frac{1}{\sqrt{k}}\sum_{j\in J}\left(\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\right)-\mathbb{E}\rho''\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(f_*)-\mathcal{L}(f_*)}{\Delta}\right)\right)\le 2\left(\frac{\sigma(\ell,f_*)}{\Delta}\sqrt{s}+\frac{s}{\sqrt{k}}\right),$$

hence the desired conclusion follows.

6.5.9 Proof of Lemma 32.

In the context of regression with quadratic loss, $\omega(\delta)$ takes the form

$$\omega(\delta)=\mathbb{E}\sup_{\ell(f)\in\mathcal{F}(\delta)}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^N\left((Y_j-f(Z_j))^2-(Y_j-f_*(Z_j))^2-\mathbb{E}\big((Y_j-f(Z_j))^2-(Y_j-f_*(Z_j))^2\big)\right)\right|.$$

In view of Bernstein's assumption verified above, $\omega(\delta)$ is bounded by

$$\mathbb{E}\sup_{\|f-f_*\|^2_{L_2(\Pi)}\le 2\delta}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^N\left((Y_j-f(Z_j))^2-(Y_j-f_*(Z_j))^2-\mathbb{E}\big((Y_j-f(Z_j))^2-(Y_j-f_*(Z_j))^2\big)\right)\right|.$$

To estimate the latter quantity, we will use the approach based on the $L_\infty(\Pi_n)$-covering numbers of the class $\mathcal{F}$ (e.g., see [BMN12]). We will also set

$$B(\mathcal{F};\tau) := \{f \in \mathcal{F} : \|f - f_*\|^2_{L_2(\Pi)} \le \tau\}.$$

It is easy to see that

$$(Y - f(X))^2 - (Y - f_*(X))^2 = (f(X) - f_*(X))^2 + 2(f(X) - f_*(X))(f_*(X) - Y),$$

hence

$$w(\delta) \le \mathbb{E} \sup_{B(\mathcal{F};2\delta)} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N (f(Z_j) - f_*(Z_j))^2 - \mathbb{E}(f(Z_j) - f_*(Z_j))^2 \right| + 2\,\mathbb{E} \sup_{B(\mathcal{F};2\delta)} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N (f(Z_j) - f_*(Z_j))(Y_j - f_*(Z_j)) \right|. \quad (6.5.22)$$

We will estimate the two terms separately. By assumption, the covering numbers of the class $\mathcal{F}$ satisfy the bound

$$N(\mathcal{F}, L_2(\Pi_N), \epsilon) \le \left(\frac{A\|F\|_{L_2(\Pi_N)}}{\epsilon} \right)^V \vee 1 \quad (6.5.23)$$

for some constants $A \ge 1$, $V \ge 1$ and all $\epsilon > 0$. We apply bound of Lemma 36 to the first term in (6.5.22) to get that

$$\mathbb{E} \sup_{B(\mathcal{F};2\delta)} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N (f(Z_j) - f_*(Z_j))^2 - \mathbb{E}(f(Z_j) - f_*(Z_j))^2 \right| \le C \left(\sqrt{2\delta} \sqrt{\Gamma_{N,\infty}(B(\mathcal{F};2\delta))} \bigvee \frac{\Gamma_{N,\infty}(B(\mathcal{F};2\delta))}{\sqrt{N}} \right).$$

To estimate $\Gamma_{n,\infty}(B(\mathcal{F};2\delta)) := \mathbb{E}\gamma_2^2(B(\mathcal{F};2\delta); L_\infty(\Pi_N))$, we will use Dudley's entropy integral bound. Observe that

$$\operatorname{diam}(B(\mathcal{F};2\delta); L_\infty(\Pi_N)) \le 2\|F\|_{L_\infty(\Pi_N)}.$$

Moreover, for any $f, g \in \mathcal{F}$,

$$\frac{1}{N} \sum_{j=1}^N (f(Z_j) - g(Z_j))^2 \ge \frac{1}{N} \max_{1 \le j \le N} (f(Z_j) - g(Z_j))^2,$$

hence $N(B(\mathcal{F};2\delta), L_\infty(\Pi_N), \epsilon) \le N\Big(B(\mathcal{F};2\delta), L_2(\Pi_N), \frac{\epsilon}{\sqrt{N}}\Big)$ and, whenever (6.5.23) holds,

$$\log N(B(\mathcal{F};2\delta), L_\infty(\Pi_N), \epsilon) \le V \log_+ \left(\frac{A\sqrt{N}\|F\|_{L_2(\Pi_N)}}{\epsilon} \right),$$

where $\log_+(x) := \max(\log x, 0)$. It yields that

$$\Gamma_{N,\infty}(B(\mathcal{F};2\delta)) \le \mathbb{E} \left(\sqrt{V} \int_0^{2\|F\|_{L_\infty(\Pi_N)}} \log_+^{1/2} \left(\frac{A\|F\|_{L_2(\Pi_N)}\sqrt{N}}{\epsilon} \right) d\epsilon \right)^2$$
$$\le C\,V \mathbb{E} \left(\|F\|^2_{L_\infty(\Pi_N)} \log \left(\frac{A\sqrt{N}\|F\|_{L_2(\Pi_N)}}{\|F\|_{L_\infty(\Pi_N)}} \vee e \right) \right) \le C\,V \log(A\sqrt{N}) \mathbb{E}\|F\|^2_{L_\infty(\Pi_N)}$$

for an absolute constant $C > 0$. Finally, since $\|F\|_{\psi_2} < \infty$,

$$\mathbb{E}\big\|F^2\big\|_{L_\infty(\Pi_N)} \le C_1 \log(N)\|F^2\|_{\psi_1} = C_1 \log(N)\|F\|_{\psi_2}^2,$$

hence

$$\mathbb{E} \sup_{B(\mathcal{F};2\delta)} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N (f(Z_j) - f_*(Z_j))^2 - \mathbb{E}(f(Z_j) - f_*(Z_j))^2 \right| \\ \le C_2 \left(\sqrt{\delta}\sqrt{V} \log(A^2 N)\|F\|_{\psi_2} \bigvee \frac{V\|F\|_{\psi_2}^2 \log^2(A^2 N)}{\sqrt{N}} \right). \quad (6.5.24)$$

Next, the multiplier inequality [vdVW96] implies that

$$\mathbb{E} \sup_{B(\mathcal{F};2\delta)} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N (f(Z_j) - f_*(Z_j))(Y_j - f_*(Z_j) \right| \\ \le C\|\eta\|_{2,1} \max_{k=1,\dots,N} \mathbb{E} \sup_{B(\mathcal{F};2\delta)} \left| \frac{1}{\sqrt{k}} \sum_{j=1}^k (f(Z_j) - f_*(Z_j)) \right|.$$

Using symmetrization inequality and applying Dudley's entropy integral bound, we deduce that for any k

$$\mathbb{E} \sup_{B(\mathcal{F};2\delta)} \left| \frac{1}{\sqrt{k}} \sum_{j=1}^k (f(Z_j) - f_*(Z_j)) \right| \le C\sqrt{V}\mathbb{E} \int_0^{\sigma_k} \log^{1/2} \left(\frac{A\|F_{2\delta}\|_{L_2(\Pi_k)}}{\epsilon} \right) d\epsilon \\ \le C_1 \sqrt{V} \mathbb{E} \left(\sigma_k \log^{1/2} \left(\frac{eA\|F_{2\delta}\|_{L_2(\Pi_k)}}{\sigma_k} \right) \right),$$

where $F_{2\delta}$ is the envelope of the class $B(\mathcal{F}; 2\delta)$ and $\sigma_k^2 := \sup_{f \in B(\mathcal{F};2\delta)} \|f - f_*\|_{L_2(\Pi_k)}^2$. Cauchy-Schwarz inequality, together with an elementary observation that $k\sigma_k^2 \ge \|F_{2\delta}\|_{L_2(\Pi_k)}^2$, gives

$$\mathbb{E} \left(\sigma_k \log^{1/2} \left(\frac{eA\|F_{2\delta}\|_{L_2(\Pi_k)}}{\sigma_k} \right) \right) \le \sqrt{\mathbb{E}\sigma_k^2} \log^{1/2}(eA\sqrt{k}).$$

According to (6.5.24),

$$\mathbb{E}\sigma_k^2 \le 2\delta + C_2 \left(\sqrt{\delta}\sqrt{\frac{V}{N}} \log(A^2 N)\|F\|_{\psi_2} \bigvee \frac{V\|F\|_{\psi_2}^2 \log^2(A^2 N)}{N} \right).$$

Simple algebra now yields that

$$\mathbb{E} \sup_{B(\mathcal{F};2\delta)} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^N (f(Z_j) - f_*(Z_j))(Y_j - f_*(Z_j) \right| \\ \le C\|\eta\|_{2,1} \sqrt{V \log(e^2 A^2 N)} \left(\sqrt{\delta} + \sqrt{\frac{V}{N}} \log(A^2 N)\|F\|_{\psi_2} \right). \quad (6.5.25)$$

Finally, combination of inequalities (6.5.24) and (6.5.25) implies that

$$w(\delta) \le \widetilde{\omega}(\delta) := C\left(\sqrt{\delta}\sqrt{V}\log(A^2N)(\|F\|_{\psi_2} + \|\eta\|_{2,1}) \bigvee \frac{V(\|F\|_{\psi_2}^2 + \|\eta\|_{2,1}^2)\log^2(A^2N)}{\sqrt{N}}\right),$$

where $\widetilde{\omega}(\delta)$ is of strictly concave type, hence

$$\bar{\delta} \le C(\rho)\frac{V\log^2(A^2N)(\|F\|_{\psi_2}^2 + \|\eta\|_{2,1}^2)}{N}$$

thus proving the claim.

6.6 Technical results for the U-statistics based estimator $\widehat{f}_N^U$.

In this section, we describe the tools necessary to extend Klein and Rio's inequality stated in Lemma 37 to nondegenerate U-statistics. First, we note that the deviation inequality (6.4.2) is a corollary of the following bound for the moment generating function (Section 12.5 in [BLM13]):

$$\log \mathbb{E}e^{\lambda\left(\sum_{j=1}^N (Z_j(f) - \mathbb{E}Z_j(f))\right)} \le \frac{e^{\lambda M} - \lambda M - 1}{M^2}\left(V^2(\mathcal{F}) + 2M\,\mathbb{E}\sup_{f\in\mathcal{F}}\left(\sum_{j=1}^N (Z_j(f) - \mathbb{E}Z_j(f))\right)\right) \tag{6.6.1}$$

that holds for all $\lambda > 0$. We use this fact to demonstrate a straightforward extension of Lemma 37 to the case of U-statistics. Let π_N be the collection of all permutations $\tau : \{1,\dots,N\} \mapsto \{1,\dots,N\}$. Given $(i_1,\dots,i_N) \in \pi_N$ and a U-statistic $U_{N,n}$ with kernel h defined in (6.1.3), let

$$T_{i_1,\dots,i_N} := \frac{1}{k}\big(h(X_{i_1},\dots,X_{i_n}) + h\big(X_{i_{n+1}},\dots,X_{i_{2n}}\big) + \dots + h\big(X_{i_{(k-1)n+1}},\dots,X_{i_{kn}}\big)\big).$$

It is well known (e.g., see Section 5 in [Hoe63]) that the following representation holds:

$$U_{N,n} = \frac{1}{N!}\sum_{(i_1,\dots,i_N)\in\pi_N} T_{i_1,\dots,i_N}. \tag{6.6.2}$$

Let $U'_{N,n}(z;f) = \frac{1}{\binom{N}{n}}\sum_{J\in\mathcal{A}_N^{(n)}}\rho'\left(\sqrt{n}\,\frac{(\overline{\mathcal{L}}(f;J) - \mathbb{E}\ell(f(X))) - z}{\Delta}\right)$. Applied to $U'_{N,n}(z;f)$, relation (6.6.2) yields that

$$U'_{N,n}(z;f) = \frac{1}{N!}\sum_{(i_1,\dots,i_N)\in\pi_N} T_{i_1,\dots,i_N}(z;f),$$

where

$$T_{i_1,\dots,i_N}(z;f) = \frac{1}{k}\left(\rho'\left(\sqrt{n}\,\frac{\overline{\mathcal{L}}(f;\{i_1,\dots,i_n\}) - \mathbb{E}\ell(f(X)) - z}{\Delta}\right) + \right.$$
$$\left. \dots + \rho'\left(\sqrt{n}\,\frac{\overline{\mathcal{L}}(f;\{i_{(k-1)n+1},\dots,i_{kn}\}) - \mathbb{E}\ell(f(X)) - z}{\Delta}\right)\right).$$

Jensen's inequality implies that for any $\lambda > 0$,

$$\mathbb{E}\exp\left(\frac{\lambda}{N!}\sum_{(i_1,\dots,i_N)\in\pi_N}(T_{i_1,\dots,i_N}(z;f)-\mathbb{E}T_{i_1,\dots,i_N}(z;f))\right) \\ \leq \frac{1}{N!}\sum_{(i_1,\dots,i_N)\in\pi_N}\mathbb{E}\exp\Big(\lambda(T_{1,\dots,N}(z;f)-\mathbb{E}T_{1,\dots,N}(z;f))\Big),$$

hence bound (6.6.1) can be applied and yields that

$$\sup_{f\in\mathcal{F}}\big(U'_{N,n}(z;f)-\mathbb{E}U'_{N,n}(z;f)\big) \leq 2\mathbb{E}\sup_{f\in\mathcal{F}}(T_{1,\dots,N}(z;f)-\mathbb{E}T_{1,\dots,N}(z;f)) \\ + \sup_{f\in\mathcal{F}}\sqrt{\mathrm{Var}\left(\rho'\left(\sqrt{n}\,\frac{\bar{\theta}(f;\{1,\dots,n\})-Pf-z}{\Delta}\right)\right)}\sqrt{\frac{2s}{k}}+\frac{8s\|\rho'\|_\infty}{3k} \quad (6.6.3)$$

with probability at least $1-e^{-s}$. The expression can be further simplified by noticing that $\|\rho'\|_\infty \leq 2$ and that

$$\mathrm{Var}\left(\rho'\left(\sqrt{n}\,\frac{\bar{\theta}(f;\{1,\dots,n\})-Pf-z}{\Delta}\right)\right) \leq \frac{\sigma^2(f)}{\Delta^2}.$$

due to Lemma 33.

6.7 Numerical algorithms and examples.

The main goal of this section is to discuss in detail the numerical algorithms[4] used to approximate estimators $\widehat{f}_N$ and $\widehat{f}_N^U$, as well as assess the quality of the resulting solutions. We will also compare our methods with the ones known previously, specifically, the median-of-means based approach proposed in [LLM20]. Finally, we perform the numerical study of dependence of the solutions on the parameters Δ and k. All evaluations are performed for logistic regression in the framework of binary classification as well as linear regression with quadratic loss using simulated data; the main definitions pertaining to the logistic and linear regression were given in Section 6.1.4. In both examples, we will assume that we are given an i.i.d. sample $(Z_1,Y_1),\dots,(Z_N,Y_N)$ having the same distribution as (Z,Y). In the end of this section, we also demonstrate applications to the real dataset from the UCI Machine Learning Repository.

Let us mention that the numerical methods for closely related approach in the special case of linear regression have been investigated in a recent work [HI17b]. Here, we focus on general algorithms that can easily be adapted to other predictions tasks and loss functions. Let us first briefly recall the formulations of both the binary classification and the linear regression problems.

6.7.1 Gradient descent algorithms.

Optimization problems (6.1.4) and (6.1.5) are not convex, so we will focus our attention of the variants of the gradient descent method employed to find local minima. We will first derive the

[4]The code used in this section is available on github at https://github.com/TimotheeMathieu/Excess-risk-bounds-in-robust-empirical-risk-minimization/

expression for $\nabla_\beta \widehat{\mathcal{L}}^{(k)}(\beta)$, the gradient of $\widehat{\mathcal{L}}^{(k)}(\beta) := \widehat{\mathcal{L}}^{(k)}(f_\beta)$, for the problems corresponding to logistic regression and regression with quadratic loss. It follows from (6.1.2) that $\widehat{\mathcal{L}}^{(k)}(\beta)$ satisfies the equation

$$\sum_{j=1}^{k} \rho'\left(\sqrt{n}\frac{\overline{\mathcal{L}}_j(\beta) - \widehat{\mathcal{L}}^{(k)}(\beta)}{\Delta}\right) = 0. \tag{6.7.1}$$

Taking the derivative in (6.7.1) with respect to β, we retrieve $\nabla_\beta \widehat{\mathcal{L}}^{(k)}(\beta)$:

$$\nabla_\beta \widehat{\mathcal{L}}^{(k)}(\beta) = \frac{\sum_{j=1}^{k}\Big(\frac{1}{n}\sum_{i\in G_j} Z_i\, \ell'(Y_i, f_\beta(Z_i))\Big)\rho''\Big(\sqrt{n}\frac{\overline{\mathcal{L}}_j(\beta)-\widehat{\mathcal{L}}^{(k)}(\beta)}{\Delta}\Big)}{\sum_{j=1}^{k}\rho''\Big(\sqrt{n}\frac{\overline{\mathcal{L}}_j(\beta)-\widehat{\mathcal{L}}^{(k)}(\beta)}{\Delta}\Big)},$$

where $\ell'(Y_i, f_\beta(Z_i))$ stands for the partial derivative $\frac{\partial \ell(y,t)}{\partial t}$ with respect to the second argument t, so that $\ell'(Y_i, f_\beta(Z_i)) = -Y_i \frac{e^{-Y_i\langle \beta, Z_i\rangle}}{1+e^{-Y_i\langle \beta, Z_i\rangle}}$ in the case of logistic regression and $\ell'(Y_i, f_\beta(Z_i)) = 2(\langle \beta, Z_i\rangle - Y_i)$ for regression with quadratic loss. In most of our numerical experiments, we choose ρ to be Huber's loss,

$$\rho(y) = \frac{y^2}{2} I\{|y| \le 1\} + \left(|y| - \frac{1}{2}\right) I\{|y| > 1\}.$$

In this case, $\rho''(y) = I\{|y| \le 1\}$ for all $y \in \mathbb{R}$, hence the expression for the gradient can be simplified to

$$\nabla_\beta \widehat{\mathcal{L}}^{(k)}(\beta) = \frac{\sum_{j=1}^{k}\Big(\frac{1}{n}\sum_{i\in G_j} Z_i\, \ell'(Y_i, f_\beta(Z_i))\Big) I\Big\{\Big|\overline{\mathcal{L}}_j(\beta) - \widehat{\mathcal{L}}^{(k)}(\beta)\Big| \le \frac{\Delta}{\sqrt{n}}\Big\}}{\#\Big\{j : \Big|\overline{\mathcal{L}}_j(\beta) - \widehat{\mathcal{L}}^{(k)}(\beta)\Big| \le \frac{\Delta}{\sqrt{n}}\Big\}}, \tag{6.7.2}$$

where we implicitly assume that Δ is chosen large enough so that the denominator is not equal to 0. To evaluate $\widehat{\mathcal{L}}^{(k)}(\beta)$, we use the "modified weights" algorithm due to Huber and Ronchetti, see Section 6.7 in [HR09]. Complete version of the gradient descent algorithm used to approximate $\widehat{\beta}_N$ (identified with the solution $\widehat{f}_N$ of the problem (6.1.4)) is presented in Figure 6.1.

Figure 6.1: Algorithm 1 – evaluation of $\widehat{\beta}_N$.

Input: the dataset $(Z_i, Y_i)_{1\le i\le N}$, number of blocks $k \in \mathbb{Z}_+$, step size parameter $\eta > 0$, maximum number of iterations M, initial guess $\beta_0 \in \mathbb{R}^d$, tuning parameter $\Delta > 0$.
Construct blocks $G_1, \dots, G_k$;
for all $t = 0, \dots, M$ **do**
  Compute $\widehat{\mathcal{L}}^{(k)}(\beta_t)$ using the Modified Weights algorithm;
  Compute $\nabla_\beta \widehat{\mathcal{L}}^{(k)}(\beta_t)$ from (6.7.2);
  Update

$$\beta_{t+1} = \beta_t - \eta \nabla_\beta \widehat{\mathcal{L}}^{(k)}(\beta_t).$$

end for
Output: β_{M+1}.

Next, we discuss a variant of a stochastic gradient descent for approximating the "permutation-invariant" estimator $\widehat{f}_N^U$ used when the subgroup size $n > 1$; in our numerical experiments (see Section 6.7.5 for the numerical comparison of two approaches), this method demonstrated

consistently superior performance. Below, we will identify $\widehat{f}_N^U$ with the vector of corresponding coefficients $\widehat{\beta}_N^U$. Recall that $\mathcal{A}_N^{(n)} := \{J : \ J \subseteq \{1,\dots,N\}, \text{Card}(J) = n\}$, and that

$$\widehat{\mathcal{L}}_U^{(k)}(\beta) = \operatorname*{argmin}_{z\in\mathbb{R}} \sum_{J\in\mathcal{A}_N^{(n)}} \rho\Big(\sqrt{n}\,\frac{\overline{\mathcal{L}}(f_\beta; J) - z}{\Delta}\Big). \tag{6.7.3}$$

Similarly to the way that we derived the expression for $\nabla_\beta \widehat{\mathcal{L}}^{(k)}(\beta)$ from (6.1.2), it follows from (6.7.3), with ρ again being the Huber's loss, that

$$\sum_{J\in\mathcal{A}_N^{(n)}} \rho'\Big(\sqrt{n}\,\frac{\overline{\mathcal{L}}(f_\beta; J) - \widehat{\mathcal{L}}_U^{(k)}(\beta)}{\Delta}\Big) = 0 \quad \text{and}$$

$$\nabla_\beta \widehat{\mathcal{L}}_U^{(k)}(\beta) = \frac{\sum_{J\in\mathcal{A}_N^{(n)}} \big(\frac{1}{n}\sum_{i\in J} Z_i\,\ell'(Y_i, f_\beta(Z_i))\big) I\Big\{\Big|\overline{\mathcal{L}}(\beta; J) - \widehat{\mathcal{L}}^{(k)}(\beta)\Big| \le \frac{\Delta}{\sqrt{n}}\Big\}}{\#\Big\{J\in\mathcal{A}_N^{(n)} : \ \Big|\overline{\mathcal{L}}(\beta; J) - \widehat{\mathcal{L}}^{(k)}(\beta)\Big| \le \frac{\Delta}{\sqrt{n}}\Big\}}. \tag{6.7.4}$$

Expressions in (6.7.4) are closely related to U-statistics, and it will be convenient to write them in a slightly different form. To this end, let π_N be the collection of all permutations $\tau : \{1,\dots,N\} \mapsto \{1,\dots,N\}$. Given $\tau = (i_1,\dots,i_N) \in \pi_N$ and an arbitrary U-statistic $U_{N,n}$ defined in (6.1.3), let

$$T_{i_1,\dots,i_N} := \frac{1}{k}\big(h(X_{i_1},\dots,X_{i_n}) + h\big(X_{i_{n+1}},\dots,X_{i_{2n}}\big) + \dots + h\big(X_{i_{(k-1)n+1}},\dots,X_{i_{kn}}\big)\big).$$

Equivalently, for $\tau = (i_1,\dots,i_N) \in \pi_N$, let

$$G_j(\tau) = \big(i_{(j-1)n+1},\dots,i_{jn}\big), \ j = 1,\dots,k = \lfloor N/n\rfloor, \tag{6.7.5}$$

which gives a compact form

$$T_\tau = \frac{1}{k}\sum_{j=1}^{k} h(X_i, i\in G_j(\tau)).$$

It is well known (Section 5 in [Hoe63]) that the following representation of the U-statistic holds:

$$U_{N,n} = \frac{1}{N!}\sum_{\tau\in\pi_N} T_\tau. \tag{6.7.6}$$

Applying representation (6.7.6) to (6.7.3), we deduce that

$$\widehat{\mathcal{L}}_U^{(k)}(\beta) = \operatorname*{argmin}_{z\in\mathbb{R}} \sum_{\tau\in\pi_N} \mathcal{R}_\tau(\beta, z), \tag{6.7.7}$$

with $\mathcal{R}_\tau(\beta,z) = \sum_{j=1}^k \rho\Big(\sqrt{n}\,\frac{\overline{\mathcal{L}}(f_\beta; G_j(\tau)) - z}{\Delta}\Big)$. Similarly, applying representation (6.7.6) to the numerator and the denominator in (6.7.4), we see that $\nabla_\beta \widehat{\mathcal{L}}_U^{(k)}(\beta)$ can be written as a weighted sum

$$\nabla_\beta \widehat{\mathcal{L}}_U^{(k)}(\beta) = \sum_{\tau\in\pi_N} \underbrace{\frac{\sum_{j=1}^k I\Big\{\Big|\overline{\mathcal{L}}(\beta; G_j(\tau)) - \widehat{\mathcal{L}}^{(k)}(\beta)\Big| \le \frac{\Delta}{\sqrt{n}}\Big\}}{\sum_{\pi\in\pi_N}\sum_{j=1}^k I\Big\{\Big|\overline{\mathcal{L}}(\beta; G_j(\pi)) - \widehat{\mathcal{L}}^{(k)}(\beta)\Big| \le \frac{\Delta}{\sqrt{n}}\Big\}}}_{=\omega_\tau,\ \text{weight corresponding to permutation } \tau} \cdot \widetilde{\Gamma}_\tau(\beta),$$

where

$$\widetilde{\Gamma}_\tau(\beta) := \frac{\sum_{j=1}^k \Big(\frac{1}{n}\sum_{i\in G_j(\tau)} Z_i\,\ell'(Y_i, f_\beta(Z_i))\Big) I\Big\{\Big|\overline{\mathcal{L}}(\beta; G_j(\tau)) - \widehat{\mathcal{L}}^{(k)}(\beta)\Big| \le \frac{\Delta}{\sqrt{n}}\Big\}}{\sum_{j=1}^k I\Big\{\Big|\overline{\mathcal{L}}(\beta; G_j(\tau)) - \widehat{\mathcal{L}}^{(k)}(\beta)\Big| \le \frac{\Delta}{\sqrt{n}}\Big\}} \tag{6.7.8}$$

is similar to the expression for the gradient of $\widehat{\mathcal{L}}^{(k)}(\beta)$ defined for a fixed partition $G_1(\tau), \dots, G_k(\tau)$, see (6.7.2). Representations in (6.7.7) and (6.7.8) can be simplified even further noting that permutations that do not alter the subgroups $G_1, \dots, G_k$ also do not change the values of $\mathcal{R}_\tau(\beta, z)$, ω_τ and $\widetilde{\Gamma}_\tau(\beta)$. To this end, let us say that $\tau_1, \tau_2 \in \pi_N$ are equivalent if $G_j(\tau_1) = G_j(\tau_2)$ for all $j = 1, \dots, k$. It is easy to see that there are $\frac{N!}{(n!)^k \cdot (N-nk)!}$ equivalence classes, and let $\pi_{N,n,k}$ be the set of permutations containing exactly one permutation from each equivalence class. We can thus write

$$\begin{aligned} \widehat{\mathcal{L}}_U^{(k)}(\beta) &= \operatorname*{argmin}_{z\in\mathbb{R}} Q(\beta, z) := \operatorname*{argmin}_{z\in\mathbb{R}} \sum_{\tau\in\pi_{N,n,k}} \mathcal{R}_\tau(\beta, z), \\ \nabla_\beta \widehat{\mathcal{L}}_U^{(k)}(\beta) &= \sum_{\tau\in\pi_{N,n,k}} \widetilde{\omega}_\tau \cdot \widetilde{\Gamma}_\tau(\beta), \end{aligned} \tag{6.7.9}$$

where $\widetilde{\omega}_\tau = (n!)^k\,(N-nk)! \cdot \omega_\tau$. Representation (6.7.9) suggests that in order to obtain an unbiased estimator of $\nabla_z Q(\beta, z)$, one can sample a permutation $\tau \in \pi_{N,n,k}$ uniformly at random, compute $\nabla_z \mathcal{R}_\tau(\beta, z)$ and use it as a descent direction. This yields a version of the stochastic gradient descent for evaluating $\widehat{\mathcal{L}}_U^{(k)}(\beta)$ presented in Figure 6.2. Once a method for computing

Figure 6.2: Algorithm 2 – evaluation of $\widehat{\mathcal{L}}_U^{(k)}(\beta)$.

Input: the dataset $(Z_i, Y_i)_{1\le i\le N}$, number of blocks $k \in \mathbb{Z}_+$, step size parameter $\eta > 0$, maximum number of iterations M, initial guess $z_0 \in \mathbb{R}$, tuning parameter $\Delta > 0$.
for all $t = 0, \dots, M$ **do**
Sample permutation τ uniformly at random from $\pi_{N,n,k}$, construct blocks $G_1(\tau), \dots, G_k(\tau)$ according to (6.7.5);
Compute $\nabla_z \mathcal{R}_\tau(\beta, z_t) = -\frac{\sqrt{n}}{\Delta}\sum_{j=1}^k \rho'\Big(\sqrt{n}\frac{\overline{\mathcal{L}}(f_\beta; G_j(\tau)) - z_t}{\Delta}\Big)$;
Update

$$z_{t+1} = z_t - \eta \nabla_z \mathcal{R}_\tau(\beta, z_t).$$

end for
Output: z_{M+1}.

$\widehat{\mathcal{L}}_U^{(k)}(\beta)$ is established, similar reasoning leads to an algorithm for finding $\widehat{f}_N^U$. Indeed, using representation (6.7.9), it is easy to see that an unbiased estimator of $\nabla_\beta \widehat{\mathcal{L}}_U^{(k)}(\beta)$ can be obtained by first sampling a permutation $\tau \in \pi_{N,n,k}$ according to the probability distribution given by the weights $\{\widetilde{\omega}_\tau,\ \tau \in \pi_{N,n,k}\}$, then evaluating $\widetilde{\Gamma}_\tau(\beta)$ via (6.7.8), and using $\widetilde{\Gamma}_\tau(\beta)$ as a direction of descent. In most typical cases, the number M of the gradient descent iterations is much smaller than $\frac{N!}{(n!)^k \cdot (N-nk)!}$, whence it is unlikely that the same permutation will be repeated twice in the sampling process. This reasoning suggests the idea of replacing the weights $\widetilde{\omega}_\tau$ by the uniform distribution over $\pi_{N,n,k}$ that leads to a much faster practical implementation which is detailed in Figure 6.3. It is easy to see that the presented gradient descent algorithms for evaluating $\widehat{f}_N$ and $\widehat{f}_N^U$ have the same numerical complexity. The following subsections provide several "proof-of-concept" examples illustrating the performance of proposed methods, as well as comparison to the existing techniques.

Figure 6.3: Algorithm 3 – evaluation of $\widehat{\beta}_N^U$.

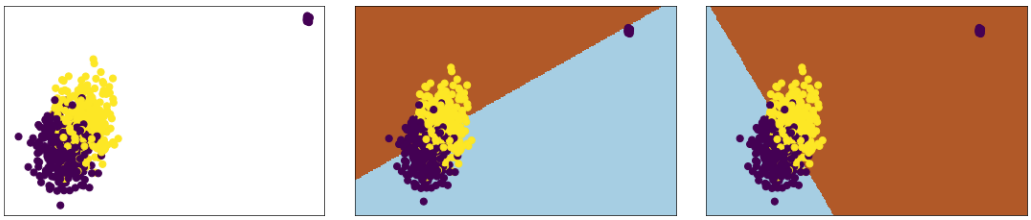
Input: the dataset $(Z_i, Y_i)_{1\le i\le N}$, number of blocks $k \in \mathbb{Z}_+$, step size parameter $\eta > 0$, maximum number of iterations M, initial guess $\beta_0 \in \mathbb{R}^d$, tuning parameter $\Delta > 0$.
for all $t = 0, \dots, M$ **do**
Sample permutation τ uniformly at random from $\pi_{N,n,k}$, construct blocks $G_1(\tau), \dots, G_k(\tau)$ according to (6.7.5);
Compute $\widehat{\mathcal{L}}_U^{(k)}(\beta_t)$ using Algorithm 2 in Figure 6.2;
Compute $\widetilde{\Gamma}_\tau(\beta_t)$ via (6.7.8);
Update

$$\beta_{t+1} = \beta_t - \eta\widetilde{\Gamma}_\tau(\beta_t).$$

end for
Output: β_{M+1}.

6.7.2 Logistic regression.

The dataset consists of pairs $(Z_j, Y_j) \in \mathbb{R}^2 \times \{\pm 1\}$, where the marginal distribution of the labels is uniform and conditional distributions of Z are normal, namely, $\mathrm{Law}(Z \,|\, Y = 1) = \mathcal{N}\big((-1,-1)^T, 1.4I_2\big)$, $\mathrm{Law}(Z \,|\, Y = -1) \sim \mathcal{N}((1,1), 1.4I_2)$, and $\Pr(Y = 1) = \Pr(Y = -1) = 1/2$; here and below, I_2 stands for the 2×2 identity matrix. The dataset includes outliers for which $Y \equiv 1$ and $Z \sim \mathcal{N}((24, 8), 0.1I_2)$. We generated $N = 600$ "informative" observations along with $\mathcal{O} = 30$ outliers, and compared the performance or our robust method (based on evaluating $\widehat{\beta}_N^U$) with the standard logistic regression that is known to be sensitive to outliers in the sample (we used implementation available in the Scikit-learn package [PVG+11]). Results of the experiment are presented in Figure 6.4 and illustrate robustness of proposed techniques. Parameters k and Δ in our implementation were tuned via cross-validation.



(a) Training Dataset

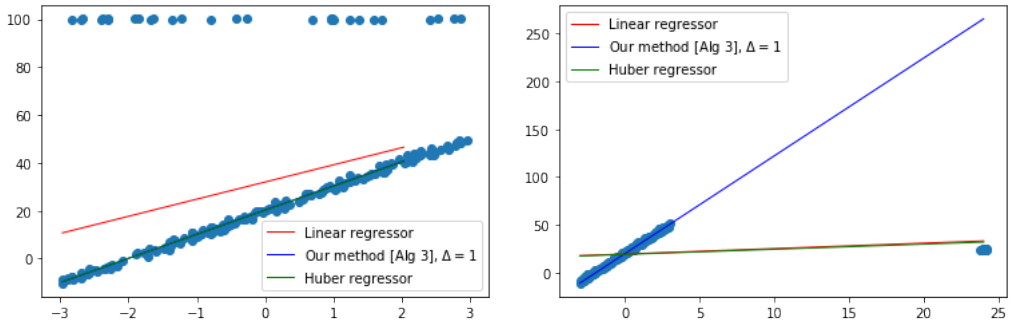
(b) Decision function – standard Logistic Regression

(c) Decision function – Algorithm 3

Figure 6.4: Scatter plot of $N^\circ = 630$ samples from the training dataset ($N = 600$ informative observations, $\mathcal{O} = 30$ outliers), the color of the points correspond to their labels and the background color – to the predicted labels (brown region corresponds to "yellow" labels and blue – to "purple").

6.7.3 Linear regression.

In this section, we compare the performance of our method (again based on evaluating $\widehat{\beta}_N^U$) with standard linear regression as well as with robust Huber's regression estimator, see Section 7 in [HR09]; linear regression and Huber's regression were implemented using 'LinearRregression' and 'HuberRegressor' functions in the Scikit-learn package [PVG+11]. As in the previous example, the dataset consists of informative observations and outliers. Informative data (Z_j, Y_j), $j = 1, \ldots, N$ for $N = 570$ are i.i.d. and satisfy the linear model $Y_j = 10Z_j + \epsilon_j + 20$ where $Z_j \sim \text{Unif}[-3, 3]$ and $\epsilon_j \sim \mathcal{N}(0, 1)$. We consider two types of outliers: (a) outliers in the response variable Y only, and (b) outliers in the predictor Z. It is well-known that standard linear regression is not robust in any of these scenarios, Huber's regression estimator is robust to outliers in response Y only, while our approach is shown to be robust to corruption of both types. In both test scenarios, we generated $\mathcal{O} = 30$ outliers. Given Z_j, the outliers Y_j of type (a) are sampled from a $\mathcal{N}(100, 0.01)$ distribution, while the outliers of type (b) are $Z_j \sim \mathcal{N}\big((24, 24)^T, 0.01\, I_2\big)$. Results are presented



(a) Outliers in response variable

(b) Outliers in predictors

Figure 6.5: Scatter plot of $N^\circ = 600$ training samples ($N = 570$ informative data and $\mathcal{O} = 30$ outliers) and the corresponding regression lines for our method, Huber's regression and regression with quadratic loss.

in Figure 6.5, and confirm the expected outcomes.

6.7.4 Choice of k and Δ.

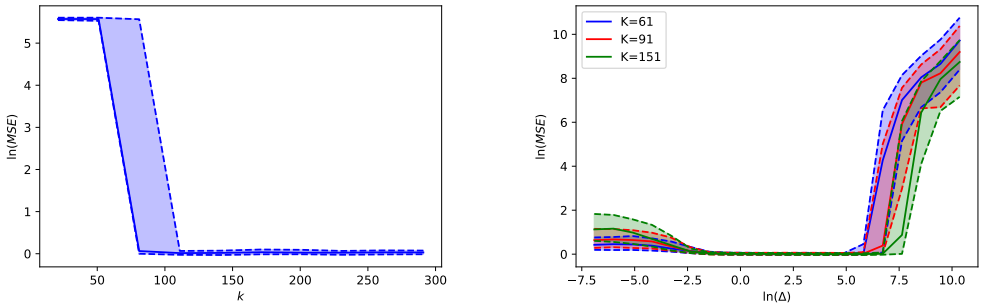
In this subsection, we evaluate the effect of different choices of k and Δ in the linear regression setting of Section 6.7.3, again with $N = 570$ informative observations and $\mathcal{O} = 30$ outliers of type (b) as described in Section 6.7.3 above. Figure 6.6a shows the plot of the resulting mean square error (MSE) against the number of subgroups k. As expected, the error decreases significantly when k exceeds 60, twice the number of outliers. At the same time, the MSE remains stable as k grows up to $k \simeq 100$, which is a desirable property for practical applications. In this experiment, Δ was set using the "median absolute deviation" (MAD) estimator defined as follows. We start with Δ_0 being a small number (e.g., $\Delta_0 = 0.1$). Given a current approximate solution β_t, a permutation τ and the corresponding subgroups $G_1(\tau), \ldots, G_k(\tau)$, set $\widehat{M}(\beta_t) :=$

$$\text{median}\Big(\widehat{\mathcal{L}}^{(k)}(\beta_t; G_1(\tau)), \dots, \widehat{\mathcal{L}}^{(k)}(\beta_t; G_k(\tau))\Big), \text{ and}$$

$$\text{MAD}(\beta_t) = \text{median}\Big(\Big|\widehat{\mathcal{L}}^{(k)}(\beta_t; G_1(\tau)) - \widehat{M}(\beta_t)\Big|, \dots, \Big|\widehat{\mathcal{L}}^{(k)}(\beta_t; G_k(\tau)) - \widehat{M}(\beta_t)\Big|\Big).$$

Finally, define $\widehat{\Delta}_{t+1} := \frac{\text{MAD}(\beta_t)}{\Phi^{-1}(3/4)}$, where Φ is the distribution function of the standard normal law. After a small number m (e.g. $m = 10$) of "burn-in" iterations of Algorithm 3, Δ is fixed at the level $\widehat{\Delta}_m$ for all the remaining iterations.

Next, we study the effect of varying Δ for different but fixed values of k. To this end, we set $k \in \{61, 91, 151\}$, and evaluated the MSE as a function of Δ. The resulting plot is presented in Figure 6.6b. The MSE achieves its minimum for $\Delta \asymp 10^2$; for larger values of Δ, the effect of outliers becomes significant as the algorithm starts to resemble regression with quadratic loss (indeed, outliers in this specific example are at a distance ≈ 100 from the bulk of the data).



(a) MSE vs k (b) MSE vs Δ (log-log scale)

Figure 6.6: Plot of the tuning parameter (x-axis) against the MSE (y-axis) obtained with Algorithm 3. The MSE was evaluated via 300 runs of the Monte-Carlo simulation with $N = 500$ samples of the data. The dotted lines show the interquartile range (25%-75%) over the runs of Monte-Carlo.

Comparison with existing methods.

In this section, we compare the performance of Algorithm 3 with a median-of-means-based robust gradient descent algorithm studied in [LLM20]. The main difference of this method is in the way the descent direction is computed at every step. Specifically, $\widetilde{\Gamma}_\tau(\beta)$ employed in Algorithm 3 is replaced by $\nabla_\beta \mathcal{L}^\diamond(\beta)$ where $\mathcal{L}^\diamond(\beta) := \text{median}\big(\overline{\mathcal{L}}(\beta; G_1(\tau)), \dots, \overline{\mathcal{L}}(\beta; G_k(\tau))\big)$, see Figure 6.7 and paper [LLM20] for the detailed description. Experiments were performed for the logistic regression problem based on the "two moons" pattern, one of the standard datasets in the Scikit-learn package [PVG+11] presented in Figure 6.8a. We performed two sets of experiments, one on the outlier-free dataset and one on the dataset consisting of 90% of informative observations and 10% of outliers, depicted as a yellow dot with coordinates $(0, 5)$ on the plot. In both scenarios, we tested the "small" ($N = 100$) and "moderate" ($N = 1000$) sample size regimes. We used standard logistic regression trained on an outlier-free sample as a benchmark; its accuracy is shown as

Figure 6.7: Algorithm 4.

Input: the dataset $(Z_i, Y_i)_{1\le i\le N}$, number of blocks $k \in \mathbb{Z}_+$, step size parameter $\eta > 0$, maximum number of iterations M, initial guess $\beta_0 \in \mathbb{R}^d$, tuning parameter $\Delta > 0$.
for all $t = 0, \dots, M$ **do**
Sample permutation τ uniformly at random from $\pi_{N,n,k}$, construct blocks $G_1(\tau), \dots, G_k(\tau)$ according to (6.7.5);
Compute $\nabla_\beta \mathcal{L}^\diamond(\beta)$;
Update

$$\beta_{t+1} = \beta_t - \eta \nabla_\beta \mathcal{L}^\diamond(\beta).$$

end for
Output: β_{M+1}.

a dotted red line on the plots. In all the cases, parameter Δ was tuned via cross-validation. In the outlier-free setting, our method (based on Algorithm 3) performed nearly as good as logistic regression; notably, performance of the method was strong even for large values of k, while classification accuracy decreased noticeably for Algorithm 4 for large k. In the presence of outliers, our method performed similar to Algorithm 4, while both methods outperformed standard logistic regression; for large values of k, our method was again slightly better. At the same time, Algorithm 4 was consistently faster than Algorithm 3 across the experiments.

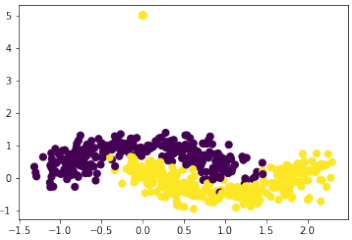
A remark on cross-validation in a corrupted setting

Cross-validation is a common way to assess the performance of a machine learning algorithm. However, cross-validation is not robust when the method itself is not robust (as it is the case here with regression with quadratic loss). For our purposes, we slightly changed the way we approach cross validation. Namely, we still partition the data into m parts used separately for training and testing, however, once we obtain the m scores associated with the m folds, we evaluate the median of these scores instead of the mean; see Figure 6.9 for the details. The rationale behind this approach is that if at least half of the folds do not contain outliers, the results of cross-validation will be robust. To use this approach, we choose m, the number of folds, to be large (in the example above, $m = 500$).

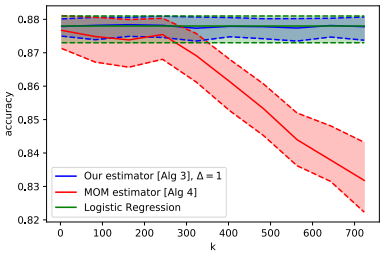
We compared the three algorithms using robust cross-validation with median described above Our method (based on Algorithm 3) yields MSE of $\simeq e^{4.2}$ while the MSE for the ordinary least squares regression is of order $e^{22.1}$, while the Huber Regression leads to MSE $\simeq e^{8.9}$. The empirical density of the logarithm of the MSE over 500 folds is shown in Figure 6.10.

6.7.5 Comparison of Algorithm 1 and Algorithm 3.

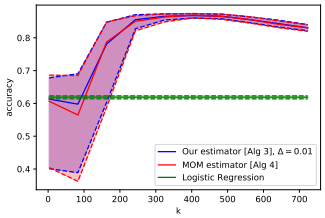
We present a numerical evidence that the permutation-invariant estimator $\widehat{f}_N^U$ is superior to the the estimator $\widehat{f}_N$ based on fixed partition of the dataset. Evaluation was performed for the regression task where the data contained outliers of type (a), as described in Section 6.7.3. Average MSE was evaluated over 500 repetitions of the experiment, and the standard deviation of the MSE was also recored. Results are presented in Figure 6.11 and confirm the significant



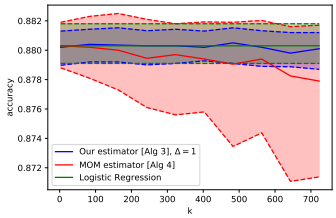
(a) "Two moons" dataset [PVG+11] with outliers.



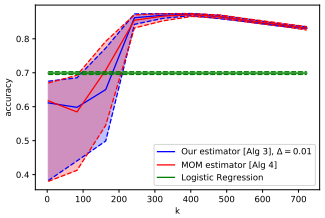
(b) $N = 100$, no outliers



(c) $N = 100$, with 10 outliers



(d) $N = 1000$, no outliers



(e) $N = 1000$, with 100 outliers

Figure 6.8: Comparison of Algorithm 3, Algorithm 4 and standard logistic regression. The accuracy was evaluated using Monte-Carlo simulation over 300 runs. The dotted lines show the interquartile range (25%-75%) over the runs of Monte-Carlo.

improvements achieved by Algorithm 3 over Algorithm 1. We set $k = 71$ and $\Delta = 1$ for both algorithms.

6.7.6 Application to the "Communities and Crime" data.

We compare the performance of our methods with the least squares regression applied to a real dataset. The dataset we chose is called "Communities and Crime Unnormalized Data Set" and is available through the UCI Machine Learning Repository. These data contain 2215 observations from a census and law enforcement records. The task we devised was to predict the crime activity (represented as the count of incidents) using the following features: the population of the area, the per capita income, the median family income, the number of vacant houses, and the land area. The choice of this specific dataset was motivated by the fact that it likely contains a non-negligible number of outliers due to the nature of the features and the fact that the data have not been preprocessed, hence the advantages of proposed approach could be highlighted. We regularized both the robust risk and the usual squared loss with $\|\cdot\|_2$ norm (ridge regression) where the regularization parameter was selected using cross-validation. Figure 6.12 presents a pairplot of the dataset; specifically, a pairplot shows all the different scatter plots of one feature versus another (hence, the diagonal consists of the histograms of an individual feature). Such a pairplot offers a visual confirmation of the fact that the data likely contains outliers. We were interested in the dependence of the MSE on the partition cardinality k. Similarly to Figure 6.6a,

Figure 6.9: Robust cross-validation with the median.

Input: the dataset $(X_i, Y_i)_{1 \le i \le N}$, the number of folds m.
Construct the blocks $G_1, \dots, G_m$, partition of $\{1, \dots, N\}$.
for all $j = 1, \dots, m$ **do**
  Train $\widehat{f}$ on the dataset $(X_l, Y_l), l \in \bigcup_{i \neq j} G_i$.
  Compute the test MSE $\text{Score}_j = \frac{1}{|G_j|} \sum_{l \in G_j} (\hat{f}(X_l) - Y_l)^2$
end for
Output: $\text{Median}(\text{Score}_1, \dots, \text{Score}_m)$.

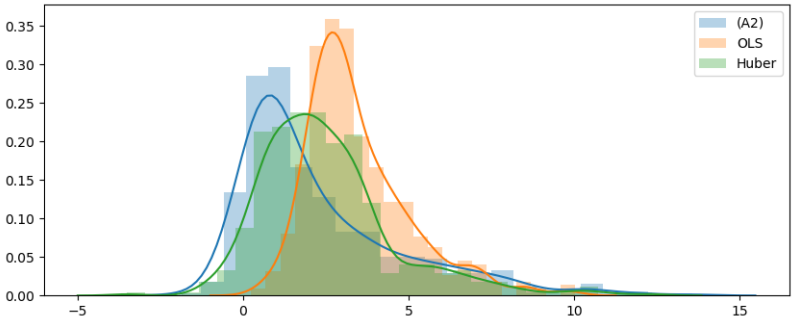


Figure 6.10: Histogram of densities of the logarithm of the MSE for the different methods (light blue corresponds to the approach of this paper (Algorithm 3), orange - to the standard least squares regression, and green - to Huber's regression).

we plotted the MSE as a function of k (Figure 6.13).

	Algorithm 1	Algorithm 3
average MSE	97.8	2
standard deviation of MSE	577.3	13

Figure 6.11: Comparison of Algorithms 1 and 3.

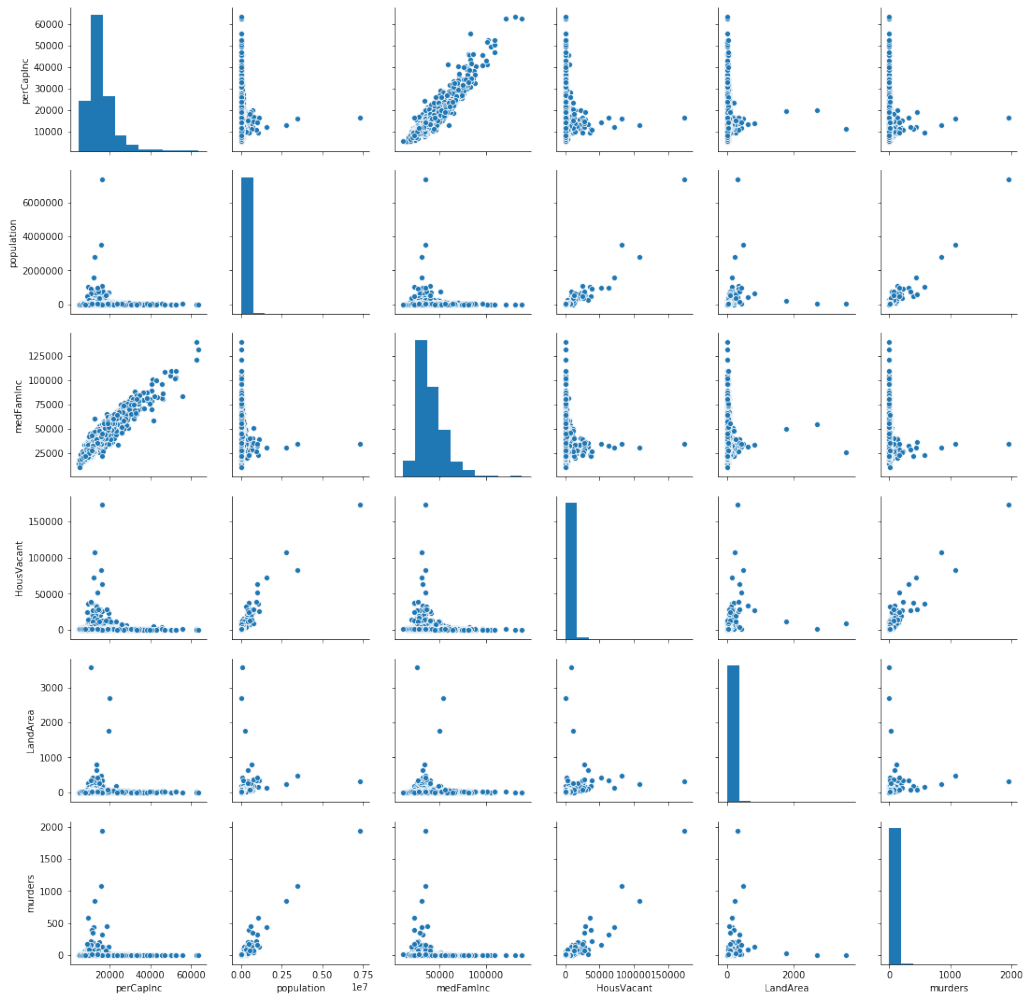


Figure 6.12: Pairplot detailing the 2D marginals of the dataset.

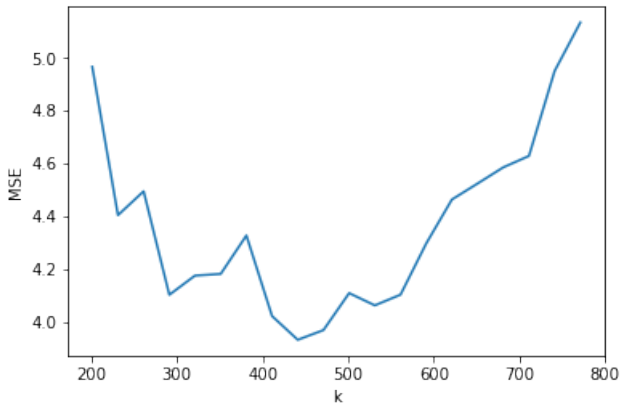


Figure 6.13: Plot of the number of blocks k (x-axis) vs the test mean squared error (y-axis) obtained with Algorithm 3 on 500 folds.

Chapter 7

MONK – Outlier-Robust Mean Embedding Estimation by Median-of-Means

Abstract

Mean embeddings provide an extremely flexible and powerful tool in machine learning and statistics to represent probability distributions and define a semi-metric (MMD, maximum mean discrepancy; also called N-distance or energy distance), with numerous successful applications. The representation is constructed as the expectation of the feature map defined by a kernel. As a mean, its classical empirical estimator, however, can be arbitrary severely affected even by a single outlier in case of unbounded features. To the best of our knowledge, unfortunately even the consistency of the existing few techniques trying to alleviate this serious sensitivity bottleneck is unknown. In this paper, we show how the recently emerged principle of median-of-means can be used to design estimators for kernel mean embedding and MMD with excessive resistance properties to outliers, and optimal sub-Gaussian deviation bounds under mild assumptions.

7.1 Introduction

Kernel methods [Aro50] form the backbone of a tremendous number of successful applications in machine learning thanks to their power in capturing complex relations [SS02, SC08]. The main idea behind these techniques is to map the data points to a feature space (RKHS, reproducing kernel Hilbert space) determined by the kernel, and apply linear methods in the feature space, without the need to explicitly compute the map.

One crucial component contributing to this flexibility and efficiency (beyond the solid theoretical foundations) is the versatility of domains where kernels exist; examples include trees [CD01a, KK02], time series [Cut11], strings [LSST+02], mixture models, hidden Markov models or linear dynamical systems [JKH04], sets [Hau99, GFKS02], fuzzy domains [GHC17], distribu-

tions [HB05, MSX+09, MFDS11], groups [CFV05] such as specific constructions on permutations [JV16], or graphs [VSKB10, KP16].

Given a kernel-enriched domain $(\mathcal{X}, K)$ one can represent probability distributions on $\mathcal{X}$ as a mean

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_K, \quad \varphi(x) := K(\cdot, x),$$

which is a point in the RKHS determined by K. This representation called *mean embedding* [BTA04, SGSS07] induces a semi-metric[1] on distributions called maximum mean discrepancy (MMD) [SGSS07, GBR+12]

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_K}. \tag{7.1.1}$$

With appropriate choice of the kernel, classical integral transforms widely used in probability theory and statistics can be recovered by $\mu_{\mathbb{P}}$; for example, if $\mathcal{X}$ equipped with the scalar product $\langle \cdot, \cdot \rangle$ is a Hilbert space, the kernel $K(x, y) = \mathrm{e}^{\langle x, y \rangle}$ gives the moment-generating function, $K(x, y) = \mathrm{e}^{\gamma \|x-y\|_2^2}$ $(\gamma > 0)$ the Weierstrass transform. As it has been shown [SKGF13] energy distance [BF04, SR04, SR05]—also known as N-distance [ZKK92, Kle05] in the statistical literature—coincides with MMD.

Mean embedding and maximum mean discrepancy have been applied successfully, in kernel Bayesian inference [SGB+11, FSG13], approximate Bayesian computation [PJS16], model criticism [LDG+14, KKK16], two-sample [BF04, SR04, SR05, HBM07, GBR+12] or its differential private variant [RLSP18], independence [GFT+08, PBSP17] and goodness-of-fit testing [JXS+17, BLY17], domain adaptation [ZSMW13] and generalization [BDD+17], change-point detection [HC07], probabilistic programming [SMF+15], post selection inference [YUFT18], distribution classification [MFDS11, ZKR+17] and regression [SBPG16, LSSF18], causal discovery [MPJ+16, PBSP17], generative adversarial networks [DRG15, LSZ15, BSAG18], understanding the dynamics of complex dynamical systems [KSM18, KBSS19], or topological data analysis [KFH16], among many others; [MFBS17] provide a recent in-depth review on the topic.

Crucial to the success of these applications is the efficient and robust approximation of the mean embedding and MMD. As a mean, the most natural approach to estimate $\mu_{\mathbb{P}}$ is the empirical average. Plugging this estimate into Eq. (7.1.1) produces directly an approximation of MMD, which can also be made unbiased (by a small correction) or approximated recursively. These are the V-statistic, U-statistic and online approaches [GBR+12]. Kernel mean shrinkage estimators [MKF+16] represent an other successful direction: they improve the efficiency of the mean embedding estimation by taking into account the Stein phenomenon. Minimax results have recently been established: the optimal rate of mean embedding estimation given N samples from $\mathbb{P}$ is $N^{-1/2}$ [TKM17] for discrete measures and the class of measures with infinitely differentiable density when K is a continuous, shift-invariant kernel on $\mathcal{X} = \mathbb{R}^d$. For MMD, using N_1 and N_2 samples from $\mathbb{P}$ and $\mathbb{Q}$, it is $N_1^{-1/2} + N_2^{-1/2}$ [TKS16] in case of radial universal kernels defined on $\mathcal{X} = \mathbb{R}^d$.

A critical property of an estimator is its robustness to contaminated data, outliers which are omnipresent in currently available massive and heterogenous datasets. To the best of our knowledge, systematically *designing outlier-robust mean embedding and MMD estimators* has hardly been touched in the literature; this is the focus of the current paper. The issue is

[1] [FGSS08, KGF+10] provide conditions when MMD is a metric, i.e. μ is injective.

particularly serious in case of unbounded kernels when for example even a single outlier can ruin completely a classical empirical average based estimator. Examples for unbounded kernels are the exponential kernel (see the example above about moment-generating functions), polynomial kernel, string, time series or graph kernels.

Existing related techniques comprise robust kernel density estimation (KDE) [KS12]: the authors elegantly combine ideas from the KDE and M-estimator literature to arrive at a robust KDE estimate of density functions. They assume that the underlying smoothing kernels[2] are shift-invariant on $\mathcal{X} = \mathbb{R}^d$ and reproducing, and interpret KDE as a weighted mean in $\mathcal{H}_K$. The idea has been (i) adapted to construct outlier-robust covariance operators in RKHSs in the context of kernel canonical correlation analysis [AFW18], and (ii) relaxed to general Hilbert spaces [SGRA18]. Unfortunately, the consistency of the investigated empirical M-estimators is unknown, except for finite-dimensional feature maps [SGRA18], or as density function estimators [VS13].

To achieve our goal, we leverage the idea of Median-Of-meaNs (MON). Intuitively, MONs replace the linear operation of expectation with the median of averages taken over non-overlapping blocks of the data, in order to get a robust estimate thanks to the median step. MONs date back to [JGV86, AMS99, NY83] for the estimation of the mean of real-valued random variables. Their concentration properties have been recently studied by [DLLO16, MS17] following the approach of [Cat12] for M-estimators. These studies focusing on the estimation of the mean of real-valued random variables are important as they can be used to tackle more general prediction problems in learning theory via the classical empirical risk minimization approach [Vap00] or by more sophisticated approach such as the minmax procedure [AC11].

In parallel to the minmax approach, there have been several attempts to extend the usage of MON estimators from $\mathbb{R}$ to more general settings. For example, [Min15, MS17] consider the problem of estimating the mean of a Banach-space valued random variable using "geometrical" MONs. The estimators constructed by [Min15, MS17] are computationally tractable but the deviation bounds are suboptimal compared to those one can prove for the empirical mean under sub-Gaussian assumptions. In regression problems, [LM19c, LL18] proposed to combine the classical MON estimators on $\mathbb{R}$ in a "test" procedure that can be seen as a Le Cam test estimator [Le 73]. The achievement in [LM19c, LL18] is that they were able to obtain optimal deviation bounds for the resulting estimator using the powerful so-called small-ball method of [KM15, Men15]. This approach was then extended to mean estimation $\mathbb{R}^d$ by [LM$^+$19d] providing the first rate-optimal sub-Gaussian deviation bounds under minimal L^2-assumptions. The constants of [LM19c, LL18, LM$^+$19d] have been improved by [CG17] for the estimation of the mean in $\mathbb{R}^d$ under L^4-moment assumption and in least-squares regression under L^4/L^2-condition that is stronger than the small-ball assumption used by [LM19c, LL18]. Unfortunately, these estimators are computationally intractable; their risk bounds however serve as an important baseline for computable estimators such as the minmax MON estimators in regression [LL20].

Motivated by the computational intractability of the tournament procedure underlying the first rate-optimal sub-Gaussian deviation bound holding under minimal assumptions in $\mathbb{R}^d$ [LM$^+$19d], [Hop20] proposed a convex relaxation with polynomial, $O(N^{24})$ complexity where N denotes the sample size. [CFB19] have recently designed an alternative convex relaxation requiring $O(N^4 + dN^2)$ computation which is still rather restrictive for large sample size and infeasible in

[2]Smoothing kernels extensively studied in the non-parametric statistical literature [GKKW02] are assumed to be non-negative functions integrating to one.

infinite dimension.

Our goal is to extend the theoretical insight of [LM+19d] from $\mathbb{R}^d$ to kernel-enriched domains. Particularly, we prove optimal sub-Gaussian deviation bounds for MON-based mean estimators in RKHS-s which hold under minimal second-order moment assumptions. In order to achieve this goal, we use a different (minmax [AC11, LL20]) construction which combined with properties specific to RKHSs (the mean-reproducing property of mean embedding and the integral probability metric representation of MMD) give rise to our practical MONK procedures. Thanks to the usage of medians the MONK estimators are also robust to contamination.

Section 7.2 contains definitions and problem formulation. Our main results are given in Section 7.3. Implementation of the MONK estimators is the focus of Section 7.4, with numerical illustrations in Section 7.5.

7.2 Definitions & Problem Formulation

In this section, we formally introduce the goal of our paper.

Notations: $\mathbb{Z}^+$ is the set of positive integers. $[M] := \{1, \dots, M\}$, $u_S := (u_m)_{m\in S}$, $S \subseteq [M]$. For a set S, $|S|$ denotes its cardinality. $\mathbb{E}$ stands for expectation. $\mathrm{med}_{q\in[Q]}\{z_q\}$ is the median of the $(z_q)_{q\in[Q]}$ numbers. Let $\mathcal{X}$ be a separable topological space endowed with the Borel σ-field, $x_{1:N}$ denotes a sequence of i.i.d. random variables on $\mathcal{X}$ with law $\mathbb{P}$ (shortly, $x_{1:N} \sim \mathbb{P}$). $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is a continuous (reproducing) kernel on $\mathcal{X}$, $\mathcal{H}_K$ is the reproducing kernel Hilbert space associated to K; $\langle\cdot,\cdot\rangle_K := \langle\cdot,\cdot\rangle_{\mathcal{H}_K}$, $\|\cdot\|_K := \|\cdot\|_{\mathcal{H}_K}$.[3] The reproducing property of the kernel means that evaluation of functions in $\mathcal{H}_K$ can be represented by inner products $f(x) = \langle f, K(\cdot,x)\rangle_K$ for all $x \in \mathcal{X}$, $f \in \mathcal{H}_K$. The mean embedding of a probability measure $\mathbb{P}$ is defined as

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} K(\cdot,x)\mathrm{d}\mathbb{P}(x) \in \mathcal{H}_K, \tag{7.2.1}$$

where the integral is meant in Bochner sense; $\mu_{\mathbb{P}}$ exists iff $\int_{\mathcal{X}}\|K(\cdot,x)\|_K\mathrm{d}\mathbb{P}(x) = \int_{\mathcal{X}}\sqrt{K(x,x)}\mathrm{d}\mathbb{P}(x) < \infty$. It is well-known that the mean embedding has mean-reproducing property $\mathbb{P}f := \mathbb{E}_{x\sim\mathbb{P}}f(x) = \langle f, \mu_{\mathbb{P}}\rangle_K$ for all $f \in \mathcal{H}_K$, and it is the unique solution of the problem:

$$\mu_{\mathbb{P}} = \operatorname*{argmin}_{f\in\mathcal{H}_K} \int_{\mathcal{X}} \|f - K(\cdot,x)\|_K^2 \mathrm{d}\mathbb{P}(x)\,. \tag{7.2.2}$$

The solution of this task can be obtained by solving the following minmax optimization

$$\mu_{\mathbb{P}} = \operatorname*{argmin}_{f\in\mathcal{H}_K} \sup_{g\in\mathcal{H}_K} J(f,g), \tag{7.2.3}$$

with $J(f,g) = \mathbb{E}_{x\sim\mathbb{P}}\Big[\|f - K(\cdot,x)\|_K^2 - \|g - K(\cdot,x)\|_K^2\Big]$. The equivalence of (7.2.2) and (7.2.3) is obvious since the expectation is linear. Nevertheless, this equivalence is essential in the construction of our estimators because we will below replace the expectation by a non-linear estimator of this quantity. More precisely, the unknown expectations are computed by using

[3] $\mathcal{H}_K$ is separable by the separability of $\mathcal{X}$ and the continuity of K [SC08, Lemma 4.33]. These assumptions on $\mathcal{X}$ and K are assumed to hold throughout the paper.

the Median-of-meaN estimator (MON). Given a partition of the dataset into blocks, the MON estimator is the median of the empirical means over each block. MON estimators are naturally robust thanks to the median step.

More precisely, the procedure goes as follows. For any map $h : \mathcal{X} \to \mathbb{R}$ and any non-empty subset $S \subseteq [N]$, denote by $\mathbb{P}_S := |S|^{-1} \sum_{i \in S} \delta_{x_i}$ the empirical measure associated to the subset x_S and $\mathbb{P}_S h = |S|^{-1} \sum_{i \in S} h(x_i)$; we will use the shorthand $\mu_S := \mu_{\mathbb{P}_S}$. Assume that $N \in \mathbb{Z}^+$ is divisible by $Q \in \mathbb{Z}^+$ and let $(S_q)_{q \in [Q]}$ denote a partition of $[N]$ into subsets with the same cardinality $|S_q| = N/Q$ ($\forall q \in [Q]$). The Median Of meaN (MON) is defined as

$$\mathrm{MON}_Q[h] = \mathrm{med}_{q \in [Q]} \left\{ \mathbb{P}_{S_q} h \right\} = \mathrm{med}_{q \in [Q]} \left\{ \left\langle h, \mu_{S_q} \right\rangle_K \right\},$$

where assuming that $h \in \mathcal{H}_K$ the second equality is a consequence of the mean-reproducing property of $\mu_{\mathbb{P}}$. Specifically, in case of $Q = 1$ the MON operation reduces to the classical mean: $\mathrm{MON}_1[h] = N^{-1} \sum_{n=1}^N h(x_n)$.

We define the minmax MON-based estimator associated to kernel K (MONK) as

$$\hat{\mu}_{\mathbb{P},Q} = \hat{\mu}_{\mathbb{P},Q}(x_{1:N}) \in \operatorname*{argmin}_{f \in \mathcal{H}_K} \sup_{g \in \mathcal{H}_K} \tilde{J}(f, g),$$

where for all $f, g \in \mathcal{H}_K$

$$\tilde{J}(f, g) = \mathrm{MON}_Q \Big[x \mapsto \|f - K(\cdot, x)\|_K^2 - \|g - K(\cdot, x)\|_K^2 \Big].$$

When $Q = 1$, since $\mathrm{MON}_1[h]$ is the empirical mean, we obtain the classical empirical mean based estimator: $\hat{\mu}_{\mathbb{P},1} = \frac{1}{N} \sum_{n=1}^N K(\cdot, x_n)$.

One can use the mean embedding (7.2.1) to get a semi-metric on probability measures: the maximum mean discrepancy (MMD) of $\mathbb{P}$ and $\mathbb{Q}$ is

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) := \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_K = \sup_{f \in B_K} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_K,$$

where $B_K = \{f \in \mathcal{H}_K : \|f\|_K \leq 1\}$ is the closed unit ball around the origin in $\mathcal{H}_K$. The second equality shows that MMD is a specific integral probability metric [Mül97, Zol83]. Assume that we have access to $x_{1:N} \sim \mathbb{P}$, $y_{1:N} \sim \mathbb{Q}$ samples, where we assumed the size of the two samples to be the same for simplicity. Denote by $\mathbb{P}_{S,x} := \frac{1}{|S|} \sum_{i \in S} \delta_{x_i}$ the empirical measure associated to the subset x_S ($\mathbb{P}_{S,y}$ is defined similarly for y), $\mu_{S_q,\mathbb{P}} := \mu_{\mathbb{P}_{S_q,x}}$, $\mu_{S_q,\mathbb{Q}} := \mu_{\mathbb{P}_{S_q,y}}$. We propose the following MON-based MMD estimator

$$\widehat{\mathrm{MMD}}_Q(\mathbb{P}, \mathbb{Q}) = \sup_{f \in B_K} \operatorname*{med}_{q \in [Q]} \left\{ \left\langle f, \mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}} \right\rangle_K \right\}. \tag{7.2.4}$$

Again, with the $Q = 1$ choice, the classical V-statistic based MMD estimator [GBR+12] is recovered:

$$\begin{aligned} \widehat{\mathrm{MMD}}(\mathbb{P}, \mathbb{Q}) &= \sup_{f \in B_K} \left[\frac{1}{N} \sum_{n \in [N]} f(x_n) - \frac{1}{N} \sum_{n \in [N]} f(y_n) \right] \\ &= \sqrt{\frac{1}{N^2} \sum_{i,j \in [N]} \left(K_{ij}^x + K_{ij}^y - 2K_{ij}^{xy} \right)}, \end{aligned} \tag{7.2.5}$$

where $K^x_{ij} = K(x_i, x_j), K^y_{ij} = K(y_i, y_j)$ and $K^{xy}_{ij} = K(x_i, y_j)$ for all $i, j \in [N]$. Changing in Eq. (7.2.5) $\sum_{i,j\in[N]}$ to $\sum_{i,j\in[N],i\neq j}$ in case of the K^x_{ij} and K^y_{ij} terms gives the (unbiased) U-statistic based MMD estimator

$$\frac{1}{N(N-1)} \sum_{\substack{i,j\in[N]\\ i\neq j}} \left(K^x_{ij} + K^y_{ij}\right) - \frac{2}{N^2} \sum_{i,j\in[N]} K^{xy}_{ij}. \tag{7.2.6}$$

Our **goal** is to lay down the theoretical foundations of the $\hat{\mu}_{\mathbb{P},Q}$ and $\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q})$ MONK estimators: study their finite-sample behaviour (prove optimal sub-Gaussian deviation bounds) and establish their outlier-robustness properties.

A **few additional notations** will be needed throughout the paper. $S_1 \backslash S_2$ is the difference of set S_1 and S_2. For any linear operator $A : \mathcal{H}_K \to \mathcal{H}_K$, denote by $\|A\| := \sup_{0\neq f\in\mathcal{H}_K} \|Af\|_K / \|f\|_K$ the operator norm of A. Let $\mathcal{L}(\mathcal{H}_K) = \{A : \mathcal{H}_K \to \mathcal{H}_K \text{ linear operator} : \|A\| < \infty\}$ be the space of bounded linear operators. For any $A \in \mathcal{L}(\mathcal{H}_K)$, let $A^* \in \mathcal{L}(\mathcal{H}_K)$ denote the adjoint of A, that is the operator such that $\langle Af, g\rangle_K = \langle f, A^*g\rangle_K$ for all $f, g \in \mathcal{H}_K$. An operator $A \in \mathcal{L}(\mathcal{H}_K)$ is called non-negative if $\langle Af, f\rangle_K \geq 0$ for all $f \in \mathcal{H}_K$. By the separability of $\mathcal{H}_K$, there exists a countable orthonormal basis (ONB) $(e_i)_{i\in I}$ in $\mathcal{H}_K$. $A \in \mathcal{L}(\mathcal{H}_K)$ is called trace-class if $\|A\|_1 := \sum_{i\in I} \left\langle (A^*A)^{1/2} e_i, e_i \right\rangle_K < \infty$ and in this case $\mathrm{Tr}(A) := \sum_{i\in I} \langle Ae_i, e_i\rangle_K < \infty$. If A is non-negative and self-adjoint, then A is trace class iff $\mathrm{Tr}(A) < \infty$; this will hold for the covariance operator ($\Sigma_{\mathbb{P}}$, see Eq. (7.2.7)). $A \in \mathcal{L}(\mathcal{H}_K)$ is called Hilbert-Schmidt if $\|A\|_2^2 := \mathrm{Tr}(A^*A) = \sum_{i\in I} \langle Ae_i, Ae_i\rangle_K < \infty$. One can show that the definitions of trace-class and Hilbert-Schmidt operators are independent of the particular choice of the ONB $(e_i)_{i\in I}$. Denote by $\mathcal{L}_1(\mathcal{H}_K) := \{A \in \mathcal{L}(\mathcal{H}_K) : \|A\|_1 < \infty\}$ and $\mathcal{L}_2(\mathcal{H}_K) := \{A \in \mathcal{L}(\mathcal{H}_K) : \|A\|_2 < \infty\}$ the class of trace-class and (Hilbert) space of Hilbert-Schmidt operators on $\mathcal{H}_K$, respectively. The tensor product of $a, b \in \mathcal{H}_K$ is $(a \otimes b)(c) = a\langle b, c\rangle_K, \quad (\forall c \in \mathcal{H}_K)$, $a \otimes b \in \mathcal{L}_2(\mathcal{H}_K)$ and $\|a \otimes b\|_2 = \|a\|_K \|b\|_K$. $\mathcal{L}_2(\mathcal{H}_K) \cong \mathcal{H}_K \otimes \mathcal{H}_K$ where the r.h.s. denotes the tensor product of Hilbert spaces defined as the closure of $\left\{ \sum_{i=1}^n a_i \otimes b_i : a_i, b_i \in \mathcal{H}_K \, (i \in [n]), \, n \in \mathbb{Z}^+ \right\}$. Whenever $\int_{\mathcal{X}} \|K(\cdot, x) \otimes K(\cdot, x)\|_2 \mathrm{d}\mathbb{P}(x) = \int_{\mathcal{X}} K(x, x) \mathrm{d}\mathbb{P}(x) < \infty$, let $\Sigma_{\mathbb{P}}$ denote the covariance operator

$$\Sigma_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}}([K(\cdot, x) - \mu_{\mathbb{P}}] \otimes [K(\cdot, x) - \mu_{\mathbb{P}}]) \in \mathcal{L}_2(\mathcal{H}_K), \tag{7.2.7}$$

where the expectation (integral) is again meant in Bochner sense. $\Sigma_{\mathbb{P}}$ is non-negative, self-adjoint, moreover it has covariance-reproducing property $\langle f, \Sigma_{\mathbb{P}} f\rangle_K = \mathbb{E}_{x\sim\mathbb{P}}[f(x) - \mathbb{P}f]^2$. It is known that $\|A\| \leq \|A\|_2 \leq \|A\|_1$.

7.3 Main Results

Below we present our main results on the MONK estimators, followed by a discussion. We allow that N_c elements$((x_{n_j})_{j=1}^{N_c}$) of the sample $x_{1:N}$ are arbitrarily corrupted (In MMD estimation $\{(x_{n_j}, y_{n_j})\}_{j=1}^{N_c}$ can be contaminated). The number of corrupted samples can be (almost) half of the number of blocks, in other words, there exists $\delta \in (0, 1/2]$ such that $N_c \leq Q(1/2 - \delta)$. If the data are free from contaminations, then $N_c = 0$ and $\delta = 1/2$. Using these notations, we can prove the following optimal sub-Gaussian deviation bounds on the MONK estimators.

Theorem 37 (Consistency & outlier-robustness of $\hat{\mu}_{\mathbb{P},Q}$)**.** *Assume that $\Sigma_{\mathbb{P}} \in \mathcal{L}_1(\mathcal{H}_K)$. Then, for any $\eta \in (0, 1)$ such that $Q = 72\delta^{-2} \ln(1/\eta)$ satisfies $Q \in (N_c/(1/2 - \delta), N/2)$, with probability at*

least $1-\eta$,

$$\|\hat{\mu}_{\mathbb{P},Q} - \mu_{\mathbb{P}}\|_K \leq \frac{12\big(1+\sqrt{2}\big)}{\delta} \max\left(\sqrt{\frac{6\|\Sigma_{\mathbb{P}}\| \ln(1/\eta)}{\delta N}}, 2\sqrt{\frac{\mathrm{Tr}\left(\Sigma_{\mathbb{P}}\right)}{N}}\right).$$

Theorem 38 (Consistency & outlier-robustness of $\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q})$)**.** *Assume that* $\Sigma_{\mathbb{P}}$ *and* $\Sigma_{\mathbb{Q}} \in \mathcal{L}_1(\mathcal{H}_K)$*. Then, for any* $\eta \in (0,1)$ *such that* $Q = 72\delta^{-2}\ln(1/\eta)$ *satisfies* $Q \in (N_c/(1/2-\delta), N/2)$*, with probability at least* $1-\eta$*,*

$$\left|\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}(\mathbb{P},\mathbb{Q})\right| \leq \frac{12\max\left(\sqrt{\frac{(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)\ln(1/\eta)}{\delta N}}, 2\sqrt{\frac{\mathrm{Tr}\left(\Sigma_{\mathbb{P}}\right)+\mathrm{Tr}\left(\Sigma_{\mathbb{Q}}\right)}{N}}\right)}{\delta}.$$

Proof (sketch). The technical challenge is to get the optimal deviation bounds under the (mild) trace-class assumption. The reasonings for the mean embedding and MMD follow a similar high-level idea; here we focus on the former. First we show that the analysis can be reduced to the unit ball in $\mathcal{H}_K$ by proving that $\|\hat{\mu}_{\mathbb{P},Q} - \mu_{\mathbb{P}}\|_K \leq (1+\sqrt{2})r_{Q,N}$, where $r_{Q,N} = \sup_{f\in B_K} \mathrm{MON}_Q\big[x \mapsto \langle f, K(\cdot,x) - \mu_{\mathbb{P}}\rangle_K\big] = \sup_{f\in B_K} \underset{q\in[Q]}{\mathrm{med}}\{r(f,q)\}$ with $r(f,q) = \big\langle f, \mu_{S_q} - \mu_{\mathbb{P}}\big\rangle_K$. The Chebyshev inequality with a Lipschitz argument allows us to control the probability of the event $\{r_{Q,N} \leq \epsilon\}$ using the variable $Z = \sup_{f\in B_K} \sum_{q\in U}[\phi(2r(f,q)/\epsilon) - \mathbb{E}\phi(2r(f,q)/\epsilon)]$, where U stands for the indices of the uncorrupted blocks and $\phi(t) = (t-1)\mathbb{I}_{1\leq t\leq 2} + \mathbb{I}_{t\geq 2}$. The bounded difference property of the Z supremum of empirical processes guarantees its concentration around the expectation by using the McDiarmid inequality. The symmetrization technique combined with the Talagrand's contraction principle of Rademacher processes (thanks to the Lipschitz property of ϕ), followed by an other symmetrization leads to the deviation bound. Details are provided in Section 7.6.1-7.6.2 (for Theorem 37-38) in the supplementary material. ■

Remarks:

- Dependence on N: These finite-sample guarantees show that the MONK estimators
  - have optimal $N^{-1/2}$-rate—by recalling [TKS16, TKM17]'s discussed results—, and
  - they are robust to outliers, providing consistent estimates with high probability even under arbitrary adversarial contamination (affecting less than half of the samples).
- Dependence on δ: Recall that larger δ corresponds to less outliers, i.e., cleaner data in which case the bounds above become tighter. In other words, making use of medians the MONK estimators show robustness to outliers; this property is a nice byproduct of our optimal sub-Gaussian deviation bound. Whether this robustness to outliers is optimal in the studied setting is an open question.
- Dependence on Σ: It is worth contrasting the rates obtained in Theorem 37 and that of the tournament procedures [LM$^+$19d] derived for the finite-dimensional case. The latter paper elegantly resolved a long-lasting open question concerning the optimal dependency in terms of Σ. Theorem 37 proves the same dependency in the infinite-dimensional case, while giving rise to computionally tractable algorithms (Section 7.4).
- Separation rate: Theorem 38 also shows that fixing the trace of the covariance operators of $\mathbb{P}$ and $\mathbb{Q}$, the MON-based MMD estimator can separate $\mathbb{P}$ and $\mathbb{Q}$ at the rate of $N^{-1/2}$.
- Breakdown point: Our finite-sample bounds imply that the proposed MONK estimators using Q blocks is resistant to $Q/2$ outliers. Since Q is allowed to grow with N (it can be can be chosen to be almost $N/2$), this specifically means that the breakdown point of our estimators can be 25%.

7.4 Computing the MONK Estimator

This section is dedicated to the computation[4] of the analyzed MONK estimators; particularly we will focus on the MMD estimator given in Eq. (7.2.4). Numerical illustrations are provided in Section 7.5. Recall that the MONK estimator for MMD [Eq. (7.2.4)] is given by

$$\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) = \sup_{f\in B_K} \operatorname*{med}_{q\in[Q]} \left\{ \frac{1}{|S_q|}\sum_{j\in S_q} f(x_j) - \frac{1}{|S_q|}\sum_{j\in S_q} f(y_j) \right\}. \tag{7.4.1}$$

By the representer theorem [SHS01], the optimal f can be expressed as

$$f(\mathbf{a},\mathbf{b}) = \sum_{n\in[N]} a_n K(\cdot,x_n) + \sum_{n\in[N]} b_n K(\cdot,y_n), \tag{7.4.2}$$

where $\mathbf{a}=(a_n)_{n\in[N]}\in\mathbb{R}^N$ and $\mathbf{b}=(b_n)_{n\in[N]}\in\mathbb{R}^N$.

Denote $\mathbf{c}=[\mathbf{a};\mathbf{b}]\in\mathbb{R}^{2N}$, $\mathbf{K}=[\mathbf{K}_{xx},\mathbf{K}_{xy};\mathbf{K}_{yx},\mathbf{K}_{yy}]\in\mathbb{R}^{2N\times 2N}$, $\mathbf{K}_{xx}=[K(x_i,x_j)]_{i,j\in[N]}\in\mathbb{R}^{N\times N}$, $\mathbf{K}_{xy}=[K(x_i,y_j)]_{i,j\in[N]}=\mathbf{K}_{yx}^*\in\mathbb{R}^{N\times N}$, $\mathbf{K}_{yy}=[K(y_i,y_j)]_{i,j\in[N]}\in\mathbb{R}^{N\times N}$. With these notations, the optimisation problem (7.4.1) can be rewritten as

$$\max_{\mathbf{c}\in\mathbb{R}^{2N}:\mathbf{c}^*\mathbf{K}\mathbf{c}\le 1} \operatorname*{med}_{q\in[Q]} \left\{ |S_q|^{-1}[\mathbf{1}_q;-\mathbf{1}_q]^*\mathbf{K}\mathbf{c} \right\}, \tag{7.4.3}$$

where $\mathbf{1}_q\in\mathbb{R}^N$ is indicator vector of the block S_q. To enable efficient optimization we follow a block-coordinate descent (BCD)-type scheme: choose the $q_\mathrm{m}\in[N]$ index for which the median is attained in (7.4.3), and solve

$$\max_{\mathbf{c}\in\mathbb{R}^{2N}:\mathbf{c}^*\mathbf{K}\mathbf{c}\le 1} |S_{q_\mathrm{m}}|^{-1}[\mathbf{1}_{q_\mathrm{m}};-\mathbf{1}_{q_\mathrm{m}}]^*\mathbf{K}\mathbf{c}.$$

This optimization problem can be solved analytically: $\mathbf{c}=\frac{[\mathbf{1}_{q_\mathrm{m}};-\mathbf{1}_{q_\mathrm{m}}]}{\|\mathbf{L}^*[\mathbf{1}_{q_\mathrm{m}};-\mathbf{1}_{q_\mathrm{m}}]\|_2}$, where $\mathbf{L}$ is the Cholesky factor of $\mathbf{K}$ ($\mathbf{K}=\mathbf{L}\mathbf{L}^*$). The observations are shuffled after each iteration. The pseudo-code of the final MONK BCD estimator is summarized in Algorithm 3.

Notice that computing $\mathbf{L}$ in MONK BCD costs $O(N^3)$, which can be prohibitive for large sample size. In order to alleviate this bottleneck we also consider an approximate version of MONK BCD (referred to as MONK BCD-Fast), where the $\sum_{n\in[N]}$ summation after plugging (7.4.2) into (7.4.1) is replaced with $\sum_{n\in S_q}$:

$$\max_{\substack{\mathbf{c}=[\mathbf{a},\mathbf{b}]\in\mathbb{R}^{2N}\\ \mathbf{c}^*\mathbf{K}\mathbf{c}\le 1}} \operatorname*{med}_{q\in[Q]} \left\{ \frac{\sum_{j,n\in S_q}[a_nK(x_j,x_n)+b_nK(x_j,y_n)]}{|S_q|} - \frac{\sum_{j,n\in S_q}[a_nK(y_j,x_n)+b_nK(y_j,y_n)]}{|S_q|} \right\}.$$

This modification allows local computations restricted to blocks and improved running time. The samples are shuffled periodically (e.g., at every 10th iterations) to renew the blocks. The resulting method is presented in Algorithm 4. The computational complexity of the different MMD estimators are summarized in Table 7.1.

[4]The Python code reproducing our numerical experiments is available at https://bitbucket.org/TimotheeMathieu/monk-mmd; it relies on the ITE toolbox [Sza14].

Algorithm 3: MONK BCD estimator for MMD

Input: Aggregated Gram matrix: $\mathbf{K}$ with Cholesky factor $\mathbf{L}$ ($\mathbf{K} = \mathbf{L}\mathbf{L}^*$).
for all $t = 1, \ldots, T$ **do**
Generate a random permutation of $[N]$: σ.
Shuffle the samples according to σ: for $\forall q \in [Q]$

$$S_q = \left\{ \sigma\left((q-1)\frac{N}{Q} + 1 \right), \ldots, \sigma\left(q\frac{N}{Q} \right) \right\}.$$

Find the block attaining the median (q_{m}):

$$\frac{[\mathbf{1}_{q_{\mathrm{m}}}; -\mathbf{1}_{q_{\mathrm{m}}}]^* \mathbf{K}\mathbf{c}}{|S_{q_{\mathrm{m}}}|} = \operatorname*{med}_{q \in [Q]} \frac{[\mathbf{1}_q; -\mathbf{1}_q]^* \mathbf{K}\mathbf{c}}{|S_q|}.$$

Compute the coefficient vector: $\mathbf{c} = \frac{[\mathbf{1}_{q_{\mathrm{m}}}; -\mathbf{1}_{q_{\mathrm{m}}}]}{\|\mathbf{L}^*[\mathbf{1}_{q_{\mathrm{m}}}; -\mathbf{1}_{q_{\mathrm{m}}}]\|_2}$.
end for
Output: $\operatorname*{med}_{q \in [Q]} \left(\frac{1}{|S_q|} [\mathbf{1}_q; -\mathbf{1}_q]^* \mathbf{K}\mathbf{c} \right)$

7.5 Numerical Illustrations

In this section, we demonstrate the performance of the proposed MONK estimators. We exemplify the idea on the MMD estimator [Eq. (7.2.4)] with the BCD optimization schemes (MONK BCD and MONK BCD-Fast) discussed in Section 7.4. Our baseline is the classical U-statistic based MMD estimator [Eq. (7.2.6); referred to as U-Stat in the sequel].

The primary goal in the first set of experiments is to understand and demonstrate various aspects of the estimators for $(K, \mathbb{P}, \mathbb{Q})$ triplets [MFBS17, Table 3.3] when analytical expression is available for MMD. This is the case for polynomial and RBF kernels (K), with Gaussian distributions ($\mathbb{P}$, $\mathbb{Q}$). Notice that in the first (second) case the features are unbounded (bounded). Our second numerical example illustrates the applicability of the studied MONK estimators in biological context, in discriminating DNA subsequences with string kernel.

Experiment-1: We used the quadratic and the RBF kernel with bandwith $\sigma = 1$ for demonstration purposes and investigated the estimation error compared to the true MMD value: $|\widehat{\mathrm{MMD}}_Q(\mathbb{P}, \mathbb{Q}) - \mathrm{MMD}(\mathbb{P}, \mathbb{Q})|$. The errors are aggregates over 100 Monte-Carlo simulations, summarized in the median and quartile values. The number of samples (N) was chosen from $\{200, 400, \ldots, 2000\}$.

We considered three different experimental settings for $(\mathbb{P}, \mathbb{Q})$ and the absence/presence of outliers:

1. Gaussian distributions with no outliers: In this case $\mathbb{P} = \mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathbb{Q} = \mathcal{N}(\mu_2, \sigma_2^2)$ were normal where $(\mu_1, \sigma_1) \neq (\mu_2, \sigma_2)$, μ_1, σ_1, μ_2, σ_2 were randomly chosen from the $[0, 1]$ interval, and then their values were fixed. The estimators had access to $(x_n)_{n=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ and $(y_n)_{n=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$.
2. Gaussian distributions with outliers: This setting is a corrupted version of the first one. Particularly, the dataset consisted of $(x_n)_{n=1}^{N-5} \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, $(y_n)_{n=1}^{N-5} \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$, while the remaining 5-5 samples were set to $x_{N-4} = \ldots = x_N = 2000$, $y_{N-4} = \cdots = y_N = 4000$.

Algorithm 4: MONK BCD-Fast estimator for MMD

Input: Aggregated Gram matrix: $\mathbf{K}$ with Cholesky factor $\mathbf{L}$ ($\mathbf{K} = \mathbf{L}\mathbf{L}^*$).
Indices at which we shuffle: J.
for all $t = 1, \ldots, T$ **do**
  if $t \in J$ **then**
    Generate a random permutation of $[N]$: σ.
    Shuffle the samples according to σ: for $\forall q \in [Q]$

$$S_q = \left\{ \sigma\left((q-1)\frac{N}{Q} + 1 \right), \ldots, \sigma\left(q\frac{N}{Q} \right) \right\}.$$

    Compute the Gram matrices and the Cholesky factors on each block $\mathbf{K}_q$ and $\mathbf{L}_q$ for $q \in [Q]$.
  end if
  Find the block[5] attaining the median (q_m):

$$\frac{[\mathbb{1}_{q_\mathrm{m}}; -\mathbb{1}_{q_\mathrm{m}}]^* \mathbf{K}_{q_\mathrm{m}} \mathbf{c}_{q_\mathrm{m}}}{|S_{q_\mathrm{m}}|} = \operatorname*{med}_{q \in [Q]} \frac{[\mathbb{1}_q; -\mathbb{1}_q]^* \mathbf{K}_q \mathbf{c}_q}{|S_q|}.$$

  Update the coefficient vector: $\mathbf{c}_{q_\mathrm{m}} = \frac{[\mathbb{1}_{q_\mathrm{m}}; -\mathbb{1}_{q_\mathrm{m}}]}{\left\| \mathbf{L}^*_{q_\mathrm{m}} [\mathbb{1}_{q_\mathrm{m}}; -\mathbb{1}_{q_\mathrm{m}}] \right\|_2}$.
end for
Output: $\operatorname*{med}_{q \in [Q]} \left(\frac{1}{|S_q|} [\mathbb{1}_q; -\mathbb{1}_q]^* \mathbf{K}_q \mathbf{c}_q \right)$

Table 7.1: Computational complexity of MMD estimators. N: sample number, Q: number of blocks, T: number of iterations.

Method	Complexity
U-Stat	$\mathcal{O}\left(N^2\right)$
MONK BCD	$\mathcal{O}\left(N^3 + T\left[N^2 + Q\log(Q)\right]\right)$
MONK BCD-Fast	$\mathcal{O}\left(\frac{N^3}{Q^2} + T\left[\frac{N^2}{Q} + Q\log(Q)\right]\right)$

3. Pareto distribution without outliers: In this case $\mathbb{P} = \mathbb{Q} = Pareto(3)$ hence $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = 0$ and the estimators used $(x_n)_{n=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ and $(y_n)_{n=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$.

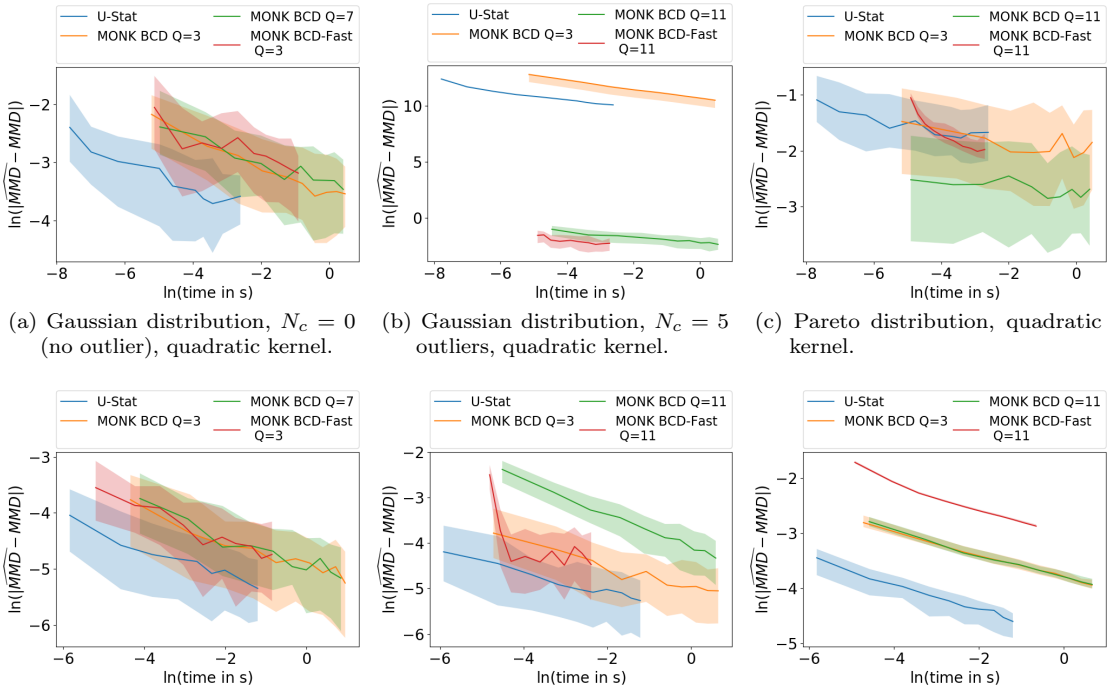
The 3 experiments were constructed to understand different aspects of the estimators: how a few outliers can ruin classical estimators (as we move from Experiment-1 to Experiment-2); in Experiment-3 the heavyness of the tail of a Pareto distribution makes the task non-trivial.

Our results on the three datasets with various Q choices are summarized in Fig. 7.1. As we can see from Fig. 7.1a and Fig. 7.1d in the outlier-free case, the MONK estimators are slower than the U-statistic based one; the accuracy is of the same order for both kernels. As demonstrated by Fig. 7.1b in the corrupted setup even a small number of outliers can completely ruin traditional MMD estimators for unbounded features while the MONK estimators are naturally robust to outliers with suitable choice of Q;[6] this is precisely the setting the MONK estimators were designed for. In case of bounded kernels (Fig. 7.1e), by construction, traditional MMD estimators are resistant to outliers; the MONK BCD-Fast method achieves comparable performance. In the final Pareto experiment (Fig. 7.1c and Fig. 7.1f) where the distribution produces "natural outliers", again MONK estimators are more robust with respect to corruption than the one relying on U-statistics in the case of polynomial kernel. These experiments illustrate the power of the studied MONK schemes: these estimators achieve comparable performance in case of bounded features, while for unbounded features they can efficiently cope with the presence of outliers.

Experiment-2 (discrimination of DNA subsequences): In order to demonstrate the applicability of our estimators in biological context, we chose a DNA benchmark from the UCI repository [DK17], the Molecular Biology (Splice-junction Gene Sequences) Data Set. The dataset consists of 3190 instances of 60-character-long DNA subsequences. The problem is to recognize, given a sequence of DNA, the boundaries between exons (the parts of the DNA sequence retained after splicing) and introns (the parts of the DNA sequence that are spliced out). This task consists of two subproblems, identifying the exon/intron boundaries (referred to as EI sites) and the intron/exon boundaries (IE sites).[7] We took 1532 of these samples by selecting 766 instances from both the EI and the IE classes (the class of those being neither EI nor IE is more heterogeneous and thus we dumped it from the study), and investigated the discriminability of the EI and IE categories. We represented the DNA sequences as strings ($\mathcal{X}$), chose K as the String Subsequence Kernel [LSST+02] to compute MMD, and performed two-sample testing based on MMD using the MONK BCD, MONK BCD-Fast and U-Stat estimators. For completeness the pseudocode of the hypothesis test is detailed in Algorithm 5 (Section 7.9). Q, the number of blocks in the MONK techniques, was equal to 5. The significance level was $\alpha = 0.05$. To assess the variability of the results 400 Monte Carlo simulations were performed, each time uniformly sampling N points without replacement resulting in $(X_n)_{n\in[N]}$ and $(Y_n)_{n\in[N]}$. To provide more detailed insights the aggregated values of $\widehat{\mathrm{MMD}}(\mathrm{EI}, \mathrm{IE}) - \hat{q}_{1-\alpha}$, $\widehat{\mathrm{MMD}}(\mathrm{EI}, \mathrm{EI}) - \hat{q}_{1-\alpha}$ and $\widehat{\mathrm{MMD}}(\mathrm{IE}, \mathrm{IE}) - \hat{q}_{1-\alpha}$ are summarized in Fig. 7.2, where $\hat{q}_{1-\alpha}$ is the estimated $(1-\alpha)$-quantile via $B = 150$ bootstrap permutations. In the ideal case, $\widehat{\mathrm{MMD}} - \hat{q}_{1-\alpha}$ is positive (negative) in the inter-class (intra-class) experiments. As Fig. 7.2 shows all 3 techniques are able to solve the task, both in the inter-class (when the null hypothesis does not hold; Fig. 7.2a) and the intra-class experiment (null holds;

[6] In case of unknown N_c, one could choose Q adaptively by the Lepski method (see for example [DLLO16]) at the price of increasing the computational effort. Though the resulting Q would increase the computational time, it would be adaptive thanks to its data-driven nature, and would benefit from the same guarantee as the fixed Q appearing in Theorem 37-38.

[7] In the biological community, IE borders are referred to as "acceptors" while EI borders are referred to as "donors".



(a) Gaussian distribution, $N_c = 0$ (no outlier), quadratic kernel.

(b) Gaussian distribution, $N_c = 5$ outliers, quadratic kernel.

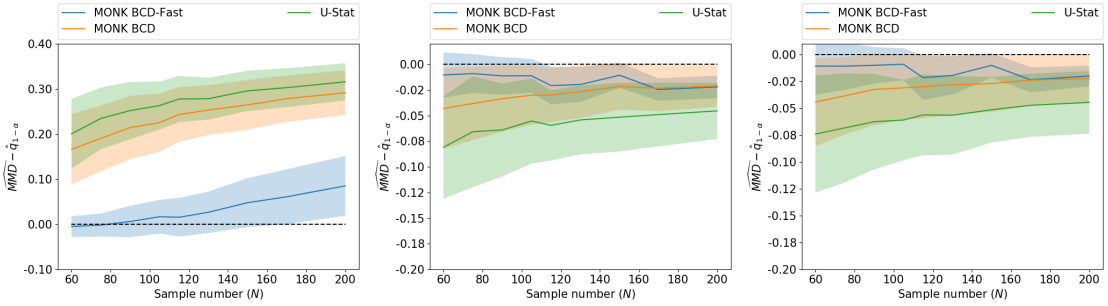
(c) Pareto distribution, quadratic kernel.

(d) Gaussian distribution, $N_c = 0$ (no outlier), RBF kernel.

(e) Gaussian distribution, $N_c = 5$ outliers, RBF kernel.

(f) Pareto distribution, RBF kernel.

Figure 7.1: Performance of the MMD estimators: median and quartiles of $\ln(|\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}(\mathbb{P},\mathbb{Q})|)$. Columns from left to right: Experiment-1 – Experiment-3. Top: quadratic kernel, bottom: RBF kernel.



(a) Inter-class: EI-IE (b) Intra-class: EI-EI (c) Intra-class: IE-IE

Figure 7.2: Inter-class and intra-class MMD estimates as a function of the sample number compared to the bootstrap-estimated $(1-\alpha)$-quantile: $\widehat{\mathrm{MMD}} - \hat{q}_{1-\alpha}$; mean $\pm$ std. The null hypothesis is rejected iff $\widehat{\mathrm{MMD}} - \hat{q}_{1-\alpha} > 0$. Notice the different scale of $\widehat{\mathrm{MMD}} - \hat{q}_{1-\alpha}$ in the inter-class and the intra-class experiments.

Fig. 7.2b and Fig. 7.2c), and they converge to a good and stable performance as a function of the sample number. It is important to note that the MONK BCD-Fast method is especially well-adapted to problems where the kernel computation (such as the String Subsequence Kernel) or the sample size is a bottleneck, as its computation is often significantly faster compared to the U-Stat technique. For example, taking all the samples ($N = 766$) in the DNA benchmark with $Q = 15$, computing MONK BCD-Fast (U-Stat) takes $32s$ ($1m28s$). These results illustrate the applicability of our estimators in gene analysis.

Acknowledgements

Guillaume Lecué is supported by a grant of the French National Research Agency (ANR), "Investissements d'Avenir" (LabEx Ecodec/ANR-11-LABX-0047).

Supplement

The supplement contains the detailed proofs of our results (Section 7.6), a few technical lemmas used during these arguments (Section 7.7), the McDiarmid inequality for self-containedness (Section 7.8), and the pseudocode of the two-sample test performed in Experiment-2 (Section 7.9).

7.6 Proofs of Theorem 37 and Theorem 38

This section contains the detailed proofs of Theorem 37 (Section 7.6.1) and Theorem 38 (Section 7.6.2).

7.6.1 Proof of Theorem 37

The structure of the proof is as follows:

1. We show that $\|\hat{\mu}_{\mathbb{P},Q} - \mu_{\mathbb{P}}\|_K \leq (1+\sqrt{2}) r_{Q,N}$, where $r_{Q,N} = \sup_{f \in B_K} \mathrm{MON}_Q\Big[\underbrace{\langle f, K(\cdot,x) - \mu_{\mathbb{P}}\rangle_K}_{f(x) - \mathbb{P}f}\Big]$, i.e. the analysis can be reduced to B_K.
2. Then $r_{Q,N}$ is bounded using empirical processes.

Step-1: Since $\mathcal{H}_K$ is an inner product space, for any $f \in \mathcal{H}_K$

$$\|f - K(\cdot,x)\|_K^2 - \|\mu_{\mathbb{P}} - K(\cdot,x)\|_K^2 = \|f - \mu_{\mathbb{P}}\|_K^2 - 2\langle f - \mu_{\mathbb{P}}, K(\cdot,x) - \mu_{\mathbb{P}}\rangle_K. \tag{7.6.1}$$

Hence, by denoting $e = \hat{\mu}_{\mathbb{P},Q} - \mu_{\mathbb{P}}$, $\tilde{g} = g - \mu_{\mathbb{P}}$ we get

$$\begin{aligned}
\|e\|_K^2 - 2r_{Q,N}\|e\|_K &\overset{(a)}{\leq} \|e\|_K^2 - 2\mathrm{MON}_Q\left[\left\langle \frac{e}{\|e\|_K}, K(\cdot,x) - \mu_{\mathbb{P}}\right\rangle_K \|e\|_K\right] \\
&\overset{(b)}{\leq} \mathrm{MON}_Q\left[\|e\|_K^2 - 2\left\langle \frac{e}{\|e\|_K}, K(\cdot,x) - \mu_{\mathbb{P}}\right\rangle_K \|e\|_K\right] \\
&\overset{(c)}{\leq} \mathrm{MON}_Q\left[\|\hat{\mu}_{\mathbb{P},Q} - K(\cdot,x)\|_K^2 - \|\mu_{\mathbb{P}} - K(\cdot,x)\|_K^2\right] \\
&\overset{(d)}{\leq} \sup_{g \in \mathcal{H}_K} \mathrm{MON}_Q\left[\|\hat{\mu}_{\mathbb{P},Q} - K(\cdot,x)\|_K^2 - \|g - K(\cdot,x)\|_K^2\right] \\
&\overset{(e)}{\leq} \sup_{g \in \mathcal{H}_K} \mathrm{MON}_Q\left[\|\mu_{\mathbb{P}} - K(\cdot,x)\|_K^2 - \|g - K(\cdot,x)\|_K^2\right] \\
&\overset{(f)}{=} \sup_{g \in \mathcal{H}_K} \left\{2\mathrm{MON}_Q\Big[\underbrace{\langle \tilde{g}, K(\cdot,x) - \mu_{\mathbb{P}}\rangle_K}_{\|\tilde{g}\|_K \left\langle \frac{\tilde{g}}{\|\tilde{g}\|_K}, K(\cdot,x) - \mu_{\mathbb{P}}\right\rangle_K}\Big] - \|\tilde{g}\|_K^2\right\} \\
&\overset{(g)}{=} \sup_{g \in \mathcal{H}_K} \left\{2\|\tilde{g}\|_K r_{Q,N} - \|\tilde{g}\|_K^2\right\} \overset{(h)}{\leq} r_{Q,N}^2,
\end{aligned} \tag{7.6.2}$$

where we used in (a) the definition of $r_{Q,N}$, (b) the linearity[8] of $\mathrm{MON}_Q[\cdot]$, (c) Eq. (7.6.1), (d) $\sup_g$, (e) the definition of $\hat{\mu}_{\mathbb{P},Q}$, (f) Eq. (7.6.1) and the linearity of $\mathrm{MON}_Q[\cdot]$, (g) the definition of $r_{Q,N}$. In step (h), by denoting $a = \|\tilde{g}\|_K$, $r = r_{Q,N}$, the argument of the sup takes the form $2ar - a^2$; $2ar - a^2 \leq r^2 \Leftrightarrow 0 \leq r^2 - 2ar + a^2 = (r-a)^2$.

In Eq. (7.6.2), we obtained an equation $a^2 - 2ra \leq r^2$ where $a := \|e\|_K \geq 0$. Hence $r^2 + 2ra - a^2 \geq 0$, $r_{1,2} = \left[-2a \pm \sqrt{4a^2 + 4a^2}\right]/2 = \left(-1 \pm \sqrt{2}\right)a$, thus by the non-negativity of a, $r \geq (-1+\sqrt{2})a$, i.e., $a \leq \frac{r}{\sqrt{2}-1} = (\sqrt{2}+1)r$. In other words, we arrived at

$$\|\hat{\mu}_{\mathbb{P},Q} - \mu_{\mathbb{P}}\|_K \leq \left(1 + \sqrt{2}\right) r_{Q,N}.$$

It remains to upper bound $r_{Q,N}$.

[8] $\mathrm{MON}_Q[c_1 + c_2 f] = c_1 + c_2 \mathrm{MON}_Q[f]$ for any $c_1, c_2 \in \mathbb{R}$.

Step-2: Our goal is to provide a probabilistic bound on

$$r_{Q,N} = \sup_{f\in B_K} \mathrm{MON}_Q[x \mapsto \langle f, K(\cdot,x) - \mu_{\mathbb{P}}\rangle_K] = \sup_{f\in B_K} \operatorname*{med}_{q\in[Q]}\{\underbrace{\langle f, \mu_{S_q} - \mu_{\mathbb{P}}\rangle_K}_{=:r(f,q)}\}.$$

The N_c corrupted samples can affect (at most) N_c of the $(S_q)_{q\in[Q]}$ blocks. Let $U := [Q]\backslash C$ stand for the indices of the uncorrupted sets, where $C := \{q \in [Q] : \exists n_j \text{ s.t. } n_j \in S_q,\, j \in [N_c]\}$ contains the indices of the corrupted sets. If

$$\forall f \in B_K : \underbrace{|\{q \in U : r(f,q) \geq \epsilon\}|}_{\sum_{q\in U}\mathbb{I}_{r(f,q)\geq\epsilon}} + N_c \leq \frac{Q}{2}, \tag{7.6.3}$$

then for $\forall f \in B_K$, $\operatorname{med}_{q\in[Q]}\{r(f,q)\} \leq \epsilon$, i.e. $\sup_{f\in B_K}\operatorname{med}_{q\in[Q]}\{r(f,q)\} \leq \epsilon$. Thus, our task boils down to controlling the event in (7.6.3) by appropriately choosing ϵ.

- **Controlling** $r(f,q)$: For any $f \in B_K$ the random variables $\langle f, k(\cdot,x_i) - \mu_{\mathbb{P}}\rangle_{\mathcal{H}_K} = f(x_i) - \mathbb{P}f$ are independent, have zero mean, and

$$\begin{aligned}\mathbb{E}_{x_i\sim\mathbb{P}}\langle f, k(\cdot,x_i) - \mu_{\mathbb{P}}\rangle_K^2 &= \langle f, \Sigma_{\mathbb{P}} f\rangle_K \\ &\leq \|f\|_K \|\Sigma_{\mathbb{P}} f\|_K \leq \|f\|_K^2 \|\Sigma_{\mathbb{P}}\| = \|\Sigma_{\mathbb{P}}\|\end{aligned} \tag{7.6.4}$$

  using the reproducing property of the kernel and the covariance operator, the Cauchy-Schwarz (CBS) inequality and $\|f\|_{\mathcal{H}_K} = 1$.
  For a zero-mean random variable z by the Chebyshev's inequality $\mathbb{P}(z > a) \leq \mathbb{P}(|z| > a) \leq \mathbb{E}(z^2)/a^2$, which implies $\mathbb{P}\Big(z > \sqrt{\mathbb{E}(z^2)/\alpha}\Big) \leq \alpha$ by a $\alpha = \mathbb{E}(z^2)/a^2$ substitution. With $z := r(f,q)$ $(q \in U)$, using $\mathbb{E}[z^2] = \mathbb{E}\langle f, \mu_{S_q} - \mu_{\mathbb{P}}\rangle_K^2 = \frac{Q}{N}\mathbb{E}_{x_i\sim\mathbb{P}}\langle f, k(\cdot,x_i) - \mu_{\mathbb{P}}\rangle_K^2$ and Eq. (7.6.4) one gets that for all $f \in B_K$, $\alpha \in (0,1)$ and $q \in U$: $\mathbb{P}\Big(r(f,q) > \sqrt{\frac{\|\Sigma_{\mathbb{P}}\|Q}{\alpha N}}\Big) \leq \alpha$. This means $\mathbb{P}\big(r(f,q) > \frac{\epsilon}{2}\big) \leq \alpha$ with $\epsilon \geq 2\sqrt{\frac{\|\Sigma_{\mathbb{P}}\|Q}{\alpha N}}$.
- **Reduction to** ϕ: As a result

$$\sum_{q\in U}\mathbb{P}\Big(r(f,q) \geq \frac{\epsilon}{2}\Big) \leq |U|\alpha$$

  happens if and only if

$$\sum_{q\in U}\mathbb{I}_{r(f,q)\geq\epsilon} \leq |U|\alpha + \sum_{q\in U}\Big[\mathbb{I}_{r(f,q)\geq\epsilon} - \underbrace{\mathbb{P}\Big(r(f,q) \geq \frac{\epsilon}{2}\Big)}_{\mathbb{E}\left[\mathbb{I}_{r(f,q)\geq\frac{\epsilon}{2}}\right]}\Big] =: A.$$

  Let us introduce $\phi : t \in \mathbb{R} \to (t-1)\mathbb{I}_{1\leq t\leq 2} + \mathbb{I}_{t\geq 2}$. ϕ is 1-Lipschitz and satisfies $\mathbb{I}_{2\leq t} \leq \phi(t) \leq \mathbb{I}_{1\leq t}$ for any $t \in \mathbb{R}$. Hence, we can upper bound A as

$$A \leq |U|\alpha + \sum_{q\in U}\Big[\phi\Big(\frac{2r(f,q)}{\epsilon}\Big) - \mathbb{E}\phi\Big(\frac{2r(f,q)}{\epsilon}\Big)\Big]$$

by noticing that $\epsilon \le r(f,q) \Leftrightarrow 2 \le 2r(f,q)/\epsilon$ and $\epsilon/2 \le r(f,q) \Leftrightarrow 1 \le 2r(f,q)/\epsilon$, and by using the $\mathbb{I}_{2\le t} \le \phi(t)$ and the $\phi(t) \le \mathbb{I}_{1\le t}$ bound, respectively. Taking supremum over B_K we arrive at

$$\sup_{f\in B_K} \sum_{q\in U} \mathbb{I}_{r(f,q)\ge\epsilon} \le |U|\alpha + \underbrace{\sup_{f\in B_K} \sum_{q\in U} \left[\phi\left(\frac{2r(f,q)}{\epsilon}\right) - \mathbb{E}\phi\left(\frac{2r(f,q)}{\epsilon}\right)\right]}_{=:Z}.$$

- **Concentration of Z around its mean:** Notice that Z is a function of x_V, the samples in the uncorrupted blocks; $V = \cup_{q\in U} S_q$. By the bounded difference property of Z (Lemma 47) for any $\beta > 0$, the McDiarmid inequality (Lemma 49; we choose $\tau := Q\beta^2/8$ to get linear scaling in Q on the r.h.s.) implies that

$$\mathbb{P}(Z < \mathbb{E}_{x_V}[Z] + Q\beta) \ge 1 - e^{-\frac{Q\beta^2}{8}}.$$

- **Bounding $\mathbb{E}_{x_V}[Z]$:** Let $M = N/Q$ denote the number of elements in S_q-s. The $\mathcal{G} = \{g_f : f \in B_K\}$ class with $g_f : \mathcal{X}^M \to \mathbb{R}$ and $\mathbb{P}_M := \frac{1}{M}\sum_{m=1}^M \delta_{u_m}$ defined as

$$g_f(u_{1:M}) = \phi\left(\frac{\langle f, \mu_{\mathbb{P}_M} - \mu_{\mathbb{P}}\rangle_K}{\epsilon}\right)$$

  is uniformly bounded separable Carathéodory (Lemma 48), hence the symmetrization technique [SC08, Prop. 7.10], [LT91] gives

$$\mathbb{E}_{x_V}[Z] \le 2\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}} \sup_{f\in B_K} \left|\sum_{q\in U} e_q \phi\left(\frac{2r(f,q)}{\epsilon}\right)\right|,$$

  where $\mathbf{e} = (e_q)_{q\in U} \in \mathbb{R}^{|U|}$ with i.i.d. Rademacher entries [$\mathbb{P}(e_q = \pm 1) = \frac{1}{2}$ $(\forall q)$].
- **Discarding ϕ:** Since $\phi(0) = 0$ and ϕ is 1-Lipschitz, by Talagrand's contraction principle of Rademacher processes [LT91], [Kol11, Theorem 2.3] one gets

$$\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}} \sup_{f\in B_K} \left|\sum_{q\in U} e_q \phi\left(\frac{2r(f,q)}{\epsilon}\right)\right| \le 2\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}} \sup_{f\in B_K} \left|\sum_{q\in U} e_q \frac{2r(f,q)}{\epsilon}\right|.$$

- **Switching from $|U|$ to N terms**: Applying an other symmetrization [(a)], the CBS inequality,

$f \in B_K$, and the Jensen inequality

$$\begin{aligned}
\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}} \sup_{f\in B_K}\left|\sum_{q=1}^{Q} e_q \frac{r(f,q)}{\epsilon}\right| &\overset{(a)}{\leq} \frac{2Q}{\epsilon N}\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}'}\left[\sup_{f\in B_K}\Big|\underbrace{\sum_{n\in V} e'_n\langle f, K(\cdot,x_n)-\mu_{\mathbb{P}}\rangle_K}_{=\left\langle f,\sum_{n\in V} e'_n[K(\cdot,x_n)-\mu_{\mathbb{P}}]\right\rangle_K}\Big|\right] \\
&\leq \frac{2Q}{\epsilon N}\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}'}\left[\sup_{f\in B_K}\underbrace{\|f\|_K}_{=1}\left\|\sum_{n\in V} e'_n[K(\cdot,x_n)-\mu_{\mathbb{P}}]\right\|_K\right] \\
&= \frac{2Q}{\epsilon N}\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}'}\left\|\sum_{n\in V} e'_n[K(\cdot,x_n)-\mu_{\mathbb{P}}]\right\|_K \\
&\leq \frac{2Q}{\epsilon N}\sqrt{\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}'}\left\|\sum_{n\in V} e'_n[K(\cdot,x_n)-\mu_{\mathbb{P}}]\right\|_K^2} \\
&\overset{(b)}{=} \frac{2Q\sqrt{|V|\operatorname{Tr}(\Sigma_{\mathbb{P}})}}{\epsilon N}.
\end{aligned}$$

In (a), we proceed as follows:

$$\begin{aligned}
\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}} \sup_{f\in B_K}\left|\sum_{q\in U} e_q \frac{r(f,q)}{\epsilon}\right| &= \mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}} \sup_{f\in B_K}\left|\sum_{q\in U} e_q \frac{\langle f, \mu_{S_q}-\mu_{\mathbb{P}}\rangle_K}{\epsilon}\right| \\
&\overset{(c)}{\leq} \frac{2Q}{N\epsilon}\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}}\mathbb{E}_{\mathbf{e}'} \sup_{f\in B_K}\left|\sum_{n\in V} e'_n e''_n\langle f, K(\cdot,x_n)-\mu_{\mathbb{P}}\rangle_K\right| \\
&= \frac{2Q}{N\epsilon}\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}'} \sup_{f\in B_K}\left|\sum_{n\in V} e'_n\langle f, K(\cdot,x_n)-\mu_{\mathbb{P}}\rangle_K\right|,
\end{aligned}$$

where in (c) we applied symmetrization, $\mathbf{e}' = (e'_n)_{n\in V} \in \mathbb{R}^{|V|}$ with i.i.d. Rademacher entries, $e''_n = e_q$ if $n \in S_q$ $(q \in U)$, and we used that $(e'_n e''_n\langle f, K(\cdot,x_n)-\mu_{\mathbb{P}}\rangle_K)_{n\in V} \overset{\text{distr}}{=} (e'_n\langle f, K(\cdot,x_n)-\mu_{\mathbb{P}}\rangle_K)_{n\in V}$.
In step (b), we had

$$\begin{aligned}
\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}'}\left\|\textstyle\sum_{n\in V} e'_n[K(\cdot,x_n)-\mu_{\mathbb{P}}]\right\|_K^2 &= \mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}'}\sum_{n\in V}[e'_n]^2\langle K(\cdot,x_n)-\mu_{\mathbb{P}}, K(\cdot,x_n)-\mu_{\mathbb{P}}\rangle_K \\
&= |V|\mathbb{E}_{x\sim\mathbb{P}}\langle K(\cdot,x)-\mu_{\mathbb{P}}, K(\cdot,x)-\mu_{\mathbb{P}}\rangle_K \\
&= |V|\mathbb{E}_{x\sim\mathbb{P}}\operatorname{Tr}([K(\cdot,x)-\mu_{\mathbb{P}}]\otimes[K(\cdot,x)-\mu_{\mathbb{P}}]) \\
&= |V|\operatorname{Tr}(\Sigma_{\mathbb{P}})
\end{aligned}$$

exploiting the independence of e'_n-s and $[e'_n]^2 = 1$.

Until this point we showed that for all $\alpha \in (0,1)$, $\beta > 0$, if $\epsilon \geq 2\sqrt{\frac{\|\Sigma_{\mathbb{P}}\|Q}{\alpha N}}$ then

$$\sup_{f\in B_K}\sum_{q=1}^{Q}\mathbb{I}_{r(f,q)\geq\epsilon} \leq |U|\alpha + Q\beta + \frac{8Q\sqrt{|V|\operatorname{Tr}(\Sigma_{\mathbb{P}})}}{\epsilon N}$$

with probability at least $1 - e^{-\frac{Q\beta^2}{8}}$. Thus, to ensure that $\sup_{f\in B_K} \sum_{q=1}^{Q} \mathbb{I}_{r(f,q)\geq\epsilon} + N_c \leq Q/2$ it is sufficient to choose $(\alpha, \beta, \epsilon)$ such that $|U|\alpha + Q\beta + \frac{8Q\sqrt{|V|\operatorname{Tr}(\Sigma_{\mathbb{P}})}}{\epsilon N} + N_c \leq \frac{Q}{2}$, and in this case $\|\hat{\mu}_{\mathbb{P},Q} - \mu_{\mathbb{P}}\|_K \leq (1+\sqrt{2})\epsilon$. Applying the $|U| \leq Q$ and $|V| \leq N$ bounds, we want to have

$$Q\alpha + Q\beta + \frac{8Q\sqrt{\operatorname{Tr}(\Sigma_{\mathbb{P}})}}{\epsilon\sqrt{N}} + N_c \leq \frac{Q}{2}. \tag{7.6.5}$$

Choosing $\alpha = \beta = \frac{\delta}{3}$ in Eq. (7.6.5), the sum of the first two terms is $Q\frac{2\delta}{3}$; $\epsilon \geq \max\left(2\sqrt{\frac{3\|\Sigma_{\mathbb{P}}\|Q}{\delta N}}, \frac{24}{\delta}\sqrt{\frac{\operatorname{Tr}(\Sigma_{\mathbb{P}})}{N}}\right)$ gives $\leq Q\frac{\delta}{3}$ for the third term. Since $N_c \leq Q(\frac{1}{2} - \delta)$, we got

$$\|\hat{\mu}_{\mathbb{P},Q} - \mu_{\mathbb{P}}\|_K \leq c_1 \max\left(\sqrt{\frac{3\|\Sigma_{\mathbb{P}}\|Q}{\delta N}}, \frac{12}{\delta}\sqrt{\frac{\operatorname{Tr}(\Sigma_{\mathbb{P}})}{N}}\right)$$

with probability at least $1-\mathrm{e}^{-\frac{Q\delta^2}{72}}$. With an $\eta = \mathrm{e}^{-\frac{Q\delta^2}{72}}$, and hence $Q = \frac{72\ln\left(\frac{1}{\eta}\right)}{\delta^2}$ reparameterization Theorem 37 follows.

7.6.2 Proof of Theorem 38

The reasoning is similar to Theorem 37; we detail the differences below. The high-level structure of the proof is as follows:

- First we prove that $\left|\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}(\mathbb{P},\mathbb{Q})\right| \leq r_{Q,N}$,
  where $r_{Q,N} = \sup_{f\in B_K}\left|\operatorname*{med}_{q\in[Q]}\left\{\left\langle f, \left(\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}\right) - \left(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right)\right\rangle_K\right\}\right|$.
- Then $r_{Q,N}$ is bounded.

Step-1:

- $\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}(\mathbb{P},\mathbb{Q}) \leq r_{Q,N}$:
  By the subadditivity of supremum $[\sup_f(a_f + b_f) \leq \sup_f a_f + \sup_f b_f]$ one gets

$$\begin{aligned}
\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) &= \sup_{f\in B_K} \operatorname*{med}_{q\in[Q]}\left\{\left\langle f, \left(\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}\right) - \left(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right)\right.\right. \\
&\qquad \left.\left. + \left(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right)\right\rangle_K\right\} \\
&\leq \sup_{f\in B_K} \operatorname*{med}_{q\in[Q]}\left\{\left\langle f, \left(\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}\right) - \left(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right)\right\rangle_K\right\} \\
&\qquad + \sup_{f\in B_K} \left\langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right\rangle_K \\
&\leq \underbrace{\sup_{f\in B_K}\left|\operatorname*{med}_{q\in[Q]}\left\{\left\langle f, \left(\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}\right) - \left(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right)\right\rangle_K\right\}\right|}_{=r_{Q,N}} + \mathrm{MMD}(\mathbb{P},\mathbb{Q}).
\end{aligned}$$

- $\mathrm{MMD}_Q(\mathbb{P},\mathbb{Q}) - \widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) \leq r_{Q,N}$:

Let $a_f := \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_K$ and $b_f := \underset{q\in[Q]}{\text{med}}\{\langle f, (\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}) - (\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}})\rangle_K\}$. Then

$$\begin{aligned} a_f - b_f &= \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_K \\ &\quad + \text{med}_{q\in[Q]}\{\langle f, (\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}) - (\mu_{\mathbb{P}} - \mu_{\mathbb{Q}})\rangle_K\} \\ &= \text{med}_{q\in[Q]}\{\langle f, \mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}} \rangle_K\} \end{aligned}$$

by $\text{med}_{q\in[Q]}\{-z_q\} = -\text{med}_{q\in[Q]}\{z_q\}$. Applying the $\sup_f(a_f - b_f) \geq \sup_f a_f - \sup_f b_f$ inequality (it follows from the subadditivity of sup):

$$\begin{aligned} \widehat{\text{MMD}}_Q(\mathbb{P},\mathbb{Q}) &\geq \text{MMD}(\mathbb{P},\mathbb{Q}) - \sup_{f\in B_K} \underbrace{\underset{q\in[Q]}{\text{med}}\{\langle f, (\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}) - (\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}})\rangle_K\}}_{-\underset{q\in[Q]}{\text{med}}\{\langle f, (\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}) - (\mu_{\mathbb{P}} - \mu_{\mathbb{Q}})\rangle_K\}} \\ &\geq \text{MMD}(\mathbb{P},\mathbb{Q}) - \underbrace{\sup_{f\in B_K} \left|\underset{q\in[Q]}{\text{med}}\{\langle f, (\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}) - (\mu_{\mathbb{P}} - \mu_{\mathbb{Q}})\rangle_K\}\right|}_{r_{Q,N}}. \end{aligned}$$

Step-2: Our goal is to control

$$\begin{aligned} r_{Q,N} &= \sup_{f\in B_K} \left|\text{med}_{q\in[Q]}\Big\{r(f,q)\Big\}\right|, \text{ where} \\ r(f,q) &:= \langle f, (\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}) - (\mu_{\mathbb{P}} - \mu_{\mathbb{Q}})\rangle_K. \end{aligned}$$

The relevant quantities which change compared to the proof of Theorem 37 are as follows.

- **Median rephrasing**:

$$\begin{aligned} \sup_{f\in B_K} \left|\underset{q\in[Q]}{\text{med}}\{r(f,q)\}\right| \leq \epsilon &\Leftrightarrow \forall f \in B_K : -\epsilon \leq \text{med}_{q\in[Q]}\{r(f,q)\} \leq \epsilon \\ &\Leftarrow \forall f \in B_K : |\{q : r(f,q) \leq -\epsilon\}| \leq Q/2 \qquad \text{and } |\{q : r(f,q) \geq \epsilon\}| \leq Q/2 \\ &\Leftarrow \forall f \in B_K : |\{q : |r(f,q)| \geq \epsilon\}| \leq Q/2. \end{aligned}$$

  Thus, $\forall f \in B_K : |\{q \in U : |r(f,q)| \geq \epsilon\}| + N_c \leq \frac{Q}{2}$, implies $\sup_{f\in B_K} \left|\text{med}_{q\in[Q]}\{r(f,q)\}\right| \leq \epsilon$.
- **Controlling** $|r(f,q)|$: For any $f \in B_K$ the random variables $[f(x_i) - f(y_i)] - [\mathbb{P}f - \mathbb{Q}f]$ are independent, zero-mean and

$$\begin{aligned} \mathbb{E}_{(x,y)\sim\mathbb{P}\otimes\mathbb{Q}}([f(x) - \mathbb{P}f] - [f(y) - \mathbb{Q}f])^2 &= \mathbb{E}_{x\sim\mathbb{P}}[f(x) - \mathbb{P}f]^2 + \mathbb{E}_{y\sim\mathbb{Q}}[f(y) - \mathbb{Q}f]^2 \\ &\leq \|\Sigma_{\mathbb{P}}\| + \|\Sigma_{\mathbb{Q}}\|, \end{aligned}$$

  where $\mathbb{P}\otimes\mathbb{Q}$ is the product measure. The Chebyshev argument with $z = |r(f,q)|$ implies that $\forall \alpha \in (0,1)$

$$(\mathbb{P}\otimes\mathbb{Q})\left(|r(f,q)| > \sqrt{\frac{(\|\Sigma_{\mathbb{P}}\| + \|\Sigma_{\mathbb{Q}}\|)Q}{\alpha N}}\right) \leq \alpha.$$

  This means $(\mathbb{P}\otimes\mathbb{Q})(|r(f,q)| > \epsilon/2) \leq \alpha$ with $\epsilon \geq 2\sqrt{\frac{(\|\Sigma_{\mathbb{P}}\| + \|\Sigma_{\mathbb{Q}}\|)Q}{\alpha N}}$.

- **Switching from $|U|$ to N terms**: With $(xy)_V = \{(x_i, y_i) : i \in V\}$, in '(b)' with $\tilde{x}_n := K(\cdot, x_n) - \mu_{\mathbb{P}}$, $\tilde{y}_n := K(\cdot, y_n) - \mu_{\mathbb{Q}}$ we arrive at

$$\begin{aligned}\mathbb{E}_{(xy)_V}\mathbb{E}_{\mathbf{e}'}\left\|\sum_{n\in V} e'_n(\tilde{x}_n - \tilde{y}_n)\right\|_K^2 &= \mathbb{E}_{(xy)_V}\mathbb{E}_{\mathbf{e}'}\sum_{n\in V}[e'_n]^2\langle\tilde{x}_n - \tilde{y}_n, \tilde{x}_n - \tilde{y}_n\rangle_K \\ &= |V|\mathbb{E}_{(xy)\sim\mathbb{P}}\|[K(\cdot,x) - \mu_{\mathbb{P}}] - [K(\cdot,y) - \mu_{\mathbb{Q}}]\|_K \\ &= |V|[\mathrm{Tr}(\Sigma_{\mathbb{P}}) + \mathrm{Tr}(\Sigma_{\mathbb{Q}})].\end{aligned}$$

- These results imply

$$Q\alpha + Q\beta + \frac{8Q\sqrt{\mathrm{Tr}(\Sigma_{\mathbb{P}}) + \mathrm{Tr}(\Sigma_{\mathbb{Q}})}}{\epsilon\sqrt{N}} + N_c \leq Q/2.$$

$\epsilon \geq \max\left(2\sqrt{\frac{3(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)Q}{\delta N}}, \frac{24}{\delta}\sqrt{\frac{\mathrm{Tr}\,(\Sigma_{\mathbb{P}})+\mathrm{Tr}\,(\Sigma_{\mathbb{Q}})}{N}}\right)$, $\alpha = \beta = \frac{\delta}{3}$ choice gives that

$$\left|\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}(\mathbb{P},\mathbb{Q})\right| \leq 2\max\left(\sqrt{\frac{3(\|\Sigma_{\mathbb{P}}\| + \|\Sigma_{\mathbb{Q}}\|)Q}{\delta N}}, \frac{12}{\delta}\sqrt{\frac{\mathrm{Tr}\,(\Sigma_{\mathbb{P}}) + \mathrm{Tr}\,(\Sigma_{\mathbb{Q}})}{N}}\right)$$

with probability at least $1 - e^{-\frac{Q\delta^2}{72}}$. $\eta = e^{-\frac{Q\delta^2}{72}}$, i.e. $Q = \frac{72\ln\left(\frac{1}{\eta}\right)}{\delta^2}$ reparameterization finishes the proof of Theorem 38.

7.7 Technical Lemmas

Lemma 46 (Supremum)**.**

$$\left|\sup_f a_f - \sup_f b_f\right| \leq \sup_f |a_f - b_f|.$$

Lemma 47 (Bounded difference property of Z)**.** *Let $N \in \mathbb{Z}^+$, $(S_q)_{q\in[Q]}$ be a partition of $[N]$, $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a kernel, μ be the mean embedding associated to K, $x_{1:N}$ be i.i.d. random variables on $\mathcal{X}$, $Z(x_V) = \sup_{f\in B_K}\sum_{q\in U}\left[\phi\left(\frac{2\langle f, \mu_{S_q} - \mu_{\mathbb{P}}\rangle_K}{\epsilon}\right) - \mathbb{E}\phi\left(\frac{2\langle f, \mu_{S_q} - \mu_{\mathbb{P}}\rangle_K}{\epsilon}\right)\right]$, where $U \subseteq [Q]$, $V = \cup_{q\in U} S_q$. Let x'_{V_i} be x_V except for the $i \in V$-th coordinate; x_i is changed to x'_i. Then*

$$\sup_{x_V\in\mathcal{X}^{|V|}, x'_i\in\mathcal{X}}\left|Z(x_V) - Z\left(x'_{V_i}\right)\right| \leq 4,\ \forall i \in V.$$

Proof. Since $(S_q)_{q\in[Q]}$ is a partition of $[Q]$, $(S_q)_{q\in U}$ forms a partition of V and there exists a unique $r \in U$ such that $i \in S_r$. Let

$$\begin{aligned}&Y_q := Y_q(f, x_V),\\ &q \in U = \phi\left(\frac{2\langle f, \mu_{S_q} - \mu_{\mathbb{P}}\rangle_K}{\epsilon}\right) - \mathbb{E}\phi\left(\frac{2\langle f, \mu_{S_q} - \mu_{\mathbb{P}}\rangle_K}{\epsilon}\right),\\ &Y'_r := Y_r(f, x'_{V_i}).\end{aligned}$$

In this case

$$\begin{aligned}\left|Z(x_V) - Z\left(x'_{V_i}\right)\right| &= \left|\sup_{f\in B_K}\sum_{q\in U} Y_q - \sup_{f\in B_K}\left(\sum_{q\in U\setminus\{r\}} Y_q + Y'_r\right)\right| \\ &\overset{(a)}{\leq} \sup_{f\in B_K}|Y_r - Y'_r| \overset{(b)}{\leq} \sup_{f\in B_K}\Big(\underbrace{|Y_r|}_{\leq 2} + \underbrace{|Y'_r|}_{\leq 2}\Big) \leq 4,\end{aligned}$$

where in (a) we used Lemma 46, (b) the triangle inequality and the boundedness of ϕ [$|\phi(t)| \leq 1$ for all t].

■

Lemma 48 (Uniformly bounded separable Carathéodory family)**.** *Let $\epsilon > 0$, $N \in \mathbb{Z}^+$, $Q \in \mathbb{Z}^+$, $M = N/Q \in \mathbb{Z}^+$, $\phi(t) = (t-1)\mathbb{I}_{1\leq t\leq 2} + \mathbb{I}_{t\geq 2}$, $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is a continuous kernel on the separable topological domain $\mathcal{X}$, μ is the mean embedding associated to K, $\mathbb{P}_M := \frac{1}{M}\sum_{m=1}^M \delta_{u_m}$, $\mathcal{G} = \{g_f : f \in B_K\}$, where $g_f : \mathcal{X}^M \to \mathbb{R}$ is defined as*

$$g_f(u_{1:M}) = \phi\left(\frac{2\langle f, \mu_{\mathbb{P}_M} - \mu_{\mathbb{P}}\rangle_K}{\epsilon}\right).$$

Then $\mathcal{G}$ is a uniformly bounded separable Carathéodory family: (i) $\sup_{f\in B_K}\|g_f\|_\infty < \infty$ where $\|g\|_\infty = \sup_{u_{1:M}\in\mathcal{X}^M}|g(u_{1:M})|$, (ii) $u_{1:M} \mapsto g_f(u_{1:M})$ is measurable for all $f \in B_K$, (iii) $f \mapsto g_f(u_{1:M})$ is continuous for all $u_{1:M} \in \mathcal{X}^M$, (iv) B_K is separable.

Proof.

(i) $|\phi(t)| \leq 1$ for any t, hence $\|g_f\|_\infty \leq 1$ for all $f \in B_K$.
(ii) Any $f \in B_K$ is continuous since $\mathcal{H}_K \subset C(\mathcal{X}) = \{h : \mathcal{X} \to \mathbb{R}$ continuous$\}$, so $u_{1:M} \mapsto (f(u_1), \ldots, f(u_M))$ is continuous. ϕ is Lipschitz, specifically continuous. The continuity of these two maps imply that of $u_{1:M} \mapsto g_f(u_{1:M})$, specifically it is Borel-measurable.
(ii) The statement follows by the continuity of $f \mapsto \langle f, h\rangle_K$ ($h = \mu_{\mathbb{P}_M} - \mu_{\mathbb{P}}$) and that of ϕ.
(iv) B_K is separable since $\mathcal{H}_K$ is so by assumption.

■

7.8 External Lemma

Below we state the McDiarmid inequality for self-containedness.

Lemma 49 (McDiarmid inequality)**.** *Let $x_{1:N}$ be $\mathcal{X}$-valued independent random variables. Assume that $f : \mathcal{X}^N \to \mathbb{R}$ satisfies the bounded difference property*

$$\sup_{u_1,\ldots,u_N,u'_r\in\mathcal{X}} |f(u_{1:N}) - f(u'_{1:N})| \leq c, \quad \forall n \in [N],$$

where $u'_{1:N} = (u_1, \ldots, u_{n-1}, u'_n, u_{n+1}, \ldots, u_N)$. Then for any $\tau > 0$

$$\mathbb{P}\left(f(x_{1:N}) < \mathbb{E}_{x_{1:N}}[f(x_{1:N})] + c\sqrt{\frac{\tau N}{2}}\right) \geq 1 - e^{-\tau}.$$

7.9 Pseudocode of Experiment-2

The pseudocode of the two-sample test conducted in Experiment-2 is summarized in Algorithm 5.

Algorithm 5: Two-sample test (Experiment-2)

Input: Two samples: $(X_n)_{n\in[N]}$, $(Y_n)_{n\in[N]}$. Number of bootstrap permutations: $B \in \mathbb{Z}^+$. Level of the test: $\alpha \in (0,1)$. Kernel function with hyperparameter $\theta \in \Theta$: K_θ.

Split the dataset randomly into 3 equal parts:

$$[N] = \dot{\bigcup_{i=1}^{3}} I_i, \quad |I_1| = |I_2| = |I_3|.$$

Tune the hyperparameters using the 1st part of the dataset:

$$\hat{\theta} = \operatorname*{argmax}_{\theta\in\Theta} J_\theta := \widehat{\mathrm{MMD}}_\theta((X_n)_{n\in I_1}, (Y_n)_{n\in I_1}).$$

Estimate the $(1-\alpha)$-quantile of $\widehat{\mathrm{MMD}}_{\hat{\theta}}$ under the null, using B bootstrap permutations from $(X_n)_{n\in I_2} \cup (Y_n)_{n\in I_2}$: $\hat{q}_{1-\alpha}$.

Compute the test statistic on the third part of the dataset:

$$T_{\hat{\theta}} = \widehat{\mathrm{MMD}}_{\hat{\theta}}((X_n)_{n\in I_3}, (Y_n)_{n\in I_3}).$$

Output: $T_{\hat{\theta}} - \hat{q}_{1-\alpha}$.

Bibliography

[A+08] Radoslaw Adamczak et al. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

[AAZL18] Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 4613–4623, 2018.

[AB09a] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

[AB09b] Sanjeev Arora and Boaz Barak. *Computational complexity*. Cambridge University Press, Cambridge, 2009. A modern approach.

[AC11] Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

[ACL19] Pierre Alquier, Vincent Cottet, and Guillaume Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *Annals of Statistics*, 47(4):2117–2144, 2019.

[AFW18] Md. Ashad Alam, Kenji Fukumizu, and Yu-Ping Wang. Influence function and robust variant of kernel canonical correlation analysis. *Neurocomputing*, 304:12–29, 2018.

[Agg13] Charu C. Aggarwal. *Outlier Analysis*. Springer Publishing Company, Incorporated, 2013.

[AH15] David F Andrews and Frank R Hampel. *Robust estimates of location: Survey and advances*. Princeton University Press, 2015.

[AMS99] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1, part 2):137–147, 1999.

[Aro50] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[Bah66] R. R. Bahadur. A note on quantiles in large samples. *Ann. Math. Statist.*, 37(3):577–580, 06 1966.

[BAM20] Victor-Emmanuel Brunel and Marco Avella-Medina. Propose, test, release: Differentially private estimation with high probability. *arXiv preprint arXiv:2002.08774*, 2020.

[BBL05] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005.

[BBM05] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[BBS17] Y. Baraud, L. Birgé, and M. Sart. A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517, 2017.

[BDD+17] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. Technical report, 2017. (https://arxiv.org/abs/1711.07910).

[BF04] Ludwig Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88:190–206, 2004.

[BGH14] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Estimator selection in the Gaussian setting. *Ann. Inst. Henri Poincaré Probab. Stat.*, 50(3):1092–1119, 2014.

[BJL15] Christian Brownlees, Emilien Joly, and Gábor Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 12 2015.

[BJM04] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Large margin classifiers: convex loss, low noise, and convergence rates. In S. Thrun, L. K. Saul, and P. B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, pages 1173–1180. MIT Press, 2004.

[BJM06] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[BJMO12] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Structured sparsity through convex optimization. *Statist. Sci.*, 27(4):450–468, 2012.

[BK07] Derya Birant and Alp Kut. St-dbscan: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208–221, 2007.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[BLY17] Krishnakumar Balasubramanian, Tong Li, and Ming Yuan. On the optimality of kernel-embedding based goodness-of-fit tests. Technical report, 2017. (https://arxiv.org/abs/1709.08148).

[BM02] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, 3(Spec. Issue Comput. Learn. Theory):463–482, 2002.

[BM06] Peter L Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.

[BMN12] Peter L Bartlett, Shahar Mendelson, and Joseph Neeman. ℓ_1-regularized linear regression: persistence and oracle inequalities. *Probability theory and related fields*, 154(1-2):193–224, 2012.

[Bre96] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[BSAG18] Mikolaj Binkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018.

[BT74] Albert E. Beaton and John W. Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185, 1974.

[BTA04] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

[Bub15] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[BY02] Peter Bühlmann and Bin Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

[Cat12] Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. H. Poincaré Probab. Statist.*, 48(4):1148–1185, 11 2012.

[CD01a] Michael Collins and Nigel Duffy. Convolution kernels for natural language. In *NIPS*, pages 625–632, 2001.

[CD01b] Christophe CROUX and Catherine DEHON. Robust linear discriminant analysis using s-estimators. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 29(3):473–493, 2001.

[CDG19] Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. *CoRR*, abs/1811.09380:2755–2771, 2019.

[CFB19] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L Bartlett. Fast mean estimation with sub-gaussian rates. *arXiv preprint arXiv:1902.01998*, 2019.

[CFV05] Marco Cuturi, Kenji Fukumizu, and Jean-Philippe Vert. Semigroup kernels on measures. *Journal of Machine Learning Research*, 6:1169–1198, 2005.

[CG17] Olivier Catoni and Ilaria Giulini. Dimension-free pac-bayesian bounds for matrices, vectors, and linear least squares regression, 2017. (https://arxiv.org/abs/1712.02747).

[CGR+18] Mengjie Chen, Chao Gao, Zhao Ren, et al. Robust covariance and scatter matrix estimation under huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

[CHK+20] Yeshwanth Cherapanamjeri, Samuel B Hopkins, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni. Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 601–609, 2020.

[CLL19a] Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust high dimensional learning for lipschitz and convex losses. *arXiv preprint arXiv:1905.04281*, 2019.

[CLL19b] Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust statistical learning with Lipschitz and convex loss functions. *Probability Theory and related fields*, pages 1–44, 2019.

[CS01] Louis HY Chen and Qi-Man Shao. A non-uniform Berry–Esseen bound via Stein's method. *Probability theory and related fields*, 120(2):236–254, 2001.

[CSX17] Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):44, 2017.

[Cut11] Marco Cuturi. Fast global alignment kernels. In *ICML*, pages 929–936, 2011.

[DG92] David L. Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.*, 20(4):1803–1827, 12 1992.

[DG12] Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, New York, 2012.

[DGL96] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, corrected 2nd edition, 1996. missing.

[DH83] D. Donoho and P. Huber. The notion of breakdown point. 1983.

[DHS08] Michiel Debruyne, Mia Hubert, and Johan AK Suykens. Model selection in kernel based regression using the influence function. *Journal of Machine Learning Research*, 9(Oct):2377–2400, 2008.

[DK17] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. (http://archive.ics.uci.edu/ml).

[DK19] Ilias Diakonikolas and Daniel M. Kane. Recent advances in algorithmic high-dimensional robust statistics, 2019.

[DKK+17] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. *Proceedings of the International Conference on Machine Learning*, pages 999–1008, 2017.

[DKK+19a] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

[DKK+19b] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. *Proceedings of the International Conference on Machine Learning*, pages 1596–1606, 2019.

[DKP20] Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability, 2020.

[DL19] Jules Depersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time, 2019.

[DLLO16] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

[DM15] D. Donoho and A. Montanari. Variance breakdown of huber (m)-estimators: $n/p \to m \in (1, +\infty)$. Technical report, Stanford University, 2015. Preprint available on arXiv:1503.02106.

[DRG15] Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *UAI*, pages 258–267, 2015.

[Dud14] Richard M Dudley. *Uniform central limit theorems*, volume 142. Cambridge University Press, 2014.

[DW83] Kjell A. Doksum and Chi-Wing Wong. Statistical tests based on transformed data. *Journal of the American Statistical Association*, 78(382):411–417, 1983.

[EP02] André Elisseeff and Massimiliano Pontil. Leave-one-out error and stability of learning algorithms with applications stability of randomized learning algorithms source. *International Journal of Systems Science - IJSySc*, 6, 01 2002.

[Fel] William Feller. An introduction to probability theory and its applications. *1957*.

[Fer83] Luisa Turrin Fernholz. *von Mises calculus for statistical functionals*, volume 19 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1983.

[FGRW12] Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM J. Comput.*, 41(6):1558–1590, 2012.

[FGSS08] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *NIPS*, pages 498–496, 2008.

[FK18] Jianqing Fan and Donggyu Kim. Robust high-dimensional volatility matrix estimation for high-frequency factor model. *J. Amer. Statist. Assoc.*, 113(523):1268–1283, 2018.

[FR22] R. A. Fisher and Edward John Russell. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922.

[FRR90] Christopher A Field, Elvezio Ronchetti, and Elvezio M Ronchetti. Small sample asymptotics. Ims, 1990.

[FSG13] Kenji Fukumizu, Le Song, and Arthur Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14:3753–3783, 2013.

[Gal14] François Le Gall. Powers of tensors and fast matrix multiplication. *CoRR*, abs/1401.7714, 2014.

[Gao17] Chao Gao. Robust regression via mutivariate regression depth. Technical report, University of Chicago, 2017. Preprint available on arXiv:1702.04656.

[GBR+12] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[GFKS02] Thomas Gärtner, Peter A. Flach, Adam Kowalczyk, and Alexander Smola. Multi-instance kernels. In *ICML*, pages 179–186, 2002.

[GFT+08] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander J. Smola. A kernel statistical test of independence. In *NIPS*, pages 585–592, 2008.

[GHC17] Jorge Guevara, Roberto Hirata, and Stéphane Canu. Cross product kernels for fuzzy set similarity. In *FUZZ-IEEE*, pages 1–6, 2017.

[GKKW02] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New-york, 2002.

[GR09] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM J. Comput.*, 39(2):742–765, 2009.

[Gut01] Peter Guttorp. *Simon Newcomb*, pages 197–199. Springer New York, New York, NY, 2001.

[Ham71] Frank R. Hampel. A general qualitative definition of robustness. *Ann. Math. Statist.*, 42(6):1887–1896, 12 1971.

[Ham74] Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

[Hau99] David Haussler. Convolution kernels on discrete structures. Technical report, Department of Computer Science, University of California at Santa Cruz, 1999. (http://cbse.soe.ucsc.edu/sites/default/files/convolutions.pdf).

[HB05] Matthias Hein and Olivier Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In *AISTATS*, pages 136–143, 2005.

[HBM07] Zaid Harchaoui, Francis Bach, and Eric Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In *NIPS*, pages 609–616, 2007.

[HC07] Zaïd Harchaoui and Olivier Cappé. Retrospective mutiple change-point estimation with kernels. In *IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 768–772, 2007.

[HD04] Mia Hubert and Katrien Van Driessen. Fast and robust discriminant analysis. *Computational Statistics & Data Analysis*, 45(2):301–320, 2004.

[HF00] Xuming He and Wing K Fung. High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis*, 72(2):151–162, 2000.

[HI17a] Matthew J Holland and Kazushi Ikeda. Efficient learning with robust gradient descent. *Machine Learning*, 108(8-9):1523–1560, Jun 2017.

[HI17b] Matthew J Holland and Kazushi Ikeda. Robust regression using biased objectives. *Machine Learning*, 106(9-10):1643–1679, 2017.

[HL19] Samuel B. Hopkins and Jerry Li. How hard is robust mean estimation? *CoRR*, abs/1903.07870, 2019.

[Hoa62] Charles AR Hoare. Quicksort. *The Computer Journal*, 5(1):10–16, 1962.

[Hoe48] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19(3):293–325, 1948.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[Hop20] Samuel B. Hopkins. Mean estimation with sub-Gaussian rates in polynomial time. *Annals of Statistics*, 48(2):1193–1213, 2020. (https://arxiv.org/abs/1809.07425).

[HR09] Peter J Huber and Elvezio M Ronchetti. *Robust statistics; 2nd ed.* Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 2009.

[HRRS86] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Statistics. Wiley, 1st edition edition, January 1986. missing.

[HS96] Xuming He and Qi-Man Shao. A general bahadur representation of m -estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.*, 24(6):2608–2630, 12 1996.

[Hub64] Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

[Hub67] Peter J Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. Berkeley, CA, 1967.

[Hub84] Peter J Huber. Finite sample breakdown of m-and p-estimators. *The Annals of Statistics*, pages 119–126, 1984.

[Hub91] Peter J Huber. Between robustness and diagnostics. In *IMA*, volume 33, page 121, 1991.

[HV10] Mia Hubert and Stephan Van Der Veeken. Robust classification for skewed data. *Adv. Data Anal. Classif.*, 4(4):239–254, December 2010.

[HW71] D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Statist.*, 42(3):1079–1083, 06 1971.

[HW17] Qiyang Han and Jon A. Wellner. A sharp multiplier inequality with applications to heavy-tailed regression problems. *arXiv:1706.02410*, 2017.

[JGV86] Mark R. Jerrum, Leslie G.Valiant, and Vijay V.Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43(2-3):169–188, 1986.

[JKH04] Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

[Jor13] Michael I. Jordan. On statistics, computation and scalability. *Bernoulli*, 19(4):1378–1390, 2013.

[JV16] Yunlong Jiao and Jean-Philippe Vert. The Kendall and Mallows kernels for permutations. In *ICML (PMLR)*, volume 37, pages 2982–2990, 2016.

[JXS+17] Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. In *NIPS*, pages 261–270, 2017.

[KB13] Vladimir S Korolyuk and Yu V Borovskich. *Theory of U-statistics*, volume 273. Springer Science & Business Media, 2013.

[KBSS19] Stefan Klus, Andreas Bittracher, Ingmar Schuster, and Christof Schütte. A kernel-based approach to molecular conformation analysis. *The Journal of Chemical Physics*, 149:244109, 2019.

[KFH16] Genki Kusano, Kenji Fukumizu, and Yasuaki Hiraoka. Persistence weighted Gaussian kernel for topological data analysis. In *ICML*, pages 2004–2013, 2016.

[KGF+10] Bharath K., Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

[KK02] Hisashi Kashima and Teruo Koyanagi. Kernels for semi-structured data. In *ICML*, pages 291–298, 2002.

[KKK16] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *NIPS*, pages 2280–2288, 2016.

[KL17] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions, 2017.

[Kle05] Lev Klebanov. *N-Distances and Their Applications*. Charles University, Prague, 2005.

[KM15] Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

[Kol11] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School].

[KP16] Risi Kondor and Horace Pan. The multiscale Laplacian graph kernel. In *NIPS*, pages 2982–2990, 2016.

[KS12] JooSeuk Kim and Clayton D. Scott. Robust kernel density estimation. *Journal of Machine Learning Research*, 13:2529–2565, 2012.

[KSM18] Stefan Klus, Ingmar Schuster, and Krikamol Muandet. Eigendecompositions of transfer operators in reproducing kernel Hilbert spaces. Technical report, 2018. (https://arxiv.org/abs/1712.01572).

[LCB19] Pierre Laforgue, Stephan Clemencon, and Patrice Bertail. On medians of (Randomized) pairwise means. volume 97 of *Proceedings of Machine Learning Research*, pages 1272–1281, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[LDG+14] James Robert Lloyd, David Duvenaud, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Automatic construction and natural-language description of nonparametric regression models. In *AAAI Conference on Artificial Intelligence*, pages 1242–1250, 2014.

[Le 73] Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1:38–53, 1973.

[Ler19] Matthieu Lerasle. Lecture notes: Selected topics on robust statistical learning theory, 2019.

[LL18] Guillaume Lecué and Matthieu Lerasle. Learning from MOM's principles: Le Cam's approach. *Stochastic Processes and their Applications*, 2018. (https://doi.org/10.1016/j.spa.2018.11.024).

[LL20] Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: theory and practice. *Annals of Statistics*, 48(2):906–931, 2020. (To appear; preprint: https://arxiv.org/abs/1711.10306).

[LLM20] Guillaume Lecué, Matthieu Lerasle, and Timothée Mathieu. Robust classification via MOM minimization. *Machine Learning*, 109(8):1635–1665, 27 April 2020.

[LM10] Guillaume Lecué and Shahar Mendelson. Sharper lower bounds on the performance of the empirical risk minimization algorithm. *Bernoulli*, pages 605–613, 2010.

[LM13] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. Technical report, CNRS, Ecole polytechnique and Technion, 2013.

[LM+18] Guillaume Lecué, Shahar Mendelson, et al. Regularization and the small-ball method i: sparse recovery. *The Annals of Statistics*, 46(2):611–641, 2018.

[LM19a] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

[LM19b] Gabor Lugosi and Shahar Mendelson. Regularization, sparse recovery, and median-of-means tournaments. *Preprint available on arXiv:1701.04112*, 25(3):2075–2106, 2019.

[LM19c] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 22(3):925–965, 2019. (To appear; preprint: https://arxiv.org/abs/1608.00757).

[LM+19d] Gábor Lugosi, Shahar Mendelson, et al. Sub-gaussian estimators of the mean of a random vector. *The annals of statistics*, 47(2):783–794, 2019.

[LO11] Matthieu Lerasle and Roberto I Oliveira. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*, 2011.

[Loh18] Po-Ling Loh. Scale calibration for high-dimensional robust regression. *arXiv preprint arXiv:1811.02096*, 2018.

[LRV16] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.

[LSC20] Pierre Laforgue, Guillaume Staerman, and Stephan Clémençon. How robust is the median-of-means? concentration bounds in presence of outliers. *arXiv preprint arXiv:2006.05240*, 2020.

[LSCB15] R. J. Lyon, B. W. Stappers, S. Cooper, and J. D. Brooke, J. M. Knowles. Fifty years of pulsar candidate selection: From simple filters to a new principled real-time classification approach. *MNRAS*, 000:000–000, 2015.

[LSML19] Matthieu Lerasle, Zoltán Szabó, Timothée Mathieu, and Guillaume Lecué. Monk outlier-robust mean embedding estimation by median-of-means. In *International Conference on Machine Learning*, pages 3782–3793, 2019.

[LSSF18] Ho Chung Leon Law, Dougal J. Sutherland, Dino Sejdinovic, and Seth Flaxman. Bayesian approaches to distribution regression. *AISTATS (PMLR)*, 84:1167–1176, 2018.

[LSST+02] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.

[LSZ15] Yujia Li, Kevin Swersky, and Richard Zemel. Generative moment matching networks. In *ICML (PMLR)*, pages 1718–1727, 2015.

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 1991. Isoperimetry and processes, Reprint of the 1991 edition.

[LTZ08] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 413–422. IEEE, 2008.

[Mat20a] Timothée Mathieu. Bound on the bias of m-estimators and application to high dimensional mean estimation, 2020.

[Mat20b] Timothée Mathieu. Robustness to outliers and concentration of M-estimators by means of influence function, 2020.

[MB11] Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

[MDC14] Dina Mayzlin, Yaniv Dover, and Judith Chevalier. Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, 104(8):2421–55, 2014.

[Men] Shahar Mendelson. An optimal unrestricted learning procedure. *Preprint available on arXiv:1707.05342*.

[Men08] Shahar Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Transactions on Information Theory*, 54(8):3797–3803, 2008.

[Men15] Shahar Mendelson. Learning without concentration. *Journal of the ACM*, 62(3):21:1–21:25, 2015.

[Men16] Shahar Mendelson. Upper bounds on product and multiplier empirical processes. *Stochastic Processes and their Applications*, 126(12):3652–3680, 2016.

[MFBS17] Krikamol Muandet, Kenji Fukumizu, Sriperumbudur Bharath, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

[MFDS11] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In *NIPS*, pages 10–18, 2011.

[Min15] Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

[Min18] Stanislav Minsker. Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523*, 2018.

[Min20] Stanislav Minsker. Asymptotic normality of robust risk minimizers. *arXiv preprint arXiv:1812.03523*, 2020.

[MKF+16] Krikamol Muandet, Bharath K., Sriperumbudur, Kenji Fukumizu, Arthur Gretton, and Bernhard Schölkopf. Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 17:1–41, 2016.

[MM19] Stanislav Minsker and Timothée Mathieu. Excess risk bounds in robust empirical risk minimization, 2019.

[MPJ+16] Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17:1–102, 2016.

[MRT12] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

[MS17] Stanislav Minsker and Nate Strawn. Distributed statistical estimation and rates of convergence in normal approximation. Technical Report 2, 2017. (https://arxiv.org/abs/1704.02658).

[MSX+09] André F. T. Martins, Noah A. Smith, Eric P. Xing, Pedro M. Q. Aguiar, and Mário A. T. Figueiredo. Nonextensive information theoretic kernels on measures. *The Journal of Machine Learning Research*, 10:935–975, 2009.

[MT99] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.

[Mül97] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 1997.

[Nec15] Ernest Fokoue Necla Gunduz. Robust classification of high dimension low sample size data. 2015.

[NY83] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[O'D14] Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, 2014.

[PBSP17] Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2017.

[PJS16] Mijung Park, Wittawat Jitkrittum, and Dino Sejdinovic. K2-ABC: Approximate Bayesian computation with kernel embeddings. In *AISTATS (PMLR)*, volume 51, pages 51:398–407, 2016.

[PSBR20] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B*, 82(3):601–627, 2020.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[Ras89] Jeffrey Lee Rasmussen. Data transformation, type i error rate and power. *British Journal of Mathematical and Statistical Psychology*, 42(2):203–213, 1989.

[RLSP18] Anant Raj, Ho Chung Leon Law, Dino Sejdinovic, and Mijung Park. A differentially private kernel two-sample test. Technical report, 2018. (https://arxiv.org/abs/1808.00380).

[Ron97] Elvezio Ronchetti. Robust inference by influence functions. *Journal of Statistical Planning and Inference*, 57(1):59–72, 1997. Robust Statistics and Data Analysis, Part I.

[Ron19] Forbes Ron Schmelzer. The Achilles' heel of AI. https://www.forbes.com/sites/cognitiveworld/2019/03/07/the-achilles-heel-of-ai, 2019.

[Ros58] Frank Rosenblatt. *The perceptron: A theory of statistical separability in cognitive systems*. Cornell Aeronautical Laboratory, Inc., Rep. No. VG-1196-G-1. U.S. Department of Commerce, Office of Technical Services, PB 151247, 1958.

[Rot01] Volker Roth. Probabilistic discriminative kernel classifiers for multi-class problems, 2001.

[Sau18] Adrien Saumard. On optimality of empirical risk minimization in linear aggregation. *Bernoulli*, 24(3):2176–2203, 2018.

[SBPG16] Zoltán Szabó, Sriperumbudur Bharath, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016.

[SC08] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

[SCV17] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. *CoRR*, abs/1703.04940, 2017.

[SGB+11] Le Song, Arthur Gretton, Danny Bickson, Yucheng Low, and Carlos Guestrin. Kernel belief propagation. In *AISTATS*, pages 707–715, 2011.

[SGRA18] Beatriz Sinova, Gil González-Rodríguez, and Stefan Van Aelst. M-estimators of location for functional data. *Bernoulli*, 24:2328–2357, 2018.

[SGSS07] Alexander Smola, Arthur Gretton, Le Song, and Bernhard Schöolkopf. A Hilbert space embedding for distributions. In *ALT*, pages 13–31, 2007.

[SHS01] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *COLT*, pages 416–426, 2001.

[SKGF13] Dino Sejdinovic, Bharath K., Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41:2263–2291, 2013.

[SMF+15] Bernhard Schölkopf, Krikamol Muandet, Kenji Fukumizu, Stefan Harmeling, and Jonas Peters. Computing functions of random variables via reproducing kernel Hilbert space representations. *Statistics and Computing*, 25(4):755–766, 2015.

[SR04] Gábor J. Székely and Maria L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.

[SR05] Gábor J. Székely and Maria L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93:58–80, 2005.

[SS02] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[Sza14] Zoltán Szabó. Information theoretical estimators toolbox. *Journal of Machine Learning Research*, 15:283–287, 2014.

[Tal14] Michel Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.

[TKM17] Ilya Tolstikhin, Bharath K., Sriperumbudur, and Krikamol Muandet. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18:1–47, 2017.

[TKS16] Ilya Tolstikhin, Bharath K., Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximal mean discrepancy with radial kernels. In *NIPS*, pages 1930–1938, 2016.

[Tsy04] Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[Tuk60] John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

[Tuk62] John W Tukey. The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67, 1962.

[Van16] Sara Van de Geer. Estimation and testing under sparsity. *Lecture Notes in Mathematics*, 2159, 2016.

[Vap98] Vladimir N. Vapnik. *Statistical learning theory.* Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., New York, 1998. A Wiley-Interscience Publication.

[Vap00] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory.* Statistics for Engineering and Information Science. Springer-Verlag, Berlin, Heidelberg, second edition, 2000.

[vdVW96] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[Vil09] Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.

[VS13] Robert Vandermeulen and Clayton Scott. Consistency of robust kernel density estimators. In *COLT (PMLR)*, volume 30, pages 568–591, 2013.

[VSKB10] S.V. N. Vishwanathan, Nicol N. Schraudolph, Risi Kondor, and Karsten M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010.

[VvdG00] Sara A Van de Geer and Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge University Press, 2000.

[YCKB18] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659, 2018.

[YUFT18] Makoto Yamada, Yuta Umezu, Kenji Fukumizu, and Ichiro Takeuchi. Post selection inference with kernels. In *AISTATS (PMLR)*, volume 84, pages 152–160, 2018.

[ZBFL18] Wen-Xin Zhou, Koushiki Bose, Jianqing Fan, and Han Liu. A new perspective on robust m-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *Ann. Statist.*, 46(5):1904–1931, 10 2018.

[ZJS19] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics. *arXiv preprint arXiv:1909.08755*, 2019.

[ZKK92] A. A. Zinger, A. V. Kakosyan, and L. B. Klebanov. A characterization of distributions by mean values of statistics and certain probabilistic metrics. *Journal of Soviet Mathematics*, 1992.

[ZKR+17] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan R. Salakhutdinov, and Alexander J. Smola. Deep sets. In *NIPS*, pages 3394–3404, 2017.

[Zol83] V. M. Zolotarev. Probability metrics. *Theory of Probability and its Applications*, 28:278–302, 1983.

[ZSMW13] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. *Journal of Machine Learning Research*, 28(3):819–827, 2013.

université PARIS-SACLAY

ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)

Titre: M-estimation et Médiane des Moyennes appliquées à l'apprentissage statistique.

Mots clés: Robustesse, apprentissage statistique, médiane des moyennes.

Résumé: Le principal objectif de cette thèse est d'étudier des méthodes d'apprentissage statistique robuste. Traditionnellement, en statistique nous utilisons des modèles ou des hypothèses simplificatrices qui nous permettent de représenter le monde réel tout en sachant l'analyser convenablement. Cependant, certaines déviations des hypothèses peuvent fortement perturber l'analyse statistique d'une base de données. Par statistiques robuste, nous entendons ici des méthodes pouvant gérer d'une part des données dites anormales (erreur de capteur, erreur humaine) mais aussi des données de nature très variables. Nous appliquons ce genre de technique à l'apprentissage statistique, donnant ainsi des assurances théoriques d'efficacité des méthodes proposées ainsi que des illustrations sur des données simulées et réelles.

Title: M-estimation and Median of Means applied to statistical learning

Keywords: Robustness, machine learning, median of means.

Abstract: The main objective of this thesis is to study methods for robust statistical learning. Traditionally, in statistics we use models or simplifying assumptions that allow us to represent the real world. However, some deviations from the hypotheses can strongly disrupt the statistical analysis of a database. By robust statistics, we mean methods that can handle on the one hand so-called abnormal data (sensor error, human error) but also data of a highly variable nature. We apply robust techniques to statistical learning, giving theoretical efficiency results of the proposed methods as well as illustrations on simulated and real data.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France