# Quelques problématiques autour du clustering : robustesse, grande dimension et détection d'intrusion

Edouard Genetay

# THESE DE DOCTORAT DE

L'ECOLE NATIONALE DE LA STATISTIQUE
ET DE L'ANALYSE DE L'INFORMATION

ECOLE DOCTORALE N° 601
*Mathématiques et Sciences et Technologies de l'Information et de la Communication*
Spécialité : Mathématiques et leurs intéractions

Par

**GENETAY Edouard**

# Quelques problématiques autour du clustering : robustesse, grande dimension et détection d'intrusion.

**Thèse présentée et soutenue à l'ENSAI, le 16 mai 2022**
**Unité de recherche : CREST-ENSAI, UMR 9194**

**Rapporteurs avant soutenance** :

| | |
|---|---|
| Erwann Le Pennec | Professeur, Ecole Polytechnique |
| Stéphane Chrétien | Professeur, Université Lyon 2 |

**Composition du Jury :**

| | | |
|---|---|---|
| Président : | Guillaume Lecué | Professeur associé, ENSAE, IP Paris |
| Examinateurs : | Claire Brécheteau | Maitre de conférences, Université Rennes 2 |
| | Stéphane Chrétien | Professeur, Université Lyon 2 |
| | Baptiste Gregorutti | Directeur technique, LumenAI |
| | Guillaume Lecué | Professeur associé, ENSAE, IP Paris |
| | Erwann Le Pennec | Professeur, Ecole Polytechnique, IP Paris |
| | Clément Levrard | Maitre de conférences, Université de Paris |
| Dir. de thèse : | Adrien Saumard | Professeur associé, ENSAI |
| Co-dir. de thèse : | Valentin Patilea | Professeur, ENSAI |

# Remerciements

Je tiens à remercier Adrien Saumard pour avoir rendu cette thèse possible du début à la fin malgré ses hauts et ses bas. On ne compte plus les difficultés rencontrées pendant ces 4 ans: stage, financement, avancement, jalons, inscription... Tant d'aléas qui auraient pu dégénérer ou même pousser la thèse à s'arrêter. Mais malgré cela, grace à tes qualités humaines indéniables, diplomatie, écoute, patience et indulgence, la collaboration s'est sublimée pour déboucher sur un an et demi stimulant, un plaisir retrouvé et un travail bien plus abouti que je n'osais l'imaginer. Merci à toi, vraiment.

J'aimerais remercier chaleureusement mon autre directeur de thèse, Valentin Patilea, pour m'avoir respecté comme il l'a fait. Jusqu'ici, dans ma vie de jeune trentenaire, peu de gens ont su en faire autant. Je t'avais déjà exprimé ma gratitude et je recommence ici.

J'aimerais avoir un mot particulier à adresser à chacun des membres de mon jury mais nous nous sommes très peu croisés alors simplement merci à vous Erwann Le Pennec et Stéphane Chrétien d'avoir accepté avec enthousiasme d'être rapporteur de ma thèse et du temps que vous y avez consacré. Merci également aux autres examinateurs, Claire Brécheteau, Clément Levrard et Guillaume lecué, d'avoir répondu présent.

Ceux qui m'ont assez cotoyé ont pu constater que la tenacité qui me permet de mener des calculs fastidieux ne m'est d'aucune utilité quand il s'agit de complèter et signer des papiers... J'aimerais donc remercier tout particulièrement Cécile Terrien de m'avoir accompagné avec rigueur et simplicité dans toutes les démarches administratives.

La dynamique haute en couleur le LumenAI m'a amené à croiser un bon nombre de personnes qui méritent d'être remerciées. Tout d'abord merci à Camille Brunet-Saumard de m'avoir formé à python et aux simulations numériques. Merci à Jonathan Labéjof et Chemsdin Naaman pour votre amitié et dévouement. Merci à Amine Medad, Anis Mokhtari et Alexandre Peter Nguema pour avoir donné cette bonne ambiance en télétravail. Merci à Benedetta Ge-

miliana Pastore d'avoir fait le choix de se battre pour LumenAI. Merci à Jeanne-Sophie Goyer pour son honneteté, maturité et simplicité. Ta as eu un role indispensable dans l'équipe. Merci enfin à Baptiste Gregorutti d'avoir fait le choix de me laisser le temps nécessaire à la réalisation des travaux académiques et d'avoir continué à me former. Ton role de manager a été très apprécié.

Merci à Carole Essirard, Corinne Barzic et Aurélie Duchesne dans l'organisation des déplacements et des TD, ainsi que pour tous ces moments de folie qui resteront associés à ce temps passé à l'ENSAI.

Le témoignage de Cédric Villani sur le fonctionnement de la recherche et la grande quantité d'articles avec plusieurs coauteurs m'avaient laissé penser que la collaboration était au coeur du métier de chercheur. Mais mon expérience personnelle s'est plutot rapprochée d'un quotidien ô combien solitaire où l'on ne peut compter que sur l'extase intellectuelle comme motivation, réconfort et satisfaction. C'est pourquoi je remercie infiniment Romaric Gaudel, Dang Hong Phuong et Basile De Loynes de m'avoir montré qu'il pouvait en être autrement. Vous êtes sortis des sentiers battus pour m'intégrer au-delà de nos statuts. Votre impact a été réel. Je ne peux que souhaiter que votre démarche formatrice et inclusive profite aux futurs doctorants comme elle m'a profité également.

Merci à mes amis de Porquerolle, Chamonix et Coulonger pour vos mots rassurants et apaisants.

Je pense que je me dois aussi de remercier Rasool Mehdizadeh, de l'école des Mines de Nancy, car si cette thèse a porté ses fruits et que j'en suis arrivé là aujourd'hui, c'est en premier lieu grace à lui. Il a su semer la première graine en me poussant vers la recherche à un moment charnière de ma vie. De plus, j'en profite pour remercier du fond du coeur Frédéric Prioa et Etienne Mann, du Larema d'Angers. Ils ont cru tous les 3 en moi et ça m'a fait du bien de me le rappeler quand je me suis senti perdu. Maintenant à l'issu de ces 3 années, on peut dire que ma production a été bien moins velue que ce qu'ils avaient anticipé mais je leur suis très reconnaissant de m'avoir permis de concrétiser ce rêve auquel j'avais bien failli renoncer.

Et bien sur, ces remerciements ne seraient pas complets si je ne te remerciais pas, toi, Juliette, qui partage ma vie depuis 13 ans inconditionnellement. Merci à toi car même ces trois années de parkour n'ont eu raison ni de ton soutien, ni de ton amour.

# Contents

# Chapter 1

# Introduction

**Résumé.** Dans ce chapitre nous présentons le contexte général de cette thèse CIFRE dédiée à l'étude de divers problèmes liés au clustering. Après avoir présenté LumenAI, l'entreprise qui a financé la thèse, nous présentons succinctement l'évolution du domaine de l'apprentissage statistique et comment ces évolutions ont poussé le clustering à s'étendre à de nouveaux types de données et de nouvelles techniques. Enfin, nous y détaillons le contexte et notre apport dans chacun des thèmes d'étude, à savoir : le développement d'un algorithme pratique de clustering robuste de vecteurs, une contribution théorique en clustering en grande dimension et la recherche d'une méthode de détection automatique de cyber attaque à l'aide de clustering sur graphe.

## 1.1 LumenAI: une société de service experte en Data Science

LumenAI a été fondée par Sébastien Loustau[1] et Wajdi Farhani[2] en 2014. C'est une entreprise de conseil en Machine Learning apportant à ses clients des compétences allant de la data science au data engineering. Les prestations se déclinent sous forme d'exploration de données, de conseils techniques, de R&D, de développement de service dédié

[1] https://sebastienloustau.github.io/ .
[2] https://www.linkedin.com/in/wajdi-farhani/?originalSubdomain=fr .

ou de mise en production. L'équipe est aujourd'hui spécialisée en traitement naturel du langage[3], en clustering et analyse de graphes.

En 2014, Sébastien Loustau était encore Maître de Conférence à l'université d'Angers. Sa motivation en créant LumenAI était principalement d'apporter à l'industrie, au sens large, les méthodes récentes de Machine Learning. Il travaillait à l'époque à l'élaboration de résultats de type PAC-Bayésiens[4] [64]. Ses travaux ont abouti notamment à deux algorithmes d'apprentissage statistique en temps réel, DistClus, permettant de classifier des vecteurs de manière non-supervisée [95], et GrphClus, effectuant une tâche de détection de communautés dans des graphes [39]. Ces deux algorithmes font aujourd'hui partie du coeur technologique propriétaire de LumenAI. GrphClus est notamment une brique du produit *Lady Of The Lake* dédié à la dédection d'intrusion dans des réseaux. Dans le cadre du partenariat CIFRE avec le CREST-ENSAI, mes encadrants à LumenAI étaient intéressés par tout apport autour du clustering ; offrant de fait, une grande latitude dans la thèse pour développer des recherches autours de ce thème, sous des aspects modernes d'analyse de données. Dans cette thèse nous nous intéressons donc au clustering dans trois contextes différents : robustesse, grande dimension et réseau.

## 1.2 Enjeux modernes du clustering

Le clustering consiste à regrouper des données en "clusters", c'est-à-dire en groupes de données présentant des similarités. Cette tâche est en général difficile parce que les similarités en question ne sont pas absolues, dans le sens où elles dépendent d'à priori métier sur les données, des méthodes utilisées pour les analyser et des finalités recherchées. Il existe donc plusieurs formalismes possibles permettant de cerner la notion de cluster.

On peut distinguer deux approches de clustering : l'une dite "hard" et l'autre "soft". Le "hard" clustering (ou dur) est un partitionnement des données observées en clusters disjoints. Un tel clustering est souvent relié à des critères géométriques (K-means, classification hiérarchique). Le "soft" clustering (ou probabiliste) attribue quant à lui à chaque point une probabilité d'appartenir à chacun des clusters présents. Ces techniques "soft" sont souvent reliées à des modèles génératifs, c'est-à-dire visant à modéliser la loi de distribution des données, à l'instar des modèles de mélanges.

[3] Natural Language Processing en anglais, souvent abbrégé en "NLP".

[4] "Probably Approximately Correct" et bayésien. Pour plus de détails, se référer au workshop de Benjamin Guedj à cette adresse https://project.inria.fr/inriacwi/files/2018/11/BenjaminGuedj-6PAC-talk-sept2018.pdf .

Le clustering est en premier lieu une méthode d'analyse de données (data mining). A ce titre les enjeux modernes dans ce domaine suivent l'évolution de leur contexte d'utilisation et de la nature des données recueillies, de plus en plus nombreuses et complexes.

### 1.2.1 Le clustering s'adapte aux types de données

On rencontre en data science un grand nombre de typologie de données : variables continues, discrètes ou catégorielles, des données relationnelles, fonctionnelles, temporelles, mais aussi des images, textes, vidéos ou enregistrements sonores.

Depuis la fin des années 90, la dimension des données et leur nombre sont devenus tels que le temps de calcul des algorithmes est une limite pratique incontournable. Les méthodes ont dues être adaptées et ont donné naissance à ce qu'on appelle aujourd'hui les statistiques en grande dimension.

Au problème de la dimension s'ajoute l'apparition de nouveaux types de données ou de corpus de données hétérogènes. Des travaux spécifiques ont vu le jour pour, notamment, adapter les métriques, définir de nouveaux critères et étudier de nouveaux modèles génératifs de données.

Pour illustrer cette complexification au cours du temps, on peut prendre l'exemple de l'indexation de page webs. Cette tâche n'utilisait à l'origine que les hyperliens[5] entre les pages web pour évaluer leur pertinence. PageRank [116], même s'il est dépassé, est l'exemple phare d'algorithme qui exploite ces relations. Les algorithmes ont évolué avec le temps pour integrer des aspects temporels (mis à jour des sites), de l'analyse de mise en page, de texte, d'image, etc. Aujourd'hui Google utiliserait plus de 200 caractéristiques[6] différentes pour indexer les pages web et on estime que le nombre de page web qu'ils indexent fluctue autour de 40 milliards[7].

On peut également prendre comme exemple les données fonctionnelles. Elles sont nées au début des années 1980 [123, 129] et n'ont pas tardé à être généralisées au cadre multivarié [124] entre 1997 et 2003. Elles ont été également indexées par le temps [141] ou encore plongées dans des espaces non-euclidiens pour suivre la migration des oiseaux [136]. Les usages et besoins actuels motivent en effet le développement des techniques. En particulier, extraire les informations disponibles dans les données fonctionnelles est un défi d'actualité pour certaines approches

[5] Les hyperliens constituent une forme de relation entre les pages webs. Cela rentre dans la catégorie des données relationnelles.

[6] d'après https://www.monsterinsights.com/google-ranking-factors/. Consulté le 24 novembre 2021.

[7] https://www.worldwidewebsize.com/ ; noter que google anonce plus de 100 milliard de pages sur ce site https://www.google.com/intl/fr/search/howsearchworks/how-search-works/organizing-information/. Consulté le 23 décembre 2021.

de développements de la voiture autonome [59].

### 1.2.2 Le clustering s'adapte aux contextes d'utilisation

Les contextes d'utilisation ont aussi évolué. Avec la croissance des capacités numériques (Figure 1.2.1 and 1.2.2), une problématique est apparue : comment traiter des jeux de données dont la taille est plus volumineuse que la mémoire disponible sur un seul ordinateur ? Cette question est connexe à l'avènement du "big data". Les techniques ont donc évolué pour permettre de distribuer le calcul sur plusieurs machines ou sur plusieurs processeurs. On peut notamment citer la stratégie "diviser pour mieux reigner"[8], le "peer-to-peer" ou le calcul avec GPU [37].

Dans le même temps, on a commencé à distinguer les algorithmes dits "online" et "offline". Le cadre classique des statistiques correspond en effet au contexte "offline", c'est-à-dire qu'elles exploitent un échantillon de données déjà disponibles. Au contraire, le paradigme "online" regroupe des méthodes permettant l'analyse d'un corpus de données en même temps qu'il se constitue.

Aujourd'hui, les thèmes émergeants foisonnent comme par exemple l'apprentissage statistique (clustering inclus) sous contraintes de confidentialité ou d'équité. Dans le premier cas, on cherche des procédures d'exploitation de données en garantissant un certain niveau d'anonymat des individus. Dans ce domaine, mentionnons notamment la "confidentialité différentielle" [46], une méthode permmettant la mise ne place efficace et flexible, d'une part, et une analyse théorique d'autre part. Et dans le second cas, il est notamment question d'apprendre malgré des biais de représentativité de certaines données, comme une sous-représentation des femmes dans un sondage par exemple, ou encore de trouver des procédures qui garantissent des critères d'équité.

En plus des thèmes émergeants, il y a de nombreux domaines de recherche très actifs dont les statistiques robustes et en grande dimension, deux thèmes étudiés pendant la thèse. La section suivante les contextualise plus en détail.

[8] 5 méthodes sont décrites ici en anglais : https://bigdata.oden.utexas.edu/project/divide-conquer-methods-for-big-data-analytics/ . Consulté le 24novembre 2021.

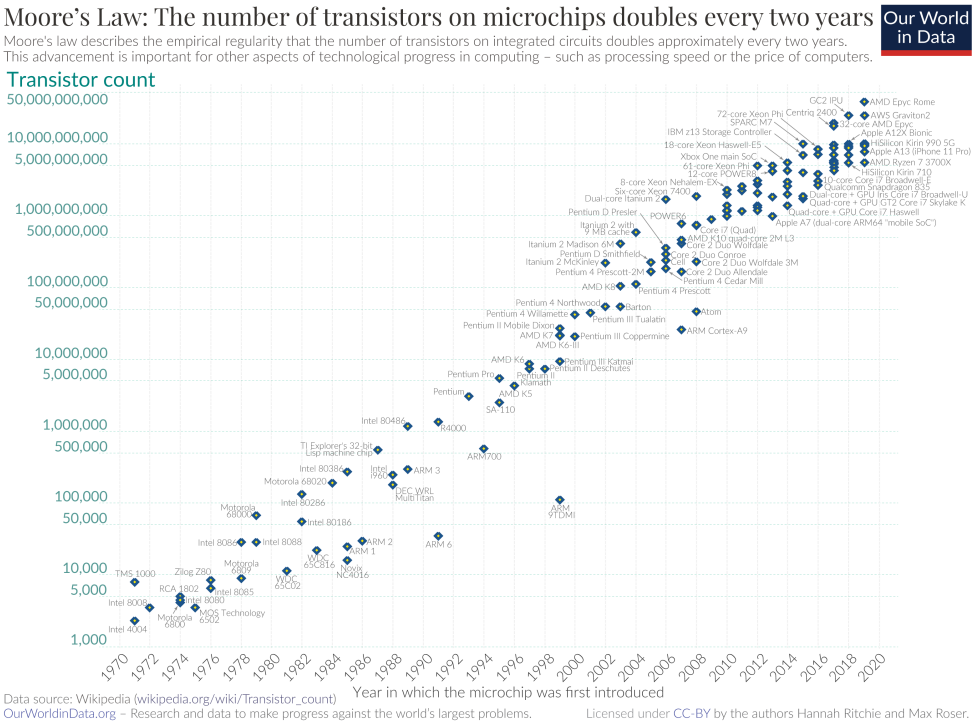



Figure 1.2.1: Evolution des capacités de miniaturisation afin d'intégrer toujours plus de transistors dans les circuits intégrés. Consultable en ligne. Il est a noté que les limites physiques sont aujourd'hui atteintes en termes de miniaturisation puisque les plus petits transistors sont un assemblage de l'ordre de quelques atomes.

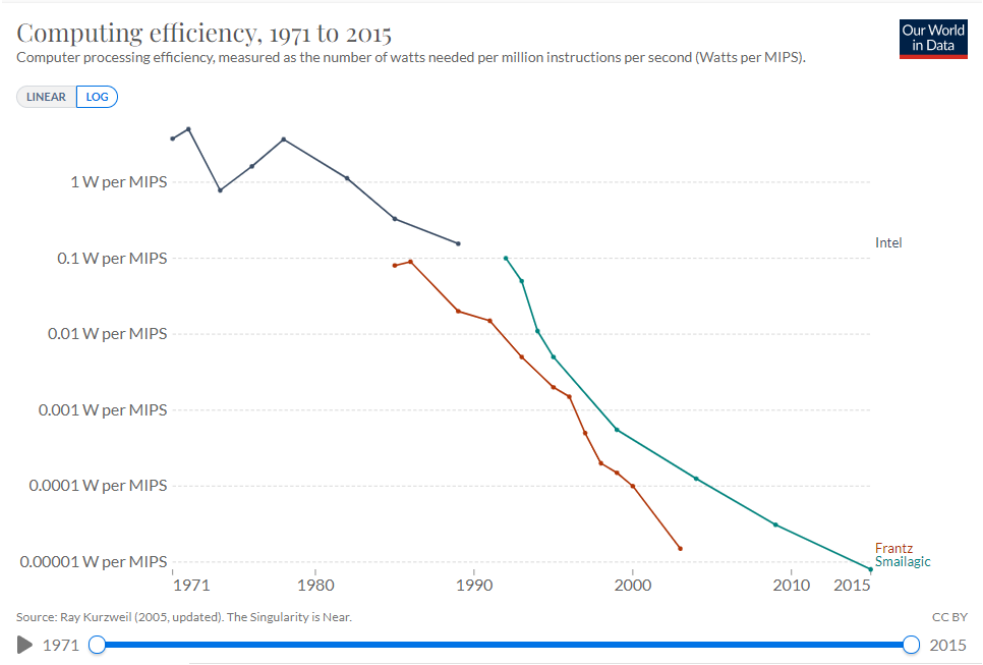



Figure 1.2.2: Evolution de l'efficacité énergétique des calculs. Notez que ce progrès tendantiel est appelé "loi de koomey" . De plus, la limite physique dans ce domaine s'appelle le principe de landauer [14, 144] ($\approx 3.10^{-21}$ Joules par bit effacé à température ambiante) et n'est pas encore atteinte.

## 1.3 Organisation de la thèse

### 1.3.1 Clustering robuste

L'apprentissage statistique robuste est un champ de recherche qui vise à créer des méthodes d'apprentissage capables de remplir leur role malgré une "pollution" dans les données. Cette "pollution" pour le statisticien peut prendre différentes formes, comme des problèmes de formats ou de données manquantes, mais aussi d'erreurs de mesure ou de report de valeur, dégradant ainsi la qualité des résultats des algorithmes.

Le pré-traitement des jeux de données par le statisticien est donc une nécessité, parfois très chronophage. C'est pourquoi l'avènement du big data, et des jeux de données digitaux de manière générale, ont induit un renouveau de la thématique de la robustesse veille d'au moins 40 ans [65, 66]. Parmi les solutions existantes, on trouve d'une part les méthodes tolérant une proportion d'outliers, c'est-à-dire des données aberrantes. On parle en fonction des contextes de contamination de Huber[9] (quand une fraction des données suit une autre distribution), de contamination paramétrique (quand une proportion des données est générée avec le même type de loi paramétrique que les inliers pour d'autres valeurs de paramètres) ou de contamination malicieuse, "adversarial" en anglais (quand une fraction des données a un comportement complètement quelconque). D'autre part, on trouve les méthodes tolérant des distributions de données à "queues lourdes". Typiquement, les distributions à queues lourdes sont les distributions dont les densités tendent moins vite vers zéro que la fonction exponentielle.

Récemment, l'estimateur de la moyenne dit "Median-of-Means" a été mis en avant, notamment pour ses propriétés théoriques intéressantes. C'est un estimateur flexible (car il permet entre autre une parallélisation des calculs), tolérant à des queues lourdes et des outliers, et il satisfait à des critères d'optimalité en termes de concentration sous-gaussienne [41]. Cet estimateur est détaillé en section 1.4.2 et 2.2.1.

### 1.3.2 Clustering en grande dimension

Le clustering en grande dimension est un domaine à la littérature déjà bien fournie puisqu'il est étudié depuis plus de 20 ans [57].

Notre contribution au domaine consiste en l'apport des premiers développements théoriques sur une méthode de

[9] Pour plus de détails sur toutes les contaminations, voir la conférence de Arnak Dalalyan: https://adalalyan.github.io/Download/Dalalyan_Frejus_2018a.pdf , consulté le 4 janvier 2021.

minimisation d'entropie apparue dans les travaux d'apprentissage statistique pour traiter des données complexes. Cet aspect est détaillé en section 1.5 de l'introduction ou dans le chapitre 3 du manuscrit.

### 1.3.3 Clustering de graphe dynamique

Le clustering de graphe, ou détection de communautés, est une technique de clustering adaptée aux graphes. Les graphes apparaissent naturellement lorsque l'on manipule des données relationnelles. En l'occurrence, nous avons travaillé dans un contexte de clustering de graphe dynamique, c'est-à-dire sur un graphe évoluant en parallèle de la tâche de clustering. Cette thématique a été investiguée dans le cadre d'un partenariat avec la Direction Générale des Armées au sein de la CyberDefense Factory : nous avons contribué à la détection d'intrusions en tirant parti de la caractéristique dynamique de données de réseau informatique.

Ce travail est détaillé en section 1.6 de l'introduction et dans le chapitre 4 du manuscrit.

## 1.4 L'algorithme K-bMOM

Dans cette partie, nous détaillons le contexte scientifique du clustering robuste ainsi que notre apport dans le domaine, qui a débouché sur un article publié au journal *Computational Statistics and Data Analysis* [26, 27].

### 1.4.1 Contexte scientifique

Les jeux de données volumineux ou complexes sont souvent pollués par des outliers, c'est-à-dire des données aberrantes. Les procedures de data mining classiques telles que K-means ou des algorithmes EM plus généraux sont cependant sensibles à la présence de tels outliers. Ecarter les outliers des données est une tâche délicate et coûteuse en temps. Dans ce contexte, des versions robustes des procedures de data mining sont particulièrement pertinentes et nous avons cherché à produire un algorithme de hard clustering, de type Lloyd, robuste à la présence d'outliers grâce à l'utilisation de Median-of-Means.

La stratégie Median-of-Means (MOM) a fait l'objet de recherches très actives ces dernières années dans les domaines de l'estimation de la moyenne, de la régression, des statistiques en grande dimension et en classification supervisée [90, 41, 84, 85, 97, 96, 98, 104]. D'autres approches de robustesse pour les K-means ont été proposées, telles que par

exemples les K-medians ou le trimming [54, 22]. La première méthode consiste à optimiser le critère des K-means, où les distances au carré sont remplacées par les distances elles-mêmes. La seconde consiste à optimiser le même critère que les K-means dans lequel une fraction données des points les plus "aberrants" est omise. Notons que très récemment, les auteurs de [81] ont produit une analyse théorique d'un estimateur mêlant K-means et Median-of-Means pour le clustering données à queues lourdes, exhibant encore des propriétés d'optimalité des bornes de risque obtenues. La conception d'estimateurs robustes avec un contrôle de la complexité algorithmique est également un thème très actif de recherche (voir l'article de survol [42]).

### 1.4.2 Apport

Nous proposons un algorithme de clustering robuste, récursif de type Lloyd, inspiré du K-means, nommé K-bMOM. Pour cela, nous avons utilisé une version bootstrap de MOM (bMOM). Lorsque l'on dispose d'un échantillon de $n$ variables aléatoires réelles indépendantes et identiquement distribuées, le MOM estime l'espérance de ces variables selon la routine suivante : on constitue d'abord $B$ sous-ensembles (appelés blocs) disjoints et de cardinal égaux – à plus ou moins un près – à partir de ces $n$ données, puis on calcule la moyenne dans chacun des blocs pour n'en retenir que leur valeur médiane. Sa définition mathématique précise est la suivante

$$\mathrm{MOM}(u_1^n, I_1^B) = \mathrm{med}\left\{ \frac{1}{|I_b|} \sum_{i \in I_b} u_i : b \in \{1, ..., B\} \right\},$$

où $I_1^B := \{I_b : b \in \{1, ..., B\}\}$ est l'ensemble des blocs constitués, $|I_b|$ leur cardinal. Puisque $\bigcup_{k=1}^n \{u_k\} = \bigcup_{i=1}^B I_i$, on a que plus $B$ est grand, plus $|I_b|$ est petit. Or dans certains contextes, un bloc trop petit peut être rédhibitoire. C'est notamment le cas en clustering où il est préférable que chacun des blocs dispose de tous les clusters, et pour cela, dispose d'un nombre suffisant de données. Nous avons ainsi introduit bMOM, dont la différence avec MOM est une génération aléatoire uniforme avec remise des blocs. Formellement, il s'agit de créer un échantillon bootstrap $v_1^q$ de taille $q = B \times n_B$, où $n_B$ est la taille des blocs et bMOM peut être exprimé comme un cas particulier de MOM:

$$\mathrm{bMOM}(u_1^n, n_B, B) = \mathrm{MOM}(v_1^q, I_1^B)\,,$$

où $I_1^B$ est fixe et telle que $I_b = \{(b-1)n_B + 1, ..., bn_B\}$.

bMOM présente aussi l'avantage d'avoir un meilleur breakdown point que MOM. Le breakdown point d'une procédure est la proportion d'outliers que la procédure peut tolérer de sorte que son résultat reste à une distance finie de sa valeur s'il n'y avait pas de pollution. Puisque bMOM et MOM prennent des réels en argument, on dira simplement que c'est le nombre maximal d'outliers possibles tels que leur résultat reste borné. On a montré que MOM a un breakdown point de $\lfloor (B-1)/2 \rfloor / n \approx 0.5/n_B$ tandis que bMOM en a un valant $\left\lfloor n\left(1-1/2^{1/n_B}\right)\right\rfloor / n \approx 0.69/n_B$, voir Section 2.3.1.

L'algorithme K-bMOM que nous proposons consiste en une stratégie de type bMOM en générant avec remise des blocs. A chaque étape de l'algorithme, on calcule dans chaque bloc les centres des clusters, puis on réassigne les données des blocs aux nouveaux centres. On sélectionne les centres du bloc qui réalise la distorsion K-means médiane, puis on continue ainsi jusqu'à convergence des centres ou jusqu'au nombre maximum d'itérations. On peut lire le pseudo-code de K-bMOM dans algorithme 1.1. Pour plus de détails, notamment sur les notations, on se reportera à la section 2.2. De manière générale, on peut observer que K-bMOM est une stratégie qui a de meilleurs résultats que les autres algorithmes quelque soit l'initialisation choisie, voir section 2.8.4.

**Algorithm 1.1** Procédure K-bMOM

A partir d'un échantillon $\{y_1, \ldots, y_n\}$, et d'une partition initiale $\mathbf{a} \in \{1, \ldots, K\}^n$ en $K$ clusters :

1. Créer $B$ blocs de taille $n_B$ à partir d'un échantillon bootstrap et en ne conservant que les blocs contenant au moins un point de chaque cluster.
2. Dans chaque bloc $b$, calculer le vecteur moyen $c_k^{(b)}$ du cluster $k$ ainsi que sa distorsion empirique $\mathcal{R}^{(b)} = \frac{1}{n_B} \sum_{k=1}^{K} \sum_{l=1}^{n_B} \left\| y_l^{(b)} - c_k^{(b)} \right\|^2 \mathbf{1}\{y_l^{(b)} \in \mathcal{C}_k^{(b)}\}$.
3. Identifier le bloc $B_{med}$ qui réalise la valeur médiane des risques empiriques.
4. Extraire les vecteurs moyens de chacun des clusters du bloc $B_{med}$ .
5. Mettre à jour $\mathbf{a}$ de sorte que chaque point de l'échantillon $\{y_1, \ldots, y_n\}$ soit affecté aux centroïde $c_k^{(B_{med})}$ le plus proche.
6. Itérer les points 1 à 5 jusqu'à convergence de la moyennes des centres sur les 10 dernières itérations ou jusqu'au nombre maximal d'itérations.

Retourner le vecteur de classification $\mathbf{a}$ et la moyenne des centres sur les 10 dernières itérations.

Nous proposons également une variante robuste de l'initialisation traditionnelle K-means++ en appliquant la stratégie bMOM. Il s'agit à l'instar de K-bMOM de constituer $B$ blocs de taille $n_B$ selon une loi uniforme et avec remise, puis dans chaque bloc d'appliquer l'initialisation K-means++ pour obtenir $K$ centroïdes. Les centroïdes

**Algorithm 1.2** Stratégie K-bMOM-km++

A partir d'un échantillon $\{y_1, \ldots, y_n\}$, d'un nombre de clusters fixé $K$, d'un nombre de blocs $B$, et d'une taille de bloc $n_B$.

1. Créer $B$ blocs aléatoirement selon une loi uniforme avec remise à partir de l'échantillon $\{y_1, \ldots, y_n\}$ et indépendamment d'un bloc à l'autre.
2. Dans chaque bloc $b \in \{1, \ldots, B\}$, appliquer l'initialisation $K$-means++ pour obtenir des centroïdes $(c_{1,++}^{(b)}, ..., c_{K,++}^{(b)})$.
3. Dans chaque bloc, calculer la distorsion $K$-means $R_{++}^{(b)} = \frac{1}{n_B} \sum_{k=1}^{K} \sum_{l=1}^{n_B} \left\| y_l^{(b)} - c_{k,++}^{(b)} \right\|^2 \mathbf{1}\{y_l^{(b)} \in \mathcal{C}_k^{(b)}\}$.
4. Identifier le bloc $B_{med}$ qui réalise la valeur médiane des risques empiriques.

Retourner $\left(\widehat{c}_{1,++}^{(B_{med})}, \ldots, \widehat{c}_{K,++}^{(B_{med})}\right)$, les centroïdes du bloc $B_{med}$.

conservés seront ceux appartenant au bloc qui réalise la valeur médiane de la distorsion K-means.

Cette stratégie a été comparée à d'autres initialisations de la littérature : l'initialisation aléatoire uniforme, K-means++, K-medians++ et ROBIN. Notre approche s'avère être généralement la meilleure pour la classification d'après le critère "Accuracy" (somme normalisée de vrais positifs et de vrais négatifs) et également pour l'estimer des centroïdes d'après le critère "RMSE" (Root Mean Square Error). On peut voir que notre approche a sensiblement les mêmes performances que les meilleures autres méthodes dans le cas sans contamination, mais que ses performances surpassent souvent celles des autres méthodes en présence d'outliers.

L'efficacité de K-bMOM a également été testée dans le cas concret de la quantification de couleurs[10] sur des images en nuances de gris[11] présentant 1% d'outliers. Cette expérience a montré que K-bMOM n'utilisait pas les outliers comme centroïdes et que sa distorsion finale était jusqu'à 32% meilleure que le K-means.

En plus de ces aspects pratiques, nous avons apporté des compléments d'analyse théorique tant sur le breakdown point que sur les vitesses d'apprentissage en présence d'outliers. L'étude est détaillée au chapitre 2.

[10] la terminologie anglophone est "Color quantization". Le mot quantization peut être considéré comme un terme dédié.

[11] L'indexation RGB des couleurs se fait par des triplets $(r, g, b) \in \{0, ..., 255\}^3$. Les nuances de gris sont le plus souvent indexées par les valeurs "diagonales" de la forme $(p, p, p)$, $p \in \{0, ..., 255\}$

## 1.5 Clustering par entropie logistique pénalisée

Dans cette partie, nous détaillons le contexte scientifique du clustering en grande dimension ainsi que notre apport dans cette thématique, qui a fait l'objet d'une prépublication [56].

### 1.5.1 Contexte scientifique

La tâche de clustering peut être décrite de la manière suivante : il s'agit de définir une fonction de classification (éventuellement aléatoire) $Y : \mathcal{X} \to \{1, ..., K\}$ à partir d'un espace mesurable $\mathcal{X}$, d'un échantillon $X_1, ..., X_n$ et d'un nombre entier $K \geq 2$. En particulier, on associe une classe (groupe) $Y_i$ à chaque donnée $X_i$. Si la fonction $Y$ est déterministe, alors le clustering est dit "dur". Si en revanche $Y$ est aléatoire, telle que la distribution des classes $Y(X)$ est caractérisée par le $K$-uplet $(\mathbb{P}(Y(X) = 1), ..., \mathbb{P}(Y(X) = K))$, alors le clustering est dit "probabiliste". Dans ce second cas, l'approche par modèle est commune ; elle consiste à modéliser la distribution des données, comme par exemple avec une distribution de mélange[12], et à mettre en relation directe les probabilités $(\mathbb{P}(Y(X) = 1), ..., \mathbb{P}(Y(X) = K))$ avec les paramètres du modèle [20]. On peut ensuite déduire de ce clustering probabiliste un clustering dur en affectant chaque donnée $X$ à la classe $k$ de probabilité $\mathbb{P}(Y(X) = k)$ maximale (ou en l'affectant au hasard uniformément parmi celles de probabilité maximale). Les algorithmes de clustering dur incluent notamment le $K$-means [93, 135, 99], le clustering hiérarchique [73], le clustering spectral [112] et bien d'autres [57].

Particulièrement développée dans la communauté de l'apprentissage automatique pour sa flexibilité d'emploi en présence de données complexes, "l'approche discriminative" vise à modéliser les probabilités de classification $(\mathbb{P}(Y(X) = 1), ..., \mathbb{P}(Y(X) = K))$, assimilables à des probabilités conditionnelles des classes sachant la position de $X$. Procéder de la sorte évite de modéliser la distribution complète des données et revient souvent à exprimer les probabilités de classification en fonction de frontières de séparation entre clusters. Ceci est généralement fait en utilisant des critères d'apprentissage classiques comme la perte logistique, la perte de Hinge ou encore la perte des champs aléatoire conditionnels [38, 60]. Plus formellement, on cherche des probabilités $\mathbb{P}(Y(X) = k)$, $k \in \{1, ..., K\}$ proportionnelles à $\exp(\ell(\beta_k, X))$ pour un vecteur $\beta_k$ et une perte $\ell$ donnés. Par exemple, la perte logistique donne des probabilités de classification proportionnelles à $\exp(w_k^t X + b_k)$, et celle de Hinge (pour deux classes, $K = 2$)

[12] une distribution de mélange est une description de la distribution globale $\mathcal{D}$ comme la somme des distributions des classes $\mathcal{D}_k$ : $\mathcal{D}(x) = \sum_{k=1}^{K} \alpha_k \mathcal{D}_k(x)$ où $\alpha_k = \mathbb{P}(Y(x) = k)$.

induit des probabilités proportionnelles à $\exp(-[1-(w_k^t\varphi(X)+b_k)]_+)$ où $\varphi$ est une fonction de transformation des co-variables et où $(w_1,b_1)=(-w_2,-b_2)$ dans ce cas à deux classes.

Ces pertes ont d'abord été introduites pour la classification supervisée et, pour pouvoir étendre leur usage au cadre de classification non-supervisée, il faut définir ce qui serait une bonne classification. Sur ce point, on peut argumenter qu'il est préférable d'obtenir des probabilités de classification les moins ambigües possible. Ceci correspond au cas où le maximum des probabilités de classification est le plus proche possible de 1. C'est pourquoi un critère naturel à considérer pour inférer une fonction de classification $\tilde{Y}$ serait d'utiliser les probabilités $\mathbb{P}(\tilde{Y}(X)=k)=Z_{\tilde{\beta}}^{-1}(X)\exp(\ell(\tilde{\beta}_k,X))$ avec une constante de normalisation $Z_{\tilde{\beta}}(X)=\sum_{k=1}^K\exp(\ell(\tilde{\beta}_k,X))$, telles que

$$(\tilde{\beta}_1,...,\tilde{\beta}_K)\in\arg\max_{(\beta_1,...,\beta_K)}\left\{\frac{1}{n}\sum_{i=1}^n\frac{1}{Z_\beta(X_i)}\max_{k\in\{1,...,K\}}[\exp(\ell(\beta_k,X_i))]\right\}. \tag{1.5.1}$$

La cible théorique à estimer est $\mathbb{P}(Y_*(X)=k)=\exp(\ell(\beta_{*,k},X))$ avec

$$(\beta_{*,1},...,\beta_{*,K})\in\arg\max_{(\beta_1,...,\beta_K)}\left\{\mathbb{E}\left[\frac{1}{Z_\beta(X)}\max_{k\in\{1,...,K\}}[\exp(\ell(\beta_k,X))]\right]\right\},$$

où $X$ suit sa distribution, inconnue et non modélisée.

Mais du fait que l'opérateur de "maximum" n'est pas une fonction lisse, il peut causer des difficultés d'optimisation de (1.5.1). L'entropie des probabilités de classification est une quantité lisse de substitution éligible puisqu'elle atteint son minimum pour des probabilités dégénérées (0 ou 1). Il est donc question de chercher une fonction de classification $\hat{Y}$ satisfaisant $\mathbb{P}(\hat{Y}(X)=k)=Z_{\hat{\beta}}^{-1}(X)\exp(\ell(\hat{\beta}_k,X))$ avec

$$(\hat{\beta}_1,...,\hat{\beta}_K)\in\arg\min_{(\beta_1,...,\beta_K)}\left\{\frac{1}{n}\sum_{i=1}^n\text{Ent}\left\{\mathbb{P}(Y_\beta(X_i)=1),...,\mathbb{P}(Y_\beta(X_i)=K)\right\}\right\}, \tag{1.5.2}$$

où

$$\text{Ent}\left\{\mathbb{P}(Y_\beta(X_i)=1),...,\mathbb{P}(Y_\beta(X_i)=K)\right\}=\sum_{k=1}^K-\frac{\exp(\ell(\beta_k,X_i))}{Z_\beta(X_i)}\log\left(\frac{\exp(\ell(\beta_k,X_i))}{Z_\beta(X_i)}\right). \tag{1.5.3}$$

Souvent, il est nécessaire, soit de restreindre la recherche du vecteur $(\beta_1,...,\beta_K)$ à un compact, soit d'ajouter à l'entropie un terme de régularisation traduisant la complexité de $(\beta_1,...,\beta_K)$ [60, 38]. Dans cette seconde formula-

tion, le vecteur cible théorique $(\beta_{0,1}, ..., \beta_{0,K})$ estimé est,

$$(\beta_{0,1}, ..., \beta_{0,K}) \in \arg \min_{(\beta_1,...,\beta_K)} \{\mathbb{E}\left[\text{Ent}\left\{\mathbb{P}(Y_\beta(X_i) = 1), ..., \mathbb{P}(Y_\beta(X_i) = K)\right\}\right]\}. \tag{1.5.4}$$

L'utilisation de termes d'entropie en apprentissage supervisé et non-supervisé est naturelle et a fait l'objet de recherches actives [63, 60, 38, 138, 137, 133, 91, 3, 106]. Ajoutons que cette approche est aussi au coeur de l'état de l'art de certaines approches de deep clustering [71]. Une autre approche fructueuse en clustering discriminative consiste à étudier une relaxation convexe de critère intractable, une méthodologie que s'est accompagnée de très bonnes garanties théoriques [51, 77, 10, 120, 57, 31, 30, 58, 105, 128, 34].

### 1.5.2 Apport

Notre travail part du constat suivant : à notre connaissance, il n'existe dans la littérature aucune garantie théorique, formulée en termes de vitesse de convergence, pour l'estimateur (régularisé) du minimum d'entropie (1.5.2). C'est une lacune par rapport à d'autres techniques comme la relaxation convexe. Alors que d'un point de vue pratique, les estimateurs de la forme (1.5.2) ont déjà prouvé leur efficacité et leur flexibilité, notamment dans des techniques de plongement de co-variables et de clustering dit "profond" (deep clustering). L'objet de notre étude est donc de fournir une première analyse théorique de ces estimateurs.

Nous considérons dans ce travail la classification non-supervisée d'un mélange de deux gaussiennes en grande dimension, de matrice de covariance égale à l'identité et séparées en moyenne par un vecteur parcimonieux. Cela se formalisé par une variable $X \sim \varepsilon\mathcal{N}(a, I_d)$ où $\varepsilon \sim \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ est une variable aléatoire de Rademacher, $\mathcal{N}(a, I_d)$ est une loi normale de vecteur moyen $a \in \mathbb{R}^d$. Ce cadre est un contexte intéressant puisqu'il est suffisamment simple pour nous permettre de comprendre la nature du minimum de l'entropie (1.5.2) $\beta_0 = (\beta_{0,1}, ..., \beta_{0,K})$, avec $K = 2$ et $\beta_{0,1} = -\beta_{0,2}$. Nous avons pris la perte logistique $\ell(\beta, X) = -\log\left(1 + e^{X^t\beta}\right)$, mais dans ce cas le critère (1.5.4) atteint son minimum en l'infini, ce qui n'est pas informatif. Pour résoudre ce problème, nous avons restreint la recherche du minimum aux vecteurs d'une demi-boule $\Psi_U := \{\beta \in B_2(0, R) : \beta^t U > 0\}$ centrée en 0, de rayon $R > 0$ et pointant dans le même demi-espace que $U \in \mathbb{R}^d$ tel que $\|U\|_2 = 1$ et choisi aléatoirement uniformément sur la

boule unité. On note

$$\begin{aligned}\mathcal{R}(\beta) &= \mathbb{E}\left[\mathrm{Ent}\left\{\mathbb{P}(Y_\beta(X)=1), \mathbb{P}(Y_\beta(X)=2)\right\}\right] \\ &= \mathbb{E}\left[-\frac{e^{X^t\beta}}{1+e^{X^t\beta}}\log\left(\frac{e^{X^t\beta}}{1+e^{X^t\beta}}\right) - \frac{1}{1+e^{X^t\beta}}\log\left(\frac{1}{1+e^{X^t\beta}}\right)\right]\end{aligned} \tag{1.5.5}$$

le risque entropique avec la perte logistique. Nous retrouvons ainsi le vecteur normal de l'hyperplan frontière entre les deux classes par minimisation du critère :

*Théorème* 1. *Il existe un unique minimiseur* $\beta_0$ *de* $\mathcal{R}(\beta)$ *où* $\beta_0 = R\frac{a}{\|a\|_2}$ *si* $a \in \Psi_U$ *et* $\beta_0 = -R\frac{a}{\|a\|_2}$ *sinon.*

Le critère d'entropie n'est pas convexe mais nous montrons dans notre deuxième résultat principal qu'il est localement convexe autour de $\beta_0$ sous de bonnes conditions :

*Théorème* 2. On définit $x_1 \approx 1.543$ comme l'unique solution positive de l'équation $1 + e^x - xe^x + x = 0$ et on pose $R := \sqrt{x_1 + 0.08}$. Si $\|a\|_2 \geq 2R$ alors $\beta \mapsto \mathcal{R}(\beta)$ est localement convexe en $\beta_0$.

Ce mélange de deux gaussiennes en grande dimension est également un contexte assez simple pour pouvoir calculer la vitesse de convergence de l'estimateur correctement renormalisé par une pénalité $\ell_1$ :

$$\widehat{\beta} \in \arg\min_{\beta\in\Psi_U}\left\{\frac{1}{n}\sum_{i=1}^{n}\rho\left(X^t\beta\right) + \lambda\|\beta\|_1\right\} \tag{1.5.6}$$

où $\forall u,\ \rho(u) := -\frac{e^u}{1+e^u}\log\left(\frac{e^u}{1+e^u}\right) - \frac{1}{1+e^u}\log\left(\frac{1}{1+e^u}\right)$ est l'entropie de l'équation (1.5.3) pour des probabilités obtenues avec la perte logistique. Notre troisième résultat exploite la convexité locale du critère pour établir l'égalité oracle suivante :

*Théorème* 3. *Si le vecteur moyen* $a \in \mathbb{R}^d$ *possède au maximum* $s$ *composantes non nulles alors il existe deux constantes* $c_{\|a\|_2,R}, \lambda_0 > 0$ *telle que* $\forall n \geq 2$, $\forall \lambda > 0$, *l'inéquation*

$$\mathcal{E}\left(\tilde{\beta}, \beta_0\right) + 4\left(\lambda - 2T\lambda_0\right)\left\|\tilde{\beta}^{S^c}\right\|_1 \leq s.c_{\|a\|_2,R}.\left(T\lambda_0 + \lambda\right)^2 \tag{1.5.7}$$

*est vraie avec une probabilité au moins égale à*

$$1 - c_1 \exp\left(-12\left(T-1\right)^2 \log\left(2d\right) \log^2 n\right) - \frac{c_2}{T^2 \log\left(2d\right).n \log^2 n},$$

*où* $c_1, c_2 > 0$. *Des valeurs admissibles des constantes sont fournies dans le théorème 1.5.4, notamment une valeur exacte de* $\lambda_0$. *On retiendra ici simplement que* $\lambda_0$ *est asymptotiquement proportionnelle à* $\|a\|_2 \left(\sqrt{\log\left(d\right)} \log n\right) n^{-1/2}$

Enfin, le contexte particulier d'un mélange de deux gaussiennes en grande dimension est intéressant parce qu'il a récemment reçu beaucoup d'attention [19, 8, 108, 92, 75, 47, 32, 9, 76, 24, 94], offrant de multiples comparaisons avec notre travail. Notez bien que le but de ce travail n'est pas de fournir estimateur aux performances comparables à l'état de l'art dans le cas de la classification de deux gaussienne en grande dimension, mais bien d'explorer pour la première fois le comportement théorique d'un estimateur discriminatif qui minimise le critère (régularisé) d'entropie de classification et comment cela s'adapte à un cas parcimonieux.

## 1.6 Détection d'intrusion par détection de communauté

Dans cette partie, nous détaillons le contexte scientifique de la détection de communauté appliquée à la cybersécurité, ainsi que notre apport, qui a débouché sur un exposé lors de la conférence CAID[13], dans le cadre de la *European Cyber Week*, et la rédaction d'un acte de conférence à paraître [102].

### 1.6.1 contexte scientifique

Le clustering sur graphe cherche à regrouper des noeuds dans un graphe sur la base des liens qui existent entre eux. Cette tâche porte aussi le nom de détection de communautés, à l'image des communautés dans les groupes sociaux. L'approche de certains travaux en détection d'intrusion est d'exploiter la structure de communauté au sein du graphe pour détecter une anomalie et lever une alerte à l'attention d'un analyste spécialisé. Nous pouvons distinguer trois approches pour y parvenir : la première consiste à édicter des règles de décision [103], la deuxième consiste à entrainer un modèle à partir d'un corpus d'attaques labellisé [88], et la troisième consiste à définir comment construire un graphe afin de mieux mettre en évidence des noeuds relatifs à une attaque (à l'aide d'un

[13] https://www.european-cyber-week.eu/conference-caid

algorithme de détection de communautés [119] par exemple). Notre approche entre dans la troisième catégorie.

### 1.6.2 Algorithme GrphClus

Sous l'impulsion des travaux de Sébastien Loustau, LumenAI a implémenté un version de l'algorithme présenté dans [39]. Nous ferons référence à cette implémentation par l'appellation GrphClus. GrphCLus est un algorithme de détection de communauté en temps réel basé sur la maximisation de la modularité. LumenAI adapte en ce moment sa technologie à la cybersécurité. En effet, les données de flux réseau ou les logs systèmes sont la plupart du temps des données relationelles et se modélisent facilement par des graphes. Deuxièmement, puisque GrphClus est un algorithme en temps réel, on peut imaginer qu'il analyse en continu des données issues d'un réseau ou d'un OS[14], tel un anti-virus.

#### 1.6.2.1 La modularité

La modularité a été introduite la première fois par Newman [111] en 2004. C'est un nombre entre $-1/2$ et 1 que l'on peut attribuer à toute partition d'un graphe. Plus ce nombre est élevé plus la partition est considérée comme satisfaisante. L'une des techniques pour trouver une partition pertinente des noeuds d'un graphe consiste donc à chercher à maximiser la modularité. Cette heuristique a été confirmée en 2016 par Newman [110] en montrant que la maximisation de la modularité est équivalente à la maximisation de vraisemblance dans le cas particulier des modèles à blocs stochastiques avec degrés corrigés[15] [78]. Ses travaux donnent donc un sens précis aux partitions obtenues par maximisation de la modularité : Cela revient à chercher quels seraient les groupes les plus pertinents dans le graphes s'ils avaient été générés par un modèles à blocs stochastiques avec degrés corrigés. Par conséquent, si le graphe suit des caractéristiques très différentes de ce modèle, il est toujours possible de déterminer de tels groupes mais leur pertinence sera discutable.

La modularité s'exprime dans un graphe non-orienté $G = (V, E, A)$ défini à partir de l'ensemble de ses $n$ noeuds $V$, de l'ensemble de ses arêtes $E \subset V \times V$ et d'une matrice d'adjacence $A \in M_n(\mathbb{R})$ qui est la matrice dont les entrées $A_{ij}$, toutes positives ou nulles, indiquent le poids associé à la liaison entre $i$ et $j$ et où $A_{ii}$ indique le double du poids de la liaision du noeud $i$ avec lui-même. On pourra aussi être amené à décrire le poids d'une arête $e = (i, j)$ par la

[14] "Operating Système". Window, iOS, Android et Linux sont des exemples d'OS.
[15] Degree-corrected stochastic block model.

notation $A_e := A_{ij}$. De plus, le degré d'un noeud $i$ dans $G$, noté $d_i = \sum_{j \in V} A_{ij}$, est la somme des poids des arêtes partant de $i$. Enfin, pour une partition $C$ des noeuds d'un graphe $G$, donnée par $K$ clusters[16] $C = \{c_1, \ldots, c_K\}$ tous disjoints deux à deux, on peut écrire la modularité du graphe $G$ partitionné selon $C$ de la manière suivante,

$$Q(C) = \frac{1}{2m} \sum_{c \in C} \sum_{i,j \in c} \left( A_{ij} - \gamma \frac{d_i d_j}{2m} \right),$$

où $m := |E| = \frac{1}{2} \sum_i d_i$ et où $\gamma$ est l'hyperparamètre de résolution ($\gamma = 1$ dans GrphClus). Plus $\gamma$ est grand plus la modularité pénalise les partitions avec de grands clusters.

#### 1.6.2.2 Partition d'un graphe par maximisation de la modularité

Depuis l'introduction de la modularité en 2004, un nombre important de procédures de recherche de la partition optimale (celle qui maximise la modularité) ont été proposées [17, 69, 43, 140, 39, 35]. La méthode de Clauset et al [35] consiste à initialiser tous les noeuds dans leur propre communauté et de raffiner cette partition de proche en proche en faisant entrer les noeuds dans les communautés voisines s'ils s'accompagnent d'un gain maximal en modularité. L'algorithme de Louvain [17] est un algorithme hiérarchique basé sur le même principe et Leiden [140] améliore Louvain en raffinant la partition à chaque itération sur des considérations de connectivité des noeuds dans leur communauté.

GrphClus maximise également la modularité en même temps qu'il met à jour le graphe arête par arête au fur et à mesure qu'elles lui sont envoyées. En parallèle de la réception des arêtes, GrphClus optimise au cours du temps, avec une méthode MCMC[17] la partition[18] $C$ du graphe. L'algorithme modifie cette partition de la façon suivante : L'algorithme tire au hasard une partition candidate $C'$ pour succéder à $C$ selon une loi de mélange $\alpha_1 p_1(\cdot \,|\, C) + \alpha_2 p_2(\cdot \,|\, C)$ où $\alpha_1$ et $\alpha_2$ sont des hyperparamètres. Les lois $p_1(\cdot \,|\, C)$ et $p_2(\cdot \,|\, C)$ sont telles que :

- sous $p_1(\cdot \,|\, C)$ : tirer aléatoirement uniformément un noeud $i$ dans l'ensemble des noeuds $V$ et une communauté $\boldsymbol{c}$ dans $C$. $C'$ est alors choisie égale à $C$ pour tous les noeuds sauf $i$ qui rejoint $\boldsymbol{c}$ s'il n'en fait pas déjà partie, sinon $i$ est placée dans une nouvelle communauté dont il est l'unique représentant.

[16] Il est à noter que contrairement aux $K$-means, $K$ n'est pas un hyperparamètre de la modularité.
[17] Markov Chain Monte Carlo.
[18] Dans les faits GrphClus calcule une partition hiérarchique mais dans ce travail nous n'avons utilisé que le premier niveau de la partition, à savoir : les communautés du graphe qu'on note $C$.

- sous $p_2(\cdot|C)$ : tirer aléatoirement uniformément un noeud $i$, puis tirer un noeud $j$ aléatoirement parmi les voisins de $i$ selon un vecteur de probabilités proportionnelles à $A_{ij}$. $C'$ est alors choisie égale à $C$ pour tous les noeuds sauf $i$ qui rejoint $\boldsymbol{c}$ s'il n'en fait pas déjà partie, sinon $i$ est placée dans une nouvelle communauté dont il est l'unique représentant.

Ensuite, cette partition candidate $C'$ est adoptée (ou rejetée) avec une probabilité $\rho$ (ou $1-\rho$). Le taux d'acceptation $\rho = \min\{1, rf(\lambda)\}$ est calculé à partir du rapport de vraisemblance[19] de $C$ par rapport à $C'$ noté $r = \frac{p(C|C')}{p(C'|C)}$ et à partir d'un facteur de qualité $f(\lambda) = e^{\lambda(Q'-Q)}$ issu du gain en modularité par le remplacement de $C$ par $C'$. $\lambda > 0$ est un hyperparamètre du modèle qui influence la stabilité et la vitesse de convergence de la partie MCMC de l'algorithme 1.3. La procédure décrite se généralise à des partitions hiérarchiques selon le même principe, moyennant une plus grande complexité dans les notations et une gestion de l'évolution de tous les niveaux impactés par l'acceptation d'une nouvelle partition (voir [39] pour plus de détails). Comme expliqué dans la section 3 de [39], l'hypothèse $\alpha_1 > 0$ est théoriquement importante car elle rend la chaîne de markov réversible et la réversibilité implique l'existence d'une loi stationnaire de la chaine de Markov d'après le théorème de Perron-Frobenius. L'existence d'une telle loi justifie alors l'utilisation d'un algorithme de type Metropolis-Hasting pour l'approcher et justifie également le choix de l'expression du taux d'acceptation $\rho$ car il a la propriété de conserver la loi stationnaire de la chaine de markov.

### 1.6.3 Apport

Partant de la conjecture qu'une attaque d'une machine informatique va déstabiliser les communautés en place et va ainsi trahir sa présence, GrphClus est un algorithme qui s'inscrit dans la stratégie de détection d'instrusion à partir du clustering d'un graphe. Il est possible que GrphClus permette de mettre en évidence une attaque comme dans [119] donc nous avons voulu tester cette hypothèse.

Nous avons tout d'abord recherché un jeu de données adapté. Nos recherches nous ont mené au jeu de données OpTC[20], mis à disposition publiquement par la DARPA[21]. Ce jeu de données dispose des logs systèmes de 1000 machines pendant 6 jours. En particulier, les machines 201, 402 et 660 pour le jours 1, et, 51 et 351 pour le jour 3,

[19] Ce rapport de vraisemblance mesure la réversibilité du passage de $C$ à $C'$ pour la chaîne de Markov sur graphe considérée. Voir article [39] pour avoir son expression littérale.
[20] Operational transparent computing
[21] https://github.com/FiveDirections/OpTC-data

**Algorithm 1.3** Algorithme simplifié de GrphClus, version pour une partition à 1 niveau

Hyperparamètres : $\lambda > 0$ (paramètre dit de "température"), $\alpha_1 > 0$ et $\alpha_2 > 0$ tels que $\alpha_1 + \alpha_2 = 1$ et la modularité est calculée avec paramètre de résolution $\gamma$ égal à 1.
Algorithme :A partir d'un graphe vierge $G = \{V = \emptyset, E = \emptyset, A = \emptyset\}$. Faire en parallèle 1 et 2 :

1. (Ecoute passive) maintien du graphe :
   (a) lorsqu'une arête $e = (i, j)$ est reçue, ajouter les noeuds $i$ ou $j$ à $V$ et $e$ à $E$ si ces éléments sont inédits et incrémenter $A_{ij}$ de 1.
2. (MCMC) faire évoluer la partition en boucle :
   (a) On tire une partition candidate $C'$ suivant une loi de mélange $\alpha_1 p_1(\cdot|C) + \alpha_2 p_2(\cdot|C)$ :
      i. sous $p_1(\cdot|C)$ : tirer aléatoirement uniformément un noeud $i$ dans l'ensemble des noeuds $V$ et une communauté $\boldsymbol{c}$ dans C.
      ii. sous $p_2(\cdot|C)$ : tirer aléatoirement uniformément un noeud $i$, puis tirer un noeud $j$ aléatoirement parmi les voisins de $i$ selon un vecteur de probabilité proportionnelle à $A_{ij}$. La communauté de $j$ est notée $c$.
      iii. $C'$ est alors choisie égale à $C$ pour tous les noeuds sauf $i$ qui rejoint $\boldsymbol{c}$ s'il n'en fait pas déjà partie, sinon, $i$ est placée dans une nouvelle communauté dont il est l'unique représentant.
   (b) Election de $C'$. Calculer:
      i. la modularité $Q'$ associée à $C'$.
      ii. le facteur de qualité $f(\lambda) = e^{\lambda(Q'-Q)}$, où $Q$ est la modularité pour la partition actuelle $C$.
      iii. et $r = \frac{p(C|C')}{p(C'|C)}$ le rapport des vraisemblance de $C'$ par rapport à $C$ (cf [39] pour la valeur de $r$).
      iv. Accepter $C'$ avec une probabilité $\rho = \min\{1, r.f(\lambda)\}$.

ont été attaquées par l'équipe "red team" de la DARPA selon différents scénarios. Nous nous sommes concentrés sur les jours 1 et 3 du fait de l'intérêt que LumenAI portait aux types d'attaques engagées. La documentation[22] du jeu de données fournit davantage de détails sur les machines attaquées et sur les intentions des attaquants. Nous avons donc mis à l'épreuve notre conjecture en utilisant GrphClus sur les graphes des processus de chacune de ces machines. On peut voir en figure 1.6.2 le type de graphe partitionné par GrphClus. L'évolution en temps réel de ce graphe est disponible dans une vidéo annexe à cette adresse[23].

Les cyber-analystes sont à la recherche de métriques efficaces et facilement interpétables. Nous avons donc monitoré le nombre de communautés au cours du temps pour chacune des machines. Nous avons obtenu un résultat corroborrant cette hypothèse : l'attaque de la machine 201 cause une rupture de tendance dans le nombre de communautés

[22] https://github.com/FiveDirections/OpTC-data/blob/master/OpTCRedTeamGroundTruth.pdf .
[23] https://www.youtube.com/watch?v=LPBuE0kBIr4 .

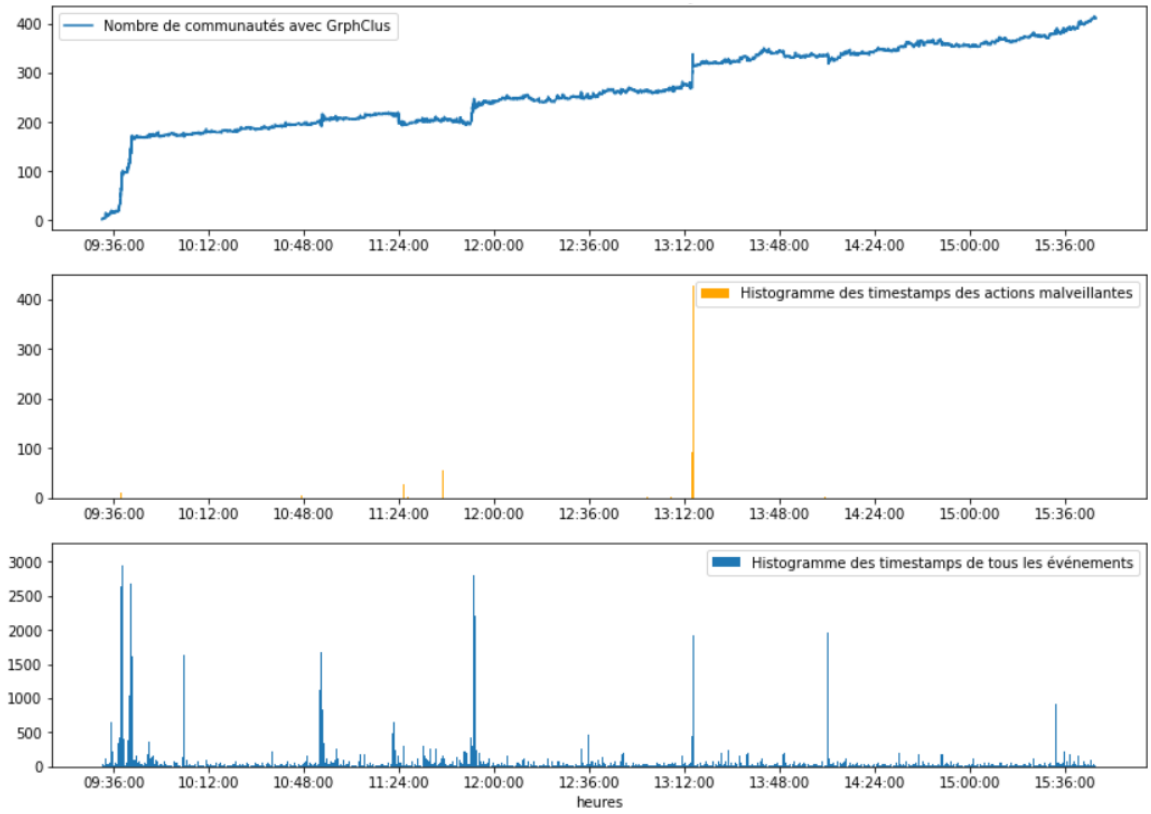



Figure 1.6.1: Coévolution du nombre de communautés et de la densité d'événements malveillants sur la machine 201 le jour 1

trahisant les événements malveillants, voir figure 1.6.1. Une analyse approfondie semble montrer que cette approche peut fonctionner pour les attaques qui modifient suffisamment la structure du graphe dans un intervalle de temps très court. Ce résultat n'a pas pu être répliqué sur les machines 402 et 660, attaquées le jour 1 également, ni sur les machines 51 et 351 attaquées le jour 3. La raison envisagée est que notre méthodologie ne fonctionne que dans les cas où l'attaque crée de nouveaux noeuds dans le graphe.

La méthodologie semble être une piste favorable pour certaines attaques mais elle doit être associée à d'autres outils. En effet, les ruptures de pente dans le nombre de communautés proposées par GrphClus seules sont un indicateur insuffisant car il aboutirait à trop de faux positifs en comparaison de la capacité d'analyse des cyber-analystes.

Pour savoir si cette méthodologie dépend de l'algorithme utilisé, nous avons également comparé les résultats de GrphClus avec d'autres algorithmes de maximisation de la modularité comme CNM, Louvain et Leiden en termes

de modularité obtenue, de comportement de la courbe du nombre de communauté et de temps de calcul.

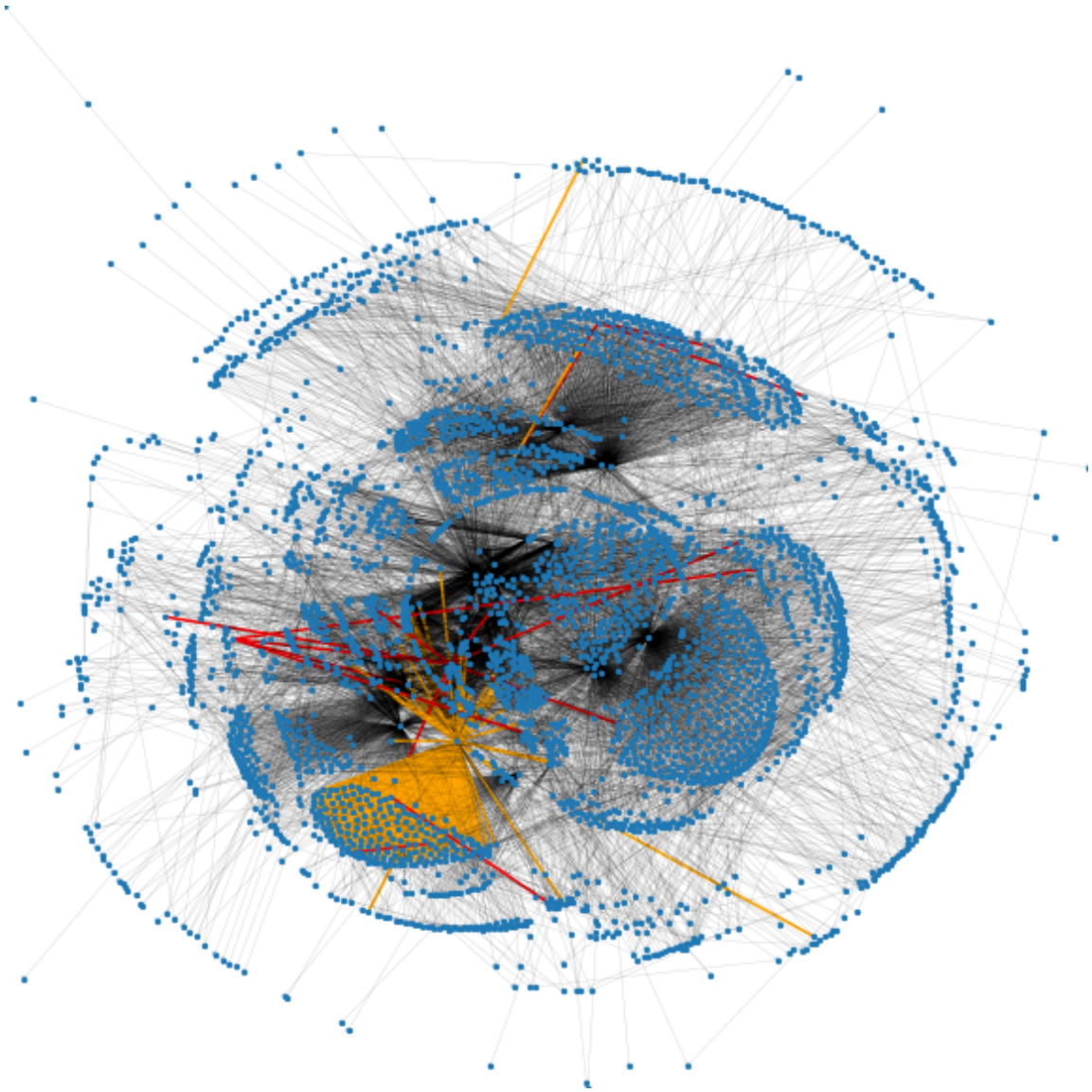


Figure 1.6.2: Etat final du graphe sur lequel travaille GrphClus pour la machine 201 du jour 1. Les noeuds représentent des processus dans la machine, les arêtes en couleur correspondent aux actions de l'attaquant : en orange, les actions liées aux PID 2952 et en rouge, celles liées au PID 5452.

# Chapter 2

# K-bMOM: a robust Lloyd-type clustering algorithm based on bootstrap Median-of-Means

**Abstract.** The median-of-means (MOM) is an estimator of the mean of a random variable that has emerged as an efficient tool to design robust learning algorithms with optimal theoretical guarantees. However, its use for the clustering task suggests dividing the dataset into blocks, which may provoke the disappearance of some clusters in some blocks and lead to bad performances. To overcome this difficulty, a procedure termed "bootstrap median-of-means" is proposed, where the blocks are generated with a replacement in the dataset. The bootstrap MOM has a better breakdown point than the median-of-means if enough blocks are generated. A clustering algorithm called K-bMOM is designed, by performing Lloyd-type iterations together with the use of the bootstrap MOM strategy. Good performances are obtained on simulated and real-world datasets for color quantization and an emphasis is put on the benefits of our robust intialization procedure. On the theoretical side, K-bMOM is also proven to have a non-trivial probabilistic breakdown point in well-clusterizable situations.

## 2.1 Introduction

Massive and complex datasets are often corrupted by outliers. Classical data mining procedures such as K-means or more general EM algorithms for instance, are however, sensitive to the presence of outliers, which can induce time consuming data pre-processing.

In this context, robust versions of data mining procedures are particularly relevant and we investigate a way to produce a Lloyd-type algorithm for hard clustering that is robust with respect to the presence of outliers. We propose more precisely using a variant of the median-of-means (MOM) strategy, that we call "bootstrap median-of-means" (bMOM). The MOM principle has been the object of recent intensive research in mean estimation, regression, high-dimensional framework and supervised classification and machine learning ([90, 41, 84, 85, 97, 96, 98, 104]). Other approaches to robustness for K-means also exist in the literature, such as for instance, K-median or trimmed K-means [54, 22] to name but a few. The design of robust estimators with a control of the algorithmic complexity has also been investigated [42].

Given a dataset, bMOM consists of first generating a (large) bootstrap sample and then performing a classical median-of-means on this bootstrap sample. This can be seen also as a so-called subragging procedure - for "sub-sample robust aggregating" - in the terminology of Bühlmann [28]. We prove in Section 2.3.1 that if enough blocks are generated from the bootstrap sampling, then for a fixed block size, bMOM has a higher breakdown point than MOM. In other words, bMOM is more robust to contamination than the classical MOM. Note that one strength of bMOM, that will be very useful in the context of clustering, is that sampling is done *with replacement* when constructing the blocks. Hence, the number of blocks for a fixed length is not limited by the amount of initial data, unlike MOM or its variant by sampling without replacement ([83]).

We propose a robust-to-outliers version of K-means, that we call K-bMOM, and that performs Lloyd-type iterations through the use of bMOM estimates for the K-means risk, as further explained in Section 2.2. In that section, a robust variant to the traditional K-means++ initialization strategy by applying the bMOM strategy is also presented.

Theoretical results are summarized in Section 2.3. We prove in particular in Section 2.3.2, that the K-bMOM algorithm is robust in a sense of a probabilistic version of a breakdown point if the initial data is in a well-clusterizable situation. This is very much in line with the results on the trimmed K-means for example ([52]).

Further theoretical results, concerning an idealized version of our algorithm, can be found in section 2.7.

In Section 2.4, the scope of application of K-bMOM is illustrated and practical considerations and guidelines are provided for choosing the number and size of the blocks. In Section 2.5, the proposed initialization procedure and the K-bMOM approach are tested in several simulation settings of outliers. It is also compared to existing robust K-means based clustering approaches in Section 2.8. And finally, this algorithm is applied to the well-known problem of color quantization in the image processing field.

Our framework is close to the recent work [81] that investigates the use of median-of-means statistics to produce a robust K-means type clustering. However, the latter work is theoretical only and the authors study probabilistic performance bounds for the minimizer of the median-of-means of the K-means distortion loss under a finite second moment assumption. In particular the authors do not discuss the use of median-of-means through Lloyd-type iterations nor a practical way to compute the estimator. Neither do they discuss the possibility of generating blocks with replacements in the dataset.

## 2.2 K-bMOM procedure

We recall first in Section 2.2.1, the Median-of-Means procedure and introduce a variant, called bootstrap Median-of-Means (bMOM), for the estimation of the mean in dimension one. We then use the latter methodology in a robust iterative clustering algorithm presented in Section 2.2.2. Our algorithm applies to multi-dimensional data, by estimating centroids according to a bMOM strategy applied to the K-means risk, that is real valued. Moreover, since most clustering approaches crucially depend on the choice of the starting centers, we propose in Section 2.2.3, a robust initialization procedure based on the bMOM principle.

### 2.2.1 Median-of-Means and bootstrap Median-of-Means

Consider a real-valued sample $u_1^n = (u_1, ..., u_n)$ and a partition $I_1^B = \{I_b : b \in \{1, ..., B\}\}$ of the set of indices $\{1, ..., n\}$. The block of index $b \in \{1, ..., B\}$ corresponds to the dataset $(u_i)_{i \in I_b}$. There are thus $B$ disjoint blocks, that form a partition of the sample. By a slight abuse of language, we sometimes refer to the "block $b$" instead of the "block of index $b$". The median-of-means (MOM) estimator of the mean in dimension one ([4, 74, 109]), subject to the partition of indices $I_1^B$, consists of taking a median of the arithmetic means computed on the collection of

blocks. The lengths of the blocks are generally taken to be equal, possibly up to one data. We thus write the MOM estimator as follows,

$$\mathrm{MOM}(u_1^n, I_1^B) = \mathrm{med}\left\{ \frac{1}{|I_b|} \sum_{i \in I_b} u_i : b \in \{1, ..., B\} \right\},$$

where $|I_b|$ denotes the cardinal of $I_b$ and med is a median, that is $|\left\{b \in \{1, ..., B\} : a_b \leq \mathrm{med}\left\{a_1^B\right\}\right\}| \geq B/2$ and $|\left\{b \in \{1, ..., B\} : a_b \geq \mathrm{med}\left\{a_1^B\right\}\right\}| \geq B/2$ for a median of a collection of real numbers $a_1^B = (a_1, ..., a_B)$. In the following, when the set of possible medians is not a singleton, we always consider its middle point as being our choice of median, that is thus uniquely defined.

We may consider that the blocks are generated according to a random sampling process, that proceeds without replacements (disjoint blocks) and according to the uniform distribution over the remaining data at each step. This formulation naturally leads to consider more general random block generating processes.

For any positive integers $n_B$ and $B$, denote $q = Bn_B$ and generate a bootstrap sample $v_1^q = (v_1, ..., v_q)$ from the dataset $u_1^n$. More precisely, each $v_l$, $l \in \{1, ..., q\}$, is taken uniformly at random from the values $(u_1, ..., u_n)$ and independently from $(v_{l'})_{l' \neq l}$. The bootstrap median-of-means (bMOM) of the dataset $u_1^n$ with parameters $n_B$ and $B$ is then the (classical) MOM estimator based on the bootstrap sample $v_1^q$ with a partition of indices $I_1^B$ given by $I_b = \{(b-1)n_B + 1, ..., bn_B\}$ for any $b \in \{1, ..., B\}$,

$$\mathrm{bMOM}(u_1^n, n_B, B) = \mathrm{MOM}(v_1^q, I_1^B). \tag{2.2.1}$$

Note that the bMOM is a randomized estimator, due to the sampling of the bootstrap tuple $(v_1, ..., v_q)$. Also, for any fixed sample size $n$, we can choose any block size $n_B$ and number of blocks $B$ to define a bMOM estimator, unlike the classical MOM, where the product of the block size with the number of blocks is equal to the sample size. This will turn out to be extremely useful in the clustering context, where we do not want too small block sizes in order to avoid the disappearance of some clusters in the blocks.

### 2.2.2 A robust Lloyd-type algorithm

In this section we propose a hard clustering algorithm based on the bMOM strategy. Unlike section 2.2.1 above, we consider multi-dimensional data, but the bMOM strategy will still be applied to some one-dimensional statistics.

Let us first introduce some notations. Let $x_1^n = (x_1, \ldots, x_n) \in \mathbb{R}^{p \times n}$ denote a dataset of $n$ observations belonging to the Euclidean space $\mathbb{R}^p$, that we want to cluster into $K$ homogeneous groups. We choose two positive integers $B$ and $n_B$, with $n_B > K$. For $b \in \{1, ..., B\}$, we denote by $(y_1^{(b)}, \ldots, y_{n_B}^{(b)})$ the block of size $n_B$ and of index $b$ generated according to the bootstrap sampling process, that selects at each step, independently from the other steps, an observation according to the uniform distribution over the sample $x_1^n$. The collection $(y_1^{(1)}, ..., y_{n_B}^{(1)}, ..., y_1^{(B)}, ..., y_{n_B}^{(B)})$, thus forms a bootstrap sample of size $q = Bn_B$ generated from the dataset $x_1^n$. Again, by a slight abuse of language, we sometimes refer to the "block $b$" instead of the "block of index $b$". We define the empirical risk - also called distortion - of the block $b$ as:

$$R^{(b)} = \frac{1}{n_B} \sum_{k=1}^{K} \sum_{l=1}^{n_B} \left\| y_l^{(b)} - c_k^{(b)} \right\|^2 \mathbf{1}\{y_l^{(b)} \in \mathcal{C}_k^{(b)}\},$$

where $y_l^{(b)}$ stands for the $l$-th datapoint contained in the $b$-th block, $\mathcal{C}_k^{(b)}$ stands for the set of datapoints belonging to the cluster $k$ in the block $b$, $\|.\|$ is the Euclidean norm in $\mathbb{R}^p$ and $\mathbf{1}\{E\}$ is the indicator of the event $E$, that equals 1 when $E$ is true and 0 otherwise. Furthermore, $c_k^{(b)}$ stands for the mean vector of the cluster $k$ in the block $b$. Finally, we denote by $\mathcal{P}(\mathbf{c}^{(b)})$, the Voronoï partition obtained from the set of centroids $\mathbf{c}^{(b)} = (c_1^{(b)}, ..., c_K^{(b)})$.

#### 2.2.2.1 The K-bMOM algorithm

The algorithm that we propose alternates three main steps iteratively. At iteration $t$, $B$ blocks each containing $n_B$ data are built by sampling uniformly with replacement, and independently from the other iterations, from the original dataset $x_1^n$. A partition per block is then computed by assigning each data point to its closest centroid given by the previous iteration. The centroids of each block of index $b$ at iteration $t$, noted as $\mathbf{c}_t^{(b)} = (c_{1,t}^{(b)}, ..., c_{K,t}^{(b)})$, are then updated according to their block partition and the empirical risk $R_t^{(b)}$ is calculated. The block with the median empirical risk, denoted by *bmed*, is selected and the centers of this median block become the current ones. The bMOM strategy is thus used here since we consider the median of the risks, that are real-valued empirical means of the K-means loss computed from the data in each block. These steps are repeated several times.

In practice, the algorithm is run through a given number of maximum iterations ($t_{max} = 25$ by default). In order to obtain a more precise estimation of the centroids, instead of retrieving the centroids of the median block computed in the last iteration, the centroids of the last 10 iterations are aggregated. The algorithm thus returns

$\bar{\mathbf{c}}^{(bmed)} = (\bar{c}_1^{(bmed)}, \ldots, \bar{c}_K^{(bmed)})$ such that $\bar{c}_k^{(bmed)} = 1/10 \sum_{t=t_{max}-10}^{t_{max}} \widehat{c}_{k,t}^{(bmed)}$ where $\widehat{c}_{k,t}^{(bmed)}$ stands for the centroid of the cluster $k$ of the median block at iteration $t$. The final partition over the whole dataset is obtained by assigning each data point to its nearest centroid in $\bar{\mathbf{c}}$. A pseudo algorithm of this procedure is detailed in Algorithm 2.1.

Our algorithm shares some similarity with the techniques of so-called consensus/ensemble clustering ([113]), since it amounts at each step, to producing a robust clustering, given by a codebook, from a collection of candidates computed on bootstrap sub-samples. However, there are also essential differences, since we select one of the candidates by a simple median criterion for dimension one statistics, whereas consensus clusterings aggregate the candidates in a more complicated fashion, using some similarity measures between clusterings. Interestingly, so-called bagged clustering ([89, 44, 45]) proposes performing clusterings on bootstrap samples and aggregating them using a hierarchical clustering on the collection of obtained centroids. Moreover, the size of the bootstrap samples are equal to the original sample size, whereas in our approach the sub-sampling is crucial and directly related to the allowed proportion of outliers (see Section 2.3.2).

A robust trimmed clustering approach for probabilities in Wasserstein space is developed in [40] and used to advantage to robustly aggregate model-based clusterings on multivariate data - each clustering being seen as a probability - that are previously learned on sub-samples of the original data. This approach, however, concerns the robust aggregation of model-based clusterings, whereas our focus is on robust hard clustering in the context of the K-means problem.

#### 2.2.2.2 Model selection

In model-based clustering, it is frequent to consider several models in order to find the most appropriate one for the considered data. In particular, for most clustering algorithms, the model is specified by its number of clusters $K$. There are lots of ad-hoc approaches in the literature to select the number of components $K$ and we can therefore think of the Gap statistics from [139], the Silhouette criterion and so one.

Furthermore, since the K-means algorithm can be seen as a hard version of an EM-like algorithm which tries to estimate a mixture of $K$ Gaussians with isotropic covariance matrices, we can therefore try to adapt classical tools for model selection including BIC, ICL criteria and the slope heuristics [12] for example. All these model selection criteria are based on a so-called penalty that is added to the empirical risk. However, it is not reasonable to use the empirical risk in our context, due to the presence of outliers. Instead, for each $K$, we could think of using

**Algorithm 2.1** Iteration phase structure

**procedure** K-BMOM($x_1^n, K, B, n_B$) ▷ ($n_B > K$)
  Let $(c_{1,0}, \ldots, c_{K,0})$ be the $K$ initial centroids and called reference centroids.
  Set $t = 1$.
  **while** $t \leq t_{max}$ **do**
    Create $B$ blocks $(y_{1,t}^{(b)}, ..., y_{n_B,t}^{(b)})$ for $b \in \{1, ..., B\}$, according to a random sampling process that at each step selects an observation uniformly over the data $x_1^n$ and independently from the other steps.
    **for all** $b \in \{1, \ldots, B\}$ **do**
      Assign each data point in the block of index $b$ to its closest reference centroid.
      Set $n_{k,t}^{(b)}$ the number of data points in the block $b$ belonging to the cluster $k$.
      **if** $n_{k,t}^{(b)} > 1, \forall k \in \{1, \ldots, K\}$ **then**
        **for all** $k \in \{1, \ldots, K\}$ **do**
          $c_{k,t}^{(b)} \leftarrow 1/n_{k,t}^{(b)} \sum_{l=1}^{n_B} y_{l,t}^{(b)} \mathbf{1}\{y_{l,t}^{(b)} \in \mathcal{C}_{k,t}^{(b)}\}$.
          $R_t^{(b)} \leftarrow \frac{1}{n_B} \sum_{k=1}^{K} \sum_{l=1}^{n_B} \left\| y_{l,t}^{(b)} - c_{k,t}^{(b)} \right\|^2 \mathbf{1}\{y_{l,t}^{(b)} \in \mathcal{C}_{k,t}^{(b)}\}$.
        **end for**
      **else**
        Skip the block.
      **end if**
    **end for**
    Get the median block *bmed* such that $R_t^{(bmed)} = \text{med}\left\{R_t^{(b)} : b \in \{1, ..., B\}\right\}$ and $\left(\widehat{c}_{1,t}^{(bmed)}, \ldots, \widehat{c}_{K,t}^{(bmed)}\right)$ the centroids assigned to the median block *bmed* at iteration $t$ becoming the reference centroids.
    $t \leftarrow t + 1$.
  **end while**
  **return** $\bar{\mathbf{c}}^{(bmed)} = \left(\bar{c}_1^{(bmed)}, \ldots, \bar{c}_K^{(bmed)}\right)$ such that $\bar{c}_k^{(bmed)} = \frac{1}{10} \sum_{t=t_{max}-10}^{t_{max}} \widehat{c}_{k,t}^{(bmed)}$ for all $k \in \{1, ..., K\}$ and $\mathcal{P}(\bar{\mathbf{c}}^{(bmed)})$.
**end procedure**

the centroids that are returned by the K-bMOM algorithm, generate blocks according to the bootstrap sampling process and use the median empirical risk over the blocks as a robust estimate of the empirical risk. The idea would then be to penalize these median empirical risks for various values of $K$ by a classical penalty, such as the one used in BIC for instance. Such robust model selection procedure would require investigations that represent an interesting direction of research for future work.

### 2.2.3 A robust initialization

It is well-known that since the clustering problem is non-convex, the initialization step is a keystone for the resulting partition. We propose therefore, a robust variant of a classical and efficient initialization procedure by applying the bMOM strategy.

More precisely, the idea is to build $B$ blocks of $n_B$ data points, with $n_B > K$, by sampling uniformly and with replacement over the dataset $x_1^n$. The K-means++ initialization [6] is then operated in each block. Recall that the latter approach proceeds in an iterative way: it starts with a centroid picked at random among the data points. Iteratively and until the number of groups $K$ is reached, a new centroid is then chosen from the data points with a probability which increases quadratically with the squared Euclidean distance to the already chosen closest centers.

In each block, the empirical risk $R_{++}^{(b)}$ is therefore computed and the centers linked to the median empirical risk, called the median block, are selected as the initial centers. This algorithm is summarized in Algorithm 2.2. The robustness of this initialization scheme is evaluated in Section 2.5.2.

**Algorithm 2.2** initialization strategy

**procedure** K-BMOM-KM++$(x_1^n, K, B, n_B)$
  Create $B$ blocks $(y_1^{(b)}, ..., y_{n_B}^{(b)})$ for $b \in \{1, ..., B\}$, according to a random sampling process that at each step selects an observation uniformly over the data $x_1^n$ and independently from the other steps.
  **for all** $b \in \{1, \dots, B\}$ **do**
    Proceed to a $K$-means++ initialization based on the sample $(y_1^{(b)}, ..., y_{n_B}^{(b)})$. This gives the centroids $(c_{1,++}^{(b)}, ..., c_{K,++}^{(b)})$.
    Compute the empirical risk $R_{++}^{(b)}$ of the block $b$:
    $R_{++}^{(b)} \leftarrow \frac{1}{n_B} \sum_{k=1}^{K} \sum_{l=1}^{n_B} \left\| y_l^{(b)} - c_{k,++}^{(b)} \right\|^2 \mathbf{1}\{y_l^{(b)} \in \mathcal{C}_k^{(b)}\}$.
  **end for**
  Select the block *bmed* having the median empirical risk.
  **return** $\left(\widehat{c}_{1,++}^{(bmed)}, \dots, \widehat{c}_{K,++}^{(bmed)}\right)$ the centroids of the median block *bmed*.
**end procedure**

## 2.3 A breakdown point analysis

We prove in Section 2.3.1 below that the bMOM estimator of the mean enables us to perform a more robust mean estimation than MOM, if enough blocks are generated, in the sense that the breakdown point of bMOM is higher. We believe that this result is of independent interest, as it shows the advantage of taking a large bootstrap sample before applying the MOM principle. This result still has some limitations however in the perspective of clustering, since K-bMOM does not correspond to the bMOM estimator for $K = 1$. In Section 2.3.2 therefore, we study a notion of breakdown point for the K-bMOM algorithm itself.

### 2.3.1 Breakdown points for mean estimation

The breakdown point is a classical concept of robust statistics ([65, 70, 100]), that gives the maximal proportion of outliers that is allowed so that the deviations of the estimator stay bounded compared to the no-corruption setting. Assume that we are given a sample $u_1^n = (u_1, ..., u_n)$ of real valued random variables.

**Definition 2.1** (Deterministic Breakdown point)**.** The (deterministic) breakdown point $\delta_n(\widehat{T}, u_1^n)$ of a real-valued estimator $\widehat{T}$ given the sample $u_1^n$, is the maximal proportion of outliers that leaves the value of the estimator bounded.

$$\delta_n\left(\widehat{T}, u_1^n\right) = \frac{1}{n} \max\left\{ m : \max_{i_1,...,i_m} \sup_{e_1,...,e_m} \left|\widehat{T}\left(s_1, ..., s_n\right)\right| < \infty \right\},$$

where the sample $(s_1, ..., s_n)$ is obtained by replacing the $m$ data points $u_{i_1}, ..., u_{i_m}$ of the sample $u_1^n$ by arbitrary values $e_1, ..., e_m$.

Definition 2.1 corresponds to a worst case analysis, the outliers potentially appearing at the worst places for the estimator $\widehat{T}$. If the estimator $\widehat{T}$ is randomized - rather denoted $\widehat{T}^\omega$ in this case -, then its breakdown point is a random variable.

For a median $\text{med}\{u_1^n\}$, it holds that $\delta_n\left(\text{med}\{u_1^n\}, u_1^n\right) = \lfloor (n-1)/2 \rfloor / n$ and for the empirical mean, $\bar{u}_n = \sum_{i=1}^n u_i/\text{n}$, $\delta_n\left(\bar{u}_n, u_1^n\right) = 0$.

**Proposition 2.1.** *The breakdown point of the median-of-means estimator of the mean in dimension one is*

$$\delta_n\left(\text{MOM}(u_1^n, I_1^B), u_1^n\right) = \frac{\left\lfloor \frac{B-1}{2} \right\rfloor}{n}.$$

The proof of Proposition 2.1 is direct since MOM diverges if and only if there is at least one outlier in a majority of blocks. Note that the same breakdown point is achieved for a more general estimator of a multi-dimensional mean called the median-of-means tournament ([126]).

Let us now consider the use of replacements while constructing the blocks and study the breakdown point of bMOM.

**Proposition 2.2.** *Assume first that the sample size satisfies $n = Bn_B$, for positive integers $B$ and $n_B$. We then have*

$$\delta_n\left(\text{bMOM}(u_1^n, n_B, B), u_1^n\right) \leq \delta_n\left(\text{MOM}(u_1^n, I_1^B), u_1^n\right) \ a.s.$$

*Secondly, fix the block size $n_B$ and the sample size $n$ and let the number of blocks taken in the bMOM, tend to infinity. It holds that*

$$\frac{\left\lceil n\left(1-\frac{1}{2^{1/n_B}}\right)\right\rceil - 1}{n} \leq \lim\inf_{B\to\infty} \delta_n\left(\text{bMOM}\left(u_1^n, n_B, B\right), u_1^n\right) \ a.s.$$

*and*

$$\lim\sup_{B\to\infty} \delta_n\left(\text{bMOM}\left(u_1^n, n_B, B\right), u_1^n\right) \leq \frac{\left\lfloor n\left(1-\frac{1}{2^{1/n_B}}\right)\right\rfloor + 1}{n} \ a.s.$$

*Note that* $1-\frac{1}{2^{1/n_B}} \sim_{n_B\to\infty} \frac{\log 2}{n_B} \approx \frac{0.69}{n_B}$.

*Proof.* The first display follows from the fact that the blocks in bMOM are built with replacement. Indeed, let us distinguish two cases: either there is not repeated datapoint in bMOM and in this case bMOM corresponds to a usual MOM, or there is at least one repeated datapoint. In this second case, one can built a set of indexes $\{i_1, ..., i_m\}$ such that the corrupted data $\{u_{i_1}, ..., u_{i_m}\}$ are present more than $m$ times in the blocks of bMOM while there will always be at most $m$ corrupted data in the usual MOM. This explains why the definitions 2.2.1 and 2.1 give $\delta_n\left(\text{bMOM}(u_1^n, n_B, B), u_1^n\right) \leq \delta_n\left(\text{MOM}(u_1^n, I_1^B), u_1^n\right)$ $a.s$ when one takes $n = Bn_B$.

For the second part of the proposition, recall that by definition,

$$\delta_n\left(\text{bMOM}\left(u_1^n, n_B, B\right), u_1^n\right) = \frac{1}{n}\max\left\{m : \max_{i_1,...,i_m} \sup_{e_1,...,e_m} \left|\text{bMOM}\left((s_1, ..., s_n), n_B, B\right)\right| < \infty\right\},$$

where the sample $(s_1, ..., s_n)$ is obtained by replacing the $m$ data points $u_{i_1}, ..., u_{i_m}$ of the sample $u_1^n$ by arbitrary values $e_1, ..., e_m$.

The set of indices of the outliers is denoted $\mathcal{O} = \{i_1, ..., i_m\}$ and its complementary, the set of indices of regular data is denoted $\mathcal{I}$. Let the set $\mathscr{A}$ parametrize the randomness of the bootstrap sampling procedure over the set $(s_1, ..., s_n)$ and set $\Omega := \mathscr{A}^{\mathbb{N}^*} = \left\{\omega = (\omega_i)_{i\geq 1} : \omega_i \in \mathscr{A}\right\}$, the set of sequences of successive bootstrap samplings. Since the blocks in the bMOM estimator are chosen by a bootstrap procedure, we denote $X^*_{i,b}(\omega_j)$ the $i$-th value of the block $b$ picked at random uniformly with replacement in $(s_1, ..., s_n)$. The block $b$ then writes

$$\left\{X^*_{1,b}(\omega_{n_B\cdot b+1}), X^*_{2,b}(\omega_{n_B\cdot b+2}), ..., X^*_{n_B-1,b}(\omega_{n_B\cdot b+n_B-1}), X^*_{n_B,b}(\omega_{n_B\cdot(b+1)})\right\}.$$

We need to know if this block is corrupted or not, so we define the quantity

$$S^{(\omega)}_{b,\mathcal{I}} := \mathbf{1}_{\left\{X^*_{1,b}(\omega_{n_B\cdot b+1})\in\mathcal{I}, X^*_{2,b}(\omega_{n_B\cdot b+2})\in\mathcal{I}, ..., X^*_{n_B,b}(\omega_{n_B\cdot(b+1)})\in\mathcal{I}\right\}},$$

which value is 1 if the block $b$ is not corrupted and 0 otherwise. $S_{b,\mathcal{I}}$ is a bernoulli random variable with parameter $\prod_{k=1}^{n_B}\mathbb{P}\left(X^*_{k,b}\in\mathcal{I}\right) = \mathbb{P}\left(X^*_{1,b}\in\mathcal{I}\right)^{n_B} = \left(1-\frac{m}{n}\right)^{n_B}$. Let us also define

$$\overline{S}^{(\omega)}_{B,\mathcal{I}} := \frac{1}{B}\sum_{b=1}^{B} S^{(\omega)}_{b,I}.$$

Take $\omega\in\Omega$. The following equivalence holds,

$$\overline{S}^{(\omega)}_{B,\mathcal{I}} > \frac{1}{2} \quad\Leftrightarrow\quad \sup_{e_1,...,e_m} |\mathrm{bMOM}((s_1, ..., s_n), n_B, B)| < \infty.$$

Furthermore, by the strong law of large numbers we have: $\exists\Omega_{\mathcal{I}}\subset\Omega$, $\mathbb{P}(\Omega_I) = 1$,

$$\forall\omega\in\Omega_{\mathcal{I}}, \lim_{B\to\infty}\overline{S}^{(\omega)}_{B,\mathcal{I}} = \mathbb{E}(S_{1,\mathcal{I}}) = \left(1-\frac{m}{n}\right)^{n_B}.$$

Set now $m_0\in\mathbb{N}$ and assume that $\left(1-\frac{m_0}{n}\right)^{n_B} > \frac{1}{2}$. By definition of the limit: $\exists\Omega_{\mathcal{I}}\subset\Omega$, $\mathbb{P}(\Omega_{\mathcal{I}}) = 1$, $\forall\omega\in\Omega$, $\exists B_{\mathcal{I},\omega}\in\mathbb{N}^*$ such that $\forall B\geq B_{\mathcal{I},\omega}$, $\overline{S}^{(\omega)}_{B,\mathcal{I}} > \frac{1}{2}$.

Let us define $\Lambda_{m_0} := \{\mathcal{I}\in\mathcal{P}(\{1, ..., n\}) : \mathrm{Card}(\mathcal{I}) = n - m_0\}$, the collection of subsets of $\{1, ..., n\}$ with cardinality $n - m_0$. As $\Lambda_{m_0}$ is finite, by setting $\Omega_0 = \bigcap_{\mathcal{I}\in\Lambda_{m_0}}\Omega_{\mathcal{I}}$, we have $\mathbb{P}(\Omega_0) = 1$. Also, $\forall\omega\in\Omega_0$, $\exists B_0 :=$

$\max_{\mathcal{I}\in\Lambda_{m_0}}(B_{\mathcal{I},\omega})\in\mathbb{N}^*$ such that $\forall B\geq B_0$, $\forall\mathcal{I}\in\Lambda_{m_0}$, $\overline{S}^{(\omega)}_{B,\mathcal{I}}>\frac{1}{2}$, that is,

$$\max_{i_1,\ldots,i_{m_0}}\sup_{e_1,\ldots,e_{m_0}}\left|\text{bMOM}\left((s_1,...,s_n),n_B,B\right)\right|<\infty.$$

Rewriting the condition $\left(1-\frac{m_0}{n}\right)^{n_B}>\frac{1}{2}$ as $m_0\leq\left\lceil n\left(1-\frac{1}{2^{1/n_B}}\right)\right\rceil-1$, we obtain

$$\liminf_{B\to\infty}\delta_n\left(\text{bMOM}\left(u_1^n,n_B,B\right),u_1^n\right)\geq\frac{\left\lceil n\left(1-\frac{1}{2^{1/n_B}}\right)\right\rceil-1}{n}\ a.s.$$

Conversely, if $\left(1-\frac{m_0}{n}\right)^{n_B}<\frac{1}{2}$, that is $m_0\geq\left\lfloor n\left(1-\frac{1}{2^{1/n_B}}\right)\right\rfloor+1$, then $\forall\omega\in\Omega_0$, $\forall\mathcal{I}\in\Lambda_{m_0}$, $\overline{S}^{(\omega)}_{B,\mathcal{I}}>\frac{1}{2}$, that is,

$$\max_{i_1,\ldots,i_{m_0}}\sup_{e_1,\ldots,e_{m_0}}\left|\text{bMOM}\left((s_1,...,s_n),n_B,B\right)\right|=\infty.$$

Therefore,

$$\limsup_{B\to\infty}\delta_n\left(\text{bMOM}\left(u_1^n,n_B,B\right),u_1^n\right)\leq\frac{\left\lfloor n\left(1-\frac{1}{2^{1/n_B}}\right)\right\rfloor+1}{n}\ a.s.$$

The proof is now complete. □

On the one hand, the first display in Proposition 2.2 states that when the number of blocks in bMOM is equal to the number of blocks in MOM, bMOM has a breakdown point that is less than or equal to the breakdown point of MOM. On the other hand, the second part of Proposition 2.2 states that for a fixed block size, when the number of blocks in bMOM tends to infinity, its breakdown point is strictly greater than the breakdown point of MOM taken with the same block size, at least for $n$ sufficiently large. Indeed, by assuming $n=Bn_B$, Proposition 2.1 gives that the breakdown point of MOM is smaller than $1/(2n_B)$, while for $n$ sufficiently large, the second part of Proposition 2.2 ensures that the breakdown point of bMOM is strictly greater than $1/(2n_B)$. This is of importance for practice, since it implies that one should consider if possible, building a bootstrap sample that is larger than the original one, and then take the MOM statistics on this bootstrap sample rather than on the original dataset.

Considering that the contaminated sample is given (fixed), it is interesting to evaluate the probability that a randomized estimator does not diverge when the outliers go to infinity. It can indeed happen that the indices of the outliers are not the worst with respect to the block sampling process. This leads to the following definition.

**Definition 2.2** (Probabilistic breakdown point)**.** The probabilistic breakdown point of a randomized estimator $\widehat{T}^{\omega}$ given the sample $u_1^n$ is

$$p_n\left(\widehat{T}^{\omega}, u_1^n, (i_1, ..., i_m)\right) = \mathbb{P}\left(\left\{\omega : \sup_{e_1,...,e_m} \left|\widehat{T}^{\omega}(s_1, ..., s_n)\right| < \infty\right\}\right). \tag{2.3.1}$$

**Definition 2.3.** where the sample $(s_1, ..., s_n)$ is obtained by replacing the $m$ data points $u_{i_1}, ..., u_{i_m}$, for some fixed indices $(i_1, ..., i_m)$, by the arbitrary values $e_1, ..., e_m$.

Note that in the probabilities of the Identity (2.3.1), the non-corrupted dataset $u_1^n$ is fixed, the indices $(i_1, ..., i_m)$ where the outliers replace the non-corrupted data $(x_{i_1}, ..., x_{i_m})$ are fixed and the outliers $(e_1, ..., e_m)$ are deterministic. The only randomness that is taken into account is the randomness induced by the randomized estimator. This may be indeed relevant in practice, where the corrupted dataset is fixed, as it allows us to discuss if a randomized estimator has a high probability of being robust to the presence of outliers, depending on the randomness of its generating process.

As $p_n\left(\text{bMOM}(u_1^n, n_B, B), u_1^n, (i_1, ..., i_m)\right)$ only depends on $m$, but not on the values of $(i_1, ..., i_m)$, we will rather denote it $p_n\left(\text{bMOM}(u_1^n, n_B, B), m\right)$. We have the following bound.

**Proposition 2.3.** *Assume that the block length $n_B$ in bMOM and the proportion of outliers $m/n$ are such that $(1 - m/n)^{n_B} > 1/2$. Then it holds that*

$$p_n\left(\text{bMOM}(u_1^n, n_B, B), m\right) \geq 1 - \exp\left(-2B\left((1 - m/n)^{n_B} - 1/2\right)^2\right).$$

*Proof.* As in the proof of Proposition 2, denote $S_b$, for $b \in \{1, ..., B\}$, the indicator that the block of index $b$ is not corrupted. Since $(1 - m/n)^{n_B} > 1/2$, we have by Hoeffding's inequality ([13, Theorem 2.27]),

$$\begin{aligned}
\mathbb{P}\left(\left\{\omega : \sup_{e_1,...,e_m} |\text{bMOM}(u_1^n, n_B, B)| = \infty\right\}\right) &= \mathbb{P}\left(\sum_{j=1}^{B}(1 - S_j) > B/2\right) \\
&= \mathbb{P}\left(\sum_{j=1}^{B}(1 - S_j) - \mathbb{E}[1 - S_j]) > B(1 - m/n)^B - B/2\right) \\
&\leq \exp\left(-2B\left((1 - m/n)^{n_B} - 1/2\right)^2\right).
\end{aligned}$$

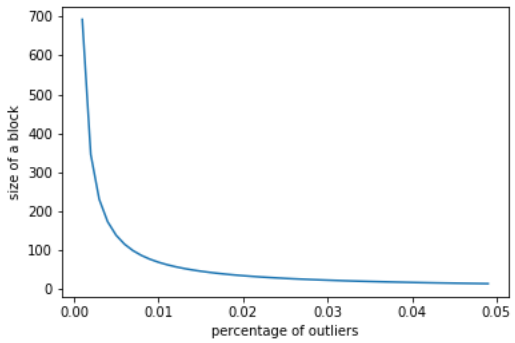



Figure 2.3.1: Maximum admissible block size $n_B$ for bMOM as a function of the proportion of outliers $p = m/n$.

as required □

According to Proposition 2.3, if the number of outliers $m$ and the sample size $n$ are fixed, then the block length $n_B$ should be chosen such that $(1 - m/n)^{n_B} > 1/2$, i.e. $n_B < \log(2)/\log(1/(1 - m/n))$ (Figure 2.3.1). One can indeed notice that the quantity $(1 - m/n)^{n_B}$ corresponds to the probability, according to the bootstrap sampling, that a block is not corrupted. Hence, in case of a large proportion of outliers $m/n$, the block length should not be taken too large. Furthermore, by denoting $D = (1 - m/n)^{n_B} - 1/2 > 0$, we have that $p_n\left(\text{bMOM}(x_1^n, n_B, B), m\right) \geq 1 - R$ if $B > \log(1/R) / \left(2D^2\right)$. Consequently, if the block size $n_B$ is chosen correctly (not too large according to the proportion of outliers, so that $D > 0$), then the probability that the bMOM remains stable under contamination, tends to one when the number of blocks $B$ tends to infinity.

### 2.3.2 Probabilistic breakdown point for the K-bMOM algorithm

Let us turn now to the study of the breakdown point of the K-bMOM algorithm presented in Section 2.2.2. It produces a randomized estimator, consisting of a set of $K$ centroids, due to the sampling of the blocks at each step of the algorithm. We thus denote it $\bar{\mathbf{c}}^{\omega}$ rather than $\bar{\mathbf{c}}$ as in Section 2.2.2, the notation $\omega$ accounting for randomization induced by the sampling of the blocks. We discuss the following notion of probabilistic breakdown point.

**Definition 2.4** (Probabilistic breakdown point for K-bMOM)**.** The probabilistic breakdown point $p_n\left(\bar{\mathbf{c}}^{\omega}, x_1^n, (i_1, ..., i_m)\right)$

of the randomized estimator $\bar{\mathbf{c}}^{\omega} = \bar{\mathbf{c}}$, defined in Section 2.2.2 as the output of the algorithm K-bMOM, is given by

$$p_n\left(\bar{\mathbf{c}}^{\omega}, x_1^n, (i_1, ..., i_m)\right) = \mathbb{P}\left(\left\{\omega : \sup_{y_1,...,y_m} \max_{c \in \bar{\mathbf{c}}^{\omega}(z_1,...,z_n)} \|c\| < \infty\right\}\right), \tag{2.3.2}$$

where the sample $(z_1, ..., z_n)$ is obtained by replacing the $m$ data points $x_{i_1}, ..., x_{i_m}$, for some fixed indices $(i_1, ..., i_m)$, by the arbitrary values $y_1, ..., y_m$ and $\bar{\mathbf{c}}^{\omega}(z_1, ..., z_n)$, consisting in a set of $K$ centroids, is the output of the K-bMOM algorithm when taking as input the dataset $(z_1, ..., z_n)$.

The probabilistic breakdown point defined in Identity (2.3.2) of Definition 2.4 corresponds to the probability that the $K$ centroids output by the K-bMOM algorithm, stay bounded when the input dataset is corrupted by outliers that can take any possible values. In this probability, the non-corrupted dataset $x_1^n$ is fixed, the indices $(i_1, ..., i_m)$ where the outliers replace the non-corrupted data $(x_{i_1}, ..., x_{i_m})$ are fixed and the outliers $(y_1, ..., y_m)$ are deterministic. The only randomness that is taken into account is the randomness induced by the sampling of the blocks at each step of the algorithm. In practice indeed, the corrupted dataset is given to the statistician and it is important to know if the randomized estimator produced by the K-bMOM algorithm has a high probability, through the sampling process, of being robust to the presence of outliers.

The K-bMOM algorithm will be proven to be robust in terms of its probabilistic breakdown point in the case of a "well-clusterizable" clustering configuration, that is a classical assumption for obtaining robustness in clustering ([54, 125]). Roughly speaking, a well-clusterizable configuration is made of "compact" clusters that are well "separated". We give the following formal definition, suitable for our needs.

**Definition 2.5** (well-clusterizableness)**.** A dataset $x_1^n$ is said to be in a well-clusterizable configuration, with compactness parameter $r$ and separation parameter $R$ satisfying $R > 2r > 0$, if the points $x_1^n$ lie in a union of $K$ disjoint balls $B(a_k, r)$, $k = 1, ..., K$, of radius $r$ with centers $a_k$ separated from each other by at least a distance $R$: $\min_{k \neq k'} ||a_k - a_{k'}|| \geq R$. Moreover, each ball $B(a_k, r)$ is assumed to contain exactly one cluster.

**Theorem 2.1** (breakdown point KbMOM)**.** *Let $\bar{\mathbf{c}}^{\omega}$ be the K-bMOM output, computed iteratively by using at each step $B$ blocks of size $n_B$. Assume that the block length $n_B$ and the proportion of outliers $m/n$ are such that $(1 - m/n)^{n_B} > 1/2$. Assume furthermore that the regular data points $x_1^n$ are in a well-clusterizable situation, with compactness and separation parameters denoted respectively $r$ and $R$, satisfying $R^2 > 16 n_B r^2$. Finally, assume that at the beginning of the last 10 iterations, the algorithm has identified the correct partition of the regular data,*

*meaning that one cluster is associated with one centroid. It holds then that* $p_n(\bar{\mathbf{c}}^\omega, x_1^n, (i_1, ..., i_m)) \geq \max\{p_1 - p_2, 0\}$ *with*

$$p_1 = \left(1 - \sum_{k=1}^{K} \left(1 - \frac{n_k^r}{n}\right)^{n_B}\right)^{10B} \tag{2.3.3}$$

*and*

$$p_2 = 10 \exp\left(-2B\left(\left(1 - \frac{m}{n}\right)^{n_B} - \frac{1}{2}\right)^2\right), \tag{2.3.4}$$

*where the quantity* $n_k^r$ *in display (2.3.3) stands for the number of regular data belonging to cluster* $k$ *in the sample* $(z_1, ..., z_n)$ *defined in (2.3.2).*

*Proof.* As described in Section 2.2.2 of the main part of the chapter, the output of the K-bMOM algorithm $\bar{\mathbf{c}}^\omega$ is given by the average of the last 10 codebooks computed through the iterations. Let us first prove the following property (**P**): if all the $K$ clusters are represented in each of the blocks generated during the last 10 iterations and if in each of these iterations, a majority of blocks are made of regular data only, the procedure does not break down.

Indeed, consider the first of the last 10 iterations, when each cluster is represented in all the $B$ blocks, a majority of which consisting of regular data only. For such regular blocks, as the algorithm has found the right partition, the new centroids are barycenters of data in one cluster, that belong to a ball of radius $r$ and so the risk of the new codebooks in regular blocks is less than $4r^2$.

Now, consider a block that contains at least one outlier. Set a constant $A > 0$ such that the regular data is contained in a ball centered at the origin and of radius $A$. After updating the centroids in this block, two cases are possible. Either one of the centroids is a barycenter between some regular data and some outliers, amongst which an outlier denoted $y$ that has the greatest norm in the dataset $z_1^n$. The risk in this configuration is thus greater than $(||y|| - A)^2/(4n_B)$, where $||y||$ is assumed to be greater than $A$ without loss of generality.

Or, the second case could be that the outlier $y$ with the greatest norm lies in a cluster that does not contain any regular data. Hence, at most $K - 1$ centroids are assigned to the regular data. Consequently, one centroid is a barycenter between regular data points of at least two clusters and the risk in this configuration is greater than $R^2/(4n_B)$. Finally, we see that if $R^2 > 16n_Br^2$ and $(||y|| - A)^2 > 16n_Br^2$, then the risk of any regular block is less than the risk of any block containing some outliers. As a majority of blocks are regular, selecting the set of centroids achieving the median of the risks along the blocks, will give a codebook corresponding to a regular block. As we already noted, these centroids induce the right partition in the sense that they are associated with data in

one cluster only. Hence, the reasoning extends to the next iterations and property (**P**) is proven.

To obtain a lower bound on the probabilistic breakdown point $p_n(\bar{\mathbf{c}}^\omega, x_1^n, (i_1, ..., i_m))$, it suffices now to provide a lower bound on the probability of the event described in property (**P**). Denote by $E_1$ the event where the $K$ clusters are represented in each block generated during the last 10 iterations and $E_2$ the event where a majority of blocks are contaminated by some outliers in at least one of the last 10 iterations. By the first part of the proof, it holds

$$p_n(\bar{\mathbf{c}}^\omega, x_1^n, (i_1, ..., i_m)) \geq \mathbb{P}\left(E_1 \bigcap E_2^c\right) \geq \mathbb{P}(E_1) - \mathbb{P}(E_2)\,.$$

By denoting $F_1$ by the event where one block, considered to be fixed, contains representatives of the $K$ clusters, we get $\mathbb{P}(E_1) = (\mathbb{P}(F_1))^{10B}$ by independence of the block generating process. Now, considering events where the data in the considered block come from outliers and all but one cluster, we obtain $\mathbb{P}(F_1) \geq 1 - \sum_{k=1}^K (1 - n_k^r/n)^{n_B}$ where $n_k^r$ is the number of regular data in cluster $k$, which gives

$$\mathbb{P}(E_1) \geq \left(1 - \sum_{k=1}^K \left(1 - \frac{n_k^r}{n}\right)^{n_B}\right)^{10B} = p_1\,.$$

To conclude, it remains to bound $\mathbb{P}(E_2)$ from above by the quantity $p_2$. Denote $S_b$ the indicator that the block $b$ is not corrupted, for some generation of B blocks of length $n_B$. A simple union bound along the 10 iterations gives

$$\mathbb{P}\left(E_2\right) \leq 10 \times \mathbb{P}\left(\sum_{b=1}^B (1 - S_b) > \frac{B}{2}\right)\,.$$

Since $(1 - m/n)^{n_B} > 1/2$, we apply Hoeffding's inequality as in the proof of Proposition 3, and get

$$\mathbb{P}\left(\sum_{b=1}^B (1 - S_b) > B/2\right) \leq \exp\left(-2B\left(\left(1 - \frac{m}{n}\right)^{n_B} - \frac{1}{2}\right)^2\right)\,.$$

Putting the two latter inequalities together gives the desired upper bound on $\mathbb{P}(E_2)$ and concludes the proof. □

We assume in Theorem 2.1 that the K-bMOM algorithm is not too far from the solution, by postulating that it has found the right partition at the beginning of the last 10 iterations. The output of the algorithm will indeed be built from the centroids obtained during these last 10 steps. This assumption seems legitimate, since analyzing the behavior of clustering algorithms in a neighborhood of the optimal solutions, by assuming a "warm start" for

instance, is very classical. We could also assume a warm start by requiring that the initialization procedure has found the right partition, at the price of considering the total number of iterations of K-bMOM instead of the last 10 iterations.

A natural and important question is to ask whether this warm start assumption is realistic in a context where outliers are present in the dataset. Even if we have no theoretical evidence of this fact, we show in our experiments that it is indeed possible to propose an initialization that is robust to the presence of some outliers, in the sense that it produces a high classification accuracy in our different experiment settings. See Section 2.5.2 for details.

Furthermore, Theorem 2.1 provides some guarantee for the use of the K-bMOM algorithm, when the latter is performed using an accurate, robust initialization. This rationale is confirmed by our experiments, that show the good behavior of K-bMOM when initialized by the bMOM strategy coupled with the algorithm K-means++ (see Section 2.6.1 below).

We also assume that the data is in a well-clusterizable configuration with compactness and separation parameters $r$ and $R$ that satisfy $R^2 > 16 n_B r^2$. This condition is surely pessimistic, but it allows us to deduce a non-trivial lower bound for the probabilistic breakdown point with a reasoning that is kept rather simple. The rationale to keep in mind for practice is that if $r$ and $R$ can be well defined, then when the ratio $R/r$ increases, the algorithm is more likely to be robust with respect to the presence of outliers.

We see from Theorem 2.1 that the K-bMOM algorithm has a high probability of being robust if the quantity $p_1$ defined in (2.3.3) is close to 1 and the quantity $p_2$ given by (2.3.4) is close to 0. To analyze $p_1$, note that if the clusters are well-balanced and if the proportion of outliers is not too large, then the quantities $n_k^r/n$ can be approximated by $1/K$. In this case, $p_1$ simplifies to $p_1 \simeq (1 - K\,(1 - 1/K)^{n_B})^{10B}$. We see from this approximation that $p_1$ can indeed be close to 1, even if our theoretical computations may be too pessimistic for being accurate in practice. For instance, if $K = 3$, $n_B = 10K = 30$ and $B = 100$, the value of $p_1$ is greater than 0.98.

Finally, for the quantity $p_2$ to be close to 0, we need the number of blocks $B$ generated in each step to be sufficiently large. This seems to corroborate our empirical conclusions related to the choice of the number of blocks, see Section 2.4.2 below.

## 2.4 Scope of K-bMOM and practical considerations

### 2.4.1 Block length, number of clusters and proportion of outliers

The purpose of this section is to clarify the scope of the K-bMOM algorithm according to the proportion of outliers $m/n$, the number of clusters $K$ and the block size $n_B$.

The reader is reminded that Theorem 2.1 on the probabilistic breakdown point of K-bMOM holds if the probability that a block is not contaminated by outliers is strictly greater than 0.5. This probability takes the value $(1-m/n)^{n_B}$. The block length $n_B$ is therefore dependent on the proportion of outliers $m/n$. Taking a block size $n_B = \gamma K$ with $\gamma \in \mathbb{N}^*$ (it can be seen roughly as the number of data points per cluster), the maximum proportion of outliers for which our proposed approach is robust can be evaluated for a given probability that a block is not contaminated by outliers.

To do so, let us take this probability equal to 0.55, such that $(1-m/n)^{n_B} = 0.55 > 1/2$. According to the previous condition, we get: $m/n = 1 - (0.55)^{1/(\gamma K)}$. Figure 2.4.1 stands for the maximum level of contamination according to the number of clusters for different values of $\gamma \in [1, 10]$.

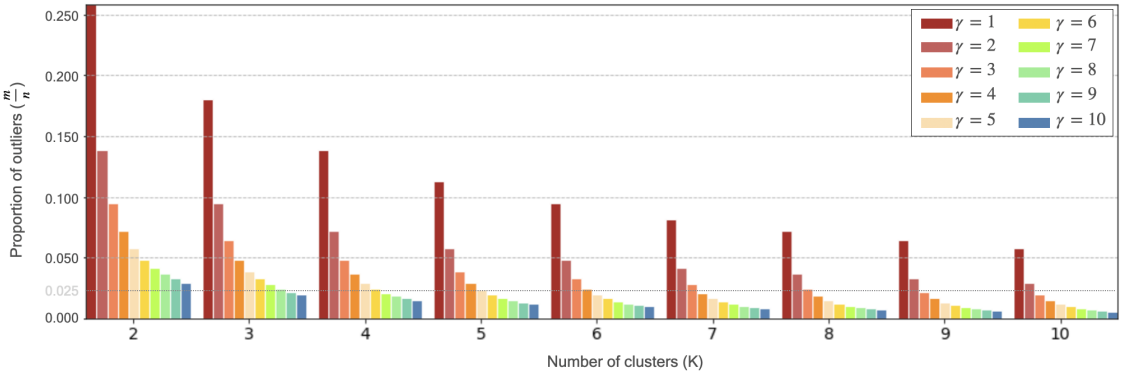



Figure 2.4.1: Evolution of the maximum proportion of outliers allowed for the probability that a block is not contaminated by outliers to be greater than 0.5, according to the number of clusters and the coefficient $\gamma$.

As can be observed, the case $\gamma = 1$ enables us to deal with a contamination level up to 10% for a number of clusters $K < 7$. However, if by chance each cluster is represented in a block, the estimation of centroids is roughly based on one data point only which may lead to inaccurate estimations. On the other hand, by taking $\gamma = 5$ (around 5 data points per cluster), the level of outliers for which the K-bMOM algorithm is robust, has to be low (under 5%

if $K > 2$) but we can expect having an accurate estimation of centroids. Therefore, there is a trade-off in practice between the number of clusters, the proportion of outliers and the accuracy of the centroid estimation.

The second illustration, proposed in Table 2.4.1, evaluates the maximum block size $n_B$ according to a range of probabilities that a block is healthy for different proportions of outliers. It can be noted that the block size $n_B$ dramatically decreases when the percentage of outliers increases. A proportion of outliers up to 0.04 leads to a majority of block sizes containing less than 10 data points. When the number of clusters remains small ($K \in \{2, 3\}$) with a uniform presence in the block ($\gamma \in \{3, 4, 5\}$), the K-bMOM algorithm should behave correctly insuring that the probability that a block is healthy is greater than 1/2 for a proportion of outliers up to 0.4 (see bold values located in the upper left diagonal in Table 2.4.1). However, if the number of groups becomes quite high, e.g. $K = 10$ with at least 5 datapoints per cluster ($\gamma = 5$), the scope of K-bMOM narrows to a proportion of outliers of 0.01 and below, as illustrated by the blue bold values in Table 2.4.1.

| | | Probability that a block is healthy | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.51 | 0.55 | 0.60 | 0.65 | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 | 0.95 |
| Proportion of outliers $m/n$ | 0.001 | **673** | **597** | **510** | **430** | **356** | **287** | **223** | **162** | **105** | **51** |
| | 0.005 | **134** | **119** | **101** | **85** | **71** | **57** | **44** | **32** | **21** | 10 |
| | 0.01 | **66** | **59** | **50** | **42** | **35** | **28** | **22** | **16** | 10 | 5 |
| | 0.02 | **33** | **29** | **25** | **21** | **17** | 14 | 11 | 8 | 5 | 2 |
| | 0.03 | **22** | **19** | **16** | 14 | 11 | 9 | 7 | 5 | 3 | 1 |
| | 0.04 | **16** | 14 | 12 | 10 | 8 | 7 | 5 | 3 | 2 | 1 |
| | 0.05 | 13 | 11 | 9 | 8 | 6 | 5 | 4 | 3 | 2 | 1 |
| | 0.1 | 6 | 5 | 4 | 4 | 3 | 2 | 2 | 1 | 1 | 0 |

Table 2.4.1: Lookup table of the maximum block size $n_B$ evaluated according to a range of proportion of outliers $m/n$ and a range of probabilities that a block is healthy. In bold, the ranges of block sizes for $K = 3$ with $\gamma = 5$. In bold blue, the possible ranges of $n_B$ for $K = 10$ with $\gamma = 5$.

In conclusion, when the number of clusters is small ($K \leq 5$) the K-bMOM algorithm should be robust with respect to a proportion of outliers up to $m/n = 0.03$ with a limited block size ($n_B \simeq 25$ and $\gamma \geq 5$). For a higher number of groups, the K-bMOM algorithm should remain accurate but for a smaller percentage of outliers (below 1%). In practice, this situation should not be too restrictive. Indeed, since an outlier is a data point that differs considerably from all or most other data in a dataset, we do not expect having a large proportion of them in a dataset (in contrast to noisy data). Section 2.6.1 evaluates the performance of the K-bMOM algorithm in different simulation contexts.

### 2.4.2 Influence of the number of blocks

In the previous section, we showed the strong influence of the block size on the maximum proportion of outliers allowed to guarantee a percentage of healthy blocks. In particular, the smaller the size, the more robust the algorithm is to a large proportion of outliers. In this section, we focus on the influence of the second hyper-parameter of the K-bMOM algorithm which is the number of blocks. To do so, we consider a 2-dimensional isotropic Gaussian mixture model of $K = 3$ components with equal size $n_1 = n_2 = n_3 = 300$. The mean vectors are set to $\mu_1 = [3, 12], \mu_2 = [6, 3]$ and $\mu_3 = [-6, 9]$ and the scaling parameter is set to $\sigma^2 = 0.6$. Twenty outliers are randomly selected from the data and their coordinates are multiplied by 10. The block size is set to $5K = 15$ to be robust to any level of outliers (see Table 2.4.1) with a sufficient number of elements per group. The number of blocks varies between 1 block up to 1000 blocks and the process is iterated 100 times. In order to evaluate the influence of the number of blocks on the robustness of the algorithm, three criteria are computed in each context:

- the accuracy computed between the partition obtained by the nearest fitted centers and the labels of regular data.

- the empirical distortion $\hat{R}$ obtained at the end of the studied process and computed over the $(1 - m/n)n$ regular data, such as:

$$\hat{R} = \frac{1}{(1 - m/n)n} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}} \left\| x_i - \bar{c}_k^{(bmed)} \right\|^2 \mathbf{1}\{x_i \in \mathcal{C}_k\}. \tag{2.4.1}$$

We recall that $\bar{\mathbf{c}}^{(bmed)} = (\bar{c}_1^{(bmed)}, \ldots, \bar{c}_K^{(bmed)})$ is the output of the K-bMOM algorithm (see Algorithm 2.1) and $\mathcal{C}_k$ is the corresponding cluster $k$.

The violin plots of these three criteria are illustrated in Figure 2.4.2. The median is shown by a bold black dash. As it can be observed, as of a number of blocks $B \geq 50$, the performance of the algorithm is ensured and remains stable.

These results confirm the choice of a small size of a block (see Section 2.4.1 and Section 2.3.1) and a high number of blocks ($B > 50$) to have a procedure that is likely to be robust. In practice, the values $n_B = 5K$ and $B = 500$ are the default parameters applied in the K-bMOM algorithm when the proportion of outliers is unknown.

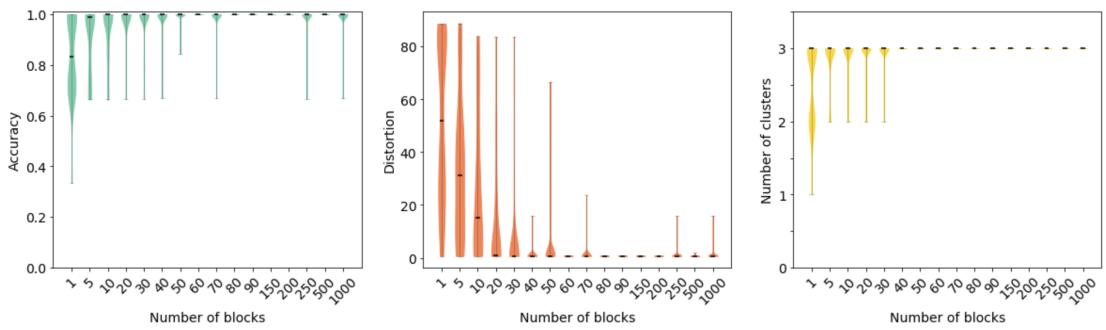



Figure 2.4.2: Violin plots of accuracy (left), distortion (middle), number of clusters (right) computed on the partition of regular data obtained by the K-bMOM algorithm as a function of the number of blocks.

## 2.5 Experimental simulations

This section aims at evaluating the scope of performances of the K-bMOM as an initialization strategy and as a clustering algorithm according to a taxonomy of different types of outliers on one hand and according to several specifications such as sample size, dimension and number of groups on the other hand.

### 2.5.1 Experimental contexts and practical considerations

The same experimental context will be addressed to evaluate the proposed robust initialization and the K-bMOM algorithm. In particular, the different situations considered will depend on the 6 different aspects listed below:

1. the outliers typology. Three types of outliers are considered: isolated multi-directional outliers, isolated oriented outliers and a cluster of outliers.
2. the proportion $m/n$ of outliers contamination,
3. the sample size $n$ of data,
4. the dimension $p$ of data. Let us note that our strategy is not designed to deal with high-dimensional data. Therefore, the number of dimensions tested will remain small and discriminant since we do not deal with the problem of noisy dimensions.
5. the separability of clusters, by varying the level of the scaling parameter $\sigma^2$ of the identity covariance matrix.
6. the number $K$ of clusters.

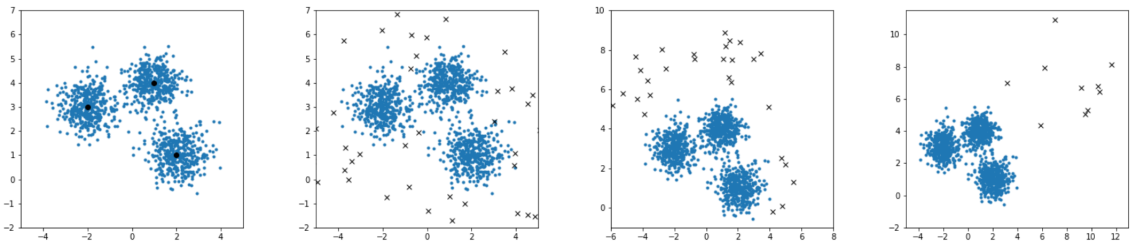
#### 2.5.1.1 Regular data and outliers generation procedures

We generate $n$ data from $K$ multivariate Gaussian distributions of dimension $p$ with equal size $n/K$, isotropic variance $\Sigma = \sigma^2 \mathbf{I}_p$ (with $\sigma^2 > 0$ and $\mathbf{I}_p$ the $p$-dimensional identity matrix) and average vectors $\mu_k$ with $k \in \{1, \ldots, K\}$. Figure 2.5.1.a illustrates one realisation of the simulated context in the case of $K = 3$ groups. Three typologies of outliers are considered and are generated as expressed below:

1. isolated outliers: they are generated uniformly in a parallelogram defined by the coordinatewise ranges of the regular data points. Using an acceptance/rejection algorithm as in [53], only data points having squared Mahalanobis distances from the centers greater than the quantile $\chi^2_{p,0.975}$ are retained and only $m$ of them are going to replace the same number of randomly selected regular data. This case is illustrated in Figure 2.5.1.b.

2. isolated oriented outliers: $m$ regular data points are randomly selected as potential outliers and their coordinates are multiplied by a constant term $\beta$ which quantifies how far these outliers are from their own distribution. Figure 2.5.1.c illustrates such a type of outliers.

3. Cluster of outliers: $m$ regular data points are randomly replaced by a cluster of outliers of the same size generated according to a 2-dimensional Gaussian distribution with average $\mu_{outlier} = \beta[1, 1]$ and $\sigma^2_{outlier}$ as a scaling parameter of the covariance matrix. This situation is depicted in Figure 2.5.1.d.



(a) Regular data (b) Punctual outliers (T1) (c) Punctual oriented outliers (T2) (d) Cluster of outliers (T3)

Figure 2.5.1: Illustrations of simulated regular data (blue points) generated according to a Gaussian Mixture Model with isotropic variance and different types of outliers (black crosses).

#### 2.5.1.2 Experimental values

The experimental values taken for each of these scenarios are detailed below:

| **aspects** | **detailed case** | **values** |
|---|---|---|
| proportion of contamination | | $m/n \in \{0, 0.001, 0.005, 0.01, ..., 0.04\}$ |
| outliers parameters | | $\beta \in \{9, 27\}$, $\sigma^2_{outlier} = 2$ |
| dimension | | $p \in \{2, 5\}$ |
| sample size | | $n \in \{120, 1200, 12000\}$ |
| separability of clusters | (high, medium, low) | $\sigma^2 \in \{0.4, 0.6, 0.8\}$ |
| number of clusters | | $K \in \{3, 5, 10\}$ |

The average vectors according to each configuration of the considered Gaussian mixture model, are defined as it follows:

- $(K = 3, p = 2)$: $\mu_1 = [1, 4], \mu_2 = [2, 1], \mu_3 = [-2, 3]$,
- $(K = 5, p = 2)$: $\mu_1, \mu_2, \mu_3$ and $\mu_4 = [0, -1], \mu_5 = [1, -3]$,
- $(K = 10, p = 2)$: $\{\mu_k\}_{k\in[1,5]}$ and $\mu_6 = [0, 7], \mu_7 = [3, 6], \mu_8 = [5, 1], \mu_9 = [7, 0], \mu_{10} = [8, 4]$,
- $(K = 3, p = 5)$: $\mu_1 = [1, 4, b, a, a], \mu_2 = [2, 1, a, b, a], \mu_3 = [-2, 3, a, a, b]$,
- $(K = 5, p = 5)$: $\{\mu_k\}_{k\in[1,3]}$ and $\mu_4 = [0, -1, b, b, b], \mu_5 = [1, -3, a, a, a]$,
- $(K = 10, p = 5)$: $\{\mu_k\}_{k\in[1,5]}$ and $\mu_6 = [0, 7, a, b, b], \mu_7 = [3, 6, b, a, a]$, $\mu_8 = [5, 1, a, b, a], \mu_9 = [7, 0, b, a, a]$, and $\mu_{10} = [8, 4, a, b, b]$,

with $a = 0$ and $b = -1$. Let us note that for all the experiments, the number of clusters (and the proportion of outliers) is supposed to be known and fixed to its true value $K$ (respectively $m/n$).

**Performance criteria**

In order to compare the different starting strategies in terms of performance, we compute three criteria: (1) the accuracy computed between the partition obtained by the nearest fitted centers and computed on the regular data, (2) the empirical distortion computed on the regular data and defined in Equation 2.4.1 and (3) the Root Mean

Square Error (RMSE) in order to evaluate the robustness of fitted centers. This criterion is calculated between the centers fitted by the studied process and those used to simulate the data:

$$\text{RMSE} = \sqrt{\frac{\sum_{k=1}^{K} \left\| \bar{c}_k^{(bmed)} - \mu_k \right\|^2}{K}}, \tag{2.5.1}$$

where $\bar{c}_k^{(bmed)}$ stands for the fitted center the most probable for the class $k$ averaged on the last 10 iterations and $\mu_k$ the average parameter of the $k$-th component.

For each simulation context, the experiment is repeated 1000 times. These criteria are averaged and standard deviations have been computed for each initialization or clustering approach.

### 2.5.2 Comparing initialization strategies for the clustering task

The experimental context of this section aims at evaluating the scope of performances of the robust initialization process that we propose. In particular, we apply the MOM principle to the most widely used initialization methods amongst which K-means++ and K-medians++ as expressed in Algorithm 2.2 in Section 2.2.2. We consider the following three traditional initialization strategies: Random initialization, K-means++ [6], K-medians++ and also a robust initialization strategy developed by [2] named ROBIN. The implementations that we used in this study for the above approaches come from SCIKIT-LEARN library which is a free software machine learning library for the Python programming language and is publicly available in [118].

The algorithmic complexities of the algorithms named above are $\mathcal{O}(K^2dn)$ (K-means++), $\mathcal{O}(K^2dn)$ (K-medians++), $\mathcal{O}\left(\left(K^2d + B\right) n_B\right)$ (K-bMOM-k-means++), $\mathcal{O}\left(K\right)$ (random) and $\mathcal{O}\left((s + K + nd)\log n\right)$ (ROBIN) where $s$ is the number of neighbours taken into account in the local outlier factor (a.k.a LOF) (see here for details about this complexity: https://towardsdatascience.com/k-nearest-neighbors-computational-complexity-502d2c440d5 and add $K \log n$ for choosing point with respect to there distance to already chosen centroids).

#### 2.5.2.1 Global comments and results

First of all, it is important to notice that the behavior of the six initialization procedures is really stable and comparable in terms of accuracy when the data are not polluted by any kind of outliers as illustrated in Table 2.5.1

(top). As expected, the accuracy decreases when the cluster separability becomes weaker (i.e. when the scaling parameter $\sigma^2$ increases). Moreover, an interesting point is the level of accuracy obtained by the different approaches to estimate the centers of each cluster. It appears that, on regular data, this is the K-bMOM-km++ and K-bMOM-kmed which on average, fits the centers of clusters better than the other methods. See Table 2.5.1 (bottom).

**Accuracies:**

| $\sigma^2$ | random | K-medians++ | K-means++ | ROBIN | K-bMOM-km++ | K-bMOM-kmed |
|---|---|---|---|---|---|---|
| 0.4 | 0.588 (0.028) | **0.995 (0.009)** | **0.995 (0.009)** | 0.979 (0.028) | **0.997 (0.015)** | 0.970 (0.027) |
| 0.6 | 0.531 (0.033) | 0.883 (0.033) | **0.895 (0.035)** | **0.898 (0.039)** | **0.902 (0.042)** | 0.886 (0.048) |
| 0.8 | 0.476 (0.047) | 0.712 (0.047) | 0.711 (0.050) | 0.735 (0.053) | 0.716 (0.053) | 0.702 (0.056) |
| **RMSE:** | | | | | | |
| 0.4 | 1.333 (0.205) | 0.612 (0.164) | 0.614 (0.179) | 0.822 (0.175) | **0.398 (0.093)** | **0.422 (0.147)** |
| 0.6 | 1.528 (0.242) | 0.955 (0.265) | 0.943 (0.254) | 1.160 (0.259) | **0.737 (0.185)** | **0.739 (0.186)** |
| 0.8 | 1.860 (0.275) | 1.267 (0.351) | 1.258 (0.350) | 1.521 (0.362) | **1.096 (0.242)** | **1.105 (0.251)** |

Table 2.5.1: Averages of accuracy (top) and RMSE (bottom) and their standard deviations (in brackets) of initialization procedures on regular data as a function of the separability of clusters.

Table 2.5.2 highlights aggregated performances on the different situations (dimension, separability, number of clusters, etc) of six initialization strategies according to the typology of outliers considered. Median and standard deviations (in brackets) of three metrics are represented. They are computed on all simulated contexts with outliers. First of all, it can be noted that in the case of isolated outliers (T1), all methods (except for the random case) perform quite well on average: their average accuracy remains above 0.83. Secondly, it is worth noting that the traditional approaches are being impacted by oriented isolated outliers (T2) and clustered outliers (T3). In particular, in the case T2, the average accuracy of these approaches are 10% (as K-medians++) to 40% (K-means++) lower than the MOM-based approaches. Moreover, the K-bMOM-km++ RMSE remains smaller compared to the rest of the approaches. Similar results are observed in the case T3.

We are going to focus on the case $\sigma^2 = 0.4$ in order to highlight the benefits and limitations of the proposed initialization approaches, depending on the number of clusters, the percentage and the degree of outliers. Moreover, since the case of isolated outliers (T1) impacts none of the initialization approaches, the specific comments will focus on types of outliers T2 and T3.

| type of outlier | | initialization | RMSE | distortion | accuracy |
|---|---|---|---|---|---|
| T1 | isolated | random | 1.643 (0.370) | 4.307 (1.700) | 0.538 (0.069) |
| | | K-medians++ | 0.934 (0.389) | 1.887 (1.656) | 0.833 (0.137) |
| | | K-means++ | 0.979 (0.405) | 1.752 (1.668) | 0.857 (0.137) |
| | | ROBIN | 1.351 (1.122) | 2.674 (2.273) | 0.847 (0.196) |
| | | K-bMOM-km++ | **0.702 (0.534)** | **1.421 (1.363)** | **0.894 (0.136)** |
| | | K-bMOM-kmed | **0.727 (0.412)** | **1.491 (1.355)** | **0.871 (0.134)** |
| T2 | oriented & isolated | random | **4.155 (5.653)** | 4.652 (1.699) | 0.708 (0.046) |
| | | K-medians++ | 39.53 (39.09) | 7.936 (5.574) | 0.412 (0.189) |
| | | K-means++ | 23.38 (33.49) | 3.458 (2.339) | 0.770 (0.146) |
| | | ROBIN | 15.95 (50.41) | 7.646 (89.79) | 0.635 (0.346) |
| | | K-bMOM-km++ | 6.552 (9.142) | **1.828 (1.491)** | **0.874 (0.085)** |
| | | K-bMOM-kmed | 7.420 (8.819) | **1.972 (1.505)** | **0.849 (0.081)** |
| T3 | cluster of outliers | random | 1.505 (0.360) | 4.157 (1.597) | 0.544 (0.066) |
| | | K-medians++ | 0.842 (0.358) | 1.872 (1.667) | 0.810 (0.152) |
| | | K-means++ | 0.880 (0.360) | 2.472 (1.755) | 0.756 (0.158) |
| | | ROBIN | 1.256 (0.817) | 3.847 (4.067) | 0.694 (0.330) |
| | | K-bMOM-km++ | **0.637 (0.429)** | **1.630 (1.523)** | **0.851 (0.153)** |
| | | K-bMOM-kmed | 0.697 (0.421) | 1.718 (1.522) | 0.800 (0.152) |

Table 2.5.2: Aggregated performances according to the typology of outliers for different strategies of initialization.

#### 2.5.2.2 Specific comments: sensibility to the type of outliers

The three barplots shown in Figure 2.5.2 indicate the average accuracies of the six initialization approaches for different proportions of type T3 outliers. They have been drawn for different number of groups: $K = 3$ (left side), $K = 5$ (middle) and $K = 10$ (right side). In the easiest configuration ($K = 3$), it is worth noting that K-bMOM-km++ and K-bMOM-kmed++ are on average, the most robust approaches to outliers whatever their proportion. Let us note that K-means++ and ROBIN remain accurate until a proportion of outliers equals 0.01 before dropping out. The same robust behavior of K-bMOM-km++ can be observed in the case T2.

## 2.6 Sensitivity of initializations

In order to have a general view of the sensitivity of initialization procedures according to experimental dimensions, an analysis of variance explaining the average accuracy has been done for each procedure. Table 2.6.1 summarizes the effect of parameters on any type of outliers. A star indicates a parameter having an impact on accuracy, corresponding to a p-value under the threshold 0.05 on at least one type of outlier. The modality of the parameter

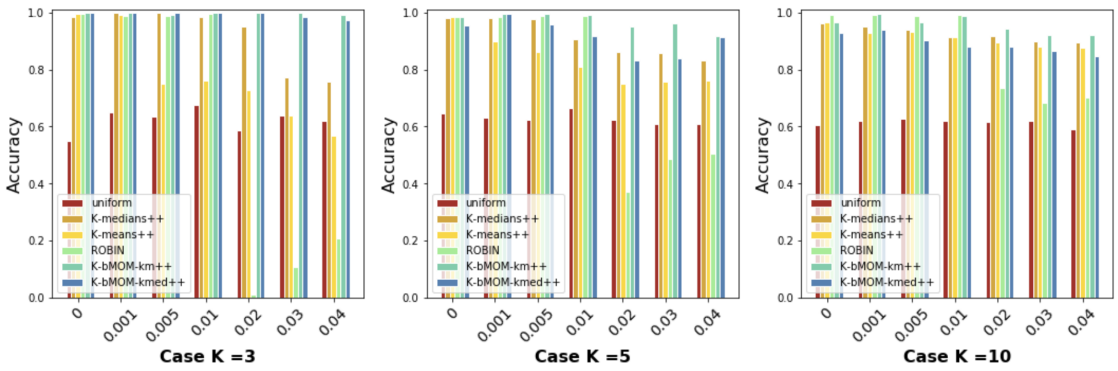



Figure 2.5.2: Case: cluster of outliers (T3). Comparison of the average accuracies per initialization approach depending on the proportion of outliers and the number of clusters ($K = 3$, $K = 5$ and $K = 10$) from the left to the right.

is filled when only its effect has an impact on the accuracy.

| initialization | $n$ | $\sigma^2$ | $m/n$ | $\beta$ | $K$ | $p$ |
|---|---|---|---|---|---|---|
| random | - | * | * | - | * | - |
| K-medians++ | - | * | * | * | * | - |
| K-means++ | - | * | * | * | * | - |
| ROBIN | $(n = 12000)$ | * | * | - | - | - |
| K-bMOM-km++ | - | * | $(m/n > 0.03)$ | - | * | - |
| K-bMOM-kmed | - | * | $(m/n > 0.03)$ | - | * | - |

Table 2.6.1: Summary of the influence of parameters on the accuracy of each initialization method. Note that when a p-value is less than 0.05, a star is indicated. A modality having a negative impact is also filled.

Table 2.6.2 reflects average accuracies of each initialization strategy according to the level of cluster separability $\sigma^2$ and the type of outliers.

The three barplots shown in Figure 2.6.1 indicate the average accuracies of 6 initialization approaches for different proportion of type T2 outliers depending on the number of simulated clusters with $K = 3$ (left side), $K = 5$ (middle) and $K = 10$ (right side) groups respectively when $\sigma^2 = 0.4$.

### 2.6.1 Evaluation of the behaviour of K-bMOM algorithm

The aim of this section is to evaluate how K-bMOM behaves according to the different scenarios detailed in Section 2.5.1. Table 2.6.3 show the mean and the standard deviation (in parenthesis) of RMSE, distortion and accuracy for each type of outlier (T1, T2, T3) with several proportions of outliers. These metrics have been

| | isolated outliers (T1) | | | oriented outliers (T2) | | | cluster of outliers (T3) | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0.4 | 0.6 | 0.8 | 0.4 | 0.6 | 0.8 | 0.4 | 0.6 | 0.8 |
| random | 0.591 | 0.545 | 0.469 | 0.586 | 0.539 | 0.477 | 0.587 | 0.540 | 0.472 |
| K-medians++ | 0.969 | 0.853 | 0.671 | 0.454 | 0.450 | 0.383 | 0.912 | 0.811 | 0.660 |
| K-means++ | 0.987 | 0.870 | 0.680 | 0.825 | 0.731 | 0.595 | 0.817 | 0.747 | 0.617 |
| ROBIN | 0.905 | 0.829 | 0.661 | 0.639 | 0.799 | 0.664 | 0.666 | 0.729 | 0.606 |
| K-bMOM-km++ | 0.989 | 0.887 | 0.711 | 0.953 | 0.867 | 0.697 | 0.958 | 0.840 | 0.673 |
| K-bMOM-kmed | 0.965 | 0.872 | 0.706 | 0.909 | 0.840 | 0.672 | 0.920 | 0.816 | 0.652 |

Table 2.6.2: Average accuracies for each initialization approaches according to the type of outliers and the level of separability.

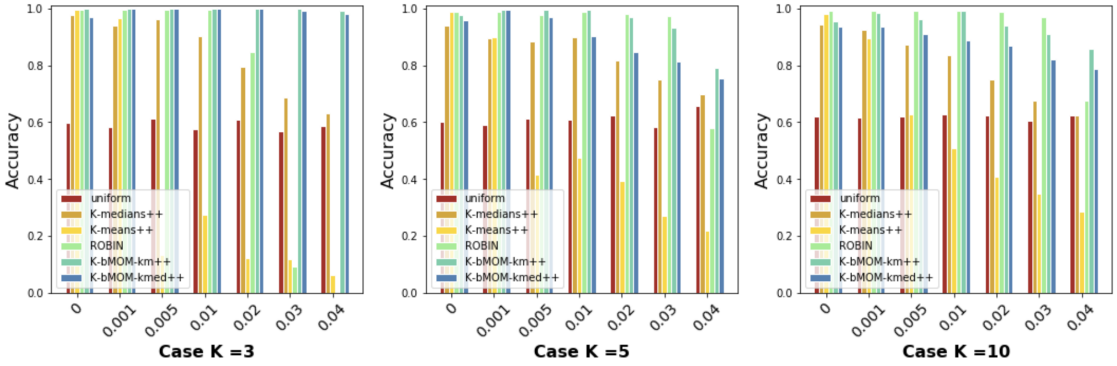



Figure 2.6.1: Case: isolated oriented outliers (T2). Comparison of the average accuracies per initialization approach as a function of the proportion of outliers, the number of clusters ($K = 3$, $K = 5$ and $K = 10$) from the left to the right.

averaged over the different combinations linked to the number of clusters, the dimension of data and the level of cluster separability.

First of all, it can be observed that isolated outliers have almost no influence on the clustering results. On average, the performances remain really close to that obtained on regular data irrespective of the percentage of outliers. This is mainly explained by the distribution of outliers which is well-dispersed around regular data from the isotropic Gaussian mixture model and therefore, it does not affect the clustering task. Secondly, concerning the T2 and T3 type outliers, it can be noted that the accuracy rate remains over 0.90 for a proportion of outliers lower than 3%. This behavior is explained by the limitations seen in Section 2.4.1. Moreover, the K-bMOM algorithm seems to be more robust in the configuration T3 of a cluster of outliers than in the configuration T2 of isolated and oriented outliers, as can be seen on the decrease of accuracies in Table 2.6.3.

The following remarks focus on the cases $\sigma^2 = 0.4$ to visualize the evolution of performances of the K-bMOM

| type of outlier | | $m/n$ | RMSE | distortion | accuracy |
|---|---|---|---|---|---|
| | regular data | 0 | 0.181 (0.086) | 0.913 (0.554) | **0.993 (0.011)** |
| T1 | isolated | 0.001 | 0.186 (0.086) | 0.914 (0.556) | **0.993 (0.010)** |
| | | 0.005 | 0.190 (0.117) | 0.911 (0.554) | **0.992 (0.015)** |
| | | 0.01 | 0.199 (0.099) | 0.904 (0.548) | **0.992 (0.015)** |
| | | 0.02 | 0.205 (0.090) | 0.912 (0.552) | **0.993 (0.011)** |
| | | 0.03 | 0.210 (0.095) | 0.922 (0.557) | **0.991 (0.016)** |
| | | 0.04 | 0.215 (0.099) | 0.930 (0.560) | **0.990 (0.020)** |
| T2 | oriented & isolated | 0.001 | 0.293 (0.366) | 1.297 (2.759) | **0.981 (0.065)** |
| | | 0.005 | 0.299 (0.362) | 1.299 (2.782) | **0.980 (0.063)** |
| | | 0.01 | 0.393 (0.515) | 1.552 (2.860) | **0.937 (0.107)** |
| | | 0.02 | 0.465 (0.575) | 1.808 (2.921) | **0.911 (0.131)** |
| | | 0.03 | 0.674 (0.709) | 2.309 (3.283) | 0.864 (0.146) |
| | | 0.04 | 0.885 (0.713) | 2.518 (2.798) | 0.814 (0.146) |
| T3 | cluster of outliers | 0.001 | 0.186 (0.086) | 0.914 (0.556) | **0.993 (0.010)** |
| | | 0.005 | 0.213 (0.148) | 0.978 (0.571) | **0.984 (0.036)** |
| | | 0.01 | 0.215 (0.151) | 0.959 (0.561) | **0.976 (0.039)** |
| | | 0.02 | 0.260 (0.177) | 1.016 (0.609) | **0.966 (0.055)** |
| | | 0.03 | 0.416 (0.221) | 1.247 (0.694) | 0.898 (0.091) |
| | | 0.04 | 0.480 (0.247) | 1.330 (0.732) | 0.882 (0.089) |

Table 2.6.3: Aggregated performance of K-bMOM as a function of the percentage of outliers for type T1 (isolated), type T2 (oriented and isolated) and type T3 (clustered) of outliers.

algorithm according to the number of clusters, the proportion and the type of outliers.

We consider on the one hand, the case of isolated oriented outliers (T2) that is displayed in Figures 2.6.2, 2.6.3 and 2.6.4. They represent violin plots of accuracies, RMSE and number of clusters fitted by the K-bMOM algorithm for a number of clusters equal to $K = 3$, $K = 5$ and $K = 10$ respectively and for several proportions of outliers. As can be observed, for a low number of groups ($K \in \{3, 5\}$), the K-bMOM algorithm is really robust to a proportion of outliers up to 3%. As expected (see Section 2.4.1), when the number of clusters is high ($K = 10$), the procedure remains robust for a lower proportion of outliers ($m/n \leq 0.005$).

On the other hand, the case of a cluster of outliers (T3) is displayed in Figures 2.6.5, 2.6.6 and 2.6.7. Again, for $K = 3$ and $K = 5$, we can observe that the procedure is really robust and stable for a proportion of outliers lower than 4%. In the case $K = 10$, the K-bMOM algorithm resists very well to the increasing proportion of outliers and appears to be more robust in this typology of outliers than for type T2.

## 2.7 Theoretical analysis of an idealized estimator based on MOM

In Section 2.7.1, we provide some deviation bounds for the performance in terms of the K-means risk of an idealized version of the estimator produced by our algorithm. Indeed, we consider a minimizer of the median-of-means of the K-means loss along possible centroids and call it K-MOM. We prove that K-MOM is robust to adversarial contamination of the dataset if the number of outliers is sufficiently small compared to the number of blocks in the MOM statistics. We also prove in Section 2.7.2, that K-MOM has a non-trivial breakdown point in any clustering configuration, which is a strong result, but note that it does directly affect the practice, because the computation of K-MOM is NP-hard like for the K-means.

### 2.7.1 Convergence rates for the K-MOM estimator

In this section, we aim at analyzing the convergence rates of an idealized version of the estimator produced by the K-bMOM algorithm presented in the main part of the chapter.

We first need to describe the setting. Our study of convergence rates requires a probabilistic framework to model the data. In this section, we thus denote the sample $(X_1, ..., X_n)$, rather than $(x_1, ..., x_n)$ as in the main part of the chapter. We assume that the dataset is made up of two disjoint components, indexed by $\mathcal{I}$ and $\mathcal{O}$ with

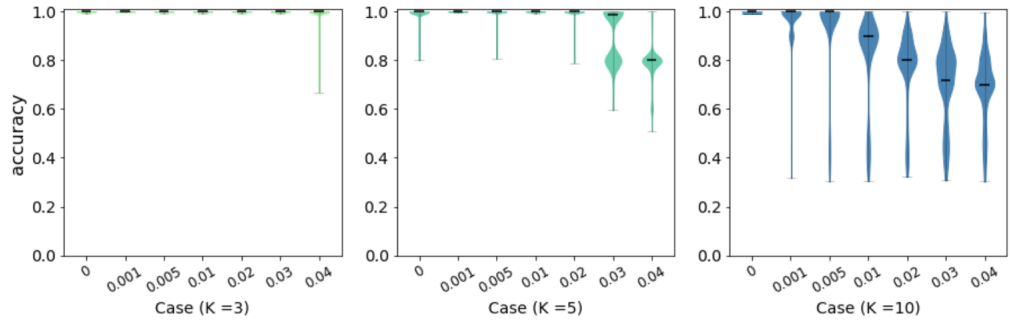



Figure 2.6.2: Violin plots of K-bMOM accuracies computed on the fitted partition among the regular data, for different proportions of T2 outliers, for different sample sizes, outlier degrees and dimensions, for $K = 3$ (left), $K = 5$ (middle) and $K = 10$ (right) with cluster separability $\sigma^2 = 0.4$.

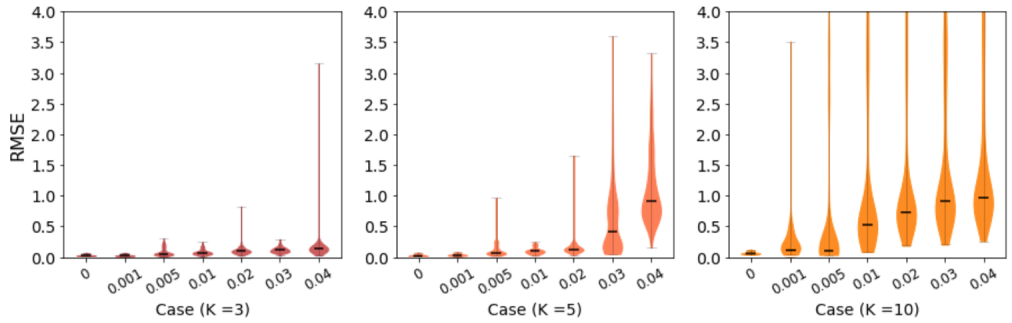



Figure 2.6.3: Violin plots of K-bMOM RMSE computed on the fitted partition among the regular data, for different proportions of T2 outliers, for different sample sizes, outlier degrees and dimensions, for $K = 3$ (left), $K = 5$ (middle) and $K = 10$ (right) with cluster separability $\sigma^2 = 0.4$.

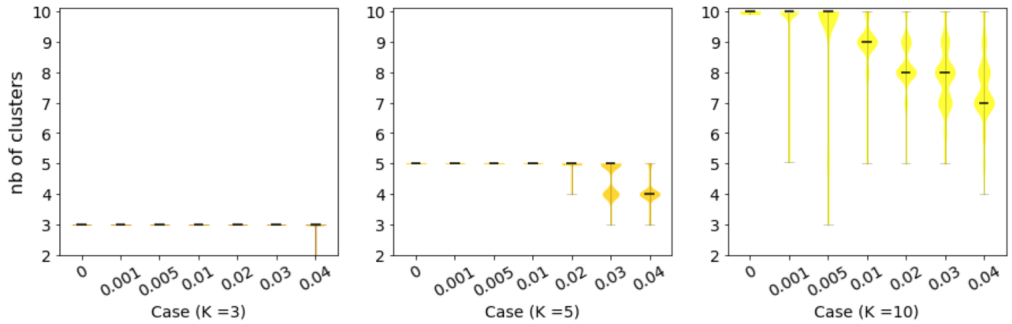



Figure 2.6.4: Violin plots of the K-bMOM number of clusters computed on the fitted partition among the regular data for different proportions of T2 outliers for $K = 3$ (left), $K = 5$ (middle) and $K = 10$ (right) with cluster separability $\sigma^2 = 0.4$, all scenarios included.

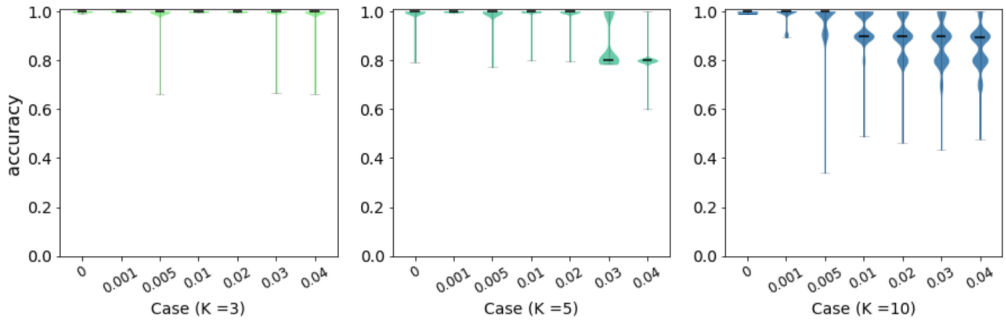



Figure 2.6.5: Violin plots of K-bMOM accuracies obtained on the fitted partition among the regular data, for different proportion of T3 outliers, for different sample sizes, outlier degrees and dimensions for $K = 3$ (left), $K = 5$ (middle) and $K = 10$ (right) with cluster separability $\sigma^2 = 0.4$.

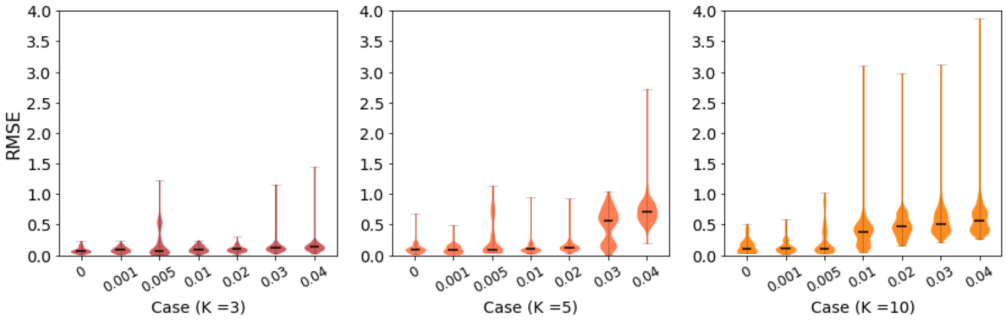



Figure 2.6.6: Violin plots of K-bMOM RMSE obtained on the fitted partition among the regular data, for different proportions of T3 outliers, for different sample sizes, outlier degrees and dimensions for $K = 3$ (left), $K = 5$ (middle) and $K = 10$ (right) with cluster separability $\sigma^2 = 0.4$.

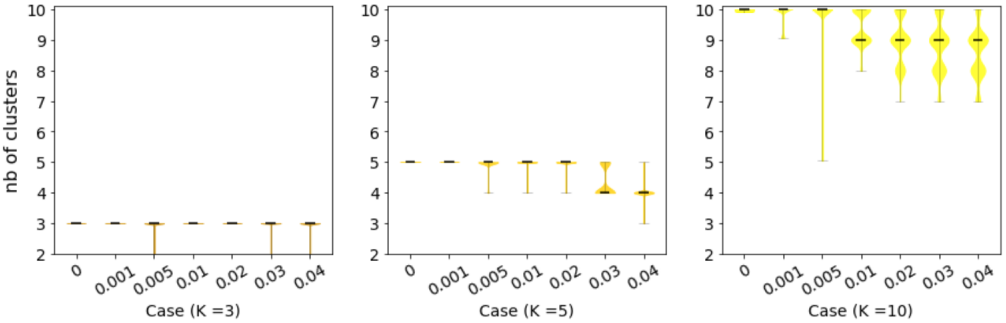



Figure 2.6.7: Violin plots of the number of clusters obtained on the fitted partition for different proportion of T3 outliers, among different sample sizes, outlier degrees and dimensions for $K = 3$ (left), $K = 5$ (middle) and $K = 10$ (right) with cluster separability $\sigma^2 = 0.4$.

$\mathcal{I} \bigcup \mathcal{O} = \{1, ..., n\}$ and $\mathcal{I} \bigcap \mathcal{O} = \emptyset$: the set of regular data $(X_i)_{i \in \mathcal{I}}$, corresponding to data that provide information and are not corrupted, and the set of outliers $(X_i)_{i \in \mathcal{O}}$, that may be completely misleading for the clustering task. The random variables $X_i$, $i = 1, ..., n$, take values in the Euclidean space $\mathbb{R}^p$ and the regular data $(X_i)_{i \in \mathcal{I}}$ are independent and identically distributed random variables. No assumption is made on the behavior of the outliers $(X_i)_{i \in \mathcal{O}}$, that may have infinite moments and depend on the regular data $(X_i)_{i \in \mathcal{I}}$. We thus study the robustness against *adversarial contamination*. Being robust in such a context indeed prevents "attacks" by a malicious adversary that would generate outliers after having seen the regular data.

We also set a generic random variable $X$, independent of the sample and of the same distribution $P$ of the regular data $X_i$, $i \in \mathcal{I}$.

Actually, it is worth noting that the results presented here in Sections 2.7.1 and 2.7.2 are valid, without modifications of the proofs, for data points belonging to general, separable Hilbert spaces rather than the Euclidean space $\mathbb{R}^p$. In particular, in the proof of Theorem 2.2 below, we use some classical results obtained in [15] for Hilbert spaces.

For any set of centroids $\mathbf{c} = \{c_1, ..., c_K\}$, also called a codebook, we denote by $\ell_{\mathbf{c}}$ a loss function on $\mathbb{R}^d$ such that $\ell_{\mathbf{c}}(x) = \min_{k=1,...,K} \left\{-2 \langle x, c_k \rangle + \|c_k\|^2\right\}$, where $\langle \cdot, \cdot \rangle$ is the scalar product associated with the Euclidean norm $\|\cdot\|$ on $\mathbb{R}^p$. Notice that $\min_{k=1,...,K} \|x - c_k\|^2 = \|x\|^2 + \ell_{\mathbf{c}}(x)$. The loss $\ell_{\mathbf{c}}$ is thus classically associated with the K-means procedure ([15]).

For any function $f$, denote $Pf := \mathbb{E}[f(X)]$. For the K-means problem to make sense, it is necessary to assume the existence of a finite second moment for the regular data, $P\|X\|^2 < \infty$. The goal is to find from the sample $(X_1, ..., X_n)$, a codebook $\hat{\mathbf{c}}$ that is close to an optimal one, denoted by $\mathbf{c}_*$, minimizing the K-means distortion risk,

$$\begin{aligned} \mathbf{c}_* &\in \arg \min_{\mathbf{c} \in \mathbb{R}^{p \times K}} \{P\ell_{\mathbf{c}}\} \\ &= \arg \min_{\mathbf{c} = \{c_1, ..., c_K\} \in \mathbb{R}^{p \times K}} \left\{ \mathbb{E}\left[ \min_{k=1,...,K} \|X - c_k\|^2 \right] \right\}. \end{aligned}$$

Also denote $\ell_* = \ell_{\mathbf{c}_*}$ the optimal distortion risk. Note that an optimal set of centroids always exists but may not be unique (see for instance [50]).

We assume furthermore, that the magnitude of an optimal codebook is known. This means that there exists a constant $M_* > 0$ such that, $\max_{k=1,...,K} \|c_{*,k}\| \leq M_*$, where $\mathbf{c}_* = (c_{*,1}, ..., c_{*,K})$. We may thus restrict our search within codebooks $\mathbf{c} = (c_1, ..., c_K)$ satisfying $\max_{k=1,...,K} \|c_k\| \leq M_*$. This assumption is rather natural (in practice,

in general we do not want to have centroids too far away) and is also considered in [81, Section 2].

Hence, we set

$$\hat{\mathbf{c}} \in \arg\min_{\mathbf{c}\in\mathcal{B}_{M_*}^K} \left\{ \mathrm{MOM}\left( (\ell_{\mathbf{c}}(X_i))_{i=1}^n, I_1^B \right) \right\}, \tag{2.7.1}$$

a codebook minimizing the median-of-means of the loss along the data, where $\mathcal{B}_{M_*} = \{x \in \mathcal{X}; \|x\| \leq M_*\}$ is the ball of radius $M_*$ in $\mathbb{R}^d$ and we recall that

$$\mathrm{MOM}\left( (\ell_{\mathbf{c}}(X_i))_{i=1}^n, I_1^B \right) = \mathrm{med}\left\{ \frac{1}{|I_b|} \sum_{i\in I_b} \ell_{\mathbf{c}}(X_i) : b \in \{1, ..., B\} \right\},$$

for a partition $I_1^B$ of the set of indices $\{1, ..., n\}$. For simplicity, we assume that the blocks have the same length $n_B$, so that $n = Bn_B$. In the minimization problem (2.7.1), the blocks are chosen once and for all and the quantities $\mathrm{MOM}\left( (\ell_{\mathbf{c}}(X_i))_{i=1}^n, I_1^B \right)$ are computed using the same blocks.

We consider that the K-bMOM algorithm, presented in Section 2.2 of the main part of the chapter, is an approximation to the minimization task defined in (2.7.1). Our algorithm indeed iteratively computes codebooks in a Lloyd-type fashion in each block of data and then at each step, keeps the codebook that achieves the median of the K-means distortion risk in each block.

Note also that in (2.7.1), we consider the "classical" MOM, instead of the bootstrap MOM. Considering a bMOM however, with the same block length and number of blocks as a MOM, should give rather similar performances. The point in using the MOM statistics is that its mathematical analysis is simpler than for bMOM, since the blocks of MOM are disjoint and so, independent. Consequently, empirical process techniques are available.

The estimators given by (2.7.1) have been studied recently in [81, Section 2], where they are proved to achieve sub-Gaussian performance bounds under a finite second moment assumption for the random variable $X$. In our result below, we take a different route by studying robustness against adversarial contamination, under the hypothesis that regular data are uniformly bounded. In the framework of robust supervised learning under adversarial contamination, Lecué and Lerasle [86] also studied estimators of the form of (2.7.1), but with different losses.

Let $O$ denote the set of indices of the blocks that contain at least one outlier and $J$ denote the set of indices of the blocks that are not corrupted, i.e. that do not contain any outlier. We thus have $|O| \leq m$, where $m$ is the number of outliers and $|O|$ denotes the cardinal of $O$, and $|J| \geq B - m$.

Denote also $R(\mathbf{c}) = P\ell_{\mathbf{c}}$, the risk of a codebook $\mathbf{c}$ and $R_* = P\ell_*$ the best possible risk, that is the minimum value of the risk over all possible codebooks $\mathbf{c} \in \mathcal{B}_{M_*}^K$. For an estimator $\hat{\mathbf{c}}$, we give probabilistic bounds on the quantity $R(\hat{\mathbf{c}}) - R_*$, also known as the excess risk.

**Theorem 2.2.** *If there exists a real number $M_X > 0$ such that $\|X\| \leq M_X$ a.s. and if the number $m$ of outliers satisfies $m \leq B/4$, then two numerical constants $h_1, h_2 > 0$ exist such that the following holds, with probability greater than $1 - 2\exp(-h_1 B)$,*

$$R(\hat{\mathbf{c}}) - R_* \leq h_2 \max\left\{ \frac{K\left[M\sqrt{\mathbb{E}\left[\|X\|^2\right]} + M^2/2\right]}{\sqrt{n}}, M\sqrt{\frac{B\mathbb{E}\left[\|X\|^2\right]}{n}} \right\}, \tag{2.7.2}$$

*where $M = \max\{M_*, M_X\}$. It can be seen from the proof given in Section 2.10 below that $h_1 = 3/64$ and $h_2 = 512$ work.*

The proof of Theorem 2.2 can be found in Section 2.10 below.

Note that in Theorem 2.2 we assume that the regular data are defined in a bounded domain and robustness is considered through the fact that there may be outliers in the dataset. If the number of outliers is sufficiently small compared to the number of blocks ($m \leq B/4$), the upper bound given in (2.7.2) for the excess K-means distortion risk, is composed of two terms. The first term in the maximum appearing at the right-hand side of (2.7.2), corresponds to the classical convergence rate of the K-means for a sample that is uniformly bounded and that do not contain any outlier, see [15]. It has been proved that in the no-contamination, bounded setting, the correct dependence in $K$ in the convergence rate is actually of the order $\sqrt{K}$ - up to logarithmic factors - rather than $K$, see [49] and [107]. This brings in many technicalities however, and we leave the optimal dependence in $K$ in our setting as an interesting open question.

The second term in the maximum appearing on the right-hand side of (2.7.2) reflects the price to pay for the presence of outliers. In particular, it does not change the rate of convergence of the no-contamination setting if the number of blocks $B$ is on the order of $K^2$. In such a case, estimators given by (2.7.1) are robust to adversarial contamination in the sense that their estimation rates are not impacted by the presence of outliers.

### 2.7.2 Breakdown point of the K-MOM estimator

In this section, we analyze the breakdown point of the K-MOM estimator. Compared to the definition given in (2.7.1) above, we make a slight change by assuming that optimizers are searched for in the whole space $\mathbb{R}^p$. In essense this does not affect the breakdown point study, but provides lighter reasoning and notations in the proofs. The estimator can be formulated as

$$\tilde{\mathbf{c}} \in \arg\min_{\mathbf{c}\in\mathbb{R}^{p\times K}} \left\{ \text{MOM}\left( (\ell_{\mathbf{c}}(x_i))_{i=1}^n, I_1^B \right) \right\}, \tag{2.7.3}$$

where $\ell_{\mathbf{c}}$ is the standard K-means loss (see Section 2.7.1 above), $(x_1, ..., x_n)$ is the dataset and $I_1^B$ forms a partition of the set $\{1, ..., n\}$. In the minimization problem (2.7.3), the blocks are chosen once and for all and the quantities $\text{MOM}(\ell_{\mathbf{c}})$ are computed using the same blocks. The dataset is denoted $x_1^n$ and in this section, is considered to be fixed. For simplicity, we assume that the blocks have the same length $n_B$, so that $n = Bn_B$, where $B$ is the number of blocks taken in the MOM estimates.

**Definition 2.6. Breakdown point for K-MOM** The breakdown point $\delta_n(\tilde{\mathbf{c}}, x_1^n)$ of the set of $K$ centroids $\tilde{\mathbf{c}}$ defined in (2.7.3), given the sample $x_1^n$, is the maximum proportion of outliers that leaves the norms of the centroids bounded,

$$\delta_n(\tilde{\mathbf{c}}, x_1^n) = \frac{1}{n} \max \left\{ m; \max_{i_1,...,i_m} \sup_{y_1,...,y_m} \max_{k\in\{1,...,K\}} \|\tilde{c}_k(z_1, ..., z_n)\| < \infty \right\},$$

where the sample $(z_1, ..., z_n)$ is obtained by replacing the $m$ data points $x_{i_1}, ..., x_{i_m}$ of the sample $x_1^n$ by arbitrary values $y_1, ..., y_m$ and $\tilde{\mathbf{c}}(z_1, ..., z_n) = (\tilde{c}_1(z_1, ..., z_n), ..., \tilde{c}_K(z_1, ..., z_n))$ is a K-MOM estimator based on the sample $(z_1, ..., z_n)$:

$$\tilde{\mathbf{c}}(z_1, ..., z_n) \in \arg\min_{\mathbf{c}\in\mathbb{R}^{p\times K}} \left\{ \text{MOM}\left( (\ell_{\mathbf{c}}(z_i))_{i=1}^n, I_1^B \right) \right\}. \tag{2.7.4}$$

**Theorem 2.3.** *Let the K-MOM estimator $\tilde{\mathbf{c}}$ be defined in (2.7.3) using $B$ blocks. For any dataset $x_1^n$, its breakdown point is given by*

$$\delta_n(\tilde{\mathbf{c}}, x_1^n) = \frac{\left\lfloor \frac{B-1}{2} \right\rfloor}{n}. \tag{2.7.5}$$

*Proof.* Let us first prove the lower bound: $\delta_n(\tilde{\mathbf{c}}, x_1^n) \geq \lfloor (B-1)/2 \rfloor / n$. Set $r > 0$ such that all regular data are in a ball $B(0, r)$ centered at the origin and of radius $r$, that is: $\max_{i=1,...,n} ||x_i|| \leq r$. If the number of outliers is less than or equal to $\lfloor (B-1)/2 \rfloor$, then a majority of blocks do not contain any outlier, so the median values that give the

quantities MOM $\left((\ell_{\mathbf{c}}(z_i))_{i=1}^n, I_1^B\right)$ are sandwiched in the risk values of some regular blocks. Consider a set of centroids $\mathbf{c}_0$ such that $\max_{k=1,...,K} ||c_{0,k}|| \leq r$. The risks the of the regular blocks for this set of centroids, are less than $4r^2$, so MOM $\left((\ell_{\mathbf{c}_0}(z_i))_{i=1}^n, I_1^B\right) \leq 4r^2$. Now, consider a set of centroids $\mathbf{c}_1$ such that $\max_{k=1,...,K} ||c_{1,k}|| > (2\sqrt{n_B}+1)r$. It is easy then to see that the risks of the regular blocks for this set of centroids, are strictly greater than $4r^2$, so MOM $\left((\ell_{\mathbf{c}_0}(z_i))_{i=1}^n, I_1^B\right) > 4r^2$ and $\mathbf{c}_1$ is not a solution of the optimization problem (2.7.4). Consequently, any solution of (2.7.4) satisfies $\max_{k=1,...,K} ||c_{1,k}|| \leq (2\sqrt{n_B}+1)r$, which proves the lower bound.

Let us now prove the upper bound: $\delta_n(\tilde{\mathbf{c}}, x_1^n) \leq \lfloor (B-1)/2 \rfloor / n$. Consider that the number of outliers is strictly larger than $\lfloor (B-1)/2 \rfloor$ and that a majority of blocks contains at least one outlier. Take all the outliers to be equal to a vector $y$ and set a constant $r > 0$ such that $\max_{i=1,...,n} ||x_i|| \leq r$. We will show that the procedure breaks down when $y$ diverges to infinity. We proceed by *reductio ad absurdum*. Let us assume that there exists a constant $t > 0$ such that for any value $y$ of the outliers, there exists a codebook $\mathbf{c}(y) = \{c_1(y), ..., c_K(y)\}$, being a solution of (2.7.4) for the contaminated sample $(z_1, ..., z_n)$ and depending on $y$, such that $\max_{k=1,...,K} ||c_k(y)|| \leq t$. Then if $||y|| > t$, the risk on each contaminated block is greater than the value $(||y|| - t)^2/n_B$ due to the contribution of $y$. As $y$ is contained in a majority of blocks, we also have MOM$((\ell_{(\mathbf{c}(y)}(z_i))_{i=1}^n, I_1^B) \geq (||y|| - t)^2/n_B$. Now, take other codebooks $\check{\mathbf{c}}(y) = (\check{c}_1(y), ..., \check{c}_K(y))$ where $\max_{k=1,...,K-1} ||\check{c}_k(y)|| \leq r$ and $\check{c}_K(y)) = y$. We get MOM$((\ell_{\check{\mathbf{c}}(y)}(z_i))_{i=1}^n, I_1^B) \leq 4r^2 < (||y|| - t)^2/n_B$, for $||y||$ sufficiently large, which gives a contradiction with $\mathbf{c}(y)$ being a solution of (2.7.4). The upper bound is thus proven, which concludes the proof. □

Theorem 2.3 is based on the fact that for a smaller amount of outliers than the value of the breakdown point in (2.7.5), a majority of blocks do not contain any outlier. Hence the medians along the blocks stay bounded for bounded centroids, whatever the values of the outliers.

According to Theorem 2.3, the K-MOM procedure (2.7.3) has the strong feature of achieving a non-trivial breakdown point *in any clustering configuration*, as opposed to other robust K-means procedures such as for instance, the trimmed K-means, that requires a so-called well-clusterizable configuration where clusters are sufficiently "compact" and sufficiently "separated" ([52] ; see also [125] and references therein).

K-MOM as defined in (2.7.3) has however, an essentially theoretical interest, since the minimization problem is intractable in general.

## 2.8 Benchmark of K-means type robust clustering algorithms

The objective of this section is to compare the performance of the K-bMOM strategy in its scope of application with the robust clustering algorithms based on K-means approaches on a specific framework with isolated oriented outliers.

### 2.8.1 Benchmark algorithms

We consider six different algorithms: our proposed robust clustering algorithm named K-bMOM with a time complexity $\mathcal{O}(Kn_BBp)$ at each iteration, the traditional K-means for comparison and also four well-known robust versions of the K-means described below:

- K-medoids which aims at finding K data points as centers such that the within inertia is minimized. This optimization process is done according to the Partition Around Medoids algorithm named PAM [79]. It has a complexity dominated by $\mathcal{O}\left(K(n-K)^2p\right)$ per iteration. Faster versions have been proposed in [130].

- K-medians which is a robust variant of the K-means algorithm [72] : in the aggregation step, instead of computing the barycenter of each group as in the K-means procedure, K-medians computes in each single dimension, the median based on the Manhattan-distance formulation. The complexity of such a procedure is dominated by $\mathcal{O}(nKp)$ per iteration as for Loyd's algorithm [67].

- trimmed-K-means which is an EM-like algorithm introduced by Cuesta et al. [36] in the late 90s. It benefits robustness properties from the trimming action during the maximisation step where only a proportion $1-m/n$ of the closest data point from their assigned centroid, is taken into account. Since the trimming needs to sort the data points according to their distance to centroid, it leads therefore to an overall complexity of $\mathcal{O}(Knp+n\log n)$ at each iteration.

- K-PDTM which is a robust quantization algorithm introduced by Brecheteau et al. [22], aims to infer the manifold from which the data points are drawn. This inference is done by means of $K$ centroids that should be on the manifold if the algorithm runs well. It is based also on a Lloyd-type algorithm where in the updating step, the centroid is computed as the barycenter of the $s$ nearest neighbours of the barycenter of the cluster. In the assignment step, the data point is assigned according to a Bregman divergence. This algorithm has two hyper-parameters: $s$, the number of neighbors used to compute the centroid and the number of clusters $K$. This leads to the following

complexity for one iteration: $\mathcal{O}(Ksp + n\log n)$.

The implementations used for the clustering approaches to compare the MOM-based ones in this experiment are publicly available. Table 2.8.1 details the programming languages and associated libraries used as well as selected hyper-parameters. Let us note that the trimming approach from the TRIMCLUSTER [68] R package has been implemented in Python language in order to be able to play easily with the initial conditions.

| Algorithm | Language | hyper-parameters |
|---|---|---|
| K-means | Python [118] | $K$, $init$ =given*,$n\_init = 1$ |
| K-medoids | Python [115] | $initial\_index\_medoids$ = given* |
| K-medians | Python [115] | $initial\_centers$ =given* |
| trimK-means | R [36, 68] | $K$, $trim = m/n$, $runs = n$, $points$ =given*,$maxit = 300$ |
| K-pdtm | Python [22, 23] | $K$, $query\_pts$ =given*,$s = K$, $k = K$, $sig = N - m$, $iter\_max = 300$, $nstart = 1$,$leaf\_size = 30$ |
| K-bMOM | Python [25] | $K$, $n_B = 5K$, $B = 500$, $iter\_max = 25$, $initial\_centers$ =given* |

*given: same centers obtained either with a random initialization or K-bMOM-km++

Table 2.8.1: Implementations and hyper-parameters

### 2.8.2 Simulation context

We dispose of $N = 1500$ points of dimension $p = 3$ which are generated according to a mixture of $K \in \{3, 4, 5\}$ multivariate Gaussian density functions with an isotropic covariance matrix. The average vectors for the $K$ components are respectively $\mu_1 = [0, 1, 4]$, $\mu_2 = [2, 1, 0]$, $\mu_3 = [0, -2, 3]$, for $K = 3$, $\mu_4 = [0, 5, -5]$ is added in the case $K = 4$, then $\mu_5 = [-1, -2, 0]$ for $K = 5$. Isolated outliers have been generated by randomly taken 30 datapoints from the regular dataset and their coordinates have been multiplied by a factor of $\pm 10$. All the algorithms have been initialized with the exact same conditions: a random initialization, a K-means++ strategy and a K-bMOM-km++ strategy presented in Section 2.5.2. These conditions have been repeated 1000 times and in order to compare the performances of these algorithms, the RMSE, the distortion and the accuracy have been computed based on the true parameters of data distribution and their label membership amongst the regular data. Moreover, the average number of clusters found amongst the regular data have also been computed.

### 2.8.3 Results and Analysis

First of all, let us note that the clustering algorithms have been executed on regular data with the same random initializations. As can be seen in Table 2.8.2, all the methods perform equally well in average and whatever the number of clusters, when there is no outlier.

| | Average performances on regular data ($m/n = 0$) | | | | | |
|---|---|---|---|---|---|---|
| | **K-means** | **K-pdtm** | **trimK-means** | **K-medians** | **K-medoids** | **K-bMOM** |
| RMSE | 0.071 (0.367) | 0.068 (0.384) | 0.071 (0.366) | 0.726 (0.379) | 0.173 (0.579) | 0.200 (0.03) |
| distortion | 1.096 (0.671) | 1.093 (0.911) | 1.144 (0.821) | 1.700 (1.241) | 1.120 (1.505) | 1.111 (0.024) |
| accuracy | 0.995 (0.111) | 0.995 (0.127) | 0.995 (0.111) | 0.989 (0.088) | 0.995 (0.121) | 0.996 (0.001) |

Table 2.8.2: Average performances and standard deviations (in brackets) of K-means and the five robust clustering approaches on regular data with the same random initializations, among $K \in \{3, 4, 5\}$.

| K | init | **K-means** | **K-pdtm** | **trimK-means** | **K-medians** | **K-medoids** | **K-bMOM** |
|---|---|---|---|---|---|---|---|
| 3 | 1 | 0.819 (0.162) | 0.938 (0.151) | 0.938 (0.156) | 0.915 (0.082) | 0.864 (0.161) | 0.943 (0.123) |
| | 2 | 0.613 (0.124) | 0.931 (0.131) | 0.833 (0.124) | 0.612 (0.123) | 0.613 (0.124) | 0.810 (0.163) |
| | 3 | **0.994 (0.006)** | 0.942 (0.123) | **0.997 (0.001)** | 0.980 (0.059) | **0.995 (0.003)** | **0.997 (0.002)** |
| 4 | 1 | 0.869 (0.124) | 0.889 (0.142) | 0.869 (0.124) | 0.903 (0.116) | 0.839 (0.120) | 0.879 (0.124) |
| | 2 | 0.502 (0.001) | 0.916 (0.116) | 0.915 (0.056) | 0.502 (0.001) | 0.502 (0.001) | 0.958 (0.001) |
| | 3 | **0.988 (0.049)** | **0.989 (0.021)** | **0.988 (0.049)** | **0.984 (0.0498)** | **0.988 (0.049)** | **0.988 (0.049)** |
| 5 | 1 | 0.820 (0.110) | 0.860 (0.155) | 0.900 (0.114) | 0.896 (0.104) | 0.854 (0.113) | **0.961 (0.075)** |
| | 2 | 0.534 (0.093) | 0.894 (0.141) | 0.777 (0.134) | 0.520 (0.097) | 0.534 (0.093) | **0.996 (0.002)** |
| | 3 | 0.965 (0.069) | 0.898 (0.132) | **0.996 (0.002)** | 0.971 (0.058) | 0.984 (0.042) | **0.996 (0.002)** |

Table 2.8.3: Average accuracies and standard deviations of clustering approaches depending on the number of clusters and the initialization scheme with type (1) random initialization, type (2) K-means++ and type (3) K-bMOM-km++.

The average accuracies calculated for the different number of groups are summarized in Table 2.8.3. As can be seen, the initialization scheme affects most of the clustering approaches and the more sensitive ones are the K-means, K-medians and K-medoids procedures with a K-means++ initialization. The average accuracies of K-bMOM and trimK-means decreases slightly when initialized by a K-means++ procedure. On the other hand, K-pdtm seems to be quite insensitive to the tested initialization schemes.

The proposed K-bMOM-km++ strategy appears to be a very good initialization strategy and K-bMOM is also a sensible strategy when only purely random initializations are considered. These two advantages are reflected in the detailed analysis of the case $K = 5$ where the average RMSE, distortions and accuracies are summarized in Table 2.8.4. Again, we can observe the strong stability of K-bMOM performances in this configuration whatever

the initialization process. The average accuracy is about 0.99 and the average RMSE and distortions are the lowest amongst the tested procedure.

| initialization | methods | RMSE | distortion | accuracy | nb K |
|---|---|---|---|---|---|
| random initialization | K-means | 1.076 (0.264) | 2.560 (0.680) | 0.820 (0.110) | 4.1 (0.5) |
| | K-pdtm | 0.508 (0.491) | 1.993 (1.199) | 0.860 (0.155) | **4.9 (0.3)** |
| | trimK-means | 0.389 (0.385) | 1.577 (0.662) | 0.900 (0.114) | **4.8 (0.3)** |
| | K-medians | 0.993 (0.559) | 2.498 (1.569) | 0.896 (0.104) | 4.7 (0.4) |
| | K-medoids | 1.068 (0.32) | 2.535 (1.042) | 0.854 (0.113) | 4.7 (0.4) |
| | K-bMOM | **0.290 (0.236)** | **1.284 (0.375)** | **0.961 (0.075)** | **4.9 (0.2)** |
| K-means++ initialization | K-means | 1.669 (0.271) | 4.881 (1.09) | 0.534 (0.093) | 2.6 (0.5) |
| | K-pdtm | 0.385 (0.393) | 1.742 (0.977) | 0.884 (0.141) | **5.0 (0.0)** |
| | trimK-means | 0.624 (0.452) | 1.836 (1.437) | 0.857 (0.134) | 4.9 (0.3) |
| | K-medians | 1.828 (0.785) | 7.650 (2.454) | 0.520 (0.097) | 2.6 (0.4) |
| | K-medoids | 1.686 (0.449) | 5.128 (1.203) | 0.534 (0.093) | 2.6 (0.4) |
| | K-bMOM | **0.186 (0.034)** | **1.105 (0.025)** | **0.996 (0.002)** | **5.0 (0.0)** |
| K-bMOM-km++ initialization | K-means | 0.882 (0.122) | 1.891 (0.286) | 0.965 (0.069) | 4.8 (0.3) |
| | K-pdtm | 0.427 (0.429) | 1.743 (0.871) | 0.888 (0.132) | 4.9 (0.2) |
| | trimK-means | **0.063 (0.011)** | **1.067 (0.019)** | **0.996 (0.002)** | **5.0 (0.0)** |
| | K-medians | 0.738 (0.228) | 1.638 (0.365) | 0.971 (0.058) | 4.9 (0.3) |
| | K-medoids | 0.819 (0.154) | 1.721 (0.205) | 0.984 (0.042) | **5.0 (0.0)** |
| | K-bMOM | 0.185 (0.025) | **1.096 (0.022)** | **0.996 (0.002)** | **5.0 (0.0)** |

Table 2.8.4: Averages and standard deviations of the performances of the five robust clustering algorithms on polluted data with a random initialization (top), a K-means++ initialization (middle) and the proposed robust initialization (bottom) in the case $K = 5$.

### 2.8.4 Further experimental results

The advantages of the k-bMOM procedure alone (whatever the initial process) as well as the proposed K-bMOM-km++ initial strategy, are reflected in Figures 2.8.1, 2.8.2, 2.8.3 and 2.8.4 for different initialization schemes in the case $K = 5$: random, Kmeans++ and K-bMOM-km++ (left, middle, right). These violin plots indicate the distribution of the accuracy (Figure 2.8.1), the RMSE (Figure 2.8.2), the distortion (Figure 2.8.3) and the number of unique groups found amongst the regular data (Figure 2.8.4) of each robust procedure. The median is depicted by a bold black dash and the average by a thin black dash.

Moreover, the detailed average performances in the cases $K = 3$ and $K = 4$ are summarized in Table 2.8.5 and Table 2.8.6.

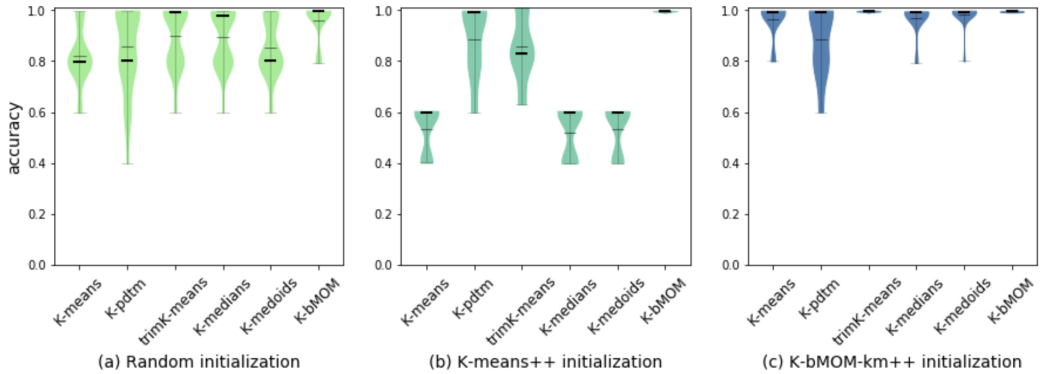



Figure 2.8.1: Violin plots of accuracies according to different initialization strategies in the case $K = 5$.

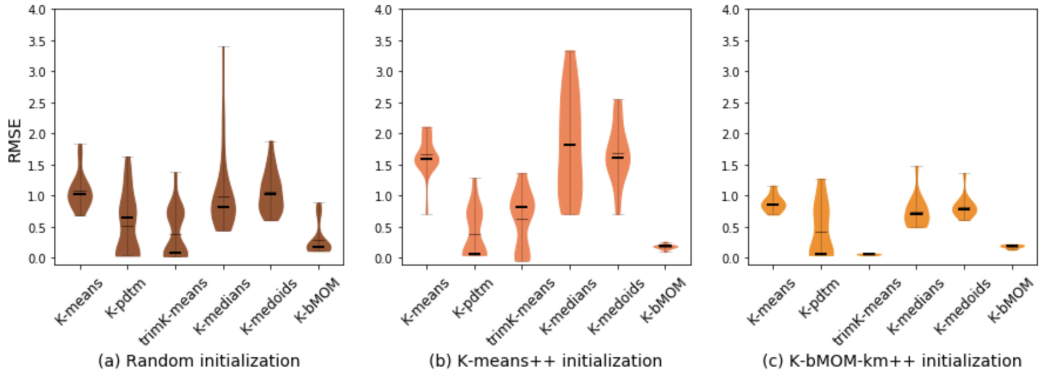



Figure 2.8.2: Violin plots of RMSE according to different initialization strategies in the case $K = 5$.

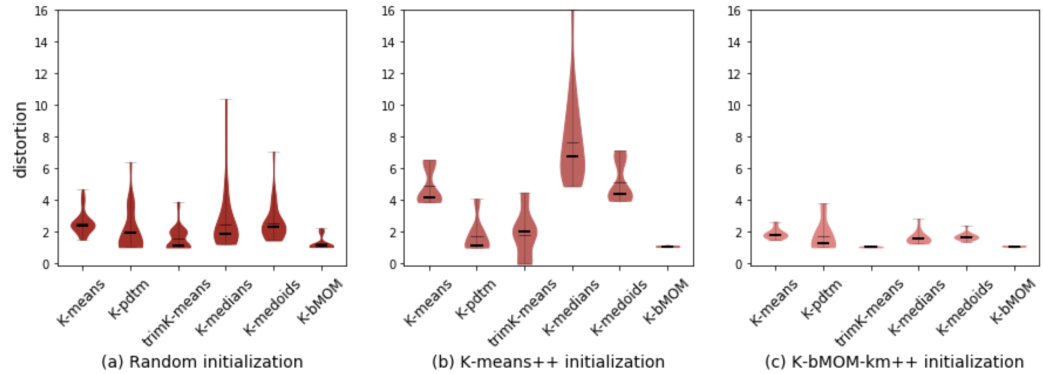



Figure 2.8.3: Violin plots of distortions according to different initialization strategies in the case $K = 5$.

| context | methods | RMSE | distortion | accuracy | nb K |
|---|---|---|---|---|---|
| random initialization | K-means | 0.949 (0.252) | 2.281 (0.705) | 0.819 (0.162) | 2.4 (0.5) |
| | K-pdtm | **0.371 (0.502)** | 1.554 (0.742) | **0.938 (0.151)** | 2.9 (0.2) |
| | trimK-means | **0.379 (0.467)** | 1.612 (0.764) | **0.938 (0.156)** | 2.9 (0.2) |
| | K-medians | 0.711 (0.299) | 1.670 (0.479) | 0.915 (0.082) | 2.9 (0.2) |
| | K-medoids | 0.997 (0.43) | 2.197 (0.877) | 0.864 (0.161) | 2.9 (0.3) |
| | K-bMOM | **0.32 (0.332)** | **1.367 (0.589)** | **0.943 (0.123)** | **3.0 (0.0)** |
| K-means++ initialization | K-means | 1.380 (0.539) | 3.489 (1.495) | 0.613 (0.124) | 1.8 (0.3) |
| | K-pdtm | **0.254 (0.414)** | **1.403 (0.645)** | **0.931 (0.131)** | **3.0 (0.0)** |
| | trimK-means | 1.359 (0.564) | 3.446 (1.512) | 0.833 (0.124) | 2.4 (0.3) |
| | K-medians | 1.778 (1.018) | 5.036 (2.200) | 0.612 (0.123) | 1.8 (0.3) |
| | K-medoids | 1.374 (0.621) | 3.577 (1.568) | 0.613 (0.124) | 1.8 (0.3) |
| | K-bMOM | 0.776 (0.536) | 2.189 (0.947) | 0.810 (0.163) | 2.4 (0.4) |
| K-bMOM-km++ initialization | K-means | 0.661 (0.079) | 1.499 (0.103) | 0.994 (0.006) | **3.0 (0.0)** |
| | K-pdtm | 0.816 (3.448) | 1.343 (0.618) | 0.942 (0.123) | **3.0 (0.0)** |
| | trimK-means | **0.050 (0.012)** | **1.071 (0.021)** | **0.997 (0.001)** | **3.0 (0.0)** |
| | K-medians | 0.710 (0.271) | 1.683 (0.593) | 0.980 (0.059) | 2.9 (0.2) |
| | K-medoids | 0.660 (0.071) | 1.499 (0.087) | 0.995 (0.003) | **3.0 (0.0)** |
| | K-bMOM | 0.176 (0.045) | **1.098 (0.025)** | **0.997 (0.002)** | **3.0 (0.0)** |

Table 2.8.5: Average and standard deviations of the performances of K-means and the five robust algorithms with different initialization strategies in the case $K = 3$.

| context | methods | RMSE | distortion | accuracy | nb K |
|---|---|---|---|---|---|
| random initialization $(m/n = 0.2)$ | K-means | **0.478 (0.404)** | **1.678 (0.594)** | **0.869 (0.124)** | **4.0 (0.0)** |
| | K-pdtm | **0.400 (0.431)** | **1.738 (1.048)** | **0.889 (0.142)** | **4.0 (0.0)** |
| | trimK-means | 0.478 (0.404) | **1.678 (0.593)** | **0.869 (0.124)** | **4.0 (0.0)** |
| | K-medians | 0.896 (0.326) | 2.342 (1.149) | **0.903 (0.116)** | **4.0 (0.0)** |
| | K-medoids | 0.759 (0.661) | 2.407 (1.889) | **0.839 (0.120)** | **4.0 (0.0)** |
| | K-bMOM | **0.510 (0.370)** | **1.661 (0.599)** | **0.879 (0.124)** | **4.0 (0.0)** |
| K-means++ initialization | K-means | 1.735 (0.210) | 5.416 (0.048) | 0.502 (0.001) | 2.0 (0.0) |
| | K-pdtm | **0.323 (0.393)** | 1.477 (0.571) | 0.916 (0.116) | 3.9 (0.2) |
| | trimK-means | 1.704 (0.224) | 1.251 (0.451) | 0.915 (0.056) | 3.9 (0.1) |
| | K-medians | 1.990 (0.755) | 8.007 (1.622) | 0.502 (0.001) | 2.0 (0.0) |
| | K-medoids | 1.810 (0.345) | 5.726 (0.135) | 0.502 (0.001) | 2.0 (0.0) |
| | K-bMOM | **0.176 (0.020)** | **1.101 (0.021)** | **0.998 (0.001)** | **4.0 (0.0)** |
| K-bMOM-km++ initialization | K-means | **0.084 (0.161)** | **1.126 (0.240)** | **0.988 (0.049)** | **4.0 (0.0)** |
| | K-pdtm | 0.378 (0.399) | 1.558 (0.592) | **0.989 (0.121)** | **4.0 (0.0)** |
| | trimK-means | **0.084 (0.161)** | **1.126 (0.240)** | **0.988 (0.049)** | **4.0 (0.0)** |
| | K-medians | 0.774 (0.208) | 1.712 (0.350) | **0.984 (0.048)** | **4.0 (0.0)** |
| | K-medoids | **0.168 (0.169)** | **1.142 (0.243)** | **0.988 (0.049)** | **4.0 (0.0)** |
| | K-bMOM | **0.201 (0.152)** | **1.151 (0.240)** | **0.988 (0.049)** | **4.0 (0.0)** |

Table 2.8.6: Average and standard deviations of the performances of K-means and the five robust algorithms with different initialization strategies in the case $K = 4$.

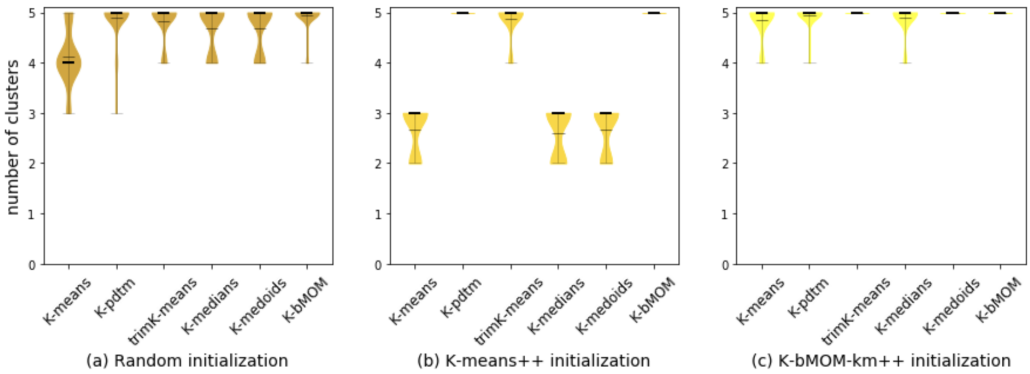



Figure 2.8.4: Violin plots of the number of clusters computed on the partition returned by K-means and the five robust algorithms among regular data according to different initialization strategies in the case $K = 5$.

## 2.9 Color quantization in image processing

We have seen in Section 2.5.2 that using a robust initialization even in the context of few outliers, is the best strategy in terms of resulting partition stability and accuracy. Given that spirit, in this last experimental section, the K-bMOM procedure is applied to the problem of color quantization addressed in image processing and computer graphics.

Color quantization (CQ) is a procedure commonly used for color analysis, image compression, segmentation, non-photorealistic rendering, etc. It is a process which aims to reduce the number of colors used in an image with the goal of keeping the same quality of visualisation as the original. CQ is a challenging problem since most real-world images contain tens of thousands of colors. CQ can be viewed as a 3-dimensional clustering problem according to the Red, Green, Blue channels of pixels of an image. A wide literature is devoted to this problem and it appears that the K-means algorithm is not used so often mainly because of its sensitivity to the initialization. We propose therefore to use the K-bMOM procedure as a robust CQ process providing confident and high-quality quantization on a noisy image.

### 2.9.1 Images and experimental setup

The K-bMOM algorithm has been used on a popular 24-bit test image, the Parrots, (768×512) coming from the Kodak Lossless True Color Image Suite database and illustrated in Figure 2.9.1(a).:

The K-bMOM procedure has been used on the grayscale image of parrots as shown in 2.9.1(b). from which 1% of

(a) Original image (b) Gray scale image (c) Outliers (1%)

Figure 2.9.1: Original Image (left), grayscale image (middle) and a filter of colored outliers applied on the grayscale image (right).

randomised pixels have retained their color from the original image. In our experiment, we consider these colored pixels, illustrated in Figure 2.9.1(c)., as outliers of the grayscale image. The objective is therefore to extract the main shades of gray from the image. The image of size $768 \times 512 \times 3$ has been shaped into a 3-dimensional matrix. The K-bMOM algorithm has been repeated 50 times for a number $K$ of gray levels (or clusters) equals to 8, 16 and 24 respectively. For these three segmentations, the number of blocks has been set to $B = 1000$ and the size of each block set to $n_B = 5 * K$.

### 2.9.2 Experimental results

In order to evaluate the quality of the quantization, the empirical distortion has been computed between the pixels $x_i \in \mathbb{R}^3$, $i \in \{1, \ldots, n\}$, of the original grayscale image and their segmented version $\bar{c}_k^{(bmed)}$ returned by the K-bMOM procedure, i.e. their nearest color. It has been averaged amongst 50 repetitions and the standard deviation has also been computed. Moreover, in order to evaluate the robustness property of the K-bMOM algorithm, the number of gray values amongst the centroids is displayed. It is expected to have $K$ levels of gray (i.e. no color amongst the centroids). In order to benchmark the K-bMOM algorithm, the traditional K-means algorithm has been executed under the same experimental conditions.

The results are summarized in Table 2.9.1. First of all, it can be noted that color quantization processed by the K-bMOM approach seems to be robust. Indeed, the centroids are in the shades of gray: the number of gray levels equals the number of clusters.

Figure 2.9.2 illustrates the quantization process on a grayscale Parrot image with outliers for $K = 8$, $K = 16$ and $K = 24$ respectively. Figures 2.9.3 and 2.9.4 show the error per pixel in a reverse grayscale mapping for the

| | | $K=8$ | $K=16$ | $K=24$ |
|---|---|---|---|---|
| K-bMOM | distortion | 171.0 (1.69) | **56.40** (27.01) | **28.70** (3.42) |
| | number of gray levels | 8.00 (0.00) | **16.00** (0.58) | **24.00** (0.00) |
| K-means | distortion | 171.5 (1.53) | 67.20 (17.77) | 42.50 (2.36) |
| | number of gray levels | 8.00 (0.00) | 15.00 (0.55) | 22.00 (0.52) |

Table 2.9.1: Median and standard deviation (in parenthesis) of the distortion and the number of gray levels obtained by the K-bMOM procedure for $K = \{8, 16, 24\}$ groups.

K-bMOM and K-means procedures respectively. The higher the error, the darker the pixel. It can be seen that the K-bMOM approach performs well in allocating $K-$representative gray levels to the different image regions while the K-means procedure fails, by selecting pixels of outliers as centers.

## 2.10 Proof of Theorem 2.2

Assume without loss of generality, that $B \geq 8$ (otherwise the bound stated in Theorem 2.2 may occur with probability zero). Denote for short $\mathrm{MOM}(\ell_{\mathbf{c}})$ instead of $\mathrm{MOM}\left((\ell_{\mathbf{c}}(X_i))_{i=1}^n, I_1^B\right)$, since the median-of-means are always taken here on the values of the K-means loss on the data $(X_1, ..., X_n)$ and according to the fixed partition $I_1^B$. Also denote $P_b = (1/|I_b|)\sum_{i\in I_b}\delta_{X_i}$ the empirical measure associated to the block indexed by $b$.

We have, by definition of $\hat{\mathbf{c}}$, for any constant $a > 0$,

$$
\begin{aligned}
&\mathbb{P}\left(R\left(\hat{\mathbf{c}}\right) - R_* > a\right)\\
\leq&\mathbb{P}\left(\inf_{\mathbf{c}\in\mathcal{F}_{>a}} \mathrm{MOM}\left(\ell_{\mathbf{c}}\right) \leq \inf_{\mathbf{c}\in\mathcal{F}_a} \mathrm{MOM}\left(\ell_{\mathbf{c}}\right)\right)\\
=&\mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{F}_{>a}} \left\{R_* - \mathrm{MOM}\left(\ell_{\mathbf{c}}\right)\right\} \geq \sup_{\mathbf{c}\in\mathcal{F}_a} \left\{R_* - \mathrm{MOM}\left(\ell_{\mathbf{c}}\right)\right\}\right)\\
\leq&\mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{F}_{>a}} \left\{R_* - \mathrm{MOM}\left(\ell_{\mathbf{c}}\right)\right\} \geq R_* - \mathrm{MOM}\left(\ell_{\mathbf{c}_*}\right)\right),
\end{aligned}
$$

where $\mathcal{F}_a = \left\{\mathbf{c} \in \mathcal{B}_{M_*}^K : R(\mathbf{c}) - R_* \leq a\right\}$ and $\mathcal{F}_{>a} = \left\{\mathbf{c} \in \mathcal{B}_{M_*}^K : R(\mathbf{c}) - R_* > a\right\} = \mathcal{B}_{M_*}^K \setminus \mathcal{F}_a$ . Now, on the one

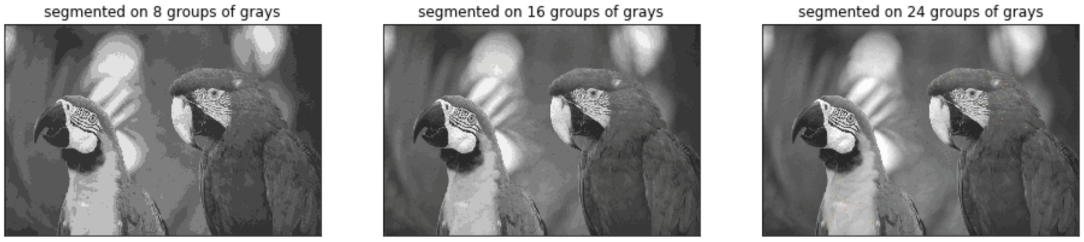



Figure 2.9.2: Sample quantization results for $K = 8$, 16 and 24 respectively from left to right on grayscale noisy Parrot image.

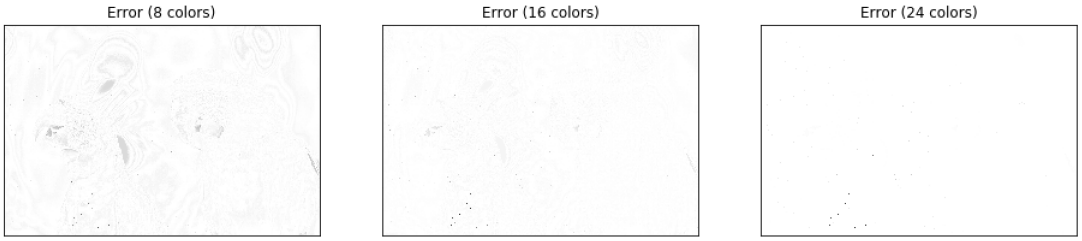



Figure 2.9.3: Full scale error images for $K = 8$, 16 and 24 respectively from left to right grayscale Parrot image with K-bMOM procedure.

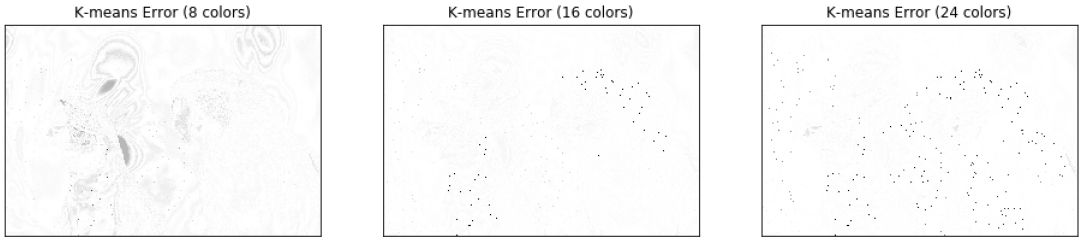



Figure 2.9.4: Full scale error images for $K = 8$, 16 and 24 respectively from left to right grayscale Parrot image with a K-means procedure.

hand, for any $x > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\mathrm{MOM}\left(\ell_{\mathbf{c}_*}\right) - R_* \geq x\right) \\
=&\mathbb{P}\left(\sum_{b=1}^{B} \mathbf{1}_{\{(P_b - P)(\ell_{\mathbf{c}_*}) \geq x\}} \geq \frac{B}{2}\right) \\
\leq&\mathbb{P}\left(\sum_{b \in J} \mathbf{1}_{\{(P_b - P)(\ell_{\mathbf{c}_*}) \geq x\}} \geq \frac{B}{2} - |O|\right) \\
\leq& \sum_{b=\lfloor B/2 - |O| \rfloor}^{B} \left( \begin{array}{c} B \\ j \end{array} \right) p^b (1-p)^{B-b} \\
&\leq p^{\lfloor B/2 - |O| \rfloor} 2^B
\end{aligned}$$

where $p = \mathbb{P}\left(\left(P_b - P\right)\left(\ell_{\mathbf{c}_*}\right) \geq x\right)$ for any $b \in \{1, ..., B\}$. In addition, by the Markov inequality,

$$p \leq \frac{B\mathrm{Var}\left(\ell_{\mathbf{c}_*}\right)}{nx^2}.$$

Hence, by choosing $x = \sqrt{64eB\mathrm{Var}\left(\ell_{\mathbf{c}_*}\right)/n}$, we get

$$\mathbb{P}\left(\mathrm{MOM}\left(\ell_{\mathbf{c}_*}\right) - R_* \geq x\right) \leq 2^B \left(\frac{1}{64e}\right)^{\lfloor B/2 - |O| \rfloor}.$$

Note that since $|O| \leq B/4$ and $B \geq 8$, we have $\lfloor B/2 - |O| \rfloor \geq \lfloor B/4 \rfloor \geq B/8$ and $2^B \leq 16^{\lfloor B/4 \rfloor + 1} \leq 64^{\lfloor B/4 \rfloor}$. This gives

$$\mathbb{P}\left(\mathrm{MOM}\left(\ell_{\mathbf{c}_*}\right) - R_* \geq x\right) \leq \exp\left(-\frac{B}{8}\right).$$

On the other hand,

$$\mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{R_* - \mathrm{MOM}\left(\ell_{\mathbf{c}}\right)\right\} \geq -x\right)$$

$$\leq \mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{B}\sum_{b=1}^{B}\mathbf{1}_{\{R_*-P_b(\ell_{\mathbf{c}})\geq -x\}}\right\} \geq \frac{1}{2}\right)$$
$$\leq \mathbb{P}\left(\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|J|}\sum_{b\in J}\mathbf{1}_{\{R_*-P_b(\ell_{\mathbf{c}})\geq -x\}}\right\} \geq \frac{B}{2|J|} - \frac{|O|}{|J|}\right)$$

Let us denote $\Delta = B/(2|J|) - |O|/|J|$ and set

$$Z\left(\mathcal{F}_{>a}, x\right) = \sup_{\mathbf{c}\in\mathcal{F}_{>a}}\frac{1}{|J|}\sum_{b\in J}\mathbf{1}_{\{R_*-P_b(\ell_{\mathbf{c}})\geq -x\}}. \tag{2.10.1}$$

By Corollary 2.1 (see Section 2.10.1 below), we have

$$\mathbb{P}\left(Z\left(\mathcal{F}_{>a}, x\right) \geq \Delta\right) \leq \exp\left(-\frac{|J|(\Delta - \mathbb{E}\left[Z\left(\mathcal{F}_{>a}, x\right)\right])^2}{2\mathbb{E}\left[Z\left(\mathcal{F}_{>a}, x\right)\right] + 2(\Delta - \mathbb{E}\left[Z\left(\mathcal{F}_{>a}, x\right)\right])/3}\right). \tag{2.10.2}$$

It remains to control $\mathbb{E}\left[Z\left(\mathcal{F}_{>a}, x\right)\right]$. Consider a function $\phi : \mathbb{R} \to \mathbb{R}$, such that $\phi(t) = (t-1)1_{\{1\leq t\leq 2\}} + 1_{\{t\geq 2\}}$. The function $\phi$ is thus 1-Lipschitz and it holds $\phi(t) \geq 1_{\{t\geq 2\}}$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[Z\left(\mathcal{F}_{>a}, x\right)\right] &= \mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|J|}\sum_{b\in J}\mathbf{1}_{\{(P-P_b)(\ell_{\mathbf{c}})\geq R(\mathbf{c})-R_*-x\}}\right\}\right] \\
&\leq \mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|J|}\sum_{b\in J}\mathbf{1}_{\{(P-P_b)(\ell_{\mathbf{c}})\geq a-x\}}\right\}\right] \\
&\leq \mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|J|}\sum_{b\in J}\phi\left(\frac{2\left(P-P_b\right)\left(\ell_{\mathbf{c}}\right)}{a-x}\right)\right\}\right]
\end{aligned} \tag{2.10.3}$$

Now, for any $b \in J$,

$$\mathbb{E}\left[\phi\left(\frac{2\left(P-P_b\right)\left(\ell_{\mathbf{c}}\right)}{a-x}\right)\right] \leq \mathbb{P}\left[\left(P-P_b\right)\left(\ell_{\mathbf{c}}\right) \geq (a-x)/2\right] \leq \frac{BL}{n\left(a-x\right)^2},$$

where the constant $L$ is such that $\sup_{\mathbf{c}} \operatorname{Var}\left(\ell_{\mathbf{c}}\right) \leq L$. More explicitly, we can choose $L = 16M^2\mathbb{E}\left[\|X\|^2\right]$. Hence, by Inequality (2.10.3) we get,

$$\mathbb{E}\left[Z\left(\mathcal{F}_{>a}, x\right)\right] \leq \frac{BL}{n\left(a-x\right)^2} + \mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|J|}\sum_{b\in J}\phi\left(\frac{2\left(P-P_b\right)\left(\ell_{\mathbf{c}}\right)}{a-x}\right) - \mathbb{E}\left[\phi\left(\frac{2\left(P-P_b\right)\left(\ell_{\mathbf{c}}\right)}{a-x}\right)\right]\right\}\right].$$

Now, by a standard symmetrisation argument, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|J|}\sum_{b\in J}\phi\left(\frac{2\left(P-P_b\right)\left(\ell_{\mathbf{c}}\right)}{a-x}\right) - \mathbb{E}\left[\phi\left(\frac{2\left(P-P_b\right)\left(\ell_{\mathbf{c}}\right)}{a-x}\right)\right]\right\}\right] \\
&\leq 2\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|J|}\sum_{b\in J}\epsilon_b\phi\left(\frac{2\left(P-P_b\right)\left(\ell_{\mathbf{c}}\right)}{a-x}\right)\right\}\right],
\end{aligned}$$

where the $\epsilon_b$'s are i.i.d. Rademacher variables (i.e. $\mathbb{P}(\epsilon_b = 1) = \mathbb{P}(\epsilon_b = -1) = 1/2$) independent of the sample. Furthermore, as the function $\phi$ is 1-Lipschitz and $\phi(0) = 0$, we can apply the so-called contraction principle ([87, Section 4]), which gives

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|J|}\sum_{b\in J}\epsilon_b\phi\left(\frac{2\left(P-P_b\right)\left(\ell_{\mathbf{c}}\right)}{a-x}\right)\right\}\right] \\
&\leq \frac{2}{a-x}\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|J|}\sum_{b\in J}\epsilon_b\left(P-P_b\right)\left(\ell_{\mathbf{c}}\right)\right\}\right]
\end{aligned}$$

and by symmetrisation again,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\frac{1}{|J|}\sum_{b\in J}\epsilon_b\left(P-P_b\right)\left(\ell_{\mathbf{c}}\right)\right\}\right] \\
&\leq \frac{2B}{|J|n}\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\sum_{i\in\mathcal{J}}\epsilon_i\ell_{\mathbf{c}}\left(X_i\right)\right\}\right],
\end{aligned}$$

where $\mathcal{J} = \bigcup_{b\in J} I_b$. By Lemma 4.3 in [15],

$$\mathbb{E}\left[\sup_{\mathbf{c}\in\mathcal{F}_{>a}}\left\{\sum_{i\in\mathcal{J}}\epsilon_i\ell_{\mathbf{c}}\left(X_i\right)\right\}\right] \leq 2K\sqrt{|\mathcal{J}|}\left[M\sqrt{\mathbb{E}\left[\|X\|^2\right]}+M^2/2\right].$$
$$\leq 2K\sqrt{n}\left[M\sqrt{\mathbb{E}\left[\|X\|^2\right]}+M^2/2\right]$$

Putting things together, we obtain

$$\mathbb{E}\left[Z\left(\mathcal{F}_{>a},x\right)\right] \leq \frac{BL}{n\left(a-x\right)^2}+\frac{8B}{\left(a-x\right)|J|\sqrt{n}}2K\left[M\sqrt{\mathbb{E}\left[\|X\|^2\right]}+M^2/2\right].$$

Now, by taking

$$a \geq \max\left\{2x, 4\sqrt{\frac{BL}{n\Delta}}, \frac{128BK\left[M\sqrt{\mathbb{E}\left[\|X\|^2\right]}+M^2/2\right]}{\Delta|J|\sqrt{n}}\right\}, \tag{2.10.4}$$

we get

$$\frac{BL}{n\left(a-x\right)^2} \leq \frac{\Delta}{4}$$

and

$$\frac{8B}{\left(a-x\right)|J|\sqrt{n}}2K\left[M\sqrt{\mathbb{E}\left[\|X\|^2\right]}+M^2/2\right] \leq \frac{\Delta}{4}.$$

This gives $\mathbb{E}\left[Z\left(\mathcal{F}_{>a},x\right)\right] \leq \Delta/2$ and so, by using Inequality (2.10.6),

$$\mathbb{P}\left(Z\left(\mathcal{F}_{>a},x\right)\geq\Delta\right) \leq \exp\left(-\frac{3|J|\Delta}{16}\right).$$

To conclude, it suffices now to notice that if $m \leq B/4$, then $|O| \leq B/4$, $|J| \geq 3B/4$ and $\Delta \geq B/(4|J|) \geq 1/4$. Indeed, in this case, Inequality (2.10.4) is achieved by choosing for instance

$$a = \max\left\{8\sqrt{\frac{eBL}{n}}, 512\frac{K\left[M\sqrt{\mathbb{E}\left[\|X\|^2\right]}+M^2/2\right]}{\sqrt{n}}\right\}.$$

### 2.10.1 A concentration inequality

We will derive here a concentration inequality for the stochastic process $Z(\mathcal{F}_{>a}, x)$ defined in (2.10.1). We proceed by proving that $Z(\mathcal{F}_{>a}, x)$ satisfies the so-called self-bounding condition, we recall now (see also [18, Theorem 6.12]).

**Definition 2.7.** A function $f$ is said to have the *self-bounding property* if, for some functions $f_i : \mathcal{Z}^{d-1} \to \mathbb{R}$, for all $z = (z_1, ..., z_d) \in \mathcal{Z}^d$ and for all $i = 1, ..., d$,

$$0 \leq f(z) - f_d\left(z^{(d)}\right) \leq 1$$

and

$$\sum_{i=1}^{d} \left(f(z) - f_i\left(z^{(i)}\right)\right) \leq f(z) \ ,$$

where $z^{(i)} = (z_1, ..., z_{i-1}, z_{i+1}, ...., z_n)$.

**Lemma 2.1.** *If $\mathcal{A}$ is a class of sets over a measurable space $(\mathcal{Z}, \mathcal{T})$, then the function $h : \mathcal{Z}^d \to \mathbb{R}$ defined by*

$$h(z_1, ..., z_d) = \sup_{A \in \mathcal{A}} \sum_{j=1}^{d} 1_A(z_j) \ ,$$

*has the self-bounding property. Consequently, if $(\xi_1, ..., \xi_d) \in \mathcal{Z}^d$ is an i.i.d. sample, then by setting $Z = h(\xi_1, ..., \xi_d)$, it holds for any $t > 0$,*

$$\mathbb{P}(Z \geq \mathbb{E}Z + t) \leq \exp\left(-\frac{t^2}{2\mathbb{E}Z + 2t/3}\right) \ . \tag{2.10.5}$$

*Proof.* Denote $h_i\left(z^{(i)}\right) = \sup_{A \in \mathcal{A}} \sum_{j \neq i} 1_A(z_j)$ . Then

$$0 \leq h(z) - h_i\left(z^{(i)}\right) \leq \sup_{A \in \mathcal{A}} 1_A(z_i) \leq 1 \ .$$

Also, assume without loss of generality that $h(z) = \sum_{j=1}^{d} 1_{A_*(z)}(z_j)$ for some $A_*(z) \in \mathcal{A}$, then

$$\begin{aligned}\sum_{i=1}^{d}\left(h(z) - h_i\left(z^{(i)}\right)\right) &\leq \sum_{i=1}^{d}\left(\sum_{j=1}^{d} 1_{A_*(z)}(z_j) - \sum_{j\neq i} 1_{A_*(z)}(z_j)\right) \\ &= \sum_{i=1}^{d} 1_{A_*(z)}(z_i) = h(z) \ .\end{aligned}$$

Hence, $h$ has the self-bounding property. Now, Inequality (2.10.5) simply follows from [18, Theorem 6.12]. □

**Corollary 2.1.** *The following process*

$$Z(\mathcal{F}_{>a}, x) = \sup_{\mathbf{c}\in\mathcal{F}_{>a}} \frac{1}{|J|}\sum_{b\in J} \mathbf{1}_{\{R_* - P_b(\ell_{\mathbf{c}}) \geq -x\}}$$

*is concentrated around its expected value according to the following inequality,*

$$\mathbb{P}\left(Z(\mathcal{F}_{>a}, x) \geq \Delta\right) \leq \exp\left(-\frac{|J|(\Delta - \mathbb{E}[Z(\mathcal{F}_{>a}, x)])^2}{2\mathbb{E}[Z(\mathcal{F}_{>a}, x)] + 2(\Delta - \mathbb{E}[Z(\mathcal{F}_{>a}, x)])/3}\right) \ . \tag{2.10.6}$$

*Proof.* It suffices to apply Lemma 2.1 with $d = |J|$, $\mathcal{Z} = \mathbb{R}^{p\times n_B}$, $\xi_b = (X_j)_{j\in I_b}$ for $b \in J$ and

$$\mathcal{A} = \left\{\left\{z = (x_1, ..., x_{n_B}) : \frac{-1}{|n_B|}\sum_{i=1}^{n_B} \ell_{\mathbf{c}}(x_i) + R_* > -x\right\} : \mathbf{c} \in \mathcal{F}_{>a}\right\} \ .$$

□

# Chapter 3

# High dimensional logistic entropy clustering

**Abstract.** Minimization of the (regularized) entropy of classification probabilities is a versatile class of discriminative clustering methods. The classification probabilities are usually defined through the use of some classical losses from supervised classification and the point is to avoid model-based techniques by just optimizing the law of the labels conditioned on the observations. We give the first theoretical study of such methods, by specializing to logistic classification probabilities. We prove that if the observations are generated from a two-component isotropic Gaussian mixture, then minimizing the entropy risk over a Euclidean ball indeed allows to identify the separation vector of the mixture. Furthermore, if this separation vector is sparse, then penalizing the empirical risk by a $\ell_1$-regularization term allows to infer the separation in a high-dimensional space and to recover its support, at standard rates of sparsity problems. Our approach is based on the local convexity of the logistic entropy risk, that occurs if the separation vector is large enough, with a condition on its norm that is independent from the space dimension. This local convexity property also guarantees fast rates in a classical, low-dimensional setting.

## 3.1 Introduction

The clustering problem can be described as follows: given a measurable space $\mathcal{X}$, a sample $(X_1, ..., X_n) \in \mathcal{X}^n$, and an integer $K \geq 2$, define a (random) labelling function $Y : \mathcal{X} \to \{1, ..., K\}$. In particular, to each data $X_i$, associate a label $Y_i$. If the function $Y$ is deterministic, then the task is termed "hard clustering". If the function $Y$ is random, the distribution of the labels $Y(x)$, for $x \in \mathcal{X}$, being characterized by the uplets $(\mathbb{P}(Y(x) = 1), ..., \mathbb{P}(Y(x) = K))$, then the clustering task is said to be "soft". In the soft clustering case, a common approach - called the modelling approach - is to model the distribution of the data, typically as a mixture distribution, and to directly relate the probabilities $(\mathbb{P}(Y(x) = 1), ..., \mathbb{P}(Y(x) = K))$ to the parameters of the mixture [20]. One can then reduce to a hard clustering by assigning each point $x$ to the maximizer of classification probabilities (or choose one at random amongst the maximizers if it is non-unique). Hard clustering algorithms include the celebrated K-means [93, 135, 99], hierachical clustering [73], spectral clustering [112]) among others.

Particularly developed in the machine learning community for its flexibility when addressing complex data, the so-called "discriminative approach" to clustering amounts to model the classification probabilities $(\mathbb{P}(Y(x) = 1), ..., \mathbb{P}(Y(x) = K))$, which can be understood as the conditional probabilities of the labels with respect to the position $x$. Proceeding this way indeed avoids the modelling of the whole distribution of data and often reduces to encode in the classification probabilities, the frontiers separating the clusters. In general, this is done through the use of classical learning losses such as the logistic, the Hinge or the Conditional Random Fields loss [38, 60]. More formally, one puts the constraint of $\mathbb{P}(Y(x) = k)$, $k \in \{1, ..., K\}$, being proportional to $\exp(-\ell(\beta_k, x))$, for a vector $\beta_k$ and a loss $\ell$. For instance the logistic loss gives classification probabilities proportional to $\exp(w_k^t x + b_k)$ and the Hinge loss (for $K = 2$) induces probabilities of a form proportional to $\exp(-[1 - (w_k^t \varphi(x) + b_k)]_+)$ for some feature map $\varphi$ and with $(w_1, b_1) = (-w_2, -b_2)$ in this binary case.

In addition, these losses were primarily introduced for supervised learning and in order to transfer them to the unsupervised setting, one has to define what would be a desirable (unobserved) label. Arguably, when classifying data, one would prefer to be as sure as possible of its cluster choice. This is equivalent to saying that the maximum of classification probabilities would be as close to one as possible. Hence, a natural criterion to infer a labelling function, would be to define $\tilde{Y}$ through the probabilities $\mathbb{P}(\tilde{Y}(x) = k) = Z_{\tilde{\beta}}^{-1}(x) \exp(-\ell(\tilde{\beta}_k, x))$, with a normalizing

constant $Z_{\tilde{\beta}}(x) = \sum_{k=1}^{K} \exp(-\ell(\tilde{\beta}_k, x))$, such that

$$(\tilde{\beta}_1, ..., \tilde{\beta}_K) \in \arg \max_{(\beta_1,...,\beta_K)} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{Z_\beta(x_i)} \max_{k\in\{1,...,K\}} [\exp(-\ell(\beta_k, x_i))] \right\}. \tag{3.1.1}$$

The associated theoretical target is $\mathbb{P}(Y_*(x) = k) = \exp(-\ell(\beta_{*,k}, x))$ with,

$$(\beta_{*,1}, ..., \beta_{*,K}) \in \arg \max_{(\beta_1,...,\beta_K)} \left\{ \mathbb{E}\left[ \frac{1}{Z_\beta(X)} \max_{k\in\{1,...,K\}} [\exp(-\ell(\beta_k, X))] \right] \right\},$$

where $X$ follows the unknown - and not modeled - distribution of data.

But the maximum is not a smooth function and it may cause difficulties when trying to optimize (3.1.1). As a smooth proxy, one can try to minimize the entropy of the classification probabilities, since it achieves its minimum value when the latter probabilities are all equal to zero or one. This amounts to search for a labelling function $\hat{Y}$ satisfying $\mathbb{P}(\hat{Y}(x) = k) = Z_{\hat{\beta}}^{-1}(x) \exp(\ell(\hat{\beta}_k, x))$ with

$$(\hat{\beta}_1, ..., \hat{\beta}_K) \in \arg \min_{(\beta_1,...,\beta_K)} \left\{ \frac{1}{n} \sum_{i=1}^{n} \text{Ent}\left\{\mathbb{P}(Y_\beta(x_i) = 1), ..., \mathbb{P}(Y_\beta(x_i) = K)\right\} \right\}, \tag{3.1.2}$$

where

$$\text{Ent}\left\{\mathbb{P}(Y_\beta(x_i) = 1), ..., \mathbb{P}(Y_\beta(x_i) = K)\right\} = \sum_{k=1}^{K} -\frac{\exp(-\ell(\beta_k, x_i))}{Z_\beta(x_i)} \log\left( \frac{\exp(-\ell(\beta_k, x_i))}{Z_\beta(x_i)} \right). \tag{3.1.3}$$

Often, one has to restrict the search among vectors $(\beta_1, ..., \beta_K)$ in a compact set, or to add to the entropy a regularization term encoding the complexity of the vectors $(\beta_1, ..., \beta_K)$ [60, 38]. In this second formulation, the theoretical target $(\beta_{0,1}, ..., \beta_{0,K})$ of estimation is,

$$(\beta_{0,1}, ..., \beta_{0,K}) \in \arg \min_{(\beta_1,...,\beta_K)} \left\{ \mathbb{E}\left[\text{Ent}\left\{\mathbb{P}(Y_\beta(x_i) = 1), ..., \mathbb{P}(Y_\beta(x_i) = K)\right\}\right] \right\}. \tag{3.1.4}$$

The use of entropy terms in semi-supervised and unsupervised learning is indeed natural and has been the object of active research [63, 60, 38, 138, 137, 133, 91, 3, 106]. Furthermore, this approach is at the core of some state-of-the-art deep clustering approaches [71]. Another fruitful approach in discriminative clustering consists in considering convex relaxations of some initial, untractable criteria and this methodology often comes with strong theoretical

guarantees [51, 77, 10, 120, 57, 31, 30, 58, 105, 128, 34].

The starting point of our work consists in the following observation: to our knowledge, no theoretical guarantee - of the type of convergence rates - exists in the literature for (regularized) minimum entropy estimators (3.1.2). This a weakness compared to other approaches, such as convex relaxations techniques for instance. But from a practical perspective, estimators of the form of (3.1.2) have already proved to be efficient and flexible - allowing for instance feature maps embedding and the use of deep architectures - and the lack of theoretical studies needs to be filled.

We consider the unsupervised classification of a bipartite high-dimensional Gaussian mixture, with sparse means. This framework is indeed a good benchmark, since on the one hand, it is sufficiently simple to allow us to understand the nature of the target $(\beta_{0,1}, ..., \beta_{0,K})$ - with $K = 2$ and $\beta_{0,1} = -\beta_{0,2}$ in our bipartite framework - and to investigate the rate of convergence of estimators of the form of (3.1.2), suitably regularized by a $\ell_1$-penalty. On the other hand, the two-component high-dimensional Gaussian mixture has received recently at lot of attention [19, 8, 108, 92, 75, 47, 32, 9, 76, 24, 94]. Let us emphasize that our goal is not *a priori* to provide a state-of-the-art method, specifically designed to solve the high-dimensional Gaussian mixture clustering, but to explore for the first time the theoretical behavior of discriminative estimators that minimize the (regularized) classification entropy and see how they can adapt to a sparse setting.

## 3.2 Some notations and definitions

Let $a := (a_1, ..., a_d) \in \mathbb{R}^d$ and $X$ be a random variable valued in $\mathbb{R}^d$, with distribution $P$. More precisely $X := \varepsilon Z$ with $\varepsilon \sim \text{Rad}\left(\frac{1}{2}\right)$ and $Z \sim \mathcal{N}(a, I_d)$ a Gaussian vector independent from $\varepsilon$, with normalized variance equal to the identity matrix $I_d$. Take $n \in \mathbb{N}^*$, $X^{(1)}, ..., X^{(n)}$ are observations of $X$ independent and identically distributed according to $P$. Our goal is to estimate the labelling function $Y_*(x) = \text{sign}(x^t a)$, or its opposite, which gives the same hard clustering. This amounts to estimate the separation vector $a$. To do this, we will use an entropy criterion.

Set the logistic probability $p_\beta(X) := 1/(1 + e^{X^t\beta})$ where $\beta \in \mathbb{R}^d$ and its complementary probability $q_\beta(X) := e^{X^t\beta}/(1 + e^{X^t\beta})$. The *logistic entropy* $\rho_\beta$ is defined as follows, $\rho_\beta(X) := \rho(\beta^t X) = -p_\beta(X)\log p_\beta(X) - q_\beta(X)\log q_\beta(X)$. The associated risk is $\mathcal{R}(\beta) := \mathbb{E}[\rho_\beta(X)]$. The latter expectation will also be denoted $P\rho_\beta$ for short. Let $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ be respectively the $L_1$,$L_2$ and $L_\infty$-norm, and denote $B_1(0, R)$, $B_2(0, R)$ and $B_\infty(0, R)$ their corresponding balls centered at 0 with radius $R$ in $\mathbb{R}^d$. We consider the minimizer $\beta_0$ of the risk

$\mathcal{R}(\beta)$ over a $L_2$-ball $B_2(0,R)$ - where the radius $R$ will be fixed latter -, $\beta_0 \in \underset{\beta \in B_2(0,R)}{\arg\min} \{\mathcal{R}(\beta)\}$, with excess risk $\mathcal{E}(\beta, \beta_0) := \mathcal{R}(\beta) - \mathcal{R}(\beta_0)$, for $\beta \in B_2(0,R)$. The empirical distribution of $X^{(1)}, ..., X^{(n)}$ is $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{X^{(i)}}$, where $\delta_{X^{(i)}}$ is the Dirac distribution on $X^{(i)}$, and the quantity$\hat{\mathcal{R}}_n(\beta) := P_n \rho_\beta = \frac{1}{n}\sum_{i=1}^n \rho_\beta\left(X^{(i)}\right)$ is the empirical counterpart of $\mathcal{R}(\beta)$, called the empirical risk.

We denote by $\gamma$ the probability density function (PDF) of a centered standard real Gaussian variable $\mathcal{N}(0,1)$. $\Phi$ is its cumulative distribution function (CDF) and $\Phi^c : t \mapsto \int_t^\infty \gamma(u)\,du$ the tail distribution of the PDF $\gamma$. In addition, we write $G$ the so-called Gaussian Mill's ratio $G(x) := \frac{\Phi^c(x)}{\gamma(x)}$. $\forall u, v \in \mathbb{R}, u \wedge v := \min(u,v)$ and $u \vee v := \max(u,v)$. In this chapter $\alpha : x \mapsto -\frac{e^x}{(1+e^x)^2}\left(1 + x\frac{1-e^x}{1+e^x}\right)$ and $x_1$ is the unique element of $\{x > 0 : \alpha(x) = 0\}$, satisfying $x_1 \in [1.54, 1.55]$.

For a vector $\beta = (\beta_1, ..., \beta_p)^t \in \mathbb{R}^p$, we define its support as the set $S$ of indices such that $S = \{i \in \{1, ..., p\}; \beta_i \neq 0\}$. The vector $\beta$ is said to be $s$-sparse if $\mathrm{Card}(S) \leq s$. Furthermore, for a set of indices $I \subset \{1, ..., p\}$, we denote $\beta^I \in \mathbb{R}^p$ the vector such that $\beta_i^I = \beta_i$ if $i \in I$ and $\beta_j^I = 0$ if $j \notin I$.

## 3.3 Minimising the risk over a $L_2$-ball

Recall that

$$\beta_0 \in \underset{\beta \in B_2(0,R)}{\arg\min} \{\mathcal{R}(\beta)\},$$

where the radius $R$ will be fixed later. Let us investigate the geometry of the risk $\mathcal{R}$ defined by the logistic entropy.

**Proposition 3.1.** *The risk is symmetric, $\mathcal{R}(\beta) = \mathcal{R}(-\beta)$, and the risk value $\mathcal{R}(\beta)$ with $\|\beta\|_2 = r$ fixed is decreasing with respect to $|\beta^t a|$.*

Proposition 3.1 states that the risk is symmetric around zero, and that its values on a sphere are increasing with respect to the distance to the line $\mathbb{R}a$. Its proof can be found in Section 3.5.1.

**Proposition 3.2.** *The function $\lambda \mapsto \mathcal{R}(\lambda\beta)$ is decreasing for $\lambda \in \mathbb{R}_+$.*

In Proposition 3.2, it is proved that the risk is decreasing on semi-lines starting at zero. For a proof of this result, see Section 3.5.1. From Propositions 3.1 and 3.2, we characterize the minimizers of the risk over a $L_2$-ball.

**Corollary 3.1.** *The minimum of $\mathcal{R}(\beta)$ on $B_2(0,R)$ is reached at $\pm\beta_0$ where $\beta_0 := Ra/\|a\|_2$.*

From Corollary 3.1, we deduce that estimating $\beta_0$ or its opposite directly gives an estimation of the best labelling function $Y_*$ for our clustering problem. A look at the proof of Propositions 3.1 and 3.2 shows that these results, and hence Corollary 3.1, hold true in the more general setting where the distribution of $Z$ is only assumed to be spherically symmetric.

In order to tackle the estimation of a sparse separation vector $a$, the following property will be helpful.

**Theorem 3.1.** *Let $\beta_0 = Ra/\|a\|_2$ and let $\Lambda_{min}$ be the smallest eigenvalue of the Hessian $d^2_{\beta_0}\mathcal{R}$ at $\beta_0$. Take a parameter $\nu = 0.95$, $R \geq \sqrt{x_1 + 0.08}$ ($R = 1.28$ for instance) and assume that $\|a\|_2 \geq 2R$, then*

$$\Lambda_{min} \geq \frac{\nu}{4}\left(\Phi^c\left(\|a\|_2 - \frac{x_1}{R}\right) - \Phi^c\left(\|a\|_2 + \frac{x_1}{R}\right)\right).$$

Theorem 3.1 states that if the radius $R$ and the mean vector $a$ are sufficiently large, then the risk defined by the logistic entropy is locally strongly convex around $\beta_0$. The risk is not convex over the whole $L_2$-ball $B_2(0, R)$, but this local convexity is very convenient, since it allows to deduce a quadratic growth of the excess risk pointed on $\beta_0$, as follows.

**Lemma 3.1.** *Set $\beta_0$ the unique minimum of $\mathcal{R}(\cdot)$ on $\Psi_U := \{\beta \in B_2(0, R) : \beta^t U > 0\}$ where $U$ is a random variable uniformly distributed on the unit $L^2$-ball. Assume that $R \geq \sqrt{x_1 + 0.08}$ and $\|a\|_2 \geq 2R$. We have*

$$\inf_{\beta \in \Psi_U} \frac{\mathcal{E}(\beta, \beta_0)}{\|\beta - \beta_0\|_2^2} \geq c_0 > 0$$

*with*

$$c_0 = L_0 \frac{(\|a\|_2 - R)^6}{\|a\|_2^8 R^2} . \exp\left(-\|a\|_2 R - 2R^2\right)$$

*for a numerical constant $L_0$ ($L_0 = 9 \times 2^{22}$ holds).*

The quadratic growth of the excess risk stated in Lemma 3.1 will turn out to be a keystone to prove the oracle inequality for the excess risk of the minimizer of empirical risk regularized by a $\ell_1$ penalty (see Section 3.4). The proof of Lemma 3.1 is postponed to Section 3.5.1.

## 3.4 An oracle inequality in high dimension

Recall that $\beta_0 = Ra/\|a\|_2$ is a minimizer of the risk over the $L_2$-ball of radius $R$: $\beta_0 \in \underset{\beta \in B_2(0,R)}{\arg\min}\, \mathcal{R}(\beta)$. Set $\Psi_U := \{\beta \in B_2(0,R) : \beta^t U > 0\}$ and where $U$ is a random variable uniformly distributed on the unit Euclidean sphere, independent from the observations. We have $\mathbb{P}(\beta_0^t U = 0) = 0$ and so $\beta_0$ or its opposite belongs to $\Psi_U$. Without loss of generality, we assume that $\beta_0 \in \Psi_U$ and analyze the situation conditionnally on the choice of $U$.

We investigate the behavior of the following estimator,

$$\hat{\beta} := \underset{\beta \in \Psi_U}{\arg\min} \left\{ \mathcal{R}_n(\beta) + \lambda \|\beta\|_1 \right\}. \tag{3.4.1}$$

Set also the empirical process $V_n(\beta) := (P_n - P)(\rho_\beta)$. For some $T > 1$, define the event

$$\mathcal{T} := \left\{ \sup_{\beta \in B_2(0,R)} \frac{|V_n(\beta) - V_n(\beta_0)|}{\|\beta - \beta_0\|_1 \vee \lambda_0} \leq 2T\lambda_0 \right\}, \tag{3.4.2}$$

where $\lambda_0 > 0$ is to be fixed in the following theorem.

**Theorem 3.2.** *Fix $n \geq 2$. Assume that $\beta_0$ is $s$-sparse, for some integer $s \geq 1$, and denote $S$ its support. Set $M_n := \|a\|_\infty + \sqrt{2\log d} + \sqrt{2\log(1+n)}$ and*

$$\lambda_0 := 3LM_n \left( 5\sqrt{3\log(2d)}\log n + 4 \right) n^{-1/2}.$$

*When the event $\mathcal{T}$ occurs, it holds: $\forall \lambda > 2T\lambda_0$,*

$$\mathcal{E}\left(\hat{\beta}, \beta_0\right) + 4(\lambda - 2T\lambda_0) \left\| \hat{\beta}^{S^c} \right\|_1 \leq A_{\|a\|_2, R} s (T\lambda_0 + \lambda)^2, \tag{3.4.3}$$

*where $A_{\|a\|_2,R}$ is a constant depending only on $\|a\|_2$ and $R$. More precisely, for a numerical constant $A_0$, one can take*

$$A_{\|a\|_2,R} = A_0 \|a\|_2^8 (\|a\|_2 - R)^{-6} R^2 e^{\|a\|_2 R + 2R^2}.$$

*Furthermore, the event $\mathcal{T}$ occurs with probability at least*

$$1-\frac{3}{4}\log\left(\frac{4R^2nd}{L^2M_n^2}\right)\exp\left(-21\left(T-1\right)^2\log\left(2d\right)\log^2 n\right)-\frac{1}{25T^2\log\left(2d\right)n\log^2 n}.$$

According to Theorem 3.2, if the regularization parameter $\lambda$ is equal for instance to $3T\lambda_0$, then the rate of convergence of the excess risk is of the order

$$\frac{s\log d\log^2 n\log\left(d\vee n\right)}{n},$$

with a pre-factor that only depends on $\|a\|_2$ and $R$. Thus the estimator $\hat{\beta}$ adapts to sparsity and is able to estimate $\beta_0$ even if $d >> n$. Furthermore, the rate of convergence of $\left\|\hat{\beta}^{S^c}\right\|_1$ would be given by

$$s\sqrt{\frac{\log d\log^2 n\log\left(d\vee n\right)}{n}},$$

with also a pre-factor that only depends on $\|a\|_2$ and $R$. This means that if $s$ and $d$ are such that this rate (for a bounded $\|a\|_2$) goes to zero with $n$ growing to infinity, then the support $S$ of $\beta_0$ is recovered in the sense that $\left\|\hat{\beta}^{S^c}\right\|_1$ goes to zero.

Note however that the dependence in $\|a\|_2$ is exponential in our bounds. This due to our argument of proof, which uses the local convexity of the risk around $\beta_0$. But when $\|a\|_2$ is large, the risk tends to be flat (see Theorem 3.1). This local convexity argument is also at the core the approach, developed in [134], to the non-convex $\ell_1$-penalized loss in mixture regression (see also [29, Chapter 9]).

It is also worth noting that in a classical, non-sparse case where the dimension is (much) smaller than the sample size, a convergence bound could also be obtained, by standard empirical process techniques. Indeed, the loss $\rho$ is bounded and Lipschitz, so the rate of convergence of the following estimator,

$$\tilde{\beta}\in\arg\min_{\beta\in B_2(0,R)}\left\{\hat{\mathcal{R}}_n(\beta)\right\},$$

is of the order

$$\sqrt{\frac{Rd}{n}}+\sqrt{\frac{\log(1/\delta)}{n}}+\frac{\log(1/\delta)}{n},$$

up to a numerical pre-factor and on an event of probability at least $1-\delta$ for $\delta\in(0,1)$. An important remark is

that the latter rate in $\sqrt{d/n}$ holds without any assumption on $R$ and $\|a\|_2$, because the local convexity of the risk on $\beta_0$ - that is Theorem 3.1 - is not needed to prove it. If Theorem 3.1 furthermore holds, it is not hard to see that the rate is actually $d/n$, up to a pre-factor. Indeed, Theorem induces a so-called margin relation for the excess risk, which in turn induces a fast rate, since the loss is bounded (see for instance [101]).

Also, one can consider the adaptive selection of the regularization parameter . For this, a sensible idea is to consider a BIC-type criterion defined with the active set of the estimators corresponding to different values of

Also, one can consider the adaptive selection of the regularization parameter. For this, a sensible idea is to consider a BIC-type criterion defined with the active set of the estimators corresponding to different values of the regularization parameter.

We postpone to a forthcoming research the practical implementation of the estimator, together with comparisons in the sparse two-component Gaussian mixture model with other available algorithms.

## 3.5 Proofs

Define the empirical process $V_n(\beta) = (P_n - P)(\rho_\beta)$ and $V_n^{trunc}(\beta) = (P_n - P)\left(\rho_\beta I_{\{G(X)\leq M_n\}}\right)$ where $G(X) := \|X\|_\infty$ and note that $\rho_\beta : \beta \mapsto \rho(\beta^t X)$, with the function $\rho$ that is $L$-lipschitz (with $L < 2.5$).

### 3.5.1 Proofs of the main results

*Proof of Proposition 3.1.* Take $X = \varepsilon Z$ where $\varepsilon \sim Rad(1/2)$ and $Z \sim \mathcal{N}(a, I_d)$, with $a \in \mathbb{R}^d$ and also $N \sim \mathcal{N}(0,1)$. Because expression (3.5.5) of Lemma 3.2 is symmetric in $X$, one has $\mathcal{R}(\beta) = \mathcal{R}(-\beta)$ and

$$\mathcal{R}(\beta) = \mathbb{E}\left[\log\left(1 + e^{Z^t\beta}\right) - \frac{Z^t\beta e^{Z^t\beta}}{1 + e^{Z^t\beta}}\right].$$

The distribution of the real-valued random variable $Z^t\beta$ is $\mathcal{N}\left(\beta^t a, \|\beta\|_2^2\right)$ and we assume that $\|\beta\|_2 = r$. The criterion can be seen as a function of $\mu := \beta^t a$ and $r$:

$$\mathcal{R}(\beta) = \mathbb{E}\left[\log\left(1+e^{\mu+rN}\right) - \frac{(\mu+rN)\,e^{\mu+rN}}{1+e^{\mu+rN}}\right] =: \mathcal{R}(\mu, r). \tag{3.5.1}$$

Its derivative with respect to $\mu$ is:

$$\begin{aligned}
\partial_\mu \mathcal{R}(\mu, r) &= \frac{d}{d\mu}\mathbb{E}\left[\log\left(1+e^{\mu+rN}\right) - \frac{(\mu+rN)\,e^{\mu+rN}}{1+e^{\mu+rN}}\right] \\
&= \mathbb{E}\left[\frac{d}{d\mu}\log\left(1+e^{\mu+rN}\right) - \frac{d}{d\mu}\frac{(\mu+rN)\,e^{\mu+rN}}{1+e^{\mu+rN}}\right] \\
&= \mathbb{E}\left[\frac{e^{\mu+rN}}{1+e^{\mu+rN}} - \left(\frac{e^{\mu+rN}}{1+e^{\mu+rN}} + \frac{(\mu+rN)\,e^{\mu+rN}}{1+e^{\mu+rN}} + (\mu+rN)\,e^{\mu+rN}\frac{-e^{\mu+rN}}{\left(1+e^{\mu+rN}\right)^2}\right)\right] \\
&= \mathbb{E}\left[-\frac{(\mu+rN)\,e^{\mu+rN}}{1+e^{\mu+rN}} + (\mu+rN)\,e^{\mu+rN}\frac{e^{\mu+rN}}{\left(1+e^{\mu+rN}\right)^2}\right] \\
&= \mathbb{E}\left[\frac{(\mu+rN)\,e^{\mu+rN}}{1+e^{\mu+rN}}\left(\frac{e^{\mu+rN}}{1+e^{\mu+rN}} - 1\right)\right] \\
\partial_\mu \mathcal{R}(\mu, r) &= -\mathbb{E}\left[\frac{(\mu+rN)\,e^{\mu+rN}}{\left(1+e^{\mu+rN}\right)^2}\right].
\end{aligned}$$

Let us define $g : x \mapsto \frac{xe^x}{(1+e^x)^2}$ so that $\partial_\mu \mathcal{R}(\mu, r) = -\mathbb{E}\left[g\left(\mu + rN\right)\right]$. We use the lemma 3.3 and the fact that $g$ is odd and positive on $(0, +\infty)$ to conclude that $\mathbb{E}\left[g\left(\mu + rN\right)\right]$ has the sign of $\mu$, which gives the result. □

*Proof of Proposition 3.2.* Take $\beta \in \mathbb{R}^d$, there is $u \in \mathbb{R}$ such that $\beta^t a = u\,\|\beta\|_2$. Recall Identity (3.5.1) above, where $\mathcal{R}$ can be seen as a function of $\mu$ and $r$ with $Z^t\beta \sim \mathcal{N}\left(\mu, r^2\right)$. Then we have

$$\frac{\partial \mathcal{R}(\lambda\beta)}{\partial \lambda} = \frac{\partial \mathcal{R}\left(\lambda\beta^t a, \|\lambda\beta\|_2\right)}{\partial \lambda} = \frac{\partial \mathcal{R}(ru, r)}{\partial r}\,\|\beta\|_2\,.$$

We set $\forall u \in \mathbb{R}, N_u \sim \mathcal{N}(u, 1)$ and Equation (3.5.1) gives:

$$
\begin{aligned}
\frac{\partial \mathcal{R}\left(ru, r\right)}{\partial r} &= \frac{\partial}{\partial r}\mathbb{E}\left[\log\left(1+e^{ru+rN_0}\right) - \frac{\left(ru+rN_0\right)e^{\left(ru+rN_0\right)}}{1+e^{\left(ru+rN_0\right)}}\right] \\
&= \mathbb{E}\left[\frac{\partial}{\partial r}\log\left(1+e^{rN_u}\right) - \frac{\partial}{\partial r}\left(\frac{rN_u e^{rN_u}}{1+e^{rN_u}}\right)\right] \\
&= \mathbb{E}\left[\frac{N_u e^{rN_u}}{1+e^{rN_u}} - \left(\frac{N_u e^{rN_u}}{1+e^{rN_u}} + \frac{rN_u\left(N_u e^{rN_u}\right)}{1+e^{rN_u}} + rN_u e^{rN_u}\frac{-N_u e^{rN_u}}{\left(1+e^{rN_u}\right)^2}\right)\right] \\
&= \mathbb{E}\left[-\frac{rN_u\left(N_u e^{rN_u}\right)}{1+e^{rN_u}} + rN_u e^{rN_u}\frac{N_u e^{rN_u}}{\left(1+e^{rN_u}\right)^2}\right] \\
&= \mathbb{E}\left[\frac{rN_u\left(N_u e^{rN_u}\right)}{1+e^{rN_u}}\left(\frac{e^{rN_u}}{1+e^{rN_u}} - 1\right)\right] \\
&= \mathbb{E}\left[\frac{rN_u\left(N_u e^{rN_u}\right)}{1+e^{rN_u}}\left(\frac{-1}{1+e^{rN_u}}\right)\right] \\
&= -\mathbb{E}\left[\frac{rN_u^2 e^{rN_u}}{\left(1+e^{rN_u}\right)^2}\right] < 0.
\end{aligned}
$$

Hence $\frac{\partial \mathcal{R}(\lambda\beta)}{\partial \lambda} < 0$ as required. □

*Proof of Theorem 3.1.* We make use of Equation (3.5.15) from Lemma 3.8: $\forall a \in \mathbb{R}^d$, $R, \nu > 0$,

$$
R\left(1 - \left(R - \|a\|_2 + \frac{x_1 + \frac{8}{100}}{R}\right) G\left(\frac{x_1}{R} + R - \|a\|_2\right)\right) \geq (1+\nu)\frac{e^{x_1}}{4} G\left(\|a\|_2 - \frac{x_1}{R}\right), \tag{3.5.2}
$$

where, see Section 3.2, $x_1$ is a positive numerical constant and the function $G$ is the so-called Gaussian Mill's ratio. By Proposition 3.4, we also have that $G$ is decreasing on $\mathbb{R}$. Hence, if Equation (3.5.2) is satisfied for some values of $\|a\|_2, R$ and $\nu$ such that $\|a\|_2 - \left(R + (x_1 + 0.08)R^{-1}\right) > 0$, then it is satisfied for any triplet $(\|a\|_2 + h, R, \nu)$ with $h > 0$. In addition, we know from Lemma 3.18 that $\|a\|_2 = 2R \approx 2.548$, $R = \sqrt{x_1 + 0.08} \approx 1.2741$ and $\nu = 0.95$ make Equation (3.5.2) hold true. Consequently, it also holds true when $\|a\|_2 \geq 2R \approx 2.548$, $R = \sqrt{x_1 + 0.08} \approx 1.2741$ and $\nu = 0.95$.

According to lemma 3.8, when Equation (3.5.2) holds, one can control from below the values of $\left(d^2_{\beta_0}\mathcal{R}\right)(h, h)$. More

precisely,

$$\begin{aligned}\Lambda_{min} &:= \inf_{\|h\|=1} \left(d^2_{\beta_0}\mathcal{R}\right)(h,h) \\ &\geq \inf_{\substack{\|h\|=1 \\ \eta=\|h_\|\|}} \frac{\nu}{4}\left(\eta^2\frac{x_1^2}{R^2}+1-\eta^2\right)\left(\Phi^c\left(\|a\|_2-\frac{x_1}{R}\right)-\Phi^c\left(\|a\|_2+\frac{x_1}{R}\right)\right) \\ &= \frac{\nu}{4}\left(\Phi^c\left(\|a\|_2-\frac{x_1}{R}\right)-\Phi^c\left(\|a\|_2+\frac{x_1}{R}\right)\right)\underbrace{\inf_{0\leq\eta\leq 1}\left(\eta^2\frac{x_1^2}{R^2}+1-\eta^2\right)}_{=1}\end{aligned}$$

because $x_1/R \geq 1$. This proves the result. □

*Proof of Lemma 3.1.* The risk $\mathcal{R}$ admits two minima $\beta_0$ and $-\beta_0$ on $B_2(0,R)$. We consider

$$\Psi_U = \left\{\beta \in B_2(0,R) : \beta^t U > 0\right\},$$

where $U$ is a random variable uniformly distributed on the unit $L^2$-ball. The probability that $U \in \beta_0^\perp$ is 0 then with probability 1 we have $U \notin \beta_0^\perp$ and there is therefore only one vector among $\beta_0$ and $-\beta_0$ that satisfies $\beta_0^t U > 0$. We call $\beta_0$ the vector satisfying both $\mathcal{R}(\beta_0)$ is the minimum of $\mathcal{R}(\cdot)$ and $\beta_0^t U > 0$.

Take $\beta \in \Psi_U$ and let $\varepsilon \in (0,R)$, we are about to control $\mathcal{E}(\beta,\beta_0)$ on $B_2(\beta_0,\varepsilon)$ and $\{\nu \in B_2(0,R) : \beta_0^t\nu > 0\} \setminus B_2(\beta_0,\varepsilon)$ but these two sets may not be included $\Psi_U$. To bypass this issue, remark that the risk $\mathcal{R}$ is symmetric with respect to 0. Hence, in the case where $\beta \notin \{\nu \in B_2(0,R) : \beta_0^t\nu > 0\}$, we will have $\mathcal{E}(\beta,\beta_0) = \mathcal{E}(-\beta,\beta_0)$ where $-\beta \in \{\nu \in B_2(0,R) : \beta_0^t\nu > 0\}$. Consequently, one can always control $\mathcal{E}(\cdot,\beta_0)$ on $\Psi_U$ with its values on $\{\nu \in B_2(0,R) : \beta_0^t\nu > 0\}$, and without loss of generality we will focus on the control of $\mathcal{E}(\cdot,\beta_0)$ on $\{\nu \in B_2(0,R) : \beta_0^t\nu > 0\}$.

Case 1: $\beta \in B_2(\beta_0,\varepsilon)$

We know from Lemma 3.21 that $\forall\beta \in B_2(\beta_0,\varepsilon)$,

$$\frac{\mathcal{E}(\beta,\beta_0)}{e^{-\left(\|a\|_2 R - R^2/2\right)}\|\beta-\beta_0\|_2^2} \geq \frac{1}{16}\left(1+\left(\|a\|_2-R\right)^2\right) - 24\|a\|^4 e^{R^2/2}e^{\varepsilon\|a\|_2}\|\beta-\beta_0\|_2.$$

When $\|\beta-\beta_0\|_2 \leq \frac{1}{2}\frac{\frac{1}{16}\left(1+\left(\|a\|_2-R\right)^2\right)}{24\|a\|^4 e^{R^2/2}e^{\varepsilon\|a\|_2}}$, one has

$$\frac{\mathcal{E}(\beta,\beta_0)}{e^{-\left(\|a\|_2 R-R^2/2\right)}\|\beta-\beta_0\|_2^2} \geq \frac{1+\left(\|a\|_2-R\right)^2}{16}.$$

In particular, the latter inequality holds when

$$\varepsilon \leq \frac{1}{768}\frac{1+\left(\|a\|_2-R\right)^2}{\|a\|_2^4 e^{R^2/2}}e^{-\varepsilon\|a\|_2},$$

which is satisfied for

$$\begin{aligned}\varepsilon &\leq \frac{1}{768}\frac{1+\left(\|a\|_2-R\right)^2}{\|a\|_2^4 e^{R^2/2}}\exp\left(-\frac{1}{384}\frac{1+\left(\|a\|_2-R\right)^2}{\|a\|_2^4 e^{R^2/2}}\|a\|_2\right)\\ &= \frac{1}{768}\frac{1+\left(\|a\|_2-R\right)^2}{\|a\|_2^4}\exp\left(-R^2/2-\frac{1+\left(\|a\|_2-R\right)^2}{384\|a\|_2^3 e^{R^2/2}}\right) =: \varepsilon_{max}.\end{aligned}$$

Then for all $\beta \in B_2(\beta_0,\varepsilon_{max})$, we have

$$\frac{\mathcal{E}(\beta,\beta_0)}{\|\beta-\beta_0\|_2^2} \geq \frac{1}{32}\left(1+\left(\|a\|_2-R\right)^2\right)e^{-\left(\|a\|_2 R-R^2/2\right)}.$$

Case 2: $\beta \in \{\nu \in B_2(0,R) : \beta_0^t\nu > 0\} \setminus B_2(\beta_0,\varepsilon_{max})$.

Lemmas 3.1 and 3.2 imply that $\forall\lambda>1, \mathcal{E}(\lambda\beta,\beta_0) < \mathcal{E}(\beta,\beta_0)$ and

$$\text{if } \nu \in \left\{\mu \in B_2(0,R) : \|\mu\| = \|\beta\| \ \ \&\ \beta_0^t\mu > \beta_0^t\beta\right\},\ \mathcal{E}(\nu,\beta_0) < \mathcal{E}(\beta,\beta_0).$$

With these two properties, we are always able to control $\mathcal{E}(\beta,\beta_0)$ with another value $\mathcal{E}(\nu,\beta_0)$ where $\nu \in B_2(\beta_0,\varepsilon_{max})$. Indeed, if $\mathbb{R}\cdot\beta$ intersects $B_2(\beta_0,\varepsilon_{max})$, there there exists $\lambda>1$ such that $\mathcal{E}(\beta,\beta_0) > \mathcal{E}(\lambda\beta,\beta_0)$, where $\lambda\beta \in B_2(\beta_0,\varepsilon_{max})$. Oherwise, we have $\mathcal{E}(\beta,\beta_0) \geq \mathcal{E}\left(R\frac{\beta}{\|\beta\|_2},\beta_0\right)$ and $\mathcal{E}\left(R\frac{\beta_0}{\|\beta_0\|_2},\beta_0\right) > \mathcal{E}(\beta_{inter},\beta_0)$ where $\beta_{inter}$ is the rotation of $R\frac{\beta_0}{\|\beta_0\|_2}$ towards $\beta_0$ so that $\beta_{inter}$ is at the frontier of $B_2(\beta_0,\varepsilon_{max})$. Moreover, $\forall\beta \in \{\nu \in B_2(0,R) : \beta_0^t\nu > 0\}$ we have $\|\beta-\beta_0\|_2^2 \leq 2R^2$. Consequently,

$$\text{for all } \beta \in \left\{\nu \in B_2(0,R) : \beta_0^t\nu > 0\right\} \setminus B_2(\beta_0,\varepsilon_{max}),$$

there exists $\nu \in B_2(\beta_0, \varepsilon_{max})$ such that $\|\nu - \beta_0\|_2 = \varepsilon_{max}$ and

$$\frac{\mathcal{E}(\beta, \beta_0)}{\|\beta - \beta_0\|_2^2} \geq \frac{\mathcal{E}(\beta, \beta_0)}{2R^2} \geq \frac{\mathcal{E}(\nu, \beta_0)}{2R^2}.$$

Furthermmore, from Case 1 above, we have that $\forall \nu \in B_2(\beta_0, \varepsilon_{max})$ such that $\|\nu - \beta_0\|_2 = \varepsilon_{max}$,

$$\mathcal{E}(\nu, \beta_0) \geq \frac{1}{32}\left(1 + (\|a\|_2 - R)^2\right) e^{-\left(\|a\|_2 R - R^2/2\right)} \varepsilon_{max}^2.$$

Hence, for all $\beta \in \{\nu \in B_2(0, R) : \beta_0^t \nu > 0\} \setminus B_2(\beta_0, \varepsilon_{max})$,

$$\frac{\mathcal{E}(\beta, \beta_0)}{\|\beta - \beta_0\|_2^2} \geq \frac{\left(1 + (\|a\|_2 - R)^2\right) e^{-\left(\|a\|_2 R - R^2/2\right)} \varepsilon_{max}^2}{64R^2}.$$

Finally, from the two cases, we get

$$\inf_{\beta \in \left\{\nu \in B_2(0,R) : \beta_0^t \nu > 0\right\}} \frac{\mathcal{E}(\beta, \beta_0)}{\|\beta - \beta_0\|_2^2} \geq \frac{\left(1 + (\|a\|_2 - R)^2\right) e^{-\left(\|a\|_2 R - R^2/2\right)} \varepsilon_{max}^2}{64R^2}.$$

Consequently, the result is also true when one takes the infimum over $\Psi_U$:

$$\begin{aligned}
\inf_{\beta \in \Psi_U} \frac{\mathcal{E}(\beta, \beta_0)}{\|\beta - \beta_0\|_2^2} &\geq \frac{\left(1 + \left(\|a\|_2 - R\right)^2\right) e^{-\left(\|a\|_2 R - R^2/2\right)}}{64R^2} \left(\frac{1}{768} \frac{1 + \left(\|a\|_2 - R\right)^2}{\|a\|_2^4} \exp\left(-R^2/2 - \frac{1 + \left(\|a\|_2 - R\right)^2}{384 \|a\|_2^3 e^{R^2/2}}\right)\right)^2 \\
&= \frac{\left(1 + \left(\|a\|_2 - R\right)^2\right)^3}{9 \cdot 2^{22} \|a\|_2^8 R^2} . \exp\left(-\|a\|_2 R - R^2/2 - R^2 - \frac{1 + \left(\|a\|_2 - R\right)^2}{384 \|a\|_2^3 e^{R^2/2}}\right) \\
&\geq \frac{\left(\|a\|_2 - R\right)^6}{9 \cdot 2^{22} \|a\|_2^8 R^2} . \exp\left(-\|a\|_2 R - R^2/2 - R^2 - R^2/2\right) \\
&\geq \frac{\left(\|a\|_2 - R\right)^6}{9 \cdot 2^{22} \|a\|_2^8 R^2} . \exp\left(-\|a\|_2 R - 2R^2\right).
\end{aligned}$$

□

We present now the proof of our main result, that is the oracle inequality stated in Section 3.4.

*Proof of Theorem 3.2.* We know from Lemma 3.4 that

$$\mathcal{E}\left(\hat{\beta},\beta_0\right)+\lambda\left\|\hat{\beta}\right\|_1\leq\left|V_n\left(\hat{\beta}\right)-V_n\left(\beta_0\right)\right|+\lambda\left\|\beta_0\right\|_1 \tag{3.5.3}$$

We set ourselves in the event $\mathcal{T}$ defined in (3.4.2). It holds

$$\sup_{\beta\in B_2(0,R)}\frac{\left|V_n\left(\beta\right)-V_n\left(\beta_0\right)\right|}{\left\|\beta-\beta_0\right\|_1\vee\lambda_0}\leq 2T\lambda_0$$

and, as $\hat{\beta}\in B_2\left(0,R\right)$, Equation (3.5.3) gives

$$\mathcal{E}\left(\hat{\beta},\beta_0\right)+\lambda\left\|\hat{\beta}\right\|_1\leq 2T\lambda_0\left\|\hat{\beta}-\beta_0\right\|_1\vee\lambda_0+\lambda\left\|\beta_0\right\|_1.$$

Case 1: $\left\|\hat{\beta}-\beta_0\right\|_1\vee\lambda_0=\lambda_0$. We successively have

$$\begin{aligned}\mathcal{E}\left(\hat{\beta},\beta_0\right)&\leq 2T\lambda_0^2+\lambda\left(\left\|\beta_0\right\|_1-\left\|\hat{\beta}\right\|_1\right)\\&\leq 2T\lambda_0^2+\lambda\left|\left\|\beta_0\right\|_1-\left\|\hat{\beta}\right\|_1\right|\\&\leq 2T\lambda_0^2+\lambda\left\|\beta_0-\hat{\beta}\right\|_1.\end{aligned}$$

Hence,

$$\begin{aligned}\mathcal{E}\left(\hat{\beta},\beta_0\right)+2\lambda\left\|\beta_0-\hat{\beta}\right\|_1&\leq 2T\lambda_0^2+3\lambda\left\|\beta_0-\hat{\beta}\right\|_1\\&\leq 2T\lambda_0^2+3\lambda\lambda_0\\&\leq 2\left(T\lambda_0+\lambda\right)^2.\end{aligned}$$

Finally, since $2\left(\lambda-2T\lambda_0\right)\leq 2\lambda$ and $\left\|\hat{\beta}^{S^c}\right\|_1=\left\|\beta_0^{S^c}-\hat{\beta}^{S^c}\right\|_1\leq\left\|\beta_0-\hat{\beta}\right\|_1$

$$\mathcal{E}\left(\hat{\beta},\beta_0\right)+2\left(\lambda-2T\lambda_0\right)\left\|\hat{\beta}^{S^c}\right\|_1\leq 3\left(T\lambda_0+\lambda\right)^2.$$

Case 2: $\left\|\hat{\beta}-\beta_0\right\|_1\vee\lambda_0=\left\|\hat{\beta}-\beta_0\right\|_1$. We have $\left\|\hat{\beta}\right\|_1=\left\|\hat{\beta}^S\right\|_1+\left\|\hat{\beta}^{S^c}\right\|_1$, $\left\|\beta_0\right\|_1=\left\|\beta_0^S\right\|_1$ and $\left\|\hat{\beta}-\beta_0\right\|_1=$

$\left\|\hat{\beta}^S - \beta_0^S\right\|_1 + \left\|\hat{\beta}^{S^c}\right\|_1 = \left\|\hat{\beta}^S - \beta_0\right\|_1 + \left\|\hat{\beta}^{S^c}\right\|_1$. Consequently, it holds successively

$$\begin{aligned}
\mathcal{E}\left(\hat{\beta}, \beta_0\right) + \lambda\left\|\hat{\beta}\right\|_1 &\leq 2T\lambda_0\left\|\hat{\beta} - \beta_0\right\|_1 + \lambda\left\|\beta_0\right\|_1, \\
\mathcal{E}\left(\hat{\beta}, \beta_0\right) + \lambda\left\|\hat{\beta}^S\right\|_1 + \lambda\left\|\hat{\beta}^{S^c}\right\|_1 &\leq 2T\lambda_0\left\|\hat{\beta}^S - \beta_0\right\|_1 + 2T\lambda_0\left\|\hat{\beta}^{S^c}\right\|_1 + \lambda\left\|\beta_0^S\right\|_1, \\
\mathcal{E}\left(\hat{\beta}, \beta_0\right) + \lambda\left\|\hat{\beta}^{S^c}\right\|_1 - 2T\lambda_0\left\|\hat{\beta}^{S^c}\right\|_1 &\leq 2T\lambda_0\left\|\hat{\beta}^S - \beta_0\right\|_1 + \lambda\left\|\beta_0^S\right\|_1 - \lambda\left\|\hat{\beta}^S\right\|_1, \\
\mathcal{E}\left(\hat{\beta}, \beta_0\right) + (\lambda - 2T\lambda_0)\left\|\hat{\beta}^{S^c}\right\|_1 &\leq 2T\lambda_0\left\|\hat{\beta}^S - \beta_0\right\|_1 + \lambda\left\|\beta_0^S - \hat{\beta}^S\right\|_1, \\
\mathcal{E}\left(\hat{\beta}, \beta_0\right) + (\lambda - 2T\lambda_0)\left\|\hat{\beta}^{S^c}\right\|_1 &\leq (2T\lambda_0 + \lambda)\left\|\beta_0 - \hat{\beta}^S\right\|_1.
\end{aligned}$$

Since $\beta_0 - \hat{\beta}^S$ has at most $s$ non-zero coordinates, one has $\left\|\beta_0 - \hat{\beta}^S\right\|_1 \leq \sqrt{s}\left\|\beta_0 - \hat{\beta}^S\right\|_2 \leq \sqrt{s}\left\|\beta_0 - \hat{\beta}\right\|_2$. Hence, for any $c_0 > 0$,

$$\mathcal{E}\left(\hat{\beta}, \beta_0\right) + (\lambda - 2T\lambda_0)\left\|\hat{\beta}^{S^c}\right\|_1 \leq (T\lambda_0 + \lambda)\sqrt{\frac{s}{c_0}}\sqrt{c_0}\left\|\beta_0 - \hat{\beta}\right\|_2.$$

Now use the fact that $\forall a, b, 2ab \leq a^2 + b^2$ to get

*Proof.*

$$\mathcal{E}\left(\hat{\beta}, \beta_0\right) + (\lambda - 2T\lambda_0)\left\|\hat{\beta}^{S^c}\right\|_1 \leq (T\lambda_0 + \lambda)^2\frac{s}{2c_0} + \frac{c_0\left\|\beta_0 - \hat{\beta}\right\|_2^2}{2}.$$

So we can use Lemma 3.1 and have $\mathcal{E}\left(\hat{\beta}, \beta_0\right) \geq c_0\left\|\beta_0 - \hat{\beta}\right\|_2^2$ where $c_0 = \frac{\left(\|a\|_2 - R\right)^6}{9 \cdot 2^{22}\|a\|_2^8 R^2} e^{-\|a\|_2 R - 2R^2}$. Consequently, for this choice of $c_0$,

$$\mathcal{E}\left(\hat{\beta}, \beta_0\right) + (\lambda - 2T\lambda_0)\left\|\hat{\beta}^{S^c}\right\|_1 \leq (T\lambda_0 + \lambda)^2\frac{s}{2c_0} + \frac{\mathcal{E}\left(\hat{\beta}, \beta_0\right)}{2},$$

which gives

$$\mathcal{E}\left(\hat{\beta}, \beta_0\right) + 2(\lambda - 2T\lambda_0)\left\|\hat{\beta}^{S^c}\right\|_1 \leq (T\lambda_0 + \lambda)^2 . \frac{s}{c_0}.$$

Finally, combining the two cases, we obtain

$$\mathcal{E}\left(\hat{\beta},\beta_0\right)+2\left(\lambda-2T\lambda_0\right)\left\|\hat{\beta}^{S^c}\right\|_1\leq\left(T\lambda_0+\lambda\right)^2.\max\left(\frac{s}{c_0},2\right).$$

The bound on the probability of the event $\mathcal{T}$ is given in Theorem 3.3. □

□

### 3.5.2 Auxiliary results

Let us first state the following basic lemma, where we compute the derivatives of the loss and its risk.

**Lemma 3.2.** *With notations of section 3.2, it holds*

$$\begin{aligned}
\rho_\beta(X) &= -\log q_\beta(X)+X^t\beta.p_\beta(X) && (3.5.4)\\
\rho_\beta(X) &= \log\left(1+e^{X^t\beta}\right)-\frac{X^t\beta e^{X^t\beta}}{1+e^{X^t\beta}} && (3.5.5)\\
\frac{\partial p_\beta(X)}{\partial\beta_u} &= -X_u p_\beta(X)\,q_\beta(X) && (3.5.6)\\
\frac{\partial q_\beta(X)}{\partial\beta_u} &= X_u p_\beta(X)\,q_\beta(X) && (3.5.7)\\
\frac{\partial\rho_\beta(X)}{\partial\beta_u} &= -X^t\beta X_u p_\beta(X)\,q_\beta(X) && (3.5.8)\\
\frac{\partial}{\partial\beta_v}\frac{\partial}{\partial\beta_u}\rho_\beta(X) &= =-X_vX_u\alpha\left(X^t\beta\right) && (3.5.9)\\
\frac{\partial}{\partial\beta_w}\frac{\partial}{\partial\beta_v}\frac{\partial}{\partial\beta_u}\rho_\beta(X) &= -X_wX_vX_u\alpha'\left(X^t\beta\right) && (3.5.10)\\
\left(d_\beta\mathcal{R}\right)(h) &= \mathbb{E}\left[-X^t\beta.p_\beta(X)\,q_\beta(X)\,X^th\right] && (3.5.11)\\
\left(d^2_\beta\mathcal{R}\right)(h,k) &= \mathbb{E}\left[X^th.X^tk.\alpha\left(X^t\beta\right)\right] && (3.5.12)\\
\left(d^3_\beta\mathcal{R}\right)(h,k,l) &= \mathbb{E}\left[X^th.X^tk.X^tl.\alpha'\left(X^t\beta\right)\right] && (3.5.13)
\end{aligned}$$

*Proof.* Consider $X,\beta\in\mathbb{R}^d$, $\rho_\beta(X)$ is defined in section 3.2. For simplicity, $p$ and $q$ stand for $p_\beta(X)$ and $q_\beta(X)$

recall that $p_\beta(X) = q_\beta(X)\, e^{-X\beta}$:

$$\begin{aligned}
\rho_\beta(X) &= -p\log p - q\log q \\
&= -\left(qe^{-X\beta}\right)\log\left(qe^{-X\beta}\right) - q\log q \\
&= -qe^{-X^t\beta}\log q + X^t\beta qe^{-X^t\beta} - q\log q \\
&= -q\left(1+e^{-X^t\beta}\right)\log q + X^t\beta qe^{-X^t\beta} \\
&= -\log q + X^t\beta p \\
&= \log\left(1+e^{X^t\beta}\right) - \frac{X^t\beta e^{X^t\beta}}{1+e^{X^t\beta}}.
\end{aligned}$$

Denote $\beta_u$ the $u$-th component of $\beta$. We have,

$$\begin{aligned}
\frac{\partial p_\beta(X)}{\partial\beta_u} &= \frac{\partial}{\partial\beta_u}\left[\frac{1}{1+e^{X^t\beta}}\right] = -\frac{X_u e^{X^t\beta}}{\left(1+e^{X^t\beta}\right)^2} \\
&= -X_u p_\beta(X)\, q_\beta(X)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial q_\beta(X)}{\partial\beta_u} &= \frac{\partial}{\partial\beta_u}\left[1-p_\beta(X)\right] = -\frac{\partial p_\beta(X)}{\partial\beta_u} \\
&= X_u p_\beta(X)\, q_\beta(X).
\end{aligned}$$

Secondly, we use Equation (3.5.5) to have

$$\begin{aligned}\frac{\partial \rho_\beta(X)}{\partial \beta_u} &= -\frac{\partial \log q}{\partial \beta_u} + \frac{\partial (X^t\beta p)}{\partial \beta_u} \\ &= -\frac{\frac{\partial q}{\partial \beta_u}}{q} + \frac{\partial (X^t\beta)}{\partial \beta_u} p + X^t\beta \frac{\partial p}{\partial \beta_u} \\ &= -\frac{(X_u pq)}{q} + X_u p + X^t\beta (-X_u pq) \\ &= -X^t\beta X_u p_\beta(X) q_\beta(X).\end{aligned}$$

The second derivatives are

$$\begin{aligned}\frac{\partial}{\partial \beta_v}\frac{\partial}{\partial \beta_u}\rho_\beta(X) &= -\frac{\partial (X\beta)}{\partial \beta_v} X_u pq - X^t\beta X_u \frac{\partial p}{\partial \beta_v} q - X^t\beta X_u p \frac{\partial q}{\partial \beta_v} \\ &= -(X_v) X_u pq - X^t\beta X_u (-X_v pq) q - X^t\beta X_u p (X_v qp) \\ &= -X_v X_u pq \left(1 - X^t\beta (q-p)\right) \\ &= -X_v X_u \frac{e^{X^t\beta}}{\left(1+e^{X^t\beta}\right)^2}\left(1 + X^t\beta\left(\frac{1-e^{X^t\beta}}{1+e^{X^t\beta}}\right)\right) \\ &= X_v X_u \alpha\left(X^t\beta\right).\end{aligned}$$

The third derivatives are

$$\begin{aligned}\frac{\partial}{\partial \beta_w}\frac{\partial}{\partial \beta_v}\frac{\partial}{\partial \beta_u}\rho_\beta(X) &= X_v X_u \frac{\partial}{\partial \beta_w}\left[\alpha\left(X^t\beta\right)\right] \\ &= X_w X_v X_u \alpha'\left(X^t\beta\right).\end{aligned}$$

As the derivatives are uniformly bounded with respect to $\beta$, the theorem of derivation under integral can be applied

and it comes that $\forall h, k, l, \in \mathbb{R}^d$,

$$\begin{aligned}
\left(d_\beta \mathcal{R}\right)(h) &= \mathbb{E}\left[-X^t\beta.p_\beta\left(X\right)q_\beta\left(X\right)X^t h\right], \\
\left(d^2_\beta \mathcal{R}\right)(h,k) &= \mathbb{E}\left[X^t h.X^t k.\alpha\left(X^t\beta\right)\right], \\
\left(d^3_\beta \mathcal{R}\right)(h,k,l) &= \mathbb{E}\left[X^t h.X^t k.X^t l.\alpha'\left(X^t\beta\right)\right].
\end{aligned}$$

□

**Lemma 3.3.** *Take $r > 0$. For any function $g$ odd on $\mathbb{R}$, positive on $(0, +\infty)$ and when $U$ is a symetric random variable with a density $\gamma$ decreasing on $\mathbb{R}_+$, the quantity $\mathbb{E}\left[g\left(\mu + rU\right)\right]$ has the sign of $\mu$.*

*Proof.* Take $r, \mu > 0$, $U_1$ and $U_2$ two independent copies of $U$. It holds

$$\begin{aligned}
\mathbb{E}\left[g\left(\mu + rU\right)\right] &= \mathbb{E}\left[g\left(\mu + rU_1\right)I_{U_1>0} + g\left(\mu - rU_2\right)I_{U_2>0}\right] \\
&= \mathbb{E}\left[g\left(\mu + rU_1\right)I_{U_1>0} + \right. \\
&\qquad \left. g\left(\mu - rU_2\right)I_{0<U_2<\frac{\mu}{r}} + g\left(\mu - rU_2\right)I_{\frac{\mu}{r}<U_2<2\frac{\mu}{r}} + g\left(\mu - rU_2\right)I_{2\frac{\mu}{r}<U_2}\right] \\
&= \mathbb{E}\left[g\left(\mu + rU_1\right)I_{U_1>0} + g\left(\mu - rU_2\right)I_{2\frac{\mu}{r}<U_2}\right. \\
&\qquad \left. + g\left(\mu - rU_1\right)I_{0<U_2<\frac{\mu}{r}} + g\left(\mu - rU_2\right)I_{\frac{\mu}{r}<U_2<2\frac{\mu}{r}}\right] \\
&= \mathbb{E}\left[g\left(\mu + rU_1\right)I_{U_1>0} + g\left(\mu - rU_2\right)I_{2\frac{\mu}{r}<U_2}\right] \\
&\qquad + \mathbb{E}\left[g\left(\mu - rU_1\right)I_{0<U_2<\frac{\mu}{r}} + g\left(\mu - rU_2\right)I_{\frac{\mu}{r}<U_2<2\frac{\mu}{r}}\right].
\end{aligned}$$

Let us compute the sign of $\mathbb{E}\left[g\left(\mu + rU_1\right)I_{U_1>0} + g\left(\mu - rU_2\right)I_{2\frac{\mu}{r}<U_2}\right]$:

$$
\begin{aligned}
&\mathbb{E}\left[g\left(\mu+rU_1\right)I_{U_1>0}+g\left(\mu-rU_2\right)I_{2\frac{\mu}{r}<U_2}\right]\\
&=\int_0^\infty g\left(\mu+rx\right)\gamma\left(x\right)dx+\underbrace{\int_{2\frac{\mu}{r}}^\infty g\left(\mu-rx\right)\gamma\left(x\right)dx}_{x=y+\frac{2\mu}{r}}\\
&=\int_0^\infty g\left(\mu+rx\right)\gamma\left(x\right)dx+\int_0^\infty g\left(\mu-r\left(y+\frac{2\mu}{r}\right)\right)\gamma\left(y+\frac{2\mu}{r}\right)dy\\
&=\int_0^\infty g\left(\mu+rx\right)\gamma\left(x\right)dx+\int_0^\infty g\left(-ry-\mu\right)\gamma\left(y+\frac{2\mu}{r}\right)dy\\
&=\int_0^\infty g\left(\mu+rx\right)\gamma\left(x\right)dx+\int_0^\infty -g\left(\mu+ry\right)\gamma\left(y+\frac{2\mu}{r}\right)dy\\
&=\int_0^\infty \underbrace{g\left(\mu+rx\right)}_{>0}\underbrace{\left[\gamma\left(x\right)-\gamma\left(x+\frac{2\mu}{r}\right)\right]}_{>0}dx\\
&>0.
\end{aligned}
$$

Let us now compute the sign of $\mathbb{E}\left[g\left(\mu-rU_1\right)I_{0<U_2<\frac{\mu}{r}}+g\left(\mu-rU_2\right)I_{\frac{\mu}{r}<U_2<2\frac{\mu}{r}}\right]$:

$$
\begin{aligned}
&\mathbb{E}\left[g\left(\mu - rU_1\right) I_{0<U_2<\frac{\mu}{r}} + g\left(\mu - rU_2\right) I_{\frac{\mu}{r}<U_2<2\frac{\mu}{r}}\right] \\
&= \int_0^{\frac{\mu}{r}} g\left(\mu - rx\right)\gamma\left(x\right)dx + \underbrace{\int_{\frac{\mu}{r}}^{\frac{2\mu}{r}} g\left(\mu - rx\right)\gamma\left(x\right)dx}_{x=\frac{2\mu}{r}-y} \\
&= \int_0^{\frac{\mu}{r}} g\left(\mu - rx\right)\gamma\left(x\right)dx + \int_0^{\frac{\mu}{r}} g\left(\mu - r\left(\frac{2\mu}{r} - y\right)\right)\gamma\left(\frac{2\mu}{r} - y\right)dy \\
&= \int_0^{\frac{\mu}{r}} g\left(\mu - rx\right)\gamma\left(x\right)dx + \int_0^{\frac{\mu}{r}} g\left(ry - \mu\right)\gamma\left(\frac{2\mu}{r} - y\right)dy \\
&= \int_0^{\frac{\mu}{r}} g\left(\mu - rx\right)\gamma\left(x\right)dx + \int_0^{\frac{\mu}{r}} -g\left(\mu - ry\right)\gamma\left(\frac{2\mu}{r} - y\right)dy \\
&= \int_0^{\frac{\mu}{r}} \underbrace{\underbrace{g(\mu - rx)}_{>0}}_{>0} \underbrace{\left[\gamma\left(x\right) - \gamma\underbrace{\left(\frac{2\mu}{r} - x\right)}_{\frac{2\mu}{r}-x>\frac{\mu}{r}>x}\right]}_{>0} dx \\
&> 0
\end{aligned}
$$

Hence $\mathbb{E}\left[g\left(\mu + rU\right)\right] > 0$. If $\mu < 0$, then one has $\mathbb{E}\left[g\left(\mu + rU\right)\right] = -\mathbb{E}\left[g\left(-\mu - rU\right)\right]$ and the previous result applies since $-\mu > 0$ and $-U \sim U$. Thus we find that if $\mu < 0, \mathbb{E}\left[g\left(\mu + rU\right)\right] < 0$. □

**Theorem 3.3.** *Set* $\Theta = B_2\left(0, R\right)$, $M_n = \left\|a\right\|_\infty + \sqrt{2\log d} + \sqrt{2\log\left(1+n\right)}$ *and*

$$\lambda_0 = 3n^{-1/2} L M_n \left(5\sqrt{3\log\left(2d\right)}\log n + 4\right).$$

*It holds* $\forall n \geq 2$, $\forall T \geq 1$,

$$P\left(\sup_{\beta\in\Theta} \frac{\left|V_n\left(\beta\right) - V_n\left(\beta_0\right)\right|}{\left\|\beta - \beta_0\right\|_1 \vee \lambda_0} > 2T\lambda_0\right) \leq \frac{3}{4}\log\left(\frac{4R^2 nd}{L^2 M_n^2}\right)\exp\left(-21\left(T-1\right)^2\log\left(2d\right)\log^2 n\right) + \frac{1}{25T^2\log\left(2d\right)n\log^2 n}.$$

*Proof.* First, the triangular inequality gives

$$\left|V_n\left(\beta\right) - V_n\left(\beta_0\right)\right| \leq \left|V_n^{trunc}\left(\beta\right) - V_n^{trunc}\left(\beta_0\right)\right| + \left|V_n^{trunc}\left(\beta\right) - V_n^{trunc}\left(\beta_0\right) - \left(V_n\left(\beta\right) - V_n\left(\beta_0\right)\right)\right|,$$

and since $\forall a, b, t > 0$ on has "$a + b > 2t$" implies "either $a > t$ or $b > t$", the probability of interest can be controlled as follows:

$$P\left(\sup_{\beta\in\Theta}\frac{|V_n(\beta)-V_n(\beta_0)|}{\|\beta-\beta_0\|_1\vee\lambda_0}>2T\lambda_0\right)\leq P\left(\sup_{\beta\in\Theta}\frac{|V_n^{trunc}(\beta)-V_n^{trunc}(\beta_0)|}{\|\beta-\beta_0\|_1\vee\lambda_0}>T\lambda_0\right)$$
$$+P\left(\sup_{\beta\in\Theta}\frac{|V_n^{trunc}(\beta)-V_n^{trunc}(\beta_0)-(V_n(\beta)-V_n(\beta_0))|}{\|\beta-\beta_0\|_1\vee\lambda_0}>T\lambda_0\right).$$

Apply now Lemma 3.29 to have:

$$P\left(\sup_{\beta\in\Theta}\frac{|V_n(\beta)-V_n(\beta_0)|}{\|\beta-\beta_0\|_1\vee\lambda_0}>2T\lambda_0\right)\leq P\left(\sup_{\beta\in\Theta}\frac{|V_n^{trunc}(\beta)-V_n^{trunc}(\beta_0)|}{\|\beta-\beta_0\|_1\vee\lambda_0}>T\lambda_0\right)$$
$$+P\left(\frac{1}{n}\sum_{i=1}^{n}F\left(X^{(i)}\right)>\frac{T\lambda_0}{L}\right).$$

From Lemmas 3.26 and 3.28, we get

$$P\left(\sup_{\beta\in\Theta}\frac{|V_n(\beta)-V_n(\beta_0)|}{\|\beta-\beta_0\|_1\vee\lambda_0}>2T\lambda_0\right)\leq\frac{3}{4}\log\left(\frac{4R^2nd}{L^2M_n^2}\right)\exp\left(-21(T-1)^2\log(2d)\log^2 n\right)+4L^2\frac{M_n^2+\|a\|_\infty+1}{n^2\lambda_0^2T^2}.$$

Furthermore,

$$\begin{aligned}4L^2\frac{M_n^2+\|a\|_\infty+1}{n^2\lambda_0^2T^2}&\leq 4L^2\frac{M_n^2+\|a\|_\infty+1}{n^2\frac{9L^2M_n^2\left(5\sqrt{3\log(2d)}\log n+4\right)^2}{n}T^2}\\&\leq 4\frac{1+\frac{\|a\|_\infty+1}{M_n^2}}{9\times25\left(3\log(2d)\log^2 n\right)nT^2}\\&\leq\frac{1}{25T^2\log(2d)\,n\log^2 n}.\end{aligned}$$

Hence,

$$P\left(\sup_{\beta\in\Theta}\frac{|V_n(\beta)-V_n(\beta_0)|}{\|\beta-\beta_0\|_1\vee\lambda_0}>2T\lambda_0\right)\leq\frac{3}{4}\log\left(\frac{4R^2nd}{L^2M_n^2}\right)\exp\left(-21(T-1)^2\log(2d)\log^2 n\right)+\frac{1}{25T^2\log(2d)\,n\log^2 n}.$$

□

**Lemma 3.4.** *Recall that* $\beta_0 = \underset{\beta \in \Psi_U}{\arg\min} \{\mathcal{R}(\beta)\}$ *and* $\hat{\beta} := \underset{\beta \in \Psi_U}{\arg\min} \left\{ \hat{\mathcal{R}}_n(\beta) + \lambda \|\beta\|_1 \right\}$. *It holds*

$$\mathcal{E}\left(\hat{\beta}, \beta_0\right) + \lambda \left\|\hat{\beta}\right\|_1 \leq \left| V_n\left(\hat{\beta}\right) - V_n(\beta_0) \right| + \lambda \|\beta_0\|_1 .$$

*Proof.* By definition of $\hat{\beta}$, we have:

$$\hat{\mathcal{R}}_n\left(\hat{\beta}\right) + \lambda \left\|\hat{\beta}\right\|_1 \leq \hat{\mathcal{R}}_n(\beta_0) + \lambda \|\beta_0\|_1 .$$

Injecting the excess risk on both sides of the inequality gives

$$\mathcal{E}\left(\hat{\beta}, \beta_0\right) + \lambda \left\|\hat{\beta}\right\|_1 \leq \mathcal{R}(\hat{\beta}) - \mathcal{R}(\beta_0) + \hat{\mathcal{R}}_n(\beta_0) - \hat{\mathcal{R}}_n\left(\hat{\beta}\right) + \lambda \|\beta_0\|_1$$

Then the result comes from the inequality:

$$\mathcal{R}(\hat{\beta}) - \mathcal{R}(\beta_0) + \hat{\mathcal{R}}_n(\beta_0) - \hat{\mathcal{R}}_n\left(\hat{\beta}\right) \leq \left| V_n\left(\hat{\beta}\right) - V_n(\beta_0) \right| .$$

□

### 3.5.3 Some further technical lemmas

**Lemma 3.5.** *Assuming* $a \in \mathbb{R}^d$, $Z \sim a + N$, $N \sim \mathcal{N}(0, I_d)$, $\beta_0 = Ra/\|a\|_2$, *and* $h_\perp \in \beta_0^\perp$, *then* $h_\perp^t Z$ *and* $Z^t \beta_0$ *are two independent Gaussian variables.*

*Proof.* Note that $h_\perp^t a = 0$. We have

$$
\begin{aligned}
Cov\left(h_\perp^t Z, Z^t\beta_0\right) &= \mathbb{E}\left[\left(h_\perp^t Z - \mathbb{E}\left[h_\perp^t Z\right]\right)\left(Z^t\beta_0 - \mathbb{E}\left[Z^t\beta_0\right]\right)\right] \\
&= \mathbb{E}\left[\left(h_\perp^t\left(a+N\right) - \mathbb{E}\left[h_\perp^t\left(a+N\right)\right]\right)\left(\left(a+N\right)^t\beta_0 - \mathbb{E}\left[\left(a+N\right)^t\beta_0\right]\right)\right] \\
&= \mathbb{E}\left[\left(h_\perp^t N - \mathbb{E}\left[h_\perp^t N\right]\right)\left(a^t\beta_0 + N^t\beta_0 - \mathbb{E}\left[a^t\beta_0 + N^t\beta_0\right]\right)\right] \\
&= \mathbb{E}\left[\left(h_\perp^t N - h_\perp^t\mathbb{E}\left[N\right]\right)\left(N^t\beta_0 - \mathbb{E}\left[N^t\right]\beta_0\right)\right] \\
&= h_\perp^t\mathbb{E}\left[NN^t\right]\beta_0 \\
&= h_\perp^t\beta_0 \\
&= 0.
\end{aligned}
$$

□

**Lemma 3.6.** *With $Z$, $\beta_0$, $\alpha$ and $\mathcal{R}$ usual notations and for all $h := h_\parallel + h_\perp \in Vect(\beta_0)\oplus\beta_0^\perp$ such that $\|h\|_2 = 1$ and with $\eta := \|h_\parallel\|_2$, we have*

$$
\left(d^2_{\beta_0}\mathcal{R}\right)(h,h) = \eta^2\mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2\alpha\left(Z^t\beta_0\right)\right] + \left(1-\eta^2\right)\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\right].
$$

*Proof.* We computed $d^2_{\beta_0}\mathcal{R}$ in Equation (3.5.12) of Lemma 3.2. The function $\alpha$ (see Section 3.2) is even, so the entries of the Hessian $d^2_{\beta_0}\mathcal{R}$ are

$$
\begin{aligned}
\forall u,v\in[\![1,d]\!],\left(d^2_{\beta_0}\mathcal{R}\right)_{u,v} &= \mathbb{E}\left[\left(\varepsilon Z_v\right)\left(\varepsilon Z_u\right)\alpha\left(\varepsilon Z^t\beta_0\right)\right] \\
&= \mathbb{E}\left[Z_v Z_u\alpha\left(Z^t\beta_0\right)\right].
\end{aligned}
$$

Now, let us use the decomposition $h = h_\parallel + h_\perp$ and remark that $h_\parallel = \epsilon\eta\frac{\beta_0}{\|\beta_0\|_2} = \epsilon\frac{\eta}{R}\beta_0$ with $\epsilon\in\{-1,1\}$. It comes

$$
\begin{aligned}
\left(d^2_{\beta_0}\mathcal{R}\right)(h,h) &= \mathbb{E}\left[\left(h^tZ\right)^2\alpha\left(Z^t\beta_0\right)\right] \\
&= \mathbb{E}\left[\left(\left(h_\parallel + h_\perp\right)^tZ\right)^2\alpha\left(Z^t\beta_0\right)\right] \\
&= \mathbb{E}\left[\left(\left(h_\parallel^tZ\right)^2 + 2.h_\parallel^tZ.h_\perp^tZ + \left(h_\perp^tZ\right)^2\right)\alpha\left(Z^t\beta_0\right)\right] \\
&= \mathbb{E}\left[\left(h_\parallel^tZ\right)^2\alpha\left(Z^t\beta_0\right)\right] + 2\mathbb{E}\left[h_\parallel^tZ.h_\perp^tZ.\alpha\left(Z^t\beta_0\right)\right] + \mathbb{E}\left[\left(h_\perp^tZ\right)^2\alpha\left(Z^t\beta_0\right)\right] \\
&= \mathbb{E}\left[\left(\epsilon\frac{\eta}{R}\beta_0^tZ\right)^2\alpha\left(Z^t\beta_0\right)\right] + 2\mathbb{E}\left[h_\parallel^tZ.h_\perp^tZ.\alpha\left(Z^t\beta_0\right)\right] + \mathbb{E}\left[\left(h_\perp^tZ\right)^2\alpha\left(Z^t\beta_0\right)\right] \\
&= \eta^2\mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2\alpha\left(Z^t\beta_0\right)\right] + 2\mathbb{E}\left[h_\parallel^tZ.h_\perp^tZ.\alpha\left(Z^t\beta_0\right)\right] + \mathbb{E}\left[\left(h_\perp^tZ\right)^2\alpha\left(Z^t\beta_0\right)\right].
\end{aligned}
$$

Also remark that $h_\perp^tZ$ and $Z^t\beta_0$ are Gaussian random variables, because $Z$ is a Gaussian vector, that are independent due to lemma 3.5.

$$
\begin{aligned}
\left(d^2_{\beta_0}\mathcal{R}\right)(h,h) &= \eta^2\mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2\alpha\left(Z^t\beta_0\right)\right] + 2\mathbb{E}\left[h_\parallel^tZ.h_\perp^tZ.\alpha\left(Z^t\beta_0\right)\right] + \mathbb{E}\left[\left(h_\perp^tZ\right)^2\alpha\left(Z^t\beta_0\right)\right] \\
&= \eta^2\mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2\alpha\left(Z^t\beta_0\right)\right] + 2\mathbb{E}\left[h_\parallel^tZ.\alpha\left(Z^t\beta_0\right)\right]\underbrace{\mathbb{E}\left[h_\perp^tZ\right]}_{=0} + \mathbb{E}\left[\left(h_\perp^tZ\right)^2\right]\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\right] \\
&= \eta^2\mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2\alpha\left(Z^t\beta_0\right)\right] + \mathbb{E}\left[\left(h_\perp^tZ\right)^2\right]\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\right].
\end{aligned}
$$

Note that $\mathbb{E}\left[\left(h_\perp^tZ\right)^2\right] = \mathbb{E}\left[\left(h_\perp^t\left(a+N\right)\right)^2\right] = \mathbb{E}\left[\left(h_\perp^tN\right)^2\right] = h_\perp^t\mathbb{E}\left[NN^t\right]h_\perp = 1-\eta^2$ hence:

$$
\left(d^2_{\beta_0}\mathcal{R}\right)(h,h) = \eta^2\mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2\alpha\left(Z^t\beta_0\right)\right] + \left(1-\eta^2\right)\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\right].
$$

□

**Lemma 3.7.** *For all $h := h_\parallel + h_\perp \in Vect(\beta_0)\oplus\beta_0^\perp$ such that $\|h\|_2 = 1$ and with $\eta := \left\|h_\parallel\right\|_2$. The two following*

*quantities*

$$A := \eta^2 \mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2 \alpha\left(Z^t\beta_0\right) \mathbb{I}_{\{Z^t\beta_0 > x_1\}}\right] + \left(1-\eta^2\right)\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0 > x_1\}}\right]$$
$$B := \eta^2 \mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2 \alpha\left(Z^t\beta_0\right) \mathbb{I}_{\{-x_1 < Z^t\beta_0 < x_1\}}\right] + \left(1-\eta^2\right)\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{-x_1 < Z^t\beta_0 < x_1\}}\right]$$

*are controlled by*

$$A > \left(\eta^2\frac{x_1^2}{R^2} + 1 - \eta^2\right)\left[R + \left[R\left(\|a\|_2 - R\right) - \left(x_1 + \frac{8}{100}\right)\right] G\left(\frac{x_1}{R} + R - \|a\|_2\right)\right]\gamma\left(\|a\|_2 - \frac{x_1}{R}\right)e^{-x_1},$$
$$B \leq \frac{1}{4}\left(\eta^2\frac{x_1^2}{R^2} + 1 - \eta^2\right)\left(\Phi^c\left(\|a\|_2 - \frac{x_1}{R}\right) - \Phi^c\left(\|a\|_2 + \frac{x_1}{R}\right)\right).$$

*Proof.* Let us first give an upper bound for the quantity $B$. Recall that, from Lemma 3.9 we have $-\alpha(x) \in \left(0, \frac{1}{4}\right)$ for $x \in [-x_1, x_1]$. It holds

$$\begin{aligned}
B &= -\eta^2\mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2 \alpha\left(Z^t\beta_0\right) \mathbb{I}_{\{-x_1 < Z^t\beta_0 < x_1\}}\right] - \left(1-\eta^2\right)\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{-x_1 < Z^t\beta_0 < x_1\}}\right] \\
&\leq \eta^2\mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2 \frac{1}{4}\mathbb{I}_{\{-x_1 < Z^t\beta_0 < x_1\}}\right] + \left(1-\eta^2\right)\mathbb{E}\left[\frac{1}{4}\mathbb{I}_{\{-x_1 < Z^t\beta_0 < x_1\}}\right] \\
&= \frac{1}{4}\left(\eta^2\frac{x_1^2}{R^2} + 1 - \eta^2\right)\mathbb{P}\left[-x_1 < Z^t\beta_0 < x_1\right] \\
&= \frac{1}{4}\left(\eta^2\frac{x_1^2}{R^2} + 1 - \eta^2\right)\mathbb{P}\left[-x_1 < R\|a\|_2 + R\mathcal{N}(0,1) < x_1\right] \\
&= \frac{1}{4}\left(\eta^2\frac{x_1^2}{R^2} + 1 - \eta^2\right)\mathbb{P}\left[-\frac{x_1}{R} - \|a\|_2 < \mathcal{N}(0,1) < \frac{x_1}{R} - \|a\|_2\right] \\
&= \frac{1}{4}\left(\eta^2\frac{x_1^2}{R^2} + 1 - \eta^2\right)\mathbb{P}\left[\|a\|_2 - \frac{x_1}{R} < \mathcal{N}(0,1) < \frac{x_1}{R} + \|a\|_2\right] \\
&= \frac{1}{4}\left(\eta^2\frac{x_1^2}{R^2} + 1 - \eta^2\right)\left(\Phi^c\left(\|a\|_2 - \frac{x_1}{R}\right) - \Phi^c\left(\|a\|_2 + \frac{x_1}{R}\right)\right).
\end{aligned}$$

Let us now turn to the lower bound for the quantity $A$:

$$\begin{aligned}
A &= \eta^2 \mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2 \alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0>x_1\}}\right] + \left(1-\eta^2\right)\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0>x_1\}}\right] \\
&\geq \eta^2 \mathbb{E}\left[\frac{1}{R^2}x_1^2\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0>x_1\}}\right] + \left(1-\eta^2\right)\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0>x_1\}}\right] \\
&= \eta^2\frac{x_1^2}{R^2}\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0>x_1\}}\right] + \left(1-\eta^2\right)\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0>x_1\}}\right] \\
&= \left(\eta^2\frac{x_1^2}{R^2} + 1 - \eta^2\right)\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0>x_1\}}\right].
\end{aligned}$$

We need now to control $\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0>x_1\}}\right]$ from below. We first use Lemma 3.13 to get:

$$\begin{aligned}
&\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_R>x_1\}}\right] \\
&\geq \int_{x_1}^{\infty}\left(\left(x - x_1 - \frac{8}{100}\right)e^{-x}\frac{e^{-\left(x/R-\|a\|_2\right)^2/2}}{\sqrt{2\pi R^2}}\right)dx \\
&= \int_{x_1}^{\infty} x e^{-x}\frac{e^{-(x/R-\|a\|)^2/2}}{\sqrt{2\pi R^2}}dx - \left(x_1+\frac{8}{100}\right)\int_{x_1}^{\infty} e^{-x}\frac{e^{-(x/R-\|a\|)^2/2}}{\sqrt{2\pi R^2}}dx.
\end{aligned}$$

Using the notations of Lemmas 3.14 and 3.15 we obtain,

$$\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0>x_1\}}\right] \geq J_{a,R}\left(1,x_1\right) - \left(x_1+\frac{8}{100}\right)K_{a,R}\left(1,x_1\right). \tag{3.5.14}$$

Hence, Lemmas 3.14 and 3.15 give:

$$\begin{aligned}
&\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0>x_1\}}\right] \\
&\geq R\left(1+\left(\|a\|-R\right)G\left(\frac{x_1}{R}+R-\|a\|\right)\right)\gamma\left(\frac{x_1}{R}-\|a\|\right)e^{-x_1} - \left(x_1+\frac{8}{100}\right)\gamma\left(\frac{x_1}{R}-\|a\|\right)G\left(\frac{x_1}{R}+R-\|a\|\right)e^{-x_1} \\
&\geq \left[R + \left[R\left(\|a\|-R\right) - \left(x_1+\frac{8}{100}\right)\right]G\left(\frac{x_1}{R}+R-\|a\|\right)\right]\gamma\left(\|a\|-\frac{x_1}{R}\right)e^{-x_1}.
\end{aligned}$$

Finally,

$$A > \left(\eta^2 \frac{x_1^2}{R^2} + 1 - \eta^2\right) \left[R + \left[R\left(\|a\| - R\right) - \left(x_1 + \frac{8}{100}\right)\right] G\left(\frac{x_1}{R} + R - \|a\|\right)\right] \gamma\left(\|a\| - \frac{x_1}{R}\right) e^{-x_1}.$$

□

**Lemma 3.8.** *Take $a \in \mathbb{R}^d$, $R, \nu > 0$ and $\beta_0 := R\frac{a}{\|a\|_2}$ if inequality*

$$R\left(1 - \left(R - \|a\|_2 + \frac{x_1 + \frac{8}{100}}{R}\right) G\left(\frac{x_1}{R} + R - \|a\|_2\right)\right) \geq (1+\nu)\frac{e^{x_1}}{4} G\left(\|a\|_2 - \frac{x_1}{R}\right) \tag{3.5.15}$$

*is true, then for all $h := h_{\|} + h_{\perp} \in Vect(\beta_0) \oplus \beta_0^{\perp}$ such that $\|h\|_2 = 1$ and $\eta := \left\|h_{\|}\right\|_2$, it also holds*

$$\left(d^2_{\beta_0}\mathcal{R}\right)(h,h) > \frac{\nu}{4}\left(\eta^2\frac{x_1^2}{R^2} + 1 - \eta^2\right)\left(\Phi^c\left(\|a\|_2 - \frac{x_1}{R}\right) - \Phi^c\left(\|a\|_2 + \frac{x_1}{R}\right)\right).$$

*Proof.* Recall that$\beta_0 := Ra/\|a\|_2$. We proved in Lemma 3.6, that $\left(d^2_{\beta_0}\mathcal{R}\right)(h,h)$ is given by the following formula:

$$\left(d^2_{\beta_0}\mathcal{R}\right)(h,h) = \eta^2 \mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2 \alpha\left(Z^t\beta_R\right)\right] + \left(1-\eta^2\right)\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\right].$$

We know from Lemma 3.9 that $\alpha$ is non-positive on the interval $[-x_1, x_1]$ and positive otherwise. Consequently, we study the sign of $\left(d^2_{\beta_0}\mathcal{R}\right)(h,h)$ on the partition $\mathbb{R} = (-\infty -, x_1) \bigcup [-x_1, x_1] \bigcup (x_1, \infty)$:

$$\begin{aligned}
\left(d^2_{\beta_0}\mathcal{R}\right)(h,h) &= \eta^2 \mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2 \alpha\left(Z^t\beta_0\right)\right] + \left(1-\eta^2\right)\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\right] \\
&= \underbrace{\eta^2 \mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2 \alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0 > x_1\}}\right] + \left(1-\eta^2\right)\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0 > x_1\}}\right]}_{A} \\
&\quad \underbrace{+ \eta^2 \mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2 \alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{-x_1 < Z^t\beta_0 < x_1\}}\right] + \left(1-\eta^2\right)\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{-x_1 < Z^t\beta_0 < x_1\}}\right]}_{-B} \\
&\quad + \eta^2 \mathbb{E}\left[\frac{1}{R^2}\left(Z^t\beta_0\right)^2 \alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0 < x_1\}}\right] + \left(1-\eta^2\right)\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0 < x_1\}}\right]
\end{aligned}$$

We have found in Lemma 3.7 two quantities $a > 0$ and $b > 0$ such that $A > a$ and $b \geq B$:

$$a := \left(\eta^2 \frac{x_1^2}{R^2} + 1 - \eta^2\right) \left[R + \left[R\left(\|a\|_2 - R\right) - \left(x_1 + \frac{8}{100}\right)\right] G\left(\frac{x_1}{R} + R - \|a\|_2\right)\right] \gamma\left(\|a\|_2 - \frac{x_1}{R}\right) e^{-x_1},$$
$$b := \frac{1}{4}\left(\eta^2 \frac{x_1^2}{R^2} + 1 - \eta^2\right) \left(\Phi^c\left(\|a\|_2 - \frac{x_1}{R}\right) - \Phi^c\left(\|a\|_2 + \frac{x_1}{R}\right)\right).$$

If $a > (1+\nu)\, b$ for some $\nu > 0$ then we have $\left(d^2_{\beta_0}\mathcal{R}\right)(h,h) > A - B > a - b > \nu b$. As

$$b < \frac{1}{4}\left(\eta^2 \frac{x_1^2}{R^2} + 1 - \eta^2\right) \Phi^c\left(\|a\|_2 - \frac{x_1}{R}\right),$$

the condition "$a > (1+\nu)\, b$" is satisfied when these successive conditions are true:

$$\left(\eta^2 \frac{x_1^2}{R^2} + 1 - \eta^2\right) \left[R + \left[R\left(\|a\|_2 - R\right) - \left(x_1 + \frac{8}{100}\right)\right] G\left(\frac{x_1}{R} + R - \|a\|_2\right)\right] \gamma\left(\|a\|_2 - \frac{x_1}{R}\right) e^{-x_1}$$
$$> (1+\nu)\, \frac{1}{4}\left(\eta^2 \frac{x_1^2}{R^2} + 1 - \eta^2\right) \Phi^c\left(\|a\|_2 - \frac{x_1}{R}\right)$$

(simplify $\left(\eta^2 \frac{x_1^2}{R^2} + 1 - \eta^2\right)$ and $R$ in factor in the left-hand side)

$$R\left[1 - \left(R - \|a\|_2 + \frac{x_1 + \frac{8}{100}}{R}\right) G\left(\frac{x_1}{R} + R - \|a\|_2\right)\right] \gamma\left(\|a\|_2 - \frac{x_1}{R}\right)$$
$$> (1+\nu)\, \frac{e^{x_1}}{4} \Phi^c\left(\|a\|_2 - \frac{x_1}{R}\right)$$

(divide by $\gamma\left(\|a\|_2 - \frac{x_1}{R}\right)$ and make Mill's ratio appear)

$$R\left(1 - \left(R - \|a\|_2 + \frac{x_1 + \frac{8}{100}}{R}\right) G\left(\frac{x_1}{R} + R - \|a\|_2\right)\right) \geq (1+\nu)\, \frac{e^{x_1}}{4} G\left(\|a\|_2 - \frac{x_1}{R}\right).$$

To conclude, when the latter inequality is true, one has $\left(d^2_{\beta_0}\mathcal{R}\right)(h,h) > \nu b$. □

**Lemma 3.9.** *Study of* $\alpha(x) = -\frac{e^x}{(1+e^x)^2}\left(1 + x\frac{1-e^x}{1+e^x}\right)$. *At* $x = 0$, $\alpha(0) = -\frac{1}{4}$ *is a global minimum,* $x_{\alpha_{max}} \in [2,3]$ *is the positive real where* $\alpha$ *is maximal with value* $\alpha_{max}$, *its derivative is bounded* $\|\alpha'\|_\infty \leq 0.22$ *and by definition of* $x_1$ *(see Section 3.2),* $\alpha(x_1) = 0$ *with*

$$x_1 \approx 1.54340463. \tag{3.5.16}$$

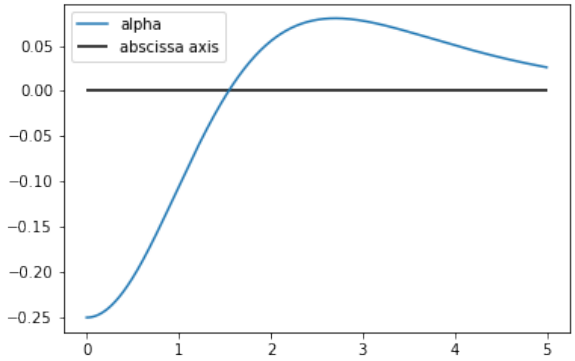
| x | 0 | | $x_2$ | | 1 | | $x_1$ | | 2 | | $x_{\alpha_{max}}$ | | $\infty$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| sign of $f''$ | $-$ | $-$ | $-$ | $-$ | | $-$ | $-$ | $-$ | $-$ | $-$ | $-$ | $-$ | $-$ |
| variations of $f'$ | 1 | $\searrow$ | 0 | $\searrow$ | | $\searrow$ | $\searrow$ | $\searrow$ | $\searrow$ | $\searrow$ | $\searrow$ | $\searrow$ | $-\infty$ |
| sign of $f'$ | $+$ | $+$ | 0 | $-$ | | $-$ | $-$ | $-$ | $-$ | $-$ | $-$ | $-$ | $-$ |
| variations of $f$ | 2 | $\nearrow$ | $f(x_2)$ | $\searrow$ | 2 | $\searrow$ | 0 | $\searrow$ | $3-e^2$ | $\searrow$ | $\searrow$ | $\searrow$ | $-\infty$ |
| sign of $f$ | $+$ | $+$ | $+$ | $+$ | $+$ | $+$ | 0 | $-$ | $-$ | $-$ | $-$ | $-$ | $-$ |
| sign of $\alpha$ | $-\frac{1}{4}$ | $-$ | $-$ | $-$ | $-$ | $-$ | 0 | $+$ | $+$ | $+$ | $\alpha_{max}$ | $+$ | 0 |

Table 3.5.1: sign and variation table of $f$, sign table of $\alpha$

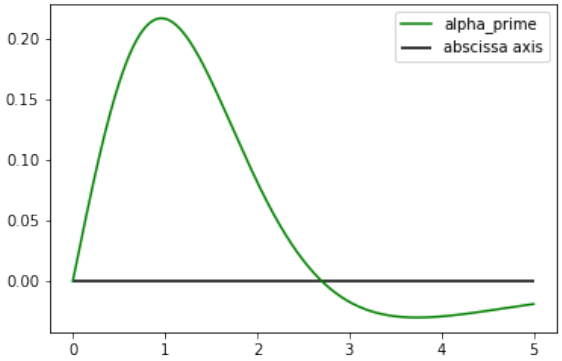
*Proof.* First, remark that $\forall x > 0$,

$$\begin{aligned}\alpha(x) \geq 0 &\Leftrightarrow 1 + x\frac{1-e^x}{1+e^x} \leq 0\\ &\Leftrightarrow x\left(1-e^x\right) \leq -\left(1+e^x\right)\\ &\Leftrightarrow 1 + e^x - xe^x + x \leq 0\\ &\Leftrightarrow f(x) \leq 0.\end{aligned}$$

We study $f : x \mapsto 1 + e^x - xe^x + x$ for $x \in \mathbb{R}_+$ since $\alpha$ is even. First of all, $f'(x) = 1 - xe^x$ and $f''(x) = -(x+1)e^x$ which gives the sign and variation table 3.5.1. It is obvious that there exists $x_1 > 0$ such that $f(x_1) = 0$. Set $p_x = (1-e^x)^{-1}$ and $q_x = e^x(1-e^x)^{-1}$. Note that $p_x - q_x = \frac{1-e^x}{1+e^x} = -\tanh\left(\frac{x}{2}\right)$ and $p_x q_x = \frac{1}{4}\left((p_x+q_x)^2 - (p_x-q_x)^2\right) = \frac{1}{4}\left(1-\tanh^2\left(\frac{x}{2}\right)\right)$.




((a)) Plot of $\alpha$ see online here




((b)) Plot of the derivative of $\alpha$ see online here

Figure 3.5.1: Plot about $\alpha$

$$
\begin{aligned}
\alpha'(x) &= \frac{d}{dx}\left[-p_x q_x \left(1 + x\left(p_x - q_x\right)\right)\right] \\
&= -\frac{d}{dx}\left[p_x\right] q_x \left(1 + x\left(p_x - q_x\right)\right) - p_x \frac{d}{dx}\left[q_x\right]\left(1 + x\left(p_x - q_x\right)\right) - p_x q_x \frac{d}{dx}\left[1 + x\left(p_x - q_x\right)\right] \\
&= p_x q_x q_x \left(1 + x\left(p_x - q_x\right)\right) - p_x p_x q_x \left(1 + x\left(p_x - q_x\right)\right) - p_x q_x \left[\left(p_x - q_x\right) + x\frac{d}{dx}\left[\left(p_x - q_x\right)\right]\right] \\
&= p_x q_x \left[q_x \left(1 + x\left(p_x - q_x\right)\right) - p_x \left(1 + x\left(p_x - q_x\right)\right) - \left(p_x - q_x\right) - x\left(-p_x q_x - p_x q_x\right)\right] \\
&= p_x q_x \left[-\left(p_x - q_x\right)\left(1 + x\left(p_x - q_x\right)\right) - \left(p_x - q_x\right) + 2x p_x q_x\right] \\
&= p_x q_x \left[2x p_x q_x - \left(p_x - q_x\right)\left(2 + x\left(p_x - q_x\right)\right)\right] \qquad (3.5.17) \\
&= p_x q_x \left[\frac{x}{2}\left(1 - \tanh^2\left(\frac{x}{2}\right)\right) + \tanh\left(\frac{x}{2}\right)\left(2 - x\tanh\left(\frac{x}{2}\right)\right)\right] \\
&= p_x q_x \left[\frac{x}{2}\left(1 - 3\tanh^2\left(\frac{x}{2}\right)\right) + 2\tanh\left(\frac{x}{2}\right)\right] \\
\alpha'(x) &= \frac{1}{4}\left(1 - \tanh^2\left(\frac{x}{2}\right)\right)\left[\frac{x}{2}\left(1 - 3\tanh^2\left(\frac{x}{2}\right)\right) + 2\tanh\left(\frac{x}{2}\right)\right].
\end{aligned}
$$

One can see on Figure 3.5.1(b) that the maximum of $\alpha$ is attained at $2 \leq x_{\alpha_{max}} \leq 3$. The function $\alpha$ is Lipschitz and one can see graphically on Figure 3.5.1(b) that $\|\alpha'\|_\infty \leq 0.22$. □

**Lemma 3.10.** $\alpha : x \mapsto -\frac{e^x}{(1+e^x)^2}\left(1 + x\frac{1-e^x}{1+e^x}\right)$ *is concave on* $[x_1, 3]$.

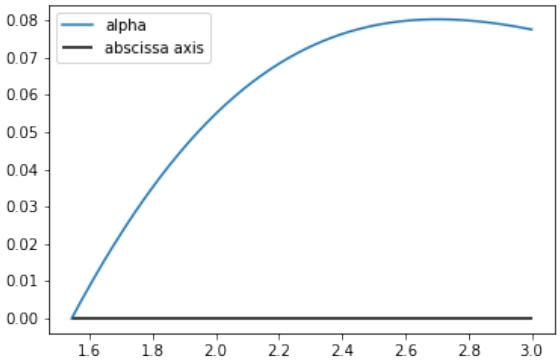



Figure 3.5.2: Plot of $\alpha$ on $[x_1, 3]$ see online here

*Proof.* The shape of $\alpha$ on $[x_1, 3]$ can be seen on figure 3.5.2.

We use the following compact notations: $p = \frac{1}{1+e^x}$, $q = 1 - p$, hence $\alpha(x) = -pq\,(1 + x\,(p - q))$. Recall that $\frac{dp}{dx} = -pq$, $\frac{dq}{dx} = pq$ and that $\forall x > 0, p < q$. We proved in Equation (3.5.17) that

$$\begin{aligned}\alpha'(x) &= pq\,[(q - p)\,(2 + x\,(p - q)) + 2xpq] \\ &= p(1 - p)\,[(1 - 2p)\,(2 + x\,(p - q)) + 2xp(1 - p)]\end{aligned}$$

In this proof we will also need the variations of $\varpi : x \mapsto 1 + x\,(p - q)$:

$$\begin{aligned}\varpi'(x) &= \frac{d}{dx}\,(1 + x\,(p - q)) \\ &= (p - q) + x\frac{d}{dx}\,(p - q) \\ &= (p - q) + x\,(-pq - pq) \\ &= (p - q) - 2xpq \\ &< 0\end{aligned}$$

We will want the sign of $\frac{d^2\alpha}{dx^2}$. First remark that

$$\begin{aligned}
\frac{d}{dx}\left[\frac{\alpha'}{pq}\right] &= \frac{d\alpha'}{dx}\frac{1}{pq} + \alpha'\frac{d}{dx}\left[\frac{1}{pq}\right] \\
&= \frac{d\alpha'}{dx}\frac{1}{pq} + \alpha'\frac{-1}{(pq)^2}\frac{d(pq)}{dx} \\
&= \alpha''\frac{1}{pq} + \alpha'\frac{-1}{(pq)^2}(-pqq + ppq) \\
\frac{d}{dx}\left[\frac{\alpha'}{pq}\right] &= \alpha''\frac{1}{pq} - \alpha'\frac{1}{pq}(p-q)
\end{aligned}$$

aalgebraic rearrangment give $\alpha'' = p(1-p)\frac{d}{dx}\left[\frac{\alpha'}{pq}\right] + \alpha'(2p-1)$. Now compute what is still missing:

$$\begin{aligned}
\frac{d}{dx}\left[\frac{\alpha'}{pq}\right] &= \frac{d}{dx}\left[\frac{pq\left[(q-p)(2+x(p-q))+2xpq\right]}{pq}\right] \\
&= \frac{d}{dx}\left[(q-p)(2+x(p-q))+2xpq\right] \\
&= \frac{d}{dx}(q-p)(2+x(p-q)) + (q-p)\frac{d}{dx}(2+x(p-q)) + 2\frac{d}{dx}(xpq) \\
&= (pq+pq)(2+x(p-q)) + (q-p)\left[(p-q)+x(-pq-pq)\right] + 2(pq - xpqq + xppq) \\
&= 2pq(2+x(p-q)) + (q-p)\left[(p-q)-2xpq\right] + 2pq(1+x(p-q)) \\
&= 2pq(3+2x(p-q)) + (q-p)(p-q) - 2xpq(q-p) \\
&= 2pq(3+2x(p-q)) - (q-p)^2 + 2xpq(p-q) \\
&= 2pq(3+2x(p-q)+x(p-q)) - (q-p)^2 \\
&= 6pq(1+x(p-q)) - (q-p)^2 \\
&= 6pq\varpi(x) - (p-q)^2 \\
\frac{d}{dx}\left[\frac{\alpha'}{pq}\right] &= 6p(1-p)\varpi(x) - (1-2p)^2
\end{aligned}$$

-Case $x \in [x_1, 2]$:

we have $p \in [0.11, 0.18]$, hence $1-2p \in [0.64, 0.78]$, $x \mapsto p$ is decreasing and $p \mapsto p(1-p)$ is increasing on this intervalles of interest then $p(1-p) \in [p_{x=2}(1-p_{x=2}), p_{x=x_1}(1-p_{x=x_1})] \subset [0.09, 0.15]$

then $\varpi$ is strictly decreasing and $\varpi(x) \in [\varpi(2), \varpi(x_1)] \subset [-0.53, 0]$ (0 occurs because by definition $x_1$ is such that $0 = \alpha(x_1) = pq\varpi(x_1)$), all intervalle put together gives in case $x \in [x_1, 2]$:

$$\begin{aligned}\alpha'(x) &= \underbrace{p(1-p)}_{\geq 0.09}\left[\underbrace{(1-2p)}_{\geq 0.64}\left(1+\underbrace{1+x\,(p-q)}_{\geq -0.53}\right)+2\,\underbrace{x_1}_{\geq 1.5435}\,\underbrace{p(1-p)}_{\geq 0.09}\right]\\ &\geq 0.052\end{aligned}$$

$$\begin{aligned}\alpha'(x) &= \underbrace{p(1-p)}_{\leq 0.15}\left[\underbrace{(1-2p)}_{\leq 0.78}\left(1+\underbrace{1+x\,(p-q)}_{\leq 0}\right)+2\,\underbrace{x_1}_{\leq 1.5436}\,\underbrace{p(1-p)}_{\leq 0.15}\right]\\ &\leq 0.19\end{aligned}$$

it holds

$$\begin{aligned}\frac{d}{dx}\left[\frac{\alpha'}{pq}\right] &= 2p(1-p)\underbrace{\varpi(x)}_{\leq 0} - \underbrace{\left(\underbrace{1-2p}_{\geq 0.64}\right)^2}_{\geq 0.4}\\ &\leq -0.4\end{aligned}$$

$$\begin{aligned}\frac{d}{dx}\left[\frac{\alpha'}{pq}\right] &= 2\underbrace{p(1-p)}_{\leq 0.15}\,\underbrace{\varpi(x)}_{\geq -0.53} - \underbrace{\left(\underbrace{1-2p}_{\leq 0.78}\right)^2}_{\leq 0.61}\\ &\geq -0.769\end{aligned}$$

And finally

$$\alpha''(x) = \underbrace{p(1-p)}_{\geq 0.09}\underbrace{\frac{d}{dx}\left[\frac{\alpha'}{pq}\right]}_{\leq -0.4} + \underbrace{\alpha'(x)}_{\geq 0.059}\underbrace{(1-2p)}_{\leq -0.64}$$

$$\leq -0.07376$$

$$\alpha''(x) = \underbrace{pq}_{\leq 0.15}\underbrace{\frac{d}{dx}\left[\frac{\alpha'}{pq}\right]}_{\geq -0.769} + \underbrace{\alpha'(x)(1-2p)}_{\leq 0.19 \ \geq -0.78}$$

$$\geq -0.26355$$

$\alpha$ is concave on $[x_1, 2]$.

-We do the same in the case $x \in [2, 2.5]$:

we have $p \in [0.075, 0.12]$, hence $1-2p \in [0.76, 0.85]$, $x \mapsto p$ is decreasing and $p \mapsto p(1-p)$ is increasing on this intervalles of interest then $p(1-p) \in [p_{x=2.5}(1-p_{x=2.5}), p_{x=2}(1-p_{x=2})] \subset [0.069, 0.11]$, $\varpi(x) \in [\varpi(3), \varpi(2)] \subset [-1.125, -0.52]$ and

$$\alpha'(x) = \underbrace{p(1-p)}_{\geq 0.069}\left[\underbrace{(1-2p)}_{\geq 0.76}\left(1 + \underbrace{1 + x\,(p-q)}_{\geq -1.125}\right) + 2\,\underbrace{x_1}_{\geq 1.5435}\,\underbrace{p(1-p)}_{\geq 0.069}\right]$$

$$\geq 0.008$$

$$\alpha'(x) = \underbrace{p(1-p)}_{\leq 0.12}\left[\underbrace{(1-2p)}_{\leq 0.85}\left(1 + \underbrace{1 + x\,(p-q)}_{\leq -0.52}\right) + 2\,\underbrace{x_1}_{\leq 1.5436}\,\underbrace{p(1-p)}_{\leq 0.12}\right]$$

$$\leq 0.094$$

We now have

$$\frac{d}{dx}\left[\frac{\alpha'}{pq}\right] = 2\underbrace{p(1-p)}_{\geq 0.069}\underbrace{\varpi(x)}_{\leq -0.52} - \underbrace{\left(\underbrace{1-2p}_{\geq 0.76}\right)^2}_{\geq 0.5776}$$

$$\leq -0.64936 \leq -0.65$$

$$\frac{d}{dx}\left[\frac{\alpha'}{pq}\right] = 2\underbrace{p(1-p)}_{\leq 0.11}\underbrace{\varpi(x)}_{\geq -0.53} - \underbrace{\left(\underbrace{1-2p}_{\leq 0.85}\right)^2}_{\leq 0.7225}$$

$$\geq -0.8391 \geq -0.84$$

And finally

$$\alpha''(x) = \underbrace{p(1-p)}_{\geq 0.069}\underbrace{\frac{d}{dx}\left[\frac{\alpha'}{pq}\right]}_{\leq -0.65} + \underbrace{\alpha'(x)}_{\geq 0.008}\underbrace{(1-2p)}_{\leq -0.76}$$

$$\leq -0.05093$$

$$\alpha''(x) = \underbrace{pq}_{\leq 0.11}\underbrace{\frac{d}{dx}\left[\frac{\alpha'}{pq}\right]}_{\geq -0.84} + \underbrace{\alpha'(x)}_{\leq 0.094}\underbrace{(1-2p)}_{\geq -0.85}$$

$$\geq -0.1723$$

Consequently $\alpha$ is concave on $[2, 2.5]$

-We do the same in the case $x \in [2.5, 3]$:

in that case $p \in [0.047, 0.076]$, hence $1 - 2p \in [0.848, 0.906]$, $x \mapsto p$ is decreasing and $p \mapsto p(1-p)$ is increasing on this intervalles of interest then $p(1-p) \in [p_{x=3}(1-p_{x=3}), p_{x=2.5}(1-p_{x=2.5})] \subset [0.0447, 0.071]$ and $\varpi(x) \in [\varpi(3), \varpi(2)] \subset [-1.72, -1.12]$.

$$\begin{aligned}\alpha'(x) &= \underbrace{p(1-p)}_{\geq 0.0447}\left[\underbrace{(1-2p)}_{\geq 0.848}\left(1+\underbrace{1+x\,(p-q)}_{\geq -1.72}\right)+2\underbrace{x_1}_{\geq 1.5435}\underbrace{p(1-p)}_{\geq 0.0447}\right]\\ &\geq -0.02113\end{aligned}$$

$$\begin{aligned}\alpha'(x) &= \underbrace{p(1-p)}_{\leq 0.071}\left[\underbrace{(1-2p)}_{\leq 0.906}\left(1+\underbrace{1+x\,(p-q)}_{\leq -1.12}\right)+2\underbrace{x_1}_{\leq 1.5436}\underbrace{p(1-p)}_{\leq 0.071}\right]\\ &\leq 0.0079\end{aligned}$$

We now have

$$\begin{aligned}\frac{d}{dx}\left[\frac{\alpha'}{pq}\right] &= 2\underbrace{p(1-p)}_{\geq 0.0447}\underbrace{\varpi(x)}_{\leq -1.12}-\underbrace{\left(\underbrace{1-2p}_{\geq 0.848}\right)^2}_{\geq 0.5776}\\ &\leq -0.83\end{aligned}$$

$$\begin{aligned}\frac{d}{dx}\left[\frac{\alpha'}{pq}\right] &= 2\underbrace{p(1-p)}_{\leq 0.076}\underbrace{\varpi(x)}_{\geq -1{,}72}-\underbrace{\left(\underbrace{1-2p}_{\leq 0.906}\right)^2}_{\leq 0.7225}\\ &\geq -1,082\end{aligned}$$

Concerning $\alpha''$, since $\alpha' \in [-0.02113, 0.0079]$, the reasonning with an intervalle containing 0 is a bit different:

$p(1-p)\frac{d}{dx}\left[\frac{\alpha'}{pq}\right] \in [-0.077, -0.039]$ and $\alpha'(x)(1-2p) \in [-0.0072, 0.0191]$, consequently

$$\begin{aligned}\alpha''(x) &= p(1-p)\frac{d}{dx}\left[\frac{\alpha'}{pq}\right] + \alpha'(x)(1-2p)\\ &\in [-0.082, -0.0199]\end{aligned}$$

Consequently $\alpha$ is concave on $[2.5, 3]$ □

**Lemma 3.11.** *$\forall x \geq 3$, $\alpha(x) - \varphi(x) \geq xe^{-x}\left(\frac{x_1+0.08-1}{x} - 4e^{-x}\right)$ where $\alpha : x \mapsto -\frac{e^x}{(1+e^x)^2}\left(1 + x\frac{1-e^x}{1+e^x}\right)$ and $\varphi : x \mapsto (x - x_1 - 0.08)e^{-x}$.*

*Proof.* Let us study $\alpha - \varphi$

$$\begin{aligned}\alpha(x) - \varphi(x) &= -\frac{e^x}{(1+e^x)^2}\left(1 + x\frac{1-e^x}{1+e^x}\right) - (x - x_1 - 0.08)e^{-x}\\ &= -\frac{e^{-2x}}{(1+e^{-x})^3}(1 + e^x + x - xe^x) - (x - x_1 - 0.08)e^{-x}\\ &= \frac{xe^{-2x}}{(1+e^{-x})^3}\left(e^x - \frac{e^x}{x} - 1 - \frac{1}{x}\right) - x\left(1 - \frac{x_1+0.08}{x}\right)e^{-x}\\ &= xe^{-x}\left[\frac{e^{-x}}{(1+e^{-x})^3}\left(e^x - \frac{e^x}{x} - 1 - \frac{1}{x}\right) - \left(1 - \frac{x_1+0.08}{x}\right)\right]\\ &= xe^{-x}\left[\frac{1}{(1+e^{-x})^3}\left(1 - \frac{1}{x} - e^{-x} - \frac{e^{-x}}{x}\right) - \left(1 - \frac{x_1+0.08}{x}\right)\right]\end{aligned}$$

Set $R(x) := \frac{1}{(1+e^{-x})^3} - 1 + 3e^{-x}$ and $\delta = x_1 + 0.08$. We get

$$\begin{aligned}\alpha(x) - \varphi(x) &= xe^{-x}\left[\left(1 - 3e^{-x} + R(x)\right)\left(1 - \frac{1}{x} - e^{-x} - \frac{e^{-x}}{x}\right) - \left(1 - \frac{\delta}{x}\right)\right]\\ &= xe^{-x}\left[(-1+\delta)\frac{1}{x} + (-1-3)e^{-x} + (-1+3)\frac{e^{-x}}{x} + 3e^{-2x} + 3\frac{e^{-2x}}{x} + R(x)\left(1 - \frac{1}{x} - e^{-x} - \frac{e^{-x}}{x}\right)\right]\\ &= xe^{-x}\left[\frac{\delta-1}{x} - 4e^{-x} + \underbrace{2\frac{e^{-x}}{x} + 3e^{-2x} + 3\frac{e^{-2x}}{x}}_{\geq 0} + R(x)\left(1 - \frac{1}{x} - e^{-x} - \frac{e^{-x}}{x}\right)\right]\\ &\geq xe^{-x}\left[\frac{\delta-1}{x} - 4e^{-x} + R(x)\left(1 - \frac{1}{x} - e^{-x} - \frac{e^{-x}}{x}\right)\right].\end{aligned}$$

Let us now discuss the sign of $R(x)\left(1 - \frac{1}{x} - e^{-x} - \frac{e^{-x}}{x}\right)$:

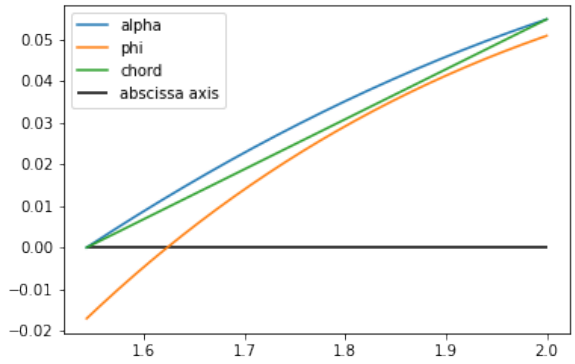



Figure 3.5.3: Plot of $\alpha$, $\varphi$ and its the chord on $[x_1, 2]$ and see online here

$R'(x) = 3e^{-x}\left(1+e^{-x}\right)^{-4} - 3e^{-x} = 3e^{-x}\left(\left(1+e^{-x}\right)^{-4} - 1\right) < 0$, $R$ is strictly decreasing. Since $\lim_{x\to+\infty} R(x) = 0$, necessarily $R \geq 0$. In addition, since $x \geq 3$,

$$1 - \frac{1}{x} - e^{-x} - \frac{e^{-x}}{x} \geq 1 - \frac{1}{3} - e^{-x} - \frac{e^{-x}}{3} = \frac{2}{3}\left(1 - 2e^{-x}\right) \geq 0.$$

Hence $R(x)\left(1 - \frac{1}{x} - e^{-x} - \frac{e^{-x}}{x}\right) \geq 0$ for $x \geq 3$ and it comes

$$\forall x \geq 3, \alpha(x) - \varphi(x) \geq xe^{-x}\left(\frac{\delta - 1}{x} - 4e^{-x}\right).$$

□

**Lemma 3.12.** *The function* $\alpha : x \mapsto -\frac{e^x}{(1+e^x)^2}\left(1 + x\frac{1-e^x}{1+e^x}\right)$ *is greater than* $\varphi : x \mapsto (x - x_1 - 0.08)\, e^{-x}$ *on* $[x_1, \infty[$.

*Proof.* Let us prove that $\alpha \geq \varphi$ by considering four intervals $[x_1, 2]$, $[2, 2.5]$, $[2.5, 3]$ and $[3, \infty]$. We know that $\alpha$ is concave on $[x_1, 3]$ according to lemma 3.10. It is also the case of $\varphi$ because $\varphi''(x) = (x - x_1 - 0.08 - 2)\, e^{-x}$ which is negative on $[x_1, 3]$ since $x_1 + 0.08 + 2 \approx 3.62$. Hence, $\alpha$ is above its geometrical chords and $\varphi$ below its tangents on $[x_1, 3]$.

Case 1 on $[x_1, 2]$:

The function $\alpha$ is above $l_1 : x \mapsto \frac{\alpha(2)}{2-x_1}(x - x_1)$ and $\varphi$ is below $l_2 : x \mapsto \varphi(1.85) + \varphi'(1.85)(x - 1.85)$. And as shown on figure 3.5.3, $l_1 \geq l_2$ and one can check it with $l_1(x_1) = 0 \geq -0.00165 \approx l_2(2)$ and $l_1(2) \approx 0.0550 \geq 0.0539 \approx l_2(2.5)$,

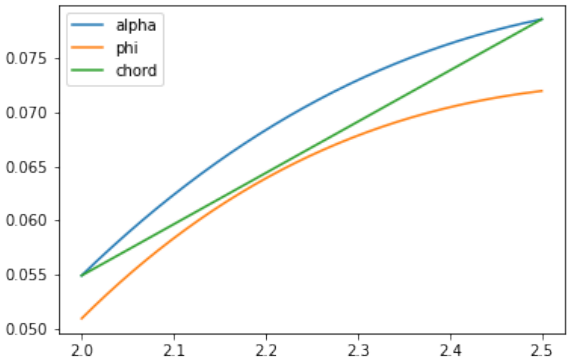



Figure 3.5.4: Plot of $\alpha$ and the chord of $\varphi$ on $[2, 2.5]$ and see online here

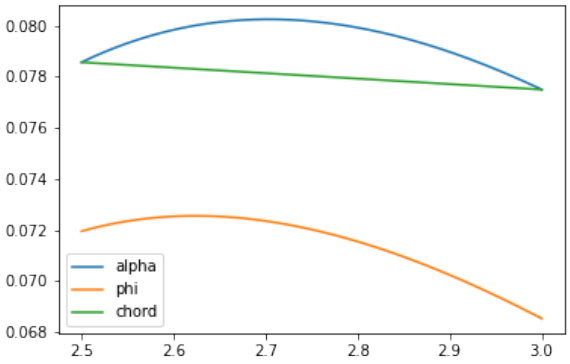



Figure 3.5.5: Plot of $\alpha$ and the chord of $\varphi$ on $[2.5, 3]$ and see online here

we can conclude that $\alpha \geq l_1 \geq l_2 \geq \varphi$ on $[x_1, 2]$.

Case 2 on $[2, 2.5]$:

The function $\alpha$ is above $l_1 : x \mapsto \alpha(2) + \frac{\alpha(2.5)-\alpha(2)}{2.5-2}(x-2)$ and $\varphi$ is below $l_2 : x \mapsto \varphi(2.2) + \varphi'(2.2)(x-2.2)$. $l_1 \geq l_2$ as well, one can check it with $l_1(2) \approx 0.05493 \geq 0.05450 \approx l_2(2)$ and $l_1(2.5) \approx 0.0785 \geq 0.0779 \approx l_2(2.5)$. Consequently on $[2, 2.5]$, $\alpha \geq l_1 \geq l_2 \geq \varphi$.

Case 3 on $[2.5, 3]$:

The function $\alpha$ is above $l_1 : x \mapsto \alpha(2.5) + \frac{\alpha(3)-\alpha(2.5)}{3-2.5}(x-2.5)$ and $\varphi$ is maximal at $x_1 + 1 + 0.08$ with approximate value 0.07256. And since $l_1(2.5) \approx 0.07856 \geq 0.07256$ and $l_1(3) \approx 0.07750 \geq 0.07256$, we can conclude that $l_1$ is

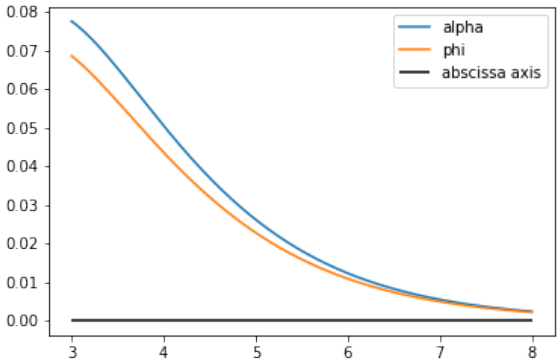



Figure 3.5.6: Plot of $\alpha$ and $\varphi$ on $[3,8]$ and see online here

above the maximum of $\varphi$. Hence, on $[2.5,3]$, $\alpha \geq l_1 \geq \varphi$.

Case 4 on $x \in [3,\infty]$:

Thanks to Lemma 3.11, we know that $\forall x \geq 3$, $\alpha(x) - \varphi(x) \geq e^{-x}\left(x_1 + 0.08 - 1 - 4xe^{-x}\right)$. Let us study the sign of $f : x \mapsto x_1 + 0.08 - 1 - 4xe^{-x}$. For any $x \geq 3$,

$$f'(x) = -4e^{-x} + 4xe^{-x} = 4\left(x-1\right)e^{-x} \geq 0.$$

We also have $f(3) \approx 0.026 > 0$. Consequently, $\forall x \geq 3, f(x) \geq 0$ and $\alpha(x) - \varphi(x) \geq 0$ as well.

This completes the proof: $\forall x \geq x_1, \alpha(x) - \varphi(x) \geq 0$. □

**Lemma 3.13.** *Recall that* $\alpha = -\frac{e^x}{(1+e^x)^2}\left(1 + x\frac{1-e^x}{1+e^x}\right)$, $Z \sim \mathcal{N}\left(a, I_d\right)$ *and* $\beta_0 = Ra/\left\|a\right\|_2$, $a \in \mathbb{R}^d$. *We have*

$$\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0 > x_1\}}\right] \geq \int_{x_1}^{\infty}\left(\left(x - x_1 - \frac{8}{100}\right)e^{-x}\frac{e^{-\left(x/R - \left\|a\right\|_2\right)^2/2}}{\sqrt{2\pi R^2}}\right)dx. \tag{3.5.18}$$

*Proof.* Recall that $\alpha(x) \geq \left(x - x_1 - \frac{8}{100}\right)e^{-x}$ on $[x_1,\infty[$ according to Lemma 3.12. Moreover $Z^t\beta_0 \sim \mathcal{N}\left(a^t\beta_0, \left\|\beta\right\|_2\right)$ with $a^t\beta_0 = R\left\|a\right\|_2$, $\left\|\beta\right\|_2 = R$. This gives

$$\begin{aligned}\mathbb{E}\left[\alpha\left(Z^t\beta_0\right)\mathbb{I}_{\{Z^t\beta_0>x_1\}}\right] &= \int_{x_1}^{\infty}\alpha(x)\frac{1}{\sqrt{2\pi R^2}}\exp\left(-\frac{(x-R\|a\|)^2}{2R^2}\right)dx \\ &\geq \int_{x_1}^{\infty}\left(x-x_1-\frac{8}{100}\right)e^{-x}\frac{1}{\sqrt{2\pi R^2}}\exp\left(-\frac{(x/R-\|a\|)^2}{2}\right)dx.\end{aligned}$$

□

**Lemma 3.14.** *For any $a, z \in \mathbb{R}^d$, $\xi \in \mathbb{R}$ and $R > 0$, it holds*

$$J_{a,R}(\xi,z) := \int_z^{\infty}\left(xe^{-\xi x}\frac{1}{\sqrt{2\pi R^2}}e^{-(x/R-\|a\|_2)^2/2}\right)dx = R\left(1+\left(\|a\|_2-R\xi\right)G\left(\frac{z}{R}+R\xi-\|a\|_2\right)\right)\gamma\left(\frac{z}{R}-\|a\|_2\right)e^{-\xi z},$$

*where $\gamma : x \to (2\pi)^{-1/2}e^{-\frac{1}{2}x^2}$ is the standard Gaussian density, $\Phi^c$ is the standard Gaussian tail function and $G : x \mapsto \Phi^c(x)/\gamma(x)$ is the Gaussian Mill's ratio.*

*Proof.* We have

$$\begin{aligned}J_{a,R}(\xi,z) &= \int_z^{\infty}xe^{-\xi x}\frac{1}{\sqrt{2\pi R^2}}\exp\left(-\frac{(x-R\|a\|)^2}{2R^2}\right)dx \\ &= \int_z^{\infty}\frac{x}{\sqrt{2\pi R^2}}\exp\left(-\frac{(x-R\|a\|)^2+2R^2\xi x}{2R^2}\right)dx. \qquad (3.5.19)\end{aligned}$$

Moreover,

$$(x-R\|a\|)^2+2R^2\xi x = \left(x+R^2\xi-R\|a\|_2\right)^2+R^2\xi\left(2R\|a\|_2-R^2\xi\right).$$

Hence,

$$\begin{aligned}J_{a,R}(\xi,z) &= \int_z^{\infty}\frac{x}{\sqrt{2\pi R^2}}\exp\left(-\frac{\left(x+R\left(R\xi-\|a\|_2\right)\right)^2+R^2\left(2\|a\|_2-R\xi\right)R\xi}{2R^2}\right)dx \\ &= e^{-\frac{(2\|a\|_2-R\xi)R\xi}{2}}\int_z^{\infty}x\frac{1}{\sqrt{2\pi R^2}}\exp\left(-\frac{\left(x/R+R\xi-\|a\|_2\right)^2}{2}\right)dx \qquad (3.5.20)\end{aligned}$$

By the change the variable $y = x/R + R\xi - \|a\|_2$, we get

$$
\begin{aligned}
J_{a,R}(\xi,z) &= e^{-\frac{(2\|a\|_2-R\xi)R\xi}{2}} \int_{z/R+R\xi-\|a\|_2}^{\infty} R\left(y-(R\xi-\|a\|_2)\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)dy \\
&= e^{-\frac{(2\|a\|_2-R\xi)R\xi}{2}} \int_{z/R+R\xi-\|a\|_2}^{\infty} Ry\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)dy \\
&\quad - e^{-\frac{(2\|a\|_2-R\xi)R\xi}{2}} \int_{z/R+R\xi-\|a\|_2}^{\infty} R\left(R\xi-\|a\|_2\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)dy \\
&= -Re^{-\frac{(2\|a\|_2-R\xi)R\xi}{2}}\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)\right]_{z/R+R\xi-\|a\|_2}^{\infty} \\
&\quad - R\left(R\xi-\|a\|_2\right)e^{-\frac{(2\|a\|_2-R\xi)R\xi}{2}}\Phi^c\left(\frac{z}{R}+R\xi-\|a\|_2\right) \\
&= Re^{-\frac{(2\|a\|_2-R\xi)R\xi}{2}}\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{z}{R}+R\xi-\|a\|_2\right)^2}+\left(\|a\|_2-R\xi\right)\Phi^c\left(\frac{z}{R}+R\xi-\|a\|_2\right)\right) \\
&= R\left[1+\left(\|a\|_2-R\xi\right)G\left(\frac{z}{R}+R\xi-\|a\|_2\right)\right]\frac{1}{\sqrt{2\pi}}e^{-\frac{R\xi(2\|a\|_2-R\xi)+\left(\frac{z}{R}+R\xi-\|a\|_2\right)^2}{2}}
\end{aligned}
$$

and since

$$
\left(2\|a\|_2-R\xi\right)R\xi+\left(z/R+R\xi-\|a\|_2\right)^2=\left(z/R-\|a\|_2\right)^2+2z\xi, \tag{3.5.21}
$$

we finally get the result. □

**Lemma 3.15.** *For any $a,z\in\mathbb{R}^d$, $\xi\in\mathbb{R}$ and $R>0$, it holds*

$$
K_{a,R}(\xi,z):=\int_z^{\infty}e^{-\xi x}\frac{1}{\sqrt{2\pi R^2}}e^{-\left(x/R-\|a\|_2\right)^2/2}dx=\gamma\left(\frac{z}{R}-\|a\|_2\right)G\left(\frac{z}{R}+R\xi-\|a\|_2\right)e^{-\xi z}
$$

*where $\gamma$ is the standard Gaussian density, $\Phi^c$ is the standard Gaussian tail function and $G:x\mapsto\Phi^c(x)/\gamma(x)$ is the Gaussian Mill's ratio.*

*Proof.* By the same calculation as in Equation 3.5.19, we can write

$$
K_{a,R}(\xi,z)=e^{-\frac{(2\|a\|_2-R\xi)R\xi}{2}}\int_z^{\infty}\frac{1}{\sqrt{2\pi R^2}}\exp\left(-\frac{\left(x/R+R\xi-\|a\|_2\right)^2}{2}\right)dx.
$$

By the change the variable $y = x/R + R\xi - \|a\|_2$, we get

$$\begin{aligned} K_{a,R}(\xi, z) &= e^{-\frac{(2\|a\|_2 - R\xi)R\xi}{2}} \int_{z/R+R\xi-\|a\|_2}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy \\ &= e^{-\frac{(2\|a\|_2 - R\xi)R\xi}{2}} \Phi^c\left(z/R + R\xi - \|a\|_2\right) \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(2\|a\|_2 - R\xi)R\xi + (z/R+R\xi-\|a\|_2)^2}{2}} .G\left(z/R + R\xi - \|a\|_2\right). \end{aligned}$$

By Identity (3.5.21), it follows that

$$K_{a,R}(\xi, z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(z/R - \|a\|_2\right)^2 - z\xi} G\left(z/R + R\xi - \|a\|_2\right),$$

as expected. □

**Lemma 3.16.** *Set $G(x) = \frac{\Phi^c(x)}{\gamma(x)}$ the Mill's ratio of the standard gaussian distribution. $G$ satisfies: $\forall x \in \mathbb{R}$*

$$\begin{aligned} xG(x) - G'(x) &= 1 \\ G''(x) - xG'(x) - G(x) &= 0 \\ G'''(x) - 2G'(x) - xG''(x) &= 0 \end{aligned}$$

*Proof.* $G(x) = \frac{\Phi^c(x)}{\gamma(x)}$ and using the fact that $\frac{d\Phi^c}{dx}(x) = -\gamma(x)$ and $\gamma'(x) = -x\gamma(x)$ it comes:

$$\begin{aligned} G'(x) &= \frac{-\gamma(x)\gamma(x) - \Phi^c(x)(-x\gamma(x))}{\gamma(x)\gamma(x)} \\ &= -1 + x\frac{\Phi^c(x)}{\gamma(x)} \\ &= -1 + xG(x) \end{aligned}$$

and $G' = xG - 1 \Rightarrow G'' = G + xG' \Rightarrow G''' = G' + G' + xG''$ □

**Proposition 3.3.** *The function $G(x) = \frac{\Phi^c(x)}{\gamma(x)}$ is known as the Gaussian Mill's ratio and $\forall x \geq 0$*

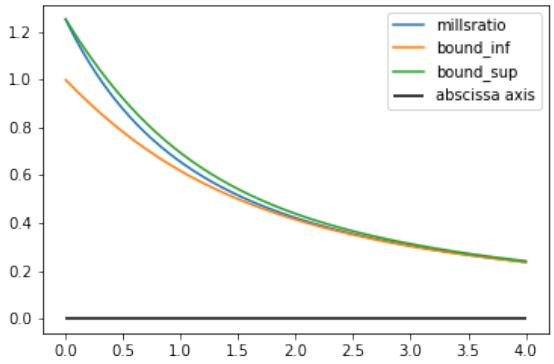



Figure 3.5.7: Plot of the function $G$ on $[0,4]$ and see online here

$$0 < \frac{2}{x+\sqrt{x^2+4}} \leq G(x) \leq \frac{2}{x+\sqrt{x^2+4.\frac{2}{\pi}}}$$

*Proof.* Focus on the first inequality: the lower bound is due to [16] and the upper bound is attributed to Pollak [121] according to [55] in which one can find the inequality in the first commentar of Remark 11 p 1848. □

**Proposition 3.4.** *The Gaussian mill's ratio function* $G : x \mapsto \frac{\Phi^c(x)}{\gamma(x)}$ *is a strictly decreasing function on* $\mathbb{R}$.

*Proof.* We have seen in Lemma 3.16 that $G' = -1 + xG$. Since $G \geq 0$, it is obvious that $G' < 0$ on $]-\infty; 0]$. Furthermore, take $x > 0$, then with proposition 3.3 we have

$$\begin{aligned} G'(x) &= -1 + xG(x) \\ &\leq -1 + x\frac{2}{x+\sqrt{x^2+4.\frac{2}{\pi}}} \\ &= \frac{2}{1+\sqrt{1+\frac{8}{\pi x^2}}} - 1 \end{aligned}$$

One can see that $\forall x > 0, \frac{2}{1+\sqrt{1+\frac{8}{\pi x^2}}} < 1$, hence $G' < 0$ everywhere on $]0, \infty[$. □

**Lemma 3.17.** *Define* $Eq_{a,b,c,d} : 1+aG(-b-a) \geq cG(a+d)$ *where* $a,b,c,d \geq 0$ *and* $G : x \mapsto \frac{\Phi^c(x)}{\gamma(x)}$ *is the Gaussian mill's ratio where* $\gamma$ *and* $\Phi^c$ *are respectively the density and the tail function of the standard univariate gaussian. If* $Eq_{a,b,c,d}$ *holds true, then* $\forall h > 0, Eq_{a+h,b,c,d}$ *holds true.*

*Proof.* Start with $(a,b,c,d)$ such that $Eq_{a,b,c,d}$ holds true and take $h > 0$. We proved in prop 3.4 that $G$ is a decreasing function,then $G(-b-a) < G(-b-(a+h))$, then one has $0 \leq aG(-b-a) < (a+h)G(-b-(a+h))$, hence

$$1 + (a+h)G(-b-(a+h)) > 1 + aG(-b-a)$$

But $(a,b,c,d)$ such that $Eq_{a,b,c,d}$ holds true, therefore

$$1 + (a+h)G(-b-(a+h)) > cG(a+d)$$

Finally, use again the fact that $G$ is decreasing, to have $G(a+d) > G((a+h)+d)$ and it comes

$$1 + (a+h)G(-b-(a+h)) > cG((a+h)+d)$$

To conclude, $Eq_{a+h,b,c,d}$ holds also true. □

**Lemma 3.18.** *The equation* $R\left(1-\left(R-\|a\|_2+\frac{x_1+\frac{8}{100}}{R}\right)G\left(\frac{x_1}{R}+R-\|a\|_2\right)\right) \geq (1+\nu)\frac{e^{x_1}}{4}G\left(\|a\|_2-\frac{x_1}{R}\right)$ *holds true, in particular, for* $R=\sqrt{x_1+0.08} \approx 1.2741$, $\|a\|_2 = 2R = R+\frac{x_1+0.08}{R} \approx 2.548$ *and* $\nu = 0.95$.

*Proof.* Replace the corresponding quantities to get as left side $R\left(1-\left(R-\|a\|_2+\frac{x_1+\frac{8}{100}}{R}\right)G\left(\frac{x_1}{R}+R-\|a\|_2\right)\right) = R = \sqrt{x_1+0.08}$ and as right side $(1+\nu)\frac{e^{x_1}}{4}G\left(R+\frac{x_1+\frac{8}{100}}{R}-\frac{x_1}{R}\right) = (1+\nu)\frac{e^{x_1}}{4}G\left(\sqrt{x_1+0.08}+\frac{0.08}{\sqrt{x_1+0.08}}\right)$. Approximation show that $\sqrt{x_1+0.08} \approx 1.2741$, $\frac{e^{x_1}}{4} \approx 1.1701$, $\sqrt{x_1+0.08}+\frac{0.08}{\sqrt{x_1+0.08}} \approx 1.33700$ and $G(1.337) \approx 0.5552$. On can see it is then enough to takes $\nu = 0.95$ because $(1+\nu)\frac{e^{x_1}}{4}G\left(\sqrt{x_1+0.08}+\frac{0.08}{\sqrt{x_1+0.08}}\right) \approx 1.2668$ (the inequality holds true because $1.2741 \geq 1.2668$). □

**Lemma 3.19.** *Recall that* $(d_\beta \mathcal{R})(\nu) = -\mathbb{E}\left[X^t \beta p_\beta(X) q_\beta(X) X^t \nu\right]$ *for* $\beta \in B_2(0,R)$ *and* $\nu \in \mathbb{R}^d$. *Assume that*

$a^t\beta - 2\|\beta\|_2^2 \geq 0$. *If* $\langle \nu, \beta \rangle \leq 0$, *it holds*

$$(d_\beta \mathcal{R})(\nu) \geq \frac{1}{8} e^{-\left(2a^t\beta - \|\beta\|_2^2\right)/2} \left\langle -\nu, \frac{\beta}{\|\beta\|_2^2} \right\rangle \left( \|\beta\|_2^2 + \left( a^t\beta - \|\beta\|_2^2 \right)^2 \right).$$

*Proof.* For $\nu \in \mathbb{R}^d$, decompose it on $\beta^\perp \oplus Vect(\beta)$ as $\nu = \nu_\perp + \nu_\parallel$, and set $\lambda_\nu := \left\langle \nu, \frac{\beta}{\|\beta\|_2} \right\rangle$ so that $\nu_\parallel = \lambda_\nu \frac{\beta}{\|\beta\|_2}$ . Recall $X \sim \varepsilon_{Rad} Z$ where $\varepsilon_{Rad}$ is a Rademacher random variable with distribution $\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$and $Z \sim \mathcal{N}(a, I_d)$. As a consequence $Z^t\beta \sim \mathcal{N}\left(a^t\beta, \|\beta\|_2^2\right)$. Set also $N \sim \mathcal{N}(0,1)$, so that $Z^t\beta = a^t\beta + \|\beta\|_2 N$. We have, by symmetry in $X$ and independence between $Z^t\beta$ and $Z^t\nu_\perp$,

$$\begin{aligned}
(d_\beta \mathcal{R})(\nu) &= -\mathbb{E}\left[ X^t\beta p_\beta(X) q_\beta(X) X^t\nu \right] \\
&= -\mathbb{E}\left[ \frac{Z^t\beta e^{-Z^t\beta}}{\left(1+e^{-Z^t\beta}\right)^2} Z^t \left(\nu_\perp + \nu_\parallel\right) \right] \\
&= -\mathbb{E}\left[ \frac{Z^t\beta e^{-Z^t\beta}}{\left(1+e^{-Z^t\beta}\right)^2} Z^t\nu_\perp \right] - \mathbb{E}\left[ \frac{Z^t\beta e^{-Z^t\beta}}{\left(1+e^{-Z^t\beta}\right)^2} Z^t\nu_\parallel \right] \\
&= -\mathbb{E}\left[ \frac{Z^t\beta e^{-X^t\beta}}{\left(1+e^{-X^t\beta}\right)^2} \right] \underbrace{\mathbb{E}\left[ Z^t\nu_\perp \right]}_{=0} - \frac{\lambda_\nu}{\|\beta\|_2} \mathbb{E}\left[ \frac{Z^t\beta e^{-Z^t\beta}}{\left(1+e^{-Z^t\beta}\right)^2} Z^t\beta \right] \\
&= -\frac{\lambda_\nu}{\|\beta\|_2} \mathbb{E}\left[ \zeta\left( a^t\beta + \|\beta\|_2 N \right) \right],
\end{aligned}$$

where $\zeta : x \mapsto \frac{x^2 e^x}{(1+e^x)^2}$. Note that the function $\zeta$ is even and that a simple calculation gives $\forall x > 0, \zeta(x) \geq \frac{x^2}{4} e^{-x}$. If $\lambda_\nu \leq 0$,

$$-\frac{\lambda_\nu}{\|\beta\|_2} \mathbb{E}\left[ \zeta\left( a^t\beta + \|\beta\|_2 N \right) \right] \geq -\frac{\lambda_\nu}{\|\beta\|_2} \int_0^{+\infty} \frac{x^2}{4} e^{-x} \frac{1}{\sqrt{2\pi\|\beta\|_2^2}} \exp\left( -\frac{\left(x - a^t\beta\right)^2}{2\|\beta\|_2^2} \right) dx.$$

Set $N_{a,\beta} \sim \mathcal{N}\left( a^t\beta - 2\|\beta\|_2^2, \|\beta\|_2^2 \right)$. This gives

$$
\begin{aligned}
\mathbb{E}\left[\zeta\left(a^t\beta+\|\beta\|_2^2 N\right)\right] &\geq \int_0^{+\infty} \frac{x^2}{4} e^{-x} \frac{1}{\sqrt{2\pi\|\beta\|_2^2}} \exp\left(-\frac{\left(x-a^t\beta\right)^2}{2\|\beta\|_2^2}\right) dx \\
&= \int_0^{+\infty} \frac{x^2}{4} \frac{1}{\sqrt{2\pi\|\beta\|_2^2}} \exp\left(-\frac{\left(x-\left(a^t\beta-\|\beta\|_2^2\right)\right)^2+\|\beta\|_2^2\left(2a^t\beta-\|\beta\|_2^2\right)}{2\|\beta\|_2^2}\right) dx \\
&= e^{-\left(2a^t\beta-\|\beta\|_2^2\right)/2} \int_0^{+\infty} \frac{x^2}{4} \frac{1}{\sqrt{2\pi\|\beta\|_2^2}} \exp\left(-\frac{\left(x-\left(a^t\beta-2\|\beta\|_2^2\right)\right)^2}{2\|\beta\|_2^2}\right) dx \\
&\geq \frac{1}{8} e^{-\left(2a^t\beta-\|\beta\|_2^2\right)/2} \mathbb{E}\left[N_{a,\beta}^2\right] \\
&= \frac{1}{8} e^{-\left(2a^t\beta-\|\beta\|_2^2\right)/2} \left(\mathbb{V}\left[N_{a,\beta}^2\right]+\mathbb{E}\left[N_{a,\beta}\right]^2\right) \\
&= \frac{1}{8} e^{-\left(2a^t\beta-\|\beta\|_2^2\right)/2} \left(\|\beta\|_2^2+\left(a^t\beta-2\|\beta\|_2^2\right)^2\right),
\end{aligned}
$$

where in the second inequality, we used the fact that $a^t\beta - 2\|\beta\|_2^2 \geq 0$. Therefore

$$
\left(d_\beta \mathcal{R}\right)(\nu) \geq -\frac{\lambda_\nu}{8\|\beta\|_2} e^{-\left(2a^t\beta-\|\beta\|_2^2\right)/2} \left(\|\beta\|_2^2+\left(a^t\beta-\|\beta\|_2^2\right)^2\right).
$$

□

**Definition 3.1.** The operator norm $\|\cdot\|_{op}$ on the trilinear symmetric operator space with respect to $\|\cdot\|_2$ is defined as: for all $T$ symetric trilinear operator, $\|T\|_{op} := \sup_{u\in\partial B_2(0,1)} |T(u,u,u)|$ as shown in equation (2) in both [145] and [122].

**Lemma 3.20.** *With trilinear symmetric operator defined above, the third derivative of the risk satisfies:* $\forall \beta \in \mathbb{R}^d$,

$$
\left\|d^3_{t\beta+(1-t)\beta_0}\mathcal{R}\right\|_{op} \leq 8e^{-\left(a^t\beta-\|\beta\|_2^2\right)} \sqrt{2\left(\|a\|^6+\mathbb{E}\left[N_0^6\right]\right)\left(\left[\|\beta\|_2^2+\left[a^t\beta-2\|\beta\|_2^2\right]^2\right]+\left[a^t\beta-2\|\beta\|_2^2\right]+1\right)},
$$

*where* $N_0 \sim \mathcal{N}(0,1)$.

*Proof.* if $u \in \partial B_2(0,1)$, then for $N \sim \mathcal{N}(0, I_d)$, $N^t u \sim N_0 \sim \mathcal{N}(0,1)$, and it is known that $\mathbb{E}[|N_0|] = \sqrt{\frac{2}{\pi}}$ and $\mathbb{E}\left[|N_0|^3\right] = 3\mathbb{E}[|N_0|]\,\mathbb{V}[|N_0|] + \mathbb{E}[|N_0|]^3$. Owing to Equation (3.5.13) and Cauchy-Schwarz inequality, we have

$$\begin{aligned}\left|d^3_\beta \mathcal{R}(u,u,u)\right| &= \left|\mathbb{E}\left[\left(X^t u\right)^3 .\alpha'\left(X^t\beta\right)\right]\right| \\ &\leq \sqrt{\mathbb{E}\left[\left(X^t u\right)^6\right]\mathbb{E}\left[\left(\alpha'\left(X^t\beta\right)\right)^2\right]}.\end{aligned}$$

On the one hand, using the fact that $\forall a,b>0, \forall n\in\mathbb{N}, (a+b)^n \leq 2^{n-1}\left(a^n+b^n\right)$ we get

$$\begin{aligned}\mathbb{E}\left[\left(X^t u\right)^6\right] &= \mathbb{E}\left[\left(a^t u + N^t u\right)^6\right] \\ &= \mathbb{E}\left[\left(a^t u + N_0\right)^6\right] \\ &\leq \mathbb{E}\left[2^5\left(\left(a^t u\right)^6 + N_0^6\right)\right] \\ &\leq 32\left(\|a\|_2^6 + \mathbb{E}\left[N_0^6\right]\right).\end{aligned}$$

On the other hand, we have already proved in Lemma 3.9 that $\forall x, \alpha'(x) = p_x q_x\left[\frac{x}{2}\left(1-3\tanh^2\left(\frac{x}{2}\right)\right)+2\tanh\left(\frac{x}{2}\right)\right]$. Hence,

$$\begin{aligned}|\alpha'(x)| &\leq p_x q_x\left|\frac{x}{2}\left(1-3\tanh^2\left(\frac{x}{2}\right)\right)+2\tanh\left(\frac{x}{2}\right)\right| \\ &\leq p_x q_x\left(\frac{x}{2}\left|1-3\tanh^2\left(\frac{x}{2}\right)\right|+2\right) \\ &\leq \frac{e^x}{\left(1+e^x\right)^2}\left(\frac{x}{2}\left(3\tanh^2\left(\frac{x}{2}\right)+1\right)+2\right) \\ &\leq \frac{e^x}{e^x\left(e^{-x/2}+e^{x/2}\right)^2}\left(\frac{x}{2}\times 4+2\right) \\ &\leq \frac{2}{\left(e^{x/2}\right)^2}(x+1) \\ &= 2e^{-x}(x+1).\end{aligned}$$

Recall $Z^t\beta \sim \mathcal{N}\left(a^t\beta, \|\beta\|_2^2\right)$,

$$\begin{aligned}\mathbb{E}\left[\left(\alpha'\left(X^t\beta\right)\right)^2\right] &= \mathbb{E}\left[\left(\alpha'\left(Z^t\beta\right)\right)^2\right] \\ &\leq \mathbb{E}\left[4e^{-2Z^t\beta}\left(Z^t\beta+1\right)^2\right] \\ &= 4\int_{\mathbb{R}}(x+1)^2 e^{-2x}\frac{1}{\sqrt{2\pi\|\beta\|_2}}\exp\left(-\frac{(x-a^t\beta)^2}{2\|\beta\|_2^2}\right)dx. \end{aligned} \tag{3.5.22}$$

By denoting $N_{a,\beta} \sim \mathcal{N}\left(a^t\beta - 2\|\beta\|_2^2, \|\beta\|_2^2\right)$, we get

$$\mathbb{E}\left[\left(\alpha'\left(X^t\beta\right)\right)^2\right] = 4\int_{\mathbb{R}}(x+1)^2\frac{1}{\sqrt{2\pi\|\beta\|_2^2}}\exp\left(-\frac{\left(x+\left(2\|\beta\|_2^2 - a^t\beta\right)\right)^2 + 2\|\beta\|_2^2\left(2a^t\beta - 2\|\beta\|_2^2\right)}{2\|\beta\|_2^2}\right)dx$$

$$\begin{aligned}&= 4e^{-2\left(a^t\beta-\|\beta\|_2^2\right)}\int_{\mathbb{R}}(x+1)^2\frac{1}{\sqrt{2\pi\|\beta\|_2^2}}\exp\left(-\frac{\left(x-\left(a^t\beta-2\|\beta\|_2^2\right)\right)^2}{2}\right)dx \\ &= 4e^{-2\left(a^t\beta-\|\beta\|_2^2\right)}\mathbb{E}\left[\left(N_{a,\beta}+1\right)^2\right] \\ &= 4e^{-2\left(a^t\beta-\|\beta\|_2^2\right)}\left(\mathbb{E}\left[N_{a,\beta}^2\right] + 2\mathbb{E}\left[N_{a,\beta}\right] + 1\right) \\ &= 4e^{-2\left(a^t\beta-\|\beta\|_2^2\right)}\left(\left[\|\beta\|_2^2 + \left[a^t\beta - 2\|\beta\|_2^2\right]^2\right] + 2\left[a^t\beta - 2\|\beta\|_2^2\right] + 1\right)\end{aligned}$$

Finally, we have

$$\left|d_\beta^3\mathcal{R}(u,u,u)\right| \leq 8e^{-\left(a^t\beta-\|\beta\|_2^2\right)}\sqrt{2\left(\|a\|^6 + \mathbb{E}\left[N_0^6\right]\right)\left(\left[\|\beta\|_2^2 + \left[a^t\beta - 2\|\beta\|_2^2\right]^2\right] + 2\left[a^t\beta - 2\|\beta\|_2^2\right] + 1\right)},$$

which gives the result, according to definition 3.1. □

**Lemma 3.21.** *Under the condition that* $\|a\|_2 \geq 2R$, $R = \sqrt{x_1 + 0.08}$, *the excess risk* $\mathcal{E}(\cdot, \beta_0)$ *satisfies around* $\beta_0$*:* $\forall\beta \in B_2(\beta_0, \varepsilon)\bigcap B_2(0, R)$,

$$\mathcal{E}(\beta,\beta_0) \geq e^{-\left(\|a\|_2 R - R^2/2\right)} \left[\frac{1}{16}\left(1+\left(\|a\|_2 - R\right)^2\right)\|\beta-\beta_0\|_2^2 - 24\,\|a\|^4\, e^{R^2/2} e^{\varepsilon\|a\|_2}\,\|\beta-\beta_0\|_2^3\right].$$

*Proof.* Fisrt note that $\mathcal{E}(\beta_0,\beta_0)=0$ by definition of $\mathcal{E}(\cdot,\beta_0)$. According to lemma 3.19 we can control $(d_\beta\mathcal{R})(\beta-\beta_0)$ from below.

Since $\langle\beta-\beta_0,\beta_0\rangle \leq 0$ and $a^t\beta_0 - 2\|\beta_0\|_2^2 \geq 0$, we have

$$\begin{aligned}(d_{\beta_0}\mathcal{R})(\beta-\beta_0) &\geq \frac{1}{8}e^{-\left(2a^t\beta_0 - \|\beta_0\|_2^2\right)/2}\left\langle \beta_0-\beta, \frac{\beta_0}{\|\beta_0\|_2^2}\right\rangle\left(\|\beta_0\|_2^2 + \left(a^t\beta_0 - \|\beta_0\|_2^2\right)^2\right)\\ &\geq \frac{1}{8}e^{-\left(\|a\|_2 R - R^2/2\right)}\langle\beta_0-\beta,\beta_0\rangle\left(1+\left(\|a\|_2-R\right)^2\right).\end{aligned}$$

Use now Lemma 3.23,

$$(d_{\beta_0}\mathcal{R})(\beta-\beta_0) \geq \frac{1}{16}e^{-\left(\|a\|_2 R - R^2/2\right)}\left(1+\left(\|a\|_2-R\right)^2\right)\|\beta-\beta_0\|_2^2$$

According to lemma 3.1, we can control $\left(d_\beta^2\mathcal{R}\right)(\beta-\beta_0,\beta-\beta_0)$ from below:

$$\left(d_\beta^2\mathcal{R}\right)(\beta-\beta_0,\beta-\beta_0) \geq \Lambda_{min}\|\beta-\beta_0\|_2^2 \geq 0.$$

In addition, we can use Lemma 3.20 to have

$$\begin{aligned}&\int_0^1 \left|\left(d^3_{t\beta+(1-t)\beta_0}\mathcal{R}\right)(\beta-\beta_0,\beta-\beta_0,\beta-\beta_0)\right| dt\\ &= \|\beta-\beta_0\|_2^3\int_0^1\left|\left(d^3_{t\beta+(1-t)\beta_0}\mathcal{R}\right)\left(\frac{\beta-\beta_0}{\|\beta-\beta_0\|_2},\frac{\beta-\beta_0}{\|\beta-\beta_0\|_2},\frac{\beta-\beta_0}{\|\beta-\beta_0\|_2}\right)\right| dt\\ &\leq \|\beta-\beta_0\|_2^3\int_0^1\left\|d^3_{t\beta+(1-t)\beta_0}\mathcal{R}\right\|_{op} dt\\ &\leq 8\,\|\beta-\beta_0\|_2^3\int_0^1 \exp\left(-\left(a^t\left(t\beta+(1-t)\beta_0\right)-\|t\beta+(1-t)\beta_0\|_2^2\right)\right)C_{3,a}\left(t\beta+(1-t)\beta_0\right)dt,\end{aligned}$$

where

$$C_{3,a}:\mu\in\mathbb{R}^d \mapsto \sqrt{2\left(\|a\|_2^6+\mathbb{E}\left[N_0^6\right]\right)\left(\left[\|\mu\|_2^2+\left[a^t\nu-2\|\mu\|_2^2\right]^2\right]+2\left[a^t\nu-2\|\mu\|_2^2\right]+1\right)}.$$

To bound $C_{3,a}$ from above, remark that $\forall \mu \in \Psi_U$, $\|\mu\|_2^2 \leq 4R^2$, $-2R^2 \leq a^t\nu - 2\|\mu\|_2^2 \leq R\|a\|_2 - 2R^2 \leq R\|a\|_2$ and remark also that owing to $\|a\| \geq 2R \approx 2.548$ and article [143], one has $\mathbb{E}\left[N_0^6\right] = 15$ with $\mathcal{N}(0,1)$, so $\mathbb{E}\left[N_0^6\right] \leq \frac{1}{18}\|a\|_2^6$. Therefore, all together this leads to

$$\begin{aligned}
C_{3,a}(\mu) &\leq \sqrt{2\left(\|a\|_2^6 + \mathbb{E}\left[N_0^6\right]\right)\left(\left[R^2 + R^2\left[\max\left(2R, \|a\|_2\right)\right]^2\right] + 2\left[R\|a\|_2 - 2R^2\right] + 1\right)} \\
&\leq \sqrt{2\left(\|a\|_2^6 + \frac{1}{18}\|a\|_2^6\right)\left(\left[R^2 + R^2\|a\|_2^2\right] + 2\left[R\|a\|_2 - 2R^2\right] + 1\right)} \\
&\leq \sqrt{\frac{19}{9}\|a\|_2^6\left(R^2\|a\|_2^2 + 2R\|a\|_2 - 3R^2 + 1\right)} \\
&\leq 3\|a\|^4. \qquad (3.5.23)
\end{aligned}$$

Hence

$$\begin{aligned}
&\int_0^1 \left|\left(d^3_{t\beta+(1-t)\beta_0}\mathcal{R}\right)(\beta-\beta_0, \beta-\beta_0, \beta-\beta_0)\right| dt \\
&\leq 24\|a\|^4\|\beta-\beta_0\|_2^3 \int_0^1 \exp\left(-\left(a^t(t\beta+(1-t)\beta_0) - \|t\beta+(1-t)\beta_0\|_2^2\right)\right) dt \\
&\leq 24\|a\|^4\|\beta-\beta_0\|_2^3 \sup_{\|\beta-\beta_0\|_2\leq\varepsilon 0\leq t\leq 1} \exp\left(-\left(a^t(t\beta+(1-t)\beta_0) - \|t\beta+(1-t)\beta_0\|_2^2\right)\right) \\
&\leq 24\|a\|^4\|\beta-\beta_0\|_2^3 \exp\left(-\inf_{\|\beta-\beta_0\|_2\leq\varepsilon 0\leq t\leq 1} a^t(t\beta+(1-t)\beta_0) + \sup_{\|\beta-\beta_0\|_2\leq\varepsilon 0\leq t\leq 1}\|t\beta+(1-t)\beta_0\|_2^2\right) \\
&\leq 24\|a\|^4\|\beta-\beta_0\|_2^3 \exp\left(-a^t\beta_0 + \|a\|_2\varepsilon + R^2\right) \\
&\leq 24\|a\|^4 e^{-\|a\|_2 R + R^2} e^{\varepsilon\|a\|_2}\|\beta-\beta_0\|_2^3.
\end{aligned}$$

Finally, this gives

$$\mathcal{E}(\beta,\beta_0) > e^{-\left(\|a\|_2 R - R^2/2\right)}\left[\frac{1}{16}\left(1 + \left(\|a\|_2 - R\right)^2\right)\|\beta-\beta_0\|_2^2 - 24\|a\|^4 e^{R^2/2} e^{\varepsilon\|a\|_2}\|\beta-\beta_0\|_2^3\right], \qquad (3.5.24)$$

as required. □

**Lemma 3.22.** *Minoration of* $\Phi^c\left(\|a\|_2 - \frac{x_1}{R}\right) - \Phi^c\left(\|a\|_2 + \frac{x_1}{R}\right)$*:*

$$\forall a, R, x_1, \Phi^c\left(\|a\|_2 - \frac{x_1}{R}\right) - \Phi^c\left(\|a\|_2 + \frac{x_1}{R}\right) \geq 2\frac{x_1}{R}\gamma\left(\|a\|_2 + \frac{x_1}{R}\right)$$

*Proof.* Simple computations give

$$\begin{aligned}\Phi^c\left(\|a\|_2 - \frac{x_1}{R}\right) - \Phi^c\left(\|a\|_2 + \frac{x_1}{R}\right) &= \int_{\|a\|_2 - \frac{x_1}{R}}^{\|a\|_2 + \frac{x_1}{R}} \gamma(x)\, d\lambda(x) \\ &\geq \int_{\|a\|_2 - \frac{x_1}{R}}^{\|a\|_2 + \frac{x_1}{R}} \gamma\left(\|a\|_2 + \frac{x_1}{R}\right) d\lambda(x) \\ &\geq 2\frac{x_1}{R}\gamma\left(\|a\|_2 + \frac{x_1}{R}\right)\end{aligned}$$

□

**Lemma 3.23.** *$\forall \beta \in B_2(0,R)$, if $\beta_0 \in \partial B_2(0,R)$ then $\langle \beta_0 - \beta, \beta_0 \rangle \geq \frac{1}{2}\|\beta - \beta_0\|_2^2$.*

*Proof.* Decompose $\beta$ as $\beta_\perp + \beta_\parallel$ on $\beta_0^\perp \oplus Vect(\beta_0)$ and note that $\exists \lambda_\beta \in [-1,1], \beta_\parallel = \lambda_\beta \beta_0$. We have

$$\frac{\langle \beta_0 - \beta, \beta_0 \rangle}{\|\beta - \beta_0\|_2^2} = \frac{\langle \beta_0 - \beta_\parallel, \beta_0 \rangle}{\|\beta_\perp\|_2^2 + \|\beta_\parallel - \beta_0\|_2^2} = \frac{(1-\lambda_\beta) R^2}{\|\beta_\perp\|_2^2 + (1-\lambda_\beta)^2 R^2}$$

Furthermore, we have $\|\beta\|_2^2 \leq R^2$ and by pythagora's theorem $\|\beta_\perp\|_2^2 \in \left[0, R^2 - \lambda_\beta^2 R^2\right]$. Therefore,

$$\begin{aligned}\frac{\langle \beta_0 - \beta, \beta_0 \rangle}{\|\beta - \beta_0\|_2^2} &\geq \frac{(1-\lambda_\beta) R^2}{\left(1-\lambda_\beta^2\right) R^2 + (1-\lambda_\beta)^2 R^2} \\ &\geq \frac{1-\lambda_\beta}{1 - \lambda_\beta^2 + \left(1 - 2\lambda_\beta + \lambda_\beta^2\right)} \\ &\geq \frac{1}{2}.\end{aligned}$$

□

**Definition 3.2.** When one has $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ with $x_1, \ldots, x_n \in \mathcal{X}$, define the following "empirical-$L^2$- norm" as:

$$\forall f : \mathcal{X} \to \mathbb{R}, \|f\|_{P_n} := \sqrt{\frac{1}{n}\sum_{i=1}^n f^2\left(X^{(i)}\right)}$$

.

**Definition 3.3.** For $\delta > 0$, the $\delta$-covering number $N(\delta, \mathscr{H}, \|\cdot\|)$ of a set $\mathscr{H}$ is the smallest number of closed balls, with respect to $\|\cdot\|$ with radius $\delta$, that covers the space. The set of the centers of the balls is called a $\delta$-covering set. The entropy of $\mathscr{H}$ with respect to a norm $\|\cdot\|$ is $H(\cdot, \mathscr{H}, \|\cdot\|) = \log N(\cdot, \mathscr{H}, \|\cdot\|)$.

.

**Lemma 3.24.** *Define* $\Theta(\varepsilon) := \{\beta \in B_2(0,R) : \|\beta - \beta_0\|_1 \leq \varepsilon\}$ *and take*

$$\mathscr{H}_{\varepsilon,M_n} := \left\{(\rho_\beta - \rho_{\beta_0}) I_{\{G \leq M_n\}} - \mathbb{E}\left[(\rho_\beta(X) - \rho_{\beta_0}(X)) I_{\{G(X) \leq M_n\}}\right] : \beta \in \Theta(\varepsilon)\right\},$$

*where* $G(X) := \|X\|_\infty$. *Recall that* $L$ *is the Lipschitz constant of* $\rho$.
*Then for all* $u > 0$ *and* $M_n > 0$, *the entropy of* $\mathscr{H}_{\varepsilon,M_n}$ *with respect to the empirical-*$L^2$*-norm* $\|\cdot\|_{P_n}$ *(see definition 3.2) satisfies*

$$H\left(u, \mathscr{H}_{\varepsilon,M_n}, \|\cdot\|_{P_n}\right) \leq \left(\frac{4L^2 M_n^2 \varepsilon^2}{u^2} + 1\right) \log(2d).$$

*Proof.* Let $\widehat{X_i}, \ldots, \widehat{X_i}$ be i.i.d copies of $X$ and set $\mathscr{B}_{\varepsilon,M_n} := \left\{f_{\beta,\beta'} : X \mapsto \frac{X^t}{M_n}(\beta - \beta') I_{\{G(X) \leq M_n\}} : \beta, \beta' \in \Theta(\varepsilon)\right\}$.
One has $\forall \beta, \beta' \in \Theta(\varepsilon)$,

$$|\rho_\beta(X) - \rho_{\beta'}(X)| = \left|\rho\left(X^t\beta\right) - \rho\left(X^t\beta'\right)\right| \leq L\left|X^t\beta - X^t\beta'\right|.$$

With $\forall a, b > 0, (a+b)^2 \leq 2\left(a^2 + b^2\right)$, it follows that

$$\begin{aligned}
&\left\|\rho_\beta I_{\{G(\cdot)\leq M_n\}} - \mathbb{E}\left[\rho_\beta I_{\{G(\cdot)\leq M_n\}}\right] - \rho_{\beta'} I_{\{G(\cdot)\leq M_n\}} + \mathbb{E}\left[\rho_{\beta'} I_{\{G(\cdot)\leq M_n\}}\right]\right\|_{P_n}^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left(\rho_\beta\left(X^{(i)}\right) I_{\left\{G\left(X^{(i)}\right)\leq M_n\right\}} - \mathbb{E}\left[\rho_\beta(X) I_{\{G(X)\leq M_n\}}\right] - \left(\rho_{\beta'}\left(X^{(i)}\right) I_{\left\{G\left(X^{(i)}\right)\leq M_n\right\}} - \mathbb{E}\left[\rho_{\beta'}(X) I_{\{G(X)\leq M_n\}}\right]\right)\right)^2 \\
&\leq \frac{2}{n}\sum_{i=1}^n \left(\rho_\beta\left(X^{(i)}\right) - \rho_{\beta'}\left(X^{(i)}\right)\right)^2 I_{\left\{G\left(X^{(i)}\right)\leq M_n\right\}} + \frac{2}{n}\sum_{i=1}^n \left(\mathbb{E}\left[\rho_{\beta'}(X) I_{\{G(X)\leq M_n\}}\right] - \mathbb{E}\left[\rho_\beta(X) I_{\{G(X)\leq M_n\}}\right]\right)^2 \\
&\leq \frac{2}{n}\sum_{i=1}^n \left(\rho_\beta\left(X^{(i)}\right) - \rho_{\beta'}\left(X^{(i)}\right)\right)^2 I_{\left\{G\left(X^{(i)}\right)\leq M_n\right\}} + 2\mathbb{E}\left[\left(\rho_{\beta'}(X) - \rho_\beta(X)\right)^2 I_{\{G(X)\leq M_n\}}\right].
\end{aligned}$$

Furthermore,

$$\begin{aligned}\frac{2}{n}\sum_{i=1}^{n}\left(\rho_\beta\left(X^{(i)}\right)-\rho_{\beta'}\left(X^{(i)}\right)\right)^2 I_{\left\{G\left(X^{(i)}\right)\leq M_n\right\}} &\leq \frac{2}{n}\sum_{i=1}^{n} L^2\left|X^{(i)t}\beta - X^{(i)t}\beta'\right|^2 I_{\left\{G\left(X^{(i)}\right)\leq M_n\right\}} \\ &\leq \frac{2}{n}L^2\sum_{i=1}^{n} G\left(X^{(i)}\right)^2\left\|\beta-\beta'\right\|_1^2 I_{\left\{G\left(X^{(i)}\right)\leq M_n\right\}} \\ &\leq 2L^2M_n^2\left\|\beta-\beta'\right\|_1^2.\end{aligned}$$

One also has

$$\mathbb{E}\left[\left(\rho_{\beta'}(X)-\rho_\beta(X)\right)^2 I_{\{G(X)\leq M_n\}}\right]\leq L^2M_n^2\left\|\beta-\beta'\right\|_1^2.$$

Hence

$$\left\|\rho_\beta I_{\{G(\cdot)\leq M_n\}}-\mathbb{E}\left[\rho_\beta I_{\{G(\cdot)\leq M_n\}}\right]+\rho_{\beta'}I_{\{G(\cdot)\leq M_n\}}-\mathbb{E}\left[\rho_{\beta'}I_{\{G(\cdot)\leq M_n\}}\right]\right\|_{P_n}^2\leq 4L^2M_n^2.\left\|\beta-\beta'\right\|_1^2. \tag{3.5.25}$$

This relation enables us to state

$$H\left(u,\mathscr{H}_{\varepsilon,M_n},\|\cdot\|_{P_n}\right)\leq H\left(\frac{u}{2LM_n},\Theta(\varepsilon),\|\cdot\|_1\right).$$

Define the convex hull of a set of vectors $\{e_j\}_{j=1}^d$ as $Conv\{e_j\}_{j=1}^d := \left\{\sum_{j=1}^d v_je_j\middle| v_i\geq 0,\|v\|_1=1\right\}$ and take in particular the vectors $\{e_j\}_{j=1}^d$ of the canonical basis in $\mathbb{R}^d$. Then

$$\Theta(\varepsilon)\subset\beta_0+\varepsilon.Conv\left\{0,\{\pm e_j\}_{j=1}^d\right\}.$$

Owing to the definition of $e_j$, we have $\forall j, \left\|e_j\right\|_1 = 1$. so we can use Lemma 14.28 in [29] to get

$$\begin{aligned} H\left(u, \Theta\left(\varepsilon\right), \left\|\cdot\right\|_1\right) &\leq H\left(u, \beta_0 + \varepsilon.Conv\left\{0, \left\{\pm e_j\right\}_{j=1}^d\right\}, \left\|\cdot\right\|_1\right) \\ &\leq H\left(\frac{u}{\varepsilon}, Conv\left\{0, \left\{\pm e_j\right\}_{j=1}^d\right\}, \left\|\cdot\right\|_1\right) \\ &\leq H\left(\frac{u}{\varepsilon}, Conv\left\{0, \left\{\pm e_j\right\}_{j=1}^d\right\}, \left\|\cdot\right\|_1\right) \\ &\leq \left\lceil \frac{\varepsilon^2}{u^2} \right\rceil \left(1 + \log\left(1 + \left(2d+1\right)\frac{u^2}{\varepsilon^2}\right)\right) \wedge \left\lceil \frac{\varepsilon^2}{u^2} \right\rceil \log\left(2d\right) \\ &\leq \left(\frac{\varepsilon^2}{u^2} + 1\right) \log\left(2d\right), \end{aligned}$$

which gives the result. □

**Lemma 3.25.** *Let $\varepsilon > 0$ and $X^{(1)}, ..., X^{(i)}, ..., X^{(n)}$ be i.i.d. copies of $X$. Let also*

$$\mathscr{H}_{\varepsilon, M_n} := \left\{\left(\rho_\beta - \rho_{\beta_0}\right) I_{\{G \leq M_n\}} - \mathbb{E}\left[\left(\rho_\beta(X) - \rho_{\beta_0}(X)\right) I_{\{G(X) \leq M_n\}}\right] : \beta \in \Theta\left(\varepsilon\right)\right\},$$

*where $G\left(X\right) := \left\|X\right\|_\infty$ and*

$$\Theta\left(\varepsilon\right) := \left\{\beta \in \mathbb{R}^d : \beta \in B_2\left(0, R\right), \left\|\beta - \beta_0\right\|_1 \leq \varepsilon\right\}.$$

*Recall that we set $L$, the Lipschitz norm of $\rho$. One has $\forall T \geq 1$, $\forall n \geq 2$,*

$$P\left(\sup_{\beta \in \Theta(\varepsilon)} \frac{\left|V_n^{trunc}\left(\beta\right) - V_n^{trunc}\left(\beta_0\right)\right|}{\varepsilon} \geq \frac{3LM_nT\left(5\sqrt{3\log\left(2d\right)}\log n + 4\right)}{\sqrt{n}}\right) < \exp\left(-21\left(T-1\right)^2\log\left(2d\right)\log^2 n\right).$$

*Proof.* According to equation (3.5.25), $\forall \widetilde{\rho}_\beta \in \mathscr{H}_{\varepsilon, M_n}$, $\left\|\widetilde{\rho}_\beta\right\|_{P_n} \leq 2LM_n\varepsilon =: R_n$ and $\mathbb{E}\left(\widetilde{\rho}_\beta\left(X\right)\right) = 0$. Hence, using Lemma 3.24 and Definition 3.3, we have

$$\begin{aligned} \log\left(1 + N\left(u, \mathscr{H}_{\varepsilon, M_n}, \left\|\cdot\right\|_{P_n}\right)\right) &\leq 1 + H\left(u, \mathscr{H}_{\varepsilon, M_n}, \left\|\cdot\right\|_{P_n}\right) \\ &\leq 1 + \left(\frac{4L^2M_n^2\varepsilon^2}{u^2} + 1\right)\log\left(2d\right) \\ &\leq \left(\frac{4L^2M_n^2\varepsilon^2}{u^2} + 2\right)\log\left(2d\right) \end{aligned}$$

Take $u := 2^{-s}R_n$ where $0 \leq s \leq S := \min\left\{s \geq 1 : 2^{-s} \leq \frac{4}{\sqrt{n}}\right\}$ (i.e. $u \in \left[\frac{2}{\sqrt{n}}R_n, R_n\right]$),

then one has $\forall 0 \leq s \leq S$,

$$\begin{aligned}\log\left(1+N\left(2^{-s}R_n, \mathscr{H}_{\varepsilon,M_n}, \|\cdot\|_{P_n}\right)\right) &\leq \left(\frac{4L^2M_n^2\varepsilon^2}{2^{-2s}R_n^2}+2\right)\log(2d)\\ &\leq \left(2^{2s}+2\right)\log(2d)\\ &\leq 2^{2s}\left(1+2^{1-2s}\right)\log(2d)\\ &\leq 2^{2s}\times 3\log(2d)\end{aligned}$$

Now one can apply [29, Corollary 14.4], where in our case $A := 3\log(2d)$. Note that $4\log n \leq 3\log_2 n \leq 5\log n$. We get

$$\mathbb{E}\left[\sup_{\widetilde{\rho}\in\mathscr{H}_{\varepsilon,M_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\rho}_\beta\left(X^{(i)}\right)\right|\right] \leq \frac{R_n}{\sqrt{n}}\left(5\sqrt{A}\log n+4\right).$$

One can apply the Massart's concentration inequality, recalled for instance in [29, Theorem 14.2]. ,Then, $\forall t>0$,

$$P\left(\sup_{\widetilde{\rho}\in\mathscr{H}_{\varepsilon,M_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\widetilde{\rho}_\beta\left(X^{(i)}\right)}{R_n}\right| \geq \mathbb{E}\left[\sup_{\widetilde{\rho}\in\mathscr{H}_{\varepsilon,M_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\widetilde{\rho}_\beta\left(X^{(i)}\right)}{R_n}\right|\right]+t\right) \leq e^{-nt^2/8},$$

which gives

$$P\left(\sup_{\widetilde{\rho}\in\mathscr{H}_{\varepsilon,M_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\rho}_\beta\left(X^{(i)}\right)\right| \geq \frac{R_n}{\sqrt{n}}\left(5\sqrt{A}\log n+4\right)+R_nt\right) \leq e^{-nt^2/8}.$$

A change of variable $t=\frac{1}{\sqrt{n}}(T-1)\left(5\sqrt{A}\log n+4\right)$ leads to: $\forall T\geq 1$,

$$P\left(\sup_{\widetilde{\rho}\in\mathscr{H}_{\varepsilon,M_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\rho}_\beta\left(X^{(i)}\right)\right| \geq \frac{R_n}{\sqrt{n}}T\left(5\sqrt{A}\log n+4\right)\right) < \exp\left(-\frac{(T-1)^2\left(5\sqrt{A}\log n+4\right)^2}{8}\right)$$

Note that $\forall\widetilde{\rho}_\beta\in\mathscr{H}_{\varepsilon,M_n}$,

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{\rho}_\beta\left(X^{(i)}\right) = V_n^{trunc}(\beta)-V_n^{trunc}(\beta_0).$$

Consequently, $\forall T\geq 1$,

$$P\left(\sup_{\beta\in\Theta(\varepsilon)}\left|V_n^{trunc}(\beta)-V_n^{trunc}(\beta_0)\right|\geq\frac{3LM_n\varepsilon T\left(5\sqrt{3\log(2d)}\log n+4\right)}{\sqrt{n}}\right)<\exp\left(-\frac{(3T/2-1)^2\left(5\sqrt{3\log(2d)}\log n+4\right)^2}{8}\right)$$
$$<\exp\left(-21(T-1)^2\log(2d)\log^2 n\right).$$

□

**Lemma 3.26.** *Grant the notations of Lemma 3.25 and set* $\lambda_0:=3LM_n\left(5\sqrt{3\log(2d)}\log n+4\right)n^{-1/2}$. *One has* $\forall T\geq 1$, $\forall n\geq 2$,

$$P\left(\sup_{\beta\in B_2(0,R)}\frac{|V_n^{trunc}(\beta)-V_n^{trunc}(\beta_0)|}{\|\beta-\beta_0\|_1\vee\lambda_0}\geq T\lambda_0\right)\leq\frac{3}{4}\log\left(\frac{4R^2nd}{L^2M_n^2}\right)\exp\left(-21(T-1)^2\log(2d)\log^2 n\right).$$

*Proof.* Let $\lambda_0>0$, $n\geq 2$ and $T\geq 1$. Let us use a peeling: define $\Theta:=B_2(0,R)$ and divide it into slices as follows:

$$\Theta_j:=\left\{\beta\in B_2(0,R):2^{-j-1}\leq\|\beta-\beta_0\|_1\leq 2^{-j}\right\}.$$

Note that $\exists j_{inf},j_{sup}\in\mathbb{Z}$, $\exists r>0$, $2^{-j_{sup}-1}\leq\lambda_0\leq 2^{-j_{sup}}$ and $2^{-j_{inf}-1}\leq r:=2R\sqrt{d}\leq 2^{-j_{inf}}$ with

$$\Theta\subset\bigcup_{j=j_{inf}}^{j_{sup}}\Theta_j\bigcup B_1\left(\beta_0,2^{-j_{sup}}\right)$$

and $\Theta\subset B_1\left(\beta_0,2^{-j_{inf}-1}\right)$. One can also prove that $j_{inf}=\lfloor-\log_2 r\rfloor=\left\lfloor-\log_2 2R\sqrt{d}\right\rfloor\leq -1$ because $R,d\geq 1$ and $j_{sup}=\lfloor-\log_2(\lambda_0)\rfloor$. Hence

$$P\left(\sup_{\beta\in\Theta}\frac{|V_n^{trunc}(\beta)-V_n^{trunc}(\beta_0)|}{\|\beta-\beta_0\|_1\vee\lambda_0}\geq T\lambda_0\right)\leq\sum_{j=j_{inf}}^{j_{sup}}P\left(\sup_{\beta\in\Theta_j}\frac{|V_n^{trunc}(\beta)-V_n^{trunc}(\beta_0)|}{\|\beta-\beta_0\|_1\vee\lambda_0}\geq T\lambda_0\right)$$
$$+P\left(\sup_{\beta\in B_1\left(\beta_0,2^{-j_{sup}}\right)}\frac{|V_n^{trunc}(\beta)-V_n^{trunc}(\beta_0)|}{\|\beta-\beta_0\|_1\vee\lambda_0}\geq T\lambda_0\right).$$

Use the fact that $\forall j\in[\![j_{inf},j_{sup}]\!]$, $\lambda_0\leq 2^{-j}$ and $\forall\beta\in B_1\left(\beta_0,2^{-j_{sup}}\right)$, $\|\beta-\beta_0\|_1\vee\lambda_0\leq 2^{-j_{sup}}$:

$$P\left(\sup_{\beta\in\Theta}\frac{\left|V_n^{trunc}(\beta)-V_n^{trunc}(\beta_0)\right|}{\|\beta-\beta_0\|_1\vee\lambda_0}\geq T\lambda_0\right)\leq\sum_{j=j_{inf}}^{j_{sup}}P\left(\sup_{\beta\in\Theta_j}\frac{\left|V_n^{trunc}(\beta)-V_n^{trunc}(\beta_0)\right|}{2^{-j}}\geq T\lambda_0\right)$$
$$+P\left(\sup_{\beta\in B_1\left(\beta_0,2^{-j_{sup}}\right)}\frac{\left|V_n^{trunc}(\beta)-V_n^{trunc}(\beta_0)\right|}{2^{-j_{sup}}}\geq T\lambda_0\right).$$

By applying Lemma 3.25 with $\lambda_0=3LM_n\left(5\sqrt{3\log(2d)}\log n+4\right)n^{-1/2}$, we get

$$P\left(\sup_{\beta\in\Theta_j}\frac{\left|V_n^{trunc}(\beta)-V_n^{trunc}(\beta_0)\right|}{2^{-j}}\geq T\lambda_0\right)<\exp\left(-21(T-1)^2\log(2d)\log^2 n\right).$$

Then

$$P\left(\sup_{\beta\in\Theta}\frac{\left|V_n^{trunc}(\beta)-V_n^{trunc}(\beta_0)\right|}{\|\beta-\beta_0\|_1\vee\lambda_0}\geq T\lambda_0\right)\leq\sum_{j=j_{inf}}^{j_{sup}}\exp\left(-21(T-1)^2\log(2d)\log^2 n\right)$$
$$+\exp\left(-21(T-1)^2\log(2d)\log^2 n\right)$$
$$\leq(j_{sup}-j_{inf}+2)\exp\left(-21(T-1)^2\log(2d)\log^2 n\right).$$

Simplify now the expression of $j_{sup}-j_{inf}+2$,

$$\begin{aligned}
j_{sup}-j_{inf}+2&=\left\lceil\log_2 2R\sqrt{d}\right\rceil-\lceil\log_2(\lambda_0)\rceil+2\\
&\leq\log_2 8R\sqrt{d}+1-\log_2(\lambda_0)\\
&\leq\log_2\frac{16R\sqrt{d}}{\lambda_0}\\
&\leq\log_2\left(\frac{2R\sqrt{nd}}{LM_n\sqrt{3\log(2d)}\log n}\right)\\
&\leq\log_2\left(\frac{2R\sqrt{nd}}{LM_n}\right)\\
&\leq\frac{3}{2}\log\left(\sqrt{\frac{4R^2nd}{L^2M_n^2}}\right)\\
&\leq\frac{3}{4}\log\left(\frac{4R^2nd}{L^2M_n^2}\right).
\end{aligned}$$

This finally gives the result. □

**Lemma 3.27.** *With $G(X) := \|X\|_\infty$ and $a$ and $X$ defined in the section Notations: if $M_n = \|a\|_\infty + \sqrt{2\log d} + \sqrt{2\log n}$ then*

$$\mathbb{E}\left(G(X) I_{\{G(X)>M_n\}}\right) \leq 2(M_n+1)\frac{e^{-2\sqrt{\log d \log(1+n)}}}{n}$$

*and*

$$\mathbb{E}\left(G(X)^2 I_{\{G(X)>M_n\}}\right) \leq 2\left(M_n^2 + \|a\|_\infty + 1\right)\frac{e^{-2\sqrt{\log d \log(1+n)}}}{n}.$$

*Proof.* First note that for $y \geq \|a\|_\infty$,

$$\begin{aligned} P(G(X) > y) &\leq P\left(\max_j |Z_j| > y - \|a\|_\infty\right) \\ &\leq dP\left(|Z_1| > y - \|a\|_\infty\right) \\ &\leq 2de^{-\frac{(y-\|a\|_\infty)^2}{2}}. \end{aligned}$$

Take now $M \geq \|a\|_\infty + 1$, we have

$$\begin{aligned} \mathbb{E}\left(G(X) I_{\{G(X)>M\}}\right) &= -\int_M^\infty y\frac{dP(G(X)>y)}{dy}dy \\ &= \left[-yP(G(X)>y)\right]_M^\infty + \int_M^\infty 1 \times P(G(X)>y)\,dy \\ &\leq 2Mde^{-\frac{(M-\|a\|_\infty)^2}{2}} + \int_M^\infty (y-\|a\|_\infty) \times P(G(X)>y)\,dy \\ &= 2Mde^{-\frac{(M-\|a\|_\infty)^2}{2}} + 2\int_M^\infty (y-\|a\|_\infty) \times de^{-\frac{(y-\|a\|_\infty)^2}{2}}dy \\ &= 2Mde^{-\frac{(M-\|a\|_\infty)^2}{2}} + 2\left[-de^{-\frac{(y-\|a\|_\infty)^2}{2}}\right]_M^\infty \\ &= 2Mde^{-\frac{(M-\|a\|_\infty)^2}{2}} + 2de^{-\frac{(M-\|a\|_\infty)^2}{2}} \\ &= 2(M+1)de^{-\frac{(M-\|a\|_\infty)^2}{2}} \end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left(G(X)^2 I_{\{G(X)>M\}}\right) &= -\int_M^\infty y^2 \frac{dP(G(X)>y)}{dy} dy \\
&= \left[-y^2 P(G(X)>y)\right]_M^\infty + \int_M^\infty y \times P(G(X)>y)\, dy \\
&= 2M^2 d e^{-\frac{(M-\|a\|_\infty)^2}{2}} + \int_M^\infty \left(y-\|a\|_\infty\right) \times P(G(X)>y)\, dy + \int_M^\infty \|a\|_\infty \times P(G(X)>y)\, dy \\
&\leq 2M^2 d e^{-\frac{(M-\|a\|_\infty)^2}{2}} + 2d e^{-\frac{(M-\|a\|_\infty)^2}{2}} + 2\|a\|_\infty \int_M^\infty \underbrace{\left(y-\|a\|_\infty\right)}_{\geq 1} \times d e^{-\frac{(y-\|a\|_\infty)^2}{2}} dy \\
&\leq M^2 d e^{-\frac{(M-\|a\|_\infty)^2}{2}} + d e^{-\frac{(M-\|a\|_\infty)^2}{2}} + 2\|a\|_\infty d e^{-\frac{(M-\|a\|_\infty)^2}{2}} \\
&= 2\left(M^2 + \|a\|_\infty + 1\right) d e^{-\frac{(M-\|a\|_\infty)^2}{2}}.
\end{aligned}$$

Hence, for $M_n := \|a\|_\infty + \sqrt{2\log d} + \sqrt{2\log(1+n)} \geq \|a\|_\infty + 1$, we have

$$\begin{aligned}
\mathbb{E}\left(G(X) I_{\{G(X)>M_n\}}\right) &\leq 2(M_n+1) d e^{-\frac{(M_n-\|a\|_\infty)^2}{2}} \\
&\leq 2(M_n+1) d e^{-\frac{\left(\sqrt{2\log d}+\sqrt{2\log(1+n)}\right)^2}{2}} \\
&\leq 2(M_n+1) d e^{-\log d - 2\sqrt{\log d \log(1+n)} - \log(1+n)} \\
&\leq 2(M_n+1) \frac{e^{-2\sqrt{\log d \log(1+n)}}}{1+n} \\
&\leq 2(M_n+1) \frac{e^{-2\sqrt{\log d \log(1+n)}}}{n}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left(G(X)^2 I_{\{G(X)>M_n\}}\right) &\leq 2\left(M_n^2 + \|a\|_\infty + 1\right) d e^{-\frac{(M_n-\|a\|_\infty)^2}{2}} \\
&= 2\left(M_n^2 + \|a\|_\infty + 1\right) \frac{e^{-2\sqrt{\log d \log(1+n)}}}{1+n} \\
&\leq 2\left(M_n^2 + \|a\|_\infty + 1\right) \frac{e^{-2\sqrt{\log d \log(1+n)}}}{n}.
\end{aligned}$$

□

**Lemma 3.28.** *Assume that* $\|a\|_2 \geq 2R \approx 2.548$. *Set*

$$F(X) := G(X) I_{\{G(X)>M_n\}} + \mathbb{E}\left[G(X) I_{\{G(X)>M_n\}}\right],$$

*where* $G(X) = \|X\|_\infty$. *Moreover, take the following constants:* $M_n := \|a\|_\infty + \sqrt{2\log d} + \sqrt{2\log(1+n)}$, $\lambda_0 := 3LM_n n^{-1/2}\left(5\sqrt{3\log(2d)}\log n + 4\right)$. *It holds:* $\forall T > 0$,

$$P\left(\frac{1}{n}\sum_{i=1}^n F\left(X^{(i)}\right) \geq \frac{\lambda_0 T}{L}\right) \leq 4\frac{L^2}{\lambda_0^2 T^2}\frac{M_n^2 + \|a\|_\infty + 1}{n^2}.$$

*Proof.* Note that with our choice of $\lambda_0$, we have by Lemma 3.27:$\lambda_0 T/L \geq 2\mathbb{E}\left[G(X) I_{\{G(X)>M_n\}}\right]$. Hence,

$$\begin{aligned}
P\left(\frac{1}{n}\sum_{i=1}^n F\left(X^{(i)}\right) \geq \frac{\lambda_0 T}{L}\right) &= P\left(\frac{1}{n}\sum_{i=1}^n \left(G\left(X^{(i)}\right) I_{\left\{G\left(X^{(i)}\right)>M_n\right\}} + \mathbb{E}\left[G(X) I_{\{G(X)>M_n\}}\right]\right) \geq \frac{\lambda_0 T}{L}\right) \\
&= P\left(\frac{1}{n}\sum_{i=1}^n \left(G\left(X^{(i)}\right) I_{\left\{G\left(X^{(i)}\right)>M_n\right\}} - \mathbb{E}\left[G(X) I_{\{G(X)>M_n\}}\right]\right) \geq \frac{\lambda_0 T}{L} - 2\mathbb{E}\left[G(X) I_{\{G(X)>M_n\}}\right]\right) \\
&\leq \frac{\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^n G\left(X^{(i)}\right) I_{\left\{G\left(X^{(i)}\right)>M_n\right\}}\right)}{\left(\frac{\lambda_0 T}{L} - 2\mathbb{E}\left[G(X) I_{\{G(X)>M_n\}}\right]\right)^2} \\
&\leq \frac{\mathbb{E}\left(G(X)^2 I_{\{G(X)>M_n\}}\right)}{n\left(\frac{\lambda_0 T}{L} - 2\mathbb{E}\left[G(X) I_{\{G(X)>M_n\}}\right]\right)^2}.
\end{aligned}$$

From Lemma 3.27, we get

$$\begin{aligned}
P\left(\frac{1}{n}\sum_{i=1}^n F\left(X^{(i)}\right) \geq \frac{\lambda_0 T}{L}\right) &\leq \frac{\mathbb{E}\left(G(X)^2 I_{\{G(X)>M_n\}}\right)}{n\left(\frac{\lambda_0 T}{L} - 2\mathbb{E}\left[G(X) I_{\{G(X)>M_n\}}\right]\right)^2} \\
&\leq \frac{2\left(M_n^2 + \|a\|_\infty + 1\right)\frac{e^{-2\sqrt{\log d \log(1+n)}}}{n}}{n\left(\frac{\lambda_0 T}{L} - 4\left(M_n+1\right)\frac{e^{-2\sqrt{\log d \log(1+n)}}}{n}\right)^2} \\
&\leq 2L^2\frac{M_n^2 + \|a\|_\infty + 1}{n^2\lambda_0^2 T^2}\frac{e^{-2\sqrt{\log d \log(1+n)}}}{\left(1 - 4L\frac{M_n+1}{n\lambda_0 T}e^{-2\sqrt{\log d \log(1+n)}}\right)^2}.
\end{aligned}$$

It holds, for $n \geq 2$,

$$
\begin{aligned}
L\frac{M_n+1}{n\lambda_0 T}e^{-2\sqrt{\log d\log(1+n)}} &\leq L\frac{M_n+1}{n\lambda_0} \\
&= L\frac{M_n+1}{nT.\frac{3LM_n\left(5\sqrt{3\log(2d)}\log n+4\right)}{\sqrt{n}}} \\
&= \frac{1+\frac{1}{M_n}}{3\sqrt{n}T\left(5\sqrt{3\log(2d)}\log n+4\right)} \\
&\leq \frac{1}{3}\frac{1+\frac{1}{\sqrt{2\log 3}}}{\sqrt{2}\left(5\sqrt{3\log 2}.\log 2+4\right)} \\
&< \frac{1}{8}.
\end{aligned}
$$

Finally, we conclude that

$$
P\left(\frac{1}{n}\sum_{i=1}^{n}F\left(X^{(i)}\right)\geq\frac{\lambda_0 T}{L}\right)<4L^2\frac{M_n^2+\|a\|_\infty+1}{n^2\lambda_0^2T^2}.
$$

□

**Lemma 3.29.** *Recall from Lemma 3.25 that* $V_n^{trunc}(\beta)=(P_n-P)(\rho_\beta I_{\{G\leq M_n\}})$ *and* $V_n(\beta)=(P_n-P)\rho_\beta$. *Recall also from Lemma 3.28 that* $F(X)=G(X)I_{\{G(X)>M_n\}}+\mathbb{E}\left[G(X)I_{\{G(X)>M_n\}}\right]$ *with* $G(X)=\|X\|_\infty$. *It holds true that* $\forall T\geq 1$,

$$
P\left(\sup_{\beta\in B_2(0,R)}\frac{|V_n^{trunc}(\beta)-V_n^{trunc}(\beta_0)-(V_n(\beta)-V_n(\beta_0))|}{\|\beta-\beta_0\|_1\vee\lambda_0}>T\lambda_0\right)\leq P\left(\frac{1}{n}\sum_{i=1}^{n}F\left(X^{(i)}\right)>\frac{T\lambda_0}{L}\right).
$$

*Proof.* Basic computations and Hölder's inequality give

$$\begin{aligned}
&\left|V_n^{trunc}(\beta) - V_n^{trunc}(\beta_0) - (V_n(\beta) - V_n(\beta_0))\right| \\
&= \left|(P_n - P)\left(\rho_\beta I_{\{G>M_n\}}\right) - (P_n - P)\left(\rho_{\beta_0} I_{\{G>M_n\}}\right)\right| \\
&\leq \left|P_n\left[(\rho_\beta - \rho_{\beta_0}) I_{\{G>M_n\}}\right]\right| + \left|P\left[(\rho_\beta(X) - \rho_{\beta_0}(X)) I_{\{G(X)>M_n\}}\right]\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} L\left|X^{(i)}(\beta - \beta_0)\right| I_{\left\{G\left(X^{(i)}\right)>M_n\right\}} + \mathbb{E}\left[L\left|X(\beta - \beta_0)\right| I_{\{G(X)>M_n\}}\right] \\
&\leq L\left(\frac{1}{n}\sum_{i=1}^{n}\left\|X^{(i)}\right\|_\infty I_{\left\{G\left(X^{(i)}\right)>M_n\right\}}\|\beta - \beta_0\|_1 + \mathbb{E}\left[\|X\|_\infty I_{\{G(X)>M_n\}}\right]\|\beta - \beta_0\|_1\right) \\
&\leq \frac{L\|\beta - \beta_0\|_1}{n}\sum_{i=1}^{n} F\left(X^{(i)}\right),
\end{aligned}$$

and the result directly follows. □

# Chapter 4

# Real-time graph clustering for network intrusion detection

**Abstract.** We propose an online graph clustering approach for detecting suspicious activities in a computing system based on the detection of communities in a network. Assuming that attacks dynamically create abnormal connections between system processes within the network, the detection of the communities will detect intrusions. We apply this approach on a publicly available dataset OpTC provided by the Defense Advanced Research Projects Agency (DARPA). Our preliminary results show the feasibility of our approach and encourages the exploration towards an intelligent attack detection end-to-end system.

## 4.1 Introduction

Anomaly detection is the identification of rare events or failures in a system. This is highly challenging especially when processing a large amount of data. According to the survey [1], various domains are considered such as health insurance [82], security [43], etc. More particularly, one can mention intrusion detection as a sub-domain of anomaly

detection applied to cybersecurity. The typology of data is numerous: DNS logs[1], authentication logs [80], system processes logs [142], network traffic logs [132], etc. The volume of data ranges from 1 Mo to several To for only a tiny fraction of anomalies [5].

Due to the network structure of the data, a graph modelling is well suited to detect anomalies. Indeed anomalies can be defined as abnormal connections between normal and suspicious objects. The graph itself carries lot of correlation information from these connections, see [1]. For instance, when considering a graph of system processes, which is the purpose of our work, the nodes of the graph are processes and the edges are the actions of processes to another one (create, open or terminate).

Community detection is one of the methods used in graph machine learning. It consists of partitioning the graph into a set of nodes that share a large number of connections between them. Rossetti [127] proposes a survey on dynamic community detection algorithms including modularity-based algorithms, see [62, 11, 131, 61, 114, 7]. Other studies focused on the analysis of large social networks or research article citations, for instance [33, 117], and more recently the analysis of emails [48]. More methods and codes are publicly available[2]. Here we propose to detect intrusions in a computing system using a dynamic MCMC algorithm introduced in [39] which is a dynamic modified version of the Louvain partitioning procedure from [17]. This is the first use of this algorithm on real data.

The chapter is organised as follows. Section 4.2 presents the motivations of our work (i.e. applying a graph clustering algorithm to system processes). Section 4.3 introduces the mathematical concepts behind the online clustering-based community detection. Section 4.4 describes the Operationally Transparent Cyber (OpTC) dataset used in our experiments. Section 4.5 is twofold. Firstly, we describe how to build a proper graph from the OpTC data. Secondly, the results of the proposed methodological approach are detailed to detect suspicious events.

## 4.2 Motivations and goals

As presented in [1, 88], detecting intrusion in a computing system should take advantage of the network structure of the data. The definition of a graph is straightforward: the nodes are the components of the system (servers, files,

[1] as provided by https://www.shodan.io/
[2] https://github.com/1172939260/community-detection

processes, etc.) and the edges are the connections between the components.

Our goal is to analyse processes in a system. In this case, the graph is defined as the actions from processes to another one: *create* or *terminate* when a process creates or terminates another process, *open* when a process accesses the memory space of another process. In [88], the authors claim that a non-supervised learning approach should be preferred to a supervised one since it does not require a labelled dataset. Similarly, we propose a clustering approach to find groups of nodes that are strongly connected together. This is commonly known as "graph clustering" or "community detection". This is motivated by the intuision that an intrusion changes the structure of the graph by creating new clusters (or communities in the sequel). In particular, processes due to an abnormal activity, create or terminate "normal" processes. Consequently, new communities are created due to new connections between nodes. On the other hand, we take into consideration that the graph changes in time, i.e. new nodes and edges are dynamically collected.

We propose to use the algorithm [39] for this purpose which is a real-time MCMC Louvain algorithm in order to dynamically detect communities within a graph representation of a computer network.

## 4.3 Online graph community detection

**Graph and community.** Let $G = (V, E, A)$ be a weighted undirected graph where $V$ is a set of $n$ nodes, $E$ is a set of edges corresponding to the relationship between the nodes and where $A \in \mathcal{M}_n(\mathbb{R})$ is the corresponding symmetric adjacency matrix such that $A_{ij}$ denotes the weight assigned to edge $(i, j) \in E$, if one refers to an edge $e_l = (i, j)$ then its weight would be also noted $A_{e_l}$. The degree $d_i$ of a node $i \in V$ is the sum of the weights of the edges incident to $i$.

There is no absolute definition of a community in a network. We choose to define a community as a set of nodes that are strongly connected together. Consequently, detecting communities in a graph is finding a partition of its nodes. One way to assess the quality of a partition and then to detect communities is to use the modularity function (see for instance [17, 69, 43]). Broadly speaking, for one partition of the graph into communities, the modularity measures the quality of the groups formed. Namely in 2016, Newman [110] found that the maximisation of the

modularity is equivalent to maximising the likelihood of a Corrected Degree Stochastic Block Model (CDSBM) [78]. This means that the partition obtained by maximisation of the modularity should be the partition that should best correspond to the communities of that particular model. Unfortunately such a partition always exists even if the model that generated the graph is not the CDSBM. If the graph has nothing to do with a CDSBM, then the interpretation of the partition obtained by maximisation of the modularity is hazardous.

More formally, let $C = \{c_1, \ldots, c_K\} \in \mathcal{C}$ be a partition of nodes (i.e. a set of communities) and $\mathcal{C}$ the set of all possible partitions. The modularity function $Q : \mathcal{C} \to [-1/2, 1]$ is defined as:

$$Q(C) = \frac{1}{2m} \sum_{k=1}^{K} \sum_{i,j \in c_k} \left( A_{ij} - \gamma \frac{d_i d_j}{2m} \right),$$

where $m := |E| = \frac{1}{2} \sum_i d_i$. We take $\gamma = 1$ as default value.

**Community detection.** The problem of detecting communities can be viewed as finding the partition of nodes $C^\star$ that maximises the modularity function:

$$C^\star \in \underset{C \in \mathcal{C}}{\operatorname{argmax}} Q(C). \tag{4.3.1}$$

This optimisation problem, also known as graph clustering, is NP-hard due to the combinatoric issues related to the set $\mathcal{C}$ [21]. For this reason, heuristic approximations was proposed such as the well-known Louvain algorithm [17]. This greedy search algorithm iterates two steps until a convergence is achieved. The optimisation step aims at maximising the modularity function by locally moving each nodes to one of its neighbours' clusters, see Algorithm 4.1. New communities are built by gathering the closest nodes together only if the modularity increases. Secondly, the aggregation step merges each cluster to one node and builds a new graph whose nodes are the communities themselves. Then, the algorithm is applied to this new graph and builds again a new graph whose nodes are communities of communities and so on.

LumenAI has implemented its own community detection algorithm. They implemented the algorithm described in [39] because it provides online community detection. We call it GraphClus. GrphClus receives the edges as they are sent and build the graph edge by edge the graph. Parallelly with the reception of edges, GrphClus optimises

**Algorithm 4.1** LOUVAIN ALGORITHM

Let $G_0$ be the initial graph.
Initialize the communities $c_i = \{i\}$, for all $i \in V$ (all node in there own community), set $G_{optim} := G$ the graph to work with and compute the initial value of the modularity $Q$.
While the number of communities in $G_{optim}$ is not 1:

- Phase 1: (Optimisation) While the modularity $Q$ increases:
  - for each node $i \in V$ of $G_{optim}$:
    - for each node $j$ adjacent to $i$:
      - Compute what would be the modularity change if the node $i$ was moved to the community of the node $j$
    - Assign the node $i$ to the community with the highest modularity increase
- Phase 2: (agregation) Create a new graph $G_{new}$ where all nodes in each community of $G_{optim}$ are merged together. Consequently, edges $\{e_1, e_2, ..., e_l\}$ within a community $C_i$ become a self-loop with a weight $\sum_{i=1}^{l} A_{e_i}$ and edges between communities become edges between nodes with the same weight. Now set $G_{optim} := G_{new}$.

Keep track of this hierarchical structure and return it

with an MCMC method[3] the partition[4] $C$ of the graph. The algorithm changes the partition as follows: GrphClus picks at random a candidate partition $C'$ to replace $C$ according to a mixture distribution $\alpha_1 p_1(\cdot|C) + \alpha_2 p_2(\cdot|C)$ where $\alpha_1$ and $\alpha_2$ are hyperparametres[5]. $p_1(\cdot|C)$ and $p_2(\cdot|C)$ are such that :

- under $p_1(\cdot|C)$ : pick uniformly at random $i$ from $V$ and a community $\boldsymbol{c}$ among those of $C$. $C'$ is then taken equal to $C$ except for the node $i$ that joins $\boldsymbol{c}$ if it was not part of it, otherwise, $i$ is placed in its own new community.
- under $p_2(\cdot|C)$ : pick uniformly at random $i$ from $V$, then pick a node $j$ randomly among the neighbours of $i$ according to probabilities proportional to $A_{ij}$. $C'$ is then taken equal to $C$ except for the node $i$ that joins $\boldsymbol{c}$ if it was not part of it, otherwise, $i$ is placed in its own new community.

The candidate partition $C'$ is accepted (or rejected) with probability $\rho$ (or $1-\rho$). The acceptation rate $\rho = \min\{1, rf(\lambda)\}$ is computed from the likelihood ratio[6] $r = \frac{p(C|C')}{p(C'|C)}$ representing the reversibility of the transition $C'$ to $C$, and from the quality factor $f(\lambda) = e^{\lambda(Q'-Q)}$ coming from the gain in modularity due to $C'$. $\lambda > 0$ is an

[3] Markov Chain Monte Carlo
[4] In fact GrphClus computes a hierarchical partition as described in [39] but for this work we only used the first level of the partition which is equivalent to a usual community detection algorithm.
[5] In practice, we used $\alpha_1 = 0.05$.
[6] This likelihood ratio measures the reversibility of the Markov chain with respect to the switch from $C$ to $C'$. For more details about the expression of $r$, refer to [39].

hyperparameter that influences the stability and the convergence rate of the MCMC part (cf. algorithm 4.2). The procedure we have just described can be generalized to hierarchical partitions with more complicated notations and a gestion of the updates of the partitions through all concerned levels in the hierarchy (see [39] for those further details). As mentioned in Section 3 in [39], the assumption $\alpha_1 > 0$ is theoretically important because it makes the Markov chain reversible and reversibility insures the existence of a stationary distribution of the Markov as stated by the Perron-Frobenius theorem[7]. The existence of such a stationary distribution motivates the use of a Metropolis-Hasting approach to approximate it. It also motivates the expression of the acceptation ratio because it is proved that it conserves the stationary distribution of the MCMC.

**Algorithm 4.2** Simplified algorithm "GrphClus" to a one-level partition

Hyperparamètres : $\lambda > 0$ (called "temperature" parameter), $\alpha_1 > 0$ and $\alpha_2 > 0$ such that $\alpha_1 + \alpha_2 = 1$.
From an empty graph $G = \{V = \emptyset, E = \emptyset, A = \emptyset\}$. Do in parallel 1 and 2 :

1. (Passive reception) build the graph as edges come :
   (a) when an edge $e = (i, j)$ is received, add the nodes $i$ or $j$ to $V$ and the edge $e$ to $E$ id these elements are seen for the first time. Increment $A_{ij}$ by 1.
2. (MCMC) evolution of the partition :
   (a) Pick a candidate partition $C'$ at random according to the mixture distribution $\alpha_1 p_1(\cdot|C) + \alpha_2 p_2(\cdot|C)$ :
      i. under $p_1(\cdot|C)$ : pick uniformly at random $i$ from $V$ and a community $\boldsymbol{c}$ among those of $C$.
      ii. under $p_2(\cdot|C)$ : pick uniformly at random $i$ from $V$, then pick a node $j$ randomly among the neighbours of $i$ according to probabilities proportional to $A_{ij}$. Note $c$ the community of $j$.
      iii. $C'$ is then taken equal to $C$ except for the node $i$ that joins $\boldsymbol{c}$ if it was not part of it, otherwise, $i$ is placed in its own new community.
   (b) Election of $C'$. Compute:
      i. the modularity $Q'$ associated to $C'$
      ii. the quality factor $f(\lambda) = e^{\lambda(Q'-Q)}$, where $Q$ is the modularity of the current partition $C$
      iii. $r = \frac{p(C|C')}{p(C'|C)}$ the ratio between the likelihoods of transition between $C'$ and $C$ (cf [39] for the exact value of $r$)
      iv. Accept $C'$ with probability $\rho = \min\{1, rf(\lambda)\}$

*Remark.* We used a slightly different implementation of the procedure described in [39]. On the one hand the article [39] mentions conditions on the communities of $i$ and $j$ that we ignored, and on the other hand the article

[7] The weakest necessary assumption to insure reversibility is the strong connexity of the graph formed by the states of the Markov chain. The transitions between the states are however difficult to study. Hence the use of $\alpha_1$ to bring reversibility in the MCMC.

explains that $j$ is chosen proportionally to the quantity $k_{i,c}^C$ (defined after) but we chose to pick it proportionally to $A_{ij}$ in order not to bias the research of a local maximum in favour of big communities. To clarify the two points of view, note that the use of the quantities $A_{ij}$ or $k_{i,c}^C$ can be equavalent. Indeed, one can either choose a node $j$ among the neighbours of $i$ according to a distribution proportional to the $A_{ij}$'s or equivalently choose a community $c_{choice}$ among all neighbour communities of $i$ with probability proportional to $k_{i,c_{choice}}^C := \sum_{j \in V} A_{ij} \delta_{C(j)=c_{choix}}$ where $C(j)$ is the community of $j$ and $\delta_{C(j)=c_{choix}} = 1$ if and only if the community of $j$ is $c_{choice}$. In both cases the objective is the same: compute the gain in modularity if $i$ joins the community of $j$ or if $i$ joins the community $c_{choix}$. It was maybe what the author of [39] meant but throroughly speaking, what is actually described in [39] is the following : imagine you have a graph, under $p_2(\cdot|C)$, if $i$ has 1 neighbour $j_1$ in community $C_1$, 10 neighbours $k_1, ..., k_{10}$ in community $C_2$ and 100 neighbours $l_1, ..., l_{100}$ in community $C_3$ then one should choose $j_1$ with probability proportional to 1, choose each $k_1, ..., k_{10}$ with propability proportional to 10 and choose each $l_1, ..., l_{100}$ with propability proportional to 100 leading to the probability triplet $\left(\frac{1*1}{10101}, \frac{10*10}{10101}, \frac{100*100}{10101}\right)$ to choose among triplet $(C_1, C_2, C_3)$. with our approach, the probability triplet is $\left(\frac{1}{111}, \frac{10}{111}, \frac{100}{111}\right)$ reflecting therefore the strength of the link between $i$ and the communities around without any particular bias in favour of $C_3$.

## 4.4 Operationally Transparent Cyber dataset

As part of the OpTC research program, DARPA has released a dataset over the past few years [142]. The objective was to provide realistic basis for cyber defence studies. The last update of the data repository proposes a new event based model, called extended Cyber Analytics Repository (eCAR). A pool of 1000 hosts with OS Windows 10 has been monitored for 6 days. During 3 of them, attacks occured with 3 different scenarios[8].

**Day 1:** To get elevated privileges and lateral movements between target hosts using Powershell Empire. To get credentials using Mimikatz. 18 machines were compromised.

**Day 2:** To extract personal data by sending phishing emails containing a malware. Once checked in, the administrator's credentials enable data exfiltration via Netcat and Remote Desktop Protocol.

[8] https://github.com/FiveDirections/OpTC-data

**Day 3:** To access personal data including credentials. A malicious binary file is downloaded during an update of Notepad Plus. After connecting to the malicious server during the update, the host opens a breach allowing the attacker to access the system's personal data, including credentials.

Each host of the experimental setup was equipped with a sensor that retrieves a multitude of events before putting them in eCAR format. The malicious actions executed by DARPA on this dataset clearly highlight the behaviour of an Advanced Persistent Threat (APT). Thus, the diversity of benign and malicious interactions makes this dataset very interesting and potentially useful for the detection of anomalies in computer networks.

The OpTC dataset contains 17 billions of events of network communications and system logs, including 292387 malicious ones. Each event describes an interaction between a process and an object with specific fields depending on the nature of the object. Therefore we find elements such as object type, action, timestamp or command lines at the origin of an event. Table 4.4.1, describes the main objects found in the dataset. Each type of object is associated with specific actions that a process can perform on it. It is important to note that the proportion of malicious events is 0.000016. The distribution of malicious and benign events is highly imbalanced and makes predictions about the minority class difficult. But this situation puts us in a realistic context in order to set up the most relevant metrics that will allow us to raise an alert in case of malicious intrusion. Also, this motivates the use of a unsupervised approach instead of a supervised prediction approach.

In section 4.5, we chose to focus on the processes of host 201 on day 1, because the attack scheme is particularly apropriated to our methodology. In this particular case, there are 86346 events concerning object "PROCESS". Among them, 73 events concern the pid 5452, and 606, the pid 2952, the two pid pointed out to be malicious action in the ground truth. That is to say 0.78% of the events are potentially malicious. In fact, the ground truth lacks a bit of precision about milicous pid because pid are not uniquely attributed to the processes. They are unique at every moment but not through time, meaning that a pid can be attributed several time on day 1 either as benign action or as malicious ones. Our investigation showed that the actions done under pid 2952, indicated as malicious, are benign ones before timestamp 11h23[9] and that they becomes malign after that. PROCESS object can correspond to 3 different actions: create, open and terminate. Our understanding of how Windows' processes[10]

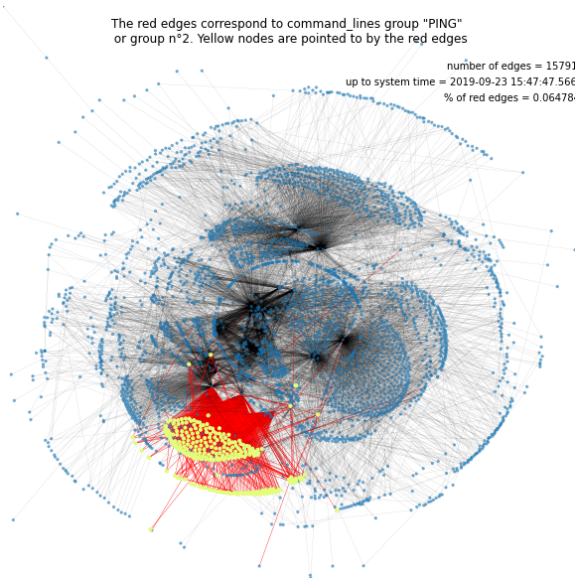
[9] and some seconds
[10] The processes in the OS Windows

work is limited, but here is a general meaning:

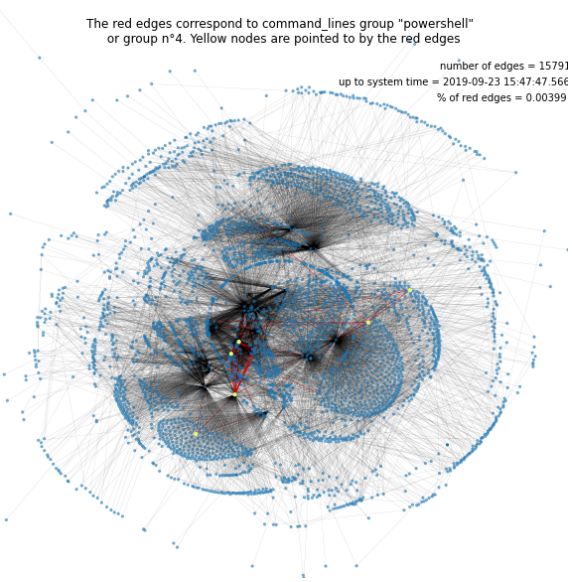
- CREATE : a process can be created be an other one.
- OPEN : a process can read the properties of an other one.
- TERMINATE : a process can terminate an other one. In practice, we observed that a PROCESS terminates itself. In other words, when an edge in the graph implies an action "terminate" a loop from the PROCESS to itself is created.

Interestingly, among those 86346 events, almost all of them come up with a command line and there are only 559 unique of them.

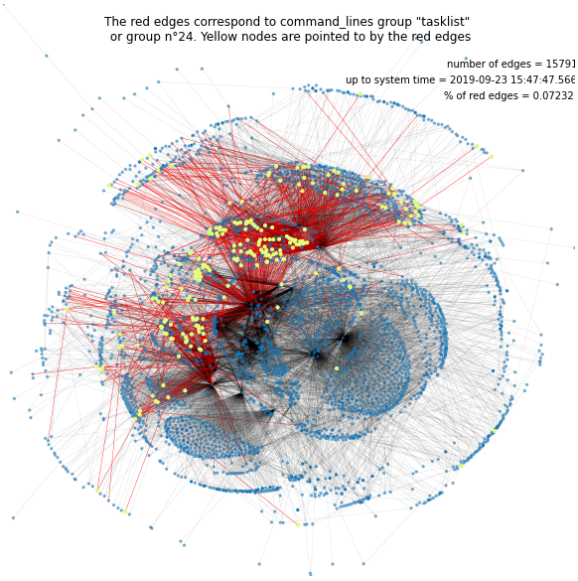
This relatively few number of command line could be manually analysed and categorized to improve the detection of attacks. We tried namely to proceed to such a manual clustering to see whether it could reveal a structure in the graph and potentially provide a meaning to the partition returned by GrphClus. One can see on figure 4.4.1 four edge clusters formed by the edges of the graph born from events whose associated command line contained a particular substring. The cluster on figure 4.4.1.a are the edges associated to command lines containing the substring "PING". It can make think of a very easy structure of the graph where all the edges associated to particular command lines would be gathered but as the three other images show, this is not the case. Indeed, substring "PING" is associated to malign actions while substring "ping" is associated to benign ones. It shows consequently that command lines as close as these ones can lie on very different location in the graph and can be gathered as well as spread out. The attacker used "powershell" technics to take control of machine 201. We show on figure 4.4.1.b the corresponding cluster. It concerns around 0.4% of edges and 20 nodes spread out in the graph with no apparent strucutre. It is then unlikely that the partition of GrphClus could reveal these actions. The figure 4.4.1.c shows how one of the biggest cluster (containing 7% of all edges) can be spread out. A priori the structure of the graph is not related to the command line. We will see how Grphclus could though detect the attack thanks to online analyses in Section 4.5.



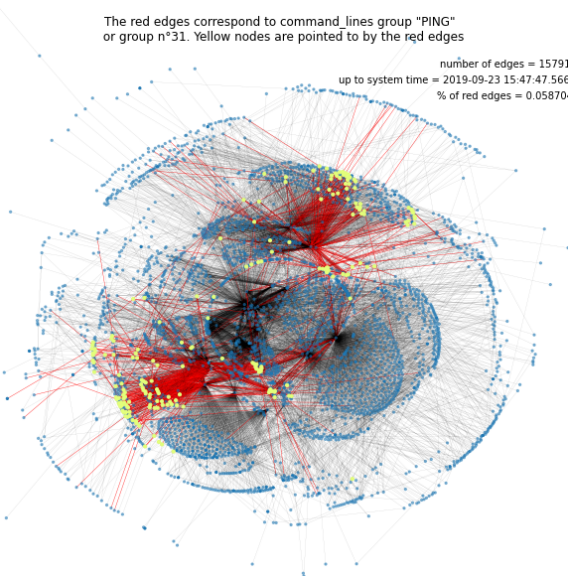

((a)) PING




((b)) powershell




((c)) tasklist




((d)) ping

Figure 4.4.1: Some example of manual clusters of edges in the graph of process of machine 201 on day 1. The name of each image is the substring contained is all the command line of the edges the cluster. All red edges form a cluster associated to a particular substring of the command line of their event. Yellow nodes are the end of red edge associated to the ObjectID while the other end is associated to the ActorID and is not highlighted (see Section 4.5 for the meaning of ObjectID and ActorID). The meta information on images are related to the dynamic evolution of the graph available in video at this link: https://www.youtube.com/watch?v=LPBuE0kBIr4

| Objects | Actions | Total % |
|---|---|---|
| FLOW | message, open, start | 71.7 |
| FILE | create, delete, modify, read, rename, write | 12.4 |
| PROCESS | create, open, terminate | 8.6 |
| MODULE | load | 3.9 |
| REGISTRY | add, edit, remove | 0.3 |
| HOST | start | 0.0 |

Table 4.4.1: Objects found in the eCAR dataset

**Algorithm 4.3** Creation of the graph in real-time

Start with a graph $G = (V, E)$ where $V = \emptyset$ and $E = \emptyset$
While one does not stop the graph-creation:

- When an edge $e = (i, j)$ comes in:
  - if $e \in E$ then $A_{ij} \leftarrow A_{ij} + 1$
  - otherwise add $i$ or $j$ to $V$ if necessary and $A_{ij} := 1$

## 4.5 Cybersecurity intrusion detection

### 4.5.1 Definition of a graph of Processes.

The notion of a process in an operating system is fundamental. Each event that occurs implies an interaction with a process that could be the cause of this event. Since operating systems are naturally process-oriented, representing the OpTC dataset as a graph of processes is an intuitive entry point. Indeed, we use the attributes "actorID" and "objectID" to build a new edge between nodes. If an edge with unknown ends come up, new nodes are created to make the edge well defined. If an edge appears for the first time, it is created with weight 1.0 otherwise the weight is incremented by 1, see algorithme 4.3. An edge represents the link between a parent process and a child process, as shown in figure 4.5.1. Given this definition, the graph represents the historical activity of the network at a time $t$.

### 4.5.2 Graph clustering and interpretation.

Our procedure continuously updates communities as explained in section 4.2 with $\lambda = 10^{-5}$ and $\alpha_1 = 0.05$. Figure 4.5.2 represents the number of the communities at each edge income (top) and the two histogram show

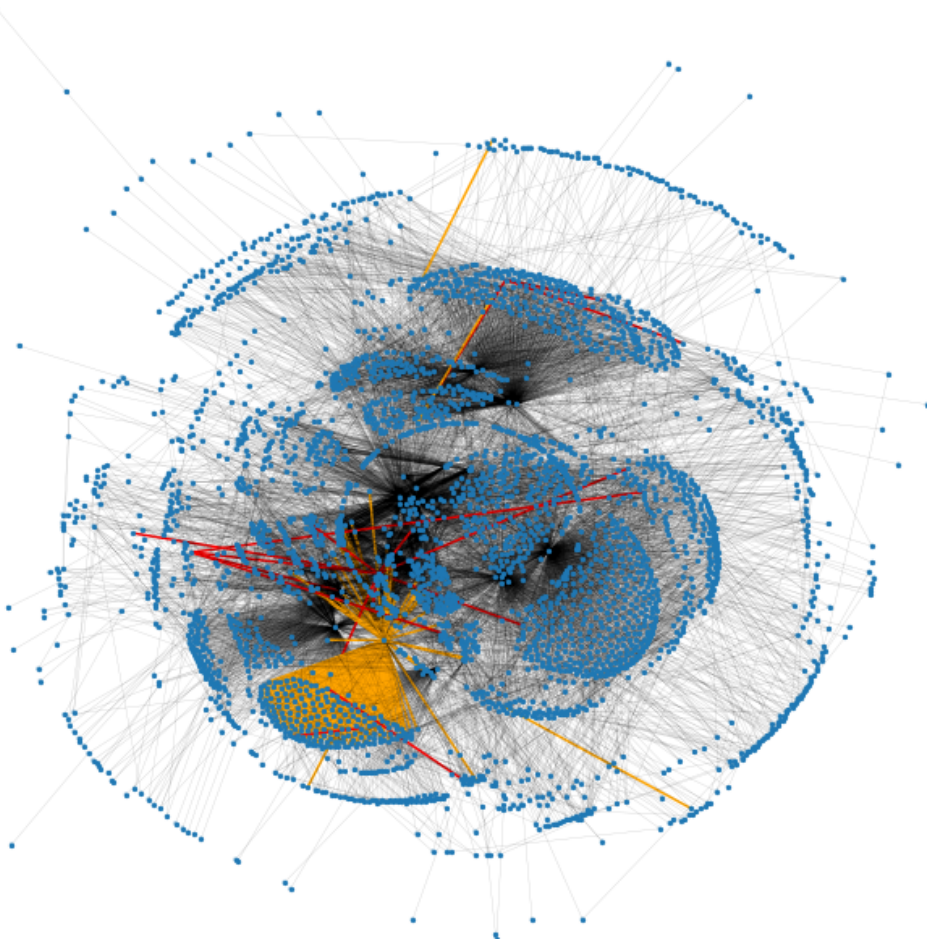


Figure 4.5.1: Kamada Kawai plot of the graph of processes (machine 201, day 1), processes with PID 2952 in orange and 5452 in red.

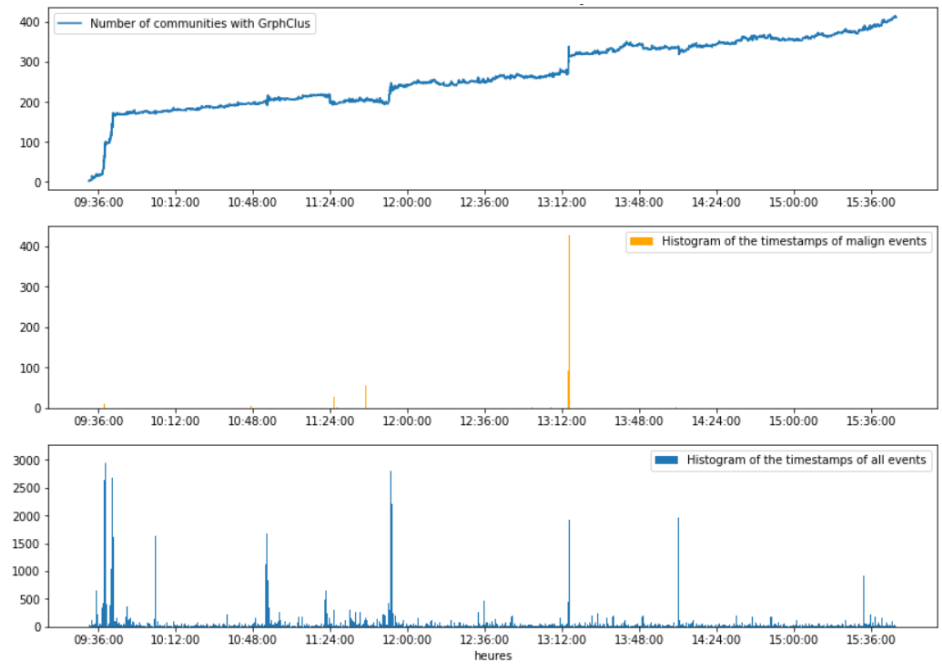



Figure 4.5.2: Number of communities (top), malign activity (middle) and all processes activity (bottom) for machine 201 on day 1.

respectively the malign[11] and benign activity of the system 201. Therefore we monitor the number of community and we observe an effect consistently with our hypothesis: the actions of the attacker distrubs the current partition so that it generates a discontinuity in the number of communities. See on Figure 4.5.2 how the histogram of malign events coincide with the discontinuity at 1:15pm on day 1.

As shown in Figure 4.5.2, the number of communities has a regular behaviour and few discontinuities. The most frank discontinuities occur around 11:50am and 1:15pm. The first big one after start, at 11:50am, is due to a normal but intensive activity of the system that change a lot the graph size and topology. You can see the impact of malign activity on the graph plot on Figure 4.5.1 and dynamically on a video in supplementary material[12]. The second big discontinuity, at 1:15pm, coincides with a high abnormal activity. The orange edges in the Figure 4.5.1 is the subgraph corresponding to PID 2952 and it coincide with the cluster named "PING" on Figure 4.4.1.a. According to our investigation, attacker scanned all the IP addresses on a sub-network leading to more or less 256 pings. Some of these pings should have given birth to their own community while others should have been attributed to

[11] malign activities are considered to be the one with pid 2952 and 5452
[12] The video is available here: https://www.youtube.com/watch?v=LPBuE0kBIr4 .

an existing one. That is why the number of community metric shows a sudden increase at the time of the attack.

On day 1, host 402 and 660 were also attacked after host 201 but the actions of the attacker were not revealed by the monitoring of the number of community. The malign actions seem to give birth of edge in already very dense region of the graph. It could be the cause of the absent reaction from GrphClus. No video are available for the hosts 402 and 660 on day 1 but we have a comparable example of that on host 51 on day 3[13].

## 4.6 Conclusion

We presented a graph based approach for the real-time detection of intrusions in a computer network. According to our observations, the number of communities seems to be a useful metric to detect some intrusion through the graph of processes. Indeed, attacks like the one done on host 201 created new processes from existing ones. It had an effect on the graph topology by creating new communities. This is in line with our choice to use the number of communities for detecting malicious actions. As a future work, we should explore additional monitoring metrics to refine the analysis, for instance the leaders of the communities. A more extensive numerical analysis will be performed to evaluate more precisely the performances of such clustering-based algorithm.

[13] You can see on a video available at https://www.youtube.com/watch?v=ZAzZly4ZZt8 that malign actions are hidden in dense regions of the graph of processes on machine 51 on day 3.

# Bibliography


[1] L. Akoglu, H. Tong, and D. Koutra. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29:626â688, 2015.

[2] M. Al Hasan, V. Chaoji, S. Salem, and M. J. Zaki. Robust partitional clustering by outlier and density insensitive seeding. *Pattern Recognition Letters*, 30(11):994–1002, 2009.

[3] E. Aldana-Bobadilla and A. Kuri-Morales. A clustering method based on the maximum entropy principle. *Entropy*, 17(1):151–180, 2015.

[4] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. volume 58, pages 137–147. 1999. Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996).

[5] Md M. Anjum, S. Iqbal, and B. Hamelin. Analyzing the usefulness of the darpa optc dataset in cyber threat detection research. In *Proc of the 26th ACM Symposium on Access Control Models and Technologies*, pages 27–32, 2021.

[6] D. Arthur and S. Vassilvitskii. K-means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, 2007.

[7] T. Aynaud and J.-L. Guillaume. Multi-step community detection and hierarchical time segmentation in evolving networks. In *Proceedings of the 5th SNA-KDD workshop*, volume 11, 2011.

[8] M. Azizyan, A. Singh, and L. Wasserman. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. 2013.

[9] M. Azizyan, A. Singh, and L. Wasserman. Efficient sparse clustering of high-dimensional non-spherical gaussian mixtures. In *Artificial Intelligence and Statistics*, pages 37–45. PMLR, 2015.

[10] F. Bach and Z. Harchaoui. Diffrac: a discriminative and flexible framework for clustering. *Advances in Neural information processing systems*, 20:49–56, 2007.

[11] S. Bansal, S. Bhowmick, and P. Paymal. Fast community detection for dynamic complex networks. In *Complex networks*, pages 196–207. Springer, 2011.

[12] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Stat. Comput.*, 22(2):455–470, 2012.

[13] B. Bercu, B. Delyon, and E. Rio. *Concentration inequalities for sums and martingales*. SpringerBriefs in Mathematics. Springer, Cham, 2015.

[14] A. Bérut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, and E. Lutz. Experimental verification of landauerâs principle linking information and thermodynamics. *Nature*, 483(7388):187–189, 2012.

[15] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Trans. Inform. Theory*, 54(2):781–790, 2008.

[16] Z. W. Birnbaum. An inequality for mill's ratio. *The Annals of Mathematical Statistics*, 13(2):245–246, 1942.

[17] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[18] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

[19] C. Bouveyron and C. Brunet-Saumard. Discriminative variable selection for clustering with the sparse fisher-em algorithm. *Computational Statistics*, 29(3):489–513, 2014.

[20] C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery. *Model-based clustering and classification for data science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2019. With applications in R.

[21] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188, 2008.

[22] C. Brécheteau. Robust shape inference from a sparse approximation of the Gaussian trimmed loglikelihood. preprint, 2018.

[23] C. Brecheteau. https://www.math.sciences.univ-nantes.fr/ brecheteau/notebooks/ Notebook_kPDTM_kPLM.html. 2020.

[24] M. Brennan and G. Bresler. Average-case lower bounds for learning sparse mixtures, robust estimation and semirandom adversaries. *arXiv preprint arXiv:1908.06130*, 2019.

[25] C. Brunet-Saumard and E. Genetay. https://github.com/csaumard/kbmom. 2021.

[26] C. Brunet-Saumard, E. Genetay, and A. Saumard. K-bMOM: a robust Lloyd-type clustering algorithm based on bootstrap Median-of-Means. 2020.

[27] C. Brunet-Saumard, E. Genetay, and A. Saumard. Supplement to: "K-bMOM: a robust Lloyd-type clustering algorithm based on bootstrap Median-of-Means". 2020.

[28] P. Bühlmann. Bagging, subagging and bragging for improving some prediction algorithms. In *Recent advances and trends in nonparametric statistics*, pages 19–34. Elsevier B. V., Amsterdam, 2003.

[29] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[30] F. Bunea, C. Giraud, X. Luo, M. Royer, and N. Verzelen. Model assisted variable clustering: minimax-optimal recovery and algorithms. *The Annals of Statistics*, 48(1):111–137, 2020.

[31] F. Bunea, C. Giraud, M. Royer, and N. Verzelen. Pecok: a convex optimization approach to variable clustering. *arXiv preprint arXiv:1606.05100*, 2016.

[32] T. T. Cai, J. Ma, and L. Zhang. Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267, 2019.

[33] B. P. Chamberlain, J. Levy-Kramer, C. Humby, and M. P. Deisenroth. Real-time community detection in full social networks on a laptop. *PLoS One*, 13(1):1–37, 2018.

[34] X. Chen and Y. Yang. Cutoff for exact recovery of gaussian mixture models. *IEEE Transactions on Information Theory*, 67(6):4223–4238, 2021.

[35] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[36] J. A. Cuesta-Albertos, A. Gordaliza, and C. Matrán. Trimmed $k$-means: An attempt to robustify quantizers. *The Annals of Statistics*, 25(2):553–576, 1997.

[37] Z. Dafir, Y. Lamari, and S. C. Slaoui. A survey on parallel clustering algorithms for big data. *Artificial Intelligence Review*, 54(4):2411–2443, 2021.

[38] Bo Dai and Baogang Hu. Minimum conditional entropy clustering: A discriminative framework for clustering. In *Proceedings of 2nd Asian Conference on Machine Learning*, pages 47–62. JMLR Workshop and Conference Proceedings, 2010.

[39] Y. Darmaillac and S. Loustau. MCMC louvain for online community detection. *CoRR*, abs/1612.01489, 2016.

[40] E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, and A. Mayo-Íscar. Robust clustering tools based on optimal transportation. *Stat. Comput.*, 29(1):139–160, 2019.

[41] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-Gaussian mean estimators. *Ann. Statist.*, 44(6):2695–2725, 2016.

[42] I. Diakonikolas and D. M. Kane. Recent Advances in Algorithmic High-Dimensional Robust Statistics, 2019.

[43] Q. Ding, N. Katenka, P. Barford, E. Kolaczyk, and M. Crovella. Intrusion as (anti) social communication: characterization and detection. In *Proc. of the 18th SIGKDD*, pages 886–894, 2012.

[44] S. Dolnicar and F. Leisch. Winter tourist segments in Austria: Identifying stable vacation styles using bagged clustering techniques. *Journal of Travel Research*, 41(3):281–292, 2003.

[45] P. D'Urso, L. De Giovanni, M. Disegna, and R. Massari. Bagged clustering and its application to tourism market segmentation. *Expert Systems with Applications*, 40(12):4944–4956, 2013.

[46] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

[47] J. Fan, H. Liu, Z. Wang, and Z. Yang. Curse of heterogeneity: Computational barriers in sparse mixture models and phase retrieval. *arXiv preprint arXiv:1808.06996*, 2018.

[48] G. Fang, O. Ward, and T. Zheng. Online community detection for event streams on networks. *CoRR*, abs/2009.01742, 2020.

[49] C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *J. Amer. Math. Soc.*, 29(4):983–1049, 2016.

[50] A. Fischer. Quantization and clustering with Bregman divergences. *J. Multivariate Anal.*, 101(9):2207–2221, 2010.

[51] N. Flammarion, B. Palaniappan, and F. Bach. Robust discriminative clustering with sparse regularizers. *The Journal of Machine Learning Research*, 18(1):2764–2813, 2017.

[52] L. Á. García-Escudero and A. Gordaliza. Robustness properties of $k$ means and trimmed $k$ means. *J. Amer. Statist. Assoc.*, 94(447):956–969, 1999.

[53] L. A. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 36(3):1324–1345, 2008.

[54] L. A. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. A review of robust clustering methods. *Adv. Data Anal. Classif.*, 4(2-3):89–109, 2010.

[55] A. Gasull and F. Utzet. Approximating mills ratio. *Journal of Mathematical Analysis and Applications*, 420(2):1832–1853, 2014.

[56] E. Genetay, A. Saumard, and R. Coulaud. High-dimensional logistic entropy clustering. *arXiv preprint arXiv:2112.08701*, 2021.

[57] C. Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021.

[58] C. Giraud and N. Verzelen. Partial recovery bounds for clustering with the relaxed $k$-means. *Mathematical Statistics and Learning*, 1(3):317–374, 2019.

[59] S. Golovkine. Statistical methods for multivariate functional data. 2021.

[60] R. Gomes, A. Krause, and P. Perona. Discriminative clustering by regularized information maximization. 2010.

[61] M. Gong, L. Ma, Q. Zhang, and L. Jiao. Community detection in networks by using multiobjective evolutionary algorithm with decomposition. *Physica A: Statistical Mechanics and its Applications*, 391(15):4050–4060, 2012.

[62] R. Görke, P. Maillard, C. Staudt, and D. Wagner. Modularity-driven clustering of dynamic graphs. In *International symposium on experimental algorithms*, pages 436–448. Springer, 2010.

[63] Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. *CAP*, 367:281–296, 2005.

[64] B. Guedj. *Aggregation of estimators and classifiers : theory and methods*. Theses, Université Pierre et Marie Curie - Paris VI, December 2013.

[65] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986. The approach based on influence functions.

[66] F. R Hampel, E. M Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

[67] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, pages 100–108, 1979.

[68] C. Hennig. *trimcluster : Cluster Analysis with Trimming*, 2021. R package version 0.1-5.

[69] J. M. Hofman and C. H. Wiggins. Bayesian approach to network modularity. *Physical Review Letters*, 100(25), 2008.

[70] P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2009.

[71] M. Jabi, M. Pedersoli, A. Mitiche, and I. B. Ayed. Deep clustering: On the link between discriminative models and k-means. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1887–1896, 2019.

[72] A. K. Jain and R. C. Dubes. Algorithms for clustering data. *Englewood Cliffs: Prentice Hall, 1988*, 1988.

[73] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[74] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.*, 43(2-3):169–188, 1986.

[75] J. Jin, Z. T. Ke, and W. Wang. Phase transitions for high dimensional clustering and related problems. *The Annals of Statistics*, 45(5):2151–2189, 2017.

[76] J. Jin and W. Wang. Influential features pca for high dimensional clustering. *The Annals of Statistics*, 44(6):2323–2359, 2016.

[77] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1943–1950. IEEE, 2010.

[78] B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

[79] L. Kaufman and P. J. Rousseeuw. Clustering by means of medoids. pages 405–416, Amsterdam: North-Holland, 1987. Statistical Data Analysis Based on the L1 Norm and Related Methods.

[80] A. D. Kent. Cybersecurity Data Sources for Dynamic Network Research. In *Dynamic Networks in Cybersecurity*. Imperial College Press, June 2015.

[81] Y. Klochkov, A. Kroshnin, and N. Zhivotovskiy. Robust k-means clustering for distributions with two moments. *arXiv preprint arXiv:2002.02339v1*, 2020.

[82] M. Kumar, R. Ghani, and Z. Mei. Data mining to predict and prevent errors in health insurance claims processing. In *Proc of the 16th SIGKDD*, pages 65–74, 2010.

[83] P. Laforgue, S. Clémençon, and P. Bertail. On Medians of (Randomized) Pairwise Means. In *36th International Conference on Machine Learning*, volume 97, 2019.

[84] G. Lecué and M. Lerasle. Learning from MOM's principles: Le Cam's approach. *Stochastic Process. Appl.*, 129(11):4385–4410, 2019.

[85] G. Lecué and M. Lerasle. Robust machine learning by median-of-means: theory and practice. *Ann. Statist.*, 48(2):906–931, 2020.

[86] G. Lecué, M. Lerasle, and T. Mathieu. Robust classification via MOM minimization. *Mach. Learn.*, 109(8):1635–1665, 2020.

[87] M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Springer, Berlin, 1991.

[88] L. Leichtnam, E. Totel, N. Prigent, and L. Mé. Novelty detection on graph structured data to detect network intrusions. In *Conference on Artificial Intelligence for Defense*, 2020.

[89] F. Leisch. Bagged Clustering. *Working Paper 51, SFB "Adaptive Information Systems and Modeling in Economics and Management Science"*, 1999.

[90] M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*, 2011.

[91] Haifeng Li, Keshu Zhang, and Tao Jiang. Minimum entropy clustering and applications to gene expression analysis. In *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.*, pages 142–151. IEEE, 2004.

[92] T. Li, X. Yi, C. Caramanis, and P. Ravikumar. Minimax gaussian classification & clustering. In *Artificial Intelligence and Statistics*, pages 1–9. PMLR, 2017.

[93] S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[94] M. Löffler, A. S. Wein, and A. S. Bandeira. Computationally efficient sparse clustering. *arXiv preprint arXiv:2005.10817*, 3 2021.

[95] S. Loustau. Online clustering of individual sequences. 2014.

[96] G. Lugosi and S. Mendelson. Mean estimation and regression under heavy-tailed distributions: a survey. *Found. Comput. Math.*, 19(5):1145–1190, 2019.

[97] G. Lugosi and S. Mendelson. Sub-Gaussian estimators of the mean of a random vector. *Ann. Statist.*, 47(2):783–794, 2019.

[98] G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *J. Eur. Math. Soc. (JEMS)*, 22(3):925–965, 2020.

[99] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[100] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2019.

[101] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[102] A. Medad, B. Gregorutti, E. Genetay, and A. P. Nguema. Real-time graph clustering for network intrusion detection. *to be published*, 2022+.

[103] S. M. Milajerdi, R. Gjomemo, B. Eshete, R. Sekar, and V. N. Venkatakrishnan. Holmes: real-time apt detection through correlation of suspicious information flows. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 1137–1152. IEEE, 2019.

[104] S. Minsker. Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523*, 2018.

[105] D. G. Mixon, S. Villar, and R. Ward. Clustering subgaussian mixtures by semidefinite programming. *Information and Inference: A Journal of the IMA*, 6(4):389–415, 2017.

[106] A. C. Müller, S. Nowozin, and C. H. Lampert. Information theoretic clustering using minimum spanning trees. In *Joint DAGM (German Association for Pattern Recognition) and OAGM Symposium*, pages 205–215. Springer, 2012.

[107] H. Narayanan and S. Mitter. Sample complexity of testing the manifold hypothesis. In *Advances in neural information processing systems*, pages 1786–1794, 2010.

[108] M. Ndaoud. Sharp optimal recovery in the two gaussian mixture model. *arXiv preprint arXiv:1812.08078*, 2018.

[109] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson.

[110] M. E. J. Newman. Community detection in networks: Modularity optimization and maximum likelihood are equivalent. *arXiv preprint arXiv:1606.02319*, 2016.

[111] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[112] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[113] N. Nguyen and R. Caruana. Consensus clusterings. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pages 607–612. IEEE, 2007.

[114] N. P. Nguyen, T. N. Dinh, Y. Xuan, and M. T. Thai. Adaptive algorithms for detecting community structure in dynamic social networks. In *2011 Proceedings IEEE INFOCOM*, pages 2282–2290. IEEE, 2011.

[115] A. Novikov. PyClustering: Data Mining Library. *Journal of Open Source Software*, 4(36):1230, apr 2019.

[116] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[117] G. Pan, W. Zhang, Z. Wu, and S. Li. Online community detection for large complex networks. *PLoS One*, 9:1–37, 2014.

[118] 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*, 12:2825–2830, 2011.

[119] K. Pei, Z. Gu, B. Saltaformaggio, S. Ma, F. Wang, Z. Zhang, L. Si, X. Zhang, and D. Xu. Hercule: Attack story reconstruction via community discovery on correlated log graph. In *Proceedings of the 32Nd Annual Conference on Computer Security Applications*, pages 583–595, 2016.

[120] J. Peng and Y. Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM journal on optimization*, 18(1):186–205, 2007.

[121] H. O. Pollak. A remark on "Elementary inequalities for Mills' ratio" by Yûsaku Komatu. *Rep. Statist. Appl. Res. Un. Japan. Sci. Engrs.*, 4:110, 1956.

[122] L. Qi. The spectral theory of tensors (rough version). *arXiv preprint arXiv:1201.3424*, 2012.

[123] J. O Ramsay. When the data are functions. *Psychometrika*, 47(4):379–396, 1982.

[124] J. O. Ramsay and B. W. Silverman. *Applied functional data analysis: methods and case studies*, volume 77. Springer, 2002.

[125] G. Ritter. *Robust cluster analysis and variable selection*, volume 137 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2015.

[126] D. Rodriguez and M. Valdora. The breakdown point of the median of means tournament. *Statist. Probab. Lett.*, 153:108–112, 2019.

[127] G. Rossetti and R. Cazabet. Community discovery in dynamic networks: a survey. *ACM Computing Surveys (CSUR)*, 51(2):1–37, 2018.

[128] M. Royer. Adaptive clustering through semidefinite programming. *arXiv preprint arXiv:1705.06615*, 2017.

[129] G. Saporta. Méthodes exploratoires d'analyse de données temporelles. *Cahiers du Bureau universitaire de recherche opérationnelle Série Recherche*, 37:7–194, 1981.

[130] E. Schubert and P. J. Rousseeuw. Faster k-medoids clustering: Improving the pam, clara, and clarans algorithms. *arXiv preprint arXiv:1810.05691*, 2018.

[131] J. Shang, L. Liu, F. Xie, Z. Chen, J. Miao, X. Fang, and C. Wu. A real-time detecting algorithm for tracking community structure of dynamic networks. *arXiv preprint arXiv:1407.2683*, 2014.

[132] I. Sharafaldin, A. H. Lashkari, and A. Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *Proc. of the 4th ICISSP*, pages 108–116, 2018.

[133] Y. Shi and F. Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. *arXiv preprint arXiv:1206.6438*, 2012.

[134] N. Städler, P. Bühlmann, and S. van de Geer. $\ell_1$-penalization for mixture regression models. *TEST*, 19(2):209–256, 2010.

[135] E. C. Steinhaus. Microbial controlâthe emergence of an idea. a brief history of insect pathology through the nineteenth century. 26(2):107–160, 1956.

[136] J. Su, S. Kurtek, E. Klassen, and A.j Srivastava. Statistical analysis of trajectories on riemannian manifolds: bird migration, hurricane tracking and video surveillance. *The Annals of Applied Statistics*, 8(1):530–552, 2014.

[137] M. Sugiyama, G. Niu, M. Yamada, M. Kimura, and H. Hachiya. Information-maximization clustering based on squared-loss mutual information. *Neural Computation*, 26(1):84–131, 2014.

[138] M. Sugiyama, M. Yamada, M. Kimura, and H. Hachiya. On information-maximization clustering: Tuning parameter selection and analytic solution. In *ICML*, 2011.

[139] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[140] V. A Traag, L. Waltman, and N. J. Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[141] A. van Delft and H. Dette. A similarity measure for second order properties of non-stationary functional time series with applications to clustering and testing. *Bernoulli*, 27(1):469–501, 2021.

[142] C. Weir, R. Arantes, H. Hannon, and M. Kulseng. Operationally Transparent Cyber (OpTC). Data retrieved from IEEE Dataport.

[143] A Wilkelbauer. Moments and absolute moments of the normal distribution.(2012). *arXiv preprint arXiv:1209.4340*, 2012.

[144] L. L. Yan, T. P. Xiong, K. Rehan, F. Zhou, D. F. Liang, L. Chen, J. Q. Zhang, W. L. Yang, Z. H. Ma, and M. Feng. Single-atom demonstration of the quantum landauer principle. *Physical review letters*, 120(21):210601, 2018.

[145] X. Zhang, C. Ling, and L. Qi. The best rank-1 approximation of a symmetric tensor and related spherical optimization problems. *SIAM Journal on Matrix Analysis and Applications*, 33(3):806–821, 2012.

DOCTORAT BRETAGNE LOIRE / MATHSTIC

ENSAI École nationale de la statistique et de l'analyse de l'information

**Titre :** Quelques problématiques autour du clustering : robustesse, grande dimension et détection d'intrusion.

**Mots clés :** Clustering, median-of-means, grande dimension, entropie conditionnelle, détection d'intrusion, nombre de communautés.

**Résumé :** Le clustering vise à regrouper les données observées en différents sous-ensembles partageant des propriétés similaires. Le plus souvent ce regroupement se fait via l'optimisation d'un critère choisi à l'avance. Dans cette thèse CIFRE, nous avons étudié le clustering sous trois aspects différents.
Dans une première partie, nous proposons une méthode d'estimation robuste de K centroïdes basé sur le critère, dit des « K-means ». Nous proposons également une méthode d'initialisation robuste de la procédure. D'une part, la robustesse des procédures proposées a été testée par de nombreuses simulations numériques. D'autre part, nous avons montré un théorème donnant la vitesse de convergence d'un estimateur idéalisé en présence d'outliers ainsi qu'un théorème donnant le breakdown point de la méthode.
Dans une seconde partie nous nous plaçons dans le cadre d'un mélange équilibré de deux gaussiennes isotropes, centré en l'origine, afin de fournir la première analyse théorique d'un estimateur de clustering basé sur un critère d'entropie conditionnelle. Nous montrons que le critère est localement convexe, offrant d'une part des vitesses d'apprentissage rapide et d'autre part une inégalité oracle en grande dimension, lorsque le vecteur moyen de séparation est sparse.
Dans une troisième partie, plus pratique et consacrée à des graphes en cybersécurité, nous regardons si l'évolution du nombre de clusters obtenus par une méthode d'optimisation de modularité peut révéler des anomalies causées par une intrusion dans un système informatique.

**Title :** Some contributions related to data clustering : robustness, high-dimensionality and intrusion detection.

**Keywords :** Clustering, median-of-means, high dimension, conditional entropy, intrusion detection, number of communities.

**Abstract :** Clustering aims at grouping observed data into different subsets sharing similar properties. Most often this clustering is done through the optimization of a criterion chosen in advance. In this CIFRE thesis, we have studied clustering under three different aspects.
In a first part, we propose a robust estimation method of K centroids based on the so-called "K-means" criterion. We also propose a robust initialization method for the procedure. On the one hand, the robustness of the proposed procedures has been tested by numerous numerical simulations. On the other hand, we have shown a theorem giving the rate of convergence of an idealized estimator in the presence of outliers and a theorem giving the breakdown point of the method.
In a second part, we place ourselves in the framework of a balanced mixture of two isotropic Gaussians, centered at the origin, in order to provide the first theoretical analysis of a clustering estimator based on a conditional entropy criterion. We show that the criterion is locally convex, offering on the one hand fast learning rates and on the other hand an oracle inequality in high dimension when the mean separation vector is sparse.
In a third part, more practical and devoted to graphs in cybersecurity, we investigate whether the evolution of the number of clusters obtained by a modularity optimization method can reveal anomalies caused by an intrusion in a computer system.