Sketching for Large-Scale Learning of Mixture Models

Nicolas Keriven

ANNÉE 2017

UNIVERSITÉ DE RENNES 1

UNIVERSITE BRETAGNE LOIRE

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Bretagne Loire

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Mathématiques et applications

Ecole doctorale MATISSE

présentée par

Nicolas Keriven

Préparée à l'unité de recherche 6074 IRISA, équipe PANAMA
(Institut de Recherche en Informatique et Systèmes Aléatoires)

Sketching for Large-Scale Learning of Mixture Models

Thèse soutenue à l'IRISA
le 12 Octobre 2017

devant le jury composé de :

Arthur GRETTON
Associate Professor, UCL
rapporteur

Romain COUILLET
Professeur des universités, CentraleSupélec
rapporteur

Francis BACH
Directeur de recherche INRIA, ENS
examinateur

Michael DAVIES
Professor, University of Edinburgh
examinateur

Karin SCHNASS
University Assistant, University of Innsbruck
examinateur

François TAIANI
Professeur des Universités, INRIA Rennes
examinateur

Rémi GRIBONVAL
Directeur de recherche INRIA
directeur de thèse

Remerciements

J'aimerais tout d'abord remercier mon directeur de thèse Rémi Gribonval, qui m'a accepté en thèse puis accompagné tout au long de ces trois années. Sa grande disponibilité, sa connaissance encyclopédique, sa confiance en moi et ses conseils toujours pertinents ont été infiniment précieux pour faire de cette thèse ce qu'elle est devenue. J'espère que ces travaux marquent le début d'une longue et fructueuse collaboration.

I gracefully thank Arthur Gretton and Romain Couillet who kindly accepted to review my thesis, and all the members of the defense committee. A lot of thanks to Francis Bach and Mike Davies for coming to my pre-defense one year ago.

J'aimerais ensuite remercier les personnes avec qui j'ai eu la chance de collaborer durant ma thèse, Anthony Bourrier, Patrick Pérèz, Yann Traonmilin, Nicolas Tremblay, Gilles Blanchard. Travailler avec vous fut un vrai plaisir, votre grande expérience m'a énormément appris.

À mes parents et mes frères et soeur, Odile, Renaud, Tibo, Katu, Colin. Merci d'avoir fait de moi qui je suis aujourd'hui, de m'avoir donné le goût de la recherche que j'espère ne jamais perdre. Merci d'avoir toujours cru en moi.

This thesis very truly started with my first research internship at Queen Mary University in London, with Pr. Mark Plumbley and Dr. Ken O'Hanlon. They encouraged me to pursue research in academics, and I sincerely thank them for that.

J'aimerais remercier mes collègues de l'IRISA, pour leur bonne humeur permanente et l'ambiance merveilleuse qui règne au sein de l'équipe PANAMA. Antoine (and Denise, my dear sweeties), avec qui j'ai exploré les confins de l'Univers, de Shanghaï et du Spotted Cat de la Nouvelle-Orléans: mess around ! Clément et Cassiãão, mes compagnons de sardines grillées à Lisbonne, Srdan, Gilles et Luc, avec qui j'ai écumé les marchés de Noël de Berlin ou les canaux de Cambridge, Corentin, Corentin, Ewen, Romain, Nancy, Fred, Stéphanie, Armelle, Nicolas (encore un !), Alexandre, Hadrien, Antoine (sketch on !), Valentin, Jérémy, Anaïk, Cagdas, Nathan, Maxime, Nam, Jean-Joseph, Mohammed... avec qui aller au bureau tous les jours et boire un nombre déraisonnable de cafés fut un vrai plaisir.

J'aimerais tout particulièrement remercier les personnes avec qui j'ai eu la chance d'organiser la Journée Science et Musique 2015, et qui ont tenu le projet à bout de bras malgré les (nombreuses) difficultés rencontrées cette année-là: Luc, co-organisateur face à l'adversité, Anaïk, Ewen et Romain, Agnès, Julie et Evelyne, Alice, les membres de l'équipe Hybrid, tous les intervenants et conférenciers.

Merci à tous les gens que j'ai pu croiser qui rendirent ces trois années à Rennes inoubliables. Merci à Carole, pour les nombreux concerts que nous avons donnés ensemble au quatre coins de la Bretagne. Nono, Raph et Didine, mes trois mousquetaires, nos souvenirs ensemble ne sont que les premiers (j'ai encore des tas de recettes à tester sur vous). Orny, qui a presque réussi à me faire aimer le blues (Mister Tchaaaang), muxu. Merci à tous les talentueux musiciens que j'ai pu croiser à l'atelier jazz du Conservatoire et au Big Band de Rennes 2, et à Erwan Boivent, excellent professeur. Merci à tous ceux que j'ai pu croiser au Fun Club 35 pour toutes ces soirées endiablées.

À mes comparses que j'ai laissé pour la plupart à Paris ou ailleurs, Matthieu, mon vieux, Vincent, mon grand, Justine, LA, Théo et Lamiae, Guillaume, Jean... et tous les autres, ne vous inquiétez pas je reviens l'année prochaine. Préparez les Berthillons.

À Morgane, merci pour toutes les merveilleuses choses que l'on a vécu ensemble à Rennes. Bonne chance pour la suite, dis bonjour aux caribous de ma part. À Sioup et François, Popie et Audrey, tout le bonheur du monde et deux fois plus.

Cette thèse a été écrite en grande partie au Relais Breton près de Dinan, un grand merci à Corinne et Dominique. J'encourage le lecteur à aller y prendre une bière locale de ma part.

Enfin, par-dessus tout, merci à ma p'tite pirate en chocolat (noir 99%), pour son soutien constant, sa bonne humeur infaillible, et simplement pour faire partie de ma vie, qui serait bien triste sans elle. Je t'embrasse fort, et plus.

Abstract

Automatic learning processes are becoming ubiquitous in many domains of science. However, nowadays databases commonly comprise millions or billions of elements, which challenge traditional learning methods. Furthermore, modern database architectures involve new difficulties: data may be seen once then discarded (a situation usually referred to as data *stream*), often databases are not stored in one single location but distributed across several storage places and it is undesirable to gather the whole database in one place for the sake of privacy and robustness to malicious attacks. It has thus become necessary to derive learning procedures that are amenable to very large databases, and to distributed and streaming computing.

A popular idea is to define an *intermediary compressed representation* of a database, which is fast to compute, adapted to streaming and distributed computing through update and merge mechanisms, preserve data privacy, and such that the desired learning task can be performed *using only this compressed representation*, with a computational complexity that is greatly reduced compared to using the full database. A popular class of such representations is called *linear sketches*: the whole database is compressed into a single fixed-size vector called *sketch*, such that the sketch of the union of two databases is the sum of their sketches. Because of this property it is obvious that linear sketches are particularly convenient for streaming, distributed and parallel computing.

In [BGP13; BGP15], Bourrier et al. introduced a learning method based on a linear sketch formed by a random sampling of the empirical characteristic function of a collection of multidimensional vectors. They showed empirically that it was possible to fit a Gaussian Mixture Model (GMM) with fixed identity covariance on the original data, using only its sketch. However, the method was restricted to GMMs with identity covariance, and theoretical justifications were still an open question. Extending this method to other models and providing a theoretical analysis of the approach is the main purpose of this thesis work.

To do so, we develop an original framework based on several different sets of mathematical tools. The expression of the sketching operator is formalized by combining *kernel mean embedding*, which allows to define tunable Hilbertian metrics on the set of probability distributions, with *Random Feature expansions*, that approximate the infinite-dimensional mapping associated with a kernel function by a finite-dimensional mapping designed randomly. Using this mathematical framework, we analyze the sketching method under the lens of *Compressive Sensing*, which states that any signal that is in some sense less complex than the ambient dimension can be successfully compressed and estimated. We adapt classic proofs for finite-dimensional settings to our generalized infinite-dimensional framework. We provide guarantees for many problems, including for that of estimating mixtures of multivariate elliptic α-stable distributions from a sketch, for which no estimator was known. We particularly extend the framework and relate it to more traditional learning in two cases: first when recovering centroids from a sketch for the k-means or k-medians problem, and for GMM estimation with known covariance.

We introduce a flexible heuristic greedy algorithm coined Compressive Learning - Orthogonal Matching Pursuit with Replacement (CL-OMPR) that can estimate any parametric mixture model from any sketch in a very wide variety of situations. Experiments are performed on real and synthetic data for three models. First, mixtures of Diracs, for which our approach is shown to be more efficient and more stable than k-means on large databases; second, GMMs with unknown diagonal covariances, where the proposed approach is seen to be faster and lighter that classic Expectation Maximization (EM). And, finally, mixtures of multivariate elliptic α-stable distributions, where our approach is the first viable algorithm of which we are aware that can perform this task.

Résumé en français

Cette section rédigée en français reprend la motivation initiale de la thèse et en résume les contributions principales. Le manuscrit principal commence à la suite de ce résumé et est rédigé en anglais.

This section written in French reviews the initial motivation of the thesis work and summarizes the main contributions. The main part of the manuscript begins after this summary and is written in English.

Les bases de données modernes peuvent contenir un très grand nombre d'éléments. De plus, pour des questions de robustesse à d'éventuelles attaques informatiques ou de préservation de la vie privée, elles sont rarement stockées intégralement en un unique endroit, mais plutôt divisées en plusieurs parties conservées séparément et pour un temps possiblement limité. On observe donc un besoin croissant pour des méthodes de traitement rapides pouvant exploiter un très grand nombre de données, et suffisamment flexibles pour s'adapter à des bases de données distribuées et constamment mises à jour.

Une méthode classique basée sur ces critères est l'utilisation d'une représentation compressée de la base de données appelée *sketch linéaire*. Un sketch linéaire est un unique vecteur contenant diverses informations sur la base de données, pouvant être récupérées en temps voulu par l'utilisateur. La propriété principale des sketchs linéaires est la suivante: *le sketch d'une union de deux bases de données est la somme de leurs sketchs*. Ainsi, les sketchs linéaires sont adaptés aux situations où la base de données est distribuée, puisqu'il suffit de calculer les sketchs de chacune des parties puis de faire la somme de tous ces sketchs pour obtenir le sketch de la base de données globale, ainsi qu'au calcul *en ligne*, c'est-à-dire lorsque les éléments de la base sont collectés de manière séquentielle et que l'on ne désire pas forcément les garder en mémoire.

Une tâche classique d'apprentissage non-supervisé est celle d'estimer la distribution de probabilité sous-jacente d'un ensemble d'objets. C'est-à-dire, en supposant que l'on dispose de n objets $\{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ tirés aléatoirement et de manière indépendante selon une certaine distribution de probabilité $\pi^\star$, on cherche à estimer $\pi^\star$. En utilisant la propriété de linéarité des sketchs, il est facile de voir qu'un sketch linéaire $\mathbf{y}$ de cet ensemble peut toujours s'écrire $\mathbf{y} = \sum_{i=1}^n \Phi(\mathbf{z}_i)$, où $\Phi(\mathbf{z}) \in \mathbb{R}^m$ est le sketch d'un élément unique. Connaissant le nombre d'éléments n dans la base données (ce nombre étant lui-même un sketch linéaire), on peut alors calculer $\hat{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^n \Phi(\mathbf{z}_i)$. Il s'agit d'une *moyenne empirique* de la function $\Phi(\cdot)$, approximativement égale à l'espérance $\mathbb{E}_{\mathbf{z}\sim\pi^\star}\Phi(\mathbf{z})$ d'après la loi des grands nombres. Ceci permet de définir un opérateur linéaire sur les distributions de probabilité $\mathcal{A}\pi = \mathbb{E}_{\mathbf{z}\sim\pi}\Phi(\mathbf{z})$, et ainsi le sketch est approximativement $\hat{\mathbf{y}} \approx \mathcal{A}\pi^\star$. Ainsi il est possible d'estimer $\pi^\star$ à partir du sketch $\hat{\mathbf{y}}$ en utilisant des méthodes de *problème inverse*: le sketch est une mesure linéaire de la distribution de probabilité sous-jacente aux données, et l'on cherche à retrouver cette distribution à partir de cette mesure.

Cette thèse se concentre sur l'étude d'une méthode pour estimer la distribution de probabilité sous-jacente à un ensemble de vecteurs à partir d'un sketch linéaire de cet ensemble, en l'approchant par un modèle de mélange, c'est-à-dire une combinaison linéaire de k éléments pris dans un ensemble de distributions basiques $\{\pi_{\boldsymbol{\theta}}\}$. Un algorithme d'estimation rapide et flexible est défini, et une étude théorique des possibilités de reconstruction est réalisée.

Garanties générales pour les problèmes inverses sur des ensembles de basses dimension. Après un chapitre introductif, notre second chapitre porte sur une première contribution indépendante de la méthode de sketching décrite ci-dessus. Elle concerne l'étude de garanties

génériques de reconstruction pour les problèmes inverses, où l'on cherche à retrouver un signal $\mathbf{x}$ à partir de mesures bruitées $\mathbf{y} = \Psi(\mathbf{x}) + \mathbf{e}$. Dans de nombreux cas l'opérateur de mesure $\Psi(\cdot)$ perd de l'information, et la reconstruction n'est faisable qu'avec un *a priori* sur le signal $\mathbf{x}$. Des études ont été réalisées prouvant que sous certaines conditions la reconstruction est possible lorsque le signal est proche (au sens d'une certaine métrique) d'un modèle de signaux de "faible complexité" $\mathfrak{S}$: l'exemple le plus classique se trouve dans le domaine de l'acquisition comprimée, où le signal $\mathbf{x}$ est un vecteur approximativement k-parcimonieux, c'est-à-dire ne possédant que k coordonnées significativement différentes de zéro, où k est bien plus faible que la dimension ambiante. Ces études montrent que la possibilité d'une reconstruction robuste est en réalité équivalente à une propriété d'*Isométrie Restreinte* (ou RIP, pour *Restricted Isometry Property*) satisfaite par l'opérateur Ψ, c'est-à-dire que cet opérateur préserve approximativement les distances sur le modèle $\mathfrak{S}$. Ces résultats, d'abord formulés lorsque l'opérateur Ψ est linéaire et le signal $\mathbf{x}$ est de dimension finie, ont été étendus à la dimension infinie. Nous étendons encore ces résultats dans plusieurs directions:

- nous démontrons que ces résultats restent valides même lorsque l'opérateur $\Psi(\cdot)$ est non-linéaire et le signal $\mathbf{x}$ appartient à un ensemble quelconque munit d'une métrique (et pas forcément un espace vectoriel);
- la RIP exacte étant parfois difficile à prouver, nous laissons la possibilité d'une erreur additive $\eta \geq 0$ dans sa formulation et démontrons que la reconstruction stable est toujours possible et peu dégradée;
- enfin, lorsque l'opérateur $\Psi(\cdot)$ est tiré aléatoirement (comme c'est souvent le cas en acquisition comprimée), nous formulons des résultats de reconstruction *non-uniformes*, c'est-à-dire exprimant la possibilité de reconstruire un signal fixé $\mathbf{x}$ avec forte probabilité, par opposition à la capacité de reconstruire tous les signaux avec forte probabilité. Nous formulons une version non-uniforme de la RIP, qui est à notre connaissance totalement nouvelle.

Techniquement, tous ces résultats ne seront pas nécessaires pour analyser la méthode d'estimation de distribution à partir d'un sketch sur laquelle porte cette thèse. En particulier, l'opérateur de mesure $\mathcal{A}$ est linéaire par rapport aux distributions de probabilités. Néanmoins, ces nouveaux résultats pourront se révéler utiles lors de développements futurs.

Un cadre pour l'estimation de distributions à partir de sketchs. Notre deuxième contribution concerne la définition d'un cadre donnant naissance à un certain type de sketchs linéaires permettant de réaliser l'estimation de distributions. Au vu des résultats du chapitre précédent, nous aimerions prouver que l'opérateur $\mathcal{A}$ satisfait la RIP, afin de garantir la possibilité d'une estimation stable de la distribution $\pi^\star$ à partir du sketch $\mathcal{A}\pi^\star$. Néanmoins, les métriques classiques sur les espaces de mesures ne semblent pas adaptées. Ainsi, nous faisons appel au domaine des *méthodes à noyaux*, qui permettent de munir n'importe quel ensemble d'objets d'une géométrie euclidienne dont les propriétés sont ajustables relativement aisément par le choix d'une fonction appelée *noyau*, qui sert de produit scalaire. Nous utilisons ainsi un outil appelé *noyau moyen*, qui à partir d'un noyau classique sur un espace mesurable permet de définir un noyau sur les *mesures* définies sur cet espace. Afin de construire l'opérateur de sketch, nous faisons appel à un dernier outil mathématique : les *descripteurs aléatoires* (RF, pour *Random Features*) de noyaux, qui permettent d'approcher cette fonction noyau par des plongements explicites dans des espaces de dimension finie, construits aléatoirement. Ainsi, dans le troisième chapitre :

- nous démontrons que sous certaines conditions portant sur la "dimension" intrinsèque du modèle $\mathfrak{S}$ et sur le comportement des descripteurs aléatoires d'un certain noyau, l'opérateur de sketch construit (aléatoirement) à partir de ces descripteurs satisfait la RIP avec forte probabilité, ce qui garantit la possibilité d'une estimation stable des distributions proches du modèle à partir de leur sketch;
- en guise de première application du résultat, nous prouvons que la reconstruction est possible pour de nombreux modèles lorsque le sketch est aussi grand que la base de donnée originale. Ces résultats préliminaires ne permettent pour le moment pas de dire

qu'utiliser le sketch est plus économe que les méthodes utilisant toute la base de données, mais ils sont relativement simples à démontrer et permettent de se familiariser avec les outils. On observera par la suite qu'ils sont fort heureusement significativement sous-optimaux par rapport aux expériences pratiques;

– nous illustrons ces premiers résultats sur deux modèles, en particulier sur le modèle de mélanges de distributions α-stables elliptiques multivariées, pour lequel il n'existait à notre connaissance pas d'estimateur avec garanties.

Ainsi nous sommes potentiellement en mesure d'estimer une distribution à partir de son sketch de manière stable. Nous devons maintenant définir un algorithme capable de réaliser cette tâche. Malheureusement la fonction de coût à minimiser ressortant de notre étude est le plus souvent non-convexe, et il semble difficile de la minimiser avec exactitude. Nous définissons dans la suite une heuristique gloutonne, afin de traiter du cas particulier où le modèle de faible complexité $\mathfrak{S}$ est un modèle de mélange.

Algorithme d'estimation glouton. Une distribution dans un modèle de mélange $\mathfrak{S}_k$ est formée par la combinaison linéaire de k éléments pris dans un ensemble de base $\{\pi_{\boldsymbol{\theta}}\}$ indexé continuement par le paramètre $\boldsymbol{\theta}$. Estimer un tel mélange à partir d'un sketch revient à un problème de moindres carrés non-linéaires $\min_{\boldsymbol{\theta},\boldsymbol{\xi}} \left\| \sum_{l=1}^{k} \xi_l \mathbf{f}(\boldsymbol{\theta}_l) - \mathbf{y} \right\|_2^2$, où un vecteur $\mathbf{y}$ est approché par une combinaison linéaires d'*atomes* choisis dans le *dictionnaire* $\{\mathbf{f}(\boldsymbol{\theta})\}$, où $\mathbf{f}(\boldsymbol{\theta}) := \mathcal{A}\pi_{\boldsymbol{\theta}}$ pour la méthode de sketch. Dans cette troisième contribution, nous étudions un algorithme pouvant être appliqué à n'importe quel problème de ce type lorsque la fonction $\mathbf{f}$ est différentiable. Ainsi dans le quatrième chapitre :

– nous commençons par démontrer que la fonction de coût est potentiellement convexe par bloc lorsque l'on se situe suffisamment proche de la solution optimale. Nous évoquons la possibilité d'utiliser un algorithme de descente de gradient par bloc, qui en pratique se révèle très limité en l'absence d'une bonne initialisation;

– nous proposons ensuite notre approche gloutonne, qui ajoute itérativement les éléments au mélange en se basant sur le principe de l'algorithme classique *Orthogonal Matching Pursuit* (OMP). Afin de gérer le dictionnaire continuement indexé, l'algorithme alterne chaque itération avec des étapes de descente de gradient classique. Nous donnons une deuxième variante de cet algorithme avec *Remplacement*, qui est capable d'ajouter des atomes mais également d'en retirer par seuillage dur;

– nous appliquons ces algorithmes à un premier exemple simple, et nous observons effectivement que l'algorithme avec remplacement surpasse les autres approches.

Bien que les garanties de cet algorithme heuristique soient encore une question ouverte, nous démontrons par la suite qu'il se révèle empiriquement efficace sur de nombreux problèmes.

Applications. Nous appliquons notre approche gloutonne à l'estimation pratique de modèles de mélange à partir d'un sketch. Nous commençons le cinquième chapitre par décrire une méthode pour apprendre à partir d'une infime fraction de la base de données certains paramètres pour ajuster la méthode de sketch, en l'absence de connaissance *a priori* sur les données. Puis nous implémentons la méthode sur trois cas.

– Dans un premier cas, le sketch est utilisé afin de retrouver des mélange de distributions de Dirac. Bien que la vraie distribution des données $\pi^\star$ n'est bien évidemment pas un mélange de Diracs, on observe en pratique que les Diracs reconstruits sont placés au centre des groupes significatifs de données. Nous comparons ainsi la méthode obtenue avec l'algorithme classique des k-moyennes, qui calcule de tels centres. Nous observons que notre méthode de sketch est bien plus rapide que cette algorithme lorsque la base de données est de grande taille, et globalement plus stable. Nous comparons également les algorithmes sur données réelles, sur une tâche de reconnaissance de chiffres manuscrits.

– Dans un deuxième temps, l'algorithme est instancié pour estimer des modèles de mélange de distributions gaussiennes avec covariance diagonale. La méthode de sketch est comparée à l'algorithme classique Espérance-Maximisation (EM), et on observe une nouvelle fois qu'elle est plus rapide et moins gourmande en mémoire sur les grandes bases de données. Les algorithmes sont comparés sur un problème de reconnaissance de locuteur.

– Enfin, la méthode de sketch est appliquée à l'estimation de mélanges de distributions α-stables elliptiques multivariées. À notre connaissance, jusqu'à présent aucun algorithme n'était capable de traiter du cas multivarié pour les mélanges de distributions α-stables.

Dans ces trois cas, on observe que la taille du sketch suffisante pour obtenir à un résultat qualitativement bon se comporte en $m \approx \mathcal{O}(kd)$, où k est le nombre de composantes dans le mélange et d est la dimension des données. Cette taille est indépendante de la taille de la base de données initiales (et intuitivement optimale puisqu'il s'agit du nombre de paramètres du problème), ce qui confirme que nos premiers résultats théoriques du troisième chapitre étaient effectivement pessimistes.

Étude plus poussée de deux modèles. Dans un dernier chapitre, nous exploitons plus en profondeur les résultats associés à notre cadre générique, dans un double but : prouver que l'estimation est possible à partir d'un sketch dont la taille ne *dépend pas* de la taille de la base de données initiales mais seulement de la complexité du modèle (tel qu'observé en pratique), et relier ces résultats non plus à une certaine métrique à noyaux, parfois difficile à interpréter, mais à des fonctions de coût plus classique en apprentissage. Ainsi :

– nous développons une analyse plus poussée dans le cas des modèles de mélange. Cette analyse est basée sur l'hypothèse-clé suivante: au sein d'un mélange donné, les composantes sont deux à deux suffisamment séparées. Sous cette hypothèse de séparation, lorsque l'on examine deux modèles de mélange proches, chaque composante de l'un peut être appariée à une unique composante de l'autre, formant ce que l'on appellera un "dipôle", tout en étant éloignée de toutes les autres composantes. La différence entre deux mélanges peut alors être décomposée en une somme de dipôles. Sous certaines hypothèses sur le noyau, ces dipôles peuvent être traités indépendamment les uns des autres, et il est alors possible d'obtenir un contrôle fin sur les différentes métriques.

– Dans un premier exemple, nous revenons sur les mélanges de Diracs implémentés dans le chapitre précédent. Nous démontrons que l'estimation est possible avec une taille de sketch $m \approx \mathcal{O}\left(k^2 d^2\right)$, qui est encore légèrement pessimiste comparée à la pratique mais effectivement indépendente de la taille de la base de donnée initiale. Par ailleurs, nous relions ces garanties aux fonctions de coût classiques pour les problèmes de k-moyennes et k-médianes.

– Nous appliquons ensuite cette analyse aux modèles de mélange de gaussiennes dont la covariance est fixée et connue. Nous démontrons que l'estimation est possible avec une taille de sketch $m \approx \mathcal{O}\left(k^2 d^2 C\right)$, où C est un paramètre dépendant de la séparation imposée sur les moyennes au sein d'un mélange. On obtient $C = \mathcal{O}(1)$ pour une "grande" séparation en $\mathcal{O}\left(\sqrt{d \log k}\right)$, et $C = \mathcal{O}\left(e^d\right)$ pour une séparation faible en $\mathcal{O}\left(\sqrt{\log k}\right)$. Ces garanties sont reliées au maximum de vraisemblance classique.

Nous finissons par donner de nombreuses pistes, théoriques et algorithmiques, soulevées par nos travaux. Soulignons parmi celles-ci la recherche d'un algorithme avec garantie, potentiellement basé sur une possible relaxation convexe du problème tel que cela est fait dans le domaine de la super-résolution, ou encore la possible analyse RIP d'opérateurs non-linéaires obtenus en cascadant des descripteurs aléatoires construisant des noyaux plus complexes, tels que les réseaux de neurones multi-couches.

Publications associated to this work

Preprint

– R. Gribonval, G. Blanchard, N. Keriven, Y. Traonmilin. Compressive Statistical Learning with Random Feature Moments. 2017, pp. 1-72. *arXiv:1706.07180*

Journal papers

– N. Keriven, A. Bourrier, R. Gribonval, P. Pérez. Sketching for Large-Scale Learning of Mixture Models. *To be published in Information & Inference*, 2016, pp. 1-47. *arXiv:1606.02838*

Conference Proceedings

– N. Keriven, N. Tremblay, Y. Traonmilin, R. Gribonval. Compressive k-means. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

– N. Keriven, A. Bourrier, R. Gribonval, P. Pérez. Sketching for Large-Scale Learning of Mixture Models. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

Workshop presentations (two-pages abstracts)

– N. Keriven, R. Gribonval, G. Blanchard, Y. Traonmilin. Random Moments for Sketched Mixture Learning. *SPARS*, 2017.

  – *Best Student Paper Award, SPARS 2017.*

– Y. Traonmilin, N. Keriven, R. Gribonval, G. Blanchard. Spikes super-resolution with random Fourier sampling. *SPARS*, 2017.

– N. Keriven, A. Bourrier, R. Gribonval, P. Pérez. Sketching for Large-Scale Learning of Mixture Models. *International Matheon Conference on Compressed Sensing and its Applications*, 2015.

– N. Keriven, R. Gribonval. Compressive Gaussian Mixture Estimation by Orthogonal Matching Pursuit with Replacement. *SPARS*, 2015.

Code

– N. Keriven. SketchMLbox: a Matlab toolbox for large-scale mixture learning. `http://sketchml.gforge.inria.fr`, 2016.

Contents

Notations

Mathematical notations

$\mathbf{x}, \mathbf{y} \ldots$	vector / generic signal
$\mathbf{X}, \mathbf{Y} \ldots$	matrix
$\lvert\mathbf{X}\rvert$	determinant of a matrix
$\odot$	element-wise multiplication between matrices or vectors of the same size
$[\![\cdot, \cdot]\!]$	integer interval
$\mathfrak{P}(\mathfrak{Z})$	set of probability distributions on a measurable space $\mathfrak{Z}$
$\mathfrak{M}(\mathfrak{Z})$	space of finite signed measures on a measurable space $\mathfrak{Z}$
$\pi \in \mathfrak{P}(\mathfrak{Z})$	probability distribution
$\mu \in \mathfrak{M}(\mathfrak{Z})$	finite signed measure
$\delta_{\mathbf{c}}$	Dirac distribution at $\mathbf{c} \in \mathbb{R}^d$
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	multivariate Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$
$\mathcal{S}_\alpha(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	multivariate elliptic stable distribution with characteristic exponent $\alpha \in (0, 2]$, mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and precision matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$
$\mathbb{E}(\cdot)$	expected value of a random variable
$\langle \cdot, \cdot \rangle$	Hermitian inner product
$\langle \mu, f \rangle$	integral of a measurable function f with respect to the measure μ
$\lVert\cdot\rVert_{\mathcal{F}}$	Integral Probability Metric $\lVert\mu\rVert_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \lvert\langle \mu, f \rangle\rvert$
$\mathbb{S}^{k-1}$	$k-1$ dimensional simplex $\mathbb{S}^{k-1} = \left\{ \boldsymbol{\xi} \in \mathbb{R}_+^k \;\middle\vert\; \sum_{l=1}^k \xi_l = 1 \right\}$
κ	positive definite kernel
$\lVert\cdot\rVert_\kappa$	Maximum Mean Discrepancy (see Chapter 1 Sec. 1.4)
$D_{\mathrm{KL}}(\cdot \lVert \cdot)$	Kullback-Leibler divergence
$\lVert\cdot\rVert_{\mathrm{TV}}$	total variation norm of measures
$\lVert\cdot\rVert_{\boldsymbol{\Sigma}}$	Mahalanobis norm $\lVert\mathbf{x}\rVert_{\boldsymbol{\Sigma}} = \sqrt{\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}}$
$\mathcal{B}_{X,d}(x, r)$	ball in a (pseudo)metric space (X, d) with center x and radius r
$\mathcal{N}(d, Y, \delta)$	covering numbers of the set Y (see Definition in Appendix A)

Manuscript conventions

d	dimension of vectors in a database
n	number of elements in a database
k	number of components in a mixture / sparsity
m	size of measurement vector / sketch size
$\mathbf{z}$	items in a database, belonging to a set $\mathfrak{Z}$ (often but not necessarily $\mathbb{R}^d$)
$\mathcal{A}$	sketching operator (linear operator on finite signed measures)
Φ	sketching function such that $\mathcal{A}\mu = \langle \mu, \Phi \rangle$
$\mathbf{y}$	sketch / measurement vector
$\boldsymbol{\theta}$	parameter of a probability distribution $\pi_{\boldsymbol{\theta}}$, often in $\mathbb{R}^q$
Θ	tuple of parameters of a k-mixture $\Theta = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k)$
$\boldsymbol{\xi}$	weights of a k-mixture $\boldsymbol{\xi} \in \mathbb{S}^{k-1}$
$\mathfrak{S}$	low-dimensional model set
$\mathcal{F}_R$	family of random features $\mathcal{F}_R = \{\phi_{\boldsymbol{\omega}}\}$
$\boldsymbol{\omega}$	"frequency" that belongs to a set Ω (often but not necessarily $\mathbb{R}^d$)
Λ	probability distribution of frequencies
Δ	decoder (solver) of an inverse problem
η	additive error in the Lower Restricted Isometry Property
τ	bias term (distance to the model $\mathfrak{S}$)

Acronyms

CS	Compressive Sensing
RKHS	Reproducing Kernel Hilbert Space
MMD	Maximum Mean Discrepancy
RF	Random Feature
RFF	Random Fourier Feature
(L)RIP	(Lower) Restricted Isometry Property
IOP	Instance Optimality Property
SSE	Sum of Squared Error
GMM	Gaussian Mixture Model
EM	Expectation Maximization

Other notations are given when appropriate. All symbols inconsistent with the notation are disambiguated in the text.

List of Figures

List of Tables

Chapter 1

Introduction

> You can't do sketches enough.
> Sketch everything and keep your curiosity fresh.
>
> *John Singer Sargent*

We live in the "Big data" era. This term is now ubiquitous in nearly all domains of industry or science. It designates a reality: with large quantities of data new opportunities arise, but databases are becoming so big that traditional methods to handle them no longer work. With the invention of the Internet, increasing use of smartphones and tablets, or large-scale scientific experiments, massive amounts of data are collected, stored and processed every minute. The ability to treat databases at such unprecedented scales is a crucial challenge in many scientific fields, while at the same time very few organizations have access to the computational power of huge corporations.

Designing light, fast and mathematically sound methods that can process vast quantities of data with controlled computational power is a paramount challenge for the future. A natural and long-studied idea is to first *pre-process* the data to reduce its size while still keeping the ability to learn the information of interest from this compressed representation. The goal of this thesis is to study one such method where a collection of items is compressed into a representation called *sketch*, which is then exploited to learn the properties of the *distribution* of these items. This introductory chapter starts by presenting the motivations for this work. It then provides the main mathematical background upon which this thesis is built, and finishes by describing the contributions of the thesis and layout of the rest of the manuscript.

1.1 Motivation

We have recently witnessed the tremendous growth of a field called *Machine Learning* (one may even distinguish the related and overlapping field of *Data Science* [Don15]), at the intersection of Computer Science and Statistics, that has impacted a great number of other areas of science. Researchers in Machine Learning develop methods and algorithms to exploit data and learn how to automatically execute certain tasks: recognize faces on images, transcribe and translate speech, recommend movies that a user may like based on the content he has previously enjoyed, but also vanquish a grand master of Go, discover the properties of molecules, and so on and so forth.

Any learning scheme usually makes use of a set of *training data*, whose properties one wants to infer, in order to exploit them later in test situations. A substantial branch of Machine Learning is inspired by Statistics, in that it incorporates *randomness* in the formulation of the learning problem. For instance, assuming each training item is drawn at random from some *probability distribution*, one may want to learn some properties of this distribution.

However, nowadays data are available in such vast quantities that traditional statistical methods are strongly challenged.

Data compression. To handle large-scale learning tasks, a very natural idea is to compress (reduce the size of) a database before performing any actual "learning" on it. This two-step approach permits to focus separately on making the compression a fast and efficient process,

while the learning step will be naturally lighter since it is performed on the compressed representation and not on the full database itself.

With modern database architectures, new challenges emerge. First, data collections are often not stored in one single location but distributed across many storage places. Therefore, when building a compressed representation of a database of objects, it is desirable that partial summaries can be computed in each of these places then merged without having to transmit original data from one place to another (parallel/distributed computation). Second, it is increasingly frequent for data to be collected and/or updated continuously, in the so-called *data stream* model. As data arrives, its compressed representation must be updated: ideally, this operation should be fast to perform, and independent of previous events. Finally, increasing efforts are dedicated to preserving data *privacy*, *i.e.* prevent a third party from examining individual data points contained in a database. One therefore looks for representations that also encrypt the data and do not allow for recovering the original data from it, while still permitting to perform the appropriate learning task on it.

Data compression has a long and rich history, attempting to give an exhaustive description of the field here would be vain. We outline three different methodologies as examples, which will naturally lead us to the notion of *sketching*, which is the main focus of this thesis.

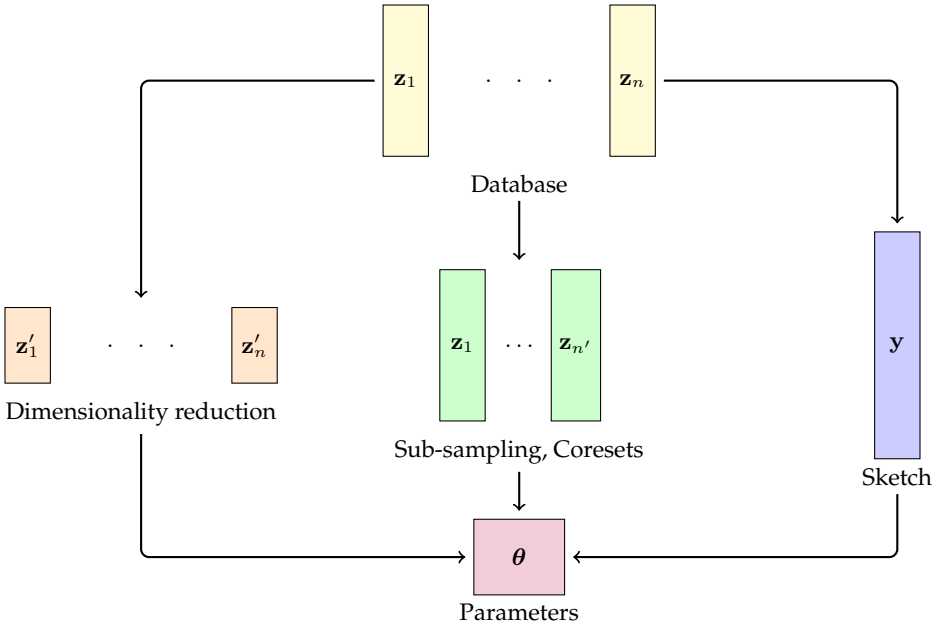

FIGURE 1.1: **Three complementary routes to compressive learning**. The training data is compressed into a smaller representation. This can either consist in reducing the dimensions of each individual entry (left), reducing the number of items (center), or in computing a more global compressed representation of the data, called *sketch* (right).

1.1.1 Three complementary compression schemes

Consider a database formed by a collection of n items $\mathbf{z}_i$. The goal is to learn some parameters $\boldsymbol{\theta}$ from this database. Compression schemes come in many different flavors, we describe three different kinds in this section, illustrated in Figure 1.1.

Dimensionality reduction. A natural idea is to compress *each individual item* in the database (left scheme in Figure 1.1). It can be referred to as *dimensionality reduction*. For instance, in a database containing images, each individual image can be stored in a compressed format with controlled loss of information (*e.g.* JPEG), the ability to learn from the full database will be only mildly affected. Dimensionality reduction is trivially amenable to streaming and distributed computing, since each data point is treated independently from all others. Data privacy may

however not be respected: if the compression process preserves enough information, then the original data points can be recovered from it.

The dimensionality reduction process can either be *learned* (a general idea often referred to as *feature selection*), or defined as a general-purpose reduction process independent of any task, for instance by *random projection*. Feature selection [TAL14; ATL16] has a long history linked to the broader problem of *model* selection. It consists in applying learning procedures to eliminate the "features" of the items $\mathbf{z}_i$ that are not relevant for the problem at hand and reduce their size. It can be a *supervised learning* method, meaning that the learning process is designed for a particular target task and it requires being able to qualitatively evaluate the execution of this task for candidate reduction schemes, or done in an *unsupervised* manner, meaning that it does not necessarily depend on the particular task that one wants to accomplish after the compression. For instance, a well-known unsupervised approach is Principal Component Analysis [Jol02] (PCA), which projects the data along dimensions where it varies the most, making the underlying assumption that these dimensions will be the most useful for the learning task. Opposite to these learned schemes, random projection [Ach01; FM03] is a general-purpose dimension-reduction method that is *not learned* from training data but simply takes random linear measurements of the data points. Under some assumptions on the items $\mathbf{z}_i$, such random measurements do not degrade the capacity to perform the learning task.

Random projection schemes has been successfully applied to many subsequent tasks such as classification [CJS09; Reb+13], regression [MM09], clustering with k-means [BZD10; Coh+15], or fitting a Gaussian Mixture Model (GMM) [Das99].

Dimensionality reduction is however challenged in the case where each individual item $\mathbf{z}_i$ is not very large but their number n is great, since the number of elements n is not reduced in the compressed version of the database.

Sub-sampling, coresets. A classic approach to treat databases with many elements is to *sub-sample* it, *i.e.* keep only a reduced number of elements from it (center-scheme of Figure 1.1). Sub-sampling has a very long history in Statistics, a good summary on sub-sampling schemes can be found *e.g.* in the book by Cormode et al. [Cor+11]. The loss of information induced by sub-sampling schemes can often be easily quantified with traditional statistical tools.

Modern sub-sampling schemes are also the basis for a class of methods called *coresets*. Coresets were initially developed for k-means [HPM04] and, more generally, subspace approximation [Fel+10; FL11]. They have been recently extended to other problems such as learning Gaussian Mixture Models [FFK11; Luc+17]. In coresets methods the number of items in the database is reduced by either sub-sampling (often weighted and adaptive [FFK11; Luc+17], similar to the k-means++ algorithm [AV07]) or construction of small local summaries using a hierarchical approach. Then classic learning algorithms are applied on this reduced database, often in a weighted version [Luc+17].

Compared to other compression methods, coresets are somehow closer to already approximately performing the learning task. For instance, the coreset described in [FS05] incorporates steps of Lloyd's k-means algorithm in its construction.

Coresets often present strong theoretical guarantees: compared to using the full data, the ability to perform the learning task from a coreset is generally precisely controlled. From a computational point of view, coresets are not specifically build for the streaming context, and they may require several passes over the data. Nevertheless they can still be adapted to streams of data, as described *e.g.* in [HPM04; FL11; Luc+17], by using a merge-and-reduce hierarchical strategy: for each batch of data that arrives sequentially, the user builds a coreset, then groups these coresets and build a coreset of coresets, and so on. However one must in general balance between keeping many coresets and letting the size of the overall summary grow with the number of points in the database, or keeping only highest-level coresets at the cost of losing precision in the theoretical guarantees each time the height of the hierarchical structure increases. Furthermore coresets methods based on a sub-sampling are obviously not privacy preserving, since some original data is kept unchanged in the reduced database.

Linear sketches. A third possibility corresponds to the right-scheme in Figure 1.1: the whole database is compacted into a single vector called *sketch*, built smartly in order to be able to learn the parameters from it. Relatively recent compared to compression of individual items

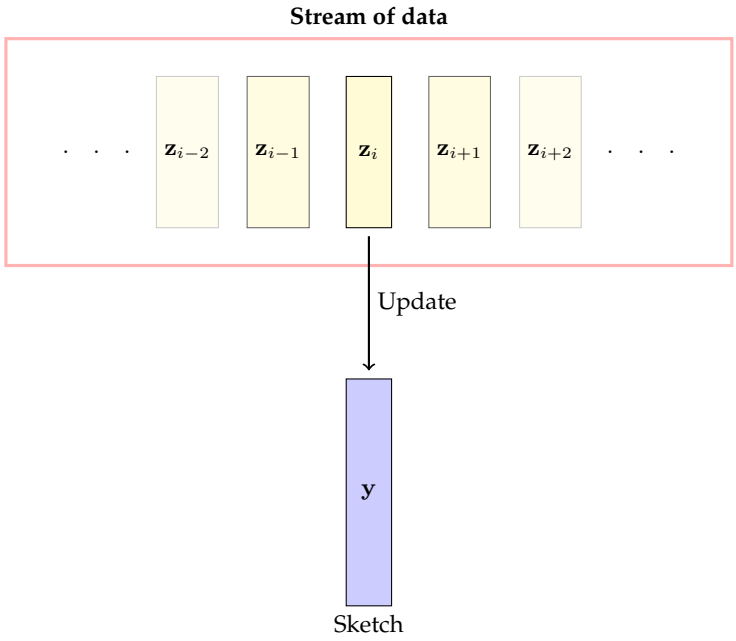

FIGURE 1.2: **Data Stream model and sketching**. In the data stream model, each data point is seen once then discarded. The sketch is updated each time a data point arrives. The update operation is ideally fast (real-time) and independent of previous events: linear sketches satisfy these requirements.

or sub-sampling, the literature on sketching has been quickly growing in the last years, due to their many computational advantages, see again [Cor+11]. In this book the notion of sketch is closely linked with the development of streaming methods. The sketch is a small summary of the data seen at a given time, that can be queried for a particular piece of information about the data. As required by the streaming context, when the database is modified, *e.g.* by inserting or deleting an element, the subsequent update of the sketch is fast and independent of previous events (see Figure 1.2).

A popular class of sketch amenable to both streaming and distributed computing is that of *linear* sketches, *i.e.* structures such that *the sketch of the union of two databases is the sum of their sketches*[1]. When a data point is added to the database, one simply adds the individual sketch of this single element (ideally, fast to compute) to the global sketch of the database. Moreover the sketch of a database divided across several storage devices is simply the sum of all the sketches of its parts, which renders distributed and parallel computing trivial to implement. In the case of distributed data, computing a sketch does not require the original data to be transmitted from one place to the other, which greatly increases robustness to malicious attacks. In the same fashion, they are extremely efficient in preserving data privacy.

The study of a particular sketching method for density estimation is the main focus of this thesis. Sketches have been used for a large variety of operations [Cor+11] such as detection of heavy-hitters [Cor+04; CM05; CH09] or, closer to our framework, approximately maintaining histograms [Tha+02] or quantiles [Gil+02]. However the latter are subject to the well-known curse of dimensionality and are unfeasible even in moderate dimension.

Compression *v.s.* Online learning. Finally, let us note that various learning algorithms have been directly adapted to streams of data without resorting to compressed representations of the database. Examples include modified versions of the Expectation-Maximization (EM) algorithm [AD03; CM09], the k-means algorithm [Guh+00; AJM09; GLA16], or Principal Component Analysis [GPP16]. In each case, the result of the algorithm is directly updated as data arrive, without resorting to maintaining an intermediate structure. However these algorithms do not fully benefit from the many advantages of sketches. Sketches are simpler to merge in a distributed context, update operations are more immediate, and the learning step can be

[1] One can think of linear sketches as a generalization of hash tables [Cor+11].

defered in time. For instance, private data can be collected and sketched on a large number of portable devices, then sketches can be transmitted to a centralized machine where they can be merged and the learning can take place. Furthermore, sketches can sometimes be queried for several task that do not need to be specified in advance, unlike when directly performing the learning as data are seen. We will see that this is the case for the sketch presented in this thesis: in Chapter 5, the *same* sketch is used to compute centroids as in the k-means problem, fit a Gaussian Mixture Model or a mixture of elliptic stable distributions[2] on the database.

1.1.2 From Sketching to linear measurements of probability distributions

Consider a database $\mathcal{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\} \subset \mathbb{R}^d$ that we want to compress into a linear sketch $\mathtt{Sk}(\mathcal{Z}) \in \mathbb{C}^m$ formed by m complex numbers (or, equivalently, $2m$ real numbers). Since the sketch is linear, *i.e.* the sketch of a union of two sets is the sum of their sketches, it is necessarily expressed as a sum of individual "sketches" for each data point:

$$\mathtt{Sk}(\mathcal{Z}) = \sum_{i=1}^{n} \Phi(\mathbf{z}_i) \ ,$$

where $\Phi : \mathbb{R}^d \to \mathbb{C}^m$ is some function. If one also maintains online the number n of elements in the database (which is also a linear sketch $n = \sum_{i=1}^{n} 1$), the following sketch can be computed:

$$\mathtt{Sk}_n(\mathcal{Z}) = \frac{1}{n} \sum_{i=1}^{n} \Phi(\mathbf{z}_i) \ , \tag{1.1}$$

which is formed by a collection of m *empirical generalized moments*, meaning that it is the empirical average of the function $\Phi(\mathbf{z})$. The term "generalized" refers here to the fact that we do not necessarily consider traditional polynomial moments $\mathbf{z}$, $\mathbf{z}^2$ etc. here, but any function of $\mathbf{z}$ instead. Strictly speaking, the sketch $\mathtt{Sk}_n(\mathcal{Z})$ is not a *linear* sketch, since it involves the normalization $1/n$ that is non-linear, however as we have seen it can be easily computed from the linear sketch $\{\mathtt{Sk}(\mathcal{Z}), n\}$.

Assume now that the items $\mathbf{z}_i$ are drawn *i.i.d.* from a *probability distribution* $\pi^\star$. Many methods in machine learning basically consist in learning some properties of this probability distribution.

The law of large numbers states that when the number of items n is large, the sketch $\mathtt{Sk}_n(\mathcal{Z})$ is approximately equal to

$$\mathtt{Sk}_n(\mathcal{Z}) \approx \mathbb{E}_{\mathbf{z} \sim \pi^\star} \Phi(\mathbf{z}) \ . \tag{1.2}$$

Therefore one can define the following *linear operator* on the set of probability distributions:

$$\mathcal{A}\pi = \mathbb{E}_{\mathbf{z} \sim \pi} \Phi(\mathbf{z}) \ ,$$

and the sketch is approximately[3]

$$\mathtt{Sk}_n(\mathcal{Z}) \approx \mathcal{A}\pi^\star \ .$$

Thus we can state:

> Any linear sketch is approximately a collection of *linear measurements of the underlying probability distribution*.

Leveraging this idea, learning from a sketch can be interpreted as an **inverse problem**: we treat a probability distribution as an object that is *encoded* in a sketch then *"decoded"* under the form of the properties we want to learn from it.

Using this paradigm, this thesis work is mainly based on the combination of three frameworks described below.

[2] Stable distributions are more often called α-stable distributions, however in the thesis we do not use this term: it somehow suggests that the so-called characteristic exponent α is fixed, and as we will see in a mixture the parameter α is *different for each component* of the mixture. Hence we simply denote "stable distribution", as in [Cas04] for instance.

[3] In fact, the sketch is exactly $\mathtt{Sk}_n(\mathcal{Z}) = \mathcal{A}\hat{\pi}_n$, where $\hat{\pi}_n = (\sum_{i=1}^{n} \delta_{\mathbf{z}_i})/n$ is the empirical probability distribution of the data. The approximation comes from the law of large numbers.

Step 1: goal of the thesis. In [BGP13], Bourrier et al. introduced an elegant sketching method that is the basis of this thesis work. They built a linear sketch (1.1) of a database of items in $\mathbb{R}^d$ by defining the function $\Phi(\mathbf{z}) = \left[e^{\mathrm{i}\boldsymbol{\omega}_j^\top \mathbf{z}}\right]_{j=1}^m$ as a collection of complex exponentials with randomly selected frequencies $\boldsymbol{\omega}_j \in \mathbb{R}^d$. They then developed an algorithm to fit a Gaussian Mixture Model (GMM) where each component has identity covariance on the original database, *using only its sketch*. It resulted in an efficient two-steps method to perform density fitting: first construct the sketch using all the advantages of a linear sketch, then learn the density using an algorithm on this sketch. They demonstrated empirically that this approach is more efficient than traditional methods that use the full database. However the method was limited to GMMs with identity covariance, and was only demonstrated to work empirically without theoretical justifications.

The main goal of this thesis is to extend this method to other models and provide theoretical justifications.

Step 2: analysis strategy For their sketching method, Bourrier et al. drew their inspiration from *Compressive Sensing* (CS) [CT04; Don06], in the field of Signal Processing. Compressive Sensing advocates that, under some hypotheses on a signal, compression can be performed at the acquisition stage (hence the name) to accelerate the measuring process and occupy less storage memory, *with controlled loss of information* about the true signal which can later be estimated. Basically, if this signal in intrinsically less complex than it seemed at first (for instance images are often formed by big blobs of uniform color, which is significantly simpler than any random color at each pixel), then one may be able to compress this signal into a representation that will be only as big as the signal's true complexity, without losing too much information. The sketching method of Bourrier et al. indeed falls into this category: it assumes that the probability distribution $\pi^\star$ is approximately a Gaussian Mixture Model, *i.e.* it has a much simpler structure than a probability distribution with no assumption.

First applied on signals such as audio or images, CS has been extended to more general notions of "signals", like functions. In this spirit, Bourrier et al. [Bou+14] derived very general conditions under which an object can be successfully compressed and recovered: if the measurement operator satisfies the so-called Restricted Isometry Property (RIP), *i.e.* it approximately preserves distance on a set of "simple" objects (like GMMs in the case of the sketching method), then robust decoding of this simple set is possible.

Studying RIP-like conditions for the linear operator $\mathcal{A}$ is the main theoretical contribution of this manuscript.

Step 3: mathematical framework Unfortunately, traditional norms on the set of probability distributions such as the total variation norm do not seem appropriate to derive the RIP: existing proofs of the RIP usually make use of the properties of the Euclidean norm, which in finite dimension is equivalent to any other norm but not in the infinite-dimensional space in which probability distributions live. The solution came when we incorporated ideas from *kernel methods* to the problem, and in particular a method called *kernel mean embedding*.

Kernel methods [Aro50; SS01] are elegant and successful examples of the paradigm which, like modern CS, consists in extending traditional methods to generalized notion of "signal". A "kernel" function between any two objects is defined, and under mild conditions it is shown to correspond to an inner product between the embeddings of these objects in a high-dimensional Hilbert space, known as the Reproducing Kernel Hilbert Space (RKHS). Therefore, *any set of objects* can be equipped with a Euclidean geometry, and classic learning algorithms can be performed with it.

This paradigm has indeed been applied to probability distributions, in a method referred to as *kernel mean embedding*. With this approach, any kernel on a measurable set of objects can be transformed into a kernel on the set of *probability distributions* on these objects.

In this thesis, we show that by using an appropriate and tunable kernel on the set of probability distributions, the analysis of the sketching operator $\mathcal{A}$ can be derived in a Compressive Sensing spirit.

The core of this thesis manuscript is therefore based on the combination of these three components:

- The main goal of this thesis is to extend, both theoretically and algorithmically, the sketching method of Bourrier et al. [BGP13]
- To provide information-preservation guarantees for the sketching method, our strategy is to prove that the sketching operator $\mathcal{A}$ satisfies the RIP.
- To do so, an appropriate mathematical framework is that of *kernel mean embedding* and related tools.

The rest of this introductory chapter is divided in three sections (Sections 1.2, 1.3 and 1.4) that can be read independently, each introducing the main mathematical tools of these three frameworks. We finish by a summary of the contributions of the thesis and layout of the manuscript in Section 1.5.

1.2 Inspiration: Sketched Gaussian Mixture modeling

In this section we describe the sketched learning method introduced by Bourrier et al. in [BGP13; Bou14; BGP15]. The principle is to first compress a database $\mathbf{z}_1, ..., \mathbf{z}_n \overset{i.i.d.}{\sim} \pi^\star$ of vectors in $\mathbb{R}^d$ into a sketch $\mathbf{y} \in \mathbb{C}^m$, with $m \ll nd$, then fit a density on the original database using only the sketch, which is faster and more memory efficient than using the full data. In this work, the sketch is formed by randomly sampling the empirical characteristic function of the data, and the density is recovered as a Gaussian Mixture Model (GMM) where each component has identity covariance. As described in the introduction, this sketching process is extremely amenable to parallel, distributed and online computing. By approximately recovering $\pi^\star$ (as a GMM) from the sketch, one therefore obtains an efficient two-step density estimation process: first compute the sketch efficiently, then estimate $\pi^\star$ with a procedure that is lighter than using the full data. As mentioned before, the main inspiration behind this work comes from the field of Compressive Sensing, of which we recall a few principles below.

1.2.1 Compressive Sensing

Compressive Sensing (CS) is based on the paradigm that when a signal has an intrinsic low-dimensional structure, it can be efficiently compressed without losing too much information and approximately recovered from its compressed representation. It is therefore useful in a number of applications where taking one measurement of the signal is costly: the compression step is applied directly at the acquisition stage (hence the name "Compressive Sensing") and the true signal is recovered later.

Somehow counter-intuitively, the acquisition rate can in fact be far below the classic Shannon-Nyquist threshold rate when the signal is *sparse*, *i.e.* it has only a few non-zero coefficients in a well-chosen basis. The field of Compressive Sensing was initiated by a quick succession of papers published between 2004 and 2006 by Candès, Tao and Romberg [CT04; CRT06b; CT06; CRT06a] and Donoho [Don06]. In 2013, Foucart and Rauhut published a book [FR13] that summarizes the main results about Compressive Sensing and gives a list of more than five hundreds references. Giving a complete overview of Compressive Sensing is of course out of scope here, we only give a quick description of the usual tools that will be of particular interest to us.

In traditional Compressive Sensing, a signal vector $\mathbf{x}^\star \in \mathbb{R}^d$ is measured through a linear operator $\mathbf{M} \in \mathbb{R}^{m \times d}$:

$$\mathbf{y} = \mathbf{M}\mathbf{x}^\star + \mathbf{e} \tag{1.3}$$

where $\mathbf{e} \in \mathbb{R}^m$ is additive noise. The goal is to (approximately) recover $\mathbf{x}^\star$ from $\mathbf{y}$.

Two key notions of Compressive Sensing will be of particular interest: *sparsity* and *random measurements*.

Sparsity. In Compressive Sensing the measurement process is most often *dimension reducing*, meaning that $m < d$, which makes the recovery problem theoretically ill-posed even in the noiseless case. Recovery is however possible when the signal $\mathbf{x} = [x_i]_{i=1}^d$ is *sparse*, meaning that it has only a few non-zero coefficients in a well-chosen basis (for simplicity, we will assume that it is sparse in the canonical basis of $\mathbb{R}^d$ here). We call *support* of the signal $\mathbf{x}$ the set of

indices where it is non-zero: $\Gamma(\mathbf{x}) = \{i \mid x_i \neq 0\}$, and $\mathbf{x}$ is called k-sparse if $|\Gamma(\mathbf{x})| \leq k$. For a set of indices Γ of cardinality k, we denote $\mathbf{M}_\Gamma \in \mathbb{R}^{m \times k}$ (resp. $\mathbf{x}_\Gamma \in \mathbb{R}^k$) the restriction of the matrix $\mathbf{M}$ (resp. the vector $\mathbf{x}$) to columns (resp. entries) in Γ. The measurement vector $\mathbf{y} \approx \sum_{i \in \Gamma(\mathbf{x})} x_i \mathbf{M}\mathbf{e}_i$ is then a sum of *atoms* in the *dictionary* $\mathcal{D} = \{\mathbf{M}\mathbf{e}_i \mid i \in [\![1, d]\!]\}$, where $\mathbf{e}_1, \ldots, \mathbf{e}_d$ is the canonical basis of $\mathbb{R}^d$.

Random Measurement Matrices. An attractive feature of Compressive Sensing is the design of a *random measurement* process. A key result states that a vector which is almost k-sparse (meaning that it is close to its best k-term approximation) can be estimated from $m = \mathcal{O}(k \log(d/k))$ measurements. One of the ways to prove this result is to show that the measurement matrix satisfies the so-called Restricted Isometry Property (RIP) for $2k$-sparse vector with constant $0 \leq \delta < 1$, which reads: for all $2k$-sparse vectors $\mathbf{x}$, we have

$$(1 - \delta) \|\mathbf{x}\|_2^2 \leq \|\mathbf{M}\mathbf{x}\|_2^2 \leq (1 + \delta) \|\mathbf{x}\|_2^2 \ . \tag{1.4}$$

The RIP states that the measurement matrix $\mathbf{M}$ is "almost" an isometry (*i.e.* it preserves distances) on the set of $2k$-sparse vectors. Intuitively, if $\mathbf{M}$ satisfies the RIP, it is able to "distinguish" two k-sparse vectors since it does not cancel on their difference, and recovering the measured k-sparse signal is possible. However designing *deterministic* matrices $\mathbf{M}$ that have the RIP in polynomial time is still an open question (see [FR13] Sec. 6.1). A key property of CS is that, by drawing the matrix $\mathbf{M}$ *randomly*, it is possible to prove that it satisfies the RIP *with high probability*. One of the first designs of this kind [CT04] was to randomly sample the Fourier transform of the signal. Modern popular choices include matrices where entries are *i.i.d.* Gaussian or Bernoulli variables (or, more generally, *sub-Gaussian* random variables), and many other random measurement processes have been studied since.

Recovery. One way [BD08b] to recover $\mathbf{x}^\star$ is to solve:

$$\arg\min_{\mathbf{z} \in \mathbb{R}^d} \|\mathbf{M}\mathbf{z} - \mathbf{y}\|_2 \text{ subject to } \|\mathbf{z}\|_0 \leq k \tag{1.5}$$

where the so-called ℓ_0 seminorm[4] $\|\cdot\|_0$ is the number of non-zero elements in a vector, *i.e.* this program returns the k-sparse vector that minimizes the measurement error. It can be shown that the solution is robust even in a noisy setting, or when the true signal $\mathbf{x}^\star$ is not exactly sparse.

Unfortunately, the minimization (1.5) is NP-complete ([FR13], Sec. 2.3), and thus unfeasible in practice. Two main approaches have been developed to solve this issue.

– *Convex relaxation*: the possibility for a convex relaxation of (1.5) is also a great achievement of Compressive Sensing. In this case, it consists in replacing the ℓ_0 norm with the ℓ_1 norm (the "convex" version of the ℓ_0 norm):

$$\arg\min_{\mathbf{z} \in \mathbb{R}^d} \|\mathbf{M}\mathbf{z} - \mathbf{y}\|_2 \text{ subject to } \|\mathbf{z}\|_1 \leq \tau. \tag{1.6}$$

Under this form, this problem is known as the LASSO [Tib11] and is more often found in Statistics, but it is equivalent ([FR13], Prop. 3.2) to other optimization problems such as Quadratically Constrained Basis Pursuit, or Basis Pursuit Denoising, which are more popular in CS.

Under some hypotheses on the matrix $\mathbf{M}$, the problem (1.6) robustly recovers the true signal, despite the NP-completeness of (1.5), with a number of measurements that still scales in $m = \mathcal{O}(k \log(d/k))$. Since (1.6) is a convex problem, many efficient solvers are available (see [FR13], Chap. 3 and 15).

– *Greedy algorithms*: greedy approaches consist in iteratively extending the support Γ of the solution according to some criterion. Since $\mathbf{y}$ is a weighted sum of atoms in the dictionary $\{\mathbf{M}\mathbf{e}_i\}_{i=1}^d$, Matching Pursuit (MP) [MZ93] and its variation Orthogonal Matching Pursuit (OMP) [PRK93] choose at each iteration the atom $\mathbf{M}\mathbf{e}_\ell$ that is most correlated with the

[4] See the definition of a seminorm in Appendix A.

residual signal $\mathbf{r} = \mathbf{y} - \mathbf{M}\mathbf{z}$, where $\mathbf{z}$ is the current solution. The OMP algorithm (Alg. 1) is the main inspiration for the CL-OMPR algorithm developed in this thesis. Under some assumptions (see *e.g.* [FR13], Sec. 3.2 and 5.3), the OMP algorithm is guaranteed to return a good approximation of the true signal $\mathbf{x}^\star$.

Compressive Sensing has since been extended in many directions. A number of works have been dedicated to design other measurement processes, in particular to increase their speed [LG16; Cha+15; KCW16]. The notion of "sparsity", or more generally low-dimensionality, has been extended to other type of signals: low-rank matrices [Can08; CT10; RFP07], cosparse vectors [Nam+13], dictionary models [CP11], signals that live in a union of subspaces [PGD13; Blu11] or more generally in a low-complexity subset of any Hilbert space [Bou+14; TG15; PDG15].

Algorithm 1: Orthogonal Matching Pursuit [MZ93; PRK93] (OMP)

Data: Measurement vector $\mathbf{y}$, measurement operator $\mathbf{M}$, sparsity k
Result: $\widetilde{\mathbf{x}}$
$\mathbf{r} \leftarrow \mathbf{y}$, $\Gamma \leftarrow \emptyset$;
for $t \leftarrow 1$ **to** k **do**
  Step 1: Find the normalized atom most correlated with residual
    $\ell^\star \leftarrow \arg\max_{i \notin \Gamma} \left| \left\langle \frac{\mathbf{M}\mathbf{e}_i}{\|\mathbf{M}\mathbf{e}_i\|_2}, \mathbf{r} \right\rangle \right|$;
  end
  Step 2: Expand sparse support
    $\Gamma \leftarrow \Gamma \cup \{i^\star\}$;
  end
  Step 3: Find coefficients by Least Squares
    $\mathbf{x} \leftarrow \arg\min_{\mathbf{z}} \|\mathbf{y} - \mathbf{M}_\Gamma \mathbf{z}\|$;
  end
  Update residual: $\mathbf{r} \leftarrow \mathbf{y} - \mathbf{M}_\Gamma \mathbf{x}$;
end
Return vector $\widetilde{\mathbf{x}}$ such that $\widetilde{\mathbf{x}}_\Gamma = \mathbf{x}$ and $\widetilde{\mathbf{x}}_{\Gamma^c} = \mathbf{0}$;

1.2.2 Sketched GMM learning

After this brief reminder of CS concepts, let us describe the sketching technique by Bourrier et al. [BGP13; Bou14; BGP15].

Denote $\mathfrak{P} = \mathfrak{P}(\mathbb{R}^d)$ the set of probability distributions on $\mathbb{R}^d$. The authors define a linear[5] operator $\mathcal{A} : \mathfrak{P} \to \mathbb{C}^m$ (referred to as *sketching operator*) that computes a collection of generalized moments of a probability distribution:

$$\mathbf{y} = \mathcal{A}\pi^\star := \mathbb{E}_{\mathbf{z} \sim \pi^\star} [\Phi(\mathbf{z})] \tag{1.7}$$

where $\Phi(\mathbf{z}) : \mathbb{R}^d \to \mathbb{C}^m$ is some non-linear map. One immediately sees the resemblance with the classic CS framework, although the reduction of dimension is here extreme: the measurement vector is of finite dimension while the encoded object is infinite dimensional.

In practice, one does not have access to the probability distribution $\pi^\star$, but to a database of vectors $\mathcal{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_n\}$ in $\mathbb{R}^d$, drawn *i.i.d.* from $\pi^\star$. The *sketch* of the database is then computed as

$$\hat{\mathbf{y}} = \mathcal{A}\hat{\pi}_n = \frac{1}{n} \sum_{i=1}^{n} \Phi(\mathbf{z}_i) \tag{1.8}$$

where $\hat{\pi}_n := n^{-1} \sum_{i=1}^n \delta_{\mathbf{z}_i}$ is the empirical distribution of the data. The goal is to recover $\pi^\star$ from $\hat{\mathbf{y}}$. By pursuing the comparison with CS the sketch can be rewritten as

$$\hat{\mathbf{y}} = \mathcal{A}\pi^\star + \mathbf{e}$$

[5] Meaning that $\mathcal{A}(t\pi + (1-t)\pi') = t\mathcal{A}\pi + (1-t)\mathcal{A}\pi'$.

where $\mathbf{e} := \mathbb{E}_{\mathbf{z}\sim\pi^\star}\left[\Phi(\mathbf{z})\right] - \frac{1}{n}\sum_{i=1}^{n}\Phi(\mathbf{z}_i)$ is a "noise" vector, which by the Law of Large Numbers have small amplitude with high probability. As mentioned before, two key notions of CS will be of interest here, random measurements and sparsity.

Random Measurements. Inspired by random Fourier sampling, the authors define the function Φ as a collection of complex exponentials evaluated at some *frequencies* $\boldsymbol{\omega}_j \in \mathbb{R}^d$, $1 \leq j \leq m$:

$$\Phi(\mathbf{z}) = \left[\exp\left(\mathrm{i}\mathbf{z}^\top\boldsymbol{\omega}_j\right)\right]_{j=1}^{m} \tag{1.9}$$

Continuing the parallel with CS and random Fourier sampling, the frequencies are chosen randomly: $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m \overset{i.i.d.}{\sim} \Lambda$ for some distribution Λ on $\mathbb{R}^d$.

With this definition, the sketching operator is a random sampling of the *characteristic function* of a probability distribution π, defined as

$$\psi_\pi(\boldsymbol{\omega}) = \mathbb{E}_{\mathbf{z}\sim\pi}\left[\exp\left(\mathrm{i}\mathbf{z}^\top\boldsymbol{\omega}\right)\right] \ . \tag{1.10}$$

Sparsity and Mixture Models. Bourrier et al. then define a notion of "sparsity" in $\mathfrak{P}$, making a parallel with the classic notion of mixture of distributions. Given $\mathfrak{T} = \{\pi_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathcal{T}\}$ a basic set of parametric distributions, a distribution π is said k-sparse if it is a *k-mixture*:

$$\pi_{\Theta,\boldsymbol{\xi}} = \sum_{l=1}^{k} \xi_l \pi_{\boldsymbol{\theta}_l} \ , \tag{1.11}$$

where $\Theta = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k) \in \mathcal{T}^k$ is a set of parameters and $\boldsymbol{\xi} \in \mathbb{S}^{k-1}$ is a weight vector with $\mathbb{S}^{k-1} = \left\{\boldsymbol{\xi} \in \mathbb{R}_+^k \mid \sum_{l=1}^{k}\xi_l = 1\right\}$ the $k-1$ dimensional simplex. To compare this definition with classic CS, a k-sparse vector is a linear combination of k elements from the canonical basis of $\mathbb{R}^d$, and a k-mixture is a combination of k elements from $\mathfrak{T}$. The *support* of a k-mixture $\pi_{\Theta,\boldsymbol{\xi}}$ is Θ and the *dictionary* of atoms is $\{\mathcal{A}\pi_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathcal{T}\}$. Notations and comparisons with traditional CS are given in Table 1.1. In Bourrier's original method [BGP13], the set $\mathfrak{T}$ is chosen as the set of Gaussian distribution with identity covariance, *i.e.* $\pi_{\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{\theta}, \mathbf{I})$ and $\boldsymbol{\theta} \in \mathcal{T} = \mathbb{R}^d$.

Recovery. Adapting the cost function (1.5), estimating a k-mixture from $\hat{\mathbf{y}}$ is done by (approximately) solving

$$\underset{\Theta\in\mathcal{T}^k,\ \boldsymbol{\xi}\in\mathbb{S}^{k-1}}{\arg\min} \left\|\mathcal{A}\pi_{\Theta,\boldsymbol{\xi}} - \hat{\mathbf{y}}\right\|_2 \tag{1.12}$$

which indeed corresponds to searching for the k-sparse distribution (*i.e.* k-mixture) that minimizes the measurement error. However, while analyzing (1.5) can be easily done with basic linear algebra (see [FR13] chap. 2), and most of the work in CS is dedicated to coping with its NP-completeness, analyzing the result produced by solving (1.12) is far from being trivial (before even wondering if it is algorithmically feasible). In [Bou14], Bourrier shows that, in the case of mixtures of Gaussians with identity covariance, for some deterministic choice of $m = \mathcal{O}\left(k^3 d\right)$ frequencies, the mapping $(\Theta, \boldsymbol{\xi}) \mapsto \mathcal{A}\pi_{\Theta,\boldsymbol{\xi}}$ is injective. Hence, if we had access to the true sketch $\mathbf{y} = \mathcal{A}\pi_{\Theta,\boldsymbol{\xi}}$, solving (1.12) with $\hat{\mathbf{y}} = \mathbf{y}$ would indeed produce the right results. However, robustness to noise (using $\hat{\mathbf{y}}$ instead of $\mathbf{y}$) and stability to modeling error (the true distribution of the data $\pi^\star$ is not exactly a GMM but close to one) are not guaranteed, in particular when using a random choice of frequencies as done in practice. In this thesis, we will provide theoretical guarantees for generalized sketch distribution estimation methods in Chapter 3, and examine more deeply the GMM with known covariances case in Chapter 6.

Algorithm. In general, the cost function (1.12) is non-convex with many spurious local minima[6], and cannot be exactly minimized numerically (this is somehow the equivalent of (1.5) being NP-complete).

The convex relaxation of such generalized problems has been formulated recently [Cha+12; TG15], however here this would be equivalent to manipulating the convex hull of the set of

[6]For instance, since permuting the components in the GMM does not change the cost function, each local (or global) minimum is found at least $k!$ times in the space of parameters.

	Usual compressive sensing	Compressive mixture estimation
Signal	$\mathbf{x} \in \mathbb{R}^d$	$\pi \in \mathfrak{P}$
Model	k-sparse vectors	k-mixtures $\pi_{\Theta,\boldsymbol{\xi}} = \sum_{l=1}^{k} \xi_l \pi_{\boldsymbol{\theta}_l}$
Measurement operator	$\mathbf{M} \in \mathbb{R}^{m \times d}$	$\mathcal{A} : \mathfrak{P} \to \mathbb{C}^m$
Support	$\Gamma(\mathbf{x}) = \{\ell \mid x_\ell \neq 0\}$	$\Gamma\left(\pi_{\Theta,\boldsymbol{\xi}}\right) = \Theta = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k\}$
Dictionary of atoms	$\{\mathbf{M}\mathbf{e}_\ell\}_{\ell=1,d}$	$\{\mathcal{A}\pi_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\mathcal{T}}$

TABLE 1.1: Correspondance between objects manipulated in usual compressive sensing of finite-dimensional signals and in compressive mixture estimation framework.

Gaussians in $\mathfrak{P}$, which seems difficult and is unlikely to yield a practical algorithm beyond very simple cases (see [DP15; DeC+15] which treats that of one-dimensional Diracs).

Another solution is to assume that the set of parameters $\mathcal{T} \subset \mathbb{R}^d$ is bounded, and discretize it to obtain a finite dictionary of atoms, from which usual methods can be applied [Bun+10]. However, it is well-known that, due to the so-called "curse of dimensionality", it becomes unfeasible when d increases to even moderate values: indeed for a precision of ε the size of the grid (and therefore of the dictionary) scales in $\mathcal{O}\left((1/\varepsilon)^d\right)$.

A final solution is to somehow derive continuous adaptations of classic CS algorithms. In this spirit, the authors [BGP13] develop an algorithm based on Iterative Hard Thresholding [BD08b; BD09], which in CS is an algorithm that alternates between a gradient descent step and a Hard Thresholding (IHT) step that ensures sparsity. Here the "gradient" step (with respect to π, not the parameters $\Theta, \boldsymbol{\xi}$) cannot by performed exactly, and is replaced by adding many atoms to the support ($2k$ in the original implementation) along which the decrease of the cost function (4.1) is maximal, then keeping only a support of size k by Hard Thresholding[7]. These atoms are found by randomly initialized numerical optimization schemes that only yield local minima. Compared to traditional IHT, this algorithm involves additional gradient descent steps of the whole cost function (4.1) at each iteration, to further decrease it.

Link with Generalized Method of Moments The recovery problem (1.12) is similar to a classic method in Statistics (and particularly Econometry) called the Generalized Method of Moments (GeMM) [Lan87; Hal05], in which the parameters $\boldsymbol{\theta}$ of a model are learned by matching a collection of theoretical generalized moments from the distribution $\pi_{\boldsymbol{\theta}}$ with empirical ones from the data. GeMM is often seen as an alternative to Maximum Likelihood estimation, to obtain different identifiability guarantees [BS10a; HK13; And+13] or for complex models for which the likelihood is not available, such as α-stable distributions [NPM01] or recursive models [Hal05, Sec. 1.3]. Traditionally, a finite number of moments is considered, but recent developments give guarantees in a theoretical framework where an infinite (integral) number of moments are available [CF00; CF02; CF14]. The collection of moments materialized by the sketch (1.7) is a sampling of the characteristic function, which has a long history of use in Statistics and GeMM, since it is a natural way of obtaining moment conditions [FM77; FM81]. It has been used for Gaussian Mixture modeling [Tra98; XK10], time series estimation [KY02], two-sample test [AJM08], estimation of α-stable distributions [NPM01], among many other problems. Following developments of GeMM with a continuum of moments instead of a finite number of them [CF00], estimators can be derived when the characteristic function is theoretically available at all frequencies simultaneously [CF02; XK10; CF14], *i.e.* by computing (approximate) integrals of the characteristic function against some operator. In practice these methods often require some prior knowledge to provide a good initialization of the optimization procedures, and because of the use of approximate methods for integration they can be time consuming.

[7] This version of the IHT algorithm is in fact more similar to CoSAMP [NT09], which is a variation of the classic OMP in which at each iteration the $2k$ atoms most correlated with the residual are added to the support, then reduced by Hard Thresholding. The "gradient step" of Bourrier indeed adds $2k$ atoms to the support by maximizing their correlation to the residual (note that only local maxima are found), then reduces it with Hard Thresholding.

GeMM *vs* Sketching. Although the recovery problem (1.12) is, strictly speaking, an instance of GeMM, the point of view of the sketching method is fairly distinct from what is usually considered in the GeMM literature. First, GeMM traditionally considers the collection of moments as a mathematical tool to obtain a statistical estimator different from maximum likelihood for instance, whereas the collection of moments that forms the sketch is considered as a *compressed representation* of the data and as a mean to achieve a learning task. As such, GeMM is rarely considered from a practical point of view: in most of the literature, the minimization (1.12) is performed with classic optimization routines and approximate numerical integration, while in the sketching context powerful algorithms are developed to handle more complex cases, with a particular attention on their computational complexity. From a theoretical point of view, GeMM studies usual statistical guarantees such as consistency and efficiency of the estimator $\hat{\boldsymbol{\theta}}$, while the results that we will obtain in this thesis are more akin to Compressive Sensing and Machine Learning. For instance, we consider robustness to modeling error (*e.g.* the true distribution of the data is not exactly a Gaussian Mixture but close to one), which is to our knowledge never a concern in GeMM. In the proof technique (Chapters 3 and 6), this is done by replacing the so-called "global identifiability condition" (*i.e.* injectivity of the moment operator, which is a classic condition in GeMM but is already difficult to prove and sometimes simply assumed by practitioners, see [NM94, p. 2127]) by the strictly stronger Restricted Isometry Property (RIP) from the Compressive Sensing literature.

1.3 Analysis strategy: robust decoding of inverse problems

In this section we present recent results [Bou+14; CDD09] that establish a connection between the Restricted Isometry Property (RIP) [Can08; Bar+08] and the existence of a robust decoder, for generalized linear inverse problems.

These results take place in a very general framework. Let E, F be two vector spaces equipped with two seminorms, respectively $\|\cdot\|_E, \|\cdot\|_F$. Consider a linear operator $\mathbf{M} : E \to F$, and suppose that we measure an object $\mathbf{x}^\star \in E$ with noise $\mathbf{e} \in F$:

$$\mathbf{y} = \mathbf{M}\mathbf{x}^\star + \mathbf{e}. \tag{1.13}$$

The linear operator $\mathbf{M}$ is usually dimension-reducing, meaning that the "dimension" of E is "larger" than that of F (note that neither of them is required to be finite-dimensional). In that case, simple arguments often prove that there is no hope of recovering all signals $\mathbf{x}^\star$ from their corresponding measurements $\mathbf{y}$ without prior knowledge. Hence we define a "low-dimensional" model $\mathfrak{S} \subset E$ (the equivalent of the set of k-sparse vectors in classic CS) that will serve as an *a priori* on $\mathbf{x}^\star$: our general goal is to guarantee that we can estimate objects in (or close to) the model $\mathfrak{S}$ from their measurements. Intuitively, as in CS the size of the measurement vector $\mathbf{y}$ required by this type of analysis will be tightly related to the "complexity" of the model $\mathfrak{S}$.

In [Bou+14], the authors establish an equivalence between two classic notions of Compressive Sensing, the Restricted Isometry Property (RIP) (or, more precisely, the *Lower* RIP, or LRIP) and the existence of an *Instance Optimal* decoder, *i.e.* a (non-linear) procedure capable of estimating $\mathbf{x}^\star$ from (1.13) while being stable with respect to modeling error and robust to noise. This connection was already noted in classic finite-dimensional CS [CDD09; FR13], its extension to general spaces was hinted in [PGD13] then formalized in [Bou+14] in the way that we are going to present here.

Lower restricted Isometry Property. The RIP (1.4) states that $\mathbf{M}$ is almost an isometry on the model $\mathfrak{S}$, meaning that it approximately preserves distance between any pair of signals in the model $\mathfrak{S}$. For the Lower RIP, only one side of the inequality remains.

Definition 1.3.1 (Lower Restricted Isometry Property)**.** *The operator $\mathbf{M}$ satisfies the Lower Restricted Isometry Property (LRIP) for the model $\mathfrak{S}$ with constant $\alpha > 0$ if: for all $\mathbf{x}, \mathbf{x}' \in \mathfrak{S}$ it holds that*

$$\|\mathbf{x} - \mathbf{x}'\|_E \leq \alpha \|\mathbf{M}(\mathbf{x} - \mathbf{x}')\|_F . \tag{1.14}$$

Remark 1.3.2. *In the original RIP (1.4), the RIP constant is placed on the left-hand side, and it is desirable to have a constant as close to* 1 *as possible, to obtain recovery guarantees on* e.g. *convex relaxations of the original inverse problem [TG15]. The corresponding constant in (1.4) is then formulated as* $\alpha = \frac{1}{\sqrt{1-\delta}}$*, with* δ *small. It is not our goal here:* any *RIP constant strictly positive will guarantee the existence of an instance optimal decoder, and in practice this constant will be fixed to an arbitrary value.*

Instance Optimal Decoder. A *decoder* Δ takes a linear operator $\mathbf{M}$, a measurement vector $\mathbf{y}$, and return a decoded signal $\widetilde{\mathbf{x}}$. It satisfies the *Instance Optimality Property* (IOP) with respect to the model $\mathfrak{S}$ if this decoding is stable to modeling error (the true signal is not in the model but close to the model) and robust to the presence of noise.

Definition 1.3.3 (Instance Optimality Property (IOP))**.** *A decoder* Δ *satisfies the Instance Optimality Property (IOP) for the operator* $\mathbf{M}$ *and model* $\mathfrak{S}$ *with constants* $A, B > 0$ *and pseudo-metric* d_E *if: for all signals* $\mathbf{x}^\star \in E$ *and noise* $\mathbf{e} \in F$*, denoting* $\widetilde{\mathbf{x}} = \Delta(\mathbf{M}, \mathbf{M}\mathbf{x}^\star + \mathbf{e})$ *the recovered signal, it holds that:*

$$\|\mathbf{x}^\star - \widetilde{\mathbf{x}}\|_E \leq A d_E(\mathbf{x}^\star, \mathfrak{S}) + B \|\mathbf{e}\|_F \tag{1.15}$$

where $d_E(\mathbf{x}, \mathfrak{S}) = \inf_{\mathbf{x}_\mathfrak{S} \in \mathfrak{S}} d_E(\mathbf{x}, \mathbf{x}_\mathfrak{S})$.

The result in [Bou+14] is the following.

Theorem 1.3.4 (Bourrier et al. [Bou+14])**.** *Consider an operator* $\mathbf{M}$ *and a model* $\mathfrak{S}$*.*

1. *If there exists a decoder* Δ *which satisfies the Instance Optimality Property for* $\mathbf{M}$ *and* $\mathfrak{S}$ *with constants* $A, B > 0$ *and pseudometric* d_E*, then the operator* $\mathbf{M}$ *satisfies the LRIP for the model* $\mathfrak{S}$ *with constant* $\alpha := B$ *(note that* d_E *does not play a role in the LRIP).*
2. *If the operator* $\mathbf{M}$ *satisfies the LRIP for the model* $\mathfrak{S}$ *with constant* α*, then assuming that it exists*[a] *the decoder defined by*

$$\Delta(\mathbf{M}, \mathbf{y}) \in \operatorname*{arg\,min}_{\mathbf{x} \in \mathfrak{S}} \|\mathbf{M}\mathbf{x} - \mathbf{y}\|_F \tag{1.16}$$

satisfies the Instance Optimality Property for the operator $\mathbf{M}$ *and model* $\mathfrak{S}$ *with constants* $A := 1,\ B := 2\alpha$ *and pseudometric* d_E *defined by* $d_E(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_E + 2\alpha \|\mathbf{M}(\mathbf{x} - \mathbf{x}')\|_F$.

[a] when it is not the case, we can define the decoder as an approximate minimization with an additional additive error as small as desired, see Chapter 2.

This result is not exactly found in this form in [Bou+14], but is proved through the use of the so-called "null space property", with slightly different constants. In Theorem 2.2.3 we prove a generalization of this theorem, hence the proof also applies to Theorem 1.3.4.

Relation with Sketching. Note that the decoder (1.16) corresponds to the cost function (1.5) in classic CS and (4.1) in sketch learning. A major theoretical contribution of this thesis is to prove that the sketching operator $\mathcal{A}$ satisfies the LRIP for some model $\mathfrak{S}$, which proves that the recovery program (4.1) is an instance optimal decoder.

Proving the LRIP: Normalized Secant Set. In traditional CS, where $\mathbf{M}$ is a random matrix, one of the most classic proofs for the LRIP [Bar07] has two steps. First, it uses concentration inequalities to prove that, for any *fixed* pair of signals $\mathbf{x}, \mathbf{x}' \in \mathfrak{S}$, with high probability on the drawing of $\mathbf{M}$, the desired inequality (1.14) is verified. Then this inequality is *uniformly extended*, *i.e.* it is shown that with high probability on $\mathbf{M}$, for *all* $\mathbf{x}, \mathbf{x}' \in \mathfrak{S}$ the inequality (1.14) is verified (note the inversion of quantifiers, which is crucial). It is done by showing that certain sets have *finite* covering numbers, meaning that they can be covered by a finite number of balls. The order of magnitude of these covering numbers largely drives the size of the measurement vector. Detailed definitions on coverings can be found in Section A.1 in Appendix A.

A key object stemming from this proof is the *normalized secant set* of $\mathfrak{S}$:

$$\mathcal{S}(\mathfrak{S}) = \left\{ \frac{\mathbf{x} - \mathbf{x}'}{\|\mathbf{x} - \mathbf{x}'\|_E} \;\middle|\; \mathbf{x}, \mathbf{x}' \in \mathfrak{S},\ \|\mathbf{x} - \mathbf{x}'\|_E > 0 \right\}. \tag{1.17}$$

Indeed, the LRIP for the operator $\mathbf{M}$ is equivalent to having: for all $\mathbf{y} \in \mathcal{S}(\mathfrak{S})$,

$$\|\mathbf{M}\mathbf{y}\|_E \geq \alpha^{-1}. \tag{1.18}$$

The normalized secant set is actually the object that needs to be "low-dimensional", and in particular we usually need to prove that it has finite covering numbers. While this is trivial in the finite-dimensional case, where it is included in a unit ball that is necessarily compact, in an infinite-dimensional framework the task is much more arduous. Interestingly, when conversely the normalized secant set is *assumed* to have finite covering numbers, then it is possible to explicitly build an operator that will satisfy the RIP with high probability (however not always implementable in practice), as described in [PDG15].

1.4 Mathematical framework: Kernel methods

Let us now turn to the third and final main inspiration for this thesis: kernel mean embedding, and Random Feature expansions of kernels.

Kernel methods have greatly gained in popularity in the last decades. Before the recent explosion of deep neural networks, they yielded state-of-the-art results on many problems. With kernel methods, any set of objects can be embedded into a Hilbert space, by defining only a *kernel* function that will serve as the inner product of this space. By the so-called *kernel trick*, the actual embedding does not need to be explicitly expressed. Hence, not only kernel methods permit to solve a more general class of problems than classic learning algorithms, they are also an elegant and powerful way to define geometries on *any* set of objects: graphs [Vis+10], text documents [Lod+02], molecules [Les+04], and so on.

Kernel methods have been applied to *probability distributions*. Kernel mean embedding [BTA04; Bor+06; Gre+06; Smo+07; Son08; Sri+10] is a method that defines a kernel between two finite signed measures over a measurable space $\mathfrak{Z}$, given a more traditional kernel on the space $\mathfrak{Z}$. It has been used for two-sample test [Gre+06; Smo+07; Chw+15], *i.e.* determining if two sets of samples come from the same distribution or not, but also as a measure of independence [Fuk+07], for classification [Mua+12; Sut+15; OSS15], for performing operations on distributions [Sch15] and, closer to the sketching method studied in this thesis, for density estimation [Son+08; Sri11]. In particular, in [Sri11] the estimation of a mixture model with respect to the metric of the RKHS is considered with a greedy algorithm. The proposed algorithm is however designed to approximate the target distribution by a large mixture with many components, resulting in an approximation error that decreases as the number of components increases, while the sketching approach considers a mixture model as a "sparse" combination of a fixed, limited number of components which we aim at identifying. Furthermore the algorithm proposed in [Sri11] does not seem to be directly implementable.

In this section, we recall the basic tools of kernel methods, present the main ideas of kernel mean embeddings, and finish by describing the so-called Random Feature expansions of kernels.

1.4.1 Reproducing Kernel Hilbert Space

The theory of Reproducing Kernel Hilbert Space (RKHS) is based on the notion of positive definite kernel.

Definition 1.4.1 (Positive definite (p.d.) kernels)**.** *Let $\mathfrak{Z}$ be an arbitrary set. A Hermitian*[a] *function, or* ***kernel****, $\kappa : \mathfrak{Z} \times \mathfrak{Z} \mapsto \mathbb{C}$ is called* ***positive definite (p.d.)*** *if, for all $n \in \mathbb{N}$, $c_1, ..., c_n \in \mathbb{C}$ and all $\mathbf{z}_1, ..., \mathbf{z}_n \in \mathfrak{Z}$, we have*

$$\sum_{i,j=1}^{n} c_i \bar{c}_j \kappa(\mathbf{z}_i, \mathbf{z}_j) \in \mathbb{R}_+ \tag{1.19}$$

[a] meaning that $\kappa(\mathbf{z}, \mathbf{z}') = \overline{\kappa(\mathbf{z}', \mathbf{z})}$

Note that strict positivity is not mandatory in the above equation. In terms of vocabulary, p.d. kernels bear connections with, e.g., positive *semi*-definite matrices (however they are indeed called positive definite kernels in the literature).

RKHS. Let $\kappa : \mathfrak{Z} \times \mathfrak{Z} \to \mathbb{C}$ be a p.d. kernel. The Moore-Aronzajn theorem [Aro50] states that we can associate to this kernel a unique functional Hilbert Space $\mathfrak{K} \subset \mathbb{C}^{\mathfrak{Z}}$ that satisfies the following properties:

- for any $\mathbf{z} \in \mathfrak{Z}$ we have $\kappa(\mathbf{z}, \cdot) \in \mathfrak{K}$;
- *reproducing property*: for all $\mathbf{z} \in \mathfrak{Z}$, $f \in \mathfrak{K}$ it holds that $f(\mathbf{z}) = \langle f, \kappa(\mathbf{z}, .) \rangle_{\mathfrak{K}}$.

This space is called the *Reproducing Kernel Hilbert Space (RKHS)* associated with κ. By the reproducing property, we have $\kappa(\mathbf{z}, \mathbf{z}') = \langle \kappa(\mathbf{z}, .), \kappa(\mathbf{z}', .) \rangle_{\mathfrak{K}}$, hence defining the mapping $\varphi : \mathfrak{Z} \to \mathfrak{K}$ as $\varphi(\mathbf{z}) = \kappa(\mathbf{z}, .)$, we have indeed defined an embedding of $\mathfrak{Z}$ into the Hilbert space $\mathfrak{K}$ such that $\langle \varphi(\mathbf{z}), \varphi(\mathbf{z}') \rangle_{\mathfrak{K}} = \kappa(\mathbf{z}, \mathbf{z}')$.

Kernel trick. The so-called *kernel trick* is linked to the *representer theorem* [KW70; SHS01], which basically states that the solutions to a large class of optimization problems over $f \in \mathfrak{K}$ that use a finite quantity of data $\mathbf{z}_1, ..., \mathbf{z}_n \in \mathfrak{Z}$ can be expressed as functions $f(\cdot) = \sum_{i=1}^{n} \alpha_i \kappa(\cdot, \mathbf{z}_i)$, and that finding the α_i can be done by using only the inner products $\kappa(\mathbf{z}_i, \mathbf{z}_j)$, meaning that the explicit high-dimensional mapping needs never be expressed. This allows for a large class of algorithm to have a "kernel" counterpart [SS01]: Principal Component Analysis (PCA), Support Vector Machine (SVM), and so on.

Feature maps. An alternative view on RKHS is to directly define *feature maps*. Let $\Pi : \mathcal{X} \to \mathcal{H}$ be a mapping from $\mathfrak{Z}$ to a Hilbert Space $\mathcal{H}$. It is then immediate that $\kappa(\mathbf{z}, \mathbf{z}') := \langle \Pi(\mathbf{z}), \Pi(\mathbf{z}') \rangle_{\mathcal{H}}$ is a positive definite kernel κ, to which is associated some RKHS $\mathfrak{K}$, which is *not* necessarily the Hilbert space $\mathcal{H}$ (in particular the space $\mathcal{H}$ is not necessarily a functional space included in $\mathbb{C}^{\mathfrak{Z}}$). In fact, we have $\mathcal{H} = \mathfrak{K}$ only when Π is the "canonical" feature map of the kernel $\Pi(\mathbf{z}) = \varphi(\mathbf{z}) = \kappa(\mathbf{z}, .)$, but this is not necessarily the case[8].

1.4.2 Mean embedding of finite signed measures

Kernel mean embedding [BTA04; Bor+06; Gre+06; Smo+07; Son08; Sri+10] defines, from a positive definite kernel on a measurable space $\mathfrak{Z}$, a positive definite kernel on the set of probability distributions over that space. Although mean kernels are usually stated on probability distributions only, we formulate here the definitions on the whole space $\mathfrak{M}$ of finite signed measures over $\mathfrak{Z}$. One can easily check that the few results that we will use also hold in that case, by noting that any finite signed measure can be decomposed as $\mu = c\pi - c'\pi'$, where π and π' are probability distributions and $c, c' \geq 0$.

[8] For instance, consider the linear kernel $\kappa(\mathbf{z}, \mathbf{z}') = \mathbf{z}^\top \mathbf{z}'$ where $\mathbf{z} \in \mathbb{R}^d$. A possible feature map is just the identity $\Pi(\mathbf{z}) = \mathbf{z}$ associated with the Hilbert space $\mathcal{H} = \mathbb{R}^d$, however the *RKHS* associated to this kernel is the dual of $\mathbb{R}^d$, and the corresponding feature map φ is the one that associates $\mathbf{z}$ with its dual element $\mathbf{z}^*$ such that $\mathbf{z}^*(\mathbf{z}') = \mathbf{z}^\top \mathbf{z}'$.

Mean kernel. Let $\mathfrak{Z}$ be any measurable space, denote $\mathfrak{M}$ the space of finite signed measures[9] over $\mathfrak{Z}$. Let κ be a p.d. kernel on $\mathfrak{Z}$, and $\mathfrak{K}$ the associated RKHS. All throughout the manuscript, for simplicity and technical (integrability) reasons, we **assume that the kernel κ is bounded**.

One can then define [Gre+06; Sri+10] the following feature map from $\mathfrak{M}$ to $\mathfrak{K}$:

$$\Pi(\mu) := \int_{\mathfrak{Z}} \kappa(\mathbf{z}, .) d\mu(\mathbf{z}) \in \mathfrak{K} \tag{1.20}$$

It naturally defines a kernel between measures, that by abuse of notation we also denote κ, called the *mean* kernel:

$$\kappa(\mu, \mu') := \langle \Pi(\mu), \Pi(\mu') \rangle_{\mathfrak{K}} \overset{(a)}{=} \iint_{\mathfrak{Z}} \kappa(\mathbf{z}, \mathbf{z}') d\mu(\mathbf{z}) d\mu'(\mathbf{z}') \tag{1.21}$$

where (a) is shown, e.g., in [Sri+10] as a direct consequence of Riesz's representation theorem and the reproducing property of κ . The careful reader would have noted that the space $\mathfrak{K}$ is *not* the RKHS associated with the *mean* kernel $\kappa(\mu, \mu)$, but the RKHS associated with the *original* kernel $\kappa(\mathbf{z}, \mathbf{z}')$. However the mean kernel can indeed be defined with a feature map Π that goes from $\mathfrak{M}$ to $\mathfrak{K}$.

Empirical approximation. Given two sets of samples $\mathbf{z}_1, ..., \mathbf{z}_n \overset{i.i.d.}{\sim} \pi$ and $\mathbf{z}'_1, ..., \mathbf{z}'_{n'} \overset{i.i.d.}{\sim} \pi'$, one can approximate the kernel $\kappa(\pi, \pi')$ by

$$\kappa(\pi, \pi') = \mathbb{E}_{\mathbf{z} \sim \pi} \mathbb{E}_{\mathbf{z}' \sim \pi'} \kappa(\mathbf{z}, \mathbf{z}') \approx \frac{1}{nn'} \sum_{i=1}^{n} \sum_{j=1}^{n'} \kappa(\mathbf{z}_i, \mathbf{z}'_j) = \kappa\left(\hat{\pi}_n, \hat{\pi'}_{n'}\right) . \tag{1.22}$$

This yields a very simple framework to apply kernel methods to objects that can be represented as sets of features (images, biological data...). It has been shown that the quality of this estimation can be improved beyond the use of simple concentration inequalities [Mua+14]. The cost of computing $\kappa\left(\hat{\pi}_n, \hat{\pi'}_{n'}\right)$ is $\mathcal{O}(nn')$, which can be prohibitive when the numbers of elements are too large. A potential solution to this computational bottleneck is to leverage techniques such as Random Feature expansions, as we will describe in the next section.

MMD. The mean kernel naturally defines a seminorm on $\mathfrak{M}$:

$$\|\mu\|_\kappa^2 = \kappa(\mu, \mu) \tag{1.23}$$

often referred to as the Maximum Mean Discrepancy (MMD) [Bor+06; Gre+06]. A large body of work has been devoted to derive conditions under which $\|\cdot\|_\kappa$ is a proper norm. Gretton et al. [Gre+06] introduce the concept of *universal* kernels κ when $\mathfrak{Z}$ is compact, while Sriperumbudur et al. [Sri+10] later extends this notion to that of *characteristic* kernels. In particular, in the latter the authors derive easily checkable conditions when κ is a translation-invariant kernel.

Translation-invariant kernels. When $\mathfrak{Z}$ is a vector space, a translation-invariant kernel is defined as $\kappa(\mathbf{z}, \mathbf{z}') = K(\mathbf{z} - \mathbf{z}')$, for some function $K : \mathfrak{Z} \to \mathbb{C}$. When $\mathfrak{Z} \subseteq \mathbb{R}^d$, the function K is characterized by Bochner's Theorem.

Theorem 1.4.2 (Bochner [Rud62], Thm. 1.4.3)**.** *Consider $K : \mathbb{R}^d \to \mathbb{C}$. A kernel of the form $\kappa(x, y) = K(x - y)$ is positive definite if and only if there exists a* probability distribution *$\Lambda \in \mathfrak{P}(\mathbb{R}^d)$ such that:*

$$K(x) = \kappa(0, 0) \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^\top x} d\Lambda(\boldsymbol{\omega}) . \tag{1.24}$$

[9] See definition in Appendix A.

Example 1.4.3. *For a Gaussian kernel* $\kappa(\mathbf{z}, \mathbf{z}') = e^{-\frac{1}{2}\mathbf{z}^\top \mathbf{\Sigma}^{-1}\mathbf{z}}$*, the frequency probability distribution is a Gaussian* $\Lambda = \mathcal{N}\left(0, \mathbf{\Sigma}^{-1}\right)$.

We can therefore express the kernel as an expectation:

$$\kappa(\mathbf{z}, \mathbf{z}') = \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda}\left(\phi_{\boldsymbol{\omega}}(\mathbf{z})\overline{\phi_{\boldsymbol{\omega}}(\mathbf{z}')}\right) , \tag{1.25}$$

where $\phi_{\boldsymbol{\omega}}(\mathbf{z}) = \sqrt{\kappa(0,0)}e^{\mathrm{i}\boldsymbol{\omega}^\top\mathbf{z}}$.

Using this expression on the mean kernel, we get the following:

$$\begin{aligned}
\kappa(\mu, \mu') &= \iint_{\mathfrak{Z}} \kappa(\mathbf{z}, \mathbf{z}')d\mu(\mathbf{z})d\mu'(\mathbf{z}') \\
&= \iint_{\mathfrak{Z}} \int_{\mathbb{R}^d} \phi_{\boldsymbol{\omega}}(\mathbf{z})\overline{\phi_{\boldsymbol{\omega}}(\mathbf{z}')}d\Lambda(\boldsymbol{\omega})d\mu(\mathbf{z})d\mu'(\mathbf{z}') \\
&= \int_{\mathbb{R}^d}\left(\int_{\mathfrak{Z}} \phi_{\boldsymbol{\omega}}(\mathbf{z})d\mu(\mathbf{z})\right)\overline{\left(\int_{\mathfrak{Z}} \phi_{\boldsymbol{\omega}}(\mathbf{z})d\mu(\mathbf{z})\right)}d\Lambda(\boldsymbol{\omega}) = \int_{\mathbb{R}^d} \psi_\mu(\boldsymbol{\omega})\overline{\psi_{\mu'}(\boldsymbol{\omega})}d\Lambda(\boldsymbol{\omega}) ,
\end{aligned} \tag{1.26}$$

where we denote

$$\psi_\mu(\boldsymbol{\omega}) := \int_{\mathfrak{Z}} \phi_{\boldsymbol{\omega}}(\mathbf{z})d\mu(\mathbf{z}) . \tag{1.27}$$

For a probability distribution $\pi \in \mathfrak{P}$, the function ψ_π is the *characteristic function* of π (multiplied by a constant).

We therefore have

$$\|\mu\|_\kappa^2 = \int_{\mathbb{R}^d} |\psi_\mu(\boldsymbol{\omega})|^2 \, d\Lambda(\boldsymbol{\omega}) . \tag{1.28}$$

Using this expression, we can characterize translation-invariant *characteristic* kernels, *i.e.* kernels for which $\|\cdot\|_\kappa$ is a proper norm on $\mathfrak{M}$. Intuitively, if there is an open set $O \subset \mathbb{R}^d$ such that $\Lambda(O) = 0$, then a measure μ such that ψ_μ is non-zero only on O but zero elsewhere (if it exists) is such that $\|\mu\|_\kappa = 0$ and $\mu \neq 0$, therefore $\|\cdot\|_\kappa$ is not a proper norm, and conversely. This is formalized by the notion of *support* of a measure[10]. The following Theorem is found in [Sri+10].

> **Theorem 1.4.4** (Sriperumbudur et al. [Sri+10], Theorem 9)**.** *A translation-invariant kernel* κ *is characteristic if and only if* $supp(\Lambda) = \mathbb{R}^d$*, where* Λ *is defined by* (1.24).

All explicit kernels considered in this thesis are translation-invariants, and therefore it is easy to check if they are characteristic kernels. However, we emphasize that **none of the results presented in this thesis technically requires the kernel to be characteristic** (of course, one could argue that estimation guarantees with respect to the MMD are only meaningful when the MMD is a proper metric, but one could also envision tasks involving a particular loss function for which this is not required).

Let us now turn to another mathematical tool that will be useful for our analysis, Random Feature (RF) expansions of kernels.

1.4.3 Random Feature expansions

Traditional kernel methods on a collection of items $\mathbf{z}_1, ..., \mathbf{z}_n$ require the computation and storage of the so-called Gram matrix $\mathbf{K} = (\kappa(\mathbf{z}_i, \mathbf{z}_j))_{ij}$, which scales in $\mathcal{O}\left(n^2\right)$. This cost becomes quickly prohibitive when the number of points n grows large.

A typical approach to address this problem is by replacing the Gram matrix with an approximate surrogate, for instance a low-rank approximation [WS01; PAB16]. This is typically referred to as "Nyström's method" and exploits, for instance, the properties of some subsampling schemes.

[10] See definition in Appendix A.

Another method consists in approximating the kernel κ by a traditional inner product between finite-dimensional mappings of the samples: $\kappa(\mathbf{z}, \mathbf{z}') \approx \Phi(\mathbf{z})^\top \overline{\Phi(\mathbf{z}')}$ where $\Phi : \mathfrak{Z} \to \mathbb{C}^m$, then perform kernel methods as traditional linear methods on the vectors $\Phi(\mathbf{z}_i)$, which can be much faster, as it usually scales linearly in the number of items n.

Random Fourier Features. Drawing the map Φ *at random* has been a popular idea in the last decade. One of the most used method of this kind is due to Rahimi and Recht [RR07; RR09], and is usually referred to as *Random Fourier Features* (RFF). It is based on Bochner's Theorem (Thm. 1.4.2), and the expression of the kernel as an expectation (1.25). Random Fourier Features consist in drawing m frequencies $\boldsymbol{\omega}_j \in \mathbb{R}^d$ *i.i.d.* from Λ, and defining the finite dimensional map[11]

$$\Phi(\mathbf{z}) := \frac{1}{\sqrt{m}} \left[\phi_{\boldsymbol{\omega}_j}(\mathbf{z})\right]_{j=1}^m \ . \tag{1.29}$$

Then, by the Law of Large Numbers, we immediately see that

$$\Phi(\mathbf{z})^\top \overline{\Phi(\mathbf{z}')} = \frac{1}{m} \sum_{j=1}^m \phi_{\boldsymbol{\omega}_j}(\mathbf{z}) \overline{\phi_{\boldsymbol{\omega}_j}(\mathbf{z}')} \approx \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda} \left(\phi_{\boldsymbol{\omega}}(\mathbf{z}) \overline{\phi_{\boldsymbol{\omega}}(\mathbf{z}')} \right) \overset{(1.25)}{=} \kappa(\mathbf{z}, \mathbf{z}') \ ,$$

which is the desired approximation. In the original paper [RR07], the authors show that when $\mathfrak{Z}$ is a compact subset of $\mathbb{R}^d$, the approximation error can be controlled uniformly over it. Refined results on the convergence rate of the approximation error of RFF are derived in [SS07; SS15].

Other Random Features. Since the original paper [RR07], the field of random features for kernel approximation has considerably grown. Beside the original RFFs, other random maps (that we will designate under the general term *Random Feature (RF) expansions*) have been designed to increase the precision of the mapping, accelerate their computation, or approximate other types of kernels. In [Xin+16; CRW17], RFFs are orthogonalized to maximize their effectiveness. In [LSS13; Yan+15; CV16], fast procedures for computing RF expansions are derived, using structured fast transforms, while in [Saa+16] Random Features corresponding to a particular kernel are computed almost instantaneously by means of an optical device. Random Features are also derived for other type of kernels such as additive kernels [VZ12] or polynomial kernels [PY15]. Finally, some approaches *learn* an RF expansion in a supervised manner [WA13; Yan+15; Muk16; SD16].

Use in kernel mean embedding. When we combine Random Fourier Features $\kappa(\mathbf{z}, \mathbf{z}') \approx \Phi(\mathbf{z})^\top \overline{\Phi(\mathbf{z}')}$ with kernel mean, we obtain exactly the sketching operator of Bourrier et al. (Section 1.2) that randomly samples the characteristic function of a distribution (with the additional normalization $m^{-\frac{1}{2}}$ here). Indeed, define Φ as (1.29) where the frequencies are drawn *i.i.d.* from Λ, and recall the sketching operator (1.7) defined as $\mathcal{A}\pi = \mathbb{E}_{\mathbf{z}\sim\pi} \Phi(\mathbf{z})$. For two probability distributions π, π' we have:

$$\begin{aligned}
(\mathcal{A}\pi)^\top \overline{(\mathcal{A}\pi')} &= \frac{1}{m} \sum_{j=1}^m \mathbb{E}_{\mathbf{z}\sim\pi} \phi_{\boldsymbol{\omega}_j}(\mathbf{z}) \overline{\mathbb{E}_{\mathbf{z}'\sim\pi'} \phi_{\boldsymbol{\omega}_j}(\mathbf{z}')} \approx \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda} \left[\mathbb{E}_{\mathbf{z}\sim\pi} \phi_{\boldsymbol{\omega}}(\mathbf{z}) \overline{\mathbb{E}_{\mathbf{z}'\sim\pi'} \phi_{\boldsymbol{\omega}}(\mathbf{z}')} \right] \\
&= \mathbb{E}_{\mathbf{z}'\sim\pi'} \mathbb{E}_{\mathbf{z}\sim\pi} \left[\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda} \phi_{\boldsymbol{\omega}}(\mathbf{z}) \overline{\phi_{\boldsymbol{\omega}}(\mathbf{z}')} \right] \overset{(1.25)}{=} \mathbb{E}_{\mathbf{z}'\sim\pi'} \mathbb{E}_{\mathbf{z}\sim\pi} \kappa(\mathbf{z}, \mathbf{z}') \overset{(1.21)}{=} \kappa(\pi, \pi')
\end{aligned}$$

This connection between the sketching operator and the geometry induced by the kernel mean embedding via Random Fourier Features is the starting point of the theoretical analysis presented in this thesis.

Structures combining RFFs with mean kernel embedding of probability distributions have also been recently used by the kernel community [BS09; Sut+15; OSS15] to accelerate methods

[11] In its original form [RR07], the RFF mapping is slightly different from (1.29), and is expressed as $\Phi(\mathbf{z}) = \left[\cos(\boldsymbol{\omega}_j^\top \mathbf{z} + b_j)\right]_{j=1,m}$, where the phases b_j's are drawn *i.i.d.* uniformly from $[0; 2\pi]$. One can show that when m grows large this mapping also approximates the kernel κ, however we choose the expression (1.29) to later link RFF with the sketch defined by Bourrier [BGP13]. An extensive comparison between the two mappings is done in [SS07].

such as classification with the so-called Support Measure Machine [Mua+12; Sut+15; OSS15] or two-sample test [ZM15; Chw+15; Jit+16; PSW16]. A theoretical analysis of the approximation power of RF expansions for kernel mean embedding is performed in [Jos+10], and a connection with quadrature rules is done in [Bac15].

1.5 Layout of the manuscript

The main goal of this Ph.D. thesis is to extend the sketching method of Bourrier beyond mixture of Gaussians with identity covariance, and provide a theoretical analysis of the method. We extend the heuristic method to general mixture of distributions with a new robust algorithm coined Compressive Learning-Orthogonal Matching Pursuit with Replacement (CL-OMPR), and apply it on three mixture models on synthetic and real data. A theoretical analysis of the method using tools from kernel mean embedding and Compressive Sensing is developed, bridging a gap between several distinct fields.

The layout of the manuscript is the following.

- In Chapter 2, we present a first contribution that is independent of sketching. Considering the general inverse problem framework of Section 1.3, we extend the results on the Restricted Isometry Property (RIP) and its link with the existence of an instance optimal decoder, in several ways.
  - we extend the results to measurement processes that are *not necessary linear*, and signals that do not necessarily live in a vector space but in any metric set;
  - since the exact LRIP is sometimes difficult to prove, we let the possibility of a small additive error $\eta \geq 0$ in its formulation, as is sometimes done in quantization [BRM15], and show that instance optimality is degraded by only this small error;
  - in a second part we assume that the measurement process is drawn at random and extend the results to *non-uniform recovery*: a given signal is recovered with high probability, in opposition to recovering all signals with high probability. A non-uniform version of the LRIP is formulated, which to our knowledge is entirely new.

  While all these new results are not, strictly speaking, necessary for analyzing the sketch procedure (in particular, the sketching operator is linear), they may prove very useful in the future for characterizing complex measurement or sketching processes (*e.g.* neural networks).
- In Chapter 3, we introduce the main analysis of the sketching method. We describe a framework in which the sketching operator is built from a RF expansion of a kernel, and aim at providing information-preservation guarantees. More precisely:
  - we prove that under two conditions the sketching operator satisfies the non-uniform LRIP with high probability, with respect to the MMD. The first condition is the domination between certain metrics and will be referred to as *admissibility* of the RF expansion of the kernel, and the second condition is finiteness of the covering numbers of some normalized secant set.
  - In a second part, we temporarily relax these strong assumptions and only assume finiteness of the covering numbers of the low-dimensional model $\mathfrak{S}$ itself, which is often relatively easy to prove. In this case we prove preliminary reconstruction results, in which the sketch needs to be as big as the original database to attain a fixed error level. These first results therefore do not guarantee that the sketching method is in that case more efficient than using the full data. However they still yield new guarantees for some models, and are useful to familiarize oneself with the mathematical tools. It is seen in the rest of the manuscript that these results are fortunately largely pessimistic compared to what is observed in practice.
  - We apply this preliminary framework with relaxed assumptions to two mixture models. In particular, the second considered model is that of mixtures of multivariate elliptic stable distributions, for which, to our knowledge, there is no estimator in the literature.

Some of these results have been published in [Ker+17b].

- In Chapter 4 we propose a greedy algorithm for learning generic mixture models from a sketch. In fact we treat general cost functions where a measurement vector $\mathbf{y}$ is approached, for the Euclidean norm, by a linear combination of atoms in a dictionary $\{\mathbf{f}(\boldsymbol{\theta})\}$ continuously indexed by $\boldsymbol{\theta}$, where $\mathbf{f}$ is differentiable. This is later applied to the sketching method by taking $\mathbf{f}(\boldsymbol{\theta}) = \mathcal{A}\pi_{\boldsymbol{\theta}}$.
  - We start by proving that the cost function is sometimes locally block convex when sufficiently close to the optimum. This encourages us to apply a simple Block Coordinate Descent algorithm, that indeed works with a good initialization. However this algorithm fails in the absence of prior knowledge.
  - We then propose a greedy approach inspired by OMP, modified to alternate between adding an atom and non-convex updates. We give a second variant with Replacement that incorporates the capacity to suppress spurious atoms by Hard Thresholding. These two algorithms are respectively coined Compressive Learning-OMP (CL-OMP) and CL-OMP with Replacement (CL-OMPR). We give details to facilitate and optimize their practical implementation.
  - We apply these algorithms to a first artificial problem independent of sketching. The Block Coordinate Descent indeed only succeeds when initialized close to the true solution, while CL-OMPR succeeds most of the time even without prior knowledge.

  The CL-OMPR algorithm is found in [Ker+16; Ker+17b; Ker+17a], and the code is available as a Matlab toolbox at [Ker16].

- In Chapter 5 the CL-OMPR algorithm is applied to the practical estimation of mixture models from a sketch. We start by describing an unsupervised learning method to learn an appropriate kernel for the method, from a fraction of training data. We then implement the method for three mixture models, illustrated in Fig. 1.3.
  - First we instantiate the CL-OMPR algorithm to recover mixture of Diracs from a sketch. Although the true distribution of the data is obviously not a mixture of Diracs, we show empirically that when data are well-clustered the locations of the recovered Diracs correspond to the center of these clusters, also called *centroids*. The method is therefore compared to a classic method for unsupervised clustering, the k-means algorithm. The sketching method is shown to be significantly faster and more memory efficient than k-means on large databases, and more stable to initialization. The algorithms are also compared on a spectral clustering task for handwritten digits recognition with the MNIST database [LCB98]. This approach is published in [Ker+17a].
  - The CL-OMPR algorithm is then applied to the estimation of GMMs with unknown diagonal covariance. It is shown to be more efficient than classic Expectation Maximization (EM) on large databases. The algorithm are compared on a speaker verification task, using the NIST05 database. These results are published in [Ker+16; Ker+17b].
  - Finally, the method is instantiated to estimate mixtures of multivariate elliptic stable distributions. To our knowledge, we obtain the only algorithm capable of performing such a task in the multivariate case. At the time of this manuscript these results are not yet published.

  In these three cases, we empirically observe on synthetic data that the sketch size necessary for the success of the estimation is as $m \approx \mathcal{O}(kd)$, where k is the number of components in the mixture and d is the dimension of the data, and independent of the number of points n in the original database. It confirms that the preliminary theoretical results obtained at the end of Chapter 3 are indeed sub-optimal.

- In Chapter 6, we greatly extend the theoretical analysis of the method, with a double purpose in mind: proving that the size of the sketch only depends on the complexity of

the model $\mathfrak{S}$ (as observed in practice), and relating the results to classic learning costs instead of the MMD, which is sometimes difficult to analyze.

- We develop a more advanced analysis of generic mixture models that fully exploits the general results presented in the first half of Chapter 3. It is based on the key restriction of the model to mixtures where components are *pairwise sufficiently separated*. Under this assumption, when two mixtures are close to each other, each component from the first mixture can be paired with a single component from the second and separated from all other components. Two such components form what we call a "dipole". Under some assumptions on the kernel, those dipoles can be treated independently, and we are able to prove "strong" versions of the admissibility condition and finiteness of the covering numbers of the normalized secant set.
- We first apply this analysis to the clustering problem by recovering mixtures of Diracs as considered in the experiments. We provide guarantees with respect to the k-means and k-medians cost. We prove that the estimation is feasible for a sketch size approximately as $m \approx \mathcal{O}\left(k^2 d^2\right)$ (up to logarithmic terms), which is indeed independent of the size of the database n, but still slightly sub-optimal compared to what is observed in practice.
- We then instantiate the analysis method to GMMs with known covariance. Results are given with respect to classic log-likelihood cost function. The necessary sketch size is as $m \approx \mathcal{O}\left(k^2 d^2 C\right)$, where C is a constant that varies with the imposed separation ε of the means between components in the model of GMMs. At one end of the spectrum, we get $C = \mathcal{O}(1)$ for a relatively large separation in $\varepsilon = \mathcal{O}\left(\sqrt{d \log k}\right)$, while at the other end we have a large $C = \mathcal{O}\left(e^d\right)$ but for a separation $\varepsilon = \mathcal{O}\left(\sqrt{\log k}\right)$ that compares favorably to existing literature.

We finish by conclusions and outlooks in Chapter 7.

In Appendix A we group some definitions and generic results. Appendix B contains the proof of Chapter 3. Appendices C, D and E contain the proofs of the results of Chapter 6.

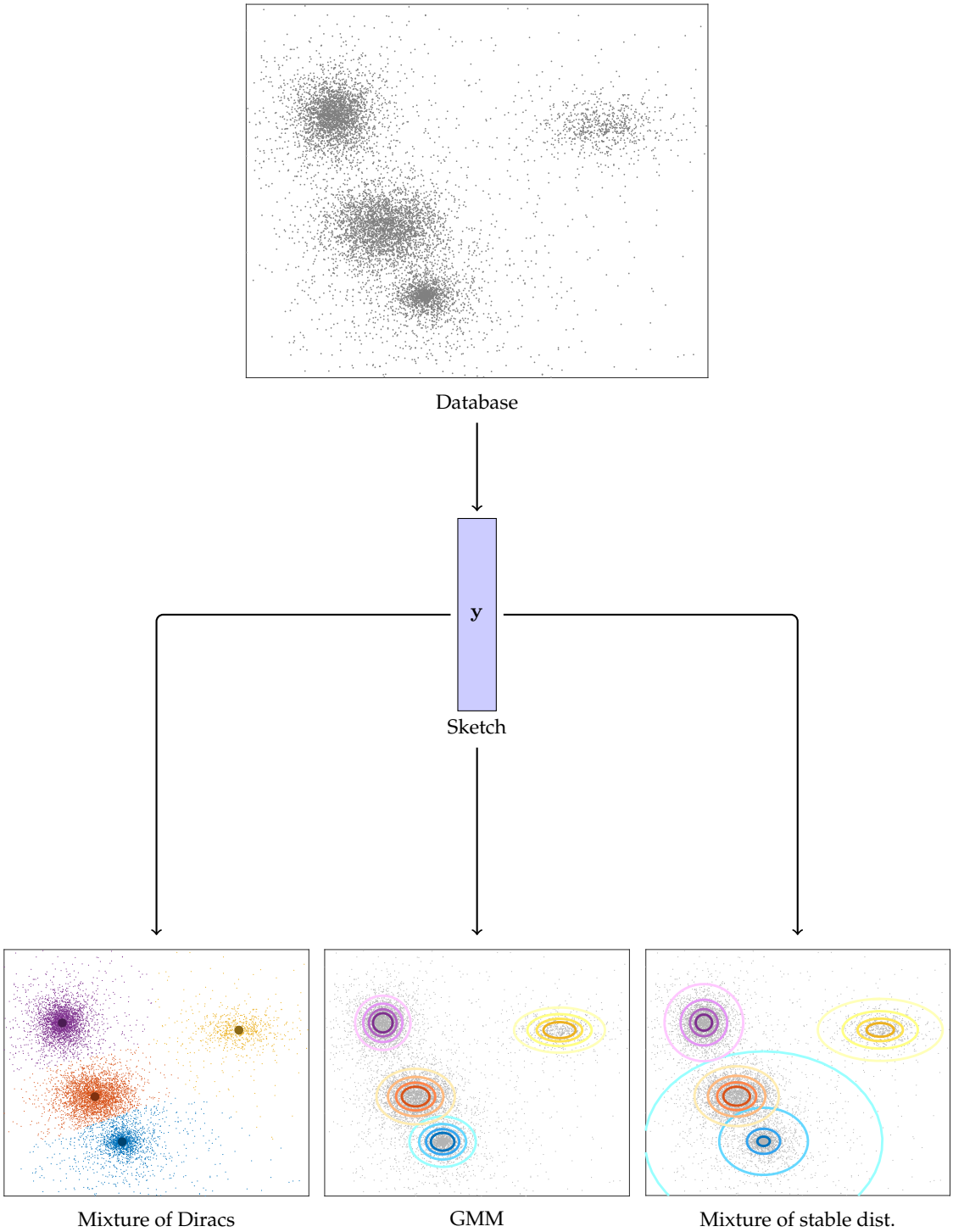

FIGURE 1.3: **Illustration of the three sketched mixture learning schemes implemented in Chapter 5.** From the same sketch, we derive either a mixture of Diracs (which is comparable to unsupervised clustering methods such as k-means), a GMM, or a mixture of elliptic stable distributions.

Chapter 2

Robust Decoding for Generalized Inverse Problems

This small chapter contains a first contribution that is not directly related to sketching, but will be used in the main sketching framework (Chapter 3). This contribution extends the results of [CDD09; Bou+14] on robust decoding in generic linear inverse problems. In particular, we generalize the results to *non-linear* inverse problems, which may be of paramount interest in the future for modern embedding architectures such as neural networks.

In Chapter 1 Section 1.3, we have reviewed the results of [Bou+14], where the authors show that the existence of a robust decoder is equivalent to the linear measurement process satisfying the Lower Restricted Isometry Property (LRIP). In this chapter these results are extended in several ways:

- we show that the equivalence between LRIP and the existence of an instance optimal decoder is still valid when the measurement process is non-linear;
- we do not assume that the signals live in a vector space, but any set;
- we allow for an additional error $\eta \geq 0$ in the LRIP and the decoding, *i.e.* when the LRIP is not exactly satisfied;
- when the measurement process is random, we extend the result to a *non-uniform* formulation similar to [CDD09]. We show that a non-uniform LRIP implies a non-uniform Instance Optimality Property, which is strictly less restrictive than the results presented in [CDD09].

All throughout the chapter, we call *signal* the measured object $\mathbf{x}$, since the RIP analysis originated from Compressive Sensing. However we emphasize that $\mathbf{x}$ can be *any* mathematical object, and that we are considering generalized inverse problems here.

We begin by some notations in Section 2.1, then turn to guarantees that are uniform in probability in Section 2.2. We then treat of non-uniform guarantees in Section 2.3.

2.1 Framework

Let us begin by introducing the problem.

2.1.1 Inverse problem

Let E be an arbitrary set[1] and F be a vector space equipped with a seminorm $\|\cdot\|_F$. Consider a mapping $\Psi : E \mapsto F$, *not necessarily linear*[2], and suppose that we measure a signal $\mathbf{x}^\star \in E$ with noise $\mathbf{e} \in F$:

$$\mathbf{y} = \Psi\left(\mathbf{x}^\star\right) + \mathbf{e}. \tag{2.1}$$

[1] Notice the difference with Section 1.3 where the signal was assumed to live in a vector space.
[2] Again, the set of signals E does not even need to be a vector space equipped with an additive operation

Our goal is to estimate the signal $\mathbf{x}^\star$ from the measurement $\mathbf{y}$. As usual the measurement process is "dimension-reducing" (taken here in a very large sense since neither E nor F are required to be finite-dimensional and E is not necessarily a vector space), we therefore introduce a "low-dimensional" subset of signals $\mathfrak{S} \subset E$ that can potentially be successfully recovered from their measurements.

2.1.2 Approximate decoder

In any realistic system one can expect both modeling error and noise. In that case, we have seen in Section 1.3 that successful recovery is linked to the notion of *instance optimal decoder* [CDD09; FR13; Bou+14], *i.e.* a decoder Δ such that the recovered signal $\Delta(\Psi, \mathbf{y})$ is close to the original signal $\mathbf{x}^\star$ even when it is not exactly in the model and in the presence of noise.

As in Section 1.3, we would like to define the "ideal" decoder as

$$\Delta(\Psi, \mathbf{y}) \in \operatorname*{arg\,min}_{\mathbf{x} \in \mathfrak{S}} \|\Psi(\mathbf{x}) - \mathbf{y}\|_F \tag{2.2}$$

However, this minimization is not guaranteed to have a solution, unless we make additional assumptions on the model $\mathfrak{S}$ and measurements function Ψ, such as compactness of the model (if E is a Banach space) and continuity of the function $\mathbf{x} \mapsto \|\Psi(\mathbf{x}) - \mathbf{y}\|_F$, both with respect to some norm $\|\cdot\|_E$. However this would unnecessarily restrict our framework: k-sparse vectors would not even be included since the set of sparse vectors is not compact.

Hence, similar to [Bou+14], we state our results for a family of decoders Δ_ι where $\iota \geq 0$, such that for all Ψ and $\mathbf{y}$, it returns an element of the model $\widetilde{\mathbf{x}} = \Delta_\iota(\Psi, \mathbf{y}) \in \mathfrak{S}$ such that

$$\|\Psi(\widetilde{\mathbf{x}}) - \mathbf{y}\|_F \leq \inf_{\mathbf{x} \in \mathfrak{S}} \|\Psi(\mathbf{x}) - \mathbf{y}\|_F + \iota. \tag{2.3}$$

In other words, it returns one signal that approaches the infimum (2.2) with a precision ι. For $\iota > 0$ the decoder Δ_ι is always defined. The ideal decoder Δ_0 only exists if we can guarantee the existence of a projection operator on the model for the appropriate metric, such that (2.2) always has a solution (not necessarily unique).

For the sake of simplicity, we will denote such an approximate infimum by

$$\Delta_\iota(\Psi, \mathbf{y}) = \operatorname*{arg\,min}_{\mathbf{x} \in \mathfrak{S},\ \text{error } \iota} \|\Psi(\mathbf{x}) - \mathbf{y}\|_F \tag{2.4}$$

and use this notation all throughout the manuscript.

While our motivation for this subtlety is mainly mathematical, we note that it can also materialize the fact that we may not be able to solve (2.2) exactly due to algorithmic limitation but only up to a precision $\iota_0 > 0$, in which case the decoder Δ_ι only exists for $\iota \geq \iota_0$.

2.2 Uniform guarantees

In this section we consider that the measurement operator Ψ is fixed, and give sufficient and necessary conditions to ensure instance optimal decoding. Inspired by the existing results presented in Section 1.3, our goal is to generalize the equivalence between the Instance Optimality Property (IOP) and the Lower Restricted Isometry Property (LRIP), schematically illustrated in Fig. 2.1.

Modified IOP and LRIP. We give our modified versions of the LRIP and IOP below. Compared to the previous definitions of the LRIP (Def. 1.3.1) and IOP (Def. 1.3.3), their most notables features are the non-linearity of Ψ and the presence of a possible additive error $\eta \geq 0$.

Definition 2.2.1 (Lower Restricted Isometry Property). *The mapping Ψ satisfies the Lower Restricted Isometry Property (LRIP) for the model $\mathfrak{S}$ with constant $\alpha > 0$, pseudometric d_E and error $\eta \geq 0$ if: for all $\mathbf{x}, \mathbf{x}' \in \mathfrak{S}$ it holds that*

$$d_E(\mathbf{x}, \mathbf{x}') \leq \alpha \|\Psi(\mathbf{x}) - \Psi(\mathbf{x}')\|_F + \eta. \tag{2.5}$$

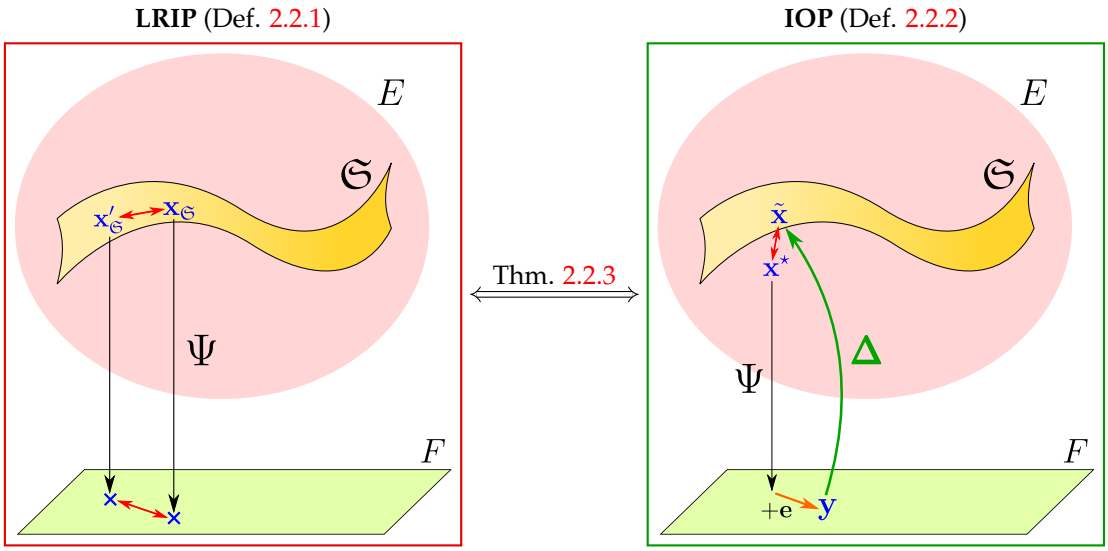

FIGURE 2.1: **Schematic illustration of the LRIP and decoder satisfying the IOP.** The LRIP (left) states that the measurement process Ψ approximately preserves the distances on the model $\mathfrak{S}$. A decoder Δ that satisfies the IOP (right) returns, from the measurement of an object, an element of the model that is almost the one closest to the original signal. The two properties are equivalent by Theorem 2.2.3.

Definition 2.2.2 (Instance Optimality Property)**.** *A decoder Δ satisfies the Instance Optimality Property for the mapping Ψ and model $\mathfrak{S}$ with constants $A, B > 0$, pseudometrics d_E, d'_E on E and error $\eta \geq 0$ if: for all signals $\mathbf{x}^\star \in E$ and noise $\mathbf{e} \in F$, denoting $\widetilde{\mathbf{x}} = \Delta(\Psi, \Psi(\mathbf{x}^\star) + \mathbf{e})$ the recovered signal, it holds that:*

$$d_E(\mathbf{x}^\star, \widetilde{\mathbf{x}}) \leq A d'_E(\mathbf{x}^\star, \mathfrak{S}) + B \|\mathbf{e}\|_F + \eta \tag{2.6}$$

where $d'_E(\mathbf{x}, \mathfrak{S}) = \inf_{\mathbf{x}_\mathfrak{S} \in \mathfrak{S}} d'_E(\mathbf{x}, \mathbf{x}_\mathfrak{S})$.

We show equivalence between IOP and LRIP in the following theorem. It has essentially the same form as Bourrier's result (Theorem 1.3.4), with the addition that Ψ is not necessarily linear and that we allow for an additive error $\eta \geq 0$.

Theorem 2.2.3 (Equivalence between IOP and LRIP.)**.** *Consider a mapping Ψ and a model $\mathfrak{S}$.*

1. *If there exists a decoder Δ which satisfies the Instance Optimality Property for Ψ and $\mathfrak{S}$ with constants $A, B > 0$, pseudometrics d_E, d'_E and error $\eta \geq 0$, then the mapping Ψ satisfies the LRIP for the model $\mathfrak{S}$ with constant $\alpha = B$, pseudometric d_E and error 2η.*
2. *If the mapping Ψ satisfies the LRIP for the model $\mathfrak{S}$ with constant α, pseudometric d_E and error $\eta \geq 0$, then the decoder Δ_ι defined by 2.4 satisfies the Instance Optimality Property for the mapping Ψ and model $\mathfrak{S}$ with constants $A = 1$ and $B = 2\alpha$, pseudometrics d_E and d'_E where d'_E is defined by $d'_E(\mathbf{x}, \mathbf{x}') := d_E(\mathbf{x}, \mathbf{x}') + 2\alpha \|\Psi(\mathbf{x}) - \Psi(\mathbf{x}')\|_F$, and error $\eta + \alpha\iota$.*

Proof. 1. Consider $\mathbf{x}, \mathbf{x}' \in \mathfrak{S}$. By triangular inequality we have

$$d_E(\mathbf{x}, \mathbf{x}') \leq d_E(\mathbf{x}, \Delta(\Psi, \Psi(\mathbf{x}'))) + d_E(\Delta(\Psi, \Psi(\mathbf{x}')), \mathbf{x}').$$

Then, by applying the Instance Optimality Property with noise $\mathbf{e} := \Psi(\mathbf{x}') - \Psi(\mathbf{x})$ we get $d_E(\mathbf{x}, \Delta(\Psi, \Psi(\mathbf{x}'))) \leq B \|\Psi(\mathbf{x}') - \Psi(\mathbf{x})\|_F + \eta$, and by applying again the Instance

Optimality Property it holds that $d_E(\Delta(\Psi, \Psi(\mathbf{x}')), \mathbf{x}') \leq \eta$, hence the result.

2. Consider any signal $\mathbf{x}^\star \in E$ and noise $\mathbf{e} \in F$, denote $\mathbf{y} = \Psi(\mathbf{x}^\star) + \mathbf{e}$ and $\widetilde{\mathbf{x}} = \Delta_\iota(\Psi, \mathbf{y})$. Let $\mathbf{x}_\mathfrak{S} \in \mathfrak{S}$ be any element of the model. We have:

$$\begin{aligned} d_E(\mathbf{x}^\star, \widetilde{\mathbf{x}}) &\leq d_E(\mathbf{x}^\star, \mathbf{x}_\mathfrak{S}) + d_E(\mathbf{x}_\mathfrak{S}, \widetilde{\mathbf{x}}) \\ &\overset{\text{LRIP}}{\leq} d_E(\mathbf{x}^\star, \mathbf{x}_\mathfrak{S}) + \alpha \left\| \Psi(\mathbf{x}_\mathfrak{S}) - \Psi(\widetilde{\mathbf{x}}) \right\|_F + \eta \\ &\leq d_E(\mathbf{x}^\star, \mathbf{x}_\mathfrak{S}) + \alpha \left\| \Psi(\mathbf{x}_\mathfrak{S}) - \mathbf{y} \right\|_F + \alpha \left\| \mathbf{y} - \Psi(\widetilde{\mathbf{x}}) \right\|_F + \eta \,. \end{aligned}$$

By definition of the decoder (2.4) we have $\left\| \Psi(\widetilde{\mathbf{x}}) - \mathbf{y} \right\|_F \leq \left\| \Psi(\mathbf{x}_\mathfrak{S}) - \mathbf{y} \right\|_F + \iota$ and therefore

$$\begin{aligned} d_E(\mathbf{x}^\star, \widetilde{\mathbf{x}}) &\leq d_E(\mathbf{x}^\star, \mathbf{x}_\mathfrak{S}) + 2\alpha \left\| \Psi(\mathbf{x}_\mathfrak{S}) - \mathbf{y} \right\|_F + \eta + \alpha\iota \\ &\leq d_E(\mathbf{x}^\star, \mathbf{x}_\mathfrak{S}) + 2\alpha \left\| \Psi(\mathbf{x}_\mathfrak{S}) - \Psi(\mathbf{x}^\star) \right\|_F + 2\alpha \left\| \Psi(\mathbf{x}^\star) - \mathbf{y} \right\|_F + \eta + \alpha\iota \\ &\leq d'_E(\mathbf{x}^\star, \mathbf{x}_\mathfrak{S}) + 2\alpha \left\| \mathbf{e} \right\|_F + \eta + \alpha\iota \end{aligned}$$

where $d'_E(\mathbf{x}, \mathbf{x}') = d_E(\mathbf{x}, \mathbf{x}') + 2\alpha \left\| \Psi(\mathbf{x}) - \Psi(\mathbf{x}') \right\|_F$. Since the result is valid for all $\mathbf{x}_\mathfrak{S} \in \mathfrak{S}$, we can take the infimum of $d'_E(\mathbf{x}, \mathbf{x}_\mathfrak{S})$ with respect to $\mathbf{x}_\mathfrak{S} \in \mathfrak{S}$ and obtain the result.

□

We now turn to our non-uniform formulation of the result.

2.3 Non-uniform guarantees

As we have seen, in Compressive Sensing the mapping Ψ is usually drawn at random. Most results state that with high probability on the drawing of the mapping Ψ, the LRIP holds and the decoder (2.4) is instance optimal. This type of guarantee is referred to as *uniform* in probability: with high probability on the mapping, the equation (2.6) is simultaneously verified for *all* signals $\mathbf{x}^\star$ and noise vectors $\mathbf{e}$. This is opposed to a *non-uniform* guarantee: for a *given* signal $\mathbf{x}^\star$, with high probability on the mapping Ψ, recovery of $\mathbf{x}^\star$ is guaranteed. This kind of non-uniform recovery result is also studied in classic Compressive Sensing [FR13], and it is known that some algorithms can only yield non-uniform recovery guarantees [Rau08].

In this thesis the non-uniform results are focused on the "LRIP implies IOP" implication (as discussed at the end of the section the converse implication seems to be less direct).

Boundedness property, non-uniform LRIP. First of all, it can be seen that the mapping Ψ itself intervenes in the definition of the metric d'_E on the right-hand side of (2.6), which is potentially undesirable since it is random. In [CDD09], the authors introduce a "boundedness property" (BP) to address this problem, of which we give a slightly modified version below. The authors in [CDD09] then show (in the usual finite-dimensional framework of Compressive Sensing) that a *uniform* LRIP and the non-uniform Boundedness Property implies a *non-uniform* IOP. We extend this result to our generalized framework where in particular the mapping Ψ is not required to be linear, and most importantly we prove that a *non-uniform* LRIP and a non-uniform boundedness property implies a non-uniform IOP, which is strictly stronger than the results of [CDD09].

Our definitions of the non-uniform Boundedness Property and non-uniform LRIP are as follows.

Definition 2.3.1 (Boundedness property)**.** *The mapping Ψ satisfies the (non-uniform) Boundedness Property for a signal $\mathbf{x} \in E$, a model $\mathfrak{S}$ and a fixed element of the model $\mathbf{x}_\mathfrak{S} \in \mathfrak{S}$ with constant β, pseudometric d_G and probability $1 - \rho$ if: with probability at least $1 - \rho$ on Ψ, we have*

$$\left\| \Psi(\mathbf{x}) - \Psi(\mathbf{x}_\mathfrak{S}) \right\|_F \leq \beta d_G(\mathbf{x}, \mathbf{x}_\mathfrak{S}) \,. \tag{2.7}$$

Definition 2.3.2 (Non-uniform LRIP). *The mapping Ψ satisfies the non-uniform LRIP for the model $\mathfrak{S}$ and a fixed element of the model $\mathbf{x}_{\mathfrak{S}} \in \mathfrak{S}$, with constant $\alpha > 0$, pseudometric d_E, probability $1-\rho$ and error $\eta \geq 0$ if: with probability at least $1-\rho$ on Ψ, for all $\mathbf{x}'_{\mathfrak{S}} \in \mathfrak{S}$ we have*

$$d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}'_{\mathfrak{S}}) \leq \alpha \left\| \Psi\left(\mathbf{x}_{\mathfrak{S}}\right) - \Psi\left(\mathbf{x}'_{\mathfrak{S}}\right) \right\|_F + \eta. \tag{2.8}$$

Remark 2.3.3 ("Semi"-uniformity of the LRIP.). *Note that our definition of the LRIP is not "fully" non-uniform but "semi"-uniform in probability: it is non-uniform with respect to the first signal $\mathbf{x}_{\mathfrak{S}} \in \mathfrak{S}$, which is given* a priori, *but uniform with respect to the second signal $\mathbf{x}'_{\mathfrak{S}} \in \mathfrak{S}$. The necessity of this unusual formulation can be understood intuitively by examining the proof of the uniform case (Theorem 2.2.3). When the LRIP is used in the proof, the first signal $\mathbf{x}_{\mathfrak{S}} \in \mathfrak{S}$ can be indeed fixed before drawing Ψ, however the second one is $\mathbf{x}'_{\mathfrak{S}} = \widetilde{\mathbf{x}} = \Delta(\Psi, \mathbf{y})$ the recovered signal, which is* itself *random since it depends on the mapping Ψ. Therefore it cannot be fixed before the drawing of Ψ.*

In the same fashion as the LRIP, the considered non-uniform IOP is non-uniform with respect to the signal $\mathbf{x} \in E$, but uniform with respect to the noise $\mathbf{e} \in F$. It is defined as follows.

Definition 2.3.4 (Non-uniform IOP). *A decoder Δ satisfies the non-uniform Instance Optimality Property for the (random) mapping Ψ, model $\mathfrak{S}$, signal $\mathbf{x}^\star \in E$ and element of the model $\mathbf{x}_{\mathfrak{S}} \in \mathfrak{S}$ with constants $A, B > 0$, pseudometrics d_E, d'_E, probability $1-\rho$ and error $\eta \geq 0$ if: with probability at least $1-\rho$ on the mapping Ψ, for all noise $\mathbf{e} \in F$, denoting $\widetilde{\mathbf{x}} = \Delta(\Psi, \Psi\left(\mathbf{x}^\star\right) + \mathbf{e})$ it holds that:*

$$d_E(\mathbf{x}^\star, \widetilde{\mathbf{x}}) \leq A d'_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + B \left\|\mathbf{e}\right\|_F + \eta \tag{2.9}$$

Remark 2.3.5 (Bias term.). *In this definition, the distance from $\mathbf{x}^\star$ to the model is replaced by its distance to* a particular *element of the model $\mathbf{x}_{\mathfrak{S}}$ fixed* a priori. *Indeed, because of the "non-uniform" flavor of the IOP, we cannot prove that the bound holds uniformly for all $\mathbf{x}_{\mathfrak{S}} \in \mathfrak{S}$ and perform a minimization* a posteriori, *as in the proof of Theorem 2.2.3. Ideally, one would therefore like to choose $\mathbf{x}_{\mathfrak{S}}$ such that $d'_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) = \inf_{\mathbf{x} \in \mathfrak{S}} d'_E(\mathbf{x}^\star, \mathbf{x}) = d'_E(\mathbf{x}^\star, \mathfrak{S})$. However, as mentioned before, such an element $\mathbf{x}_{\mathfrak{S}}$ does not always exist. We therefore leave the term $d'_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}})$ in the instance optimality bound, and in some cases we will be able to exhibit a $\mathbf{x}_{\mathfrak{S}}$ to express this term more explicitly.*

In the rest of the manuscript, we will refer to this term $d'_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}})$ as "bias" term, i.e. *deviation in the recovery due to modeling error.*

Our main result is the following.

Theorem 2.3.6 (The non-uniform LRIP and BP implies the non-uniform IOP). *Consider a random mapping Ψ, a model $\mathfrak{S}$, a signal $\mathbf{x}^\star \in E$ and a fixed element of the model $\mathbf{x}_{\mathfrak{S}} \in \mathfrak{S}$. Assume that:*

i) the mapping Ψ satisfies the non-uniform LRIP for the model $\mathfrak{S}$ and element of the model $\mathbf{x}_{\mathfrak{S}}$ with constant $\alpha > 0$, pseudometric d_E, probability $1-\rho_1$ and error $\eta \geq 0$;

ii) the mapping Ψ satisfies the non-uniform Boundedness Property for $\mathbf{x}^\star$ and $\mathbf{x}_{\mathfrak{S}}$ with constant β, pseudometric d_G and probability $1-\rho_2$;

Then, the decoder Δ_ι defined by (2.3) satisfies the non-uniform Instance Optimality Property for the mapping Ψ, model $\mathfrak{S}$ signal $\mathbf{x}^\star$ and element of the model $\mathbf{x}_{\mathfrak{S}}$ with constants $A := 1$, $B := 2\alpha$, pseudometrics d_E and $d'_E := d_E + 2\alpha\beta d_G$, probability $1-\rho_1-\rho_2$ and error $\eta + \alpha\iota$.

Proof. By assumption $i)$, we apply the non-uniform LRIP on $\mathbf{x}_{\mathfrak{S}}$, and with probability at least $1-\rho_1$ on the mapping Ψ we have

$$\forall \mathbf{x}'_{\mathfrak{S}} \in \mathfrak{S},\ d_E(\mathbf{x}_{\mathfrak{S}}, \mathbf{x}'_{\mathfrak{S}}) \leq \alpha \left\| \Psi\left(\mathbf{x}_{\mathfrak{S}}\right) - \Psi\left(\mathbf{x}'_{\mathfrak{S}}\right) \right\|_F + \eta. \tag{2.10}$$

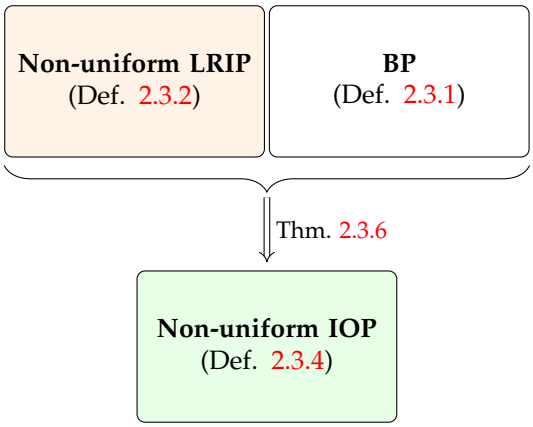

FIGURE 2.2: Illustration of Theorem 2.3.6. The non-uniform LRIP and the BP imply the non-uniform IOP.

In the same fashion, by assumption $ii)$, we apply the non-uniform Boundedness Property, and with probability at least $1-\rho_2$ on the mapping Ψ we have

$$\|\Psi(\mathbf{x}^\star) - \Psi(\mathbf{x}_{\mathfrak{S}})\|_F \leq \beta d_G(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}). \tag{2.11}$$

Therefore, by a union bound, with probability at least $1-\rho_1-\rho_2$ on the mapping Ψ both (2.10) and (2.11) are satisfied, and for all noise element $\mathbf{e} \in F$, denoting $\mathbf{y} = \Psi(\mathbf{x}^\star) + \mathbf{e}$ and $\widetilde{\mathbf{x}} = \Delta(\Psi, \mathbf{y})$, we have:

$$\begin{aligned} d_E(\mathbf{x}^\star, \widetilde{\mathbf{x}}) &\leq d_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + d_E(\mathbf{x}_{\mathfrak{S}}, \widetilde{\mathbf{x}}) \\ &\leq d_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + \alpha \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\widetilde{\mathbf{x}})\|_F + \eta \\ &\leq d_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + \alpha \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \mathbf{y}\|_F + \alpha \|\mathbf{y} - \Psi(\widetilde{\mathbf{x}})\|_F + \eta. \end{aligned}$$

by (2.10). Once again by definition of the decoder (2.4) we have $\|\Psi(\widetilde{\mathbf{x}}) - \mathbf{y}\|_F \leq \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \mathbf{y}\|_F + \iota$ and therefore

$$\begin{aligned} d_E(\mathbf{x}^\star, \widetilde{\mathbf{x}}) &\leq d_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + 2\alpha \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \mathbf{y}\|_F + \eta + \alpha\iota \\ &\leq d_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + 2\alpha \|\Psi(\mathbf{x}_{\mathfrak{S}}) - \Psi(\mathbf{x}^\star)\|_F + 2\alpha \|\Psi(\mathbf{x}^\star) - \mathbf{y}\|_F + \eta + \alpha\iota \\ &\leq d'_E(\mathbf{x}^\star, \mathbf{x}_{\mathfrak{S}}) + 2\alpha \|\mathbf{e}\|_F + \eta + \alpha\iota. \end{aligned}$$

by applying (2.11). □

We schematically illustrate this theorem in Fig. 2.2. In general, proving the non-uniform LRIP is much more difficult than showing the non-uniform BP. In this thesis, we will mostly use uniform bound $\|\Psi(\mathbf{x}) - \Psi(\mathbf{x}')\|_F \lesssim d_G(\mathbf{x}, \mathbf{x}')$ valid for any pair of signals $\mathbf{x}, \mathbf{x}'$, meaning that the BP is satisfied with probability 1.

Converse implication. Like the uniform case, it would be desirable to have a converse implication "the non-uniform IOP implies the non-uniform LRIP", to obtain equivalence between the two properties. Unfortunately, it seems that a supplementary hypothesis on the decoder is required (namely, that it decodes *uniformly* the signals that are exactly in the model) for this implication. Since this side of the implication is not crucial for our goal where we aim at proving that the sketching operator $\mathcal{A}$ satisfies the LRIP, we elected to exclude it from this manuscript. Deriving a more satisfying formulation of the necessity of the LRIP in a non-uniform context will be subject to future investigations.

2.4 Conclusion

In this chapter, we considered inverse problems for dimension-reduction measurement processes, and showed that the LRIP is necessary and sufficient for the existence of instance optimal decoders, in frameworks much more general than traditional Compressive Sensing. In particular, the signal does not need to live in a vector space but only in any metric set, and the measurement operator does not need to be linear. When this operator is designed at random we showed the LRIP can be relaxed to a non-uniform formulation, and still guarantees non-uniform instance optimal recovery. These results may have strong implications for many problems where dimension reduction processes are used, especially intricate non-linear schemes.

Chapter 3

A Framework for Sketched Learning

This chapter presents the main analysis of the sketching method in a very general framework, and constitutes the core of this manuscript. Using tools from kernel mean embedding and RF expansions, we generalize the sketching method of Bourrier et al. [BGP13] (Section 1.2) and provide recovery guarantees under general assumptions.

The chapter is divided in three parts.

- We start by describing the framework in Section 3.1. We generalize the notations and definitions of Section 1.2, and define the sketch recovery problem.
- In Section 3.2, we establish conditions for the sketching operator to verify the non-uniform LRIP. They involve a domination condition between some metrics related to the considered random features (that will be referred to as *admissibility* of the RF expansion), and a condition of finiteness of the covering numbers of the normalized secant set, similar to the classic CS settings described in Section 1.3. Our main result for the LRIP is Theorem 3.2.5. Once the LRIP and the IOP are guaranteed, the empirical error can be controlled with a generalized Hoeffding's inequality, and we obtain our main recovery result in Theorem 3.2.7.
- In Section 3.3 we formulate a first application of the Theorem 3.2.7 under weaker assumptions, namely only assuming that the model $\mathfrak{S}$ itself has finite covering numbers (instead of the normalized secant set). In that case, we obtain estimation guarantees with a precision that scales in $\mathcal{O}\left(1/\sqrt{n} + 1/\sqrt{m}\right)$, where n is the size of the original database and m is the size of the sketch. This is obviously of limited interest, since it does not prove that the method is effective with a sketch that is significantly smaller than the original database (but is fortunately seen to be largely pessimistic in practice, see Chapter 5), and will be improved upon in Chapter 6. Nevertheless, we show that this result applies to a large variety of models without requiring the advanced analysis of Chapter 6. We give two examples: first GMM with diagonal covariance, and second mixtures of multivariate elliptic stable distributions, for which, unlike GMMs, this result constitutes (to our knowledge) the only known estimator with provable guarantees.

3.1 Framework, definitions

Let κ be a p.d. kernel on a measurable space $\mathfrak{Z}$.

Remark 3.1.1. *Unlike the previous sketching framework introduced by Bourrier (Section 1.2), our results apply to any measurable space $(\mathfrak{Z}, \Sigma)$ equipped with a p.d. kernel, and not only $\mathbb{R}^d$. In practice, our experiments are still performed with $\mathfrak{Z} = \mathbb{R}^d$ (Chapter 5).*

Denote $\mathfrak{M} = \mathfrak{M}(\mathfrak{Z})$ the space of finite signed measures over $\mathfrak{Z}$, and $\mathfrak{P} \subset \mathfrak{M}$ the space of probability distributions, *i.e.* nonnegative measures π such that $\|\pi\|_{\mathrm{TV}} = 1$ (see Appendix A.1.2 for detailed definitions).

Random Feature Expansion. As mentioned in Section 1.4.3, the proposed construction of the sketching operator is based on the important notion of *Random Feature expansion* (which is a generalization of Random Fourier Features), that we formally define below.

Definition 3.1.2 (Random Feature expansion, uniformly bounded features.). *A pair $(\mathcal{F}_R, \Lambda)$ is a **Random Feature (RF) expansion** of the kernel if:*

$$\kappa(\mathbf{z}, \mathbf{z}') = \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda} \phi_{\boldsymbol{\omega}}(\mathbf{z})\overline{\phi_{\boldsymbol{\omega}}(\mathbf{z}')}\,, \tag{3.1}$$

where $\mathcal{F}_R = \{\phi_{\boldsymbol{\omega}} : \mathfrak{Z} \to \mathbb{C} \mid \boldsymbol{\omega} \in \Omega\}$ is a set of bounded continuous feature functions from $\mathfrak{Z}$ to $\mathbb{C}$ parameterized by $\boldsymbol{\omega} \in \Omega$, and Λ is a probability distribution over a measurable space Ω.

*The Random Features are said **uniformly bounded** by $B_{\mathcal{F}_R} < \infty$ if*

$$\sup_{\boldsymbol{\omega}\in\Omega, \mathbf{z}\in\mathfrak{Z}} |\phi_{\boldsymbol{\omega}}(\mathbf{z})| \leq B_{\mathcal{F}_R}\,. \tag{3.2}$$

By similarity with Random Fourier Features (that arise from Bochner's Theorem and are defined as $\phi_{\boldsymbol{\omega}}(\mathbf{z}) = e^{i\boldsymbol{\omega}^\top \mathbf{z}}$ with $\mathbf{z}, \boldsymbol{\omega} \in \mathbb{R}^d$, see Section 1.4.3), the parameters $\boldsymbol{\omega} \in \Omega$ will be called *frequencies*, even if in theory they can be very general objects. The only requirement is for Ω to be a measurable space.

For a given kernel κ there is usually an infinity of pairs $(\mathcal{F}_R, \Lambda)$ for which (3.1) is satisfied. A simple but important example is to take existing features and re-weight them.

Example 3.1.3 (Reweighting of features). *Consider a Random Features expansion of the kernel $(\mathcal{F}_R, \Lambda)$ (based on Bochner's Theorem for instance). Then, for any function $c(\boldsymbol{\omega}) > 0$ such that $C_\Lambda = \sqrt{\mathbb{E}_{\boldsymbol{\omega}\sim\Lambda} c(\boldsymbol{\omega})^2} < \infty$, we can define $\widetilde{\mathcal{F}_R} = \left\{\widetilde{\phi_{\boldsymbol{\omega}}} = C_\Lambda \phi_{\boldsymbol{\omega}}/c(\boldsymbol{\omega}) \;\middle|\; \phi_{\boldsymbol{\omega}} \in \mathcal{F}_R\right\}$ and $\widetilde{\Lambda}$ such that $d\widetilde{\Lambda}(\boldsymbol{\omega}) = c(\boldsymbol{\omega})^2\, d\Lambda(\boldsymbol{\omega})/C_\Lambda^2$. One can then easily verify that $(\widetilde{\mathcal{F}_R}, \widetilde{\Lambda})$ is also a RF expansion of the kernel κ. This reweighting of features will be of importance in Chapter 6, where some results will only be valid after re-weighting usual Random Fourier Features $\phi_{\boldsymbol{\omega}}(\mathbf{z}) = e^{i\boldsymbol{\omega}^\top \mathbf{z}}$.*

RF expansion and kernel mean embedding. In the following paragraph we recall some properties and definitions that emerge when we associate RF expansions and the kernel mean embedding methodology.

The following notations will be encountered all throughout the thesis.

- For a bounded continuous function $f : \mathfrak{Z} \to \mathbb{C}$ and a finite signed measure $\mu \in \mathfrak{M}$ we denote
$$\langle \mu, f \rangle := \int_{\mathfrak{Z}} f(\mathbf{z}) d\mu(\mathbf{z}) \in \mathbb{C}\,. \tag{3.3}$$
Similarly, for a multivariate function $\mathbf{f} : \mathfrak{Z} \to \mathbb{C}^m$, we denote $\langle \mu, \mathbf{f} \rangle := [\langle \mu, f_j \rangle]_{j=1}^m \in \mathbb{C}^m$.
- *Integral Probability Metric [Mul97].* For any family of functions $\mathcal{F} = \{f : \mathfrak{Z} \mapsto \mathbb{C}\}$, we define the seminorm
$$\|\mu\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |\langle \mu, f \rangle| \tag{3.4}$$
- *Maximum Mean Discrepancy.* For any RF expansion $(\mathcal{F}_R, \Lambda)$ of the kernel κ, recall (see Section 1.4.2) that by a simple computation the MMD $\|\mu\|_\kappa = |\kappa(\mu, \mu)|^{\frac{1}{2}}$ can be expressed as
$$\|\mu\|_\kappa^2 = \int_\Omega |\psi_\mu(\boldsymbol{\omega})|^2\, d\Lambda(\boldsymbol{\omega}) \tag{3.5}$$
where we define
$$\psi_\mu(\boldsymbol{\omega}) := \langle \mu, \phi_{\boldsymbol{\omega}} \rangle\,, \tag{3.6}$$
which can be thought of as a generalization of the characteristic function of probability distributions : the function ψ_μ is indeed the characteristic function of a probability distribution π when $\phi_{\boldsymbol{\omega}}(\mathbf{z}) = e^{i\boldsymbol{\omega}^\top \mathbf{z}}$ are Random Fourier features and $\mu = \pi$ is a probability distribution.

Note that, when the function f is a so-called "test function" the definition $\langle \mu, f \rangle$ (and all the definitions that follow) can be extended to μ being a tempered distribution instead of a finite signed measure. This will only be briefly needed in the proofs of Chapter 6, hence we do not elaborate further on the subtle interplay between measures and distributions here. See Rudin's books [Rud87; Rud91] for a complete review of these notions.

Remark 3.1.4. *Often the definitions above correspond to objects with which the reader may be more familiar. For instance, the so-called Integral Probability Metric $\|\cdot\|_{\mathcal{F}}$ will be often used when $\mathcal{F}$ is a set of feature functions $\mathcal{F} = \mathcal{F}_R = \{\phi_{\boldsymbol{\omega}} \mid \boldsymbol{\omega} \in \Omega\}$, where $(\mathcal{F}_R, \Lambda)$ is a RF expansion of a kernel κ. In that case, again denoting $\psi_\mu(\boldsymbol{\omega}) = \langle \mu, \phi_{\boldsymbol{\omega}} \rangle$, we have (in a loose sense, assuming these norms are well defined):*

$$\|\mu\|_{\mathcal{F}_R} = \sup_{\boldsymbol{\omega} \in \Omega} |\langle \mu, \phi_{\boldsymbol{\omega}} \rangle| = \|\psi_\mu\|_{L^\infty}$$
$$\|\mu\|_{\kappa} = \left(\int_\Omega |\langle \mu, \phi_{\boldsymbol{\omega}} \rangle|^2 \, d\Lambda(\boldsymbol{\omega}) \right)^{\frac{1}{2}} = \|\psi_\mu\|_{L^2(\Lambda)} \ ,$$

i.e. *these norms are respectively the L^∞ and L^2 norms of the same function. The interplay between the two will often be at the core of our analysis, and it is useful to keep in mind these representations that use the generalized characteristic function during the proofs.*

Sketching Operator. We construct the sketching operator $\mathcal{A} : \mathfrak{M} \to \mathbb{C}^m$ by combining RF expansions with the kernel mean embedding framework. Let $(\mathcal{F}_R, \Lambda)$ be an RF expansion of a kernel κ. Drawing $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m \overset{i.i.d.}{\sim} \Lambda$, define $\Phi(\mathbf{z}) := \left[\frac{1}{\sqrt{m}} \phi_{\boldsymbol{\omega}_j}(\mathbf{z}) \right]_{j=1}^m$ and

$$\mathcal{A}\mu := \langle \mu, \Phi \rangle = \frac{1}{\sqrt{m}} \left[\psi_\mu(\boldsymbol{\omega}_j) \right]_{j=1}^m \in \mathbb{C}^m \ , \tag{3.7}$$

where $\psi_\mu(\boldsymbol{\omega})$ is defined by (3.6) for the family of feature functions $\mathcal{F}_R$.

This sketching operator has a similar form compared to the one proposed by Bourrier et al. [BGP13] (Section 1.2), with generic RF expansions instead of specifically Random Fourier Features.

Useful properties. In the course of the thesis, we will make heavy use of simple domination properties between the metrics (3.4) and (3.5), summarized in the following Lemma.

Lemma 3.1.5. *For any uniformly bounded RF expansion $(\mathcal{F}_R, \Lambda)$, any draw of the sketching operator (3.7) and any finite signed measure $\mu \in \mathfrak{M}$, we have*

$$\left. \begin{matrix} \|\mu\|_\kappa \\ \|\mathcal{A}\mu\|_2 \end{matrix} \right\} \leq \|\mu\|_{\mathcal{F}_R} \leq B_{\mathcal{F}_R} \|\mu\|_{TV} \tag{3.8}$$

Proof. All inequalities are immediate, using simple manipulations of the definitions above. □

Several consequences arise from this simple Lemma:

- If the kernel is characteristic (*i.e.* $\|\cdot\|_\kappa$ is a proper norm), then $\|\cdot\|_{\mathcal{F}_R}$ is also a proper norm, since $\|\mu\|_{\mathcal{F}_R} = 0$ implies $\|\mu\|_\kappa = 0$ which implies $\mu = 0$. However, we outline that ***none* of our results require the kernel to be characteristic**[1].
- For any model $\mathfrak{S}$, the sketch operator **always satisfies the Boundedness Property** (Def. 2.3.1) with the metric $d_G := \|\cdot\|_{\mathcal{F}_R}$ and probability 1, since $\|\mathcal{A}\mu\|_2 \leq \|\mu\|_{\mathcal{F}_R}$ is always

[1] As mentioned in the introduction, one might argue that estimation guarantees with respect to the MMD are only meaningful when the MMD is a proper metric, but it is not technically required.

true. In practice this is the property we will use most of the time , and our focus will be on proving the LRIP, which combined to this always-true Boundedness Property yields the instance-optimality of the decoder Δ_ι.

– When the RFs are uniformly bounded, all (semi)norms are well-defined, since $\|\mu\|_{\mathrm{TV}}$ is finite. **In the rest of the thesis we will always consider families of RFs that are uniformly bounded.**

Sketched distribution learning: toward information-preservation guarantees. We generalize the sketching method of Bourrier et al. (Section 1.2). Given a database $\mathcal{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_n\}$ of items in $\mathfrak{Z}$ drawn *i.i.d.* from a probability distribution $\pi^\star \in \mathfrak{P}$, we compute the sketch of this database by

$$\hat{\mathbf{y}} := \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{z}_i) = \mathcal{A}\hat{\pi}_n \in \mathbb{C}^m$$

and aim at estimating $\pi^\star$ from $\hat{\mathbf{y}}$. Since the sketching operator embeds infinite-dimensional objects into a finite-dimensional space, as usual we have to recourse to a low-complexity model $\mathfrak{S} \subset \mathfrak{P}$. In practice the definition of the model often naturally emerges from the estimation task we want to perform (*e.g.* it is defined as the set of mixtures of Gaussians to perform GMM estimation). As in Chapter 2 the decoder is expressed as

$$\Delta_\iota(\mathcal{A}, \mathbf{y}) = \underset{\pi \in \mathfrak{S},\ \text{error } \iota}{\arg\min} \|\mathcal{A}\pi - \mathbf{y}\|_2 \tag{3.9}$$

with the presence of an error $\iota \geq 0$ that can be chosen as small as desired but strictly positive when the argmin is not guaranteed to exist (see Chapter 2 Section 2.1.2).

We have shown in the previous chapter that if $\mathcal{A}$ satisfies the non-uniform LRIP then this decoder verifies the non-uniform IOP. Our main contribution stems from the connection with kernel mean embedding: we have indeed

$$\|\mathcal{A}\mu\|_2^2 = \frac{1}{m}\sum_{j=1}^{m} |\psi_\mu(\boldsymbol{\omega}_j)|^2 \approx \mathbb{E}_{\boldsymbol{\omega}\sim\Lambda} |\psi_\mu(\boldsymbol{\omega})|^2 \overset{(3.5)}{=} \|\mu\|_\kappa^2$$

In particular, when $\mu = \pi - \pi'$ is a difference between probability distributions, with high probability we have

$$\|\mathcal{A}(\pi - \pi')\|_2^2 \approx \|\pi - \pi'\|_\kappa^2 \tag{3.10}$$

and this fact will be at the base of our strategy to prove that $\mathcal{A}$ satisfies the LRIP.

3.2 Information-preservation guarantees

In this section, core of the chapter, we provide information-preservation guarantees for the sketching method. Our strategy is to prove that the sketching operator satisfies the non-uniform LRIP with respect to the MMD, then use the results of our study of generalized inverse problems in Chapter 2.

We want to show that there exist constants $\alpha > 0$ and $\eta \geq 0$ such that for any fixed $\pi_{\mathfrak{S}} \in \mathfrak{S}$, with high probability on $\mathcal{A}$ (*i.e.* on the drawing of the frequencies $\boldsymbol{\omega}_j$) we have

$$\forall \pi'_{\mathfrak{S}} \in \mathfrak{S},\ \|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_\kappa \leq \alpha \|\mathcal{A}(\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}})\|_2 + \eta. \tag{3.11}$$

We will give sufficient conditions on the model $\mathfrak{S}$, the kernel κ and its RF expansion $(\mathcal{F}_R, \Lambda)$ to prove that the sketching operator satisfies the non-uniform LRIP. The proofs presented in the rest of this chapter make heavy use of the definitions and properties in Appendix A, especially the lemmas on covering numbers of Section A.3. We recommend that the reader interested in the proofs and unfamiliar with covering numbers reads this appendix first.

3.2.1 Non-uniform version of the normalized secant set

As in Section 1.3, we will see that a key object to prove the LRIP is the normalized secant set, of which we introduce a non-uniform version with error $\eta \geq 0$ below. Given a distribution in the model $\pi_{\mathfrak{S}} \in \mathfrak{S}$, it is defined as:

$$\mathcal{S}^{\eta}(\pi_{\mathfrak{S}}, \mathfrak{S}) := \left\{ \frac{\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}}{\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_{\kappa}} \;\middle|\; \pi'_{\mathfrak{S}} \in \mathfrak{S},\ \|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_{\kappa} > \eta \right\}. \tag{3.12}$$

It is the set of normalized differences of distributions $\frac{\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}}{\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_{\kappa}}$, where $\pi_{\mathfrak{S}} \in \mathfrak{S}$ is fixed and $\pi'_{\mathfrak{S}} \in \mathfrak{S}$ varies. Similar to classic LRIP proofs [Bar07] (Section 1.3), the "size" (*i.e.* the covering numbers) of this set represents the "dimensionality" of the problem and largely drives the sketch size with which we will obtain estimation guarantees.

As we will see immediately after, the additive error $\eta \geq 0$ in the LRIP is reflected in the condition $\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_{\kappa} > \eta$ in the above definition. When $\eta > 0$, we will see that it greatly facilitates the control over the behavior of $\frac{\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}}{\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_{\kappa}} \in \mathcal{S}^{\eta}(\pi_{\mathfrak{S}}, \mathfrak{S})$ since the denominator cannot go to 0. However, this comes at the price of having a potentially harmful additive error $\eta > 0$ in the LRIP and therefore in the decoding.

When $\eta > 0$ we say that the normalized secant set is *"extruded"* at level η, in the sense that it does not take into account the distributions $\pi'_{\mathfrak{S}}$ that are in a ball of radius η around $\pi_{\mathfrak{S}}$.

By Lemma 3.1.5 we have $\|\pi - \pi'\|_{\kappa} \leq \|\pi - \pi'\|_{\mathcal{F}_R} \leq B_{\mathcal{F}_R} \|\pi - \pi'\|_{\mathrm{TV}} \leq 2B_{\mathcal{F}_R}$, and therefore when $\eta \geq 2B_{\mathcal{F}_R}$ the normalized secant set is empty since the condition $\|\pi - \pi'\|_{\kappa} > \eta$ is never satisfied. Of course in reality one wishes η to be small while $B_{\mathcal{F}_R}$ is in general in $\mathcal{O}(1)$, nevertheless for simplicity in the rest of the manuscript we have to make the following (mostly technical) supposition.

To avoid trivial situations where the normalized secant set is empty, we always assume $\eta < 2B_{\mathcal{F}_R}$.

Strategy to prove the LRIP. For a fixed $\pi_{\mathfrak{S}} \in \mathfrak{S}$, *if* we can prove that with high probability on the sketching operator, for *all* $\mu \in \mathcal{S}^{\eta}(\pi_{\mathfrak{S}}, \mathfrak{S})$ it holds that

$$\|\mathcal{A}\mu\|_2 \geq \alpha^{-1} \tag{3.13}$$

for some constant $\alpha > 0$, then for all $\pi'_{\mathfrak{S}} \in \mathfrak{S}$ such that $\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_{\kappa} > \eta$ we have

$$\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_{\kappa} \leq \alpha \|\mathcal{A}(\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}})\|_2 \ .$$

Therefore for all $\pi'_{\mathfrak{S}} \in \mathfrak{S}$ we have

$$\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_{\kappa} \leq \max(\alpha \|\mathcal{A}(\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}})\|_2, \eta) \leq \alpha \|\mathcal{A}(\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}})\|_2 + \eta.$$

which is indeed the desired LRIP. Our main goal is therefore to prove that with high probability (3.13) hold for all $\mu \in \mathcal{S}^{\eta}(\pi_{\mathfrak{S}}, \mathfrak{S})$.

Our strategy to prove (3.13) and therefore the LRIP follows a two-step methodology inspired by CS [Bar07] (also evoked in Section 1.3):

1. first prove that for any *fixed* $\mu \in \mathcal{S}^{\eta}(\pi_{\mathfrak{S}}, \mathfrak{S})$, the desired inequality (3.13) holds with high probability,
2. then extend to a result valid for *all* $\mu \in \mathcal{S}^{\eta}(\pi_{\mathfrak{S}}, \mathfrak{S})$ with high probability, by assuming that the normalized secant set (3.12) has finite covering numbers.

3.2.2 Admissibility

We will deal with the first step by controlling the approximation (3.10) with Bernstein's concentration inequality (Lemma A.2.1). To apply it we need to uniformly bound the elements in the normalized secant set. We define the notion of *admissible* RF expansion of the kernel.

Definition 3.2.1 (Admissibility). *The Random Feature expansion $(\mathcal{F}_R, \Lambda)$ of the kernel κ is* ***admissible*** *for the model $\mathfrak{S}$ with constant $W_\Lambda < \infty$ and error $\eta \geq 0$ if: for all $\pi_{\mathfrak{S}} \in \mathfrak{S}$ and all $\mu \in \mathcal{S}^\eta(\pi_{\mathfrak{S}}, \mathfrak{S})$, we have*

$$\|\mu\|_{\mathcal{F}_R} \leq W_\Lambda \tag{3.14}$$

First note that, by Lemma 3.1.5, we have necessarily $W_\Lambda \geq 1$ since $\|\cdot\|_\kappa \leq \|\cdot\|_{\mathcal{F}_R}$. As noted before, a kernel often has an infinity of possible RF expansions, with potentially different admissibility constants W_Λ and error η.

Remark 3.2.2. *To better understand the admissibility condition, recall the expression of the (semi)norms: $\|\mu\|_{\mathcal{F}_R} = \|\psi_\mu\|_{L^\infty}$ and $\|\mu\|_\kappa = \|\psi_\mu\|_{L^2(\Lambda)}$. Hence, proving that the admissibility condition holds amounts to showing that for $\frac{\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}}{\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_\kappa} \in \mathcal{S}^\eta(\pi_{\mathfrak{S}}, \mathfrak{S})$ it holds that:*

$$\left\|\psi_{\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}}\right\|_{L^\infty} \lesssim \left\|\psi_{\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}}\right\|_{L^2(\Lambda)} .$$

One can immediately see that this is in general not true, and even if true, potentially challenging to prove. In the simple examples presented at the end of this chapter, we will use a "weak" admissibility condition, introduced below, that is always true when $\eta > 0$. It will however come at the price of obtaining sub-optimal results in terms of sketch size. In the more advanced analysis of Chapter 6, we will fully exploit the geometry of the model $\mathfrak{S}$ instead, to prove admissibility conditions with $\eta = 0$.

Weak admissibility. Since all the considered features are uniformly bounded, we have the following "weak" admissibility result.

Lemma 3.2.3 (Weak admissibility.). *For any model $\mathfrak{S}$ and strictly positive error $\eta > 0$,* any RF expansion $(\mathcal{F}_R, \Lambda)$ *is admissible for $\mathfrak{S}$ with error η and constant $W_\Lambda = 2B_{\mathcal{F}_R}/\eta$.*

Proof. Consider any model $\mathfrak{S}$, error $\eta > 0$ and RF expansion $(\mathcal{F}_R, \Lambda)$. Let $\pi_{\mathfrak{S}} \in \mathfrak{S}$ be an element of the model. Any measure $\mu \in \mathcal{S}^\eta(\pi_{\mathfrak{S}}, \mathfrak{S})$ can be decomposed as $\mu = (\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}})/\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_\kappa$ where $\pi'_{\mathfrak{S}} \in \mathfrak{S}$ and $\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_\kappa > \eta$. Then we have

$$\|\mu\|_{\mathcal{F}_R} = \frac{\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_{\mathcal{F}_R}}{\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_\kappa} \overset{\text{Lem. 3.1.5}}{\leq} \frac{B_{\mathcal{F}_R}\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_{\text{TV}}}{\eta} \leq \frac{2B_{\mathcal{F}_R}}{\eta},$$

since $\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_{\text{TV}} \leq 2$. □

Remark 3.2.4 (Recovery results with weak admissibility). *As mentioned above, the weak admissibility condition is always true but leads to sub-optimal results. More precisely, we will see in the next theorem that our analysis yields a sufficient sketch size for the LRIP that is at least as $m \geq \mathcal{O}\left(W_\Lambda^2\right)$, and that such an expression leads to a recovery result where the additive error behaves as $\mathcal{O}\left(1/\sqrt{n} + 1/\sqrt{m}\right)$, which only proves that sketch learning is possible with a sketch size of the order of that of the original database. In Chapter 6 we prove the existence of admissibility constants that produce better estimation results.*

3.2.3 Proof of the LRIP

Before proving the LRIP, we describe a possible route to relate the problem to other metrics instead of the MMD, as will be done in Chapter 6 to link the sketching method to more traditional learning.

Replacing the MMD with a more meaningful metric. By combining the LRIP (3.11) with the Theorem 2.2.3, we guarantee that the decoder (3.9) is instance optimal with respect to the MMD $\|\cdot\|_\kappa$. We could be satisfied with this result: the MMD has indeed been used as an evaluation metric in many problems, including density fitting [Sri+09]. However, it may be unclear whether this result yields guarantees with respect to more traditional metrics [Red+15]. In the formulation of our main result below (assumption iii) in Theorem 3.2.5), we allow for replacing the MMD with another metric $d_{\mathcal{L}}$, by assuming that it is dominated by the MMD, to eventually obtain recovery results with respect to $d_{\mathcal{L}}$ if so desired. As outlined this assumption is "optional" in the sense that one can always select $d_{\mathcal{L}} := \|\cdot\|_\kappa$ to obtain recovery result with respect to the MMD, with no additional assumption.

Let us now formulate our result for the LRIP.

Theorem 3.2.5 (LRIP for the sketching operator)**.** *Consider a model $\mathfrak{S} \subset \mathfrak{P}$, a fixed distribution in the model $\pi_{\mathfrak{S}} \in \mathfrak{S}$, a kernel κ with a RF expansion $(\mathcal{F}_R, \Lambda)$, an error $\eta \geq 0$ and a sketch size m such that:*

i) the RF expansion $(\mathcal{F}_R, \Lambda)$ is admissible (Def. 3.2.1) for the model $\mathfrak{S}$ with constant $W_\Lambda < \infty$ and error η;

ii) the normalized secant set $\mathcal{S}^\eta(\pi_{\mathfrak{S}}, \mathfrak{S})$ has finite covering numbers (see definition in Appendix A.1.1). Denote

$$N := \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{S}^\eta(\pi_{\mathfrak{S}}, \mathfrak{S}), \frac{1}{4}\right) < \infty; \tag{3.15}$$

iii) there is a pseudometric $d_{\mathcal{L}}$ and a constant $W_{\mathcal{L}} < \infty$ such that for all $\pi'_{\mathfrak{S}} \in \mathfrak{S}$,

$$d_{\mathcal{L}}(\pi_{\mathfrak{S}}, \pi'_{\mathfrak{S}}) \leq W_{\mathcal{L}} \|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_\kappa \tag{3.16}$$

(as mentioned before this assumption is optional, one can take $d_{\mathcal{L}} := \|\cdot\|_\kappa$ and $W_{\mathcal{L}} := 1$);

iv) the sketch size satisfies

$$m \geq c W_\Lambda^2 \log\left(\frac{N}{\rho}\right), \tag{3.17}$$

for some $\rho > 0$, where $c = 1760/147$ is a universal constant.

Then, the sketching operator (3.7) satisfies the non-uniform LRIP for the model $\mathfrak{S}$ and element of the model $\pi_{\mathfrak{S}}$ with constant $\alpha = 2W_{\mathcal{L}}$, pseudometric $d_{\mathcal{L}}$, probability $1 - \rho$ and error $W_{\mathcal{L}}\eta$. Meaning that, with probability at least $1 - \rho$ on the drawing of the $\boldsymbol{\omega}_j$'s that define the sketching operator $\mathcal{A}$ by (3.7), we have for all $\pi'_{\mathfrak{S}} \in \mathfrak{S}$:

$$d_{\mathcal{L}}(\pi_{\mathfrak{S}}, \pi'_{\mathfrak{S}}) \leq 2W_{\mathcal{L}} \|\mathcal{A}\pi_{\mathfrak{S}} - \mathcal{A}\pi'_{\mathfrak{S}}\|_2 + W_{\mathcal{L}}\eta \tag{3.18}$$

Proof. Denote $\mathcal{S} = \mathcal{S}^\eta(\pi_{\mathfrak{S}}, \mathfrak{S})$. We are going to prove that with high probability on the sketching operator $\mathcal{A}$, for all $\mu \in \mathcal{S}$ it holds that

$$\|\mathcal{A}\mu\|_2 \geq \frac{1}{2}, \tag{3.19}$$

from which we have seen that we can deduce that for all $\pi'_{\mathfrak{S}} \in \mathfrak{S}$ we have

$$\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_\kappa \leq \max(2\|\mathcal{A}(\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}})\|_2, \eta) \leq 2\|\mathcal{A}(\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}})\|_2 + \eta,$$

which combined with (3.16) yields the desired result.

To prove that (3.19) holds with high probability for all $\mu \in \mathcal{S}$, we prove that it holds for a *finite* number of μ_i's in $\mathcal{S}$ with high probability, then use a δ-covering to extend the result to the whole set.

Let $\pi'_{\mathfrak{S}} \in \mathfrak{S}$ be such that $\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_\kappa > \eta$. Define $\mu = \frac{\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}}{\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_\kappa} \in \mathcal{S}$. Draw $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m \in \Omega$ *i.i.d.* from Λ and denote $Y_j = 1 - |\psi_\mu(\boldsymbol{\omega}_j)|^2$. We have:

- the Y_j's are *i.i.d.* and $\mathbb{E}Y_j = 1 - \mathbb{E}_{\boldsymbol{\omega}_j \sim \Lambda} \left|\psi_{\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}}(\boldsymbol{\omega}_j)\right|^2 / \|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_\kappa^2 = 0$;
- by admissibility,
$$|\psi_\mu(\boldsymbol{\omega}_j)|^2 = \left|\langle \mu, \phi_{\boldsymbol{\omega}_j} \rangle\right|^2 \leq \|\mu\|_{\mathcal{F}_R}^2 \leq W_\Lambda^2 ,$$
and therefore $Y_j \in [1 - W_\Lambda^2, 1]$ which, since by Lemma 3.1.5 we have necessarily $W_\Lambda \geq 1$, implies
$$|Y_j| \leq W_\Lambda^2 ;$$
- the variance of the Y_j's satisfies
$$\begin{aligned}\mathrm{Var}(Y_j) = \mathrm{Var}\left(|\psi_\mu(\boldsymbol{\omega}_j)|^2\right) \leq \mathbb{E}\,|\psi_\mu(\boldsymbol{\omega}_j)|^4 &= \frac{\mathbb{E}\left|\psi_{\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}}(\boldsymbol{\omega}_j)\right|^4}{\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_\kappa^4} \\ &\leq \frac{\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_{\mathcal{F}_R}^2 \,\mathbb{E}\left|\psi_{\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}}(\boldsymbol{\omega}_j)\right|^2}{\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_\kappa^4} = \frac{\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_{\mathcal{F}_R}^2}{\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_\kappa^2} \leq W_\Lambda^2 .\end{aligned}$$
- defining the sketching operator as (3.7), we have $\frac{1}{m}\sum_{j=1}^m Y_j = 1 - \|\mathcal{A}\mu\|_2^2$

Hence, we can apply Bernstein's inequality (Lemma A.2.1 in Appendix A), we obtain that:

$$P\left(1 - \|\mathcal{A}\mu\|_2^2 \geq \frac{7}{16}\right) \leq \exp\left(-\frac{m}{cW_\Lambda^2}\right)$$

with $c = \frac{1760}{147}$.

Denote $N = \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{S}, \frac{1}{4}\right)$ and let $\mu_1, ..., \mu_N$ be a $1/4$-covering of $\mathcal{S}$. A union bound yields that with probability at least $1 - N\exp\left(-\frac{m}{cW_\Lambda^2}\right)$, we have:

$$\forall \mu_i,\ \|\mathcal{A}\mu_i\|_2 \geq \sqrt{1 - \frac{7}{16}} = \frac{3}{4}. \tag{3.20}$$

Assuming now that (3.20) is satisfied, for all $\mu \in \mathcal{S}$ there exists μ_i such that $\|\mu - \mu_i\|_{\mathcal{F}_R} \leq 1/4$ and we have

$$1 - \|\mathcal{A}\mu\|_2 = 1 - \|\mathcal{A}\mu_i\|_2 + \|\mathcal{A}\mu_i\|_2 - \|\mathcal{A}\mu\|_2 \overset{(3.20)}{\leq} 1 - \frac{3}{4} + \|\mathcal{A}\mu_i\|_2 - \|\mathcal{A}\mu\|_2$$

Then using the reverse triangular inequality

$$1 - \|\mathcal{A}\mu\|_2 \leq \frac{1}{4} + \|\mathcal{A}(\mu - \mu_i)\|_2 \overset{\text{Lem. 3.1.5}}{\leq} \frac{1}{4} + \|\mu - \mu_i\|_{\mathcal{F}_R} \leq \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

which is the desired result. We then denote $\rho = N\exp\left(-\frac{m}{cW_\Lambda^2}\right)$ the probability of failure and solve for m to obtain the bound (3.17). □

Hypotheses. As we discussed, this theorem shows that the LRIP holds under two main hypotheses: the admissibility condition and the assumption of finiteness of the covering numbers of the normalized secant set. As we will further observe in Chapter 6, both can be relatively technical to prove, depending on the allowed additive error $\eta \geq 0$.

Sketch size. The bound (3.17) on the sufficient sketch size involves two terms, coming from our two-parts proof strategy. The first is the squared admissibility constant W_Λ^2, which reflects how well the LRIP inequality holds for a fixed pair of distributions, with high probability (*i.e.* it reflects the quality of the concentration result obtained with Bernstein's inequality). This

pointwise concentration is the first part of our proof. The second term is the logarithm of the covering numbers of the normalized secant set, $\log(N)$. It arises from the second part of the proof, which is to extend the pointwise concentration result to a *uniform* bound valid for all element of the normalized secant set with high probability. Both terms are important and must be carefully controlled in each particular instantiation of the sketching method.

Usual expression of the covering numbers. We will see that the normalized secant set usually has a finite *upper box counting dimension* $q \in \mathbb{N}$ [Rob11; PDG15], meaning that its covering numbers are as $\mathcal{N}(\|\cdot\|, \mathcal{S}, \delta) \propto (C/\delta)^q$. In that case, the logarithm of the covering numbers is as $q\log(C/\delta)$, and the sketch size (3.17) indeed scales, up to logarithmic terms, in $\mathcal{O}\left(W_\Lambda^2 q\right)$. Here q reflects the complexity (the "dimensionality") of the problem.

3.2.4 Bounding the empirical error

The LRIP implies that the decoder (3.9) is instance optimal. When the decoding is done from the empirical sketch $\hat{\mathbf{y}} = \mathcal{A}\hat{\pi}_n$ the "noise" is

$$\mathbf{e} = \mathcal{A}\pi - \mathcal{A}\hat{\pi}_n \tag{3.21}$$

which, given that the operator $\mathcal{A}$ computes a collection of moments, can be bounded using a generalized Hoeffding's inequality (Lemma A.2.2).

Lemma 3.2.6 (Bounding the empirical error.). *Consider $\mathbf{z}_1, ..., \mathbf{z}_n \in \mathfrak{Z}$ drawn* i.i.d. *from $\pi^\star$ and let $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m \in \Omega$ be m frequencies drawn* i.i.d. *from Λ. Then, with probability at least $1-\rho$ on the drawing of both $\mathbf{z}_i$'s and $\boldsymbol{\omega}_j$'s we have*

$$\|\mathcal{A}(\pi^\star - \hat{\pi}_n)\|_2 \leq \frac{B_{\mathcal{F}_R}\left(1 + \sqrt{2\log(1/\rho)}\right)}{\sqrt{n}} \tag{3.22}$$

Proof. Consider for now a *fixed* set of frequencies $\{\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m\}$ that define an operator $\mathcal{A}$. Draw $\mathbf{z}_1, ..., \mathbf{z}_n$ *i.i.d.* from π, and denote $Y_i = m^{-\frac{1}{2}}\left[\phi_{\boldsymbol{\omega}_j}(\mathbf{z}_i)\right]_{j=1}^m$, which are drawn *i.i.d.* in $\mathbb{C}^m$. We have $\mathcal{A}\hat{\pi}_n = \frac{1}{n}\sum_{i=1}^n Y_i$, $\mathcal{A}\pi = \mathbb{E}Y_i$, and $\|Y_i\|_2 \leq B_{\mathcal{F}_R}$ since the Random Features are uniformly bounded by $B_{\mathcal{F}_R}$. We can therefore apply Lemma A.2.2, which yields that:

$$P_{\mathbf{z}}\left(\|\mathcal{A}(\pi - \hat{\pi}_n)\|_2 \leq \frac{B_{\mathcal{F}_R}\left(1 + \sqrt{2\log(1/\rho)}\right)}{\sqrt{n}}\right) \geq 1 - \rho.$$

Then, since the samples $\mathbf{z}_i$ and frequencies $\boldsymbol{\omega}_j$'s are drawn independently and this property is valid with probability $1-\rho$ for all sets of $\boldsymbol{\omega}_j$'s, it is valid with probability $1-\rho$ on the drawing of *both*[a] the $\mathbf{z}_i$'s and $\boldsymbol{\omega}_j$'s. □

[a] Indeed, if an event A depends on two independent random variables X and Y and we have that for all fixed y, $P_X(A(X,y)) = \int_x 1_A(x,y)dP_X(x) \geq 1-\rho$, then the joint probability is such that $P_{X,Y}(A(X,Y)) = \iint_{x,y} 1_A(x,y)dP_{X,Y}(x,y) = \int_y \left(\int_x 1_A(x,y)dP_X(x)\right)dP_Y(y) \geq (1-\rho)\int_y dP_Y(y) = 1-\rho$.

3.2.5 Main result

Using the LRIP and the bound on the empirical error, we obtain the following theorem, which is our main recovery result.

Theorem 3.2.7 (Information-preservation guarantees for the sketching method)**.** *Consider a model $\mathfrak{S} \subset \mathfrak{P}$, a distribution $\pi^\star \in \mathfrak{P}$, a distribution in the model $\pi_\mathfrak{S} \in \mathfrak{S}$, a kernel κ with a RF expansion $(\mathcal{F}_R, \Lambda)$, an error $\eta \geq 0$ and a sketch size m such that:*

- *the hypotheses of Theorem 3.2.5 hold:*
  - *i) the RF expansion $(\mathcal{F}, \Lambda)$ is admissible (Def. 3.2.1) for the model $\mathfrak{S}$ with constant $W_\Lambda < \infty$ and error η;*
  - *ii) the normalized secant set $\mathcal{S}^\eta(\pi_\mathfrak{S}, \mathfrak{S})$ has finite covering numbers. Denote*

$$N := \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{S}^\eta(\pi_\mathfrak{S}, \mathfrak{S}), \frac{1}{4}\right) < \infty \; ; \tag{3.23}$$

  - *iii) there is a pseudometric $d_\mathcal{L}$ and a constant $W_\mathcal{L} < \infty$ such that for all $\pi'_\mathfrak{S} \in \mathfrak{S}$,*

$$d_\mathcal{L}(\pi_\mathfrak{S}, \pi'_\mathfrak{S}) \leq W_\mathcal{L} \left\|\pi_\mathfrak{S} - \pi'_\mathfrak{S}\right\|_\kappa \; ; \tag{3.24}$$

  - *iv) the sketch size satisfies*

$$m \geq c W_\Lambda^2 \log\left(\frac{N}{\rho}\right) \; , \tag{3.25}$$

  for some $\rho > 0$, where $c = 1760/147$ is a universal constant.

- *the sketching operator satisfies the non-uniform Boundedness Property (Def. 2.3.1) for $\pi^\star$ and $\pi_\mathfrak{S}$ with constant β, pseudometric d_G and probability $1 - \rho'$.*

Denote the (ideally small) bias term

$$\tau = d'_E(\pi^\star, \pi_\mathfrak{S}) \; , \tag{3.26}$$

where $d'_E = d_\mathcal{L} + 4\beta W_\mathcal{L} d_G$.

Consider items $\mathbf{z}_1, ..., \mathbf{z}_n \in \mathfrak{Z}$ drawn i.i.d. *from $\pi^\star$ and frequencies $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m \in \Omega$ drawn* i.i.d. *from Λ, which define the sketching operator $\mathcal{A}$ by (3.7). Denote $\widetilde{\pi} = \Delta_\iota(\mathcal{A}, \mathcal{A}\hat{\pi}_n)$ the probability distribution recovered from the empirical sketch. Then, with probability at least $1 - (\rho + \rho' + \rho'')$ on the drawing of the $\mathbf{z}_i$'s and $\boldsymbol{\omega}_j$'s, it holds that*

$$d_\mathcal{L}(\pi^\star, \widetilde{\pi}) \leq \tau + \frac{4 W_\mathcal{L} B_{\mathcal{F}_R}\left(1 + \sqrt{2\log(1/\rho'')}\right)}{\sqrt{n}} + W_\mathcal{L}(\eta + 2\iota). \tag{3.27}$$

Proof. Since the hypotheses of Theorem 3.2.5 hold, the sketching operator satisfies the non-uniform LRIP (Def. 2.3.2) for the model $\mathfrak{S}$ and distribution $\pi_\mathfrak{S}$ with constant $\alpha := 2W_\mathcal{L}$, pseudometric $d_\mathcal{L}$, probability $1 - \rho$ and error $W_\mathcal{L}\eta$.

By hypothesis the sketching operator satisfies the non-uniform Boundedness Property, and we can therefore apply Theorem 2.3.6, which shows that the decoder satisfies the non-uniform IOP for the sketching operator $\mathcal{A}$, model $\mathfrak{S}$, signal $\pi^\star$ and distribution in the model $\pi_\mathfrak{S}$ with constant $A := 1$ and $B := 4W_\mathcal{L}$, pseudometrics $d_\mathcal{L}$ and $d'_E := d_\mathcal{L} + 4\beta W_\mathcal{L} d_G$, probability $1 - \rho - \rho'$ and error $W_\mathcal{L}(\eta + 2\iota)$. Meaning that, with probability at least $1 - \rho - \rho'$ on the drawing of the frequencies $\boldsymbol{\omega}_j$, for *all* sets of items $\mathbf{z}_1, ..., \mathbf{z}_n$, denoting $\widetilde{\pi} = \Delta_\iota(\mathcal{A}, \mathcal{A}\hat{\pi}_n)$ we have

$$d_\mathcal{L}(\pi^\star, \widetilde{\pi}) \leq d'_E(\pi^\star, \pi_\mathfrak{S}) + \left\|\mathcal{A}(\pi^\star - \hat{\pi}_n)\right\|_2 + W_\mathcal{L}(\eta + 2\iota),$$

We use Lemma 3.2.6 to bound $\left\|\mathcal{A}(\pi^\star - \hat{\pi}_n)\right\|_2$, with a union bound to get the desired result. □

The proof mechanisms of this theorem are illustrated is Fig. 3.1, where we emphasize our use of the previous results of Chapter 2.

Precision of the estimation. Theorem 3.2.7 shows that, under assumptions that are essentially that of Theorem 3.2.5 (admissibility and finiteness of covering numbers), the distribution

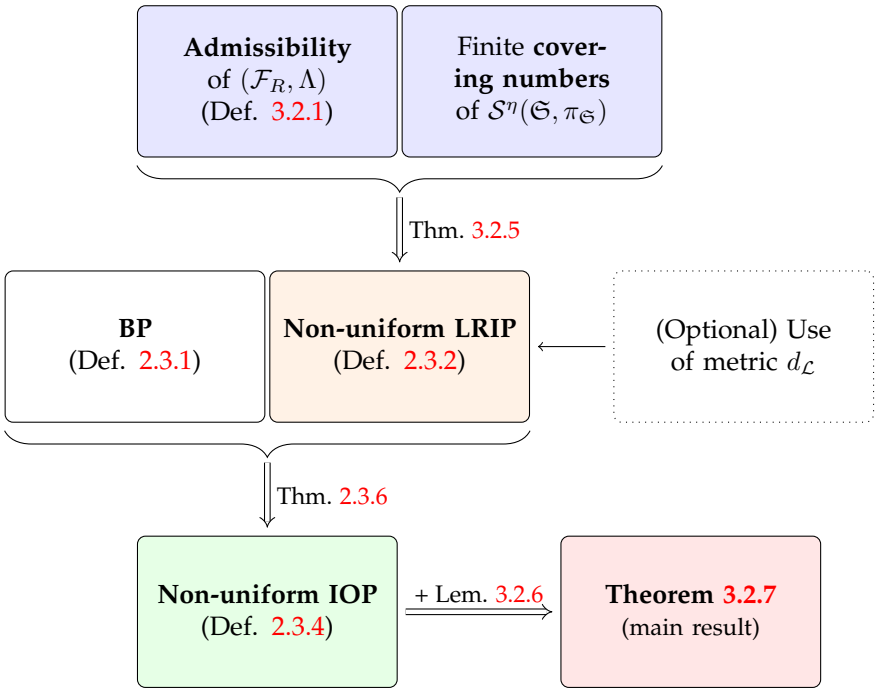

FIGURE 3.1: Illustration of the proof of Theorem 3.2.7. The admissibility condition and finiteness of the covering numbers of the normalized secant sets are used to prove the LRIP, then the previous results of Chapter 3 show the IOP. With the additional control on the empirical error, it yields Theorem 3.2.7.

$\widetilde{\pi}$ recovered from the empirical sketch is close to the true distribution of the data $\pi^\star$, with a precision that involves several terms:

- the bias term $\tau = d'_E(\pi^\star, \pi_{\mathfrak{S}})$, which indicates how well the true distribution of the data is modeled by $\mathfrak{S}$, such that the recovery is stable to modelization error. If the distribution of the data is in the model, this term is zero.
- the empirical error in $\mathcal{O}(1/\sqrt{n})$. This is a classic learning rate when using empirical data, that arises from well-known concentration inequalities such as the Hoeffding's inequality that we used in Lemma 3.2.6. That this learning rate is not degraded by the sketch-and-learn process is somewhat fortuitous, if not relatively expected given that our proof strategy considers the empirical error as the amplitude of a "noise", and that an instance optimal decoder is robust to noise.
- two additional errors, η and ι. As described in Chapter 2 Section 2.1.2, the latter mostly exists for technical reasons and can be chosen as small as desired. On the contrary, the additive error $\eta \geq 0$ is crucial, as it may largely drive the admissibility condition and the covering numbers of $\mathcal{S}^\eta(\pi_{\mathfrak{S}}, \mathfrak{S})$. The ideal case is $\eta = 0$, and we will exhibit some examples that satisfy this condition in Chapter 6. As we will see in the next section, at the other end of the spectrum, the "worst" case (in the sense that it is almost always true) is an error that behaves as $\eta = \mathcal{O}(1/\sqrt{m})$. In this "worst" case, the total error is as $\mathcal{O}(1/\sqrt{n} + 1/\sqrt{m})$, and the obtained recovery result does not prove that the sketching method is efficient when using a sketch that is significantly smaller than the original database.

3.3 First applications of Theorem 3.2.7

As mentioned earlier, the admissibility hypothesis and finiteness of the covering numbers of the normalized secant set can be technical to obtain, especially in the case where a true LRIP is desired with no additive error ($\eta = 0$). In this section, we will derive preliminary results

under weaker hypotheses that will apply to a large class of models. We will however obtain an additional error η that is significantly sub-optimal compared to what is observed in practice.

The proofs of this section are given in Appendix B.

3.3.1 Weak assumptions

Here we just assume that *the model* $\mathfrak{S}$ *itself* has finite covering numbers (instead of the normalized secant set), which is often relatively easy to prove (for instance in the next sections we will show this property for GMMs and mixtures of elliptic stable distributions). In that case we show that for all *strictly positive* error levels ($\eta > 0$) the normalized secant set $\mathcal{S}^\eta(\pi_{\mathfrak{S}}, \mathfrak{S})$ has finite covering numbers.

Lemma 3.3.1 (Covering numbers of the extruded secant set)**.** *Let $(\mathcal{F}_R, \Lambda)$ be a Random Feature expansion of the kernel κ that is admissible for the model $\mathfrak{S}$ with constant $W_\Lambda > 0$ and* strictly positive *error $\eta > 0$. Assume the model $\mathfrak{S}$ has finite covering numbers with respect to the norm $\|\cdot\|_{\mathcal{F}_R}$. Then, for any $\pi \in \mathfrak{S}$, the normalized secant set $\mathcal{S}^\eta(\pi, \mathfrak{S})$ has finite covering numbers for the norm $\|\cdot\|_{\mathcal{F}_R}$ and we have:*

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{S}^\eta(\pi, \mathfrak{S}), \delta\right) \leq \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{S}, \frac{\delta\eta}{8W_\Lambda}\right) \tag{3.28}$$

Proof. This is a particular case of Lemma A.3.5 which proves this result for more general normalized secant sets. We apply with $\|\cdot\|_a := \|\cdot\|_{\mathcal{F}_R}$, $\|\cdot\|_b := \|\cdot\|_\kappa$, $A := 1$ and $B := W_\Lambda$. □

Then, with no other assumption, we obtain the following corollary.

Corollary 3.3.2 (Information-preservation guarantees with weak assumptions.)**.** *Consider a model $\mathfrak{S}$, a distribution $\pi^\star \in \mathfrak{P}$, a distribution in the model $\pi_{\mathfrak{S}} \in \mathfrak{S}$, a kernel κ with a RF expansion $(\mathcal{F}_R, \Lambda)$, an error $\eta > 0$ and a sketch size m such that:*

- *the model has finite covering numbers with respect to the norm $\|\cdot\|_{\mathcal{F}_R}$;*
- *the sketch size satisfies*

$$m \geq 4cB_{\mathcal{F}_R}^2 \eta^{-2} \log\left(\frac{\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{S}, \frac{\eta^2}{64B_{\mathcal{F}_R}}\right)}{\rho}\right), \tag{3.29}$$

for some $\rho > 0$, where $c = 1760/147$.

Denote the bias term

$$\tau = d'_E(\pi^\star, \pi_{\mathfrak{S}}),$$

where $d'_E := \|\cdot\|_\kappa + 4\|\cdot\|_{\mathcal{F}_R}$.

Consider items $\mathbf{z}_1, ..., \mathbf{z}_n \in \mathfrak{Z}$ drawn i.i.d. *from $\pi^\star$ and frequencies $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m \in \Omega$ drawn* i.i.d. *from Λ, which define the sketching operator $\mathcal{A}$ by (3.7). Denote $\widetilde{\pi} = \Delta_\iota(\mathcal{A}, \mathcal{A}\hat{\pi}_n)$ the probability distribution recovered from the empirical sketch. Then, with probability at least $1 - (\rho + \rho')$ on the drawing of the $\mathbf{z}_i$'s and $\boldsymbol{\omega}_j$'s, it holds that*

$$\|\pi^\star - \widetilde{\pi}\|_\kappa \leq \tau + \frac{4B_{\mathcal{F}_R}\left(1 + \sqrt{2\log(1/\rho')}\right)}{\sqrt{n}} + \eta + 2\iota, \tag{3.30}$$

Proof. The result is obtained by directly using Theorem 3.2.7 with assumptions that are simplified as much as possible:

- the covering numbers of the normalized secant sets are bounded using Lemma 3.3.1;

– we use "weak" admissibility (Lemma 3.2.3);

– we do not suppose the existence of the optional metric $d_{\mathcal{L}}$ but use the fact that Theorem 3.2.7 is applicable with $d_{\mathcal{L}} := \|\cdot\|_{\kappa}$ and $W_{\mathcal{L}} := 1$;

– we also use the fact that the sketching operator satisfies the Boundedness Property for $d_G := \|\cdot\|_{\mathcal{F}_R}$ and probability 1 (see the discussion after Lemma 3.1.5).

□

Sub-optimal sketch size. If the model has an upper box counting dimension q, the error η in this corollary scales approximately in $\mathcal{O}\left(\sqrt{q/m}\right)$. As mentioned earlier, since the empirical error in Theorem 3.2.7 behaves as $\mathcal{O}\left(1/\sqrt{n}\right)$, the result indeed suggests that attaining a fixed error level requires a sketch as big as the original database. While this corollary does not show that the sketching method is significantly faster than using the full data, it proves nevertheless that the sketch-and-recover procedure is asymptotically stable, when both database and sketch sizes are large, which is a first step toward more optimal results. Furthermore, there may be cases where a reduced sketch size is somehow not our primary concern. For instance, in the case of mixture of multivariate stable distributions described after, to our knowledge the proposed sketching method is the first estimator with guarantees.

Fortunately, these results are largely pessimistic compared to what is observed in practice. In numerical experiments we obtain excellent recovery results for sketches whose size does not seem to depend on the size of the database but rather on the number of parameters in the model. In Chapter 6 we will push further the theoretical analysis and obtain this kind of results.

Result with admissibility. Let us examine the case where we still only assume finiteness of the covering numbers of the model instead of the normalized secant set, but where we assume additionally that we have an RF expansion of the kernel that is admissible with a constant $W_\Lambda < \infty$ that does not depend on $\eta > 0$. In that case we can change (3.29) by

$$m \geq cW_\Lambda^2 \log\left(\frac{\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{S}, \frac{\eta}{32W_\Lambda}\right)}{\rho}\right). \tag{3.31}$$

When the model has a finite upper box counting dimension q the additive error scales as $\eta = \mathcal{O}\left(e^{-m/q}\right)$, which is far better than the result in $\eta = \mathcal{O}\left(1/\sqrt{m}\right)$ obtained when we used the weak admissibility. In that case, the sketching method is indeed effective with a sketch significantly smaller than the full database. In some sense, the admissibility condition is therefore the first and most crucial condition to obtain an efficient sketching method, since it drives the very concentration result at the heart of the LRIP.

Such importance of the admissibility condition is due to our use of (this version of) Bernstein's inequality, which requires a uniform bound. In future work we will examine if other concentration inequalities can be used to refine this presently quite rough condition.

In Chapter 6 we will prove a "strong" admissibility condition with $\eta = 0$ in two particular cases.

3.3.2 Mixture model

This thesis work is mainly oriented toward mixture models. We show that the covering numbers of a basic set of distributions can be used to bound the covering numbers of the set of *mixtures* of distributions from this set.

Consider a set of basic probability distributions $\mathfrak{T} \subset \mathfrak{P}$ and an integer $k > 0$. The *mixture model* $\mathfrak{S}_k(\mathfrak{T})$ is defined as

$$\mathfrak{S}_k(\mathfrak{T}) = \left\{ \sum_{l=1}^{k} \boldsymbol{\xi}_l \pi_l \;\middle|\; \boldsymbol{\xi} \in \mathbb{S}^{k-1},\ \pi_l \in \mathfrak{T} \right\} \tag{3.32}$$

where $\mathbb{S}^{k-1} = \left\{ \boldsymbol{\xi} \in \mathbb{R}^d_+ \mid \sum_{l=1}^k \xi_l = 1 \right\}$ is the $k-1$ dimensional simplex.

Lemma 3.3.3. *Consider any RF expansion $(\mathcal{F}_R, \Lambda)$. For all $0 < \delta \leq 16 B_{\mathcal{F}_R}$ the set $\mathfrak{S}_k(\mathfrak{T})$ satisfies*

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{S}_k(\mathfrak{T}), \delta\right) \leq \left(\frac{16 B_{\mathcal{F}_R} \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{T}, \delta/2\right)}{\delta} \right)^k . \tag{3.33}$$

Hence for any mixture models we can very quickly obtain (sub-optimal) recovery guarantees by:

1. bounding the covering numbers of the set $\mathfrak{T}$ of basic distributions;
2. applying Lemma 3.3.3 to bound the covering numbers of the set of mixtures $\mathfrak{S}_k(\mathfrak{T})$;
3. applying Corollary 3.3.2.

Bounding the covering numbers of the set of basic distributions $\mathfrak{T}$ can often be done easily enough. In general, it is a parametric set of distributions $\mathfrak{T} = \{\pi_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathcal{T}\}$ where the parameter $\boldsymbol{\theta}$ lives in a finite-dimensional set $\mathcal{T} \subset \mathbb{R}^q$ assumed bounded[2]. Then our strategy is to prove that the embedding $\boldsymbol{\theta} \mapsto \pi_{\boldsymbol{\theta}}$ is Lipschitz continuous, and use Lemma A.3.2 to bound the covering numbers of $\mathfrak{T}$ by these of $\mathcal{T}$.

We will now give two examples of this application.

3.3.3 Example 1: Gaussian Mixture model

We first illustrate Corollary 3.3.2 on the problem of recovering a Gaussian Mixture Model with unknown covariance. We restrict to diagonal covariance for simplicity.

In this example $\mathfrak{Z} := \mathbb{R}^d$. The basic set of distributions is defined as a parametric set $\mathfrak{T} := \{\pi_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathcal{T}\}$, where $\boldsymbol{\theta} := (\boldsymbol{\mu}, \boldsymbol{\sigma})$ contains the mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and diagonal of the covariance $\boldsymbol{\sigma} = \left[\sigma_l^2\right]_{l=1}^d \in \mathbb{R}^d_+$ of a Gaussian $\pi_{\boldsymbol{\theta}} := \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}))$.

We restrict to a parameter set $\mathcal{T} = \mathcal{D}_{\boldsymbol{\mu}} \times \mathcal{D}_{\boldsymbol{\sigma}}$ where $\mathcal{D}_{\boldsymbol{\mu}} \subset \mathbb{R}^d$ is the set of means that is defined as a Euclidean ball[3]:

$$\mathcal{D}_{\boldsymbol{\mu}} := \mathcal{B}_{\mathbb{R}^d, \|\cdot\|_2}(0, R_{\boldsymbol{\mu}}) = \left\{ \boldsymbol{\mu} \in \mathbb{R}^d \mid \|\boldsymbol{\mu}\|_2 \leq R_{\boldsymbol{\mu}} \right\} , \tag{3.34}$$

and $\mathcal{D}_{\boldsymbol{\sigma}}$ is a bounded set of vectors in $\mathbb{R}^d_+$ such that

$$R_{\boldsymbol{\sigma}} := \max_{\boldsymbol{\sigma} \in \mathcal{D}_{\boldsymbol{\sigma}}} \|\boldsymbol{\sigma}\|_2 < \infty , \tag{3.35}$$

$$\sigma^2_{\min} := \min_{\boldsymbol{\sigma} \in \mathcal{D}_{\boldsymbol{\sigma}}} \min_{1 \leq i \leq d} \sigma_i^2 > 0 , \tag{3.36}$$

meaning that the Euclidean norm of the vector $\boldsymbol{\sigma}$ is bounded by $R_{\boldsymbol{\sigma}}$ and each of its entries is bounded away from 0 by $\sigma^2_{\min}$.

As we have said, our strategy is to prove that the mapping $\boldsymbol{\theta} \mapsto \pi_{\boldsymbol{\theta}}$ is Lipschitz. We use the following lemma (all proofs are in Appendix B).

Lemma 3.3.4. *Consider two Gaussians $\pi_1 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\pi_2 = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. We have*

$$\|\pi_1 - \pi_2\|_{TV} \leq \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}} + \left(\frac{1}{2} \|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_F \left\|\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\right\|_F \right)^{\frac{1}{2}} \tag{3.37}$$

where $\boldsymbol{\Sigma} = \left(\frac{\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1}}{2} \right)^{-1}$, $\|\cdot\|_{\boldsymbol{\Sigma}}$ is the Mahalanobis norm defined by $\|\mathbf{x}\|^2_{\boldsymbol{\Sigma}} = \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}$ and $\|\cdot\|_F$ is the Frobenius norm.

[2] Note that the sample domain $\mathfrak{Z}$ is *not* assumed compact or bounded, but the set of *parameters* $\mathcal{T}$ is in general assumed bounded.

[3] Notation in Appendix A.

We can now apply Lemma A.3.2 to bound the covering numbers of the set of Gaussians.

Lemma 3.3.5. *For any family of random feature functions $\mathcal{F}_R$, we have: for all $\delta > 0$,*

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{T}, \delta\right) \leq \max\left(\left(\frac{A}{\delta}\right)^d, 1\right) \cdot \max\left(\left(\frac{B}{\delta}\right)^d, 1\right). \tag{3.38}$$

with $A = 8B_{\mathcal{F}_R}R_{\boldsymbol{\mu}}/\sigma_{\min}$ and $B = 8B_{\mathcal{F}_R}\sqrt{2}R_{\boldsymbol{\sigma}}/\sigma_{\min}^2$.

Now that we have shown that the covering numbers of the set of single Gaussians are bounded, we can use Lemma 3.3.3 to bound the covering numbers of the set of GMMs

$$\mathfrak{S}_k\left(\mathfrak{T}\right) = \left\{ \pi_{\Theta,\boldsymbol{\xi}} = \sum_{l=1}^{k} \boldsymbol{\xi}_l \pi_{\boldsymbol{\theta}_l} \;\middle|\; \boldsymbol{\xi} \in \mathbb{S}^{k-1},\ \boldsymbol{\theta}_l \in \mathcal{T} \right\}$$

where the *support* of the mixture is denoted $\Theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$. Then we use Corollary 3.3.2 to directly obtain a recovery result.

Corollary 3.3.6. *Consider $\pi^\star \in \mathfrak{P}$ and $\pi_{\mathfrak{S}} \in \mathfrak{S}_k\left(\mathfrak{T}\right)$ which is a good approximation of $\pi^\star$ in terms of Kullback-Leibler (KL) divergence (Definition A.1.7 in Appendix A), denote the bias*

$$\tau = \left(D_{KL}\left(\pi^\star || \pi_{\mathfrak{S}}\right) + D_{KL}\left(\pi_{\mathfrak{S}} || \pi^\star\right)\right)^{\frac{1}{2}} \tag{3.39}$$

(note that we use a symmetrized version of the KL divergence).

Consider any kernel κ with a RF expansion $(\mathcal{F}_R, \Lambda)$. Let $\rho, \eta > 0$ be two constants, assume the sketch size satisfies

$$m \geq 4cB_{\mathcal{F}_R}^2\eta^{-2}\left[dk\left(\log_+\left(\frac{128B_{\mathcal{F}_R}A}{\eta^2}\right) + \log_+\left(\frac{128B_{\mathcal{F}_R}B}{\eta^2}\right)\right) + 2k\log\left(\frac{32B_{\mathcal{F}_R}}{\eta}\right) + \log\left(\frac{1}{\rho}\right)\right]$$

where $c = 1760/147$, $\log_+ = \max(\log, 0)$, and A, B are defined as in Lemma 3.3.5.

Consider items $\mathbf{z}_1, ..., \mathbf{z}_n \in \mathbb{R}^d$ drawn i.i.d. *from $\pi^\star$ and frequencies $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m \in \Omega$ drawn* i.i.d. *from Λ, which define the sketching operator $\mathcal{A}$ by (3.7). Denote $\pi_{\widetilde{\Theta},\widetilde{\boldsymbol{\xi}}} = \Delta_\iota(\mathcal{A}, \mathcal{A}\hat{\pi}_n)$ the GMM recovered from the empirical sketch. Then, with probability at least $1 - (\rho + \rho')$ on the drawing of the $\mathbf{z}_i$'s and $\boldsymbol{\omega}_j$'s, it holds that*

$$\left\|\pi^\star - \pi_{\widetilde{\Theta},\widetilde{\boldsymbol{\xi}}}\right\|_\kappa \leq 5B_{\mathcal{F}_R}\tau + \frac{4B_{\mathcal{F}_R}\left(1 + \sqrt{2\log(1/\rho')}\right)}{\sqrt{n}} + \eta + 2\iota\,. \tag{3.40}$$

Proof. Using Lemma 3.3.5 with Lemma 3.3.3, for all $0 < \delta \leq 16B_{\mathcal{F}_R}$ we can bound

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{S}_k\left(\mathfrak{T}\right), \delta\right) \leq \left(\frac{16B_{\mathcal{F}_R}}{\delta}\right)^k \cdot \max\left(\left(\frac{2A}{\delta}\right)^{dk}, 1\right) \cdot \max\left(\left(\frac{2B}{\delta}\right)^{dk}, 1\right)$$

where A, B are defined as in Lemma 3.3.5.

The desired result is then directly obtained as a consequence of Corollary 3.3.2, where the covering numbers of $\mathfrak{S}_k\left(\mathfrak{T}\right)$ are taken with $\delta := \frac{\eta^2}{64B_{\mathcal{F}_R}}$ (such that $\delta \leq B_{\mathcal{F}_R}/16$ is indeed verified since we assumed $\eta \leq 2B_{\mathcal{F}_R}$ in the whole chapter).

We use the additional bound on the bias term with Lemma 3.1.5:

$$\|\pi^\star - \pi_{\mathfrak{S}}\|_\kappa + 4\|\pi^\star - \pi_{\mathfrak{S}}\|_{\mathcal{F}_R} \leq 5\|\pi^\star - \pi_{\mathfrak{S}}\|_{\mathcal{F}_R} \leq 5B_{\mathcal{F}_R}\|\pi^\star - \pi_{\mathfrak{S}}\|_{\mathrm{TV}}$$

Then apply Pinsker's inequality [FHT03]:

$$\|\pi - \pi'\|_{\mathrm{TV}}^2 \leq 2D_{\mathrm{KL}}\left(\pi||\pi'\right)$$

from which

$$\|\pi^\star - \pi_{\mathfrak{S}}\|_{\mathrm{TV}} \leq \sqrt{D_{\mathrm{KL}}\left(\pi^\star||\pi_{\mathfrak{S}}\right) + D_{\mathrm{KL}}\left(\pi_{\mathfrak{S}}||\pi^\star\right)} = \tau \ .$$

□

As mentioned before, here we have used the "weak" version of the admissibility condition, and therefore obtained recovery results with an additive error in $\eta = \mathcal{O}\left(1/\sqrt{m}\right)$. It only proves that stable recovery is possible when the sketch is as large as the original database, and is strongly sub-optimal compared to what we could expect from a compressive GMM estimation method (as we will see, fortunately we can do much better in practice). Nevertheless, we obtain a novel GMM estimator that is in particular stable to modeling error.

Remark 3.3.7 (Role of characteristic kernels.)**.** *The result above is valid for* any *RF expansion* $(\mathcal{F}_R, \Lambda)$ *with bounded family of features* $\mathcal{F}_R$*. This can seem absurd: does it imply that taking* e.g. *a singleton family* $\mathcal{F}_R = \{1\}$ *allows for recovery? This is of course not the case: in this case the kernel is not* characteristic, i.e. *the MMD* $\|\cdot\|_\kappa$ *is not a proper metric, and may be meaningless in terms of recovery. This illustrates the importance of characteristic kernels to avoid trivial results, even though no result in this thesis requires the kernel to be characteristic.*

3.3.4 Example 2: Mixture of elliptic stable distributions

Let us now turn to a second example: estimation of mixtures of multivariate elliptic stable distributions.

> Stable distributions are also called α-*stable* distributions, however in the thesis we do not use this term: it somehow suggests that the so-called characteristic exponent α is fixed, while on the contrary in the considered mixtures the parameter α is *different for each component*. Hence we simply say "stable distribution".

Stable distributions have been proven to be useful to model asymmetric and/or heavy-tailed distributions. Although first and foremost used in practice in economy [GK99; NPM01; Nol03], they have also been introduced in signal processing [Cas04; BKK13]. However their use remains very scarce in this domain: they are indeed notoriously hard to estimate, mainly due to the intractability of the likelihood except for few special cases. In particular, they have been limited by a) the heavy use of approximate integrals which are time consuming and make the estimation intractable on large databases and b) the lack of available estimators for mixtures of *multivariate* data.

Although multimodal distributions are very common in some domains, very few works address the problem of estimating *mixtures* of stable distributions [SGKR09; SGKR10; Sho+10], and to our knowledge existing methods are restricted to the univariate case. The theoretical results presented here and the implementation described in Chapter 5 constitute the first valid method of which we are aware.

Elliptic stable distributions. Generic multivariate stable distributions are not parametrized by finite dimensional quantities but by a measure called the spectral measure, which renders their estimation extremely arduous. As often done in the literature, we thus restrict to the so-called *elliptic stable distributions* [Ome15; Nol13; Kri+09].

> In the literature (*e.g.* in [Ome15]), elliptic stable distributions are sometimes referred to as *subgaussian* stable distributions. However we do not use this term to avoid the confusion with *subgaussian random variables*, which is an unrelated notion, and for instance often encountered in Compressive Sensing.

An elliptic stable distribution $\pi_{\boldsymbol{\theta}}$ is parameterized by $\boldsymbol{\theta} := (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha)$ where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean of the distribution, $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is a positive definite matrix referred to as the dispersion

matrix, and $\alpha \in (0, 2]$ is the so-called characteristic exponent of the distribution. In general, stable distributions do not have explicit likelihood, or even second order moments. They are defined by their characteristic function $\psi_\pi(\boldsymbol{\omega}) := \mathbb{E}_{\mathbf{z}\sim\pi} e^{\mathrm{i}\boldsymbol{\omega}^\top \mathbf{z}}$, which in the case of elliptic stable distribution has the form:

$$\psi_{\pi_{\boldsymbol{\theta}}}(\boldsymbol{\omega}) = e^{\mathrm{i}\boldsymbol{\omega}^\top \boldsymbol{\mu}} e^{-\left(\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}\boldsymbol{\omega}\right)^{\alpha/2}} \tag{3.41}$$

When $\alpha = 2$ the distribution $\pi_{\boldsymbol{\theta}}$ is a Gaussian, with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. This is the only particular case for which the distribution has a second order moment.

Bounded parameters. Similar to the GMM case, we will restrict to diagonal matrices $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\sigma})$, and restrict the parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \alpha)$ to a set $\mathcal{T} := \mathcal{D}_{\boldsymbol{\mu}} \times \mathcal{D}_{\boldsymbol{\sigma}} \times \mathcal{D}_\alpha$ such that (3.34) holds, *i.e.* $\mathcal{D}_{\boldsymbol{\mu}}$ is a ball of radius $R_{\boldsymbol{\mu}}$, the set $\mathcal{D}_{\boldsymbol{\sigma}}$ is such that (3.35) and (3.36) hold, and $\mathcal{D}_\alpha = [\alpha_{\min}, 2]$ where $\alpha_{\min} > 0$. The set of stable distributions is denoted by $\mathfrak{T} = \{\pi_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathcal{T}\}$.

Like the previous case we begin by proving that the embedding $\boldsymbol{\theta} \mapsto \pi_{\boldsymbol{\theta}}$ is Lipschitz continuous. However, unlike GMMs proving this property for the TV norm seems difficult due to the lack of closed form expressions. Since the characteristic function of stable distribution is convenient to work with, in this section we directly work with the norm $\|\cdot\|_{\mathcal{F}_R}$ where the feature family $\mathcal{F}_R$ is the family of *Fourier* features (unlike GMMs where our results hold for all RF expansions), defined as:

$$\mathcal{F}_R := \left\{ \phi_{\boldsymbol{\omega}} : \mathbf{z} \mapsto e^{\mathrm{i}\boldsymbol{\omega}^\top \mathbf{z}} \;\middle|\; \boldsymbol{\omega} \in \mathbb{R}^d \right\}$$

We prove the following Lipschitz continuity property.

Lemma 3.3.8. *Consider $\pi_{\boldsymbol{\theta}_1}, \pi_{\boldsymbol{\theta}_2} \in \mathfrak{T}$. We have*

$$\|\pi_{\boldsymbol{\theta}_1} - \pi_{\boldsymbol{\theta}_2}\|_{\mathcal{F}_R} \leq L_1 \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2 + L_2 \|\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2\|_2 + L_3 |\alpha_1 - \alpha_2| \tag{3.42}$$

with

$$L_1 = \sup_{x>0,\ \alpha\in[\alpha_{\min},2]} x e^{-\left(\frac{\sigma_{\min}^2}{2} x^2\right)^{\alpha/2}}$$

$$L_2 = \sup_{x>0,\ \alpha\in[\alpha_{\min},2]} \frac{\alpha}{2(\sigma_{\min}^2)^{1-\alpha/2}} x^\alpha e^{-(x\sigma_{\min})^\alpha}$$

$$L_3 = \sup_{x>0,\ \alpha,\alpha'\in[\alpha_{\min},2]} \left|\log\left(\sqrt{R_{\boldsymbol{\sigma}}} x\right)\right| \left(\sqrt{R_{\boldsymbol{\sigma}}} x\right)^{\alpha'} e^{-(x\sigma_{\min})^\alpha}$$

Note that it is easy to prove that these constants are finite, we leave this proof for the reader (explicit closed-form expressions are not crucial here).

With this Lipschitz property we can bound the covering numbers of the set of stable distributions.

Lemma 3.3.9. *For all δ we have*

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{T}, \delta\right) \leq \max\left(\left(\frac{A}{\delta}\right)^d, 1\right) \cdot \max\left(\left(\frac{B}{\delta}\right)^d, 1\right) \cdot \max\left(\frac{C}{\delta}, 1\right) \tag{3.43}$$

where $A = 12 R_{\boldsymbol{\mu}} L_1$, $B = 12 R_{\boldsymbol{\sigma}} L_2$, $C = 3(2 - \alpha_{\min}) L_3$ where L_1, L_2, L_3 are defined as in Lemma 3.3.8.

Finally, we obtain the same corollary as in the Gaussian case.

Corollary 3.3.10. *Consider $\pi^\star \in \mathfrak{P}$ and a mixture of α-stable $\pi_{\mathfrak{S}} \in \mathfrak{S}_k(\mathfrak{T})$ which is a good approximation of $\pi^\star$ in terms of KL-divergence, denote*

$$\tau := \left(D_{KL}\left(\pi^\star || \pi_{\mathfrak{S}}\right) + D_{KL}\left(\pi_{\mathfrak{S}} || \pi^\star\right)\right)^{\frac{1}{2}} \tag{3.44}$$

(note that for any distribution $\pi^\star$ the bias term can be chosen at least as small as in the GMM case, since the model of mixtures of elliptic stable distributions strictly includes that of GMMs).

Consider any kernel κ with a RF expansion $(\mathcal{F}_R, \Lambda)$ where $\mathcal{F}_R$ are the Fourier features. Let $\rho, \eta > 0$ be two constants, assume the sketch size satisfies

$$m \geq 4c\eta^{-2}\left[dk\left(\log_+\left(\frac{128A}{\eta^2}\right) + \log_+\left(\frac{128B}{\eta^2}\right)\right) + k\left(\log_+\left(\frac{128C}{\eta^2}\right) + 2\log\left(\frac{32}{\eta}\right)\right) + \log\left(\frac{1}{\rho}\right)\right]$$

where $c = 1760/147$ and A, B, C is defined as in Lemma 3.3.9.

Consider items $\mathbf{z}_1, ..., \mathbf{z}_n \in \mathbb{R}^d$ drawn i.i.d. *from $\pi^\star$ and frequencies $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m \in \mathbb{R}^d$ drawn* i.i.d. *from Λ, which define the sketching operator $\mathcal{A}$ by (3.7). Denote $\pi_{\widetilde{\Theta},\widetilde{\boldsymbol{\xi}}} = \Delta_\iota(\mathcal{A}, \mathcal{A}\hat{\pi}_n)$ the probability distribution recovered from the empirical sketch. Then, with probability at least $1 - (\rho + \rho')$ on the drawing of the $\mathbf{z}_i$'s and $\boldsymbol{\omega}_j$'s, it holds that*

$$\left\|\pi^\star - \pi_{\widetilde{\Theta},\widetilde{\boldsymbol{\xi}}}\right\|_\kappa \leq 5\tau + \frac{4\left(1 + \sqrt{2\log(1/\rho')}\right)}{\sqrt{n}} + \eta + 2\iota\,. \tag{3.45}$$

Proof. The proof is exactly the same than that of Corollary 3.3.6, but with $B_{\mathcal{F}_R} = 1$. □

As in the previous case for GMM estimation, here the weak admissibility condition yields results that hold only when the sketch is as large as the database, due to the presence of the error $\eta = \mathcal{O}\left(1/\sqrt{m}\right)$. However, if in the previous case methods that use the full database do exist and these sub-optimal theoretical results were of limited interest, here on the contrary we obtain, to our knowledge, the first estimator for mixtures of multivariate elliptic stable distribution with provable guarantees. As we will see in Chapter 5, this theoretical sufficient sketch size is also sub-optimal compared to what is observed in practice.

3.4 Conclusion

In this chapter, we presented the general principles of our main theoretical contribution.

In Section 3.1 and 3.2, we established the connection between the sketching methodology and kernel mean embedding, using Random Features expansions. Using a strategy inspired by Compressive Sensing in infinite-dimensional spaces, we proved that the LRIP holds when the RF expansion satisfies some admissibility condition, and the normalized secant sets of the low-dimensional model have finite covering numbers. These somewhat technical assumptions will be developed in Chapter 6.

We proved that sub-optimal recovery results can already be obtained with simpler assumptions, as soon as the model itself has finite covering numbers. We applied this result to two models: GMM with diagonal covariance and mixtures of elliptic stable distributions. In this second case in particular, even with sub-optimal sketch size the obtained result constitutes the first estimator with provable guarantees.

We therefore showed that the parallel between the sketching method and Compressive Sensing initiated by Bourrier et al. can be extended to obtain theoretical guarantees, by reworking usual proof techniques and using adapted tools from kernel mean embedding and Random Features. By seeing the RF expansion as a *compressed representation* of the distribution instead of a finite mapping that aims at approximating the high-dimensional embedding induced by the kernel, we make new connections between these fields.

Chapter 4

A Greedy Algorithm for Learning Mixture Models

Context. In Chapter 3, we introduced conditions under which the following minimization problem yields an instance optimal decoder:

$$\arg\min_{\pi \in \mathfrak{S}} \|\mathcal{A}\pi - \hat{\mathbf{y}}\|_2 \tag{4.1}$$

where $\hat{\mathbf{y}} \in \mathbb{C}^m$ is the pre-computed sketch of a database, $\mathfrak{S} \subset \mathfrak{P}(\mathfrak{Z})$ is a low-dimensional model of probability distributions and $\mathcal{A} : \mathfrak{M} \to \mathbb{C}^m$ is a linear sketching operator, built in our analysis by combining RF expansions and kernel mean embedding.

Consider the case where the model is formed by mixtures of distributions. Denote $\mathfrak{T} = \{\pi_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathcal{T}\}$ a parametric set of distributions, where $\mathcal{T} \subset \mathbb{R}^q$. For $k > 0$ the mixture model is defined as

$$\mathfrak{S} := \mathfrak{S}_k(\mathfrak{T}) = \left\{ \pi_{\Theta,\boldsymbol{\xi}} = \sum_{l=1}^{k} \xi_l \pi_{\boldsymbol{\theta}_l} \;\middle|\; \boldsymbol{\theta}_l \in \mathcal{T},\ \boldsymbol{\xi} \in \mathbb{S}^{k-1} \right\} . \tag{4.2}$$

where $\Theta = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k) \in \mathfrak{T}^k$ denotes the tuple of parameters. In that case the sketch is a linear combination of k *atoms* in the dictionary $\{\mathcal{A}\pi_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}}$, continuously indexed by $\boldsymbol{\theta}$.

This chapter. In this chapter, we develop a greedy approach, inspired by Compressive Sensing, to handle the estimation problem (4.1) in the case of mixture models. In fact, the proposed algorithm is applicable to any generic minimization problems of the form

$$\min_{\Theta \in \mathcal{T}^k,\ \boldsymbol{\xi} \in \mathbb{S}^{k-1}} G(\Theta, \boldsymbol{\xi}) \tag{4.3}$$

where, given a vector $\mathbf{y} \in \mathbb{C}^m$ and defining a smooth function $\mathbf{f} : \mathcal{T} \to \mathbb{C}^m$, the cost function has the form

$$G(\Theta, \boldsymbol{\xi}) = \left\| \mathbf{y} - \sum_{l=1}^{k} \xi_l \mathbf{f}(\boldsymbol{\theta}_l) \right\|_2^2 , \tag{4.4}$$

When cast for a mixture model $\mathfrak{S} = \mathfrak{S}_k(\mathfrak{T})$, the estimation problem (4.1) is indeed of the form (4.4) by defining

$$\mathbf{f}(\boldsymbol{\theta}_l) := \mathcal{A}\pi_{\boldsymbol{\theta}_l} , \tag{4.5}$$

and therefore the proposed algorithm is applicable to *any mixture model and sketching operator*, as long as $\boldsymbol{\theta} \mapsto \mathcal{A}\pi_{\boldsymbol{\theta}}$ and its derivative have a closed-form expression.

The layout of the chapter is the following.

- In Section 4.1 we realize a preliminary study of the cost function, and introduce a naive approach to minimize it.
- In Section 4.2 we describe the proposed greedy approach, inspired by the OMP algorithm adapted to continuous settings.
- In Section 4.3 we give some implementation details about the proposed algorithm.

– we finish by a comparison of the algorithm on a simple example problem in Section 4.4, before the applications to the sketched mixture model estimation problem in the next chapter.

4.1 Brief study of the cost function, naive algorithm

As mentioned in Section 1.2, a simple method to handle 4.3 consists in discretizing the parameter space $\mathcal{T} \subset \mathbb{R}^q$ and optimizing using this finite grid either by brute-force or, *e.g.* , using methods inspired by CS [Bun+10]. However this approach quickly fails in moderate or high dimension q. We directly derive an algorithm in the continuous settings instead. Let us first briefly examine the cost function G and its derivatives.

Notations. For a univariate function $f : \mathbb{R}^q \to \mathbb{R}$, we denote $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = \left[\frac{\partial}{\partial \theta_i} f(\boldsymbol{\theta})\right]_{i=1}^{q} \in \mathbb{R}^q$ its gradient, and $\mathbf{H}_f(\boldsymbol{\theta}) = \left[\frac{\partial^2}{\partial \theta_i \partial \theta_l} f(\boldsymbol{\theta})\right]_{\substack{i=1,q \\ l=1,q}} \in \mathbb{R}^{q \times q}$ its Hessian matrix. For a multivariate function $\mathbf{f} : \mathbb{R}^q \to \mathbb{R}^m$, we denote $f_j : \mathbb{R}^q \to \mathbb{R}$ its j^{th} coordinate and $\mathbf{J}_{\mathbf{f}}(\boldsymbol{\theta}) = \left[\frac{\partial}{\partial \theta_i} f_j(\boldsymbol{\theta})\right]_{\substack{j=1,m \\ i=1,q}} \in \mathbb{R}^{m \times q}$ its Jacobian matrix (*i.e.* whose rows are $\nabla_{\boldsymbol{\theta}} f_j(\boldsymbol{\theta})^\top$). When a function $f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ depends on several parameters, the Hessian may be only derived with respect to the parameters $\boldsymbol{\theta}_1$ and written $\mathbf{H}_f(\boldsymbol{\theta}_1)$ (which is here evaluated in $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ where $\boldsymbol{\theta}_2$ is defined with no ambiguity in the text) and similarly for the Jacobian matrix of $\mathbf{f}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Finally, for a vector $\mathbf{x} \in \mathbb{R}^d$ the notations $\mathrm{Re}(\mathbf{x}) \in \mathbb{R}^d$ and $\mathrm{Im}(\mathbf{x}) \in \mathbb{R}^d$ indicate the vectors formed respectively by the real part and imaginary part of its entries, and similarly for a complex multivariate function $\mathbf{f} : \mathbb{R}^q \to \mathbb{C}^m$, the notation $\mathrm{Re}(\mathbf{f})$ (resp. $\mathrm{Im}(\mathbf{f})$) indicates the function from $\mathbb{R}^q$ to $\mathbb{R}^m$ formed by the real part (resp. imaginary part) of its coordinates.

Minimization with respect to the weights. When all parameters $\boldsymbol{\theta}_l$ are fixed, minimizing $\boldsymbol{\xi} \mapsto G(\Theta, \boldsymbol{\xi})$ is a convex problem, since it is a simple Least-Square problem restricted to the convex domain $\mathbb{S}^{k-1}$. There is therefore a whole range of methods to solve this problem, see *e.g.* the book by Boyd and Vandenberghe [BV04].

Considering this, a good intuitive strategy is to alternatively minimize $G(\Theta, \boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}$ and Θ. Similar to usual CS, finding the true support $\Theta^\star$ is much more challenging than estimating the weights.

Non-linear Least Squares. Consider now that the weights $\boldsymbol{\xi}$ are fixed. With respect to the hyper parameter Θ, the minimization of the cost function (4.4) is a Non-Linear Least Squares (NLLS) problem [Mar63; MNT04]. The gradient of the cost function reads $\nabla_\Theta G(\Theta, \boldsymbol{\xi}) = [\nabla_{\boldsymbol{\theta}_l} G(\Theta, \boldsymbol{\xi})]_{l=1}^k \in \mathbb{R}^{kq}$, with

$$\nabla_{\boldsymbol{\theta}_l} G(\Theta, \boldsymbol{\xi}) = -2\xi_l \left(\mathbf{J}_{\mathrm{Re}(\mathbf{f})}(\boldsymbol{\theta}_l)^\top \mathrm{Re}(\mathbf{r}) + \mathbf{J}_{\mathrm{Im}(\mathbf{f})}(\boldsymbol{\theta}_l)^\top \mathrm{Im}(\mathbf{r}) \right) , \tag{4.6}$$

where $\mathbf{r} = \mathbf{r}(\Theta, \boldsymbol{\xi}) = \mathbf{y} - \sum_{l=1}^k \xi_l \mathbf{f}(\boldsymbol{\theta}_l)$ is the residual in $\mathbb{C}^m$ (which also depends on Θ). Putting all these gradients to zero leads to the *normal equations* that characterize a stationary point of the problem. Nevertheless, unlike classic Least Squares, in general stationary points cannot be explicitly characterized, hence the need for iterative procedures. Although there is a growing literature on non-convex optimization problems [SQW15] and their properties, we leave this type of study of the cost function for future work and focus on developing a heuristic that works well in practice.

Numerous algorithms can be applied to the NLLS problem, from simple gradient descents to more advanced Gauss-Newton type methods such as the well-known Levenberg-Marquardt algorithm [Mar63]. See [MNT04] for a good summary. Nevertheless, NLLS problems are generally non-convex and notoriously hard to initialize. A rule of thumb is that the initial guess must be somewhat close to the optimum: to illustrate this claim let us attempt to prove that G is convex with respect to $\Theta \in \mathcal{T}^k$. Denote $\mathbf{F} : \Theta \mapsto \sum_{l=1}^k \xi_l \mathbf{f}(\boldsymbol{\theta}_l)$ (recall that the

weights are fixed), the Hessian matrix of the cost function G with respect to the parameter Θ reads

$$\mathbf{H}_G(\Theta) = 2\left(\mathbf{J}_{\mathrm{Re}(\mathbf{F})}(\Theta)^\top \mathbf{J}_{\mathrm{Re}(\mathbf{F})}(\Theta) + \mathbf{J}_{\mathrm{Im}(\mathbf{F})}(\Theta)^\top \mathbf{J}_{\mathrm{Im}(\mathbf{F})}(\Theta)\right) - 2\sum_{j=1}^{m}\left(\mathrm{Re}(r_j)\mathbf{H}_{\mathrm{Re}(F_j)}(\Theta) + \mathrm{Im}(r_j)\mathbf{H}_{\mathrm{Im}(F_j)}(\Theta)\right) . \tag{4.7}$$

The right-hand side of this expression is a sum of two terms. The first is necessarily positive semi-definite, while the second is not. Nevertheless, when the residual $\mathbf{r}$ goes to 0 the second term may become negligible compared to the first one, in which case the problem is locally convex around the optimum.

However, in practical situations this condition might still be arduous to derive with respect to the full parameter Θ.

Block convexity. Although it is in general difficult to guarantee local convexity with respect to the whole parameter Θ, it is sometimes possible to characterize local convexity with respect to *one* parameter $\boldsymbol{\theta}_l$, which is referred to as *block* convexity in the literature [Tse01; XY13]. The Hessian matrix of the cost function G with respect to the parameter $\boldsymbol{\theta}_l$ reads:

$$\mathbf{H}_G(\boldsymbol{\theta}_l) = 2\xi_l^2\left(\mathbf{J}_{\mathrm{Re}(\mathbf{f})}(\boldsymbol{\theta}_l)^\top \mathbf{J}_{\mathrm{Re}(\mathbf{f})}(\boldsymbol{\theta}_l) + \mathbf{J}_{\mathrm{Im}(\mathbf{f})}(\boldsymbol{\theta}_l)^\top \mathbf{J}_{\mathrm{Im}(\mathbf{f})}(\boldsymbol{\theta}_l)\right) - 2\xi_l\sum_{j=1}^{m}\left(\mathrm{Re}(r_j)\mathbf{H}_{\mathrm{Re}(f_j)}(\boldsymbol{\theta}_l) + \mathrm{Im}(r_j)\mathbf{H}_{\mathrm{Im}(f_j)}(\boldsymbol{\theta}_l)\right) , \tag{4.8}$$

hence, since $\xi_l \geq 0$, it is positive semi-definite if and only if: for all vectors $\mathbf{x} \in \mathbb{R}^q$ we have

$$\xi_l\left(\left\|\mathbf{J}_{\mathrm{Re}(\mathbf{f})}(\boldsymbol{\theta}_l)\,\mathbf{x}\right\|_2^2 + \left\|\mathbf{J}_{\mathrm{Im}(\mathbf{f})}(\boldsymbol{\theta}_l)\,\mathbf{x}\right\|_2^2\right) - \sum_{j=1}^{m}\left(\mathrm{Re}(r_j)\mathbf{x}^\top\mathbf{H}_{\mathrm{Re}(f_j)}(\boldsymbol{\theta}_l)\,\mathbf{x} + \mathrm{Im}(r_j)\mathbf{x}^\top\mathbf{H}_{\mathrm{Im}(f_j)}(\boldsymbol{\theta}_l)\,\mathbf{x}\right) \geq 0 . \tag{4.9}$$

Once again this condition might be satisfied when the residual signal is sufficiently small. This is best illustrated on an example, see also Fig. 4.1.

Example 4.1.1. *Define m frequency vectors $\boldsymbol{\omega}_j \in \mathbb{R}^q$, and consider the complex exponential functions*

$$f_j(\boldsymbol{\theta}) := e^{i\boldsymbol{\omega}_j^\top\boldsymbol{\theta}} . \tag{4.10}$$

In that case, we have

$$\nabla\mathrm{Re}(f_j)(\boldsymbol{\theta}) = -\sin(\boldsymbol{\omega}_j^\top\boldsymbol{\theta})\boldsymbol{\omega}_j , \qquad \nabla\mathrm{Im}(f_j)(\boldsymbol{\theta}) = \cos(\boldsymbol{\omega}_j^\top\boldsymbol{\theta})\boldsymbol{\omega}_j$$

and

$$\mathbf{H}_{\mathrm{Re}(f_j)}(\boldsymbol{\theta}) = -\cos(\boldsymbol{\omega}_j^\top\boldsymbol{\theta})\boldsymbol{\omega}_j\boldsymbol{\omega}_j^\top , \qquad \mathbf{H}_{\mathrm{Im}(f_j)}(\boldsymbol{\theta}) = -\sin(\boldsymbol{\omega}_j^\top\boldsymbol{\theta})\boldsymbol{\omega}_j\boldsymbol{\omega}_j^\top .$$

Hence, denoting $C_j = \cos(\boldsymbol{\omega}_j^\top\boldsymbol{\theta}_l)$ and $S_j = \sin(\boldsymbol{\omega}_j^\top\boldsymbol{\theta}_l)$, the left hand side of (4.9) reads

$$\begin{aligned}\xi_l\sum_{j=1}^{m}\left(S_j^2\left(\boldsymbol{\omega}_j^\top\mathbf{x}\right)^2 + C_j^2\left(\boldsymbol{\omega}_j^\top\mathbf{x}\right)^2\right) + \sum_{j=1}^{m}\left(C_j\mathrm{Re}(r_j)\left(\boldsymbol{\omega}_j^\top\mathbf{x}\right)^2 + S_j\mathrm{Im}(r_j)\left(\boldsymbol{\omega}_j^\top\mathbf{x}\right)^2\right) &= \sum_{j=1}^{m}\left(\boldsymbol{\omega}_j^\top\mathbf{x}\right)^2\left(\xi_l + C_j\mathrm{Re}(r_j) + S_j\mathrm{Im}(r_j)\right) \\ &\geq \sum_{j=1}^{m}\left(\boldsymbol{\omega}_j^\top\mathbf{x}\right)^2\left(\xi_l - |r_j|\right) .\end{aligned}$$

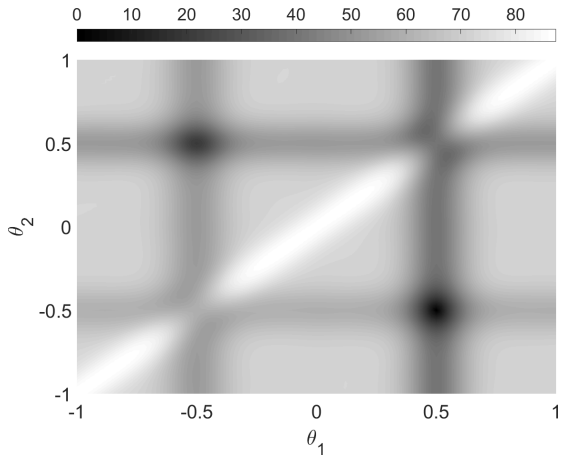

FIGURE 4.1: Value of the cost function for a toy instantiation of Example 4.1.1, for $k = 2$ components with parameter dimension $q = 1$ and measurement vector dimension $m = 5000$. We first draw $\omega_1, \dots, \omega_m$ from $\mathcal{N}(0, 25)$, and define the functions $f_j(\theta) := e^{i\omega_j \theta}$. The measurement vector is defined as $\mathbf{y} = 0.6\mathbf{f}(0.5) + 0.4\mathbf{f}(-0.5)$, and we display the cost function with the weights put at their true value $G(\theta_1, \theta_2) = \|0.6\mathbf{f}(\theta_1) + 0.4\mathbf{f}(\theta_2) - \mathbf{y}\|_2^2$. We can clearly see that the cost function is indeed not globally convex. We observe a basin of attraction at the true values of the parameters $\theta_1 = 0.5, \theta_2 = -0.5$, but also when the values of the parameters are exchanged (which is equivalent to approaching the vector $\mathbf{y}$ with the right parameters θ_l but exchanged weights).

Hence, when each coordinate of the residual is such that $|r_j| \leq \xi_l$, the cost function G is convex with respect to $\boldsymbol{\theta}_l$. Furthermore, if $m \geq q$ and the m frequency vectors $\boldsymbol{\omega}_j \in \mathbb{R}^q$ span the whole space $\mathbb{R}^q$, this convexity is strict.

Note that this somehow supports the intuitive fact that components with a higher weight ξ_l are more easily identified than components with a low weight, since the above condition is valid for a larger range of residuals.

Block coordinate descent algorithm. This possible block convexity encourages us to apply a *block coordinate* descent (BCD) algorithm [XY13], which is simply a minimization of the cost function G with respect to each $\boldsymbol{\theta}_l$ in turn, treated in a random order at each cycle, while all the other parameters are kept fixed. We describe this block coordinate descent algorithm in Algorithm 2. Note that the minimization with respect to $\boldsymbol{\theta}_l$, denoted by $\texttt{minimize}_{\boldsymbol{\theta}_l}$, can be done with any algorithm for Non-Linear Least Squares. In the absence of block convexity, it yields only a local minimum. After each cycle of going through each $\boldsymbol{\theta}_l$, we minimize the cost function with respect to $\boldsymbol{\xi}$ with a Non-Negative Least Squares (NNLS) [LH95] minimization algorithm, denoted $\texttt{minimize}_{\boldsymbol{\xi} \geq 0}$. Note that we do not enforce normalization $\sum_{l=1}^{k} \xi_l = 1$ at each iteration. Instead, a normalization of $\boldsymbol{\xi}$ is performed at the end of the algorithm[1].

There have been studies of the properties of the block coordinate descent algorithm [Tse01; XY13] in various mathematical contexts. However, we will see that in practice this approach has a very limited efficiency, and conditions for block convexity seem to be rarely met unless very close to the optimal solution. Instead, similar to Bourrier et al. [BGP13] (see Sec. 1.2), in the next section we propose a strategy inspired by classic Compressive Sensing and develop a greedy algorithm similar to Orthogonal Matching Pursuit and its variant OMP with Replacement (OMPR).

[1] Enforcing the normalization constraint at each iteration was found on initial experiments to have a negligible effect while increasing computation time

Algorithm 2: Block Coordinate Descent algorithm (BCD)

Data: Sketch $\mathbf{y}$, function $\mathbf{f}$, number of iterations T, initial parameters $(\Theta^0, \boldsymbol{\xi}^0)$
Result: Parameters $(\Theta, \boldsymbol{\xi})$
$(\Theta, \boldsymbol{\xi}) \leftarrow (\Theta^0, \boldsymbol{\xi}^0)$;
for $t \leftarrow 1$ **to** T **do**
  Draw a random permutation of $[\![1, k]\!]$ noted $\boldsymbol{\sigma}$;
  for $p \leftarrow 1$ **to** k **do**
    $l \leftarrow \boldsymbol{\sigma}(p)$;
    $\boldsymbol{\theta}_l \leftarrow \texttt{minimize}_{\boldsymbol{\theta} \in \mathcal{T}} \left(\left\| \mathbf{y} - \sum_{u \neq l} \xi_u \mathbf{f}(\boldsymbol{\theta}_u) - \xi_l \mathbf{f}(\boldsymbol{\theta}) \right\|_2^2, \texttt{init} = \boldsymbol{\theta}_l \right)$;
  end
  $\boldsymbol{\xi} \leftarrow \texttt{minimize}_{\boldsymbol{\xi} \geq 0} \left(\left\| \mathbf{y} - \sum_{l=1}^{k} \xi_l \mathbf{f}(\boldsymbol{\theta}_l) \right\|_2^2, \texttt{init} = \boldsymbol{\xi} \right)$;
end
Normalize $\boldsymbol{\xi}$ such that $\sum_{l=1}^{k} \xi_l = 1$;

4.2 Proposed greedy approach

The cost function (4.4) is a particular case of NLLS. Its main notable feature is the form of the function with which the vector $\mathbf{y}$ is approached:

$$\mathbf{F}_k(\Theta, \boldsymbol{\xi}) := \sum_{l=1}^{k} \xi_l \mathbf{f}(\boldsymbol{\theta}_l) \,. \tag{4.11}$$

Going back to the vocabulary of sparse recovery and Compressive Sensing of Section 1.2, the function $\mathbf{F}$ is a linear combination of a limited number of *atoms* in the *dictionary* $\mathcal{D} = \{\mathbf{f}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \mathcal{T}\}$. Like the usual Compressive Sensing settings, the main difficulty in solving (4.3) is to localize the *support* $\Theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$. Once this is done we have already mentioned that finding the weights $\boldsymbol{\xi}$ is less challenging.

Similar to the original work by Bourrier et al. [BGP13], this encourages us to adapt classic CS algorithms for sparse recovery to this continuous settings. In this fashion, Bourrier et al. develop an algorithm inspired by Iterative Hard Thresholding (IHT) [BD09], where the gradient step is approximated by an extension of the support in many directions[2], then reduce the support to a size k with Hard Thresholding, and reiterate. Although originally developed for the case of GMM with identity covariance estimation, this IHT algorithm can easily be adapted to any differentiable function $\mathbf{f}$. However we observed that, for certain richer mixture models like GMMs with unknown covariances, this IHT algorithm has limited efficiency and can easily get stuck into spurious local minima , due to the particular form of its approximate gradient step that adds many atoms at once (see Fig. 4.2). We found instead that it was preferable to proceed more carefully and add atoms only one at a time to the support, in a strategy similar to Matching Pursuit - based methods.

4.2.1 Compressive Learning - OMP (CL-OMP)

Matching Pursuit [MZ93] and Orthonormal Matching Pursuit (OMP) [PRK93] (see Alg. 1) deal with general sparse approximation problems. They gradually extend the sparse support by selecting atoms most correlated with the residual signal, until the desired sparsity is attained.

Adapting OMP to the considered framework requires several modifications. The main challenge stems from the continuous indexation of the dictionary (instead of having a finite dictionary as in classic CS), which *e.g.* prevents us from exactly maximizing the correlation between an atom and the residual signal and requires the use of a continuous optimization scheme instead.

We detail the modifications brought to OMP below, and summarize them in Algorithm 3. We call the resulting algorithm Compressive Learning - OMP (CL-OMP).

[2] Due to the form of this approximate gradient step, this IHT algorithm is in fact also similar to CoSAMP [FR13].

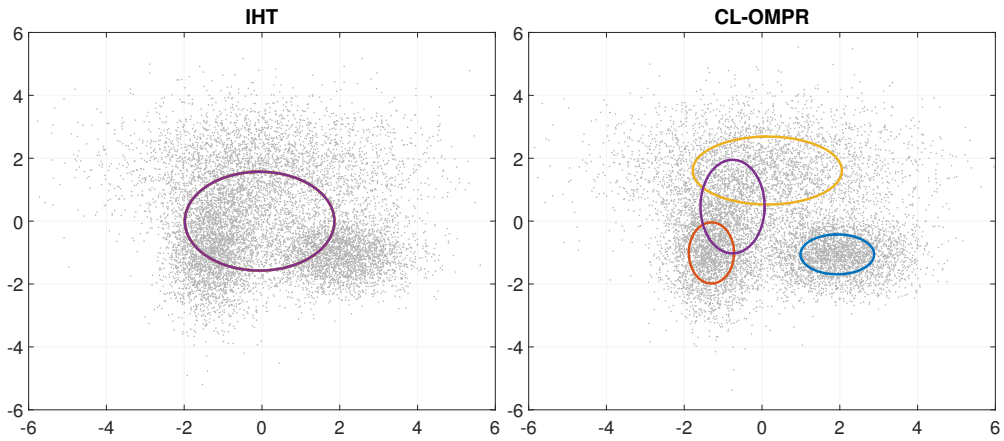

FIGURE 4.2: Comparison of the result obtained with IHT [BGP13] (left) and the proposed CL-OMPR (right) on the GMM with unknown covariance estimation problem of Chapter 5. During its step that adds many atoms at once, the IHT algorithm falls into a local minimum of the cost function where all Gaussians in the mixture are equal.

– **Non-negativity.** The compressive mixture estimation framework imposes a non-negativity constraint on the weights $\boldsymbol{\xi}$, that we enforce at each iteration. Thus Step 1 is modified compared to classic OMPR by replacing the *modulus* of the correlation by its *real part*, to avoid negative correlation between atom and residual. Similarly, in Step 4 we perform a Non-Negative Least-Squares (NNLS) [LH95] instead of a classic Least-Squares, like in the BCD algorithm.

– **Continuous dictionary.** As mentioned before the set $\mathcal{T}$ of parameters is continuously indexed and cannot be exhaustively searched. Instead we propose to replace the maximization in Step 1 of classic OMP with a randomly initialized gradient descent, denoted by a call to a sub-routine $\texttt{maximize}_{\boldsymbol{\theta}}$, leading to a – local – maximum of the correlation between atom and residual. Note that the atoms are normalized during the search, as is often the case with OMP.

– **Additional non-convex optimization step to handle coherent dictionaries.** Compared to classic OMP, the proposed algorithm includes a new step at each iteration (Step 4), which further reduces the cost function with a few gradient descent steps initialized with the *current* parameters $(\Theta, \boldsymbol{\xi})$. This is denoted by a call to the sub-routine $\texttt{minimize}_{\Theta,\boldsymbol{\xi}}$. The need for this additional step stems from the lack of incoherence between the elements of the uncountable dictionary. For instance, in the case of GMM estimation, *i.e.* $\mathbf{f}(\boldsymbol{\theta}) = \mathcal{A}\pi_{\boldsymbol{\theta}}$ where $\pi_{\boldsymbol{\theta}}$ is a GMM with unknown covariance, a $(k+1)$-GMM approximation of a distribution cannot be directly derived from a k-GMM by simply adding a Gaussian. An example run of CL-OMP for GMM estimation is shown in Figure 4.3, where the need for this additional non-convex update after each extension of the support is clearly visible.

  This is reminiscent of a similar problem handled in High Resolution Matching Pursuit (HRMP) [Gri+96], which uses a multi-scale decomposition of atoms, while we handle here the more general case of a continuous dictionary using a global gradient descent that adjusts all atoms.

In difficult cases, the CL-OMP algorithm can still be challenged: in particular, if during one iteration it has mistakingly placed a component $\boldsymbol{\theta}_l$ far from the true support, then even the gradient descent of Step 4 might not be enough to modify it significantly before the algorithm stops. Thus we also study a variant of OMP called OMP with Replacement (OMPR) [JTD11], that allows for the suppression of spurious atoms.

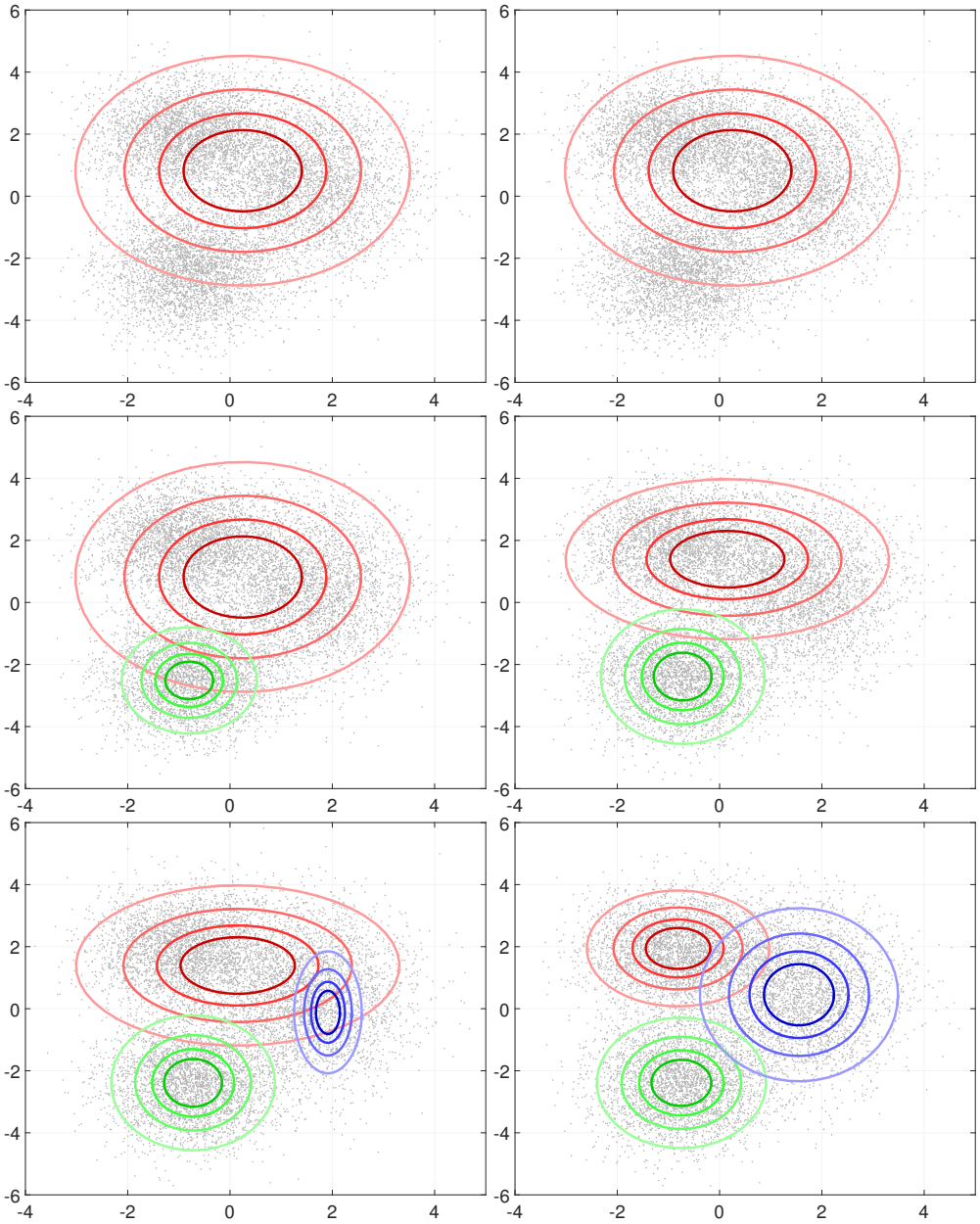

FIGURE 4.3: Step-by-step execution of the CL-OMP algorithm on a GMM estimation problem (Chapter 5) with three components in dimension $d = 2$. Although the data points are displayed here, recall that the algorithm has only access to the sketch of the data. The three iterations of CL-OMP are displayed from top to bottom, just after adding an atom (Step 1) on the left, and just after the gradient descent on all parameters (Step 4) on the right. The necessity for Step 4 is here clearly outlined: each time a Gaussian component is added, the previous ones must be displaced to accommodate for it.

Algorithm 3: Compressive Learning OMP (CL-OMP)

Data: Measurement vector $\mathbf{y} \in \mathbb{C}^m$, function $\mathbf{f} : \mathbb{R}^q \to \mathbb{C}^m$, sparsity $k > 0$
Result: Parameters $(\Theta, \boldsymbol{\xi})$
$\mathbf{r} \leftarrow \mathbf{y}$; $\Theta \leftarrow \emptyset$;
for $t \leftarrow 1$ **to** k **do**
  Step 1: Find a normalized atom highly correlated with the residual with a gradient descent (local maximum)
  $\boldsymbol{\theta} \leftarrow \texttt{maximize}_{\boldsymbol{\theta}'} \left(\mathrm{Re}\left(\left\langle \frac{\mathbf{f}(\boldsymbol{\theta}')}{\|\mathbf{f}(\boldsymbol{\theta}')\|_2}, \mathbf{r} \right\rangle \right), \texttt{init} = \texttt{rand} \right)$;
  end
  Step 2: Expand support
  $\Theta \leftarrow \Theta \cup \{\boldsymbol{\theta}\}$;
  end
  Step 3: Project to find weights
  $\boldsymbol{\xi} \leftarrow \texttt{minimize}_{\boldsymbol{\xi} \geq 0} \left\| \mathbf{y} - \sum_{l=1}^{t} \xi_l \mathbf{f}(\boldsymbol{\theta}_l) \right\|_2$;
  end
  Step 4: Perform a gradient descent *initialized with current parameters*
  $\Theta, \boldsymbol{\xi} \leftarrow \texttt{minimize}_{\Theta,\, \boldsymbol{\xi} \geq 0} \left(\left\| \mathbf{y} - \sum_{l=1}^{t} \xi_l \mathbf{f}(\boldsymbol{\theta}_l) \right\|_2, \texttt{init} . = (\Theta, \boldsymbol{\xi}) \right)$;
  end
  Update residual: $\mathbf{r} \leftarrow \mathbf{y} - \sum_{l=1}^{t} \xi_l \mathbf{f}(\boldsymbol{\theta}_l)$;
end
Normalize $\boldsymbol{\xi}$ such that $\sum_{l=1}^{k} \xi_l = 1$;

4.2.2 CL-OMP with Replacement

In classic CS, an efficient variation of OMP called OMP with Replacement (OMPR) [JTD11] exhibits better reconstruction guarantees. Inspired by IHT [BD09], and similar to CoSAMP or Subspace Pursuit [FR13], it increases the number of iterations of OMP, extends the size of the support *further than* the desired sparsity and reduces it with Hard Thresholding to suppress spurious atoms.

Our adaptation of OMPR is very simple: first we perform CL-OMP to find initial parameters. Then we continue to perform the same steps as CL-OMP for k additional iterations, with the addition of a Hard Thresholding step between Step 2 and Step 3 of CL-OMP to suppress one atom.

We obtain an algorithm coined CL-OMP with Replacement (CL-OMPR), described in Algo. 4. Although the modification from CL-OMP to CL-OMPR seems exceedingly simple, we will see that on a number of problems the CL-OMPR algorithm perform far better than CL-OMP.

Function HardThres$(\Theta, \mathbf{y}, \mathbf{f}, k)$: reduction of the support by Hard Thresholding

Data: Current support Θ, measurement vector $\mathbf{y} \in \mathbb{C}^m$, function $\mathbf{f} : \mathbb{R}^q \to \mathbb{C}^m$, desired sparsity k
Result: Reduced support Θ
$\widetilde{\boldsymbol{\xi}} \leftarrow \arg\min_{\boldsymbol{\xi} \geq 0} \left\| \mathbf{y} - \sum_{l=1}^{|\Theta|} \xi_l \frac{\mathbf{f}(\boldsymbol{\theta}_l)}{\|\mathbf{f}(\boldsymbol{\theta}_l)\|_2} \right\|_2$;
Select the k largest entries $\xi_{i_1} \geq \ldots \geq \xi_{i_k}$;
Reduce the support $\Theta \leftarrow (\boldsymbol{\theta}_{i_1}, \ldots, \boldsymbol{\theta}_{i_k})$;

Learning the number of components? In the proposed framework, the number of components k is known in advance and provided by the user. However, it is known that greedy approaches such as OMP are convenient to derive stopping conditions, that could be readily applied to CL-OMP: when the residual falls below a fixed (or adaptive) threshold, stop the algorithm (adapted strategies could be derived for CL-OMPR). In this thesis however, we only compare the proposed method with classic approaches such as the Expectation Maximization (EM) algorithm, that also consider the number of components k to be known in advance. We

Algorithm 4: Compressive Learning OMP with Replacement (CL-OMPR)

Data: Measurement vector $\mathbf{y} \in \mathbb{C}^m$, function $\mathbf{f} : \mathbb{R}^q \to \mathbb{C}^m$, sparsity k, additional number of iterations $T > 0$ (usually $T = k$)
Result: Parameters $(\Theta, \boldsymbol{\xi})$
First initialize with CL-OMP (Alg. 3): $(\Theta, \boldsymbol{\xi}) \leftarrow \texttt{CL-OMP}(\mathbf{y}, \mathbf{f}, k)$;
$\mathbf{r} \leftarrow \mathbf{y} - \sum_{l=1}^{k} \xi_l \mathbf{f}(\boldsymbol{\theta}_l)$;
for $t \leftarrow 1$ **to** T **do**
  Perform **Step 1** and **Step 2** of Alg. 3 (the support has now size $k+1$);
  Additional step : Suppress one atom by Hard Thresholding
    $\Theta \leftarrow \texttt{HardThres}(\Theta, \mathbf{y}, \mathbf{f}, k)$;
  end
  Perform **Step 3** and **Step 4** of Alg. 3;
  Update residual: $\mathbf{r} \leftarrow \mathbf{y} - \sum_{l=1}^{k} \xi_l \mathbf{f}(\boldsymbol{\theta}_l)$;
end
Normalize $\boldsymbol{\xi}$ such that $\sum_{l=1}^{k} \xi_l = 1$;

leave the implementation of a stopping condition for CL-OMP(R) and comparison with existing methods for model selection for future work.

Guarantees for CL-OMP(R). It is not surprising that theoretical guarantees for the CL-OMP algorithm seem complicated to obtain: it involves a number of random steps and non-convex gradient descent steps for which no guarantees can be provided in general.

At the time of our first publication of the CL-OMPR algorithm, we were made aware of the parallel work of Boyd et al. [BSR15], where the authors develop an algorithm coined *Alternating Descent Conditional Gradient Method*, for generalized sparse inverse problems in continuous domain. Surprisingly enough, although the interpretations of the two approaches are fairly different this algorithm is extremely similar to CL-OMP, in that it progressively extends a sparse support and alternates with non-convex updates[3]. In fact a similar algorithm also appeared in [BP12]. The two main differences between these algorithms and the proposed CL-OMP(R) are a) the normalization of atoms by their Euclidean norm in the maximization of correlation step of CL-OMP(R) (step 1 of Alg. 3), which is seen to be empirically necessary when dealing *e.g.* with GMMs with unknown covariance, for numerical stability reasons, and b) the Hard Thresholding step (for CL-OMPR), which as we will see can dramatically improves performance.

In these papers, the authors provide theoretical guarantees that hold even without the non-convex update step. They are of the form:

> "when k grows to infinity, the cost function (different from ours) can be reduced with a provable rate".

This is unfortunately fundamentally different from the type of guarantees that would be of interest for our problem, which is that we aim at identifying a *fixed* number of *exactly* k sparse components. In particular, for the sketching method of Chapter 3, this number k is related to the complexity of the problem and our analysis fails if k is allowed to grow indefinitely.

When the number of components is fixed, provable guarantees for CL-OMP(R) are still an open question. In this thesis we have focused in Chapter 3 on providing *feasibility* and *information-preservation* guarantees, proving that approximately minimizing the cost function indeed solves the problem. In the next chapter we demonstrate empirically that CL-OMPR performs excellently on a number of problems. A paramount question for the future is to prove that there is an algorithm which, possibly under some additional hypotheses, approximately solves (4.3).

[3] Performed by the BCD algorithm!

4.3 Implementation and complexity

In this section, we briefly study the complexity of the algorithms and give a few implementation details on the CL-OMP(R) algorithm. We released a Matlab implementation of CL-OMP(R) at [Ker16]. The algorithm can be implemented for every differentiable function $\mathbf{f}$, and the code is written so that users can easily implement their own models.

4.3.1 Complexity of the algorithms

Assuming that the complexity of the optimization scheme $\texttt{minimize}_{\boldsymbol{\theta}_l}$ with respect to $\boldsymbol{\theta}_l \in \mathbb{R}^q$ scales linearly in $\mathcal{O}(q)$, the BCD algorithm simply scales in $\mathcal{O}(Tmqk)$, where T is the (fixed) number of iterations.

Just as OMP, whose complexity scales quadratically with the sparsity parameter k, the proposed greedy approaches CL-OMP or CL-OMPR have a computational cost of the order of $\mathcal{O}(mqk^2)$. This is potentially a limiting factor for the estimation of mixtures with many basic components (large k). In classic sparse approximation, approximate least squares approaches such as Gradient Pursuit [BD08a] or LocOMP [MG09] have been developed to overcome this computational bottleneck. One could probably get inspiration from these approaches to further scale up compressive mixture estimation, which is left for future work. For some particular models in Chapter 5 however, we will leverage some ideas based on hierarchical constructions to reduced the quadratic cost to $k \log k$.

4.3.2 Implementing the optimization procedures of CL-OMP(R)

The CL-OMP(R) algorithm requires the implementation of optimization schemes for two functions:

$$v(\boldsymbol{\theta}) = \operatorname{Re}\left(\left\langle \frac{\mathbf{f}(\boldsymbol{\theta})}{\|\mathbf{f}(\boldsymbol{\theta})\|_2}, \mathbf{r} \right\rangle\right), \quad G(\Theta, \boldsymbol{\xi}) = \left\| \mathbf{y} - \sum_{l=1}^{k} \xi_l \mathbf{f}(\boldsymbol{\theta}_l) \right\|_2^2.$$

The first function is the correlation between a normalized atom and the residual that must be approximately maximized in Step 1, and the second is the global cost function that must be further reduced in Step 4. We therefore have to compute the gradients of these functions to implement numerical optimization schemes. Let us show that this can be done if one is able to numerically compute two functions: the feature function $\mathbf{f}(\boldsymbol{\theta}) \in \mathbb{C}^m$ and the function $\mathbf{g} : \mathbb{R}^q \times \mathbb{C}^m \mapsto \mathbb{R}^q$ defined by

$$\mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{J}_{\operatorname{Re}(\mathbf{f})}(\boldsymbol{\theta})^\top \operatorname{Re}(\mathbf{x}) + \mathbf{J}_{\operatorname{Im}(\mathbf{f})}(\boldsymbol{\theta})^\top \operatorname{Im}(\mathbf{x}). \tag{4.12}$$

In the implementation of CL-OMP(R) available at [Ker16], the user can thus instantiate the method on a new model by simply providing expressions for $\mathbf{f}$ and $\mathbf{g}$.

Global cost function. We have already computed the gradient of the cost function G in Section 4.1. It takes the form:

$$\nabla_{\boldsymbol{\xi}} G = -2\left[\operatorname{Re}\left(\langle \mathbf{f}(\boldsymbol{\theta}_l), \mathbf{r} \rangle\right)\right]_{l=1}^{k} \in \mathbb{R}^k, \qquad \nabla_{\boldsymbol{\theta}_l} G = -2\xi_l \mathbf{g}(\boldsymbol{\theta}_l, \mathbf{r}) \in \mathbb{R}^q \tag{4.13}$$

where $\mathbf{r} = \mathbf{y} - \sum_{l=1}^{k} \xi_l \mathbf{f}(\boldsymbol{\theta}_l)$ is the residual. These expressions can indeed be derived using only the functions $\mathbf{f}$ and $\mathbf{g}$.

Normalized correlation. Let us now examine the function maximized in the first step of CL-OMP(R), the correlation between a normalized atom and the residual signal. For simplicity of computation, denote by $\widetilde{\mathbf{f}}(\boldsymbol{\theta}) \in \mathbb{R}^{2q}$ (resp. $\widetilde{\mathbf{r}} \in \mathbb{R}^{2q}$) the function (resp. vector) obtained by stacking the real and imaginary parts of $\mathbf{f}(\boldsymbol{\theta}) \in \mathbb{C}^m$ (resp. $\mathbf{r} \in \mathbb{C}^m$). The function v then takes

the form $v(\boldsymbol{\theta}) = \frac{\widetilde{\mathbf{f}}(\boldsymbol{\theta})^\top \widetilde{\mathbf{r}}}{\|\widetilde{\mathbf{f}}(\boldsymbol{\theta})\|_2}$. Its gradient is:

$$\nabla_{\boldsymbol{\theta}} v(\boldsymbol{\theta}) = \frac{\mathbf{J}_{\widetilde{\mathbf{f}}}(\boldsymbol{\theta})^\top \widetilde{\mathbf{r}}}{\left\|\widetilde{\mathbf{f}}(\boldsymbol{\theta})\right\|_2} - \frac{\widetilde{\mathbf{f}}(\boldsymbol{\theta})^\top \widetilde{\mathbf{r}}}{\left\|\widetilde{\mathbf{f}}(\boldsymbol{\theta})\right\|_2^3} \mathbf{J}_{\widetilde{\mathbf{f}}}(\boldsymbol{\theta})^\top \widetilde{\mathbf{f}}(\boldsymbol{\theta}) = \frac{1}{\left\|\widetilde{\mathbf{f}}(\boldsymbol{\theta})\right\|_2} \cdot \mathbf{J}_{\widetilde{\mathbf{f}}}(\boldsymbol{\theta})^\top \left(\widetilde{\mathbf{r}} - \frac{v(\boldsymbol{\theta})}{\left\|\widetilde{\mathbf{f}}(\boldsymbol{\theta})\right\|_2} \widetilde{\mathbf{f}}(\boldsymbol{\theta}) \right)$$

and therefore

$$\nabla_{\boldsymbol{\theta}} v(\boldsymbol{\theta}) = \frac{1}{\|\mathbf{f}(\boldsymbol{\theta})\|_2} \cdot \mathbf{g}\left(\boldsymbol{\theta}, \mathbf{r} - \frac{v(\boldsymbol{\theta})}{\|\mathbf{f}(\boldsymbol{\theta})\|_2} \mathbf{f}(\boldsymbol{\theta})\right) \tag{4.14}$$

which can indeed be expressed only in terms of $\mathbf{f}$ and $\mathbf{g}$.

4.3.3 Possibility for fast transforms

We thus showed the following:

> Since all optimization schemes in CL-OMP(R) can be performed only by computing $\mathbf{f}$ and $\mathbf{g}$, **there is in general no need to compute the whole matrix $\mathbf{J}_{\mathbf{f}}(\boldsymbol{\theta})^\top$, but only its multiplication by a vector.**

A very interesting possibility is then to replace the matrix-vector multiplication by a faster operation, sometimes with sub-linear cost (a so-called *fast transform*).

Let us illustrate this by going back to example 4.1.1. Denoting $\mathbf{W} = [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m] \in \mathbb{R}^{q \times m}$ the matrix containing the frequency vectors, the computation of the function $\mathbf{f}$ can be done as

$$\mathbf{f}(\boldsymbol{\theta}) = \rho_{\text{im.}}\left(\mathbf{W}^\top \boldsymbol{\theta}\right) \tag{4.15}$$

where $\rho_{\text{im.}}(\cdot)$ is the pointwise application of $x \mapsto e^{\mathrm{i}x}$. Once $\mathbf{f}(\boldsymbol{\theta})$ has been computed, one can apply the function $\mathbf{g}$ by:

$$\begin{aligned} \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) &= \sum_{j=1}^{m} \left(-\sin(\boldsymbol{\omega}_j^\top \boldsymbol{\theta}) \mathrm{Re}\,(x_j) \boldsymbol{\omega}_j + \cos(\boldsymbol{\omega}_j^\top \boldsymbol{\theta}) \mathrm{Im}\,(x_j) \boldsymbol{\omega}_j \right) \\ &= \mathbf{W}\Big(-\mathrm{Im}\,(\mathbf{f}(\boldsymbol{\theta})) \odot \mathrm{Re}\,(\mathbf{x}) + \mathrm{Re}\,(\mathbf{f}(\boldsymbol{\theta})) \odot \mathrm{Im}\,(\mathbf{x}) \Big) \end{aligned} \tag{4.16}$$

where $\odot$ is the Hadamard product between matrices or vectors, *i.e.* element-by-element multiplication.

In these expression, the matrix-vector multiplications $\boldsymbol{\theta} \mapsto \mathbf{W}^\top \boldsymbol{\theta}$ and $\mathbf{x} \mapsto \mathbf{W}\mathbf{x}$ have, in theory, a computational cost of $\mathcal{O}(mq)$. However, it has been shown that the matrix $\mathbf{W}$ can be defined with an underlying *structure* such that these operations are made much faster, and such that there is no need to store the entire matrix $\mathbf{W}$ but only the few parameters that define it, which makes the method not only faster but also more memory efficient.

A paramount example of such fast transforms is the well-known Fast Fourier Transform (FFT) algorithm: when the matrix $\mathbf{W}$ is the squared Fourier matrix of size d, the FFT algorithm performs its multiplication by a vector in time $\mathcal{O}(d \log d)$ instead of $\mathcal{O}(d^2)$. Generalizing this principle, some approaches define efficient matrices while preserving the important properties of usual random matrices [LSS13; Yan+15] or deep networks [SS16; CS16], while other methods attempt to *learn* an efficient factorization of any matrices [LeM16; Cha+15].

In this manuscript we do not experiment with fast transforms, and leave this interesting idea that would allow for an even more efficient sketching method for future investigations. Some works [Cha17] have already begun to explore this direction.

4.4 Experimental illustration

In this last section, we briefly compare the proposed algorithms by implementing the framework of Example 4.1.1, and show empirically that the CL-OMPR algorithm performs incredibly better than the BCD and CL-OMP algorithms (also recall that the algorithm in [BSR15] is

roughly similar to CL-OMP without Replacement).

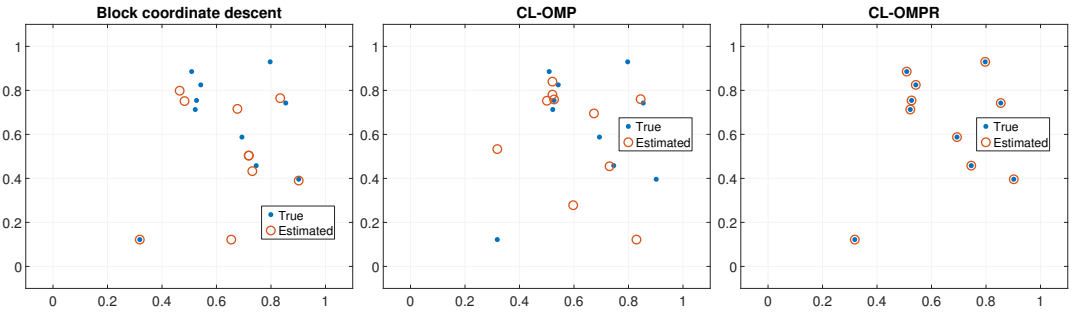

FIGURE 4.4: Example of true parameters $\Theta^\star$ and recovered parameters $\widetilde{\Theta}$ along the first two dimensions, for the three algorithms BCD (left), CL-OMP (center) and CL-OMPR (right).

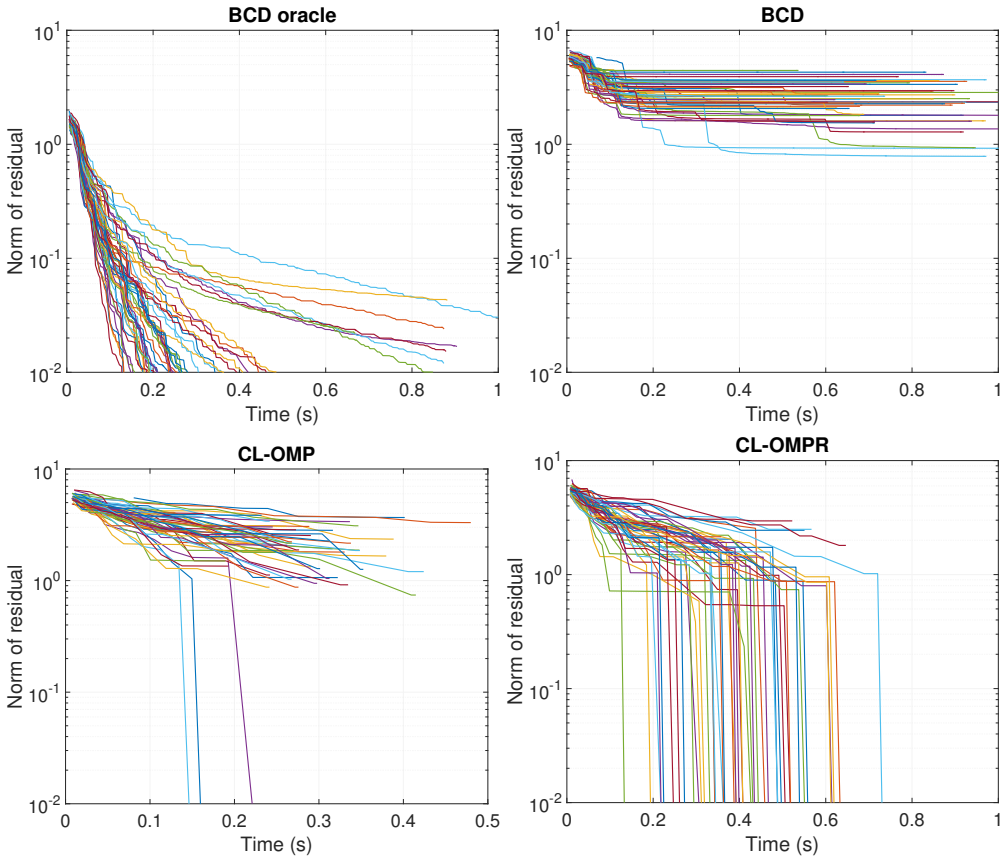

FIGURE 4.5: Evolution of the cost function G with time, for 30 experiments, for the four algorithms BCD oracle (top left), BCD (top right), CL-OMP (bottom left) and CL-OMPR (bottom right).

Setup. In this experiment, we consider a mixture of $k = 10$ components, with parameters $\boldsymbol{\theta}_l$ of size $q = 5$ (recall that the total parameter size of the mixture, counting the weights, is $k(q+1)$), for a measurement vector of size $m = 300$. First we generate the true parameter $\Theta^\star$ by drawing k points $\boldsymbol{\theta}_l^\star$ uniformly from $\mathcal{T} = [0,1]^q$. The weights $\boldsymbol{\xi}^\star$ are chosen by first drawing a vector of k numbers between 0.5 and 1.5 uniformly, then normalizing such that $\sum_{l=1}^k \xi_l^\star = 1$. We draw the frequency vectors $\boldsymbol{\omega}_j$ randomly from a Gaussian distribution $\mathcal{N}(0, \sigma^2 \mathbf{I})$ with $\sigma^2 = 25$, which defines the function $\mathbf{f}(\boldsymbol{\theta}) = \left[e^{\mathrm{i}\boldsymbol{\omega}_j^\top \boldsymbol{\theta}}\right]_{j=1}^m$. The measurement vector is then computed without noise as $\mathbf{y} = \sum_{l=1}^k \xi_l^\star \mathbf{f}(\boldsymbol{\theta}_l^\star)$.

Parameters $(\widetilde{\Theta}, \widetilde{\boldsymbol{\xi}})$ are then recovered using four different algorithms. First, to test the local convexity when close to the optimum, the BCD algorithm is performed with initial parameters Θ^0 that are close to the true parameters, defined as $\Theta^0 = \Theta^\star + 0.05 * \bar{\Theta}$ where $\bar{\Theta}$ are k points drawn *i.i.d.* from $[0,1]^q$. This is referred to as BCD *oracle*. Second, the BCD algorithm is performed with no *a priori* knowledge, with initial parameters $\boldsymbol{\theta}_l^0$ directly drawn *i.i.d.* from $[0,1]^q$. In both cases the weights are initialized uniformly as $\xi_l^0 = 1/k$. Finally, we perform CL-OMP (Alg. 3) and CL-OMPR (Alg. 4), where the gradient descent of Step 1 is initialized with a random point in $[0,1]^q$. An example of true and recovered parameters is given in Figure 4.4.

Minimization of the cost function. We repeat this experiment 30 times. We measure the evolution of the cost function $G(\Theta, \boldsymbol{\xi})$ with respect to the time of execution in Figure 4.5. In this figure we consider that the minimization is "successful" when the cost function falls below an arbitrary threshold of 10^{-2} on the y axis. It is seen that BCD *with oracle initialization* manage to attain this goal almost at each run, while the BCD algorithm is never successful. This confirms that NLLS problems can be dealt with using a simple (naive) approach with a good oracle initialization, and are difficult with no *a priori* knowledge.

Even without prior knowledge, CL-OMP is successful in a few runs of the experiment, and *CL-OMPR is successful in a great majority of the cases*.

	BCD	CL-OMP	CL-OMPR	BCD oracle
Mean SSE	1.06	0.794	0.0825	$3.80e^{-5}$
Var. SSE	0.453	0.259	0.0845	$1.99e^{-8}$
Median SSE	1.01	0.742	$5.65e^{-13}$	$7.40e^{-12}$

TABLE 4.1: Mean, variance and median of the SSE (lower is better) over 30 experiments for the BCD, BCD oracle, CL-OMP and CL-OMPR algorithm.

Recovery of the parameters: a first step toward Compressive k-means. To confirm the recovery of the parameters $\Theta^\star$ by minimization of the cost function, we examine the sum of squared distances of each parameter $\boldsymbol{\theta}_l^\star$ to its closest recovered point, referred to as the Sum of Squared Error (SSE):

$$SSE(\widetilde{\Theta}) = \sum_{l=1}^{k} \min_{1 \leq p \leq k} \left\| \boldsymbol{\theta}_l^\star - \widetilde{\boldsymbol{\theta}}_p \right\|_2^2 .$$

We show the mean, the variance and the median of the SSE over 30 experiments in Table 4.1 . All three quantities indicate that the BCD algorithm *with oracle initialization* indeed almost always yields an SSE very close to 0, which indicates that in nearly all cases the true parameters are *exactly* recovered. The CL-OMP performs slightly better than BCD without oracle, however, it yields a rather elevated median SSE, which indicates that in the majority of the cases the parameters $\Theta^\star$ are not recovered. The CL-OMPR algorithm has a mean SSE that is slightly degraded by the few failure cases, but exhibits a median SSE that is close to 0, similar to BCD with oracle initialization. Among the three algorithms with no oracle knowledge, CL-OMPR indeed seems to be the only one capable of recovering $\Theta^\star$ in the vast majority of the cases.

We observe that minimizing the cost function G indeed leads to recovering the true parameters $\Theta^\star$. The problem of exactly recovering points (centroids) in $\mathbb{R}^q$ with a Random Fourier Sampling is akin to the Compressive k-means problem that will be examined in the next chapter, and for which theoretical guarantees of success are given in Chapter 6.

4.5 Conclusion

In this chapter, we developed a greedy approach for handling a particular Non-Linear Least Squares problem, where a measurement vector is approached by a linear combination of atoms chosen in an infinite, continuously indexed dictionary. This flexible algorithm can be applied as soon as the mapping which associates a parameter with an atom is differentiable.

We showed that the considered optimization problem is sometimes locally block convex when close to the optimum. This was illustrated on an example where a simple block coordinate descent algorithm was shown to perform well when initialized close to the true solution. In the absence of prior knowledge, however, this naive approach was observed to fail. On the contrary, the proposed CL-OMPR algorithm successfully minimized the cost function in the vast majority of the experiments, *without requiring a good initialization*.

Future work will aim at experimenting with fast transforms to speed-up the greedy method, and most of all provide theoretical guarantees for CL-OMP(R).

In the next chapter, we apply the CL-OMPR algorithm to the sketching problem defined in Chapter 3, for the estimation of several mixture models, and examine in more details the role of each parameters in the algorithm.

Chapter 5

Application: Sketched estimation of three mixture models

Context. In the previous chapter, we defined a flexible heuristic algorithm to deal with the minimization of any function of the form

$$G(\Theta, \xi) = \left\| \sum_{l=1}^{k} \xi_l \mathbf{f}(\boldsymbol{\theta}_l) - \mathbf{y} \right\|_2^2 \tag{5.1}$$

where $\mathbf{y} \in \mathbb{C}^m$ is a measurement vector, $\boldsymbol{\theta}_l \in \mathcal{T} \subset \mathbb{R}^q$ are parameters, $\boldsymbol{\xi} \in \mathbb{S}^{k-1}$ are normalized weights and $\mathbf{f} : \mathbb{R}^q \mapsto \mathbb{C}^m$ is a differentiable feature function. We briefly applied it on a simple example.

This chapter. In this chapter we go back to the sketching problem defined in Chapter 3 in the case of mixture model estimation (recalled below), and apply the CL-OMP(R) algorithm to this difficult non-convex optimization problem, by defining $\mathbf{f}(\boldsymbol{\theta}) := \mathcal{A}\pi_{\boldsymbol{\theta}}$, where $\pi_{\boldsymbol{\theta}}$ is a basic component in a mixture and $\mathcal{A}$ is a linear operator randomly designed.

The outline of the chapter is the following.

- In Section 5.1, we develop a simple unsupervised method to learn the parameters necessary for the design of the sketching operator $\mathcal{A}$ before the actual sketching takes place, using a very small sample of training data. The proposed approach is relatively simple compared to existing methods to design kernels, but it is nevertheless shown to yield good results in practice.
- In Section 5.2 we summarize the methodology and discuss some implementation details.
- In Section 5.3 we describe a first implementation of CL-OMPR, where $\pi_{\boldsymbol{\theta}} = \delta_{\boldsymbol{\theta}}$ is a Dirac distribution. This implementation is compared to the k-means algorithm, and in the next chapter we will give related theoretical results for this method.
- In Section 5.4 we implement CL-OMP(R) for Gaussian Mixture Modeling with diagonal covariance, *i.e.* $\pi_{\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}))$. We introduce an algorithm, alternative to CL-OMP(R) but specific to GMM, that scales in $k \log k$ instead of k^2. The sketching approach is compared to the traditional Expectation-Maximization (EM) algorithm.
- In Section 5.5 we briefly compare the proposed sketching method with the coreset method described in [Luc+17], for both the k-means and GMM estimation problems.
- In Section 5.6 we instantiate CL-OMP(R) for the problem of estimating mixtures of multivariate elliptic stable distributions with diagonal precision matrix $\pi_{\boldsymbol{\theta}} = \mathcal{S}_\alpha(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}))$. To our knowledge, our approach is the first algorithm capable of performing this estimation.

Sketching method with RFFs. Let us briefly recall the sketching method, as it will be instantiated in this chapter. Let $\mathfrak{T} = \{\pi_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathcal{T}\} \subset \mathfrak{P}(\mathbb{R}^d)$ be a set of probability distributions over

$\mathbb{R}^d$. Define a sketching function $\Phi : \mathbb{R}^d \mapsto \mathbb{C}^m$, and the sketching operator as:

$$\mathcal{A}\mu := \langle \mu, \Phi \rangle \ . \tag{5.2}$$

Given a database of samples $\mathbf{z}_1, \dots, \mathbf{z}_n \in \mathbb{R}^d$, the sketch of the database is computed as

$$\hat{\mathbf{y}} := \mathcal{A}\hat{\pi}_n = \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{z}_i) \tag{5.3}$$

Fitting a k-mixture on the database can then be done (Chapter 3) as:

$$\min_{\Theta \in \mathcal{T}^k,\ \boldsymbol{\xi} \in \mathbb{S}^{k-1}} \|\mathcal{A}\pi_{\Theta,\boldsymbol{\xi}} - \hat{\mathbf{y}}\|_2^2 \tag{5.4}$$

where $\pi_{\Theta,\boldsymbol{\xi}} = \sum_{l=1}^{k} \xi_l \pi_{\boldsymbol{\theta}_l}$. This problem is indeed of the form (5.1) with $\mathbf{f}(\boldsymbol{\theta}) := \mathcal{A}\pi_{\boldsymbol{\theta}}$, and we can apply the CL-OMP(R) algorithm as long as $\boldsymbol{\theta} \mapsto \mathcal{A}\pi_{\boldsymbol{\theta}}$ and its gradient have a closed-form expression.

In the applications presented in this chapter we will implement the method with the function Φ defined as a collection of complex exponentials at randomly drawn frequencies, which is an instantiation of the sketching method of Chapter 3 with the kernel RF expansion taken as Random *Fourier* Features [RR07], and similar to Bourrier's original framework [BGP13] (Chapter 1 Section 1.2). Indeed, since we aim at spatially localizing several components, a Fourier sampling seems natural[1]. Given frequency vectors $\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_m \in \mathbb{R}^d$ (in practice drawn *i.i.d.* from a distribution Λ) the sketching function is defined as

$$\Phi(\mathbf{z}) = \left[e^{\mathrm{i}\boldsymbol{\omega}_j^\top \mathbf{z}}\right]_{j=1}^{m} \ . \tag{5.5}$$

where, compared to the sketch (3.7) studied in Chapter 3, we omit the normalization $\frac{1}{\sqrt{m}}$ since it has no influence on the recovery procedure. The sketching operator $\mathcal{A}$ is then a sampling of the *characteristic function* of the distribution π, defined as

$$\psi_\pi(\boldsymbol{\omega}) = \mathbb{E}_{\mathbf{z}\sim\pi} e^{\mathrm{i}\boldsymbol{\omega}_j^\top \mathbf{z}} \ . \tag{5.6}$$

For many important parametric sets of distributions $\mathfrak{T} = \{\pi_{\boldsymbol{\theta}}\}$ the characteristic function has a differentiable, closed-form expression with respect to $\boldsymbol{\theta}$, such that the CL-OMP(R) algorithm can be applied.

5.1 Kernel choice

The selection of an appropriate Random Feature expansion $(\mathcal{F}_R, \Lambda)$ amounts to choosing a kernel κ. Given some learning task, the selection of an appropriate kernel, or *kernel design*, is known to be a difficult problem. The theoretical results of Chapter 6 will be derived for a Gaussian kernel for the simplicity of its expression, however in practice one must adjust the kernel to the given data less the method may perform very poorly.

With the Fourier features that we use in this chapter, it is seen that the frequencies $\boldsymbol{\omega}_j$ used to build the sketching operator have to be (randomly) selected at the right *scale* to obtain good results, *i.e.* with the right balance between "low" and " high" frequencies.

In this section we describe a simple unsupervised procedure to learn this scale parameter using a fraction of training data, prior to the actual sketching. We also derive an innovative radial kernel whose expression is based on a heuristic and which seems to perform well in practice.

Note that, by Theorem 1.4.4, it is easy to check that all RF expansions presented here yield characteristic kernels, *i.e.* the induced MMD is a proper norm. Let us first give some references on kernel design.

[1] In the next Chapter, we will provide guarantees for *reweighted* versions of the Fourier sampling (Example 3.1.3). In practice we find that adding these weights has very little effect and increases computation time.

5.1.1 A few general principles for kernel design

Kernel design/learning is often done in a supervised manner with respect to a particular learning task. A natural idea is to define a parametric family of kernels (*e.g.* Gaussian kernels parametrized by a bandwidth) and apply a supervised parameter learning scheme. Modern architectures often learn the kernel as a convex combination of many kernels [BLJ04; Son+06].

With Random Features. Since supervised kernel design can be computationally intensive, one can use acceleration techniques like RF expansions to leverage the computational gain of explicit embeddings. Often the considered frequency distribution Λ is derived *from* a kernel κ chosen in a parametric family endowed with closed-form expressions for both κ and Λ [RR07]. However an increasingly popular idea is to directly learn the RF expansion of the kernel, without resorting to a closed-form expression for the kernel κ. Researchers have explored the possibility of modifying the matrix of frequencies to obtain a better approximation quality [Xin+16] or to accelerate the computation of the kernel [LSS13; Yan+15]. Both ideas have been exploited for learning an appropriate frequency distribution Λ, for instance modeled as a mixture of Gaussians [WA13; Yan+15; Oli+15], or by optimizing weights over a finite dictionary of many distributions [SD16].

For kernel mean embedding. In the context of kernel mean embedding, learning a kernel has mainly been explored for the two-sample test problem [Sri+09; Gre+12], *i.e.* the problem of deciding if two sets of samples are drawn from the same distribution. The main idea of these methods is to maximize the "discriminative" power of the MMD, *i.e.* its ability to distinguish distributions of interest that are different. In [Sri+09], the authors introduce a new metric defined as the maximal MMD over a whole family of kernels. In [Gre+12] a learning procedure is described to derive a kernel with maximal testing power, as a convex combination over a finite family of kernels. In [Chw+15; Jit+16] the acceleration provided by RF expansions is exploited to learn kernels, also for two-sample test, while in [PSW16] the learning is done in a streaming context.

In this chapter. In the next section, we describe a simple method to learn a frequency distribution Λ for RFFs, adapted to compute the sketch of some data in practice. Most approaches described in the literature cannot be readily applied in this context, since we sketch only a *single* distribution that we aim at recovering, while the two-sample test aims at distinguishing the distributions of several sets of samples. However, the general idea of maximizing the capacity of the MMD to discriminate distributions that are close to the true distribution of the data will still be our main inspiration. The biggest difference is that the proposed method is done in a *completely unsupervised* manner, based on an approximate theoretical expression of the MMD for clustered data, instead of a thorough statistical analysis of the kernel learning problem. We leave experimentation with more exotic kernels and more advanced learning methods for future work.

5.1.2 Oracle frequency sampling pattern for clustered data

In this thesis, the sketching method is applied to mixture model learning. The data are therefore supposed to be approximately *clustered*, *i.e.* shaped as several localized groups of data. Hence we develop our heuristic for kernel learning based on the supposition that the probability distribution of the data $\pi^\star$ is a mixture of Gaussians. In practice, the learning method derived in the next section works well even when $\pi^\star$ is not a GMM (*e.g.* on speech data in Section 5.4.6): it learns a "scale" at which patterns are potentially interesting in the data.

We denote a Gaussian distribution in $\mathbb{R}^d$ by $\pi_{\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean of the Gaussian, the positive definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is its covariance, and $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the set of both parameters. A k-GMM is denoted $\pi_{\Theta,\boldsymbol{\xi}} = \sum_{\ell=1}^{k} \xi_\ell \pi_{\boldsymbol{\theta}_\ell}$ where $\boldsymbol{\xi} \in \mathbb{S}^{k-1}$ and $\Theta = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k)$.

Recall that in this chapter we consider usual Random Fourier Features $\mathcal{F}_R = \left\{ e^{\mathrm{i}\boldsymbol{\omega}^\top \mathbf{z}} \mid \boldsymbol{\omega} \in \mathbb{R}^d \right\}$, such that the sketch $\mathcal{A}\pi$ is a random sampling of the characteristic function of the probability

distribution π. For a Gaussian $\pi_{\boldsymbol{\theta}}$, the characteristic function is

$$\psi_{\pi_{\boldsymbol{\theta}}}(\boldsymbol{\omega}) = e^{\mathrm{i}\boldsymbol{\omega}^\top \boldsymbol{\mu}} e^{-\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma} \boldsymbol{\omega}} . \tag{5.7}$$

which we also denote $\psi_{\boldsymbol{\theta}}$. The characteristic function of a GMM is denoted by $\psi_{\Theta,\boldsymbol{\xi}}$.

To derive a method to automatically learn a kernel, we proceed step by step. We first describe a heuristic to choose an appropriate frequency distribution Λ when the distribution $\pi^\star$ is a single Gaussian with *known* parameter $\boldsymbol{\theta}^\star$ (which is of course not the case in practice). We refer to this "ideal" distribution as oracle distribution. We then extend to mixtures of Gaussians with known parameters. Finally, inspired by this study we derive our learning method.

Oracle frequency sampling pattern for a single known Gaussian

We start by designing a heuristic for choosing frequencies adapted to the estimation of a single Gaussian $\pi_{\boldsymbol{\theta}^\star}$, assuming the parameter $\boldsymbol{\theta}^\star = (\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star)$ is available.

Gaussian frequency distribution. Consider the expression (5.7) of the characteristic function of the Gaussian $\pi_{\boldsymbol{\theta}^\star}$. It is an oscillating function with Gaussian amplitude of inverted variance with respect to the original Gaussian. Given that $|\psi_{\boldsymbol{\theta}^\star}| \propto \mathcal{N}\left(0, (\boldsymbol{\Sigma}^\star)^{-1}\right)$, choosing a Gaussian frequency distribution denoted by $\Lambda^{(G)}_{\boldsymbol{\Sigma}^\star} = \mathcal{N}\left(0, (\boldsymbol{\Sigma}^\star)^{-1}\right)$ is a possible, intuitive choice [BGP13; BGP15] to sample the characteristic function. It concentrates frequencies in the regions where the sampled characteristic function has high amplitude.

However, points drawn from a high-dimensional Gaussian concentrate on an ellipsoid which moves away from the origin as the dimension d increases. Such a Gaussian sampling therefore "undersamples" low or even middle frequencies (Fig. 5.1). This phenomenon has long been one of the reasons for using dimensionality reduction for GMM estimation [Das99]. Hence, in high dimension the amplitude of the characteristic function becomes negligible (with high probability) at all selected frequencies. We will see in Chapter 6 that reconstruction is still possible in theory, however in practice the CL-OMP(R) algorithm of Chapter 4 is very unstable in that case.

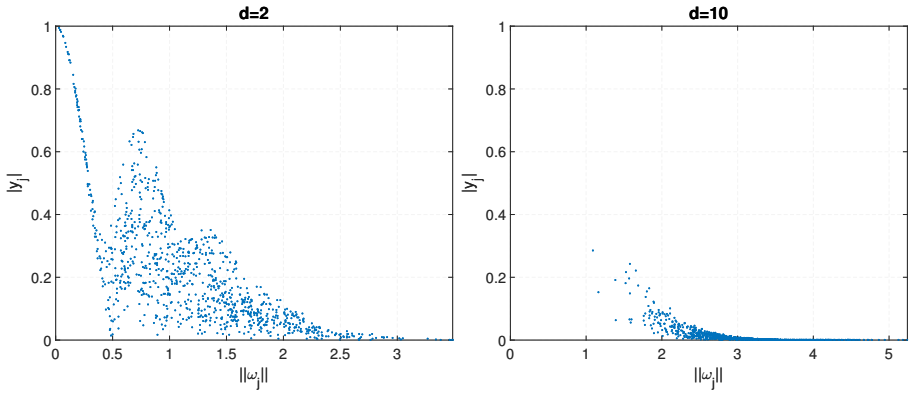

FIGURE 5.1: Modulus of the entries of a sketch $\mathbf{y} = \mathcal{A}\pi_{\Theta,\boldsymbol{\xi}}$ of a mixture of $k = 5$ isotropic Gaussians computed with sampling frequencies $\{\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m\}$ drawn from a normal distribution $\Lambda^{(G)}_{\mathbf{I}}$, for dimensions $d = 2$ (left) and $d = 10$ (right). Each point $|\mathbf{y}_j|$ is placed with respect to the norm of the corresponding frequency $\|\boldsymbol{\omega}_j\|_2$. As expected, in high dimension only the "tail" of the characteristic function $\psi_{\Theta,\boldsymbol{\xi}}$ is captured.

Folded Gaussian radial frequency distribution. In light of the problem observed with the Gaussian frequency distribution, we propose to draw frequencies from a distribution that allows for an accurate control of the quantity $\boldsymbol{\omega}^\top \boldsymbol{\Sigma}^\star \boldsymbol{\omega}$, and thus of the amplitude $e^{-\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}^\star \boldsymbol{\omega}}$ of

the characteristic function. This is achieved by drawing

$$\boldsymbol{\omega} = R\left(\boldsymbol{\Sigma}^\star\right)^{-\frac{1}{2}}\boldsymbol{\rho}\,, \tag{5.8}$$

where $\boldsymbol{\rho} \in \mathbb{R}^d$ is uniformly distributed on the ℓ_2 unit sphere $\mathcal{S}_{d-1}$, and $R \in \mathbb{R}_+$ is a radius chosen independently from $\boldsymbol{\rho}$ with a distribution π_R we will now specify.

With the decomposition (5.8), the characteristic function $\psi_{\boldsymbol{\theta}^\star}$ is now expressed as

$$\psi_{\boldsymbol{\theta}^\star}\left(R\left(\boldsymbol{\Sigma}^\star\right)^{-\frac{1}{2}}\boldsymbol{\rho}\right) = \exp\left(\mathrm{i}R\boldsymbol{\rho}^\top\left(\boldsymbol{\Sigma}^\star\right)^{-\frac{1}{2}}\boldsymbol{\mu}^\star\right)\exp\left(-\tfrac{1}{2}R^2\right) = \psi_\theta(R),$$

where ψ_θ is the characteristic function of a *one-dimensional* Gaussian with mean $\mu = \boldsymbol{\rho}^\top\boldsymbol{\Sigma}^{\star-\frac{1}{2}}\boldsymbol{\mu}$ and variance $\sigma^2 = 1$. We thus consider the estimation of a one-dimensional Gaussian $\pi_{\theta^\star} = \mathcal{N}\left(\mu^\star, \left(\sigma^\star\right)^2\right)$, with any mean but variance $\sigma^\star = 1$, as our baseline to design a radius distribution π_R.

In this setting, we no longer suffer from unwanted concentration phenomena and can resort to the intuitive Gaussian radius distribution to sample $\psi_{\theta^\star}$. It corresponds to a radius density function $\pi_R = \mathcal{N}^+(0, \frac{1}{\sigma^{\star 2}}) = \mathcal{N}^+(0,1)$ (*i.e.* Gaussian with absolute value, referred to as *folded* Gaussian). Using this radius distribution with the decomposition (5.8) yields a frequency distribution $\Lambda^{(FGr)}_{\boldsymbol{\Sigma}^\star}$ referred to as *Folded Gaussian radius* frequency distribution. Note that, similar to the Gaussian frequency distribution, the Folded Gaussian radius distribution only depends on the (oracle) covariance $\boldsymbol{\Sigma}^\star$ of the sketched distribution $\pi_{\boldsymbol{\theta}^\star}$.

Adapted radius distribution Though we will see it yields decent results, the Folded Gaussian radius frequency distribution somehow produces too many frequencies with a low radius R. These carry a limited quantity of information about the original distribution, since all characteristic functions equal 1 at the origin[2]. We now present a heuristic that may avoid this "waste" of frequencies.

As described in Section 5.1.1, maximizing the capacity of the MMD to discriminate between different distributions of interest is a classic strategy for two-sample tests using kernel mean embeddings. Although only a single distribution is sketched here, intuitively the MMD should distinguish the true distribution from distributions that are "close" to it, to be able to recover it from the sketch with precision. Thus the chosen frequencies should properly discriminate Gaussians with different parameters in the neighborhood of the true parameter $\theta^\star = (\mu^\star, 1)$. This corresponds to promoting frequencies $\boldsymbol{\omega}$ leading to a large difference $|\psi_\theta(\omega) - \psi_{\theta^\star}(\omega)|$ for parameters θ close to $\theta^\star$. A way to achieve this is to promote frequencies where the norm of the gradient $\|\nabla_\theta\psi_\theta(\omega)\|$ is large.

Recall that for a one-dimensional Gaussian $\psi_\theta(\omega) = e^{\mathrm{i}\mu\omega}e^{-\frac{1}{2}\sigma^2\omega^2}$. The norm of the gradient is expressed as:

$$\begin{aligned}\left\|\nabla_\theta\psi_\theta(\omega)\right\|_2^2 &= \left|\nabla_\mu\psi_\theta(\omega)\right|^2 + \left|\nabla_{\sigma^2}\psi_\theta(\omega)\right|^2 \\ &= \left|\mathrm{i}\omega\psi_\theta(\theta)\right|^2 + \left|-\frac{1}{2}\omega^2\psi_\theta(\omega)\right|^2 = \left(R^2 + \frac{R^4}{4}\right)e^{-\sigma^2R^2}\end{aligned}$$

where $R = |\omega|$, and therefore $\left\|\nabla_\theta\psi_{\theta^\star}(\omega)\right\|_2 = \left(R^2 + \frac{R^4}{4}\right)^{\frac{1}{2}}e^{-\frac{1}{2}R^2}$ since $\left(\sigma^\star\right)^2 = 1$. This expression still has a Gaussian decrease (up to polynomial factors), and indeed avoids very low frequencies (Figure 5.2). It can be normalized to a density function:

$$\pi_R = C\left(R^2 + \tfrac{R^4}{4}\right)^{\frac{1}{2}}e^{-\frac{1}{2}R^2}, \tag{5.9}$$

with C some normalization constant. Using this radius distribution with the decomposition (5.8) yields a distribution $\Lambda^{(Ar)}_{\boldsymbol{\Sigma}}$ referred to as *Adapted radius* frequency distribution. Once again, this distribution only depends on the covariance $\boldsymbol{\Sigma}$.

[2] In a way, numerous measures of the characteristic function near the origin essentially measure its derivatives at various orders, which are associated to classic polynomial moments.

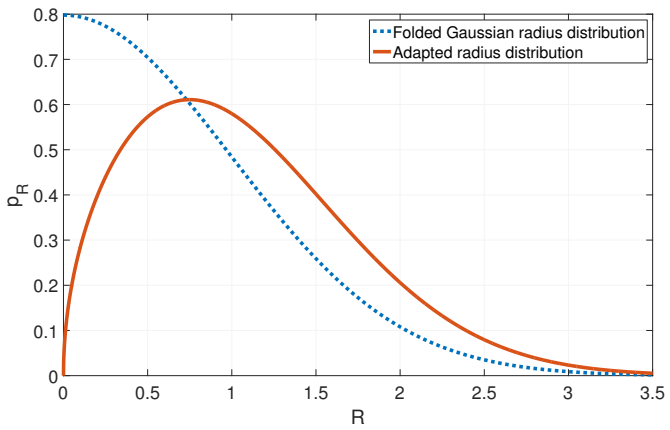

FIGURE 5.2: Folded Gaussian radius (FGr) and Adapted radius (Ar) distributions.

Oracle frequency sampling pattern for a known mixture of Gaussians

Any frequency distribution $\Lambda^{(.)}_{\mathbf{\Sigma}}$ selected for sampling the characteristic function of a single known Gaussian $\pi_{\boldsymbol{\theta}^\star}$ can be immediately and naturally extended to a frequency distribution $\Lambda^{(.)}_{\Theta^\star,\boldsymbol{\xi}^\star}$ to sample the characteristic function of a known GMM $\pi_{\Theta^\star,\boldsymbol{\xi}^\star} = \sum_{l=1}^k \xi_l^\star \pi_{\boldsymbol{\theta}_l^\star}$, by mixing the frequency distributions corresponding to each Gaussian:

$$\Lambda^{(.)}_{\Theta^\star,\boldsymbol{\xi}^\star} = \sum_{l=1}^k \xi_l^\star \Lambda^{(.)}_{\mathbf{\Sigma}_l^\star}. \tag{5.10}$$

Each component $\Lambda^{(.)}_{\mathbf{\Sigma}_l^\star}$ has the same weight as its corresponding Gaussian $\pi_{\boldsymbol{\theta}_l^\star}$. Indeed, a Gaussian with a high weight must be precisely estimated, as its influence on the overall reconstruction error (*e.g.* in terms of Kullback-Leibler divergence) is more important than that of components with low weights. Thus more frequencies adapted to this Gaussian are selected.

The draw of frequencies with an oracle distribution $\Lambda^{(.)}_{\Theta,\boldsymbol{\xi}}$ is summarized in Function `DrawFreq`.

5.1.3 Learning the frequency sampling pattern in practice

Inspired by the study of the previous section for known GMM, we now derive a method to automatically learn a kernel that works well in practice with no a priori knowledge. The reader should also keep in mind that it is still very easy to integrate some prior knowledge in this design, especially since the proposed frequency distributions only take into account the variances of the GMM components, not their means.

Given a database $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{R}^d$ that we want to sketch, this method uses a fraction of the database $\mathbf{z}_1, \ldots, \mathbf{z}_{n_0}$ with $n_0 \ll n$ to learn an appropriate frequency distribution with a light procedure, before the actual sketch is computed on the database.

The idea is to estimate the average variance $\bar{\sigma}^2 = \frac{1}{kd}\sum_{l=1}^k \sum_{i=1}^d \sigma_{l,i}^2$ of the components in the GMM, where $\sigma_{l,1}^2, ..., \sigma_{l,d}^2$ are the eigenvalues of $\mathbf{\Sigma}_l$. Note that this parameter may be significantly different from the global variance of the data, for instance in the case of well-separated components with small variances. This estimation is performed using a light sketch $\mathbf{y}_0$ with m_0 frequencies, computed on a small subset of n_0 items from the database, then a frequency distribution corresponding to a single isotropic Gaussian $\Lambda^{(.)}_{\bar{\sigma}^2\mathbf{I}}$ is selected.

Indeed, if the variances $\sigma_{l,i}^2$'s are not too different from each other, the amplitude of the empirical characteristic function $\left|\frac{1}{n_0}\sum_{i=1}^{n_0} e^{\mathrm{i}\boldsymbol{\omega}^\top \mathbf{z}_i}\right|$ *approximately* follows $e^{-\frac{1}{2}\|\boldsymbol{\omega}\|_2^2\bar{\sigma}^2}$ with high oscillations, allowing for a very simple amplitude estimation process: assuming the m_0 frequencies used to compute the sketch $\mathbf{y}_0$ are ordered by increasing norm, the sketch $\mathbf{y}_0$ is divided into consecutive blocks, maximal peaks of its modulus are identified within each block forming a curve that approximately follows $e^{-\frac{1}{2}R^2\bar{\sigma}^2}$, then a simple regression is used to estimate $\bar{\sigma}^2$. This process is illustrated in Figure 5.3.

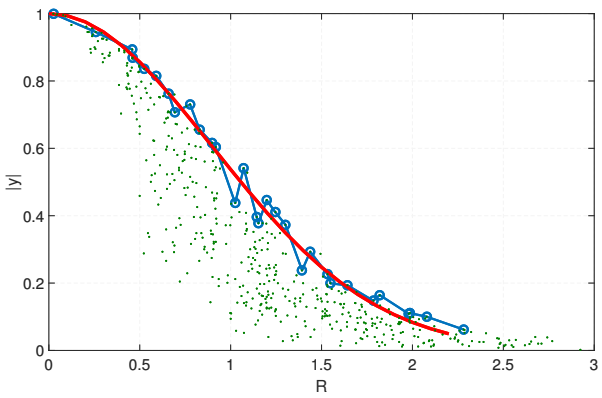

FIGURE 5.3: Illustration of the estimation of $\bar{\sigma}^2$ (Function `EstimMeanSigma`), for $d = 10$, $k = 5$, $m_0 = 500$ and $n_0 = 5000$. Green dots: modulus of the sketch with respect to the norm of frequencies (ordered by increasing radius). Blue lines: visualization of the peaks in each block of 20 consecutive values. Red curve: fitted curve $e^{-\frac{1}{2}R^2\bar{\sigma}^2}$ for the estimated $\bar{\sigma}^2$.

To cope with the fact that the "range" of frequencies that must be considered to compute $\mathbf{y}_0$ is also not known beforehand, we initialize $\bar{\sigma}^2 = 1$ and reiterate this procedure several times, each time drawing m_0 frequencies adapted to the current estimate of $\bar{\sigma}^2$, *i.e.* with some choice of frequency distribution $\Lambda^{(.)}_{\bar{\sigma}^2\mathbf{I}}$, and update $\bar{\sigma}^2$ at each iteration. In practice, the procedure quickly converges in a few iterations. The entire process is summarized in detail in Function `EstimMeanSigma`.

Function DrawFreq($\left(\{\boldsymbol{\Sigma}_l\}_{\ell=1}^k, \boldsymbol{\xi}\right), m,$ `distr`): drawing frequencies for a GMM with known variances and weights, choosing one of the three distributions described in Section 5.1.2

Data: Set of variances and weights of a GMM $\left(\{\boldsymbol{\Sigma}_l\}_{l=1}^k, \boldsymbol{\xi}\right)$, number of frequencies m, type of frequency distribution `distr` $\in \{(G), (FGr), (Ar)\}$
Result: Set of frequencies $\mathcal{W} = \{\boldsymbol{\omega}_1, ldots, \boldsymbol{\omega}_m\}$
for $j \leftarrow 1$ **to** m **do**
  Draw a label according to the weights of the GMM $l_j \sim \sum_{l=1}^k \xi_l \delta_l$;
  if `distr` $= (G)$ **then**
    $\boldsymbol{\omega}_j \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}_{l_j}^{-1}\right)$; // Gaussian
  end
  else
    Draw a direction $\boldsymbol{\rho} \sim \mathcal{U}(\mathcal{S}_{d-1})$;
    if `distr` $= (FGr)$ **then**
      $R \sim \mathcal{N}^+(0, 1)$; // Folded Gaussian radius
    end
    else if `distr` $= (Ar)$ **then**
      $R \sim \pi_R$ with π_R defined by (5.9) ; // Adapted radius
    end
    $\boldsymbol{\omega}_j \leftarrow R\boldsymbol{\Sigma}_{l_j}^{-\frac{1}{2}}\boldsymbol{\rho}$;
  end
end

Connections with Distilled sensing. The reader may note that designing a measurement operator *adapted* to some particular data does not fit the classic paradigm of Compressive Sensing. The two-stage approaches used to choose the frequency distribution presented above can be related to a line of work referred to as adaptive (or *distilled*) sensing [Hau+11], in which

Function EstimMeanSigma($\mathcal{Z}, m_0, c, T$): Estimation of the mean variance $\bar{\sigma}^2$

Data: Small training dataset $\mathcal{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_{n_0}\}$, small number of frequencies m_0, number of blocks $c \in \mathbb{N}_+^*$, number of iterations T
Result: Estimated mean variance $\bar{\sigma}^2$
begin Initialize
  $\bar{\sigma}^2 \leftarrow 1$;
end
for $t \leftarrow 1$ **to** T **do**
  begin Draw some frequencies adapted to the current $\bar{\sigma}^2$
    $\{\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_{m_0}\} \leftarrow \texttt{DrawFreq}(\bar{\sigma}^2\mathbf{I}, 1, m_0, (Ar))$;
    Sort the frequencies $\{\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_{m_0}\}$ by increasing radius $\|\boldsymbol{\omega}_j\|_2$;
  end
  begin Compute small empirical sketch (Figure 5.3, green dots)
    $\hat{\mathbf{y}}_0 \leftarrow \left[\frac{1}{n_0}\sum_{i=1}^{n_0} e^{\mathrm{i}\boldsymbol{\omega}_j^\top \mathbf{z}_i}\right]_{j=1}^{m_0}$;
  end
  begin Divide sketch into blocks, find maximum peak in each block (Figure 5.3, blue line)
    $s \leftarrow \lfloor m_0/c \rfloor$;
    for $q \leftarrow 1$ **to** c **do**
      $j_q = \arg\max_{j \in [(q-1)s+1; qs]} |\hat{y}_{0,j}|$;
    end
  end
  begin Update $\bar{\sigma}^2$ (Figure 5.3, red curve)
    $\hat{\mathbf{e}} \leftarrow \left[\hat{y}_{0,j_q}\right]_{q=1}^{c}$;
    $\bar{\sigma}^2 = \arg\min_{\sigma^2 > 0} \left\| \hat{\mathbf{e}} - \left[e^{-\frac{1}{2}R_{j_q}^2 \sigma^2}\right]_{q=1}^{c} \right\|_2$;
  end
end

		(G)	(FGr)	(Ar)
$d = 2$	$\Lambda^{(\cdot)}_{\Theta^\star,\boldsymbol{\xi}^\star}$	-7.88	-7.90	$\mathbf{-8.24}$
	$\Lambda^{(\cdot)}_{\bar{\sigma}^2\mathbf{I}}$	-7.61	-6.91	$\mathbf{-7.77}$
$d = 20$	$\Lambda^{(\cdot)}_{\Theta^\star,\boldsymbol{\xi}^\star}$	11.96	-5.45	$\mathbf{-5.72}$
	$\Lambda^{(\cdot)}_{\bar{\sigma}^2\mathbf{I}}$	32.44	-5.28	$\mathbf{-5.62}$

TABLE 5.1: Log-KL-divergence on synthetic data using the CL-OMPR algorithm on the GMM estimation problem with synthetic data (see Section 5.4), for $k = 5$ components, $m = 10dk$ frequencies and $n = 200\,000$ items. We compare the three proposed frequency distributions: Gaussian [BGP13] (G), Folded Gaussian radius (FGr) or Adapted radius (Ar), using either the oracle distribution defined in Section 5.1.2 or the approximate distribution used in practice, learned with `EstimMeanSigma`.

a portion of the computational budget is used to crudely design the measurement operator while the rest is used to actually measure the signal. Most often these methods are extended to multi-stage approaches, where the measurement operator is refined at each iteration, and have been used in machine learning [CGJ96] or signal processing [BRH08].

5.1.4 Experiments

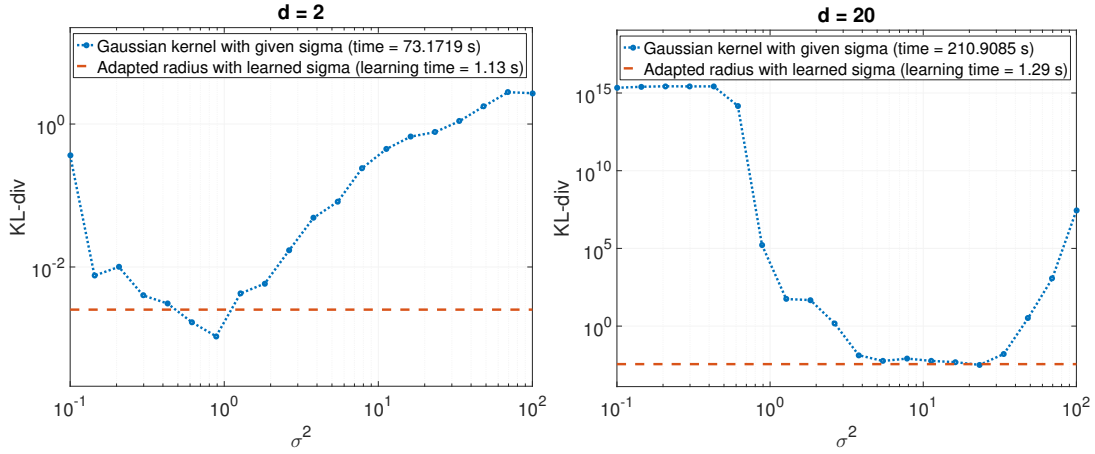

FIGURE 5.4: KL-divergence results (lower is better) of CL-OMPR for the GMM estimation problem of Section 5.4 for $k = 5$ components, $m = 10dk$ frequencies and $n = 2 \cdot 10^5$ items. We compare a Gaussian frequency distribution with a varying bandwidth σ^2 (blue dotted curve), and the proposed learned frequency distribution $\Lambda^{(Ar)}_{\bar{\sigma}^2\mathbf{I}}$ (red dashed line). In each case we outline the time taken to "learn" the frequency distribution: for the Gaussian kernel it corresponds to trying every value of σ^2, for the proposed method it is the execution time of `EstimMeanSigma`.

We conduct a small experiment to validate the proposed design and automatic learning of the distribution of frequencies, on the GMM learning problem that will be studied in Section 5.4 (however there should be no need to read Section 5.4 to understand the few experiments of the current section).

The parameters of a GMM $(\Theta^\star, \boldsymbol{\xi}^\star)$ are generated as described in Section 5.4.3, then n items $\mathbf{z}_i$ are drawn *i.i.d.* from it. The frequency distribution is either chosen as the oracle $\Lambda^{(\cdot)}_{\Theta^\star,\boldsymbol{\xi}^\star}$ to serve as a baseline (*i.e.* using Function `DrawFreq` with the true parameters of the GMM) – we remind the reader that this setting unrealistically assumes that the variances and weights of the GMM are known beforehand –, or taken as $\Lambda^{(\cdot)}_{\bar{\sigma}^2\mathbf{I}}$ where $\bar{\sigma}^2$ is learned on a small portion of the data with `EstimMeanSigma`. Then m frequencies are drawn *i.i.d.* from this distribution, the sketch is computed and a GMM is recovered using the CL-OMPR algorithm. The precision

of the recovery is measured using the KL-divergence, as described in Section 5.4.3: lower is better.

We compare different choices of frequency distributions. We draw $n = 2 \cdot 10^5$ items in dimension $d = 2$ or $d = 20$, with $k = 5$ components in the GMM. In each setting we construct the sketch with $m = 10kd$ frequencies.

In Table 5.1, we compare the three frequency distributions introduced in Section 5.1.2, both with the oracle frequency distribution $\Lambda^{(.)}_{\Theta^\star,\boldsymbol{\xi}^\star}$, and with the approximate one $\Lambda^{(.)}_{\bar{\sigma}^2\mathbf{I}}$. The results show that the Gaussian frequency distribution indeed yields poor reconstruction results in high dimension ($d = 20$), while the Adapted radius frequency distribution outperforms the two others. The use of the approximate $\Lambda^{(.)}_{\bar{\sigma}^2\mathbf{I}}$ instead of the oracle $\Lambda^{(.)}_{\Theta^\star,\boldsymbol{\xi}^\star}$ is shown to have little effect.

In Figure 5.4 we compare the reconstruction results obtained either by learning the scale parameter $\bar{\sigma}^2$ with the proposed unsupervised method and selecting the Adapted radius frequency distribution $\Lambda^{(Ar)}_{\bar{\sigma}^2\mathbf{I}}$, or by performing a supervised learning method where we try many Gaussian frequency distributions $\Lambda = \mathcal{N}(0, \mathbf{I}\sigma^2)$ with different values of σ^2 and select the one that yields the best reconstruction results. This second method roughly corresponds to the supervised learning procedure that is often done in practice [RR07; Sut+15], *i.e.* each evaluation requires the CL-OMPR algorithm to be applied and we assume that *a method to evaluate the quality of the estimated GMM* is available. It is seen that the Adapted radius distribution approaches the result of the best Gaussian distribution in dimension $d = 2$ and reaches it in dimension $d = 20$. It is also seen that the supervised learning of the best Gaussian distribution (*i.e.* trying every value of σ^2) is far more time consuming that the proposed unsupervised learning procedure.

5.2 Summary of the method, computational details

At this point, it seems useful to summarize the sketching methodology before diving into the experiments. This is done in Algorithm 5. The code is available as a Matlab toolbox in [Ker16].

Algorithm 5: Summary of the sketching method.

Data: Dataset $\mathcal{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$, parametric model $\mathfrak{T} = \{\pi_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathcal{T}\}$, sketch size m, number of components k, (Optional) frequency distribution Λ
Result: Mixture parameters $(\Theta, \boldsymbol{\xi})$
if *the distribution* Λ *is not provided* **then**
  Estimate $\bar{\sigma}^2$ with function `EstimMeanSigma`;
  Set $\Lambda = \Lambda^{(Ar)}_{\bar{\sigma}^2\mathbf{I}}$;
end
begin
  Draw frequencies $\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_m \overset{i.i.d.}{\sim} \Lambda$;
end
begin
  Compute the sketch $\hat{\mathbf{y}} = \frac{1}{n}\left[e^{\mathrm{i}\boldsymbol{\omega}_j^\top \mathbf{z}_i}\right]_{j=1}^{m}$; `// Streaming, distributed, etc.`
end
begin
  Recover $(\Theta, \boldsymbol{\xi})$ by approximately minimizing (5.4) with CL-OMP(R);
end

Frequency distribution. The frequency distribution Λ can either be provided by the user, or learned with `EstimMeanSigma` and selected as $\Lambda = \Lambda^{(Ar)}_{\bar{\sigma}^2\mathbf{I}}$. In our experiments the function `EstimMeanSigma` is performed with parameters $n_0 = \min(5000, n)$, $m_0 = 500$, $c = 30$, $T = 3$. Although originally designed for data drawn from a GMM, the proposed unsupervised learning method proves to yield good results for many types of data[3].

[3] In practice we use `EstimMeanSigma` for all experiments except for the MNIST data of Section 5.3.

Computation of the sketch. Computing the sketch has a cost that scales in $\mathcal{O}(mnd)$, which is linear in the size of the database n. However, as mentioned throughout the whole thesis, since $\hat{\mathbf{y}}$ is a linear sketch (with additional normalization $1/n$) it can be computed extremely efficiently in a streaming, distributed and parallel context, in only one pass over the data. Once the sketch is computed, the complexity of the CL-OMPR algorithm is independent of the size of the database.

In the present experiments, we did not implement parallel or distributed computations of the sketch (except for the speaker verification experiment in Section 5.4.6 where the database is distributed), since such implementations are highly dependent on the user's available hardware. For real-life applications of our method, practitioners are however encouraged to do so.

Fast transforms. Observe that computing the sketch can be expressed as:

1. Compute the matrix $\mathbf{U} = \mathbf{W}^\top \mathbf{Z}$, where $\mathbf{W} = [\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_m] \in \mathbb{R}^{d \times m}$ and $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_n] \in \mathbb{R}^{d \times n}$;
2. Apply non-linearity $\mathbf{V} = \rho_{\text{im.}}(\mathbf{U})$ where $\rho_{\text{im.}}(\cdot)$ is the pointwise application of $x \mapsto e^{\mathrm{i}x}$;
3. Average the columns of $\mathbf{V}$.

Hence it would be possible to replace the full matrix-vector multiplication $\mathbf{z} \mapsto \mathbf{W}^\top \mathbf{z}$ by a structured frequency matrix corresponding to a fast transform, to accelerate the computation of the sketch, as is described in Section 4.3.3 for the implementation of CL-OMP(R). The use of fast transforms is left for future work here, a first adaptation of the sketching method with fast transforms is done in [Cha17].

Link with neural networks. Readers familiar with neural networks may notice that it is possible to draw connections between the proposed sketching operation and a simple one-layer network, formed by the successive application of a linear transformation, a pointwise non-linearity $\rho_{\text{im.}}(\cdot)$ and an average-pooling. Neural networks with weights $\mathbf{W}$ chosen at random rather than learned on training data have been studied in the so-called *Random Kitchen Sinks* [RR09], which is a follow-up of RFFs by the same authors, or in the context of Deep Neural Networks (DNN) with Gaussian weights [GSB15]. In the latter, they have been shown to perform a stable embedding of the input $\mathbf{z}$ when it lives on a low-dimensional manifold. In a similar fashion, we show in this thesis that with high probability the sketching procedure is a stable embedding *of the probability distribution of* $\mathbf{z}$ when this distribution belongs to a low-dimensional model.

Execution of CL-OMP(R). All continuous optimization schemes in CL-OMPR are performed with Stephen Becker's adaptation of the L-BFGS-B algorithm [Byr+95] in C, with Matlab wrappers [Bec13]. In some cases (which will be detailed in due time) we enforce "box"-constraints, *i.e.* all parameters $\boldsymbol{\theta}_l \in \mathbb{R}^q$ are constrained to $\boldsymbol{\ell} \leq \boldsymbol{\theta}_l \leq \mathbf{u}$ where $\boldsymbol{\ell}, \mathbf{u} \in \mathbb{R}^q$ are some vectors and $\leq$ is element-by-element comparison. The initialization of Step 1 is also detailed on a case-by-case basis.

5.3 Compressive k-means

Let us turn to the first proposed implementation of the CL-OMPR algorithm where basic distributions $\pi_{\boldsymbol{\theta}}$ are defined as Dirac distributions $\pi_{\boldsymbol{\theta}} = \delta_{\boldsymbol{\theta}}$. The resulting method is somehow not that of the traditional *density fitting* paradigm: the data are obviously not drawn from a mixture of Diracs. However, if intuitively the data are well *clustered*, they are intuitively drawn from a "noisy" mixture of Diracs. Recovering a mixture of Diracs from the sketch is therefore identified with finding **the centers of significant clusters**. Thus we draw a connection between this first application of the sketching framework and the classic *k-means* clustering method, recalled below.

Note that, since the sketching method discards the training data after computing the sketch, in real life applications it would be used for clustering only *test data*, that have a similar behavior to the training data. On the contrary, the k-means algorithm is often used to cluster the

data on which it has been trained. In our experiments, we still evaluate both algorithms on the training data, using the cost function that k-means aims at minimizing (the SSE, see below).

5.3.1 Framework

Consider a database $\mathcal{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\} \subset \mathbb{R}^d$. The k-means problem consists in finding a set $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k)$ of k points in $\mathbb{R}^d$ called *centroids*, that minimizes the *Sum of Squared Errors*

$$\mathrm{SSE}(\Theta) = \sum_{i=1}^{n} \min_{1 \leq \ell \leq k} \|\mathbf{z}_i - \boldsymbol{\theta}_\ell\|_2^2 \tag{5.11}$$

which is the sum of squared distances between each point in the database and its closest centroid. Although simple in its formulation, this problem is known to be NP-hard [Alo+09].

k-means. The most well-known heuristic is known as Lloyd-Max algorithm [Ste56; Llo82], which iterates over two simple steps:

- Assign each point to its closest centroids: $l_i \leftarrow \arg\min_l \|\mathbf{z}_i - \boldsymbol{\theta}_l\|_2^2$
- update each centroid by taking the mean of its assigned points: $\boldsymbol{\theta}_l \leftarrow \frac{1}{|\{\#i, l_i = l\}|} \sum_{i,\ l_i = l} \mathbf{z}_i$

Indeed, it is easy to show that the solution that minimizes (5.11), although NP-hard to compute, is such that each centroid is the empirical average of the points closest to it, which is the main inspiration behind the Lloyd-Max algorithm. In fact this algorithm is often directly referred to as *k-means* (henceforth when we say k-means algorithm we will refer to Lloyd-Max's algorithm).

Sketching method. In this application the sketching approach and the CL-OMPR algorithm are used to recover mixtures of Diracs $\pi_{\boldsymbol{\theta}} = \delta_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} \in \mathbb{R}^d$ from the sketch. Considering the Fourier features (5.5), once the sketch is computed recovering $\pi_{\Theta,\boldsymbol{\xi}}$ corresponds to minimizing the cost function (5.1) with the feature function

$$\mathbf{f}(\boldsymbol{\theta}) := \mathcal{A}\delta_{\boldsymbol{\theta}} = \left[e^{\mathrm{i}\boldsymbol{\omega}_j^\top \boldsymbol{\theta}}\right]_{j=1}^{m} . \tag{5.12}$$

where the $\boldsymbol{\omega}_j$'s are drawn *i.i.d.* from some distribution Λ. This framework is exactly that of Example 4.1.1 in Chapter 4, therefore all details of implementation of CL-OMP(R) for this case have already been derived in this previous chapter. In particular, we have outlined that the simple expressions of the functions $\mathbf{f}(\boldsymbol{\theta})$ and $\mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{J}_{\mathrm{Re}(\mathbf{f})}(\boldsymbol{\theta})^\top \mathrm{Re}(\mathbf{x}) + \mathbf{J}_{\mathrm{Im}(\mathbf{f})}(\boldsymbol{\theta})^\top \mathrm{Im}(\mathbf{x})$ (which, remember, are the only expressions required for the implementation of the method) would allow for efficient implementations using fast transforms instead of the full frequency matrix.

5.3.2 Setup

We compare k-means and CL-OMPR on synthetic and real data.

Generating synthetic data. Given d, k, the data points $\mathbf{z}_i \in \mathbb{R}^d$ are drawn *i.i.d.* from a well-clustered GMM generated as follows.

The weights $\boldsymbol{\xi}$ are drawn on the $k-1$-simplex according to a Dirichlet distribution $\mathtt{Dir}(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha}$ is a vector of so-called *concentration parameters* set to $\boldsymbol{\alpha} = \alpha\mathbf{1}$ is our experiments, with parameter $\alpha = 5$. Basically, the higher α is, the higher the probability for $\boldsymbol{\xi}$ to be close to a uniform weight vector $\mathbf{1}/k$ is.

The means $\boldsymbol{\mu}_l \in \mathbb{R}^d$ are drawn according to a Gaussian distribution $\mathcal{N}(0, c^2\sigma_{\boldsymbol{\mu}}^2\mathbf{I})$ with a constant $c = 1.5$. The $\sigma_{\boldsymbol{\mu}}^2$ is defined as $\sigma_{\boldsymbol{\mu}} := k^{1/d}$, using volumetric considerations: considering that a ball of radius r has volume $D_d r^d$ (with D_d a constant that depends on d), an isotropic Gaussian with covariance $\sigma^2\mathbf{I}$ "occupies" a volume $V_{\sigma^2} = \sigma^d V_1$, where V_1 is a reference volume for $\sigma = 1$, $\sigma_{\boldsymbol{\mu}}^2$ is defined such that $V_{\sigma_{\boldsymbol{\mu}}^2}$ contains k volumes V_1.

Real data. We will also test our method on real data. The considered problem consists in performing spectral clustering [NJW01] on the MNIST dataset [LCB98]. To test our method's performance on a large dataset, we use the original $7 \cdot 10^4$ images, that we complete with images artificially created by distortion of the original ones using the toolbox infMNIST proposed in [LCB07]. In our experiments we use up to $n = 10^6$ images. For each dataset, we extract SIFT [VF10] descriptors of each image, and compute the k-nearest neighbours adjacency matrix (with $k = 10$) using FLANN [ML09]. As we know there are ten classes, we compute the first ten eigenvectors of the associated normalized Laplacian matrix, and run CKM on these n 10-dimensional feature vectors.

Remark 5.3.1. *Note that spectral clustering requires the first few eigenvectors of the global Laplacian matrix, of size n^2, which becomes prohibitive for large n and does not fit with our compressive, streaming approach. There exist indeed compressive versions of spectral clustering [Tre+16a; Tre+16b] or efficient kernel methods such as in [CJJ12], combining them with the sketching method is left for future investigations.*

Computing the sketch. The frequency distribution is set as $\Lambda := \Lambda^{(Ar)}_{\sigma^2_{\text{freq.}}\mathbf{I}}$, for some parameter $\sigma^2_{\text{freq.}}$ either defined as $\sigma^2_{\text{freq.}} := \bar{\sigma}^2$ where $\bar{\sigma}^2$ is learned with `EstimMeanSigma`, or set manually (see next section). Then, as described in Algorithm 5, m frequency vectors $\boldsymbol{\omega}_j$ are drawn *i.i.d.* from Λ, and the sketch of the database is computed as (5.3) with the sketching function (5.5). For this application, during the computation of the sketch $\hat{\mathbf{y}}$ we also compute the maximum and minimum value of the data along each dimension (which can also be done in a streaming or distributed context): *i.e.* we store $\boldsymbol{\ell}, \mathbf{u} \in \mathbb{R}^d$ such that for all $1 \leq i \leq n$ we have $\boldsymbol{\ell} \leq \mathbf{z}_i \leq \mathbf{u}$.

CL-OMPR. We then recover parameters $(\Theta, \boldsymbol{\xi})$ with the CL-OMPR algorithm[4]. The optimization schemes of CL-OMPR are here constrained so that centroids $\boldsymbol{\theta}_l$ lie in $\boldsymbol{\ell} \leq \boldsymbol{\theta}_l \leq \mathbf{u}$). Unless otherwise specified, the maximization of the correlation in Step 1 of CL-OMP(R) is initialized by a point drawn uniformly between $\boldsymbol{\ell}$ and $\mathbf{u}$.

For this application the weights $\boldsymbol{\xi}$ are not used, only the centroids Θ are kept.

k-means. We compared our sketching approach with Matlab's `kmeans` function. The number of iterations is limited to 1000. By default the k-means algorithm is also initialized with points drawn uniformly between $\boldsymbol{\ell}$ and $\mathbf{u}$.

For both methods the quality of the reconstruction is then evaluated using the SSE (5.11). For the spectral clustering problem for handwritten digits recognition, we also use the **Adjusted Rand Index** [Ran71], which evaluates the difference between the clustering produced by the algorithm and the ground truth.

Results are averaged over 50 experiments, using the geometric mean for the SSE and arithmetic mean for the Adjusted Rand Index, computing time or memory consumption.

5.3.3 Choosing the parameter $\sigma^2_{\text{freq.}}$

For synthetic data drawn from a GMM, we use the function `EstimMeanSigma` to learn $\bar{\sigma}^2$ and put the parameter of the frequency distribution to $\sigma^2_{\text{freq}} = \bar{\sigma}^2$. Preliminary experiments similar to the experiments performed in Section 5.1.4 show that it performs as well as the best σ^2 chosen among a large range of parameters. Is it also true for real data ?

In some cases such as the speaker verification experiments that will be performed in Section 5.4.6, we will see that this learned parameter yields good results.

For the MNIST data, we performed a first experiment to assess the quality of the results when using $\sigma^2_{\text{freq.}} = \bar{\sigma}^2$. In Fig. 5.5 we show the SSE obtained by varying the parameter $\sigma^2_{\text{freq.}}$, and outline the value of the parameter learned with `EstimMeanSigma`. It is seen that this value yields decent results, but is a bit low compared to the best choice for $\sigma^2_{\text{freq.}}$. In practice,

[4] In Chapter 4 we have already established the superiority of CL-OMPR over CL-OMP and BCD for the task of recovering Diracs from Fourier measurements.

when we use this learned value, the CL-OMPR algorithm is seen to be quite unstable and have a non-negligible rate of "failure" between runs (recall that the results are averaged over 50 experiments).

Hence in the rest of the experiments, for the spectral clustering problem we manually choose $\sigma^2_{\text{freq.}} = 0.2$, which is seen to yield very stable results. We leave the search for a kernel learning method that would produce this parameter for this type of data for future work.

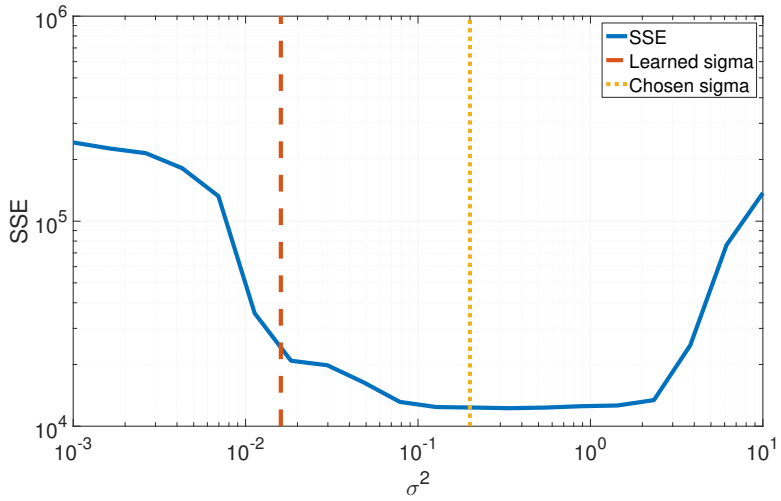

FIGURE 5.5: SSE result (lower is better) of CL-OMPR on the spectral clustering problem with MNIST data ($d = 10, k = 10$) with $n = 3 \cdot 10^5$, with respect to σ^2 for a frequency distribution $\Lambda^{(Ar)}_{\sigma^2\mathbf{I}}$. The parameter learned by `EstimMeanSigma` and the one chosen in practice are outlined.

5.3.4 Role of initialization

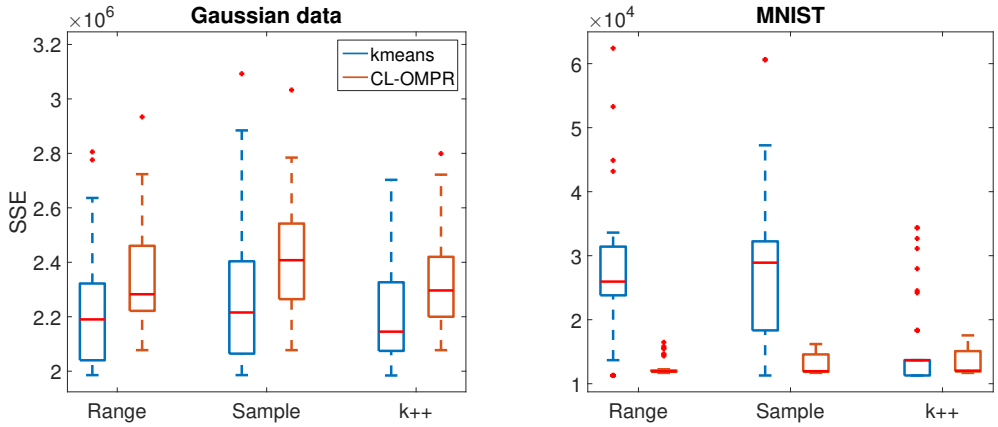

FIGURE 5.6: Comparison of initialization methods for k-means and CL-OMPR, with Gaussian synthetic data ($d = 10, k = 10, n = 2 \cdot 10^5$) or MNIST data (with the same d, k, n). We display box plots over 50 experiments, that indicates median value, 25th and 75th percentiles, maximum and minimum values (outsiders are indicated as points). For the sketched approach, only the "range" method actually fits in the compressive framework.

The k-means algorithm admits several initialization methods that can greatly influence the produced result. In CL-OMPR, the only step for which an initialization method must be defined is Step 1, the gradient descent that seeks for a maximum correlation between the residual signal and an atom.

By default, the CL-OMPR and k-means algorithms are initialized with points drawn uniformly between $\boldsymbol{\ell}$ and $\mathbf{u}$ (what we call "range" below). We perform an experiment to test this approach against two other initialization methods:

- **Range:** for CL-OMPR, Step 1 is initialized by a point drawn uniformly between $\boldsymbol{\ell}$ and $\mathbf{u}$; for `kmeans`, select k such points. This is the default initialization method in the rest of the experiments.
- **Sample:** for CL-OMPR, select a point $\boldsymbol{\theta} = \mathbf{z}_i$ from the data at random; for `kmeans`, select k such points.
- **k++,** a strategy analog to the k-means++ algorithm [AV07]: for CL-OMPR, select $\boldsymbol{\theta} = \mathbf{z}_i$ from the data with a probability inversely proportional to its distance to the *current* set of centroids Θ; for `kmeans`, run exactly the K++ algorithm [AV07].

Note that the last two strategies do *not* fit in the sketching framework, since they still require access to the data. They are implemented for testing purpose. In the rest of the experiments the "Range" strategy is always adopted.

Fig. 5.6 shows box plots for SSE results over 50 experiments, for both synthetic and real data. The k++ approach is seen to significantly improve `kmeans` on real data, while the CL-OMPR algorithm seems to be more robust to initialization strategy. The CL-OMPR algorithm also seems to perform surprisingly well on MNIST data.

5.3.5 Time and memory use on large-scale databases

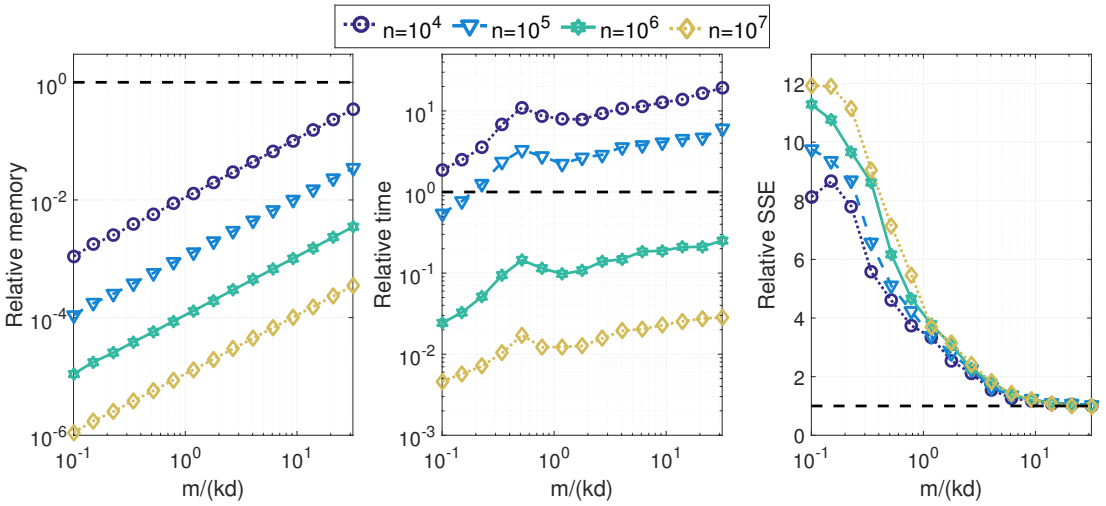

FIGURE 5.7: Time, memory and SSE of CL-OMPR divided by those of Matlab's `kmeans` function (black dotted line) on Gaussian synthetic data with $d = 10, k = 10$.

Next we examine the gains of the proposed sketching method compared to Matlab's `kmeans` function in terms of computation time and memory requirement, to confirm that the compressive approach is more efficient than using the full data for large databases.

Experiment. In Fig. 5.7, relative computation time, memory requirement and SSE of the sketching method (*i.e.* time, memory and SSE of CL-OMPR divided by those of Matlab's `kmeans` function) are showed, for four increasingly large databases with $n = 10^4, 10^5, 10^6, 10^7$ items. For the computation time, we do *not* take into account the computation of the sketch here but only the CL-OMPR algorithm, since we suppose the sketch to be computed beforehand in a streaming, distributed and parallel context.

It is seen that, for a sketch size that *does not seem to depend on the size of the database* n of approximately $m = 10kd = 1000$, the compressive approach reaches the precision of `kmeans` in terms of SSE, while being much more efficient on large databases, up to 10^4 times more memory efficient and 10^2 times faster for $n = 10^7$, for the same SSE. The reader must also keep in mind that `kmeans` is an extremely optimized implementation of Lloyd-Max's algorithm, while CL-OMPR is written in Matlab [Ker16].

5.3.6 Empirically sufficient sketch size

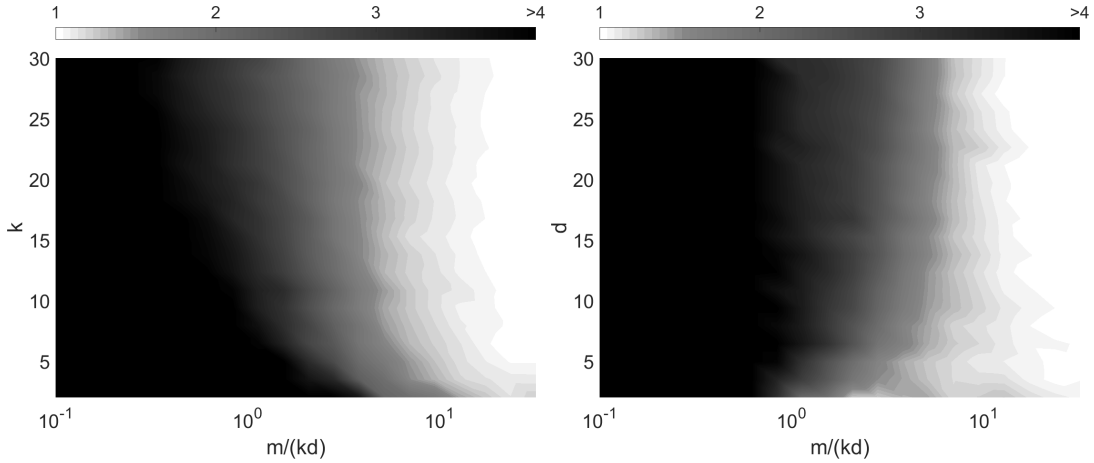

FIGURE 5.8: Relative SSE (SSE of CL-OMPR divided by that of Matlab's `kmeans`) on synthetic data with $n = 10^5$, with respect to relative sketch size $m/(kd)$, for many dimension d and number of components k. On the left $k = 10$ and on the right $d = 10$.

Since we observed on the previous experiment that the sketch size m empirically sufficient to obtain a good result does not seem to depend on the number of element n in the database, one guesses that it depends on the complexity of the problem instead, *i.e.* the dimension d and number of components k.

Experiment. In Fig. 5.8 we draw so-called phase transition diagrams, where we examine the relative SSE of the sketching method with respect to the relative sketch size $m/(kd)$, for many numbers of components k and dimensions d. It seems that a sketch size of the order of the numbers of parameters $m \approx \mathcal{O}(kd)$ is sufficient to obtain an SSE of the order of that produced by `kmeans`. Intuitively, this result is "optimal", by dimensionality arguments. In the next chapter (Chapter 6) our theoretical study will lead to a complexity $m \approx \mathcal{O}\left(d^2k^2\right)$ with additional logarithmic terms, which is still polynomial in d and k but sub-optimal compared to what is observed in practice. Further work will aim at bridging this gap.

5.3.7 Influence of the number of replicates

Often the k-means algorithm is repeated several times with random initializations, and the set of centroids yielding the lowest SSE is kept. One can wonder if increasing the number of replicates for the compressive approach also improve the results. However, in the compressive approach, we do not have access to the SSE in practice since the data are discarded after computation of the sketch. Hence, **when several replicates of CL-OMPR are performed,** ***we select instead the set of centroids that minimizes the cost function*** (5.1) (although we will see below that several replicates of CL-OMPR are in general not required).

Experiments. In Fig. 5.9, we compare k-means and CL-OMPR when using 1 or 5 replicates on the spectral clustering problem, in terms of SSE and Adjusted Rand Index. It is seen that CL-OMPR performs excessively well on this problem: it outperforms k-means in almost all settings. In particular, it is extremely *stable* between runs. While k-means is, as expected,

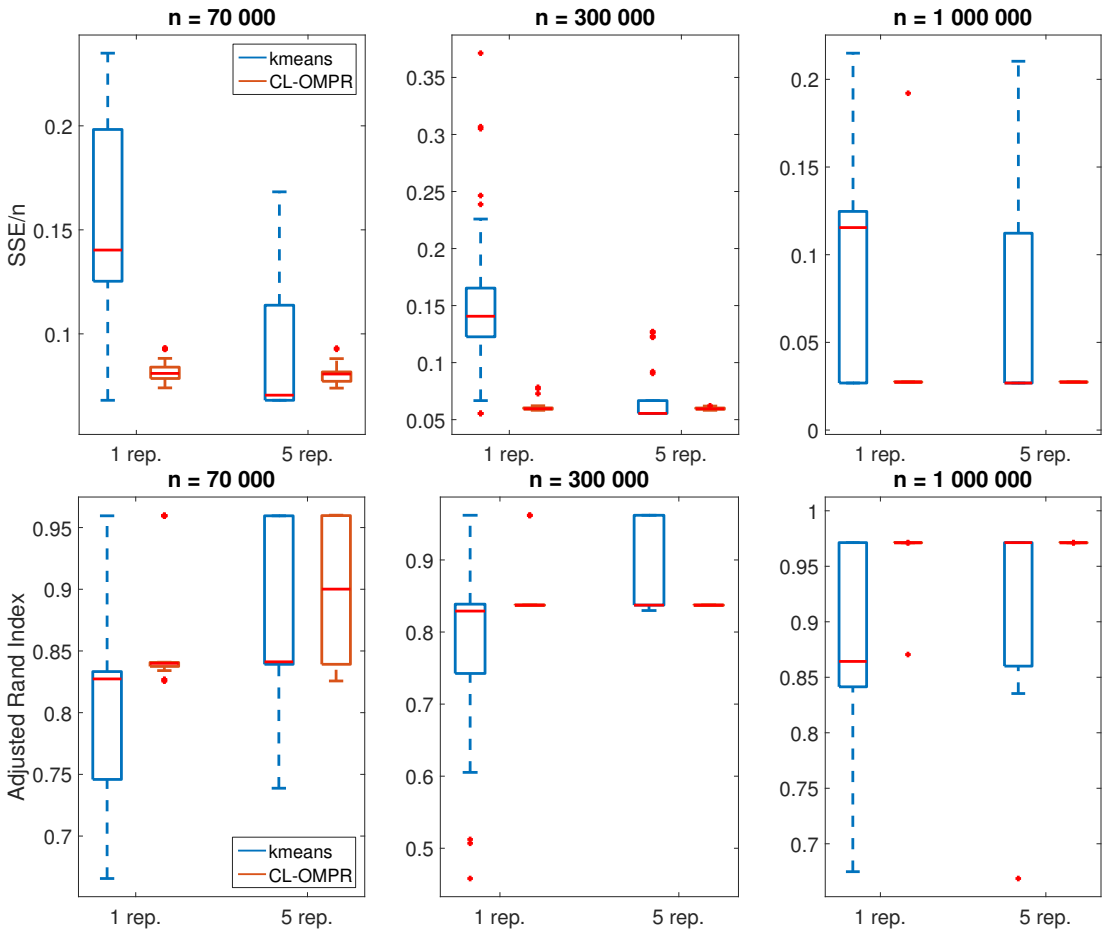

FIGURE 5.9: SSE divided by n (top, lower is better) and clustering performance with Adjusted Rand Index [Ran71] (bottom, higher is better) for k-means and CL-OMPR, on MNIST data ($d = 10, k = 10$), for three database sizes.

greatly improved by performing five replicates instead of one, there is almost no difference for CL-OMPR between one and five replicates.

5.3.8 Conclusion on mixtures of Diracs

In this first Section, we have shown that, by instantiating the CL-OMPR algorithm to recover mixtures of Diracs, we can efficiently recover centroids of clustered data from their sketch.

Compared to Matlab's `kmeans` implementation, our approach is several orders of magnitude faster and more memory efficient on large databases. It also performs surprisingly well on real data in a spectral clustering problems, where it is seen to be much more stable than `kmeans` with respect to the number of replicates and the initialization strategy.

The sketch size empirically sufficient to guarantee success of the CL-OMPR algorithm does not seem to depend on the size of the database, but rather on the complexity of the problem, with an evolution roughly proportional to the number of parameters $m \approx \mathcal{O}(kd)$.

5.4 Gaussian Mixture Models

We now implement the sketching method for estimating GMMs with unknown diagonal covariances, which is an extension of the original framework of Bourrier et al. [BGP13]. In Chapter 3, we obtained information preservation guarantees with a sufficient sketch size that was as large as the original database. Fortunately, in practice we are going to see that the sketch size seems to depend only on the complexity of the problem.

In the GMM with diagonal covariance framework, a mixture component is a Gaussian distribution

$$\pi_{\boldsymbol{\theta}} = \mathcal{N}\left(\boldsymbol{\mu}, \operatorname{diag}\left(\boldsymbol{\sigma}\right)\right) \tag{5.13}$$

with $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}) \in \mathbb{R}^{2d}$, where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean of the distribution and $\boldsymbol{\sigma} \in \mathbb{R}^d_+$ is the diagonal of the covariance of the Gaussian, with entries $\sigma_i^2 > 0$.

Given a database $\mathbf{z}_1, \ldots, \mathbf{z}_n$, learning a GMM $\pi_{\Theta,\boldsymbol{\xi}}$ is usually done by minimizing the negative log-likelihood:

$$\min_{\Theta,\ \boldsymbol{\xi}} \left(-\sum_{i=1}^{n} \log \pi_{\Theta,\boldsymbol{\xi}}(\mathbf{z}_i) \right) \tag{5.14}$$

The most classic heuristic to perform this task is the Expectation Maximization (EM) algorithm [DLR77], against which we are going to compare our sketching approach.

5.4.1 Implementation of CL-OMP(R)

Given randomly drawn frequencies $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m \in \mathbb{R}^d$, the feature function $\mathbf{f}$ is a sampling of the characteristic function of a Gaussian, which is:

$$\mathbf{f}(\boldsymbol{\theta}) = \left[e^{\mathrm{i}\boldsymbol{\omega}_j^\top \boldsymbol{\mu}} e^{-\frac{1}{2}\sum_{i=1}^d \omega_{j,i}^2 \sigma_i^2} \right]_{j=1}^m \tag{5.15}$$

As usual (see Section 4.3.2), implementing CL-OMP(R) only requires being able to compute $\mathbf{f}$ and the function $\mathbf{g}$ defined by (4.12). Similar to the Dirac case, in the Gaussian case we can obtain simplified expressions:

$$\mathbf{f}(\boldsymbol{\theta}) = \rho_{\text{im.}}\left(\mathbf{W}^\top \boldsymbol{\mu}\right) \odot \rho_{\text{re.}}\left(-\frac{1}{2}\mathbf{W}_2^\top \boldsymbol{\sigma}\right) \tag{5.16}$$

where $\mathbf{W}_2 = \mathbf{W} \odot \mathbf{W}$ with $\odot$ the Hadamard product (element by element multiplication), $\rho_{\text{im.}}(\cdot)$ is the pointwise application of $x \mapsto e^{\mathrm{i}x}$ and $\rho_{\text{re.}}(\cdot)$ is the pointwise application of $x \mapsto e^x$. For the function $\mathbf{g}(\boldsymbol{\theta}, \mathbf{x})$ we get:

$$\begin{aligned} \mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) &= \begin{bmatrix} \sum_{j=1}^m \left(-\sin(\boldsymbol{\omega}_j^\top \boldsymbol{\theta})\operatorname{Re}\left(x_j\right)\boldsymbol{\omega}_j + \cos(\boldsymbol{\omega}_j^\top \boldsymbol{\theta})\operatorname{Im}\left(x_j\right)\boldsymbol{\omega}_j\right) e^{-\frac{1}{2}\sum_{i=1}^d \omega_{j,i}^2 \sigma_i^2} \\ \sum_{j=1}^m \left(-\frac{1}{2}\cos(\boldsymbol{\omega}_j^\top \boldsymbol{\theta})\operatorname{Re}\left(x_j\right)\boldsymbol{\omega}_j^{\odot 2} - \frac{1}{2}\sin(\boldsymbol{\omega}_j^\top \boldsymbol{\theta})\operatorname{Im}\left(x_j\right)\boldsymbol{\omega}_j^{\odot 2}\right) e^{-\frac{1}{2}\sum_{i=1}^d \omega_{j,i}^2 \sigma_i^2} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{W}\Big(-\operatorname{Im}\left(\mathbf{f}(\boldsymbol{\theta})\right) \odot \operatorname{Re}\left(\mathbf{x}\right) + \operatorname{Re}\left(\mathbf{f}(\boldsymbol{\theta})\right) \odot \operatorname{Im}\left(\mathbf{x}\right)\Big) \\ -\frac{1}{2}\mathbf{W}_2\Big(\operatorname{Re}\left(\mathbf{f}(\boldsymbol{\theta})\right) \odot \operatorname{Re}\left(\mathbf{x}\right) + \operatorname{Im}\left(\mathbf{f}(\boldsymbol{\theta})\right) \odot \operatorname{Im}\left(\mathbf{x}\right)\Big) \end{bmatrix} \begin{matrix} \} \text{ size } d \\ \} \text{ size } d \end{matrix} \end{aligned} \tag{5.17}$$

where $\mathbf{x}^{\odot 2} = \mathbf{x} \odot \mathbf{x}$.

Like the Dirac case these expressions may potentially allow for the use of fast transforms instead of full matrix-vector multiplications (see Section 4.3.3). However in that case one must not only implement fast versions of $\boldsymbol{\mu} \mapsto \mathbf{W}^\top \boldsymbol{\mu}$ and $\mathbf{x} \mapsto \mathbf{W}\mathbf{x}$, but also $\boldsymbol{\sigma} \mapsto \mathbf{W}_2^\top \boldsymbol{\sigma}$ and $\mathbf{x} \mapsto \mathbf{W}_2\mathbf{x}$, where $\mathbf{W}_2 = \mathbf{W} \odot \mathbf{W}$. Methods to *learn* fast factorization of matrices such as [LG16] may thus become all the more useful.

5.4.2 A fast hierarchical alternative to CL-OMP(R)

As mentioned before, the greedy CL-OMP(R) scales quadratically in the number of components k, which is potentially limiting. In this section we describe an alternative to CL-OMP(R), *specific to GMMs*, that scales in $k \log k$ instead of k^2.

Inspired by hierarchical versions of the classic Expectation Maximization (EM) algorithm [SSH13], this algorithm replaces the greedy step by a division of each Gaussian component along its direction of highest variance. The algorithm in [SSH13] alternates this division step with a few EM steps, while in the compressive approach we alternate with the gradient descent (Step 4) already present in CL-OMPR.

This algorithm, coined Hierarchical Compressive GMM estimation (HCGMM), is described in Algorithm 6. The binary split is performed by calling the function `Split`. In the case where

the targeted number of components k is not a power of 2, we split the GMM until the support reaches a size $2^{\lceil \log_2 k \rceil} > k$, then reduce it with a Hard Thresholding.

Function Split(Θ): split each Gaussian in the support along its dimension of highest variance

Data: Support $\Theta = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k\}$ where $\boldsymbol{\theta}_l = (\boldsymbol{\mu}_l, \boldsymbol{\sigma}_l)$
Result: New support Θ^{new} of size $|\Theta^{new}| = 2k$
$\Theta^{new} \leftarrow \emptyset$;
for $l \leftarrow 1$ **to** k **do**
  $i_0 \leftarrow \arg\max_{1 \leq i \leq d} \sigma_{l,i}^2$;
  $\Theta^{new} \leftarrow \Theta^{new} \cup \{(\boldsymbol{\mu}_l - \sigma_{l,i_0}\mathbf{e}_{i_0}, \boldsymbol{\sigma}_l), (\boldsymbol{\mu}_l + \sigma_{l,i_0}\mathbf{e}_{i_0}, \boldsymbol{\sigma}_l)\}$;
end

Algorithm 6: Hierarchical Compressive GMM estimation (HCGMM)

Data: Measurement vector $\mathbf{y} \in \mathbb{C}^m$, function $\mathbf{f} : \mathbb{R}^q \to \mathbb{C}^m$, sparsity $k > 0$
Result: Parameters $(\Theta, \boldsymbol{\xi})$
$\Theta \leftarrow \emptyset$;
begin Initialize with *one* atom highly correlated with the sketch
  Perform **Step 1** and **Step 2** of Alg. 3;
end
for $t \leftarrow 1$ **to** $\lceil \log_2 k \rceil$ **do**
  begin Split each Gaussian in the support along its dimension of highest variance
    $\Theta \leftarrow$ `Split`(Θ);
  end
  if $|\Theta| > k$ **then**
    $\Theta \leftarrow$ `HardThres`$(\Theta, \mathbf{y}, \mathbf{f}, k)$;
  end
  Perform **Step 3** and **Step 4** of Algorithm 3;
end
Normalize $\boldsymbol{\xi}$ such that $\sum_{l=1}^k \xi_l = 1$

Let us now turn to our experiments.

5.4.3 Setup

We first test the estimation of GMMs from a sketch on synthetic data.

Generating the data. Given a number of components k and a dimension d, we randomly generate the parameters of a GMM as follows, inspired by classic Bayesian prior distributions.

Like in the previous section, the weights $\boldsymbol{\xi}$ are generated according to a Dirichlet distribution `Dir`$(\alpha \mathbf{1})$ with parameter $\alpha = 5$, and the means are drawn from a distribution $\mathcal{N}\left(0, c^2 k^{1/d} \mathbf{I}\right)$, with a parameter $c = 1$ (recall that in the previous section we chose $c = 1.5$).

Each variance parameter $\sigma_{l,i}^2$, for $1 \geq l \geq k$ and $1 \leq i \leq d$, is drawn according to an inverse-Wishart distribution with $p > 2$ degrees of freedom scaled such that $\mathbb{E}(\sigma_{l,i}^2) = 1$, *i.e.* it is drawn as $\sigma_{l,i}^2 = (p-2)/x$, where $x \sim \chi^2(p)$ is drawn according to a χ^2 distribution with p degrees of freedom. Basically, the higher p is, the more concentrated $\sigma_{l,i}^2$ is around 1. In practice we put $p = 10$, which yields the histogram showed in Fig. 5.10.

Computing the sketch. In all experiments we use the frequency distribution defined as $\Lambda := \Lambda_{\bar{\sigma}^2}^{(Ar)}$ with $\bar{\sigma}^2$ learned by `EstimMeanSigma`, including on real data (Section 5.4.6). Note that, since we know that $\mathbb{E}(\sigma_{l,i}^2) = 1$ here, the learning procedure `EstimMeanSigma` should learn a parameter $\bar{\sigma}^2$ close to 1. However for fairness we do not assume prior knowledge of this, and let the algorithm run with the value $\bar{\sigma}^2$ learned by `EstimMeanSigma`.

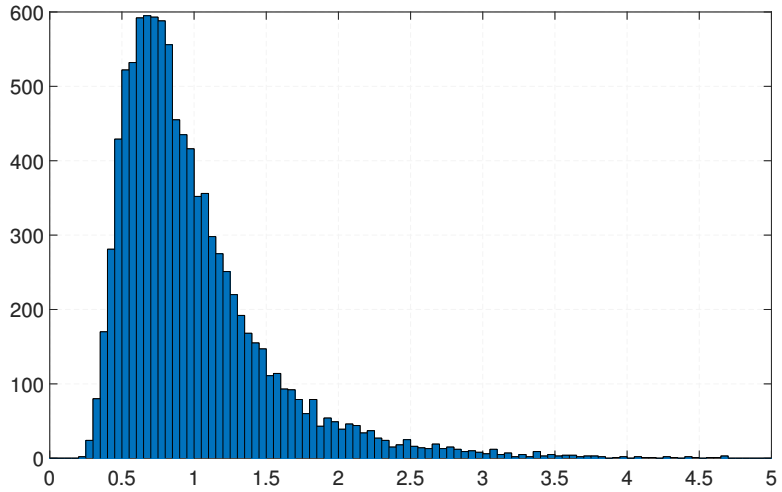

FIGURE 5.10: Histogram of 5000 draws of σ^2 according to a normalized inverse-Wishart distribution with $p = 10$ degrees of freedom.

Then m frequency vectors $\boldsymbol{\omega}_j$ are drawn *i.i.d.* from Λ, and the sketch of the database is computed as described in Algorithm 5.

Compressive algorithms. The CL-OMP, CL-OMPR or HCGMM algorithm is then performed to recover a GMM from the sketch. Step 1 in all algorithms is initialized with a centered isotropic Gaussian $\mathcal{N}\left(0, b\bar{\sigma}^2\mathbf{I}\right)$, where $\bar{\sigma}^2$ is the parameter that has been learned for the frequency distribution and b is a random variable drawn uniformly between 0.5 and 1.5. During the optimization schemes of the algorithms, the means are unconstrained, and the variances $\sigma^2_{l,i}$ are constrained to be above a value 10^{-15} for numerical stability.

Expectation Maximization. We compare our approach with the classic EM [DLR77] algorithm for GMM estimation. We use the `gmm` function of the toolbox VLFeat [VF10], which by default already returns a GMM with diagonal covariances. Each run is limited to 100 iterations. The EM algorithm is either performed only once and denoted "EM1", or repeated 10 times with random initializations and the result that yields the best log-likelihood is selected, which is denoted "EM10". As we observed in the previous experiments that the sketching method is usually stable between runs, we always perform only one replicate of the compressive approaches.

Evaluation measure. Since we know the ground truth GMM here, we use a distance measure that does not depend on the training data (as would traditional log-likelihood for instance) but directly quantifies the difference between the ground truth GMM and the recovered one. We use a symmetrized version of the Kullback-Leibler divergence (Definition A.1.7), still referred to as "KL-divergence" in practice:

$$\begin{aligned} d(\pi, \pi') &= D_{\mathrm{KL}}\left(\pi||\pi'\right) + D_{\mathrm{KL}}\left(\pi'||\pi\right) \\ &= \mathbb{E}_{\mathbf{z}\sim\pi} \log \pi(\mathbf{z}) + \mathbb{E}_{\mathbf{z}\sim\pi'} \log \pi(\mathbf{z}') - \mathbb{E}_{\mathbf{z}\sim\pi'} \log \pi(\mathbf{z}') - \mathbb{E}_{\mathbf{z}\sim\pi} \log \pi'(\mathbf{z}) \end{aligned}$$

To compute it in practice, given two GMMs π and π', we draw $n = 3 \cdot 10^5$ samples $\mathbf{z}_i$ from π and $\mathbf{z}'_i$ from π' (independent of the training data used in the algorithm), then compute

$$d(\pi, \pi') \approx \frac{1}{n} \sum_{i=1}^{n} \left(\log \pi(\mathbf{z}) + \log \pi'(\mathbf{z}') - \log \pi(\mathbf{z}') - \log \pi'(\mathbf{z})\right) \tag{5.18}$$

Results are averaged over 50 experiments, using the geometric mean for the KL-divergence and the arithmetic mean for the computation time and the memory consumption.

5.4.4 Role of database size

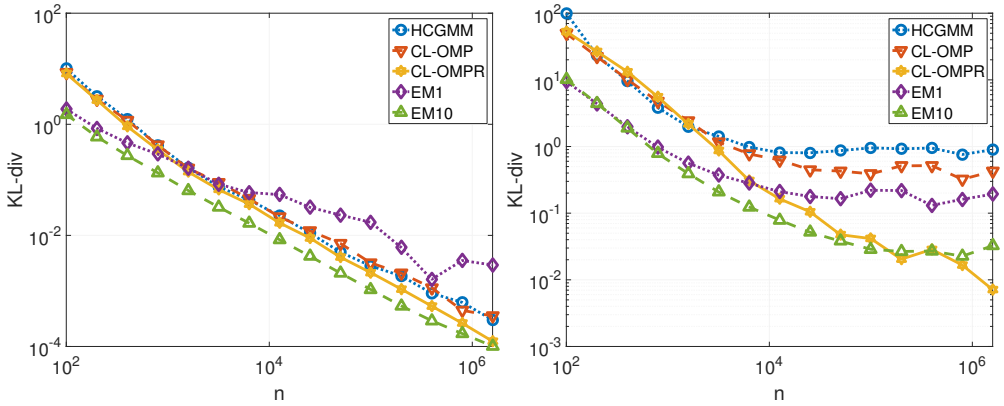

FIGURE 5.11: Comparison of the different algorithms: KL-divergence result with respect to n, in dimension $d = 10$, with $k = 5$ (left) or $k = 20$ (right), using $m = 15kd$ frequencies for the sketching method.

We first compare the proposed compressive algorithms against the classic EM algorithm for increasingly large database sizes, to assess precision of the compressive methods, and time and memory savings.

Precision of the estimation. In Fig. 5.11, KL-divergence result are shown for all algorithms, with respect to the number of items n in the database. Two problems, $k = 5$ and $k = 20$ in dimension $d = 10$ are shown. For $k = 5$, all compressive algorithms yield the same precision for all database sizes, which would suggest using the fastest one (HCGMM) in practice. For $k = 20$ however, CL-OMPR significantly outperforms other approaches for large databases, confirming its capacity to handle more difficult problems. In both $k = 5$ and $k = 20$, EM10 is seen to significantly outperform EM1 at large n, which is to be expected. And, while it is performed for only one run, the CL-OMPR algorithm is on par with EM10 in each case, which confirms its stability compared to classic approaches, as already observed for the Dirac case.

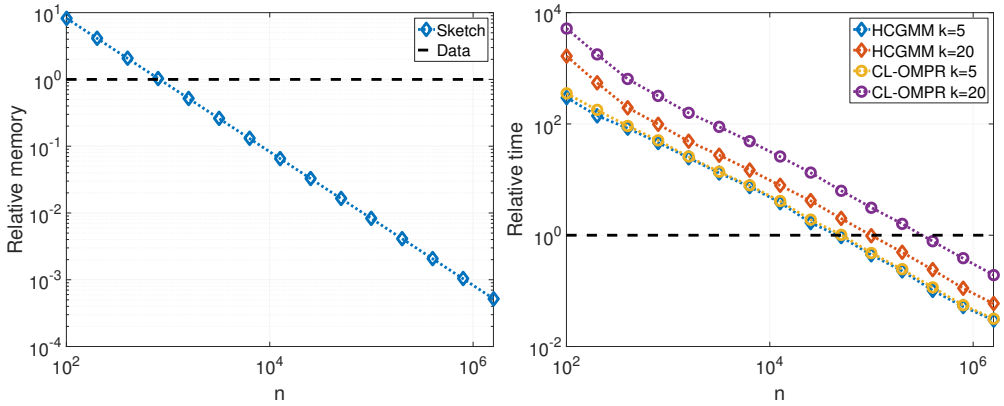

FIGURE 5.12: Relative time and memory consumption of the sketch algorithms with respect to the EM10 algorithm, in dimension $d = 10$, with $m = 15kd$. The relative memory used by the sketching method also includes the frequency matrix.

Time and memory savings. In Fig. 5.12 we examine relative time and memory consumption of the compressive approaches, compared to EM10. Once again, for the time curve we do *not* outline the time taken to compute the sketch, since we suppose it has been done beforehand in an streaming or distributed context. As expected, for large n the compressive approaches are

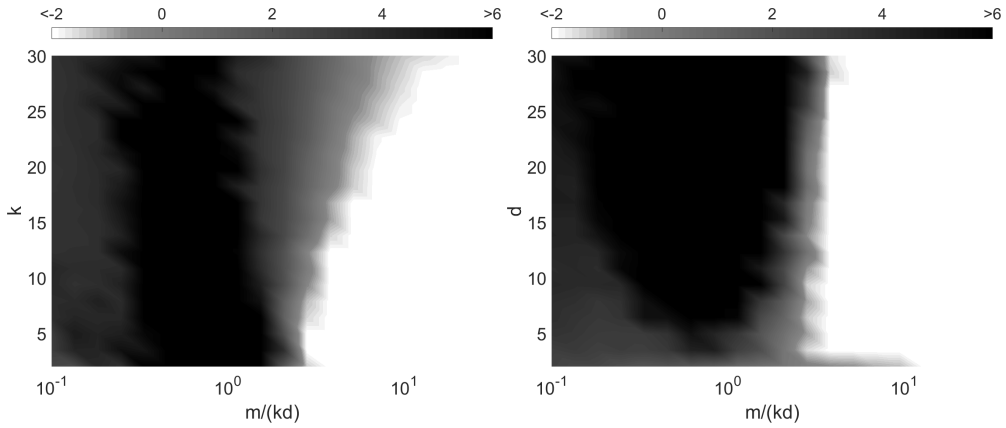

FIGURE 5.13: Log-KL divergence result for CL-OMPR, with respect to the relative sketch size $m/(kd)$. On the left, $d = 10$, on the right $k = 10$.

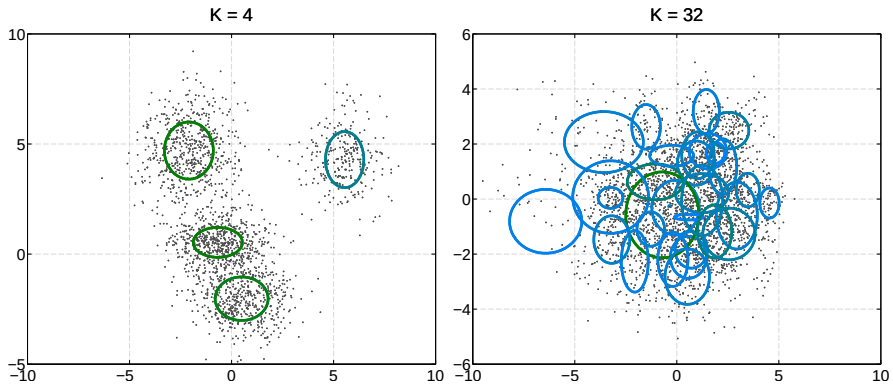

FIGURE 5.14: (Left) Clustering example: partitioning synthetic data into a small number of meaningful groups; (Right) Example of density estimation problem: using EM to fit a GMM with many components to speech data (MFCC features). The HCGMM algorithm is expected to be a suitable alternative to CL-OMP and EM for the latter problem on the right.

order of magnitudes faster and more memory efficient than VLFeat's state-of-art C++ implementation of EM. It is also seen that when the number of components k increases, HCGMM is faster than CL-OMPR, since it scales in $k \log k$ instead of k^2.

5.4.5 Empirically sufficient sketch size

Like in the case of mixtures of Diracs, we evaluate the sketch size m empirically sufficient to obtain acceptable estimation results, by constructing transitions diagrams for the CL-OMPR algorithm (Fig. 5.13). Despite a little deviation for high number of components k (that might be an artifact from using the CL-OMPR algorithm or the definition of the KL-divergence itself), it is seen that, like in the Dirac case a sketch size in $m \approx \mathcal{O}(kd)$ is approximately sufficient.

It confirms that the theoretical results of Chapter 3 indeed seems to be largely pessimistic. As we will see, the analysis that will be presented in Chapter 6 will unfortunately not apply to GMMs with unknown covariances, and improving upon the current theoretical results in this case is left for future investigations.

5.4.6 Large-scale proof of concept: Speaker verification

Gaussians Mixture Models are popular for their capacity to smoothly approximate *any* distribution [RR95] by a large number of Gaussians. This is often the case with real data, and the problem of *fitting* a large GMM to data drawn from some distribution is somewhat different

from that of *clustering* data and identifying reasonably well separated components, as presented in the previous section (Figure 5.14). In order to try out the sketching method on this challenging task, we test it on a speaker verification problem, with a classic approach requiring GMM referred to as Universal Background Model (GMM-UBM) [RQD00].

Overview of Speaker Verification

Given a fragment of speech and a candidate speaker, the goal is to assess if the fragment was indeed spoken by that person.

We quickly describe GMM-UBM in this section. For more details we refer the reader to the original paper [RQD00]. Similar to many speech processing tasks, this approach uses Mel Frequency Cepstrum Coefficients (MFCC) and their derivatives (Δ-MFCC) as features $\mathbf{z}_i$. Those features have been often modeled with GMMs or more advanced Markov models. However, in our framework we do *not* use Δ-MFCC; indeed those coefficients typically have a negligible range in dynamic compared to the MFCC, which results in a difficult and unstable choice of frequencies. This problem may be potentially solved by a pre-whitening of the data, which we leave for future work. In this configuration, the speaker verification results will indeed not be state-of-the-art, but our goal is mainly to test our compressive approach on a different type of problem than that of clustering synthetic data, for which we have already observed excellent results.

In the GMM-UBM model, each speaker S is represented by one GMM $(\Theta_S, \boldsymbol{\xi}_S)$. The key point is the introduction of a model $(\Theta_{UBM}, \boldsymbol{\xi}_{UBM})$ that represents a "generic" speaker, referred to as Universal Background Model (UBM). Given speech data $\mathcal{Z}$ and a candidate speaker S, the statistic used for hypothesis testing is a likelihood ratio between the speaker and the generic model:

$$T(\mathcal{Z}) = \frac{\pi_{\Theta_S, \boldsymbol{\xi}_S}(\mathcal{Z})}{\pi_{\Theta_{UBM}, \boldsymbol{\xi}_{UBM}}(\mathcal{Z})}. \tag{5.19}$$

If $T(\mathcal{Z})$ exceeds a threshold τ, the data $\mathcal{Z}$ are considered to be uttered by the speaker S.

The GMMs corresponding to each speaker must somehow be "comparable" to each other and to the UBM. Therefore, the UBM is learned *prior* to individual speaker models, using a large database of speech data uttered by many speakers. Then, given training data $\mathcal{Z}_S$ specific to one speaker, one M-step from the EM algorithm *initialized with the UBM* is used to adapt the UBM and derive the model $(\Theta_S, \boldsymbol{\xi}_S)$. We refer the reader to [RQD00] for more details on this procedure.

In our framework, the EM or compressive estimation algorithms are used to learn the UBM.

We note that this type of signal processing task may fully benefit from the advantages of the sketch structure. For instance, in practice one can imagine collecting bit by bit the data to train the UBM in a real-life environment, in which case the sketch and the UBM may be progressively updated without having to keep the spoken fragments, possibly of sensitive nature.

Setup

The experiments were performed on the NIST05 speaker verification database. Both training and testing fragments are 5-minutes conversations between two speakers. The database contains approximately 650 speakers, and 30 000 trials.

The MFCCs are computed using the Voicebox toolbox [Bro05]. After filtering the audio data by a speech activity detector, the MFCCs are computed on 23ms frames with a 50% overlap. The first coefficient is removed and we obtain 12-dimensional features ($d = 12$).

Results are presented by choosing the threshold τ that yields the same rates of false alarm and missed detection, referred to as Equal Error Rate (EER). Each result is obtained as the mean of five experiments.

In all experiments, except when indicated otherwise, the compressive methods are performed using a sketch obtained by compressing the entire database of $n = 2.10^8$ MFCC vectors after voice activity detection. Here the computation of the sketch is performed taking advantage of distributed computing, by dividing the database into 200 parts that are then

compressed simultaneously on a computer cluster. Hence, even for a high number of frequencies $m = 10^5$ the compression of the $n = 2.10^8$ items takes less than an hour.

Results

	EER (%)		Time (s)	
	CL-OMPR	HCGMM	CL-OMPR	HCGMM
$m = 10^3$	40.3	32.5	7.10^2	5.10
$m = 10^4$	29.4	29.0	7.10^3	5.10^2
$m = 10^5$	28.8	28.6	7.10^4	5.10^3

TABLE 5.2: Comparison between CL-OMPR and HCGMM for speaker verification, with $k = 64$.

Hierarchical algorithm In the previous section, HCGMM was observed to be less accurate than CL-OMPR. However, as mentioned before the estimation problem considered here is somehow not to identify well-separated components, but rather to fit a GMM with a large number of components to a smooth probability density. In the first case, on synthetic data, HCGMM is indeed expected to sometimes yield poor results: unlike a Matching Pursuit-based approach such as CL-OMPR, at each iteration it locally divides the current Gaussians rather than "exploring" elsewhere. In the second case however, HCGMM may yield a correct approximation of the smooth density, by successively approaching it with GMMs at increasingly finer scales.

In Table 5.2, we compare the results obtained with CL-OMPR and HCGMM on the speaker verification task using $k = 64$ Gaussians in the UBM. Results are indeed similar when the number of frequencies m is large, and even surprisingly better with HCGMM for a low number of frequencies $m = 1000$. Naturally, HCGMM is much faster than CL-OMPR, with more than a 10 times speedup.

Sketching a large database In Table 5.3, we compare EER results when using either $n_1 = 3.10^5$ items uniformly selected in the database to cover all speakers, or all $n_2 = 2.10^8$ items in the database. The compressive HCGMM is performed at both scales, while EM is only performed with n_1 items, since the whole database is too large to be handled by the VLFeat toolbox on a machine with $8\,GB$ of RAM. For the compressive approach, the use of the entire database indeed improves the results when compared to using only n_1 items to compute the sketch. At low $k = 8$ or $k = 64$ and high number of frequencies m, the compressive approach using n_2 items outperforms EM using only n_1 items.

		$K = 8$		$K = 64$		$K = 512$	
		n_1	n_2	n_1	n_2	n_1	n_2
EM		31.4	n/a	29.5	n/a	27.5	n/a
Alg. 2	$m = 10^3$	32.5	**31.2**	31.1	32.5	31.2	29.4
	$m = 10^4$	32.1	**30.7**	30.2	**29.0**	30.3	29.1
	$m = 10^5$	32.5	**30.7**	29.8	**28.6**	29.4	29.2

TABLE 5.3: Comparison between EM and HCGMM for speaker verification, in terms of EER, for $n_1 = 3 \cdot 10^5$ or $n = 2 \cdot 10^8$. For HCGMM, results that outperform those of EM are outlined.

Limitations due to coherence. While increasing the number of components k seems to consistently improves the results of EM, it is not the case with the compressive method for a fixed sketch size m. A possible intuitive explanation could be that, by increasing the number of components we also increase the coherence between them – *i.e.* the Gaussians in the GMM are increasingly overlapping each other – which makes it more and more difficult to handle for any sparsity-based approach. In practice, it results in many components in the GMM having

weights $\xi_l \approx 0$. In other words, the algorithm outputs a k'-GMM with $k' < k$: there seems to be a "limit" number of components above which additional Gaussians are useless. It may be possible to deal with a higher level of sparsity by drastically increasing the number of frequencies m, at the cost of higher compression and estimation times.

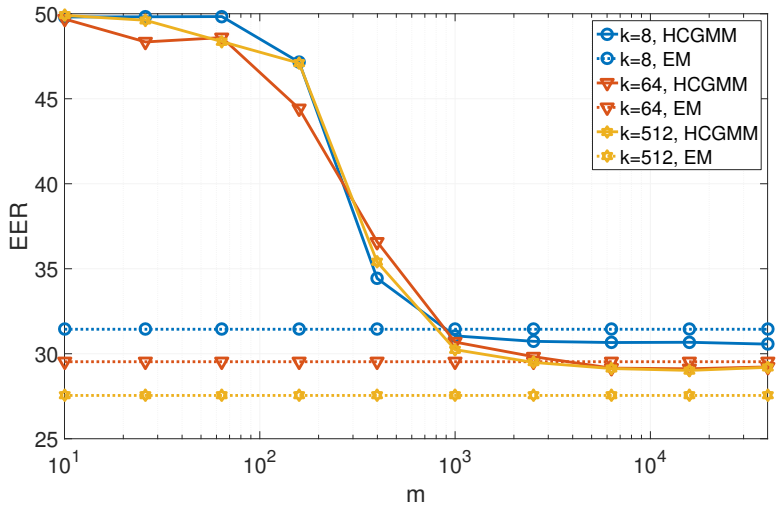

FIGURE 5.15: Equal Error Rate (lower is better) with respect to sketch size, using $n_1 = 3 \cdot 10^5$ items for EM and $n_2 = 2 \cdot 10^8$ items for HCGMM.

Number of components k and compression. In Figure 5.15 we study the effect of m for various numbers of components $k = 8$, 64 and 512. In each case we observe a sharp phase transition going from an EER of 50%, which corresponds to random guessing, to the results observed in Table 5.3. Somehow surprisingly, this phase transition does not seem to depend on k, unlike the one observed on synthetic data. As mentioned before it could be interesting to drastically increase m to see if the gap between results obtained EM and those obtained with HCGMM can be bridged in the $k = 512$ case, however the phase transition pattern does not support this idea but rather a limitation of the method itself, maybe in the algorithmic approach.

Overall, results on synthetic and real data show that the *fitting* problem is, as expected, more challenging than the *clustering* problem for the proposed sparsity-based approach. Indeed, while the clustering problem (synthetic data) is that of identifying well-separated components of a sparse distribution, the fitting problem is similar to a *sparse approximation* task, which is known to be challenging when the "signal" (*i.e.* the true distribution of the data) is *not* sparse. Nevertheless, let us point out that in Figure 5.15, results approaching those of EM are obtained for $m = 3000$ frequencies only, which corresponds to a whopping 33000-fold compression of the database.

5.4.7 Conclusion on compressive GMM estimation

In this application the compressive method was used to estimate GMMs with unknown diagonal covariances, for which it was naturally seen to be significantly more efficient than EM on large databases. In addition to the greedy approaches proposed in Chapter 4, we developed a hierarchical algorithm that scales in $k \log k$ instead of k^2 with respect to the number of components. Both CL-OMP and this algorithm were observed to perform on par with CL-OMPR on simple problems, while CL-OMPR outperforms them on more complex problems.

Similar to the previous application a sketch size in $m \approx \mathcal{O}(kd)$ was also observed to be sufficient, and furthermore this sketch size is independent from the number of points n in the database, which further indicates that the theoretical results obtained at the end of Chapter 3 seem to be sub-optimal.

The method was applied to a speaker verification problem, where the problem is more that of computing a smooth approximation of a distribution with many components than that of identifying separated components. On this problem the sparsity-based sketching method was

naturally observed to be more challenged, and the hierarchical approach was seen to be more effective than the greedy method. However the sketching method is easily applicable to a quantity of data for which classic approaches would require huge computational power.

5.5 Comparison with coresets

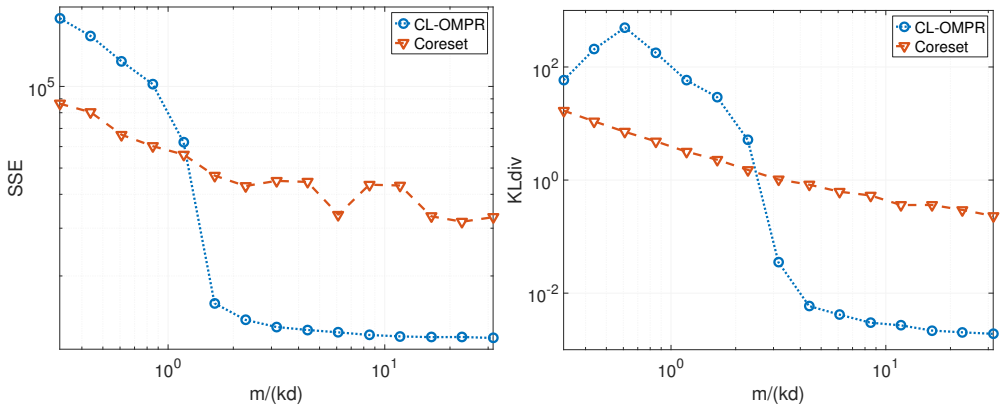

FIGURE 5.16: Comparison of the sketching method and the coreset approach in [Luc+17]. Left: SSE result for the k-means problem, which for the sketching method corresponds to recovering mixtures of Diracs, on MNIST data ($k = 10,\ d = 10$) with $n = 2 \cdot 10^5$. Right: KL-divergence result for GMM estimation on synthetic data with $k = 10,\ d = 10,\ n = 2 \cdot 10^5$.

In this short section, we briefly compare our sketching method with the coreset method described in [Luc+17] which is very simple to implement, for both the k-means problem (Section 5.3) or learning GMMs with diagonal covariances (Section 5.4).

Coresets. As described in the introduction of the thesis, a coreset is a summary of a database usually formed by a reduced number of weighted points, often taken as an adaptive subsampling of the data or a hierarchical construction. The method described in [Luc+17] is seeded with k-means++ [AV07], then performs an adaptive sampling of the database to return a collection of m weighted items from the database, where points far from the seeds are sampled with more probability but have lower weight. Then a weighted version of k-means or EM is performed on the coreset to learn either centroids or a GMM. In [Luc+17], guarantees are given when learning a GMM on this coreset[5].

Note that, as such, the global memory occupied by the coreset is $m(d+1)$ (since it is formed by m points and their weights), which is exactly the memory occupied by the sketch plus the collection of frequencies (although the sketching method has the advantage that the frequency matrix can be shared by several instantiations of the method). However we have seen that the sketching method could leverage the use of fast transforms for speed-up and reduced memory use.

Experiment. In Fig. 5.16 we compare results for the coreset method and the sketching method with respect to their size m. We consider MNIST data for the k-means problem, and synthetic data for the GMM estimation problem. In each case, it is seen that the results obtained with the coreset method improve regularly and steadily when the size of the coreset increases. For the sketching method however, we witness the "phase transition" pattern already observed is past experiments: experiments are unsuccessful until a certain sketch size is reached then they become fully successful. Hence it is seen that, *before* the phase transition, coresets outperform the sketching method, while *after* the phase transition, the sketching method is more precise than the coreset method.

[5] Although in the original paper [Luc+17] guarantees are only given for GMM for this particular coreset, we also adopt the same method for k-means, for the simplicity of its implementation. It is very similar to other coresets for which guarantees are given for k-means.

5.6 Mixtures of elliptic stable distribution

We now instantiate the CL-OMPR algorithm to estimate mixtures of elliptic stable distributions. In Chapter 3 Section 3.3.4, we have obtained theoretical guarantees similar to those for GMMs with diagonal covariances, for which the sufficient sketch size is as large as the original database. However in that case these results yield, to our knowledge, the first estimator with guarantees for multivariate mixtures of stable distributions.

As described in Section 3.3.4, mixtures of multivariate stable distributions have been scarcely used in practice, due to the lack of algorithms to estimate them. In the univariate case, a few methods do exist, mainly based on Bayesian approaches [SGKR10; Cas04], using computationally intensive tools such as MCMC simulations. Interestingly, some works have tried to abstract from these demanding and unpractical approaches by *defining a new class* of distributions inspired by stable distributions [Sho+10; Bro+13], and our sketching method may also apply in this case.

None of these approaches work in the multivariate case however, our instantiation of CL-OMPR might be the first algorithm able to handle this problem. Furthermore, in most existing methods [Cas04; Sho+10; Bro+13] practitioners limit themselves to $k = 2$ or $k = 3$ components for univariate variables, while CL-OMPR easily handles $k = 10$ or more components in dimension $d = 10$ or more.

Brief comparison with reported results. We leave comparison of CL-OMPR to existing approaches (in the univariate case) for future work. However, as an illustration, let us already point out a first comparison with results reported in the literature: in [Cas04], an execution time of $t = 9249s$ is reported for estimating a mixture of $k = 2$ distributions in dimension $d = 1$ using $n = 15000$ runs of a Gibbs sampler, while the entire sketching method (including computation of the sketch here, *without* parallel computing nor fast transforms) for estimating $k = 10$ components in dimension $d = 10$ on $n = 2 \cdot 10^5$ points with $m = 2000$ frequencies takes a mere $t = 80s$.

Elliptic stable distributions. Recall the parameterization of an elliptic stable distribution with diagonal dispersion matrix (Sec. 3.3.4), denoted by

$$\pi_{\boldsymbol{\theta}} = \mathcal{S}_\alpha \left(\boldsymbol{\mu}, \operatorname{diag}(\boldsymbol{\sigma}) \right) \tag{5.20}$$

with $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \alpha) \in \mathbb{R}^{2d+1}$, where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean of the distribution, $\boldsymbol{\sigma} \in \mathbb{R}^d_+$ is the diagonal of the dispersion matrix with coordinates $\sigma_i^2 > 0$, and $\alpha \in (0, 2]$ is the characteristic exponent of the distribution.

When $\alpha = 2$, the distribution is a Gaussian. When $\alpha < 2$, the distribution does not have a second order moment. The lower α is, the heavier the tail of the distribution is: in Fig. 5.17 we compare a random draw of $n = 10^4$ points from a mixture of $k = 3$ components in dimension $d = 10$ (used in Section 5.6.3 as a toy example), either with all characteristic exponents put to $\alpha_l = 2$ (*i.e.* it is a GMM) or with lower characteristic exponents. It is indeed seen that the data are much more spread in the second case. In particular, we outline the norm of the sample that is furthest from the origin $\max_i \|\mathbf{z}_i\|$. In the Gaussian case this maximal radius is of the order of 8.5, while in the stable distribution case this radius is approximately $2.9 \cdot 10^8$.

5.6.1 Implementation of CL-OMP(R)

Given randomly drawn frequencies $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m \in \mathbb{R}^d$, the feature function $\mathbf{f}$ is a sampling of the characteristic function of an elliptic stable distribution (3.41):

$$\mathbf{f}(\boldsymbol{\theta}) := \left[e^{\mathrm{i}\boldsymbol{\omega}_j^\top \boldsymbol{\mu}} \exp\left(- \left(\frac{1}{2} \sum_{i=1}^d \omega_{j,i}^2 \sigma_i^2 \right)^{\frac{\alpha}{2}} \right) \right]_{j=1}^m \tag{5.21}$$

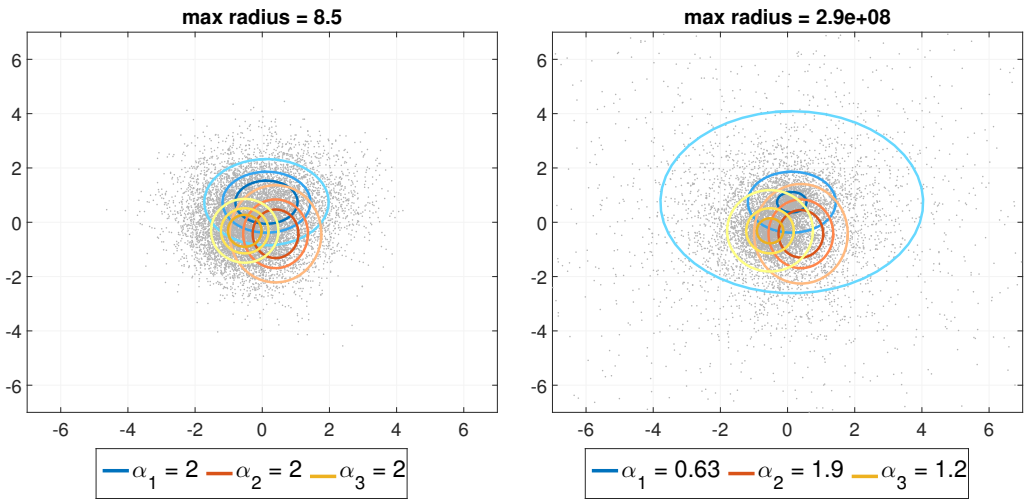

FIGURE 5.17: Comparison of data points drawn from a GMM or a mixture of elliptic stable distributions. The means are the same between the two mixtures, and the covariances of the GMM are the precision matrices of the mixture of stable distributions. In other words, the GMM can be considered as the same mixture of stable distributions with only the characteristic components put to $\alpha_l = 2$. The axes on the right have been clipped, since the furthest point has a norm in the hundreds of millions.

Similar to the two previous cases, we give convenient expressions to implement CL-OMPR in practice. The function $\mathbf{f}$ itself is

$$\mathbf{f}(\boldsymbol{\theta}) = \rho_{\text{im.}}\left(\mathbf{W}^\top \boldsymbol{\mu}\right) \odot \rho_{\text{re.}}\left(-\left(\frac{1}{2}\mathbf{W}_2^\top \boldsymbol{\sigma}\right)^{\odot(\alpha/2)}\right) \tag{5.22}$$

where $\mathbf{W}_2 = \mathbf{W} \odot \mathbf{W}$, $\mathbf{x}^{\odot\alpha} = [x_i^\alpha]$ is the element-by-element power operation, $\rho_{\text{im.}}(\cdot)$ is the point wise application of $x \mapsto e^{\mathrm{i}x}$ and $\rho_{\text{re.}}(\cdot)$ is the pointwise application of $x \mapsto e^x$.

The function $\mathbf{g}(\boldsymbol{\theta}, \mathbf{x})$ is expressed here:

$$\mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) = \begin{bmatrix} \mathbf{W}\Big(-\operatorname{Im}\left(\mathbf{f}(\boldsymbol{\theta})\right) \odot \operatorname{Re}\left(\mathbf{x}\right) + \operatorname{Re}\left(\mathbf{f}(\boldsymbol{\theta})\right) \odot \operatorname{Im}\left(\mathbf{x}\right)\Big) \\ -\frac{\alpha}{4}\mathbf{W}_2\Big[\left(\mathbf{s}^{\odot\left(\frac{\alpha}{2}-1\right)}\right) \odot \Big(\operatorname{Re}\left(\mathbf{f}(\boldsymbol{\theta})\right) \odot \operatorname{Re}\left(\mathbf{x}\right) + \operatorname{Im}\left(\mathbf{f}(\boldsymbol{\theta})\right) \odot \operatorname{Im}\left(\mathbf{x}\right)\Big)\Big] \\ -\frac{1}{2}\left(\log(\mathbf{s}) \odot \mathbf{s}^{\odot\alpha/2}\right)^\top \Big(\operatorname{Re}\left(\mathbf{f}(\boldsymbol{\theta})\right) \odot \operatorname{Re}\left(\mathbf{x}\right) + \operatorname{Im}\left(\mathbf{f}(\boldsymbol{\theta})\right) \odot \operatorname{Im}\left(\mathbf{x}\right)\Big) \end{bmatrix} \begin{matrix} \} \text{ size } d \\ \} \text{ size } d \\ \} \text{ size } 1 \end{matrix} \tag{5.23}$$

where $\mathbf{s} = \frac{1}{2}\mathbf{W}_2^\top \boldsymbol{\sigma} \in \mathbb{R}^m$. Hence, like in the GMM case one can hope to leverage fast transforms if both $\mathbf{W}$ and $\mathbf{W} \odot \mathbf{W}$ can be replaced with efficient structured matrices.

5.6.2 Setup

We mainly test our approach on synthetic data.

Generating the parameters. For tests on synthetic data, we generate the parameters of a mixture as such: we first draw the parameters of a k-GMM $(\xi_l, \boldsymbol{\mu}_l, \boldsymbol{\sigma}_l)_{l=1}^k$ in dimension d exactly as in Section 5.4.3, but with a parameter $c = 1.5$ (instead of $c = 1$) for the drawing of the means[6] $\boldsymbol{\mu}_l \sim \mathcal{N}\left(0, ck^{\frac{1}{d}}\mathbf{I}\right)$, then draw the parameters α_l uniformly in $[0.5, 2]$.

[6]Since we expect each component to be more heavy tailed, we spread their means a little more.

Drawing samples. Drawing a sample $\mathbf{z}$ from a multivariate elliptic stable distribution $\mathcal{S}_\alpha(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is not immediate, unlike Gaussians for which many routines exist. Following the method described in [Nol13], a sample $\mathbf{z}$ can be drawn by being defined as:

$$\mathbf{z} = \boldsymbol{\mu} + \sqrt{y/2}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{x} \tag{5.24}$$

where $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$, and y is drawn according to an univariate general $\alpha/2$-stable distribution with mean $\mu = 0$, so-called "skewness" parameter $\beta = 1$ (such that it is asymmetric and non-negative) and "scale" parameter $\gamma = 2\cos\left(\frac{\pi\alpha}{4}\right)^{2/\alpha}$ (se [Nol13] for details). In practice we draw y using the `stbl` toolbox [Vei12].

Computing the sketch. The sketch is computed as in Section 5.4.3, by first designing the frequency distribution $\Lambda = \Lambda^{(Ar)}_{\bar{\sigma}^2\mathbf{I}}$ with `EstimMeanSigma`, drawing m frequencies then computing the sketch with (5.3) and (5.5).

CL-OMPR. For these experiments we only test the CL-OMPR algorithm. During the execution we enforce $\sigma^2_{l,i} \geq 10^{-15}$ and $\alpha_l \geq 10^{-2}$ for numerical stability. For the initialization of the gradient descent of Step 1 the parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ are randomly drawn as in the GMM case (Section 5.4.3), and the parameter α is initialized at $\alpha = 1.5$.

Evaluation distance. In general the likelihood of stable distributions is not explicitely computable, it is therefore impossible to use a traditional distance measure like the KL-divergence as we did for GMMs. In the first toy example (next section), we directly look at the values of each recovered parameter and compare it to their true values, as is usually done in the literature [Cas04]. For the experiments of Section 5.6.4, we use an approximation of the MMD $\|\cdot\|_\kappa$ using its expression (1.28): given the true parameters $(\Theta^\star, \boldsymbol{\xi}^\star)$ and recovered parameters $(\widetilde{\Theta}, \widetilde{\boldsymbol{\xi}})$, we draw $m' = 2 \cdot 10^5$ frequencies $\boldsymbol{\omega}'_j$ *i.i.d.* from Λ (*distinct* from the first draw of m frequencies used to compute the sketch) and compute

$$\left\|\pi_{\Theta^\star,\boldsymbol{\xi}^\star} - \pi_{\widetilde{\Theta},\widetilde{\boldsymbol{\xi}}}\right\|_\kappa \approx \left(\frac{1}{m'}\sum_{j=1}^{m'}\left|\psi_{\pi_{\Theta^\star,\boldsymbol{\xi}^\star}}(\boldsymbol{\omega}'_j) - \psi_{\pi_{\widetilde{\Theta},\widetilde{\boldsymbol{\xi}}}}(\boldsymbol{\omega}'_j)\right|^2\right)^{\frac{1}{2}} \tag{5.25}$$

As usual results are averaged over 50 experiments.

5.6.3 Toy example: parameter recovery

The most powerful feature of our algorithm is arguably the ability to precisely estimate the different characteristic exponents α_l of each component. As said earlier there are not many evaluation distances available to measure the quality of the estimation, and furthermore when using a global distance between the true $\pi_{\Theta^\star,\boldsymbol{\xi}^\star}$ and the estimated $\pi_{\widetilde{\Theta},\widetilde{\boldsymbol{\xi}}}$ it does not necessarily guarantee that the parameters are precisely recovered.

Hence we test our algorithm against a toy example and evaluate the estimation error of each individual parameter, as is often done in the literature [Cas04]. For this experiment we generate a mixture of $k = 3$ components in dimension $d = 10$ illustrated in Fig. 5.17. In particular, the randomly drawn characteristic exponents are $\alpha_1 = 0.63$, $\alpha_2 = 1.9$ and $\alpha_3 = 1.2$ in this example. When evaluating the recovery of the parameters of each component, we solve the permutation indeterminacy between components by comparing the parameters of each true components $\boldsymbol{\theta}^\star_l$ with those of the recovered component $\widetilde{\boldsymbol{\theta}_p}$ that has the mean $\widetilde{\boldsymbol{\mu}_p}$ closest to $\boldsymbol{\mu}^\star_l$.

In Fig. 5.18 we examine the recovery of the parameters with respect to the relative sketch size $m/(kd)$. It is seen that for a sketch size of approximately $m = 10kd$ all parameters are precisely recovered, including the characteristic exponents α_l with an absolute precision of 10^{-2}, which is significantly more precise than other results reported in the literature [Cas04; SGKR09].

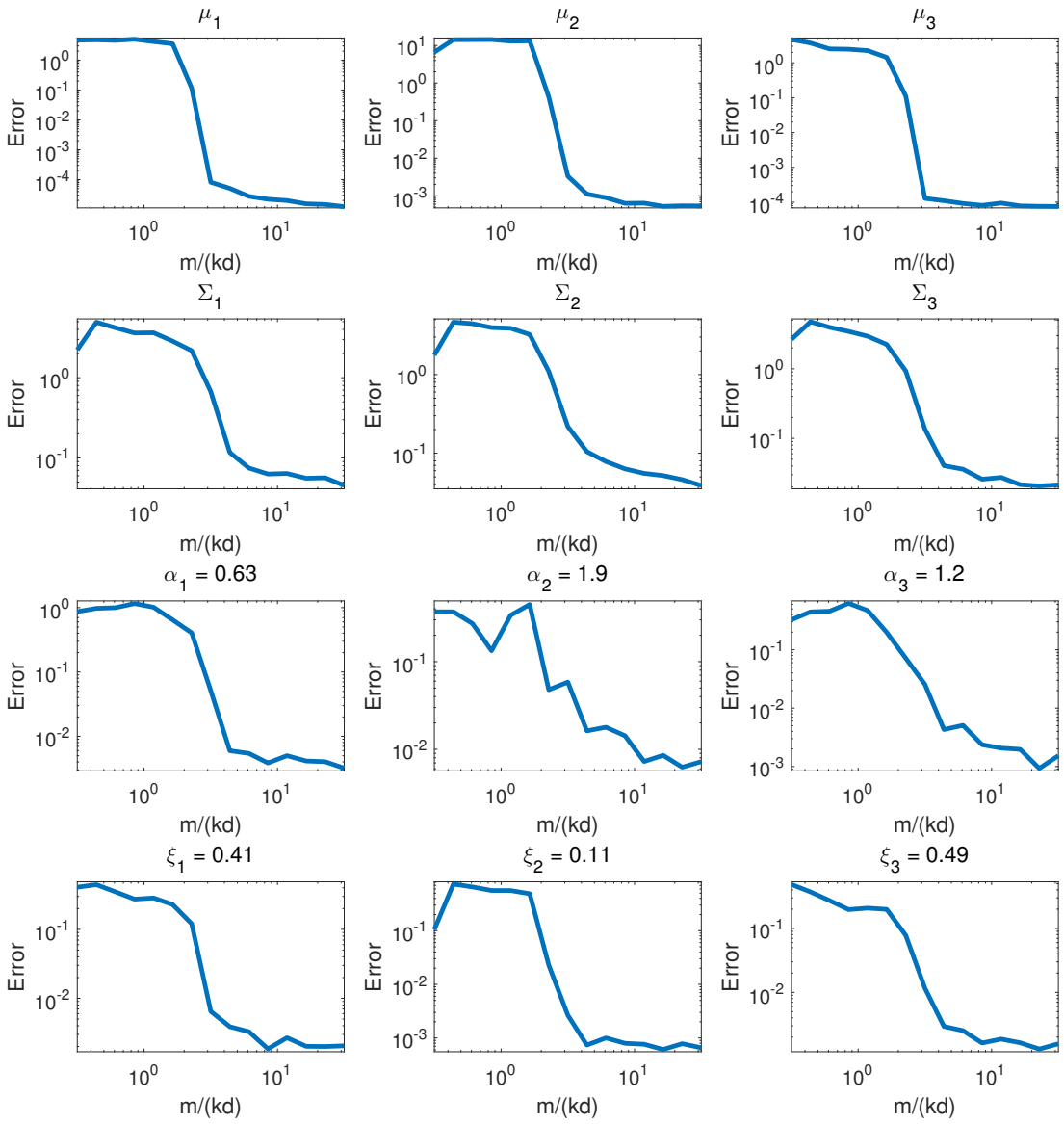

FIGURE 5.18: Estimation error of each individual parameter of a mixture of stable distributions with $k = 3$ components with respect to sketch size m, in dimension $d = 10$, on a database with $n = 5 \cdot 10^5$ vectors.

In Fig. 5.19 we fix the sketch size at $m = 15kd$ and report the precision of the recovery with respect to the number of items n in the database. As expected, as the database size increases the recovery becomes more and more precise.

5.6.4 Empirically sufficient sketch size

Like in the previous cases, we examine the sketch size required by the method. In Fig. 5.20 we draw transition diagrams, measuring our approximation of the MMD 5.25 with respect to the relative sketch size $m/(kd)$ for many number of components k and dimension d. As in the two previous applications with mixtures of Diracs and GMMs, it is seen that a sketch size $m \approx \mathcal{O}(kd)$ is sufficient for successful estimation.

Once again, the theoretical results obtained in Chapter 3 for mixtures of elliptic stable distributions indeed seem to be sub-optimal compared to what is observed in practice.

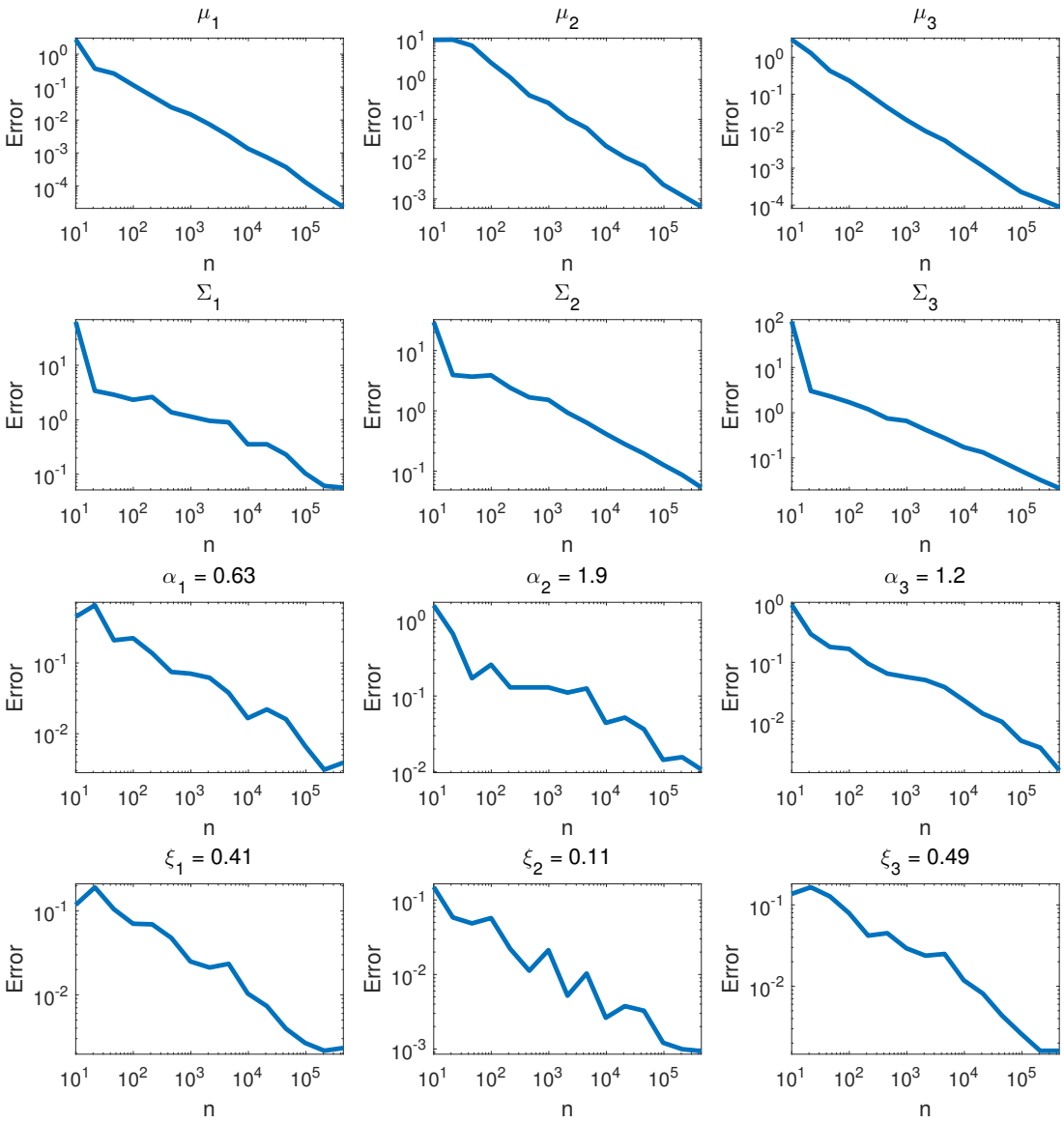

FIGURE 5.19: Estimation error of each individual parameter of a mixture of stable distributions with $k = 3$ components with respect to the size of the database n, in dimension $d = 10$, with a sketch size $m = 15kd$.

5.6.5 Comparison with GMMs

The model of mixtures of Gaussians with diagonal covariance is *strictly included* in the model of mixtures of elliptic stable distributions with diagonal precision matrix. Hence the optimal value of the cost function 5.4 is theoretically lower when the model over which it is minimized is the set of mixtures of stable distributions than when it is the set of GMMs.

Experiment. We verify this fact by performing an experiment in Fig. 5.21, on two types of data: either MFCCs coming from the speaker verification experiment performed in Section 5.4.6, or synthetic data drawn from a true mixture of stable distributions. For the speech data, the cost function is indeed slightly lower when minimized on mixtures of stable distributions than when minimized on the set of GMMs. The discrepancy is however very small, which shows than GMMs are already well-adapted to speech data. On synthetic data drawn from stable distributions however, the cost function is as expected far more minimized when using the adapted model of mixtures of stable distributions rather than GMMs.

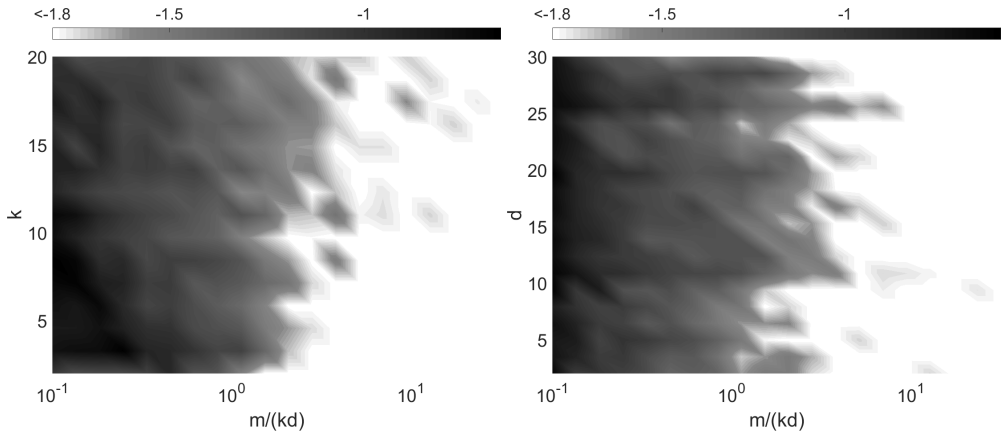

FIGURE 5.20: Log-MMD result for CL-OMPR for the estimation of mixtures of elliptic stable distributions, with respect to the relative sketch size $m/(kd)$. On the left, $d = 10$, on the right $k = 10$.

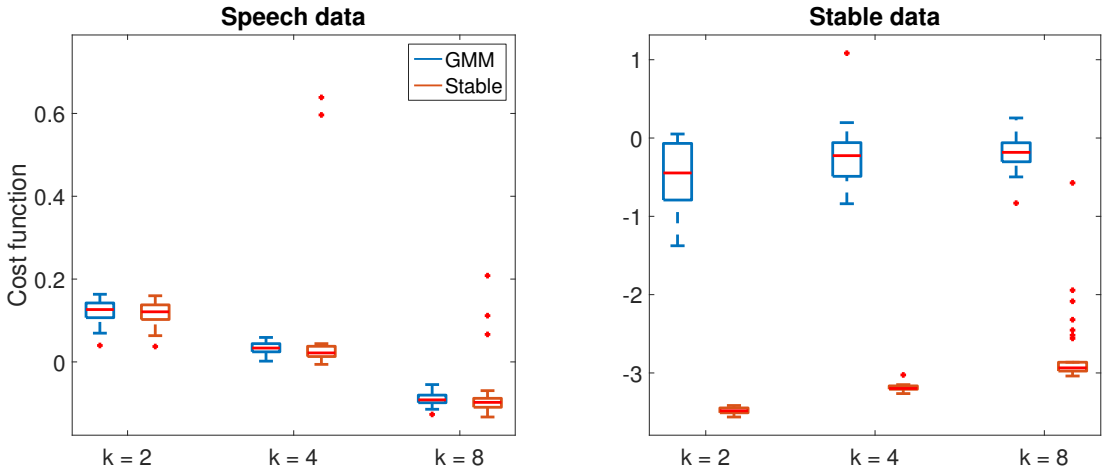

FIGURE 5.21: Comparison of the value of the cost function 5.4 after being minimized with CL-OMPR to recover either a GMM or a mixture of stable distributions, on $n = 2 \cdot 10^5$ MFCCs ($d = 12$) randomly selected from the NIST05 database used in the speaker verification experiment of Section 5.4.6, or synthetic data drawn from a mixture of stable distribution in dimension $d = 10$.

5.6.6 Conclusion on mixtures of stable distributions

In this application we have shown that CL-OMPR is capable of precisely recovering the parameters of a mixture of multivariate elliptic stable distributions, where each component has a different characteristic exponent. We performed experiments on synthetic data, for which CL-OMPR was observed to be orders of magnitude faster and more precise than results reported in the literature (for methods that are only available for the univariate case), by avoiding use of heavy tools such as MCMC simulations. A more extensive comparison with these methods on real data will be done in the future.

Mixtures of stable distributions have been very scarcely used in the past, due to the lack of estimators that were both efficient and able to handle multivariate data. The CL-OMPR algorithm may well be one of the first algorithm for this task, and we hope that it will open the way for a new class of methods that use mixtures of stable distributions in real life situations.

5.7 General conclusion

We started this chapter by describing a light unsupervised method to learn an appropriate frequency distribution for the sketching method, when the sketching operator is a random

sampling of the characteristic function of distributions. Although our underlying assumption was that data were drawn from a GMM, the proposed method also proved to be efficient for speech data and for data drawn from a mixture of elliptic stable distributions, while being faster than traditional supervised learning methods. Future work will aim at generalizing this strategy to other types of data (including the MNIST data for spectral clustering for which we have seen that choosing the frequency by hand was more effective).

Then the sketching method was applied to the compressive estimation of three mixture models.

First for mixtures of Diracs, it was shown to yield a clustering method that attains the precision of k-means while being more efficient on large databases. The CL-OMP(R) was shown to be more stable than k-means between runs and less sensitive to the initialization method.

The sketching method was then applied to the estimation of GMMs with diagonal covariances. A hierarchical algorithm faster than CL-OMPR for large number of components was defined, and observed to perform well on certain tasks such as speaker verification with MFCCs. Again, the CL-OMPR algorithm was seen to reach the precision of EM while being more efficient on large databases.

Finally, the sketching method was applied to the estimation of mixtures of multivariate elliptic stable distributions, for which no algorithm existed in the literature in the multivariate case. The proposed method was observed to be able to precisely estimate the parameters of a mixture, even for significantly different characteristic exponents α_l between components. As expected, the cost function reached a lower value when minimized on the set of mixtures of stable distributions rather than the set of GMMs, since the former includes the latter.

In each three cases an empirically sufficient sketch size in $m \approx \mathcal{O}(kd)$ was observed, which confirms that the preliminary theoretical results obtained in Chapter 3 for GMMs and mixtures of stable distributions were probably pessimistic.

Finally the proposed method was also briefly compared on a simple coreset approach, for k-means and GMM estimation. It was observed that, when the sketch size is insufficient (*i.e.* before the phase transition empirically observed in $\mathcal{O}(kd)$), a coreset of the same size outperforms the proposed method, however after the phase transition the proposed approach significantly outperforms coresets.

In the next chapter, we go back to the theoretical analysis of the method and provide guarantees for sufficient sketch size that only depends on k, d.

Chapter 6

Sketching and Statistical Learning

Context. In Chapter 3, we introduced generic conditions under which a sketching operator $\mathcal{A}$ satisfies the Lower Restricted Isometry Property (LRIP) with high probability, leading to guarantees of robust reconstruction of a low-complexity model $\mathfrak{S}$. We showed that the distance between sketches approximates the MMD defined by a kernel, through the use of Random Feature expansions. Two assumptions were at the core of our analysis:

i) an *admissibility condition* (Def. 3.2.1), that formulates a certain domination property between two norms, with a possible additional error $\eta \geq 0$;

ii) finiteness of the covering numbers of the non-uniform *normalized secant set* of the model, with a possible "extrusion" at level $\eta \geq 0$.

These hypotheses are in general difficult to prove. We nevertheless showed at the end of Chapter 3 that they can be satisfied for many models as soon as *the model itself* has finite covering numbers, and gave two examples. However, the recovery results obtained with this first approach exhibited two main problems:

– the precision of the recovery was expressed with respect to the MMD. Ideally, one would like to obtain guarantees of success for more traditional problems in machine learning;

– more importantly, in these results the error level were as $\eta = \mathcal{O}\left(1/\sqrt{m}\right)$, which yields a sufficient sketch size on par with that of the original database to attain a given error level (as we have seen this is largely pessimistic compared to what is observed in practice).

This chapter. In this chapter we address these problems. The outline is the following.

– In Section 6.1 we recall the main objective of statistical learning. Replacing the MMD by metrics linked to classic statistical learning tasks is the first goal of this chapter.

– In Section 6.2 we introduce a general analysis strategy for the estimation of mixture models. Given a certain *separation* assumption and some hypotheses on the kernel, we are able to provide results with no additional error $\eta = 0$.

– In Section 6.3 and 6.4, we apply this analysis strategy to two cases. First we show that recovering mixtures of Diracs can be done with a sketch size that only depends on k, d, and that guarantees can be obtained with respect to the traditional costs for k-means and k-medians. Second we show that recovery of Gaussian Mixture Models with fixed known covariance can also be done efficiently, and provide guarantees with respect to traditional log-likelihood.

This work is the result of a collaboration with Rémi Gribonval, Gilles Blanchard and Yann Traonmilin, where each author contributed equally[1]. It resulted in a paper [Gri+17] where we introduce a framework slightly more general than that described in this thesis.

[1]My own contributions include: the original framework of [Ker+17b] that introduces kernel mean embedding and random features for analyzing the sketching method, a simplification of the proof, and the application of the framework to GMMs.

6.1 Statistical learning

Statistical Learning consists in deriving a *hypothesis* $h \in \mathcal{H}$ from training data $\mathbf{z}_1, \dots, \mathbf{z}_n \in \mathfrak{Z}$. A *learning task* is a pair $(\ell, \mathcal{H})$ where $\mathcal{H}$ is a class of hypotheses and $\ell : \mathfrak{Z} \times \mathcal{H} \mapsto \mathbb{R}$ is a *loss function*. Denote

$$\mathcal{L}(\mathcal{H}) := \{\ell(\cdot, h) \mid h \in \mathcal{H}\} \tag{6.1}$$

the family of loss functions.

Expected risk minimization. Assuming the training samples $\mathbf{z}_1, \dots, \mathbf{z}_n$ are drawn *i.i.d.* from a distribution $\pi^\star$, to ensure good generalization property one usually aims at finding the hypothesis $h^\star$ that minimizes the *expected risk*:

$$\mathcal{R}_{\pi^\star}(h) := \mathbb{E}_{\mathbf{z}\sim\pi^\star}\ell(\mathbf{z}, h) \tag{6.2}$$

$$h^\star \in \operatorname*{arg\,min}_{h\in\mathcal{H}} \mathcal{R}_{\pi^\star}(h) \tag{6.3}$$

Empirical risk minimization. Of course, in practice one does not have access to the true risk function $\mathcal{R}_{\pi^\star}$, but only to the training data. Hence what is traditionally done is to simply replace the expectation by the empirical average over the training data, to obtain the *empirical risk*:

$$\mathcal{R}_{\hat{\pi}_n}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{z}_i, h) \tag{6.4}$$

Finding the hypothesis that minimizes the empirical risk is referred to as *Empirical Risk Minimization* (ERM) and is one of the foundation of statistical learning [Vap95]. ERM is naturally challenged when the number of elements in the database is large, and we hope to prove that the sketching method is more efficient than ERM in this context.

Sketching and risk minimization. Recall the sketching method introduced in Chapter 3. Given a model $\mathfrak{S}$ and a RF expansion $(\mathcal{F}_R, \Lambda)$, it consists in the following:

1. Draw m frequencies $\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_m \overset{i.i.d.}{\sim} \Lambda$, define the sketching function $\Phi(\mathbf{z}) := \frac{1}{\sqrt{m}}\left[\phi_{\boldsymbol{\omega}_j}(\mathbf{z})\right]_{j=1}^{m}$ and the sketching operator $\mathcal{A}\mu := \langle \mu, \Phi \rangle$;
2. Given a database $\mathbf{z}_1, \dots, \mathbf{z}_n \overset{i.i.d.}{\sim} \pi^\star$ compute the sketch
$$\hat{\mathbf{y}} = \mathcal{A}\hat{\pi}_n = \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{z}_i)$$
3. Recover a probability distribution in the model from the sketch with:
$$\widetilde{\pi} := \Delta_\iota(\mathcal{A}, \hat{\mathbf{y}}) = \operatorname*{arg\,min}_{\pi\in\mathfrak{S},\ \text{error } \iota} \|\mathcal{A}\pi - \hat{\mathbf{y}}\|_2$$

In Chapter 3 we showed that, under hypotheses, with high probability the decoder Δ_ι is instance-optimal and the recovered distribution $\widetilde{\pi}$ is close to the true distribution $\pi^\star$ with respect to some metric $d_{\mathcal{L}}$. Can we leverage this kind of result in the context of statistical learning?

For that we define the notion of *acceptable* models $\mathfrak{S}$.

Definition 6.1.1 (Acceptability)**.** *The model $\mathfrak{S}$ is **acceptable** for the learning task $(\ell, \mathcal{H})$ if for all $\pi_{\mathfrak{S}} \in \mathfrak{S}$ one has a closed-form expression to derive:*

$$h_{\mathfrak{S}} \in \operatorname*{arg\,min}_{h\in\mathcal{H}} \mathcal{R}_{\pi_{\mathfrak{S}}}(h)\ . \tag{6.5}$$

Remark 6.1.2. *Acceptability is not a precise mathematical notion. It expresses the fact that, for the sketching method to be computationally efficient, it should be easier to recover a hypothesis from a distribution* ***in the model*** *(itself estimated using sketches), than from the original empirical distribution* $\hat{\pi}_n$.

A simple but frequent case is when the recovered probability distribution is directly parameterized *by the hypothesis that satisfies* (6.5)*, which will be the case in the two examples developed afterward in this chapter. In these examples, the model will be a parametric set of distributions* $\mathfrak{S} = \{\pi_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathcal{T}\}$*, the class of hypotheses* h *a set of parameters* $\boldsymbol{\theta}$*, and the loss function such that: if the recovered distribution is denoted by* $\pi_{\widehat{\mathfrak{S}}} = \pi_{\widehat{\boldsymbol{\theta}}}$*, equation* (6.5) *yields* $h_{\widehat{\mathfrak{S}}} = \widehat{\boldsymbol{\theta}}$*. In that case acceptability is immediate.*

From sketch recovery to excess risk control. Now, recall the notation

$$\|\mu\|_{\mathcal{L}(\mathcal{H})} = \sup_{h \in \mathcal{H}} |\langle \mu, \ell(\cdot, h) \rangle| \tag{6.6}$$

where $\mathcal{L}(\mathcal{H})$ is defined by (6.1). It is a case of Integral Probability Metric as we defined in (3.4).

Given a distribution $\widetilde{\pi} = \Delta(\mathcal{A}, \hat{\mathbf{y}}) \in \mathfrak{S}$ estimated from the sketch, *using acceptability of the model we can compute*

$$\widetilde{h} = \arg\min_{h \in \mathcal{H}} \mathcal{R}_{\widetilde{\pi}}(h) \; . \tag{6.7}$$

Then, if we are able to control

$$\|\pi^\star - \widetilde{\pi}\|_{\mathcal{L}(\mathcal{H})} = \sup_{h \in \mathcal{H}} |\mathcal{R}_{\pi^\star}(h) - \mathcal{R}_{\widetilde{\pi}}(h)| \leq \varphi \; , \tag{6.8}$$

for a small φ, we can control the *excess risk*:

$$\begin{aligned} \mathcal{R}_{\pi^\star}\left(\widetilde{h}\right) - \mathcal{R}_{\pi^\star}\left(h^\star\right) &= \mathcal{R}_{\pi^\star}\left(\widetilde{h}\right) - \mathcal{R}_{\widetilde{\pi}}\left(\widetilde{h}\right) + \mathcal{R}_{\widetilde{\pi}}\left(\widetilde{h}\right) - \mathcal{R}_{\pi^\star}\left(h^\star\right) \\ &\overset{(6.7)}{\leq} \mathcal{R}_{\pi^\star}\left(\widetilde{h}\right) - \mathcal{R}_{\widetilde{\pi}}\left(\widetilde{h}\right) + \mathcal{R}_{\widetilde{\pi}}\left(h^\star\right) - \mathcal{R}_{\pi^\star}\left(h^\star\right) \\ &\leq 2\left\|\pi^\star - \widetilde{\pi}\right\|_{\mathcal{L}(\mathcal{H})} \overset{(6.8)}{\leq} 2\varphi \end{aligned} \tag{6.9}$$

which is the guarantee we want to obtain. Therefore:

> We aim at providing information-preservation guarantees (Theorem 3.2.7) with the norm $d_{\mathcal{L}} := \|\cdot\|_{\mathcal{L}(\mathcal{H})}$, to obtain a bound of the type (6.8).

For that, we need to replace the MMD $\|\cdot\|_\kappa$ with the metric $d_{\mathcal{L}} = \|\cdot\|_{\mathcal{L}(\mathcal{H})}$ in the LRIP (*i.e.* in hypothesis iii of Theorem 3.2.5). We formalize this property with the notion of *compatible* kernel.

Definition 6.1.3 (Compatibility)**.** *A kernel* κ *is* ***compatible*** *for the learning task* $(\ell, \mathcal{H})$ *and model* $\mathfrak{S}$ *with compatibility constant* $W_{\mathcal{L}} > 0$ *if for all* $\pi_{\mathfrak{S}}, \pi'_{\mathfrak{S}} \in \mathfrak{S}$ *we have*

$$\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_{\mathcal{L}(\mathcal{H})} \leq W_{\mathcal{L}} \|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_\kappa \; . \tag{6.10}$$

One can see that this condition is somewhat similar to the admissibility condition (Def. 3.2.1). It is essentially the same supposition that

$$\|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_{\mathcal{F}} \lesssim \|\pi_{\mathfrak{S}} - \pi'_{\mathfrak{S}}\|_\kappa$$

with $\mathcal{F} = \mathcal{F}_R$ for the admissibility condition and $\mathcal{F} = \mathcal{L}(\mathcal{H})$ for the compatibility condition. As we will see, the strategy developed in this chapter is to derive a generic analysis to treat both at once.

Summary. Hence, given a learning task, our analysis strategy lies on the definition of the following objects and their respective properties:

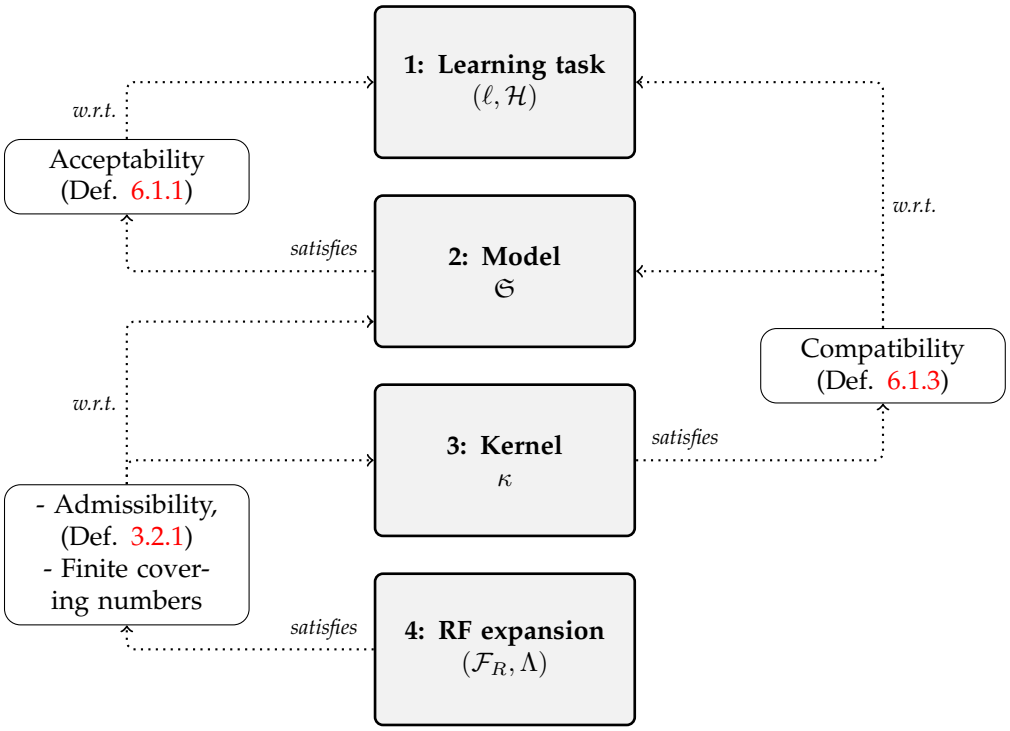

FIGURE 6.1: **Illustration of our approach for sketched statistical learning.** Our analysis lies on the definition of four objects that satisfies different relations between them: a learning task $(\mathcal{H}, \ell)$, a model set $\mathfrak{S}$ that is acceptable for the learning task, a kernel that is compatible with the model and the learning task, and an RF expansion of this kernel that is admissible for the model and such that the covering numbers of the normalized secant sets are finite.

1. a model $\mathfrak{S}$ **acceptable** for the learning task;
   - In the two examples developed at the end of the chapter, it arises naturally from the definition of the problem.
2. a kernel κ **compatible** with the model and the learning task;
   - This is similar to admissibility but with an IPM $\|\cdot\|_{\mathcal{F}} = \|\cdot\|_{\mathcal{L}(\mathcal{H})}$ that corresponds to the learning task, and somewhat of equal difficulty to prove.
3. a RF expansion of the kernel $(\mathcal{F}_R, \Lambda)$ **admissible** for the model, such that the normalized secant sets have **finite covering numbers**.
   - This is often the most substantial part of the proof.

We schematically illustrate the relations between these objects in Fig. 6.1.

Rest of the chapter. In this chapter, we are going to examine the k-means (and k-medians) problem by recovering mixtures of Diracs, as implemented in Chapter 5, and the problem of estimating GMMs with known covariance, which was the original empirical framework of Bourrier et al. [BGP13]. We will obtain results with no additive error in the LRIP ($\eta = 0$). In Chapter 5, we have seen that the sketching method is also experimentally efficient with richer models, however the analysis presented in the present chapter does not seem to apply, and we have not yet been able to improve the results obtained at the end of Chapter 3.

To treat the k-means and GMM problems, we will develop a general analysis for mixture models, described in the next section.

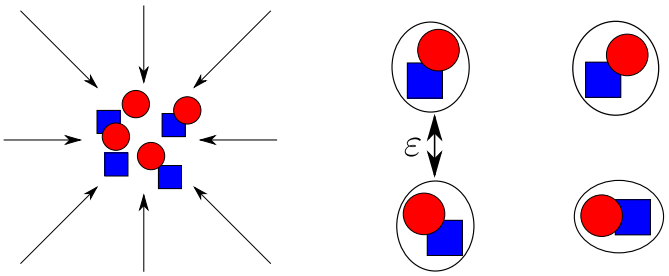

FIGURE 6.2: Illustration of our strategy with separation. Red circles represent a first mixture, blues squares a second one. Suppose the norm between the two mixtures goes to zero. Without separation hypothesis (left) the situation is difficult to describe accurately. With the assumption that all circles are pairwise separated and all squares are pairwise separated, each circle is close to at most one square and separated from all other circles and squares. We witness the apparition of dipoles pairwise separated, which is the basis for our analysis.

6.2 Generic analysis of Mixture Models

In this section, we develop a generic analysis of the problem when the model is that of mixtures of parametric distributions in $\mathfrak{T} = \{\pi_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathcal{T}\}$, denoted by

$$\mathfrak{S}_k(\mathfrak{T}) := \left\{ \sum_{l=1}^{k} \xi_l \pi_{\boldsymbol{\theta}_l} \;\middle|\; \boldsymbol{\theta}_l \in \mathcal{T}, \xi_l \geq 0, \sum_{l=1}^{k} \xi_l = 1 \right\}$$

Acceptability of mixture models will be proven on a case-by-case basis, we focus on the other hypotheses: finding a compatible kernel and an admissible RF expansion such that the normalized secant sets have finite covering numbers.

Analysis strategy: ε-separation. Noting the similarity between admissibility $\|\pi - \pi'\|_{\mathcal{F}_R} \lesssim \|\pi - \pi'\|_\kappa$ (Def. 3.2.1) and compatibility $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \lesssim \|\pi - \pi'\|_\kappa$ (Def. 6.1.3), we treat them with the same set of mathematical tools. We have seen that all hypotheses, including finiteness of the covering numbers of the normalized secant sets, can be handled when $\|\pi - \pi'\|_\kappa \geq \eta > 0$. Intuitively, the main technical difficulty is therefore to analyze the behavior of $\|\pi - \pi'\|_\kappa$ when $\pi, \pi' \in \mathfrak{S}$ get "close" to each other: are the other norms $\|\cdot\|$ sufficiently regular so that $\|\pi - \pi'\| / \|\pi - \pi'\|_\kappa$ is bounded when $\|\pi - \pi'\|_\kappa \to 0$?

Proving these hypotheses is often feasible when $k = 1$ and the model $\mathfrak{S} := \mathfrak{S}_1(\mathfrak{T}) = \mathfrak{T}$ is formed by simple single distributions (Gaussians, Diracs...), by obtaining closed-form expressions for all norms with respect to the parameters $\boldsymbol{\theta}, \boldsymbol{\theta}'$. However, it becomes much more complicated when the model is that of mixtures of distributions with $k > 1$.

Our strategy to treat mixtures of distributions is to impose that all the components of a mixture are *pairwise sufficiently separated*. By doing so, we can ensure that each component $\pi_{\boldsymbol{\theta}_l}$ of π is close to *at most one* component $\pi_{\boldsymbol{\theta}'_p}$ of π', and sufficiently separated from all other components. This is illustrated in Fig. 6.2, where the components of π are represented by red circles and these of π' by blue squares. In the left schematic picture, no hypothesis is made, and when the two mixtures get close to each other it is difficult to control the behavior of each of their individual components. In the right picture, the red circles are pairwise separated from each other, and so are the blue squares. Therefore, when the mixtures get close to each other each red circle is paired with one blue square, and these pairs are separated from each other.

A pair of two components close to each other will be called a *dipole*. We can then make sure that the kernel κ has the ability to "distinguish" dipoles that are sufficiently separated, and obtain the desired results.

All proofs are in Appendix C.

6.2.1 Framework

Consider $\mathfrak{T} = \{\pi_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathcal{T}\} \subset \mathfrak{P}$ a parametric family of probability distributions, and assume that the parameter space $\mathcal{T}$ is equipped with some metric

$$\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}') := \|\chi(\boldsymbol{\theta}) - \chi(\boldsymbol{\theta}')\|_2 \tag{6.11}$$

where $\chi(\cdot)$ denotes an (implicit) feature mapping. A pair of distributions whose parameters are close with respect to this metric is called a dipole.

Definition 6.2.1 (Dipoles, separation of dipoles)**.** *A measure μ is called a **dipole** with respect to the metric ϱ if it can be decomposed as $\mu = \xi_1 \pi_{\boldsymbol{\theta}_1} - \xi_2 \pi_{\boldsymbol{\theta}_2}$ where $\varrho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \leq 1$ and $0 \leq \xi_i \leq 1$. Note that the ξ_i's are not normalized to 1 here, and that they can be put to 0 to characterize single distributions.*

*Two dipoles $\mu = \xi_1 \pi_{\boldsymbol{\theta}_1} - \xi_2 \pi_{\boldsymbol{\theta}_2}$ and $\mu' = \xi'_1 \pi_{\boldsymbol{\theta}'_1} - \xi'_2 \pi_{\boldsymbol{\theta}'_2}$ are said 1-**separated** if $\varrho(\boldsymbol{\theta}_i, \boldsymbol{\theta}'_j) \geq 1$ for all $i, j \in \{1, 2\}$.*

The model of ε-separated mixtures $\mathfrak{S}_{k,\varepsilon,\varrho}(\mathfrak{T})$ is defined as

$$\mathfrak{S}_{k,\varepsilon,\varrho}(\mathfrak{T}) := \left\{ \sum_{l=1}^{k} \xi_l \pi_{\boldsymbol{\theta}_l} \;\middle|\; \boldsymbol{\xi} \in \mathbb{S}^{k-1},\ \pi_{\boldsymbol{\theta}_l} \in \mathfrak{T},\ \varrho(\boldsymbol{\theta}_l, \boldsymbol{\theta}_p) \geq \varepsilon\ \forall l \neq p \right\} \tag{6.12}$$

Depending on the context, it can be denoted $\mathfrak{S}_{k,\varepsilon}(\mathfrak{T})$ or even $\mathfrak{S}_{k,\varepsilon}$ when there is no ambiguity.

Remark 6.2.2. *In the following, we will impose a constant 2-separation and work with $\mathfrak{S}_{k,2,\varrho}(\mathfrak{T})$. Then it is easy to note that by defining $\widetilde{\varrho} = \varepsilon\varrho/2$ we have $\mathfrak{S}_{k,2,\varrho}(\mathfrak{T}) = \mathfrak{S}_{k,\varepsilon,\widetilde{\varrho}}(\mathfrak{T})$, with adjustable ε-separation of components.*

In this chapter we consider real-valued kernels κ such that the mean kernel is expressed as[2]:

$$\kappa(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) = K(\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}')), \quad \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{T} \tag{6.13}$$

where $K : \mathbb{R}_+ \to \mathbb{R}_+$.

We will see that the kernel will be able to "distinguish" separated dipoles if K decays sufficiently fast. More precisely, we will impose that the function K belongs to the following family of functions.

Definition 6.2.3. *The class $\mathcal{E}(A, B, C, \gamma)$ are functions $K : \mathbb{R}_+ \to \mathbb{R}_+$ that satisfy*

i) over the interval $[0, 1]$:

- *$K(0) = 1$;*
- *$K(u) \leq 1 - \frac{\gamma u^2}{2}$ for all $u \leq 1$;*

ii) over the interval $[1, \infty)$:

- *K is bounded: $0 \leq K(u) \leq A$, for all $u \geq 1$;*
- *K is differentiable with bounded derivative: $|K'(u)| \leq B$, for all $u \geq 1$;*
- *K' is C-Lipschitz: $|K'(u) - K'(v)| \leq C\,|u - v|$, for all $u, v \geq 1$.*

6.2.2 Incoherence of dipoles

We first show that for the right choice of kernel, the *coherence* between any two 1-separated dipoles is bounded for the inner product that corresponds to the kernel κ, in the sense that their product is small compared to the product of their norms.

[2] Such kernels may not always exist!

Lemma 6.2.4. *Consider a kernel $\kappa(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) = K(\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}'))$, with a pseudo metric ϱ of the form (6.11) and a function $K \in \mathcal{E}(A, B, C, \gamma)$.*

Then, for two dipoles that are 1-separated μ, μ', we have

$$|\kappa(\mu, \mu')| \leq \frac{8 \max(A, 2(B+C))}{\gamma} \|\mu\|_\kappa \|\mu'\|_\kappa \ . \tag{6.14}$$

Such an *incoherence* property is a classic tool in Compressive Sensing (see [FR13] Chap. 5), although it is expressed here under an original form between *dipoles* instead of vectors, and with the inner product materialized by the kernel κ.

It naturally leads to the following Theorem which is somehow interpreted as similar to Pythagore's theorem, in the sense that the sum of the squared norm of separated dipoles is upper bounded by the squared norm of their sum.

Theorem 6.2.5. *Consider a kernel $\kappa(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) = K(\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}'))$, with a pseudo metric ϱ of the form (6.11) and a function $K \in \mathcal{E}(A, B, C, \gamma)$. Consider k such that*

$$1 \leq k \leq \frac{3\gamma}{32 \max(A, 2(B+C))} \tag{6.15}$$

For k dipoles μ_i that are pairwise 1-separated, the following holds:

$$\sum_{i=1}^k \|\mu_i\|_\kappa^2 \leq 4 \left\| \sum_{i=1}^k \mu_i \right\|_\kappa^2 . \tag{6.16}$$

Proof. We have

$$\begin{aligned}
\left| \left\| \sum_{i=1}^k \mu_i \right\|_\kappa^2 - \sum_{i=1}^k \|\mu_i\|_\kappa^2 \right| &\leq \sum_i \sum_{j,\, j \neq i} |\kappa(\mu_i, \mu_j)| \leq \frac{8 \max(A, 2(B+C))}{\gamma} \sum_i \sum_{j,\, j \neq i} \|\mu_i\|_\kappa \|\mu_j\|_\kappa \\
&\leq \frac{8 \max(A, 2(B+C))}{\gamma} \left(\sum_{i=1}^k \|\mu_i\|_\kappa \right) \sqrt{k} \sqrt{\sum_{j=1}^k \|\mu_j\|_\kappa^2} \\
&\leq \frac{3}{4} \sum_{i=1}^k \|\mu_i\|_\kappa^2
\end{aligned}$$

Therefore $\left| \frac{\left\| \sum_{i=1}^k \mu_i \right\|_\kappa^2}{\sum_{i=1}^k \|\mu_i\|_\kappa^2} - 1 \right| \leq 3/4$, which proves the desired property. □

As we will see, Theorem 6.2.5 is one of the main tool that will allow us to handle sums of separated dipoles.

6.2.3 Admissibility, compatibility

We now turn to proving admissibility and compatibility, which both can be summarized as: for a certain class of functions $\mathcal{Q}$, a model $\mathfrak{S}$ and a kernel κ, there exists a constant $W > 0$ such that for all all $\pi, \pi' \in \mathfrak{S}$:

$$\|\pi - \pi'\|_{\mathcal{Q}} \leq W \|\pi - \pi'\|_\kappa \tag{6.17}$$

We will treat both properties in the same fashion, by proving this property for the family of feature functions $\mathcal{Q} = \mathcal{F}_R$, and the family of loss functions $\mathcal{Q} = \mathcal{L}(\mathcal{H})$. For that we define the following class of functions.

Definition 6.2.6 ("Bounded and Lipschitz in expectation" functions). *A function $f : \mathfrak{Z} \to \mathbb{C}$ is "bounded and Lipschitz (with respect to the metric $\varrho(\cdot,\cdot)$) in expectation" on the basic set $\mathfrak{T} = \{\pi_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \mathcal{T}\}$ if there exists $D, L < \infty$ such that for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{T}$,*

$$|\langle \pi_{\boldsymbol{\theta}}, f\rangle| \leq D, \tag{6.18}$$

$$|\langle \pi_{\boldsymbol{\theta}}, f\rangle - \langle \pi_{\boldsymbol{\theta}'}, f\rangle| \leq L\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}'). \tag{6.19}$$

We denote by $\mathcal{Q}(D, L, \mathcal{T}, \varrho)$ (or in short $\mathcal{Q}(D, L)$) the set of all functions satisfying (6.18)-(6.19).

Let us begin by treating one dipole and prove the desired property.

Lemma 6.2.7. *Consider a kernel $\kappa(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) = K(\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}'))$, with a pseudo metric ϱ of the form (6.11) and a function $K \in \mathcal{E}(A, B, C, \gamma)$.*

For all dipoles μ, we have

$$\|\mu\|_{\mathcal{Q}(D,L)} \leq W_0 \|\mu\|_\kappa \tag{6.20}$$

with $W_0 = \left(L^2/\gamma + 2D^2\right)^{\frac{1}{2}}$.

Using incoherence between separated dipoles, we can then go from one dipole to proving the result for differences of mixtures with separation assumption (which is a sum of separated dipoles, see Fig. 1.3), which is the desired property for admissibility and compatibility.

Theorem 6.2.8. *Consider a kernel $\kappa(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) = K(\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}'))$, with a pseudo metric ϱ of the form (6.11) and a function $K \in \mathcal{E}(A, B, C, \gamma)$. Consider k such that*

$$1 \leq k \leq \frac{3\gamma}{64 \max(A, 2(B + C))} \tag{6.21}$$

Then, for all $\pi,\ \pi' \in \mathfrak{S}_{k,2,\varrho}(\mathfrak{T})$ the following holds:

$$\|\pi - \pi'\|_{\mathcal{Q}(D,L)} \leq 2W_0\sqrt{2k} \|\pi - \pi'\|_\kappa \tag{6.22}$$

with $W_0 = \left(L^2/\gamma + 2D^2\right)^{\frac{1}{2}}$.

Proof. Since π and π' are mixtures of distributions that are 2-separated, we can decompose:

$$\pi - \pi' = \sum_{l=1}^{2k} \mu_l$$

where μ_l are dipoles that are pairwise 1-separated, with some dipoles that are single distributions, and some that are zero.

Then we get we get

$$\begin{aligned}
\|\pi - \pi'\|_{\mathcal{Q}(D,L)} = \left\|\sum_{l=1}^{2k} \mu_l\right\|_{\mathcal{Q}(D,L)} &\leq \sum_{l=1}^{2k} \|\mu_l\|_{\mathcal{Q}(D,L)} \\
&\overset{\text{Lem. 6.2.7}}{\leq} W_0 \sum_{l=1}^{2k} \|\mu_l\|_\kappa \leq W_0\sqrt{2k}\sqrt{\sum_{l=1}^{2k} \|\mu_l\|_\kappa^2} \\
&\overset{\text{Thm. 6.2.5}}{\leq} 2W_0\sqrt{2k}\left\|\sum_{l=1}^{2k} \mu_l\right\|_\kappa = 2W_0\sqrt{2k}\|\pi - \pi'\|_\kappa,
\end{aligned}$$

which is the desired result. □

To summarize, assuming that the kernel has the right form and the model is that of mixtures that are sufficiently separated: if the feature functions in the set $\mathcal{F}_R$ are bounded and

Lipschitz "in expectation" (respectively, if the loss functions in the set $\mathcal{L}$ are bounded and Lipschitz in expectation), then the admissibility condition is satisfied with no error $\eta = 0$ and an admissibility constant that only depends on the smoothness of the features in $\mathcal{F}_R$ and the number of components k (respectively, the compatibility condition is satisfied with a constant expressed in a similar way).

6.2.4 Covering numbers of the secant set

Let us now prove that, under some additional hypotheses, when the model $\mathfrak{S}$ is included in the set of 2-separated mixtures, for any $\pi \in \mathfrak{S}$ the covering numbers of the normalized secant set $\mathcal{S}^0(\pi, \mathfrak{S})$ are finite.

For $\xi \in [0,1]$ and $\pi_{\boldsymbol{\theta}} \in \mathfrak{T}$, denote the non-uniform set of normalized dipoles:

$$\mathcal{D}(\xi\pi_{\boldsymbol{\theta}}) := \left\{ \frac{\xi\pi_{\boldsymbol{\theta}} - \xi'\pi_{\boldsymbol{\theta}'}}{\|\xi\pi_{\boldsymbol{\theta}} - \xi'\pi_{\boldsymbol{\theta}'}\|_\kappa} \;\middle|\; \xi' \in [0;1],\ \boldsymbol{\theta}' \in \mathcal{T},\ \varrho(\boldsymbol{\theta}, \boldsymbol{\theta}') \leq 1,\ \|\xi\pi_{\boldsymbol{\theta}} - \xi'\pi_{\boldsymbol{\theta}'}\|_\kappa > 0 \right\} \tag{6.23}$$

Then, again using incoherence between dipoles, elements of the normalized secant set can be decomposed into a sum of normalized dipoles.

Lemma 6.2.9. *Consider a kernel $\kappa(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) = K(\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}'))$, with a pseudo metric ϱ of the form (6.11) and a function $K \in \mathcal{E}(A, B, C, \gamma)$. Consider k such that (6.21) is satisfied. Define a model $\mathfrak{S} \subset \mathfrak{S}_{k,2,\varrho}(\mathfrak{T})$. Let $(\mathcal{F}_R, \Lambda)$ be a RF expansion of the kernel κ such that $\mathcal{F}_R \subset \mathcal{Q}(D, L)$.*

Consider $\pi = \sum_{l=1}^k \xi_l \pi_{\boldsymbol{\theta}_l} \in \mathfrak{S}$. Then we have

$$\begin{aligned} \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{S}^0(\pi, \mathfrak{S}), \delta\right) \leq \max\left(\left(\frac{16W_0 r}{\delta}\right)^{2k}, 1\right) \mathcal{N}^k\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{T}, \frac{\delta}{4r}\right) \\ \prod_{l=1}^k \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{D}(\xi_l \pi_{\boldsymbol{\theta}_l}), \frac{\delta}{4r}\right) \end{aligned} \tag{6.24}$$

where $W_0 = (L^2/\gamma + 2D^2)^{\frac{1}{2}}$ and $r = 2\sqrt{2k}$.

As we have seen it is often relatively immediate to prove that the set of basic distributions $\mathfrak{T}$ has finite covering numbers, therefore the only thing left to prove is that the normalized sets of dipoles $\mathcal{D}(\cdot)$ have finite covering numbers. We can further decompose them into an "extruded" part and a tangent part, the latter is then approached by a set of tempered distributions with controllable covering numbers.

Lemma 6.2.10. *Consider a kernel $\kappa(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) = K(\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}'))$, with a pseudo metric ϱ of the form (6.11) and a function $K \in \mathcal{E}(A, B, C, \gamma)$. Let $(\mathcal{F}_R, \Lambda)$ be a RF expansion of the kernel κ such that $\mathcal{F}_R \subset \mathcal{Q}(D, L)$.*

Consider $\pi_{\boldsymbol{\theta}} \in \mathfrak{T}$. Assume that there exists a set $\mathcal{V}_{\boldsymbol{\theta}}$ of tempered distributions(we assume that the random features are smooth) and two constants $M > 0$, $\eta_{\max} > 0$ such that for all $0 < \eta \leq \eta_{\max}$ and all non-zero dipoles μ of the form $\mu = \pi_{\boldsymbol{\theta}} - a\pi_{\boldsymbol{\theta}'}$ or $\mu = a\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}$ with $a \in [0;1]$ and $\boldsymbol{\theta}' \in \mathcal{T}$ (such that $\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}') \leq 1$ since μ is a dipole) and $\|\mu\|_\kappa \leq \eta$, there exists $\nu \in \mathcal{V}_{\boldsymbol{\theta}}$ such that

$$\left\| \frac{\mu}{\|\mu\|_\kappa} - \nu \right\|_{\mathcal{F}_R} \leq M\eta. \tag{6.25}$$

Then for any weight $\xi \in [0;1]$ and all $\delta \leq 4M\eta_{\max}$

$$\begin{aligned} \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{D}(\xi\pi_{\boldsymbol{\theta}}), \delta\right) \leq \max\left(\frac{96MW_0D}{\delta^2}, \frac{3}{2}\right) \cdot \mathcal{N}\left(\varrho, \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1), \frac{\delta^2}{64MW_0L}\right) \\ + \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{V}_{\boldsymbol{\theta}}, \frac{\delta}{4}\right) \end{aligned} \tag{6.26}$$

6.2.5 Choice of kernel

Before summarizing our findings and closing this section, we give the expression of a kernel that we are going to use in the instantiation of the analysis. The reasoning is here backward: we fix a given k, and design a function $K \in \mathcal{E}(A, B, C, \gamma)$ such that (6.21) is satisfied. We base our results on the Gaussian kernel for simplicity.

Lemma 6.2.11. *Define*

$$\sigma_k^2 = \frac{1}{2.4(\ln k + 10)} \tag{6.27}$$

Then $K : x \mapsto e^{-\frac{x^2}{2\sigma_k^2}} \in \mathcal{E}(A, B, C, \gamma = 1)$ *with* A, B, C *such that* (6.21) *is satisfied.*

6.2.6 Summary: Main result

Let us now combine this analysis with the results from Chapter 3. The following theorem is the main result of the chapter.

Theorem 6.2.12. *Consider a learning task $(\ell, \mathcal{H})$, a set of basic distributions $\mathfrak{T} = \{\pi_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathcal{T}\}$, a kernel $\kappa(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) = K(\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}'))$, with a pseudo metric ϱ of the form (6.11), with a random feature expansion $(\mathcal{F}_R, \Lambda)$. Consider $k > 1$ and define the model as $\mathfrak{S} \subset \mathfrak{S}_{k,2,\varrho}(\mathfrak{T})$.*

Assume that:

i) the model $\mathfrak{S}$ is acceptable for the learning task;

ii) the kernel is such that $K \in \mathcal{E}(A, B, C, \gamma)$ such that (6.21) is satisfied;

iii) the family of loss functions $\mathcal{L}(\mathcal{H}) = \{\ell(\cdot, h) \mid h \in \mathcal{H}\}$ is such that $\mathcal{L}(\mathcal{H}) \subset \mathcal{Q}(D_\ell, L_\ell)$;

iv) the features are such that $\mathcal{F}_R \subset \mathcal{Q}(D_{\mathcal{F}_R}, L_{\mathcal{F}_R})$;

v) the basic set $\mathfrak{T}$ has finite covering numbers for the norm $\|\cdot\|_{\mathcal{F}_R}$;

vi) there are constants $M > 0$ and $\eta_{\max} > 0$, and for all $\boldsymbol{\theta} \in \mathcal{T}$ a tangent set $\mathcal{V}_{\boldsymbol{\theta}}$ with finite covering numbers with respect to $\|\cdot\|_{\mathcal{F}_R}$, such that the hypotheses of Lemma 6.2.10 are satisfied. Suppose $M\eta_{\max} \geq 1/128$ and $M \geq 1$ for simplicity.

Define $W_0 = \left(L_{\mathcal{F}_R}^2/\gamma + 2D_{\mathcal{F}_R}^2\right)^{\frac{1}{2}}$, $W_\Lambda = 2\sqrt{2k}W_0$ and $W_{\mathcal{L}} = 2\sqrt{2k}\left(L_\ell^2/\gamma + 2D_\ell^2\right)^{\frac{1}{2}}$. Consider a distribution $\pi^\star \in \mathfrak{P}$, denote

$$\mathcal{R}^\star := \min_{h \in \mathcal{H}} \mathcal{R}_{\pi^\star}(h)$$

the minimum of the expected risk.

Let $\pi_{\mathfrak{S}} \in \mathfrak{S}$ be an approximation of $\pi^\star$ in the model, denote the (ideally small) bias term

$$\tau := \|\pi^\star - \pi_{\mathfrak{S}}\|_{\mathcal{L}(\mathcal{H})} + 4W_{\mathcal{L}} \|\pi^\star - \pi_{\mathfrak{S}}\|_{\mathcal{F}_R} \ . \tag{6.28}$$

Let m be a sketch size that satisfies

$$m \geq cW_\Lambda^2 \log(N/\rho) \tag{6.29}$$

for some $\rho > 0$, where $c = 1760/147$ is a universal constant and

$$N := \left(2^{14}kW_0^2\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{T}, \frac{1}{32\sqrt{2k}}\right)\right)^k \prod_{l=1}^{k}\left(3\cdot 2^{16}kMW_0D_{\mathcal{F}_R}\mathcal{N}\left(\varrho, \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}_l, 1), \frac{1}{2^{17}kMW_0L_{\mathcal{F}_R}}\right) + \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{V}_{\boldsymbol{\theta}_l}, \frac{1}{128\sqrt{2k}}\right)\right) \tag{6.30}$$

Consider items $\mathbf{z}_1, ..., \mathbf{z}_n \in \mathfrak{Z}$ drawn i.i.d. *from $\pi^\star$ and frequencies $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m \in \Omega$ drawn* i.i.d. *from Λ, which define the sketching operator $\mathcal{A}$ by (3.7). Denote $\widetilde{\pi} = \Delta_\iota(\mathcal{A}, \mathcal{A}\hat{\pi}_n) \in \mathfrak{S}$ the probability distribution recovered from the empirical sketch and*

$$\widetilde{h} = \arg\min_{h \in \mathcal{H}} \mathcal{R}_{\widetilde{\pi}}(h)$$

which requires no computation since the model is acceptable for the learning task.

Then, with probability at least $1 - (\rho + \rho')$ on the drawing of the $\mathbf{z}_i$'s and $\boldsymbol{\omega}_j$'s, the excess risk satisfies

$$\mathcal{R}_{\pi^\star}\left(\widetilde{h}\right) - \mathcal{R}^\star \leq 2\varphi$$

where

$$\varphi = \tau + \frac{4W_{\mathcal{L}}B_{\mathcal{F}_R}\left(1 + \sqrt{2\log(1/\rho')}\right)}{\sqrt{n}} + 2W_{\mathcal{L}}\iota \ . \tag{6.31}$$

Proof. Our goal is to prove that the hypotheses necessary to apply Theorem 3.2.7 hold with $\eta = 0$.

- **Admissibility**: by assumption $ii)$ the feature functions are bounded and Lipschitz in expectation. We can therefore apply Theorem 6.2.8 to prove that the RF expansion is admissible with constant $W_\Lambda = 2\sqrt{2k}\left(L^2_{\mathcal{F}_R}/\gamma + 2D^2_{\mathcal{F}_R}\right)^{\frac{1}{2}}$.
- **Compatibility**: by assumption $iii)$ the loss functions are bounded and Lipschitz in expectation. We apply Theorem 6.2.8 to prove that the kernel is compatible with constant $W_\mathcal{L} = 2\sqrt{2k}\left(L^2_\ell/\gamma + 2D^2_\ell\right)^{\frac{1}{2}}$.
- **Covering numbers**: given assumptions $iv)$ and $v)$ we can apply Lemma 6.2.10 to prove finiteness of the covering numbers of the set of normalized dipoles (and it is applied with $\delta := 1/(32\sqrt{2k})$ that indeed satisfies $\delta \leq 4M\eta_{\max}$ since we assumed $M\eta_{\max} \geq 1/128$), then Lemma 6.2.9 to bound the covering numbers of the normalized secant set. Note that during these computations all the $\max(C, 1)$ are resolved as $\max(C, 1) = C$ (recalling that necessarily $W_0 = W_\Lambda/(2\sqrt{2k}) \geq 1/(2\sqrt{2k})$).
- **Boundedness property**: we just use Lemma 3.1.5 to prove boundedness property with respect to $d_G := \|\cdot\|_{\mathcal{F}_R}$ and probability 1.

We can therefore apply Theorem 3.2.7 and obtain a bound with high probability on $\|\pi^\star - \widetilde{\pi}\|_{\mathcal{L}(\mathcal{H})}$. Then, using (6.9) we can control the excess risk with the desired quantity. □

Theorem 6.2.12 indeed address the problems that the first results obtained at the end of Chapter 3 exhibited. It yields a direct bound on the excess risk of the sketching method instead of the MMD between true and recovered distribution, and eliminates the additional error η from the bound, which now only includes the bias τ and the empirical error in $\mathcal{O}\left(1/\sqrt{n}\right)$.

Remark 6.2.13. *The bias term (6.28) is crucial, and by no mean "always" small. To make it as small as possible, one must choose the model $\mathfrak{S}$ as large as possible (with consequences on the required sketch size), since it reduces the distance between the true distribution and the model, but also the class of hypotheses $\mathcal{H}$ as small as possible (while of course keeping the learning task meaningful), since the norm $\|\cdot\|_{\mathcal{L}(\mathcal{H})}$ is increasing with $\mathcal{H}$.*

In particular, one may wrongly think that as long as the model is acceptable for the learning task, the hypothesis class $\mathcal{H}$ can be as large as desired, which reduces the optimal risk $\mathcal{R}^\star$, while keeping the excess risk small. This is of course not true, since by increasing the hypothesis class the bias may become large, and the excess risk as well.

Remark 6.2.14. *As in Chapter, 3 the lower bound on the sufficient sketch size involves two terms, the admissibility constant W_Λ that reflects our use of Bernstein's inequality, and the logarithm of the covering numbers of the normalized secant set that reflects the dimensionality of the problem. Unlike the widely applicable but sub-optimal results derived at the end of Chapter 3 that were quite generic, here our analysis makes large use of the properties of each of the various objects involved in the method. It is reflected in the presence of,* e.g. *, the Lipschitz constant of the random features in the admissibility constant, or the covering numbers of the "tangent" set of dipoles $\mathcal{V}$ in the covering numbers of the normalized secant set. We believe that the approach described in this chapter, that uses separation of components of mixture models and controlled smoothness of the kernel and feature functions, is but one example of the kind of analysis that can be done to prove the initial and more general conditions of admissibility and finite covering numbers described in Chapter 3.*

In the next two sections we instantiate this analysis for the k-means (and k-medians) and GMM with known covariance problems. In each case we will prove that:

1. the model $\mathfrak{S} \subset \mathfrak{S}_{k,2,\varrho}\left(\mathfrak{T}\right)$ is acceptable for the learning task;
2. the kernel has the right form based on Lemma 6.2.11: $\kappa(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) = \exp\left(-\frac{\varrho(\boldsymbol{\theta},\boldsymbol{\theta}')^2}{2\sigma_k^2}\right)$;

3. the loss functions are bounded and Lipschitz in expectation;
4. the random features are bounded and Lipschitz in expectation;
5. the basic set has finite covering numbers;
6. there exist tangent sets with finite covering numbers.

In the k-means and k-medians cases, we will also prove an additional bound on the bias term.

We illustrate the proof of Theorem 6.2.12 with these hypotheses in Fig. 6.3.

6.3 Application to Mixture of Diracs

We now apply the framework of the previous section to the k-medians and k-means problems. As we will see, the results obtained will be somehow more satisfying and require less assumptions in the k-medians problem compared to k-means, although in the experiments we compared our method to the classic Lloyd's algorithm for k-means. The proofs are in Appendix D.

6.3.1 Framework

Fix the sample space $\mathfrak{Z} \subset \mathbb{R}^d$, and the number of components $k > 0$. Let us define all the following objects.

Learning task. In the k-means (or k-medians) problem, a hypothesis is a set of centroids: $h := (\mathbf{c}_1, \dots, \mathbf{c}_k) \in (\mathbb{R}^d)^k$. We assume that the hypothesis class is such that centroids are bounded:

$$\mathcal{H} \subset \left\{ h = (\mathbf{c}_1, \dots, \mathbf{c}_k) \;\middle|\; \|\mathbf{c}_l\|_2 \leq R_{\mathbf{c}} \right\} \tag{6.32}$$

Note that we do not suppose that the samples themselves are bounded, only the centroids with which we are going to approach them. Also, note that the sets of centroids in the hypothesis class are not necessarily ε-separated, unlike the mixture of Diracs in the model that we will consider (see below).

The loss function is defined as

$$\ell(\mathbf{z}, h) := \min_{1 \leq l \leq k} \|\mathbf{z} - \mathbf{c}_l\|_2^b \tag{6.33}$$

where $b = 2$ for the k-means problem and $b = 1$ for the k-medians problem. In the k-means case, one recognizes the SSE defined in the previous chapter as the empirical risk: $\mathcal{R}_{\hat{\pi}_n}(h) = SSE(h)$.

Basic set of individual components. The basic set of distributions is that of Diracs whose location is a bounded vector (with the same radius $R_{\mathbf{c}}$ than the hypothesis class):

$$\mathfrak{T} := \{\pi_{\boldsymbol{\theta}} = \delta_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \mathcal{T}\} \tag{6.34}$$

where $\mathcal{T} = \mathcal{B}_{\mathbb{R}^d, \|\cdot\|_2}(0, R_{\mathbf{c}})$ is the ball of radius $R_{\mathbf{c}} > 0$.

Kernel. We define the kernel as

$$\kappa(\mathbf{z}, \mathbf{z}') := \exp\left(-\frac{\|\mathbf{z} - \mathbf{z}'\|_2^2}{2\lambda^2} \right) , \tag{6.35}$$

for some adjustable scale parameter $\lambda > 0$.

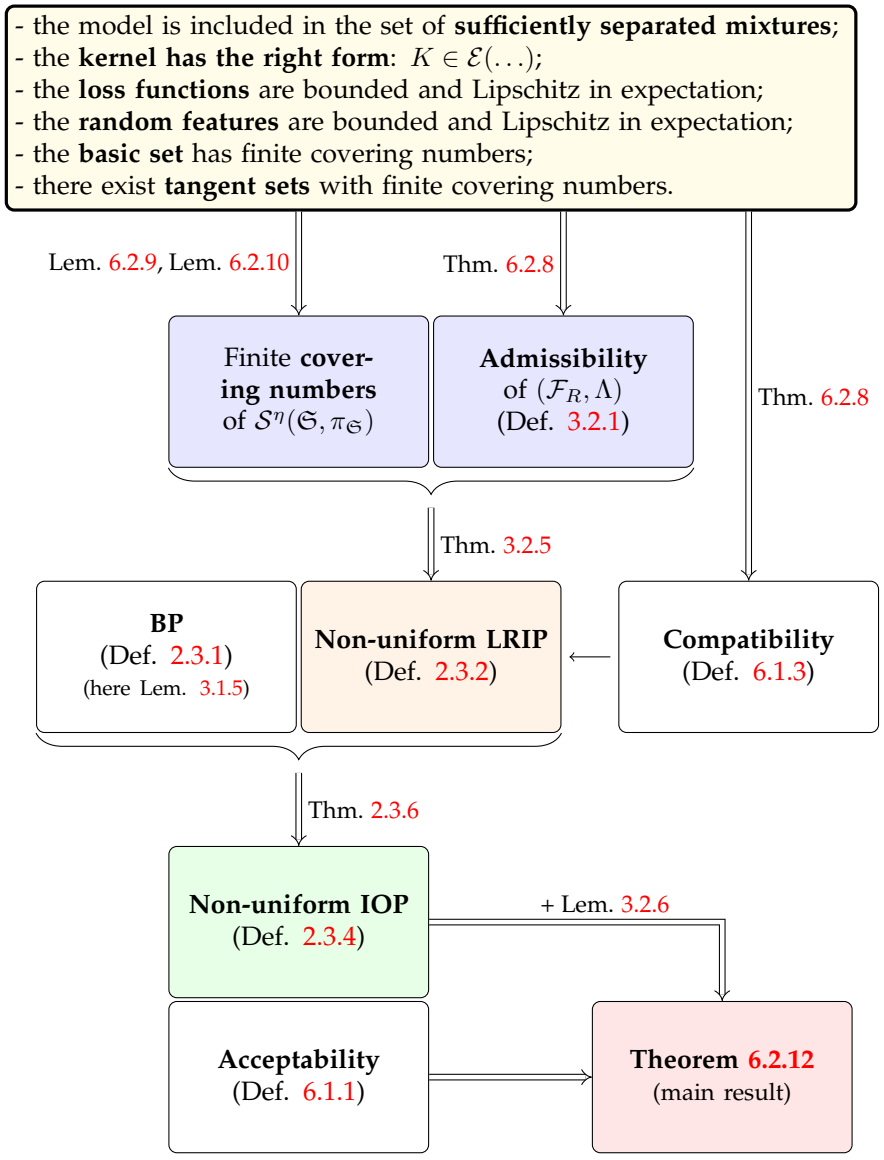

FIGURE 6.3: Illustration of the proof of Theorem 6.2.12. We use a set of hypotheses to prove admissibility and finiteness of the covering numbers of the secant sets, and therefore LRIP and IOP. The notions of compatibility of the kernel and acceptability of the model allow to relate the sketching method to statistical learning.

Metric. Define the metric ϱ as

$$\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}') := \frac{\sigma_k}{\lambda} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \tag{6.36}$$

where σ_k is defined by (6.27).

Model. The considered model is that of mixture of Diracs whose set of parameters Θ *is in the hypothesis class* $\mathcal{H}$ (as we will see, this is necessary for acceptability), *with the additional separation hypothesis*. Denote the separation:

$$\varepsilon_\lambda := \frac{2\lambda}{\sigma_k} \tag{6.37}$$

such that $\mathfrak{S}_{k,2,\varrho}(\mathfrak{T}) = \mathfrak{S}_{k,\varepsilon_\lambda,\|\cdot\|_2}(\mathfrak{T})$. The model is:

$$\mathfrak{S} := \mathfrak{S}_{\mathcal{H},\varepsilon_\lambda} = \left\{ \pi_{\Theta,\boldsymbol{\xi}} = \sum_{l=1}^{k} \xi_l \pi_{\boldsymbol{\theta}_l} \;\middle|\; \Theta \in \mathcal{H},\ \|\boldsymbol{\theta}_l - \boldsymbol{\theta}_p\|_2 \geq \varepsilon_\lambda,\ \boldsymbol{\xi} \in \mathbb{S}^{k-1} \right\} \subset \mathfrak{S}_{k,2,\varrho}(\mathfrak{T}) \ . \tag{6.38}$$

Note that, **to avoid an empty model, we assume λ is selected such that $\varepsilon_\lambda \leq 2R_{\mathbf{c}}$.**

Random Feature expansion. The chosen RF expansion of the kernel is based on a re-weighting of traditional Fourier Features $\left\{ \mathbf{z} \mapsto e^{\mathrm{i}\boldsymbol{\omega}^\top \mathbf{z}} \mid \boldsymbol{\omega} \in \mathbb{R}^d \right\}$, as described in Example 3.1.3. It basically weights down the high frequencies, while shifting the distribution of the frequencies toward high frequencies to compensate. Early experiments with the CL-OMPR algorithm showed this re-weighting to have little visible effect, and we leave more thorough experimentations with such modified Random Features for future work.

For the kernel (6.35), the distribution of frequencies for the random Fourier features is a Gaussian $\Lambda_0 = \mathcal{N}(0, \lambda^{-2}\mathbf{I})$. We define the re-weighting coefficients as

$$c(\boldsymbol{\omega}) := \sqrt{2 + \frac{\lambda^2}{d} \|\boldsymbol{\omega}\|_2^2 + \frac{\lambda^4}{d(d+2)} \|\boldsymbol{\omega}\|_4^4} \tag{6.39}$$

and in that case the constant C_{Λ_0} is

$$C_{\Lambda_0} := \sqrt{\mathbb{E}_{\boldsymbol{\omega} \sim \Lambda_0} c(\boldsymbol{\omega})^2} = \sqrt{2 + 1 + 1} = 2 \ , \tag{6.40}$$

using the moments of the χ^2-distribution. Then, the considered RF expansion $(\mathcal{F}_R, \Lambda)$ of the kernel is defined as:

$$\mathcal{F}_R := \left\{ \phi_{\boldsymbol{\omega}}(\mathbf{z}) = \frac{2}{c(\boldsymbol{\omega})} e^{\mathrm{i}\mathbf{z}^\top \boldsymbol{\omega}} \mid \boldsymbol{\omega} \in \mathbb{R}^d \right\} \qquad d\Lambda(\boldsymbol{\omega}) := \frac{c(\boldsymbol{\omega})^2}{4} \mathcal{N}\left(\boldsymbol{\omega}; 0, \lambda^{-2}\mathbf{I}\right) \ . \tag{6.41}$$

Of course all these definitions are not the only possible ones, and we have tailored the different expressions to simplify the calculations.

6.3.2 Main properties

Let us now prove the six properties required by our analysis. All proofs that are not given here are in Appendix D.

Step 1: the model is acceptable for the learning task.

Given a mixture of Diracs $\pi_{\Theta,\boldsymbol{\xi}} = \sum_{l=1}^{k} \xi_l \delta_{\boldsymbol{\theta}_l}$ in the model, by definition of the model we have $\Theta \in \mathcal{H}$. For both k-means and k-medians the expected risk $\mathcal{R}_{\pi_{\Theta,\boldsymbol{\xi}}}(h)$ *can be put to zero* by choosing $h = \Theta$, which is obviously its minimal value since the loss function is positive.

Hence the model is acceptable for the learning task, and as expected the locations of the Diracs in the recovered mixture *are* the desired centroids.

Step 2: the kernel has the right form.

Given the definition of the kernel (6.35) and metric (6.36), it is immediate that we have

$$\kappa(\delta_{\boldsymbol{\theta}}, \delta_{\boldsymbol{\theta}'}) = \exp\left(-\frac{\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2^2}{2\lambda^2}\right) = \exp\left(-\frac{\varrho(\boldsymbol{\theta},\boldsymbol{\theta}')^2}{2\sigma_k^2}\right)$$

which is the desired kernel $\kappa(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) = K(\varrho(\boldsymbol{\theta},\boldsymbol{\theta}'))$ where, by Lemma 6.2.11, $K \in \mathcal{E}(A,B,C,1)$ such that (6.21) is satisfied.

Step 3: the loss functions are bounded and Lipschitz

Lemma 6.3.1. *The loss class $\mathcal{L}$ satisfies*

$$\mathcal{L} \subset \mathcal{Q}(D_\ell, L_\ell, \mathcal{T}, \varrho) \tag{6.42}$$

where $D_\ell = (2R_{\mathbf{c}})^b$ and $L_\ell = (\lambda/\sigma_k)(4R_{\mathbf{c}})^{b-1}$, with $b=2$ for k-means and $b=1$ for k-medians.

Proof. For all $h \in \mathcal{H}$ and $\boldsymbol{\theta} \in \mathcal{T}$, by triangular inequality it is immediate that $\mathbb{E}_{\mathbf{z}\sim\delta_{\boldsymbol{\theta}}}\ell(\mathbf{z},h) \le (2R_{\mathbf{c}})^b$, where $b=2$ for k-means and $b=1$ for k-medians.

Consider $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathfrak{Z}$. Consider $h = (\mathbf{c}_1, \dots, \mathbf{c}_k) \in \mathcal{H}$, and let $l^\star$ be an index such that $\ell(\boldsymbol{\theta}_2, h) = \min_{1\le l\le k} \ell(\boldsymbol{\theta}_2, \{\mathbf{c}_l\}) = \ell(\boldsymbol{\theta}_2, \{\mathbf{c}_{l^\star}\})$. By definition of $\ell(\boldsymbol{\theta}_1, h)$ we have

$$\begin{aligned}\mathbb{E}_{\mathbf{z}\sim\delta_{\boldsymbol{\theta}_1}}\ell(\mathbf{z},h) - \mathbb{E}_{\mathbf{z}\sim\delta_{\boldsymbol{\theta}_2}}\ell(\mathbf{z},h) &= \ell(\boldsymbol{\theta}_1,h) - \ell(\boldsymbol{\theta}_2,h) \le \ell(\boldsymbol{\theta}_1,\{\mathbf{c}_{l^\star}\}) - \ell(\boldsymbol{\theta}_2,h)\\ &= \ell(\boldsymbol{\theta}_1,\{\mathbf{c}_{l^\star}\}) - \ell(\boldsymbol{\theta}_2,\{\mathbf{c}_{l^\star}\})\\ &\le (4R_{\mathbf{c}})^{b-1}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_2 = (4R_{\mathbf{c}})^{b-1}(\lambda/\sigma_k)\varrho(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\end{aligned}$$

where in the case of k-means we have used

$$\begin{aligned}\|\boldsymbol{\theta}_1-\mathbf{c}_{l^\star}\|_2^2 - \|\boldsymbol{\theta}_2-\mathbf{c}_{l^\star}\|_2^2 &= (\|\boldsymbol{\theta}_1-\mathbf{c}_{l^\star}\|_2 + \|\boldsymbol{\theta}_2-\mathbf{c}_{l^\star}\|_2)(\|\boldsymbol{\theta}_1-\mathbf{c}_{l^\star}\|_2 - \|\boldsymbol{\theta}_2-\mathbf{c}_{l^\star}\|_2)\\ &\le 4R_{\mathbf{c}}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_2.\end{aligned}$$

By symmetry we obtain $\left|\mathbb{E}_{\mathbf{z}\sim\delta_{\boldsymbol{\theta}_1}}\ell(\mathbf{z},h) - \mathbb{E}_{\mathbf{z}\sim\delta_{\boldsymbol{\theta}_2}}\ell(\mathbf{z},h)\right| \le (4R_{\mathbf{c}})^{b-1}(\lambda/\sigma_k)\varrho(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$. □

Step 4: the random features are bounded and Lipschitz

Lemma 6.3.2. *We have $\mathcal{F}_R \subset \mathcal{Q}(D_{\mathcal{F}_R}, L_{\mathcal{F}_R}, \mathcal{T}, \varrho)$ where $D_{\mathcal{F}_R} = \sqrt{2}$ and $L_{\mathcal{F}_R} = 2\sqrt{d}/\sigma_k$.*

Proof. It is immediate that

$$|\mathbb{E}_{\mathbf{z}\sim\delta_{\boldsymbol{\theta}}}\phi_{\boldsymbol{\omega}}(\mathbf{z})| = |\phi_{\boldsymbol{\omega}}(\boldsymbol{\theta})| = \frac{2\left|e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}}\right|}{c(\boldsymbol{\omega})} \le 2/\sqrt{2} = \sqrt{2}$$

and by a simple Taylor expansion

$$\begin{aligned}\left|\mathbb{E}_{\mathbf{z}\sim\delta_{\boldsymbol{\theta}}}\phi_{\boldsymbol{\omega}}(\mathbf{z}) - \mathbb{E}_{\mathbf{z}\sim\delta_{\boldsymbol{\theta}'}}\phi_{\boldsymbol{\omega}}(\mathbf{z})\right| &= \frac{2}{c(\boldsymbol{\omega})}\left|e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}} - e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}'}\right|\\ &\overset{c(\boldsymbol{\omega})\ge\frac{\lambda\|\boldsymbol{\omega}\|_2}{\sqrt{d}}}{\le} \frac{2\sqrt{d}}{\lambda\|\boldsymbol{\omega}\|_2}\|\boldsymbol{\omega}\|_2\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2 = \frac{2\sqrt{d}}{\sigma_k}\varrho(\boldsymbol{\theta},\boldsymbol{\theta}')\end{aligned}$$

which proves the result. □

Step 5: the basic set has finite covering numbers

Lemma 6.3.3. *For all $\delta > 0$ we have*

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{T}, \delta\right) \leq \max\left(\left(\frac{B}{\delta}\right)^d, 1\right) \tag{6.43}$$

where $B = 8R_{\mathbf{c}}\sqrt{d}/\lambda$.

The proof is in Appendix D.

Step 6: there exist tangent sets with finite covering numbers

Lemma 6.3.4. *For all $\boldsymbol{\theta} \in \mathcal{T}$, there exists a set $\mathcal{V}_{\boldsymbol{\theta}}$ that satisfies the requirements of Lemma 6.2.10 with $M = \frac{2\sqrt{d(d+2)}}{\sigma_k^2}$, $\eta_{\max} = 1/2$ such that for all $0 < \delta \leq 1$ (where the r.h.s. bound is here for simplicity)*

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{V}_{\boldsymbol{\theta}}, \delta\right) \leq \frac{6\sqrt{2}B^d}{\delta^{2d+1}} \tag{6.44}$$

where $B = 3^2 2^7 d\sqrt{d+2}/\sigma_k^3$.

The proof can be found in Appendix D.

6.3.3 Summary and main result for mixtures of Diracs

Combining all these properties, we get the following result.

Theorem 6.3.5. *Adopt the definitions of Section 6.3.1.*

Consider a distribution $\pi^\star \in \mathfrak{P}$, denote

$$\mathcal{R}^\star := \min_{h \in \mathcal{H}} \mathcal{R}_{\pi^\star}(h)$$

Define

$$W_{\mathcal{L}} := 2\sqrt{2k\left(2(2R_{\mathbf{c}})^{2b} + \frac{\lambda^2}{\sigma_k^2}(4R_{\mathbf{c}})^{2(b-1)}\right)} = \mathcal{O}\left(\sqrt{k}R_{\mathbf{c}}^b\right) \tag{6.45}$$

Let $\pi_{\mathfrak{S}} \in \mathfrak{S} = \mathfrak{S}_{\mathcal{H},\varepsilon_\lambda}$ be an approximation of $\pi^\star$ in the model, denote the bias term

$$\tau := \|\pi^\star - \pi_{\mathfrak{S}}\|_{\mathcal{L}(\mathcal{H})} + 4W_{\mathcal{L}}\|\pi^\star - \pi_{\mathfrak{S}}\|_{\mathcal{F}_R} \ . \tag{6.46}$$

Let m be a sketch size that satisfies

$$m \geq \mathcal{O}\left(k^2 d^2 \texttt{polylog}\left(k, d, \frac{R_{\mathbf{c}}}{\varepsilon_\lambda}, \frac{1}{\rho}\right)\right) \tag{6.47}$$

where `polylog` is a polynomial expression containing only logarithmic terms (see proof in Appendix D for detailed expression).

Consider items $\mathbf{z}_1, ..., \mathbf{z}_n \in \mathfrak{Z}$ drawn i.i.d. *from $\pi^\star$ and frequencies $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m \in \mathbb{R}^d$ drawn* i.i.d. *from Λ, which define the sketching operator $\mathcal{A}$ by (3.7). Denote $\pi_{\widetilde{\Theta},\widetilde{\boldsymbol{\xi}}} = \Delta_\iota(\mathcal{A}, \mathcal{A}\hat{\pi}_n) \in \mathfrak{S}$ the probability distribution recovered from the empirical sketch and $\widetilde{h} = \widetilde{\Theta}$.*

Then, with probability at least $1 - (\rho + \rho')$ on the drawing of the $\mathbf{z}_i$'s and $\boldsymbol{\omega}_j$'s, the excess risk satisfies

$$\mathcal{R}_{\pi^\star}\left(\widetilde{h}\right) - \mathcal{R}^\star \leq 2\varphi$$

where

$$\varphi = \tau + \frac{4\sqrt{2}W_{\mathcal{L}}\left(1 + \sqrt{2\log(1/\rho')}\right)}{\sqrt{n}} + 2W_{\mathcal{L}}\iota \ . \tag{6.48}$$

Discussion. Up to logarithmic terms, the sketch size scales as $m \geq \mathcal{O}\left(k^2 d^2\right)$, which, unlike the results obtained in Chapter 3 for other mixture models, does not depend on the size of the original database but only on the complexity of the problem, which was the desired property. However it is still sub-optimal compared to the $m \approx \mathcal{O}(kd)$ observed in practice. When examining our proof in details, it is seen that the logarithm of the covering numbers is as $\log(N) = \mathcal{O}(kd)$ (with additional logarithmic terms), which is to be expected since this is the dimensionality of the parametric model. Hence the spurious additional kd comes from the admissibility constant W_Λ^2, and its use in Bernstein's inequality. A path to improve our result would thus be to either tighten the admissibility constant, or use more refined concentration inequalities.

By definition, the recovered set of centroids $\widetilde{h}$ has ε_λ-separation, while this is *not* necessarily the case of the optimal set of centroids $h^\star = \arg\min_{h \in \mathcal{H}} \mathcal{R}_{\pi^\star}(h)$, depending on the true distribution $\pi^\star$ and the hypothesis class $\mathcal{H}$. As mentioned before, this "gap" is materialized in the bias term τ, that we further control in the next section.

6.3.4 Bounding the bias

In density fitting cases such as GMM estimation, it is intuitive that $\pi^\star$ may be close to the model, or even *in* the model as is often assumed in Generalized Method of Moments for instance, which naturally yields a small bias τ. Here however the model is formed by mixtures of Diracs, and it is of course implausible that $\pi^\star$ is exactly, or even approximately, a mixture of Diracs.

In the following Lemma, we are in fact able to relate the bias term τ to the optimal risk $\mathcal{R}^\star$, by defining an appropriate distribution $\pi_{\mathfrak{S}}$ in the model. Two cases are to be distinguished: if the optimal sets of centroids $h^\star \in \mathcal{H}$ is *already* such that centroids are ε_λ-separated (either by assumption on the distribution $\pi^\star$ or by restricting the hypothesis class $\mathcal{H}$), then the bias can be directly bounded by the optimal risk , if not then we get an additional error ε_λ.

Lemma 6.3.6. *Consider a distribution $\pi^\star$. Define $W_{\mathcal{L}}$ as in Theorem 6.3.5. We have the following.*

- ***k-medians**: Denote $h^\star = \arg\min_{h \in \mathcal{H}} \mathcal{R}_{\pi^\star}(h)$ and $\mathcal{R}^\star = \mathcal{R}_{\pi^\star}(h^\star)$. There exists a distribution $\pi_{\mathfrak{S}} \in \mathfrak{S} = \mathfrak{S}_{\mathcal{H},\varepsilon_\lambda}$ in the model such that the bias term τ defined by (6.28) satisfies*

$$\tau \leq L(\mathcal{R}^\star + \varepsilon') \tag{6.49}$$

  where $L = 1 + \frac{8 W_{\mathcal{L}} \sqrt{d}}{\sigma_k}$ and

  - *$\varepsilon' = 0$ if $h^\star = (\mathbf{c}_1^\star, \dots, \mathbf{c}_k^\star)$ is such that $\left\|\mathbf{c}_l^\star - \mathbf{c}_p^\star\right\|_2 \geq \varepsilon_\lambda$;*
  - *$\varepsilon' = \varepsilon_\lambda$ otherwise.*

- ***k-means**: Assume that the sample space is restricted so that $\|\mathbf{z}\|_2 \leq R_{\mathbf{c}}$. Denote $h^\star = \arg\min_{h \in \mathcal{H}} \mathcal{R}_{\pi^\star}(h)$ and $\mathcal{R}^\star = \mathcal{R}_{\pi^\star}(h^\star)$. There exists a distribution $\pi_{\mathfrak{S}} \in \mathfrak{S}$ in the model such that the bias term τ defined by (6.28) satisfies*

$$\tau \leq L(\sqrt{\mathcal{R}^\star} + \varepsilon') \tag{6.50}$$

  where $L = 4R_{\mathbf{c}} + \frac{8 W_{\mathcal{L}} \sqrt{d}}{\sigma_k}$ and ε' is defined as in the k-medians case.

Remark 6.3.7. *The k-means framework is seen to be somehow less convenient than the k-medians one in our analysis: not only we obtain a bound $\sqrt{\mathcal{R}^\star}$ instead of directly $\mathcal{R}^\star$, it seems that we must assume that the* sample space *itself is bounded, while this is not the case in the k-medians case where the sample space is unrestricted.*

Discussion. By Lemma 6.3.6 we obtain a distribution-free version of Theorem 6.3.5, where the bias $\tau = d(\pi^\star, \pi_{\mathfrak{S}})$ does not need to be optimized with respect to $\pi_{\mathfrak{S}} \in \mathfrak{S}$ but is directly

bounded by the optimal risk. In particular, for k-medians we obtain an oracle bound of the form

$$\mathcal{R}_{\pi^\star}(\widetilde{h}) \lesssim \mathcal{R}^\star + \dots$$

with a non-tight constant. Note that it is possible to tighten a bit further the constant L in the expression of the bias term by using matrix concentration inequalities, as is written in our publication [Gri+17]. As this result is not crucial for the theory and not one of my personal contribution, I elected to exclude it from the present manuscript. In any case, further tightening of the multiplicative constant will have to be envisioned in the future.

Consequence for super resolution Our results have interesting consequences for super-resolution [CFG12; DeC+15; DP15], which is the problem of recovering linear combinations of Diracs from noisy Fourier measurements, which is exactly the framework of this section[3], using a convex cost function however.

One of the main difference is that we draw the frequencies *randomly*, in a Compressive Sensing spirit, while usually in super-resolution the frequencies are deterministic (they stem from the use of an ideal low-pass filter on the torus). In this spirit, the most interesting feature in our results is that the sketch size is as $m = \mathcal{O}(\log(1/\varepsilon))$ with respect to the separation, while most results in super-resolution consider $m = \mathcal{O}(1/\varepsilon)$ equally spaced Fourier measurements, to reach the Shannon-Nyquist cutoff frequency. Or course, since our frequencies are drawn from a Gaussian, technically speaking they can be infinitely high, which may not be feasible with practical acquisition devices. Nevertheless, to our knowledge the incorporation or *random* Fourier measurements in a super-resolution context is fairly new, as is our analysis inspired by a Compressive Sensing strategy.

6.4 Gaussian Mixture Model

We now turn to GMM estimation with known covariance, which is the original framework by Bourrier et al. [BGP13]. All proofs are in Appendix E.

6.4.1 Framework

Consider the sample space $\mathcal{Z} = \mathbb{R}^d$, and a fixed, known positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. Recall the Mahalanobis norm:

$$\|\mathbf{x}\|_{\boldsymbol{\Sigma}} = \sqrt{\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}}$$

Fix the number of components $k > 0$.

Learning task. In the context of GMM with known covariance a hypothesis is a set of means and weights $h = ((\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k), \boldsymbol{\xi})$, where $\boldsymbol{\mu}_l \in \mathbb{R}^d$ are the means of the Gaussians and $\boldsymbol{\xi} \in \mathbb{S}^{k-1}$ are the weights. The hypothesis class is such that means are bounded for the Mahalanobis norm:

$$\mathcal{H} \subset \left\{ h = ((\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k), \boldsymbol{\xi}) \;\middle|\; \|\boldsymbol{\mu}_l\|_{\boldsymbol{\Sigma}} \leq R_{\boldsymbol{\mu}},\ \boldsymbol{\xi} \in \mathbb{S}^{k-1} \right\} \tag{6.51}$$

Once again we outline that we do not make any assumption on the samples $\mathbf{z}_i$ or their distribution $\pi^\star$, only on the GMMs that we are going to fit on them.

For a hypothesis h, denote the GMM

$$\pi_h = \sum_{l=1}^{k} \xi_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}) \ .$$

For density fitting the usual loss function is the negative log-likelihood:

$$\ell(\mathbf{z}, h) = -\log \pi_h(\mathbf{z}) \ . \tag{6.52}$$

[3] Where we used a reweighting on the features however.

Basic set of individual components. The basic set of distribution $\pi_{\boldsymbol{\theta}}$ is naturally that of Gaussians with mean $\boldsymbol{\mu} = \boldsymbol{\theta}$, also bounded, and covariance $\boldsymbol{\Sigma}$:

$$\mathfrak{T} = \{\pi_{\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \mid \boldsymbol{\theta} \in \mathcal{T}\} \tag{6.53}$$

where $\mathcal{T} = \mathcal{B}_{\mathbb{R}^d, \|\cdot\|_{\boldsymbol{\Sigma}}}(0, R_{\boldsymbol{\mu}})$ is the ball of radius $R_{\boldsymbol{\mu}} > 0$ for the Mahalanobis norm.

Kernel. We define the kernel as a gaussian kernel $\kappa \propto \mathcal{N}(0, \lambda^2 \boldsymbol{\Sigma})$ with covariance proportional to $\boldsymbol{\Sigma}$. As we will see, to satisfy the conditions required by our analysis we need to normalize this kernel by a constant C_λ:

$$\kappa(\mathbf{z}, \mathbf{z}') := C_\lambda^2 \exp\left(-\frac{\|\mathbf{z} - \mathbf{z}'\|_{\boldsymbol{\Sigma}}^2}{2\lambda^2}\right), \tag{6.54}$$

for some adjustable scale parameter $\lambda > 0$. The constant C_λ is defined as

$$C_\lambda := \left(\frac{2}{\lambda^2} + 1\right)^{\frac{d}{4}} = \mathcal{O}\left(e^{\frac{d}{2\lambda^2}}\right) \tag{6.55}$$

Metric. Define the metric ϱ as

$$\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}') := \left(\frac{\sigma_k^2}{2 + \lambda^2}\right)^{\frac{1}{2}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_{\boldsymbol{\Sigma}} \tag{6.56}$$

where σ_k is defined by (6.27).

Model. The considered model is that of GMMs whose parameters $(\Theta, \boldsymbol{\xi})$ are in the hypothesis class $\mathcal{H}$, *with the additional separation hypothesis*. Denote the separation:

$$\varepsilon_\lambda := 2\sqrt{\frac{2 + \lambda^2}{\sigma_k^2}} = \mathcal{O}\left(\sqrt{(1 + \lambda^2)\log k}\right) \tag{6.57}$$

such that $\mathfrak{S}_{k,2,\varrho}(\mathfrak{T}) = \mathfrak{S}_{k,\varepsilon_\lambda,\|\cdot\|_{\boldsymbol{\Sigma}}}(\mathfrak{T})$. The model is:

$$\mathfrak{S} := \mathfrak{S}_{\mathcal{H},\varepsilon_\lambda} = \left\{\pi_{\Theta,\boldsymbol{\xi}} = \sum_{l=1}^{k} \xi_l \pi_{\boldsymbol{\theta}_l} \;\middle|\; (\Theta, \boldsymbol{\xi}) \in \mathcal{H},\ \|\boldsymbol{\theta}_l - \boldsymbol{\theta}_p\|_{\boldsymbol{\Sigma}} \geq \varepsilon_\lambda\right\} \subset \mathfrak{S}_{k,2,\varrho}(\mathfrak{T}) \ . \tag{6.58}$$

Unlike the Dirac case, the separation ε_λ of the means of the GMMs in the model cannot be chosen as small as desired by adjusting λ, its minimal value scales in $\varepsilon_\lambda \geq \varepsilon_0 = \mathcal{O}\left(\sqrt{\log k}\right)$. As we will see, a "small" separation will induce a large sketch size, and vice-versa. This is naturally expected to have consequence on the bias: in particular if $\pi^\star$ is exactly a ε-separated GMM the bias can be zero only if $\varepsilon_\lambda \leq \varepsilon$.

Like in the Dirac case, **we assume that the radius $R_{\boldsymbol{\mu}}$ is big enough and λ is set such that $\varepsilon_\lambda \leq 2R_{\boldsymbol{\mu}}$, so that the model is not empty.**

Random Feature expansion. The chosen features are directly the classic Random Fourier features [RR07] for the Gaussian kernel κ (unlike the mixture of Diracs case we do not need to reweight them):

$$\mathcal{F}_R = \left\{\phi_{\boldsymbol{\omega}}(\mathbf{z}) = C_\lambda e^{\mathrm{i}\mathbf{z}^\top \boldsymbol{\omega}} \;\middle|\; \boldsymbol{\omega} \in \mathbb{R}^d\right\} \qquad \Lambda = \mathcal{N}\left(0, \lambda^{-2}\boldsymbol{\Sigma}^{-1}\right) \ . \tag{6.59}$$

6.4.2 Main properties

Let us now turn to proving the six properties required for our analysis.

Step 1: the model is acceptable

For any pair of distributions $\pi, \pi' \in \mathfrak{P}$, the log-likelihood has the following property

$$\mathbb{E}_{\mathbf{z}\sim\pi}\left(-\log\pi'(\mathbf{z})\right) = \mathbb{E}_{\mathbf{z}\sim\pi}\log\left(\frac{\pi(\mathbf{z})}{\pi'(\mathbf{z})}\right) - \mathbb{E}_{\mathbf{z}\sim\pi}\log\pi(\mathbf{z}) = D_{\mathrm{KL}}\left(\pi||\pi'\right) + \mathrm{H}\left(\pi\right) \tag{6.60}$$

where $D_{\mathrm{KL}}\left(\cdot||\cdot\right)$ is the Kullback Leibler divergence and $\mathrm{H}\left(\pi\right) = \mathbb{E}_{\mathbf{z}\sim\pi}\left(-\log\pi(\mathbf{z})\right)$ is the differential entropy of π.

It is well-known that the KL divergence $D_{\mathrm{KL}}\left(\pi||\pi'\right)$ is non-negative, and zero if and only if $\pi = \pi'$. Hence for $\pi_{\Theta,\boldsymbol{\xi}} \in \mathfrak{S} = \mathfrak{S}_{\mathcal{H},\varepsilon_\lambda}$ the expected risk

$$\mathcal{R}_{\pi_{\Theta,\boldsymbol{\xi}}}(h) = D_{\mathrm{KL}}\left(\pi_{\Theta,\boldsymbol{\xi}}||\pi_h\right) + \mathrm{H}\left(\pi_h\right)$$

is indeed minimized by choosing $h = (\Theta, \boldsymbol{\xi})$, which is in the hypothesis class $\mathcal{H}$ by definition of the model. Thus the model is acceptable for the learning task, and as expected the parameters of the GMM recovered from the sketch are directly the desired hypothesis.

Step 2: the kernel has the right form

The expression of the kernel is based on the following generic lemma.

Lemma 6.4.1. *Define a Gaussian kernel* $\kappa(\mathbf{z},\mathbf{z}') = \exp\left(-\frac{1}{2}\left\|\mathbf{z}-\mathbf{z}'\right\|^2_{\boldsymbol{\Sigma}_\kappa}\right)$.

For two Gaussians $\pi_1 = \mathcal{N}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right)$, $\pi_2 = \mathcal{N}\left(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right)$, *the mean kernel* $\kappa(\pi_1,\pi_2) = \mathbb{E}_{\mathbf{z}_1\sim\pi_1}\mathbb{E}_{\mathbf{z}_2\sim\pi_2}\kappa(\mathbf{z}_1,\mathbf{z}_2)$ *expresses*

$$\kappa(\pi_1,\pi_2) = \frac{\left|\boldsymbol{\Sigma}_\kappa\right|^{\frac{1}{2}}}{\left|\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2+\boldsymbol{\Sigma}_\kappa\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\left\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right\|^2_{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2+\boldsymbol{\Sigma}_\kappa}\right) \tag{6.61}$$

where $|\cdot|$ *denotes the determinant of matrices.*

Using this closed-form expression on the kernel (6.54) we get here:

$$\begin{aligned}\kappa(\pi_{\boldsymbol{\theta}},\pi_{\boldsymbol{\theta}'}) &= C_\lambda^2\frac{\left|\lambda^2\boldsymbol{\Sigma}\right|^{\frac{1}{2}}}{\left|(2+\lambda^2)\boldsymbol{\Sigma}\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\cdot\frac{1}{2+\lambda^2}\left\|\boldsymbol{\theta}-\boldsymbol{\theta}'\right\|^2_{\boldsymbol{\Sigma}}\right)\\ &= C_\lambda^2\left(\frac{\lambda^2}{2+\lambda^2}\right)^{d/2}\exp\left(-\frac{1}{2\sigma_k^2}\cdot\frac{\sigma_k^2}{2+\lambda^2}\left\|\boldsymbol{\theta}-\boldsymbol{\theta}'\right\|^2_{\boldsymbol{\Sigma}}\right) = K(\varrho(\boldsymbol{\theta},\boldsymbol{\theta}'))\end{aligned}$$

with $K(x) = e^{-\frac{x^2}{2\sigma_k^2}}$, which is the desired expression $K \in \mathcal{E}(A,B,C,1)$.

Step 3: the loss functions are bounded and Lipschitz

Lemma 6.4.2. *The loss class* $\mathcal{L}$ *satisfies*

$$\mathcal{L} \subset \mathcal{Q}(D_\ell, L_\ell, \mathcal{T}, \varrho) \tag{6.62}$$

where $D_\ell = 2R_{\boldsymbol{\mu}}^2 + \frac{1}{2}\left|\log\left|2\pi e\boldsymbol{\Sigma}\right|\right|$ *and* $L_\ell = 2R_{\boldsymbol{\mu}}\varepsilon_\lambda$.

Step 4: the random features are Bounded Lipschitz

Lemma 6.4.3. *We have* $\mathcal{F}_R \subset \mathcal{Q}(D_{\mathcal{F}_R}, L_{\mathcal{F}_R}, \mathcal{T}, \varrho)$ *with* $D_{\mathcal{F}_R} = C_\lambda$ *and* $L_{\mathcal{F}_R} = C_\lambda\sqrt{\frac{2+\lambda^2}{\sigma_k^2}}$.

Proof. It is immediate that

$$\mathbb{E}_{\mathbf{z}\sim\pi_{\boldsymbol{\theta}}} |\phi_{\boldsymbol{\omega}}(\mathbf{z})| \leq C_\lambda$$

and

$$|\langle \pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}, \phi_{\boldsymbol{\omega}} \rangle| \leq C_\lambda \|\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}\|_{\mathrm{TV}} \overset{\text{Lem. 3.3.4}}{\leq} C_\lambda \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_{\boldsymbol{\Sigma}} \leq C_\lambda \sqrt{\frac{2+\lambda^2}{\sigma_k^2}} \varrho(\boldsymbol{\theta}, \boldsymbol{\theta}')$$

which proves the result. □

Step 5: the basic set has finite covering numbers

Lemma 6.4.4. *For all $\delta > 0$, we have*

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{T}, \delta\right) \leq \max\left(\left(\frac{B}{\delta}\right)^d, 1\right). \tag{6.63}$$

with $B = 4C_\lambda R_{\boldsymbol{\mu}}$.

Proof. Consider the embedding $\varphi : \mathcal{T} \to \mathfrak{T}$ defined as $\varphi(\boldsymbol{\theta}) = \pi_{\boldsymbol{\theta}}$, which is surjective by definition of $\mathfrak{T}$.

Consider $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{T}$. By Lemma 3.3.4 we have $\|\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}\|_{\mathcal{F}_R} \leq C_\lambda \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_{\boldsymbol{\Sigma}}$. Therefore φ is C_λ-Lipschitz. Then, we have

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{T}, \delta\right) \overset{\text{Lem. A.3.2}}{\leq} \mathcal{N}\left(\|\cdot\|_{\boldsymbol{\Sigma}}, \mathcal{T}, \frac{\delta}{C_\lambda}\right) \overset{\text{Lem. A.3.4}}{\leq} \max\left(\left(\frac{4C_\lambda R_{\boldsymbol{\mu}}}{\delta}\right)^d, 1\right). \tag{6.64}$$

□

Step 6: there exist tangent sets with finite covering numbers

Let us prove the existence of tangent sets $\mathcal{V}$ with finite covering numbers.

Lemma 6.4.5. *For all $\boldsymbol{\theta} \in \mathcal{T}$, there exists a set $\mathcal{V}_{\boldsymbol{\theta}}$ that satisfies the requirements of Lemma 6.2.10 with $M = \frac{2C_\lambda(2+\lambda^2)}{e\sigma_k^2}$, $\eta_{\max} = 1/2$ such that for all $0 < \delta \leq 1$ we have*

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{V}_{\boldsymbol{\theta}}, \delta\right) \leq \frac{6C_\lambda B^d}{\delta^{2d+1}} \tag{6.65}$$

where $B = 3^2 2^6 C_\lambda^2 \left(\frac{2+\lambda^2}{e\sigma_k^2}\right)^{\frac{3}{2}}$.

6.4.3 Summary and main result for mixtures of Gaussians

Combining all these properties, we get the following result.

Theorem 6.4.6. *Adopt the definitions of Section 6.4.1.*
Consider a distribution $\pi^\star \in \mathfrak{P}$, denote

$$\mathcal{R}^\star := \min_{h\in\mathcal{H}} \mathcal{R}_{\pi^\star}(h)$$

Define

$$W_{\mathcal{L}} := 2\sqrt{2k\left(2D_\ell^2 + L_\ell^2\right)} \tag{6.66}$$

where D_ℓ, L_ℓ are defined as in Lemma 6.4.2.
Let $\pi_{\mathfrak{S}} \in \mathfrak{S}$ be an approximation of $\pi^\star$ in the model, denote the bias

$$\tau := \|\pi^\star - \pi_{\mathfrak{S}}\|_{\mathcal{L}(\mathcal{H})} + 4W_{\mathcal{L}} \|\pi^\star - \pi_{\mathfrak{S}}\|_{\mathcal{F}_R} \tag{6.67}$$

Let m be a sketch size that satisfies

$$m \geq \mathcal{O}\left(e^{\frac{d}{\lambda^2}} k^2 d^2 \texttt{polylog}\left(k, d, R_{\boldsymbol{\mu}}, \frac{1}{\rho}\right)\right) \tag{6.68}$$

where `polylog` *is a polynomial expression containing only logarithmic terms (see proof in Appendix E for detailed expression).*
Consider items $\mathbf{z}_1, ..., \mathbf{z}_n \in \mathfrak{Z}$ drawn i.i.d. *from $\pi^\star$ and frequencies $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_m \in \mathbb{R}^d$ drawn* i.i.d. *from Λ, which define the sketching operator $\mathcal{A}$ by (3.7). Denote $\pi_{\widetilde{\Theta},\widetilde{\boldsymbol{\xi}}} = \Delta_\iota(\mathcal{A}, \mathcal{A}\hat{\pi}_n) \in \mathfrak{S}$ the probability distribution recovered from the empirical sketch and $\widetilde{h} = \widetilde{\Theta}$.*
Then, with probability at least $1 - (\rho + \rho')$ on the drawing of the $\mathbf{z}_i$'s and $\boldsymbol{\omega}_j$'s, the excess risk satisfies

$$\mathcal{R}_{\pi^\star}\left(\widetilde{h}\right) - \mathcal{R}^\star \leq 2\varphi$$

where

$$\varphi = \tau + \frac{4C_\lambda W_{\mathcal{L}}\left(1 + \sqrt{2\log(1/\rho')}\right)}{\sqrt{n}} + 2W_{\mathcal{L}}\iota\,. \tag{6.69}$$

Bias. Unlike the k-means and k-medians case, we do not have further control on the bias term yet. One term uses the norm $\|\cdot\|_{\mathcal{F}_R}$, which can be bounded by the TV-norm (Lemma 3.1.5) then by Pinsker's inequality [FHT03] by the square root of the Kullback-Leibler divergence for instance, however the second term with the norm $\|\cdot\|_{\mathcal{L}(\mathcal{H})}$ may be technical to relate to other quantities without additional assumptions on the true distribution of the data $\pi^\star$. Let us nevertheless note that, unlike the previous case where the true distribution of the data was obviously not a mixture of Diracs, here it is plausible that the true distribution $\pi^\star$ is exactly or "close" to a GMM in the model (*i.e.* with ε_λ-separation of means). We leave further control of the bias for future work.

Separation and sketch size. Similar to the Dirac case, by adjusting the parameter λ one can see a trade-off between the required sketch size (6.68) and the separation ε_λ (6.57) of components in the model: as λ decreases, the kernel κ is more "precise", higher frequencies are sampled (but more of them are required) and the separation in the model can be smaller. However, as mentioned before, while still adjustable by λ, unlike the Dirac case the separation ε_λ cannot go to zero. At one end of the spectrum, it scales in $\mathcal{O}\left(\sqrt{\log k}\right)$, which compares favorably to existing literature[4] [AM05; VW04] but the sketch size is polynomial in k and exponential in d, while at the other end the sketch size is polynomial in k and d but the required separation is in $\mathcal{O}\left(\sqrt{d\log k}\right)$. A few values are summarized in Table 6.1.

[4] Recent works make use of more complex conditions that theoretically permits arbitrary separation [BS10b], however all these approaches use the full data while we consider a compressive approach that uses only a sketch of the data.

λ^2	C_λ	ε_λ	m
$\mathcal{O}(1)$	$\mathcal{O}\left(e^{d/2}\right)$	$\mathcal{O}\left(\sqrt{\log k}\right)$	$\mathcal{O}\left(e^d k^2 d^2 \texttt{polylog}\left(k, d, R_{\boldsymbol{\mu}}, \frac{1}{\rho}\right)\right)$
$\mathcal{O}\left(\sqrt{d/\log k}\right)$	$\mathcal{O}\left(\sqrt{k}\right)$	$\mathcal{O}\left(\sqrt{d+\log k}\right)$	$\mathcal{O}\left(k^3 d^2 \texttt{polylog}\left(k, d, R_{\boldsymbol{\mu}}, \frac{1}{\rho}\right)\right)$
$\mathcal{O}\left(\sqrt{d}\right)$	$\mathcal{O}(1)$	$\mathcal{O}\left(\sqrt{d \log k}\right)$	$\mathcal{O}\left(k^2 d^2 \texttt{polylog}\left(k, d, R_{\boldsymbol{\mu}}, \frac{1}{\rho}\right)\right)$

TABLE 6.1: Trade-off between separation ε_λ of means in the model and required sketch size m. The sketch size does not seem to be improved by choosing λ greater than $\sqrt{d}$.

6.5 Conclusion

In this chapter, we derive an advanced analysis of mixture models to fully exploit the results of Chapter 3. Our goal was to prove strong conditions to obtain information-preservation guarantees with no additive error ($\eta = 0$), and relate the guarantees to usual learning and expected risk control. Our analysis of mixture models includes a key assumption of separation of components, from which we can decompose a difference of mixtures into a sum of separated dipoles, controlled independently with usual tools such as incoherence. The link with risk control was made through the definition of appropriate metrics (notion of *compatibility* of the kernel) and low-dimensional models that are directly related to the hypothesis class (notion of *acceptable* model).

This set of results was applied to two mixture learning problems. The first was to relate the recovery of mixtures of Diracs from the sketch with the k-means and k-medians problems. It was seen that separation of Diracs can be controlled by directly tuning the precision of the kernel. The sketch size resulting from our analysis is somehow better than usual results in super-resolution[5] with respect to this separation, which is certainly due to the introduction of randomness in the design of the measurement operator, as is done in Compressive Sensing.

The second case was that of GMM estimation with known covariance, which is the original framework that was the starting point of this thesis. We provided guarantees with usual log-likelihood, and observed an interesting trade-off between required separation of means and sketch size. The remaining bias term was not subject to the same analysis than in the case of mixtures of Diracs and will be the subject of future investigations.

In both applications a sketch size at least quadratic in the number of components and dimension of the data was obtained, which is polynomial, but still seem to be sub-optimal compared to the linear dependence observed in practice. Since our control of covering numbers reflects the right dimensionality of the problem, we conjectured that potential sub-optimality was due to our current strategy for concentration of measure.

[5] Although, once again, the frameworks and guarantees are fairly different.

Chapter 7

Conclusion

This thesis work focused on the theoretical analysis and practical implementation of a sketching method for efficient learning of mixture models. The notion of linear sketches was revisited under the lens of Compressive Sensing and inverse problems, based on the idea that any linear sketch is a collection of generalized moments with respect to the underlying probability distribution of the data. The method was analyzed by drawing innovative connections between generalized Compressive Sensing, kernel mean embedding and Random Feature expansions. A versatile greedy algorithm was developed and applied to three different problems on synthetic and real data.

7.1 Summary of the contributions

We summarize our main contributions below.

7.1.1 Theoretical contributions

In Chapter 2, we described a first contribution independent of sketching. We extended previous results for linear inverse problems that stated that the existence of a decoder that is instance optimal with respect to a low-dimensional model is equivalent to the measurement process satisfying a certain Lower Restricted Isometry Property on this model. We proved that this result holds for non-linear measurement operators in any metric set, and extended the formulation to guarantees that are non-uniform in probability.

In Chapter 3, we described the main framework for the proposed sketching method. We defined the sketching operator as a collection of generalized moments with respect to a finite measure. These moments were defined using Random Feature expansions of positive definite kernels, and the resulting operator was shown to approximate the infinite-dimensional mapping induced by the kernel mean embedding methodology. We used this mathematical framework to provide conditions under which this operator satisfies the non-uniform LRIP with high probability. Then, using results from the previous chapter, we provided information-preservation guarantees for the sketching method.

In a second part of the chapter, we relaxed most of the technical assumptions used to prove the LRIP and provided preliminary guarantees that are only meaningful when the sketch is as large as the original database (fortunately theses results are seen in Chapter 5 to be largely pessimistic in practice). We showed that this analysis applies to a large class of models with no difficulty. We provided two examples: GMMs with diagonal covariance and mixtures of elliptic stable distributions. The latter is the most interesting, since despite the sub-optimality of the sketch size it is the only known estimator for mixtures of multivariate stable distributions. In the course of these first proofs, we manipulated several useful tools on covering numbers and provided intermediary results on Gaussians and stable distributions.

This analysis was greatly extended in Chapter 6, with two purposes: improving over the sub-optimal preliminary guarantees obtained at the end of Chapter 3 and relating the obtained information-preservation guarantees to more traditional learning metrics. We analyzed generic mixture models based on the key assumptions that the components of a mixture are pairwise sufficiently separated and that the mean kernel between two components is sufficiently decreasing around the origin. Under these assumptions, the difference between two

mixtures can be decomposed into a sum of incoherent "dipoles". We instantiated this framework on two examples. First mixtures of Diracs, for which we provided guarantees with respect to the classic k-means and k-medians loss functions. And second GMMs with fixed known covariance, where an interesting trade-off between sketch size and required separation of components was observed. In both cases the required sketch size is at least $m \geq \mathcal{O}\left(k^2 d^2\right)$, where k is the number of components and d is the dimension, which is indeed independent of the size of the original database but still sub-optimal compared to the $m \approx \mathcal{O}\left(kd\right)$ observed in practice.

7.1.2 Algorithmic contributions

Chapter 4 was devoted to defining a greedy heuristic to estimate a mixture model from a sketch. We first analyzed the cost function and showed in a simple case that it may be convex by block with respect to each component when the residual signal is sufficiently close to zero. A block coordinate gradient descent was indeed observed to perform well when initialized close to the optimum, but to fail in the absence of prior knowledge. We then introduced the proposed greedy strategy. Inspired by Orthogonal Matching Pursuit, it alternates between adding a component and performing non-convex updates. In its variant with Replacement, it performs more iterations than the desired number of components and suppress some of them with Hard Thresholding. Although it requires *no prior knowledge*, this algorithm was shown to perform on par with the block coordinate descent algorithm *initialized close to the optimum*. Finally, we outlined the simplicity of its implementation and possibility for optimization at key steps of the algorithm.

In Chapter 5 we applied the proposed greedy strategy to the sketched mixture model estimation problem. We first described an unsupervised method to learn an appropriate kernel, that requires only a fraction of training data. Although relatively simple compared to existing literature on the matter, it was shown to be extremely fast compared to a supervised approach and to perform well on a wide range of problems.

The sketch mixture model estimation method was then instantiated on three problems.

The first consists in recovering a mixture of Diracs from the sketch. We showed that it compares favorably with the classic k-means algorithm. The sketching method was observed to be much more efficient on large databases, and we showed that unlike k-means it does not necessitate several replicates with random initializations. The method was applied to a spectral clustering problem for handwritten digits recognition.

Then the sketching method was applied to the estimation of GMMs with unknown diagonal covariances. In this specific case we defined an additional algorithm that has a better cost than the greedy approaches with respect to the number of components, although it was sometimes seen to yield results that are less precise than the ones obtained with CL-OMPR. The latter was observed to reach the precision of EM even when EM is repeated 10 times with random initializations, while being faster and more memory efficient on large databases. The method was applied to a speaker verification problem on a database comprising hundreds of millions of elements.

Finally the proposed method was instantiated for the estimation of mixtures of multivariate elliptic stable distributions. It was shown to be orders of magnitudes faster and more precise than the few results reported in the literature for the univariate case. To our knowledge, this is the first algorithm capable of handling the multivariate case.

In all applications the sketch size required to obtain good estimation results was approximately observed to scale as $m \approx \mathcal{O}\left(kd\right)$, which is intuitively an optimal rate since it is the number of parameters.

7.2 Perspectives

Let us now outline some interesting perspectives that arise from our work, some of which have already been evoked in the course of the manuscript.

7.2.1 Short-term perspectives

Extension of CL-OMPR. Many immediate extensions can be envisioned to the CL-OMPR algorithm. In case where the number of atoms is unknown , integrating a stopping condition when the residual falls below a threshold requires no effort, and model selection is somehow made even easier than with classic algorithms like k-means or EM for which it is known to be a difficult problem. A remaining issue with the greedy approach is its quadratic cost in the number of components. In the specific case of GMMs, we defined a hierarchical algorithm that is faster at large k, and it would be interesting to generalize such a strategy to any mixture model.

In [BSR15] Boyd an al. use an algorithm similar to CL-OMP for many problems other than mixture model estimation, hence the CL-OMPR (*with* Replacement) can also be applied to these problems and may improve over CL-OMP without Replacement.

Use of fast transforms. All throughout our description of the implementation of the sketching method, we have outlined the possibility to use fast transforms to accelerate some computations (see Section 4.3.3). The most prominent example is that of mixture of Diracs, where both the computation of the sketch with complex exponentials and the CL-OMPR algorithm can directly benefit from replacing the matrix of frequencies $\mathbf{W}$ by a structured fast version. Other models such as GMMs with diagonal covariance require fast versions of $\mathbf{W} \odot \mathbf{W}$, which is a less common but interesting challenge.

A paramount question is then to examine the possible integration of such transforms into theoretical proofs. Indeed, many structured transforms are shown to have similar properties as *e.g.* random Gaussian matrix, and are already integrated in fast versions of RF expansions of kernels [LSS13; Yan+15]. Hence if we were able to define an admissible RF expansion (*i.e.* such that $\|\cdot\|_{\mathcal{F}_R} \lesssim \|\cdot\|_\kappa$ on differences of distributions in the model) that integrates such structures, our theoretical analysis would apply.

Combination with dimension-reduction. As we advocated in the introduction of the thesis, the proposed sketching method is tailored for databases with a large number of elements n, and not so much for element that have a high dimension d. In particular, we observe empirically that the sketch has a size that is linear in d, and provided theoretical guarantees with a sketch size at least quadratic in d. However, in the introduction we also mentioned a number of existing methods that reduce the dimension of each individual data point while retaining the ability to perform the learning task, like k-means [BZD10] or GMM estimation [Das99].

It would be feasible to combine these approaches with the sketching method: first reduce the dimension of the data, *then* sketch them and learn. Implementing such a scheme is immediate, and it may very well be possible to simply "plug" the theoretical guarantees of these dimension reduction methods in our RIP analysis. Hence we hope to develop more "complete" compression methods in the future, that would handle data that are both large in number and high-dimensional.

Practical extension to other mixture models and sketching operators. In Chapters 4 and 5 we observed the CL-OMPR algorithm to be extremely versatile and outlined that it is applicable as long as $\boldsymbol{\theta} \mapsto \mathcal{A}\pi_{\boldsymbol{\theta}}$ is differentiable. It could therefore be applied to many other models of distributions $\pi_{\boldsymbol{\theta}}$ or sketching operators $\mathcal{A}$.

Preliminary experiments with Gaussian Locally Linear Mapping (GLLiM) [DFH14], which is another restriction of the full GMM framework whose purpose is to perform high-dimensional regression, are very encouraging. The CL-OMPR algorithm also opens the way for estimating more exotic mixture models that could not be envisioned until now [Das+05] just by providing the expression of their characteristic functions (when using the Fourier sketch), as we did for mixtures of stable distributions. In these cases practical applications may still be scarce due to the lack of estimators, but we hope that our work, among others, opens the way to a new class of methods that use such mixture models.

Other sketching operators can also be envisioned. An interesting idea is to replace the complex exponential used in the experiments with other non-linearities $\rho_{\text{im.}}(\cdot)$, more attuned to physical devices, or comparable to classic non-linearities used *e.g.* in the Neural Networks

literature. Based on our work, the method described in [Sch17] replaces the complex exponential by the binary output of a periodic universal quantifier, and shows that the performance of the k-means method can be preserved for sketches comprising more measurements, but stored more efficiently in memory. We expect more sketching operators to emerge in the future.

7.2.2 Mid-term perspectives

Bridging the gap between theory and practice. As mentioned before, the sketch size obtained in theory $m \approx \mathcal{O}\left(k^2 d^2\right)$ does not match that observed in practice $m \approx \mathcal{O}\left(kd\right)$, which may be due to an artifact of the proof. As described in Chapter 6, the size of the sketch arises from two factors: the logarithm of the covering numbers of normalized secant sets, which represents the "dimensionality" of the problem, and the admissibility constant, which is used in Bernstein's concentration inequality. In the two considered cases the covering numbers are probably optimal: we indeed show that the normalized secant sets have an upper box counting dimension in $\mathcal{O}\left(kd\right)$, which is the number of parameters in the problem. Hence it is most probable that the sub-optimality of the results comes from the way we exploited Bernstein's inequality. With the technique employed, that controls dipoles by bounding their coherence, it seems at least difficult to avoid the factor $\sqrt{k}$ that appears in Theorem 6.2.5 and in the admissibility constant. Thus a potential solution would probably come by using a more powerful tool than Bernstein's inequality, or changing our analysis strategy completely.

Extension of the analysis to other mixtures of separated components. In Chapter 6 we derived sufficient conditions for the sketched estimation of generic mixture models, based on a separation assumption. In our analysis, the mean kernel must be as $\kappa(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'}) = K(\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}'))$, where ϱ is some metric and K is *sufficiently decreasing* around the origin (in order to distinguish separated components). We showed that the Fourier sketch was well-adapted for mixtures of components that are *localized in space* (mixtures of Diracs and GMMs with known covariance), which is intuitive. A paramount question is to define other sketching operators and mixture models that yield appropriate expressions of the kernel.

As an illustration, consider a particular case of GMM with unknown covariance: mixtures of two-dimensional centered Gaussians with "flat" covariance, defined as a diagonal matrix rotated by an angle θ (Fig. 7.1, left): $\pi_\theta = \mathcal{N}\left(0, \mathbf{R}_\theta \mathrm{diag}\left([\sigma_1^2, \sigma_2^2]\right) \mathbf{R}_\theta^\top\right)$ where

$$\mathbf{R}_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

In that case, it is possible to show that with a Gaussian kernel $\kappa(\mathbf{z}, \mathbf{z}') = e^{-\frac{1}{2}\left\|\mathbf{z}-\mathbf{z}'\right\|_2^2}$ the mean kernel expresses

$$\kappa(\pi_\theta, \pi_{\theta'}) = \frac{1}{\sqrt{A + B\sin^2(\theta - \theta')}} \tag{7.1}$$

where $A = (1+2\sigma_1^2)(1+2\sigma_2^2)$ and $B = (\sigma_1^2 - \sigma_2^2)^2$ (proof in Appendix E). This kernel, illustrated in Fig. 7.1, seems to have the desired shape to apply our analysis: it is translation-invariant with respect to θ, with a decrease around the origin that becomes sharper as the Gaussians become "flatter" and $B = (\sigma_1^2 - \sigma_2^2)^2$ increases. Therefore we could potentially apply the results of Chapter 6 to learn mixtures of flat rotated Gaussians, although we have not completed the proof. Interestingly, learning such flat Gaussians is similar to learning the directions of several subspaces, which can be related to dictionary learning. Hence we expect other mixture models and non-linearities to emerge such that an analysis similar to ours applies.

Necessary conditions. Traditional Fourier analysis includes proofs of necessity of various conditions, such as the Shannon-Nyquist universal threshold. We have seen that our analysis of mixture models in Chapter 6 rests on an assumption of separation of components, and of sufficient precision of the kernel. It would be interesting to determine which proportion of these conditions is necessary. Such a study could also help us design other kernels, by realizing that the present conditions are either too strong or somewhat tight.

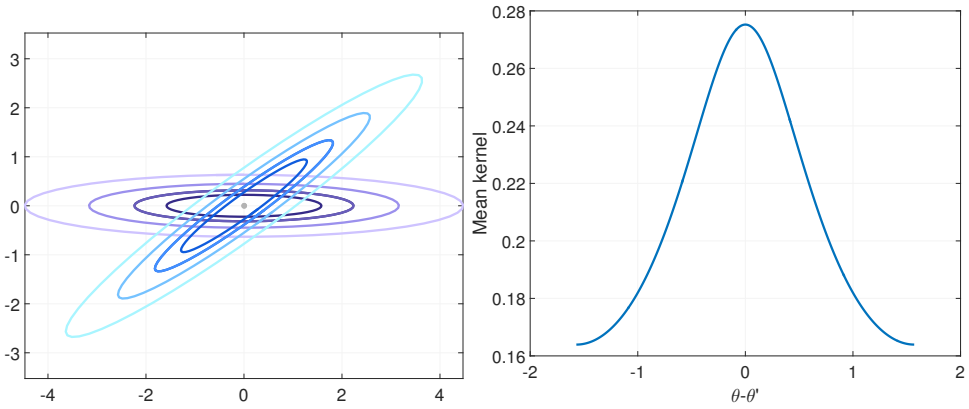

FIGURE 7.1: **Left:** illustration of two 2-dimensional centered Gaussians with different rotated covariances (with shared eigenvalues $\sigma_1^2 = 5$, $\sigma_2^2 = 0.1$, and respective angles of rotation $\theta = 0$, $\theta' = \pi/5$). **Right:** corresponding mean kernel with respect to the difference in angles of rotation $\theta - \theta'$, with a Gaussian kernel.

7.2.3 Long-term perspectives

Algorithmic guarantees, convex relaxation. A paramount question is of course to obtain an algorithm with provable guarantees. In this thesis we focused on proving that minimizing the cost function indeed permits recovery (as we have seen, a somewhat technical question), and providing a good heuristic algorithm which was thoroughly tested in practical situations. However we are still missing an algorithm with guarantees. Existing analyses for similar methods [BSR15] guarantee that when the number of components k grows to infinity, the true signal is approximated. However they do not really apply in our context, where we are looking for a mixture of exactly k components, else our theoretical analysis fails. We have quickly shown in simple situations that local block convexity sometimes occurs when close to the optimum, a possibility would be to see if under some assumptions the CL-OMPR algorithm (or any other algorithm) indeed falls into this basin of attraction with high probability.

A long-term perspective would be to examine the potential implications of the LRIP or even a full RIP on the cost function itself: can it be used to prove stronger convexity properties? A final possibility would then be to define convex relaxations of the cost function, as is done in Compressive Sensing in a finite-dimensional framework. In the context of mixtures of Diracs, such convex relaxation is performed in recent advances in super-resolution [DP15; DeC+15]. Their analysis however does not use RIP-like conditions, it is intriguing to see if some of our results can be integrated in this context.

Extension to other learning tasks. By drawing connections with inverse problems, we have designed a mixture estimation method from a sketch. This is by no mean the only envisioned use of sketches. As mentioned before such mean kernel embeddings (with or without random features) have been used for two-sample tests, classification of distributions, independence testing, and so on. A particularity of our analysis is that we prove that with a *limited number* of random features the sketch contains approximately all the information on the probability distribution of the data for certain learning tasks. It would be extremely interesting to see if this potentially applies to performing other learning tasks with sketches[1].

Existence of non-uniform problems. We formulated a non-uniform version of the LRIP that is entirely new. Although we elected to prove the non-uniform version of the LRIP for the sketch problem in this manuscript, it is possible to show that the uniform version also holds[2],

[1] In particular, in the two cases studied in Chapter 6 the link between the model $\mathfrak{S}$ and the learning task at hand was almost tautological: the notion of *acceptability* of the model was trivial to prove and not really exploited. Other learning tasks that are fundamentally different might ask for a redefinition of the notion of low-dimensional model and/or non-trivial acceptability proofs.

[2] We chose the non-uniform version in this thesis in the interest of avoiding redundancy with our paper [Gri+17] and introducing slightly different mathematical tools. The interested reader can read the uniform version in our paper.

as we do in [Gri+17]. To the best of our knowledge it is still an open question to prove the existence of cases where the non-uniform version holds and the uniform version *does not*. Note that it is already known that uniform results are impossible to meet for certain practical algorithms [Rau08].

Link with Neural Networks. In Chapter 5 we outlined the computational similarity between the linear sketch and a one-layer Neural Network (NN). Deep Neural Networks are now state-of-the-art in a large number of learning tasks but it is well-known that they still await a thorough mathematical explanation of these surprisingly good results. A great effort is dedicated today to providing such an analysis. It may be possible to draw further connections between some of the tools described in this thesis work and the methodology of neural networks.

It has already been noted that a network with random weights may give rise to a well-behaved embedding of data points [GSB15], which can be analyzed in terms of kernels and RF expansions [RR09]. For instance, define as usual the sketching operator $\mathcal{A}\mu$ such that $\|\mathcal{A}\mu\|_2 \approx \|\mu\|_\kappa$ for some kernel κ. Then, consider the realization of a RF expansion of another translation-invariant kernel such that $K(\|y - y'\|_2) \approx \langle \Phi(y), \Phi(y') \rangle$ where $y \in \mathbb{C}^m$. We can define the "level-2" mapping as $\Phi_2(\mu) = \Phi(\mathcal{A}\mu)$. It corresponds to defining a finite-dimensional approximate mapping for a "level-2" kernel [Mua+12; OSS15]:

$$\langle \Phi_2(\pi), \Phi_2(\pi') \rangle \approx K(\|\mathcal{A}\pi - \mathcal{A}\pi'\|_2) \approx K(\|\pi - \pi'\|_\kappa) = \kappa_2(\pi, \pi')$$

Using such a kernel, the "sketch" of a database $\Phi_2(\hat{\pi}_n)$ (which is no longer a linear sketch) may be more powerful than traditional linear sketches to perform complex tasks. The attentive reader would have recognized that such a sketch is often a two-layer NN: for instance, in the case of Random Fourier features, it is built by multiplying the matrix of data by the matrix of frequencies, taking pointwise complex exponential, pooling the columns (which at this point is the normal linear sketch), but then again multiplying by some frequencies and taking pointwise complex exponentials. Such intricate operators may be analyzed using the results of Chapter 2, which provide information-preservation guarantees even when the measurement operator is non-linear. These guarantees for finite or infinite-dimensional objects such as distributions would be useful to further understand the behavior of these networks. Proving (or disproving!) the LRIP for such constructions can no longer rely on normalized secant sets, which are useful only when the measurement operator is linear, and it is probable that an entirely new set of tools will have to be defined.

Appendix A

Definitions, Preliminary results

In this first Appendix we group some definitions, notations, and preliminary results that will be useful.

A.1 Notations, definitions

A.1.1 Metrics and covering numbers

Definition A.1.1. *A **pseudometric** d over a set X satisfies all the axioms of a metric, except that $d(x,y) = 0$ does not necessarily imply $x = y$. Similarly, a **seminorm** $\|\cdot\|$ over a vector space X satisfies the axioms of a norm except that $\|x\| = 0$ does not necessarily imply $x = 0$.*

The radius of a seminormed vector space $(X, \|\cdot\|)$ is denoted $\mathrm{rad}_{\|\cdot\|}(X) = \sup_{x\in X} \|x\|$. The diameter of a pseudometric set (X, d) is denoted $\mathrm{diam}_d(X) = \sup_{x,x'\in X} d(x,x')$.

Definition A.1.2 (Ball, δ-covering, covering number)**.** *Let (X, d) be a pseudometric space. For any $\delta > 0$ and $x \in X$, we denote $\mathcal{B}_{X,d}(x,\delta)$ the **ball** of radius δ centered at the point x:*

$$\mathcal{B}_{X,d}(x,\delta) = \{y \in X,\ d(x,y) \leq \delta\}.$$

*Let $Y \subset X$ be a subset of X. A subset $Z \subset Y$ is a δ-**covering** of Y if $Y \subset \bigcup_{z\in Z} \mathcal{B}_{X,d}(z,\delta)$.*
*The **covering number** $\mathcal{N}(d, Y, \delta) \in \mathbb{N} \cup \{+\infty\}$ is the smallest number of points $y_i \in Y$ such that the set $\{y_i\}$ is a δ-covering of Y.*

Remark A.1.3. *Our definition of covering numbers is that of* internal *covering numbers, meaning that the centers of the covering balls are required to be included in the set being covered. Somehow counter-intuitively these covering numbers (for a fixed radius δ) are not necessarily increasing with the inclusion of sets: for instance, consider a set A formed by two points, included in set B which is a ball of radius δ. Suppose those two points diametrically opposed in B. We have $A \subset B$, but two balls of radius δ are required to cover A (since their centers have to be in A), while only one such ball is sufficient to cover B. See Lemma A.3.1.*

A.1.2 Measures

Definition A.1.4 (Nonnegative measure)**.** *A measure μ over a measurable space $(X, \mathcal{B})$ is said to be **nonnegative** if:*

$$\forall B \in \mathcal{B},\ \mu(B) \geq 0.$$

Most often, a measure is by definition nonnegative, and called *signed* measure when it is not. However to avoid confusion we sometimes specify that a measure is indeed nonnegative.

Definition A.1.5 (Support of a measure). *The **support** of a signed measure μ over a measurable, topological space X is defined to be the closed set,*

$$\operatorname{supp}(\mu) := X \backslash \bigcup \{U \subset X : \ U \text{ is open}, \ \mu(U) = 0\} .$$

Definition A.1.6 (Total variation norm, Finite measure). *Let μ be a signed measure over a measurable space $(X, \mathcal{B})$. Define the Jordan decomposition (μ^-, μ^+) of μ where μ^+ and μ^- are positive measures (see [Fis12] and [Rud87] Chap. 6 for more details). Denote $|\mu| = \mu^+ + \mu^-$. The **total variation norm** of μ is defined as:*

$$\|\mu\|_{TV} = |\mu|(X) = \int_X d|\mu|(x).$$

*The measure μ is said **finite** if $\|\mu\| < \infty$.*

Note that if μ is totally continuous with respect to the Lebesgue measure, i.e. *if there exists an integrable function f such that $d\mu(x) = f(x)dx$, then the total variation norm is the classic L^1-norm of this function: $\|\mu\| = \|f\|_{L^1}$.*

Definition A.1.7 (Kullback-Leibler divergence [KL51]). *Let $\pi, \widetilde{\pi}$ be two probability measures* (i.e. *nonnegative and such that their total variation norm is equal to 1) over a measurable space $(X, \mathcal{B})$. Provided π is absolutely continuous with respect to $\widetilde{\pi}$, the **Kullback-Leibler** divergence from $\widetilde{\pi}$ to π is defined as*

$$D_{KL}(\pi||\pi') = \int_X \log \frac{d\pi}{d\widetilde{\pi}} d\pi \ . \tag{A.1}$$

If π and π' are both continuous with respect to e.g. *the Lebesgue measure, with density f and $\widetilde{f}$, then*

$$D_{KL}(\pi||\pi') = \int_X \log \frac{f(x)}{\widetilde{f}(x)} f(x)dx \ . \tag{A.2}$$

A.1.3 Sets and models

Consider $(X, \|\cdot\|)$ a seminormed real vector space.

Given an integer $k > 0$, sets $Y_l \subset X$, $l = 1, ..., k$, and a set $\mathcal{W} \subset \mathbb{R}^k$, we define

$$Y_{\mathcal{W}}^{(k)} := \left\{ \sum_{l=1}^{k} \xi_l y_l \ \middle| \ \boldsymbol{\xi} = (\xi_l)_{l=1}^k \in \mathcal{W}, \ y_l \in Y_l \right\} \tag{A.3}$$

In particular, when $Y_1 = ... = Y_k = Y$, we obtain a *mixture set* denoted $Y_{\mathcal{W}}$.

Given $\eta \geq 0$ and $y \in Y$, the *normalized secant set* $\mathcal{S}^{\eta}_{\|\cdot\|}(y, Y)$ is defined as

$$\mathcal{S}^{\eta}_{\|\cdot\|}(y, Y) := \left\{ \frac{y - y'}{\|y - y'\|} \ \middle| \ y' \in Y, \ \|y - y'\| > \eta \right\} \tag{A.4}$$

When $\eta > 0$, we say that the normalized secant set is *extruded*. These definitions are generalizations of those used in the core of the thesis.

A.2 Measure concentration

In Theorem 3.2.5 we use Bernstein's inequality in the following simple version [Sri02]:

Lemma A.2.1 (Bernstein's inequality ([Sri02], Thm. 6)). *Let $x_1, \dots, x_n \in \mathbb{R}$ be i.i.d. bounded random variables such that $\mathbb{E}x_i = 0$, $|x_i| \leq M$ and $Var(x_i) \leq \sigma^2$ for all i's.*

Then for all $t > 0$ we have

$$P\left(\frac{1}{n}\sum_{i=1}^{n} x_i \geq t\right) \leq \exp\left(-\frac{nt^2}{2\sigma^2 + 2Mt/3}\right). \tag{A.5}$$

We also report a concentration result in Hilbert spaces from [RR09].

Lemma A.2.2 ([RR09], Lemma 4). *Let $\mathbf{z}_i, \dots, \mathbf{z}_n \in \mathcal{H}$ be i.i.d. random variables in a Hilbert Space $(\mathcal{H}, \|\cdot\|)$ such that $\|\mathbf{z}_i\| \leq M$ with probability one. Denote $\bar{\mathbf{z}}$ their empirical average $\bar{\mathbf{z}} = \left(\sum_{i=1}^{n} \mathbf{z}_i\right)/n$. Then for any $\rho > 0$, with probability at least $1 - \rho$,*

$$\|\bar{\mathbf{z}} - \mathbb{E}\bar{\mathbf{z}}\| \leq \frac{M}{\sqrt{n}}\left(1 + \sqrt{2\log\frac{1}{\rho}}\right). \tag{A.6}$$

A.3 Generalities on covering numbers

In this section we formulate generic results on covering numbers.

A.3.1 Basic properties

Lemma A.3.1. *Let $A \subset B \subset X$ be subsets of a pseudometric set (X, d), and $\delta > 0$. Then,*

$$\mathcal{N}(d, A, \delta) \leq \mathcal{N}(d, B, \delta/2)\ . \tag{A.7}$$

Proof. Let $b_1, ..., b_N$ be a $\delta/2$-covering of B. We construct a δ-covering a_i of A in the following way. Each b_i is either: a) in the set A, in which case we take $a_i = b_i$, b) at distance less than $\delta/2$ of a point $a \in A$, in which case we take $a_i = a$ and note that the ball centered on a_i covers at least as much as the ball centered in b_i, i.e. $\mathcal{B}_{X,d}(b_i, \delta/2) \subset \mathcal{B}_{X,d}(a_i, \delta)$, c) in none of these cases and we discard it. There are less a_i's than b_i's, and the union of balls of radius δ with centers a_i covers at least as much as the balls of radius $\delta/2$ with centers b_i, and therefore the set of a_i's is a δ-covering of B and of A. □

Lemma A.3.2. *Let (X, d) and (X', d') be two pseudometric sets, and $Y \subset X$, $Y' \subset X'$. If there exists a surjective function $f : Y \to Y'$ which is L-Lipschitz with $L > 0$, i.e. such that*

$$\forall x, y \in Y,\ d'(f(x), f(y)) \leq Ld(x, y),$$

then for all $\delta > 0$ we have

$$\mathcal{N}(d', Y', \delta) \leq \mathcal{N}(d, Y, \delta/L)\ . \tag{A.8}$$

Proof. Define $\delta_2 = \delta/L$, denote $N = \mathcal{N}(d, Y, \delta_2)$, and let $y_i \in Y$, $i = 1, ..., N$ be a δ_2-covering of Y. Consider any $y' \in Y'$. There exists $y \in Y$ such that $f(y) = y'$ since f is surjective. For some $1 \leq i \leq N$ we have $d(y, y_i) \leq \delta_2$, hence we have

$$d'(y', f(y_i)) = d'(f(y), f(y_i)) \leq Ld(y, y_i) \leq L\delta_2 = \delta.$$

Thus $\{f(y_i)\}_{i=1,...,N}$ is an δ-covering of Y', and we have $\mathcal{N}(d', Y', \delta) \leq N$. □

Lemma A.3.3. *Let Y, Z be two subsets of a pseudometric set (X, d) such that the following holds:*

$$\forall z \in Z,\ \exists y \in Y,\ d(z, y) \leq \eta \tag{A.9}$$

where $\eta \geq 0$.

Then for all $\delta > 0$

$$\mathcal{N}(d, Z, 2(\delta + \eta)) \leq \mathcal{N}(d, Y, \delta). \tag{A.10}$$

Proof. Denote $N = \mathcal{N}(d, Y, \delta)$ and let $y_1, ..., y_N \in Y$ be an δ-covering of Y.

For all $z \in Z$, given (A.9) there is a $y \in Y$ such that $d(z, y) \leq \eta$, and subsequently a y_i such that

$$d(z, y_i) \leq d(z, y) + d(y, y_i) \leq \delta + \eta$$

Hence $Z \subset \bigcup_{i=1}^{N} \mathcal{B}_{X,d}(y_i, \delta + \eta)$ and applying Lemma A.3.1 yields

$$\mathcal{N}(d, Z, 2(\delta + \eta)) \leq \mathcal{N}\left(d, \bigcup_{i=1}^{N} \mathcal{B}_{X,d}(y_i, \delta + \eta), \delta + \eta\right) \leq N.$$

□

Lemma A.3.4 ([CS02], Prop. 5)**.** *Let $(X, \|\cdot\|)$ be a Banach space of finite dimension d. Then for any $\delta > 0$, $x \in X$ and $R > 0$ we have*

$$\mathcal{N}\left(\|\cdot\|, \mathcal{B}_{X,\|\cdot\|}(x, R), \delta\right) \leq \max\left(\left(\frac{4R}{\delta}\right)^d, 1\right) \tag{A.11}$$

In all applications we will consider for simplicity that δ is sufficiently small so that $\left(\frac{4R}{\delta}\right)^d \geq 1$.

A.3.2 Extruded Secant set

The covering numbers of the extruded normalized secant set of a set Y can be controlled by those of Y itself when $\eta > 0$. It is based on the following Lemma.

Lemma A.3.5. *Let X be a vector space and consider two subset $Y, Z \subset X$ and two seminorms $\|\cdot\|_a, \|\cdot\|_b$ such that, for some constants $0 < A \leq B < \infty$,*

$$\forall v, v' \in V \text{ where } V = Y \text{ or } V = Z,\ A\|v - v'\|_b \leq \|v - v'\|_a \tag{A.12}$$

$$\forall y \in Y, z \in Z \text{ s.t. } \|y - z\|_b > \eta,\ \|y - z\|_a \leq B\|y - z\|_b \tag{A.13}$$

Consider $\eta > 0$, and a set

$$\mathcal{S} \subset \left\{ \frac{y - z}{\|y - z\|_b} \,\middle|\, y \in Y,\ z \in Z,\ \|y - z\|_b > \eta \right\} \tag{A.14}$$

For any $\delta > 0$ we have

$$\mathcal{N}(\|\cdot\|_a, \mathcal{S}, \delta) \leq \mathcal{N}(\|\cdot\|_a, Y, \delta') \cdot \mathcal{N}(\|\cdot\|_a, Z, \delta') \tag{A.15}$$

with $\delta' = \frac{\delta\eta}{4(1+B/A)}$.

Proof. Define the (semi)norm on $Y \times Z$:

$$\|(y, z) - (y', z')\|_a = \|y - y'\|_a + \|z - z'\|_a$$

and note that we have $\mathcal{N}(\|\cdot\|_a, Y \times Z, \delta) \leq \mathcal{N}(\|\cdot\|_a, Y, \delta/2)\mathcal{N}(\|\cdot\|_a, Z, \delta/2)$. Consider the set:

$$V = \left\{(y,z) \in Y \times Z,\ \frac{y-z}{\|y-z\|_b} \in \mathcal{S}\right\} \subset Y \times Z, \tag{A.16}$$

and the function $f : V \to \mathcal{S}$ defined by $f(y,z) = \frac{y-z}{\|y-z\|_b}$, which is by definition surjective. Let us show that f is Lipschitz continuous for the norm $\|\cdot\|_a$, and conclude with Lemma A.3.2.

For any $(y,z),(y',z') \in V$, we have

$$\begin{aligned}\|f(y,z) - f(y',z')\|_a &= \left\| \frac{y-z}{\|y-z\|_b} - \frac{y'-z'}{\|y'-z'\|_b} \right\|_a, \\ &\leq \left\| \frac{y-z}{\|y-z\|_b} - \frac{y'-z'}{\|y-z\|_b} \right\|_a + \left\| \frac{y'-z'}{\|y-z\|_b} - \frac{y'-z'}{\|y'-z'\|_b} \right\|_a.\end{aligned}$$

Since $\|y-z\|_b > \eta$, the first term is bounded by

$$\frac{1}{\eta}\Big(\|y-y'\|_a + \|z-z'\|_a \Big) = \frac{1}{\eta}\|(y,z)-(y',z')\|_a,$$

while the second term is bounded by

$$\begin{aligned}\|y'-z'\|_a \left| \frac{1}{\|y-z\|_b} - \frac{1}{\|y'-z'\|_b} \right| &\leq \frac{\|y'-z'\|_a}{\|y'-z'\|_b} \frac{1}{\|y-z\|_b} \Big| \|y'-z'\|_b - \|y-z\|_b \Big|, \\ &\overset{(A.13)}{\leq} \frac{B}{\eta} \Big| \|y'-z'\|_b - \|y-z\|_b \Big|, \\ &\leq \frac{B}{\eta}\Big(\|y-y'\|_b + \|z-z'\|_b \Big), \\ &\overset{(A.12)}{\leq} \frac{B}{A\eta}\left(\|y-y'\|_a + \|z-z'\|_a \right), \\ &= \frac{B}{A\eta}\|(y,z)-(y',z')\|_a.\end{aligned}$$

Hence we have

$$\|f(y,z) - f(y',z')\|_a \leq \frac{1+B/A}{\eta}\|(y,z)-(y',z')\|_a.$$

The function f is Lipschitz continuous with constant $L = (1+B/A)/\eta$, and therefore for all $\delta > 0$:

$$\begin{aligned}\mathcal{N}(\|\cdot\|_a, \mathcal{S}, \delta) &\overset{\text{Lem. A.3.2}}{\leq} \mathcal{N}(\|\cdot\|_a, V, \delta/L) \overset{\text{Lem. A.3.1}}{\leq} \mathcal{N}\left(\|\cdot\|_a, Y \times Z, \frac{\delta}{2L}\right) \\ &\leq \mathcal{N}\left(\|\cdot\|_a, Y, \frac{\delta}{4L}\right) \cdot \mathcal{N}\left(\|\cdot\|_a, Z, \frac{\delta}{4L}\right).\end{aligned}$$

□

A.3.3 Mixture set

Let $(X, \|\cdot\|)$ be a (semi)normed vector space and $Y_1, ..., Y_k \subset X$ non-empty subsets of X. Consider $\mathcal{W} \subset \mathbb{R}^k$.

Lemma A.3.6. *For all $\delta > 0$ the set $Y_{\mathcal{W}}^{(k)}$ satisfies*

$$\mathcal{N}\left(\|\cdot\|, Y_{\mathcal{W}}^{(k)}, \delta\right) \leq \mathcal{N}\left(\|\cdot\|_1, \mathcal{W}, \frac{\delta}{2\max_l \operatorname{rad}_{\|\cdot\|}(Y_l)}\right) \cdot \prod_{l=1}^{k} \mathcal{N}\left(\|\cdot\|, Y_l, \frac{\delta}{2\operatorname{rad}_{\|\cdot\|_1}(\mathcal{W})}\right) \tag{A.17}$$

Proof. Consider $\delta > 0$. Denote $\delta_Y = \tau\delta/\operatorname{rad}_{\|\cdot\|_1}(\mathcal{W})$ and $\delta_{\mathcal{W}} = (1-\tau)\delta/\operatorname{rad}_{\|\cdot\|}(Y)$. For $l = 1, ..., k$ denote $N_l = \mathcal{N}(\|\cdot\|, Y_l, \delta_Y)$ and let $\mathcal{C}_l = \{y_{l,1}, ..., y_{l,N_1}\}$ be a δ_Y-covering of Y_l. Similarly, denote $N_{\mathcal{W}} = \mathcal{N}(\|\cdot\|_1, \mathcal{W}, \delta_{\mathcal{W}})$, let $\mathcal{C}_{\mathcal{W}} = \{\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_{N_{\mathcal{W}}}\}$ be a $\delta_{\mathcal{W}}$-covering of $\mathcal{W}$.

Define the set

$$Z = \left\{ \sum_{l=1}^{k} \xi_l y_l \;\middle|\; y_l \in \mathcal{C}_l,\ \boldsymbol{\xi} \in \mathcal{C}_{\mathcal{W}} \right\} \tag{A.18}$$

The cardinality of this set verifies $|Z| \leq |\mathcal{C}_{\mathcal{W}}| \prod_{l=1}^{k} |\mathcal{C}_l| = N_{\mathcal{W}} \prod_{l=1}^{k} N_l$.

Let us show that Z is a δ-covering of $Y_{\mathcal{W}}$. Consider $y = \sum_{l=1}^{k} \xi_l y_l \in Y_{\mathcal{W}}$. For all $l = 1...k$, let $\bar{y}_l \in \mathcal{C}_l$ be a element in $\mathcal{C}_l$ which is closest to y_l, and $\bar{\boldsymbol{\xi}} \in \mathcal{C}_{\mathcal{W}}$ be a vector in $\mathcal{C}_{\mathcal{W}}$ which is closest to $\boldsymbol{\xi}$ for the norm $\|\cdot\|_1$. Denote $\bar{y} = \sum_{l=1}^{k} \bar{\xi}_l \bar{y}_l \in Z$. We have, using $\|y_l - \bar{y}_l\| \leq \delta_Y$ and $\left\|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}\right\|_1 \leq \delta_{\mathcal{W}}$,

$$\begin{aligned}
\|y - \bar{y}\| &= \left\| \sum_{l=1}^{k} \xi_l y_l - \sum_{l=1}^{k} \bar{\xi}_l \bar{y}_l \right\|, \\
&\leq \left\| \sum_{l=1}^{k} \xi_l y_l - \sum_{l=1}^{k} \xi_l \bar{y}_l \right\| + \left\| \sum_{l=1}^{k} \xi_l \bar{y}_l - \sum_{l=1}^{k} \bar{\xi}_l \bar{y}_l \right\|, \\
&\leq \sum_{l=1}^{k} |\xi_l| \, \|y_l - \bar{y}_l\| + \sum_{l=1}^{k} \left|\xi_l - \bar{\xi}_l\right| \|\bar{y}_l\|, \\
&\leq \|\boldsymbol{\xi}\|_1 \, \delta_Y + \left\|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}\right\|_1 \operatorname{rad}_{\|\cdot\|}(Y) \leq \operatorname{rad}_{\|\cdot\|_1}(\mathcal{W}) \, \delta_Y + \delta_{\mathcal{W}} \operatorname{rad}_{\|\cdot\|}(Y) = \delta,
\end{aligned} \tag{A.19}$$

and Z is indeed a δ-covering of $Y_{\mathcal{W}}^{(k)}$. Therefore, we have the bound (for all τ)

$$\mathcal{N}\left(\|\cdot\|, Y_{\mathcal{W}}^{(k)}, \delta\right) \leq |Z| \leq N_{\mathcal{W}} \prod_{l=1}^{k} N_l$$

□

Appendix B

Proof of Chapter 3

In this appendix we group the proofs of the results found in Chapter 3.

B.1 Proof of Lemma 3.3.3

Proof of Lemma 3.3.3. This is a particular case of Lemma A.3.6, which handles more general mixture sets. We apply it with the weight set $\mathcal{W} = \mathbb{S}^{k-1} \subset \mathcal{B}_{\mathbb{R}^k, \|\cdot\|_1}(0,1)$ the $k-1$ dimensional simplex and for $l = 1...k$ the sets $Y_l = \mathfrak{T}$. We obtain:

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{S}_k(\mathfrak{T}), \delta\right) \leq \mathcal{N}\left(\|\cdot\|_1, \mathbb{S}^{k-1}, \frac{\delta}{2B_{\mathcal{F}_R}}\right) \left(\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{T}, \delta/2\right)\right)^k$$

We then use Lemma A.3.1 to bound the covering numbers of $\mathbb{S}^{k-1}$ by those of $\mathcal{B}_{\mathbb{R}^k, \|\cdot\|_1}(0,1)$:

$$\mathcal{N}\left(\|\cdot\|_1, \mathbb{S}^{k-1}, \delta\right) \leq \mathcal{N}\left(\|\cdot\|_1, \mathcal{B}_{\mathbb{R}^k, \|\cdot\|_1}(0,1), \delta/2\right),$$

then Lemma A.3.4 to bound the covering numbers of $\mathcal{B}_{\mathbb{R}^k, \|\cdot\|_1}(0,1)$:

$$\mathcal{N}\left(\|\cdot\|_1, \mathcal{B}_{\mathbb{R}^k, \|\cdot\|_1}(0,1), \delta\right) \leq \max\left(\left(\frac{4}{\delta}\right)^k, 1\right)$$

such that $\mathcal{N}\left(\|\cdot\|_1, \mathbb{S}^{k-1}, \frac{\delta}{2B_{\mathcal{F}_R}}\right) \leq \max\left(\left(\frac{16B_{\mathcal{F}_R}}{\delta}\right)^k, 1\right) = \left(\frac{16B_{\mathcal{F}_R}}{\delta}\right)^k$ since $\delta \leq 16B_{\mathcal{F}_R}$. We obtain the desired result. □

B.2 Gaussian distributions

Proof of Lemma 3.3.4. We begin by Pinsker's inequality [FHT03]:

$$\|\pi_1 - \pi_2\|_{\mathrm{TV}} \leq \sqrt{2D_{\mathrm{KL}}(\pi_1||\pi_2)}, \tag{B.1}$$

where D_{KL} is the Kullback-Leibler divergence. By symmetry, we get

$$\|\pi_1 - \pi_2\|_{\mathrm{TV}}^2 \leq D_{\mathrm{KL}}(\pi_1||\pi_2) + D_{\mathrm{KL}}(\pi_2||\pi_1) \tag{B.2}$$

The Kullback-Leibler divergence has a closed form expression in the case of multivariate Gaussians [Duc07]:

$$D_{\mathrm{KL}}(\pi_1||\pi_2) = \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} + \mathrm{tr}\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right) - d + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]. \tag{B.3}$$

and therefore

$$\|\pi_1 - \pi_2\|_{\mathrm{TV}} \leq \left(\frac{\mathrm{tr}\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right) + \mathrm{tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2\right) - 2d}{2} + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2_{\left(\frac{\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1}}{2}\right)^{-1}} \right)^{\frac{1}{2}}$$
$$\leq \left(\frac{\mathrm{tr}\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right) + \mathrm{tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2\right) - 2d}{2} \right)^{\frac{1}{2}} + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\left(\frac{\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1}}{2}\right)^{-1}}$$

We have

$$\mathrm{tr}\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right) + \mathrm{tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2\right) - 2d = \mathrm{tr}\left(\left(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}\right)\left(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\right)\right) \leq \left\|\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}\right\|_{\mathrm{F}} \|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_{\mathrm{F}}$$

by Cauchy-Schwartz inequality. □

Proof of Lemma 3.3.5. Consider $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{T} = \mathcal{D}_{\boldsymbol{\mu}} \times \mathcal{D}_{\boldsymbol{\sigma}}$. Denote $\sigma^2_{\min}$ the minimum eigenvalue of $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$. Note that we have

$$\left\|\mathrm{diag}\left(\boldsymbol{\sigma}_1\right)^{-1} - \mathrm{diag}\left(\boldsymbol{\sigma}_2\right)^{-1}\right\|_{\mathrm{F}} = \left(\sum_{i=1}^{d} \left(\frac{1}{\sigma^2_{1,i}} - \frac{1}{\sigma^2_{2,i}} \right)^2 \right)^{\frac{1}{2}}$$
$$= \left(\sum_{i=1}^{d} \left(\frac{\sigma^2_{2,i} - \sigma^2_{1,i}}{\sigma^2_{2,i}\sigma^2_{1,i}} \right)^2 \right)^{\frac{1}{2}} \leq \frac{1}{\sigma^4_{\min}} \|\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2\|_{\mathrm{F}}$$

By Lemma 3.1.5 we have $\|\cdot\|_{\mathcal{F}_R} \leq B_{\mathcal{F}_R} \|\cdot\|_{\mathrm{TV}}$, hence by Lemma 3.3.4 we have

$$\|\pi_{\boldsymbol{\theta}_1} - \pi_{\boldsymbol{\theta}_2}\|_{\mathcal{F}_R} \leq B_{\mathcal{F}_R} \sqrt{\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^\top \left(\frac{\mathrm{diag}\left(\boldsymbol{\sigma}_1\right)^{-1} + \mathrm{diag}\left(\boldsymbol{\sigma}_2\right)^{-1}}{2} \right) \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)}$$
$$+ \frac{B_{\mathcal{F}_R}}{\sigma^2_{\min}\sqrt{2}} \|\mathrm{diag}\left(\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2\right)\|_{\mathrm{F}}$$
$$\leq \frac{B_{\mathcal{F}_R}}{\sigma_{\min}} \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2 + \frac{B_{\mathcal{F}_R}}{\sigma^2_{\min}\sqrt{2}} \|\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2\|_2 \ . \tag{B.4}$$

Consider $\delta > 0$. Define $\mathcal{C}_{\boldsymbol{\mu}}$ a $\frac{\delta\sigma_{\min}}{2B_{\mathcal{F}_R}}$-covering of $\mathcal{D}_{\boldsymbol{\mu}} = \mathcal{B}_{\mathbb{R}^d, \|\cdot\|_2}(0, R_{\boldsymbol{\mu}})$ and $\mathcal{C}_{\boldsymbol{\sigma}}$ a $\frac{\delta\sigma^2_{\min}}{\sqrt{2}B_{\mathcal{F}_R}}$-covering of $\mathcal{D}_{\boldsymbol{\sigma}} \subset \mathcal{B}_{\mathbb{R}^d, \|\cdot\|_2}(0, R_{\boldsymbol{\sigma}})$, both with respect to the Euclidean norm. Given (B.4), for all $\boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\mu}} \times \mathcal{D}_{\boldsymbol{\sigma}}$ there is $\bar{\boldsymbol{\theta}} \in \mathcal{C}_{\boldsymbol{\mu}} \times \mathcal{C}_{\boldsymbol{\sigma}}$ such that $\|\pi_{\boldsymbol{\theta}} - \pi_{\bar{\boldsymbol{\theta}}}\|_{\mathcal{F}_R} \leq \delta$. Thus

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{T}, \delta\right) \leq |\mathcal{C}_{\boldsymbol{\mu}}| \cdot |\mathcal{C}_{\boldsymbol{\sigma}}| \ .$$

Then, we conclude with

$$|\mathcal{C}_{\boldsymbol{\mu}}| \overset{\text{Lem. A.3.4}}{\leq} \max\left(\left(\frac{8B_{\mathcal{F}_R}R_{\boldsymbol{\mu}}}{\sigma_{\min}\delta} \right)^d, 1 \right)$$

and

$$|\mathcal{C}_{\boldsymbol{\sigma}}| \overset{\text{Lem. A.3.1}}{\leq} \mathcal{N}\left(\|\cdot\|_2, \mathcal{B}_{\mathbb{R}^d, \|\cdot\|_2}(0, R_{\boldsymbol{\sigma}}), \frac{\delta\sigma^2_{\min}}{2\sqrt{2}B_{\mathcal{F}_R}} \right) \overset{\text{Lem. A.3.4}}{\leq} \max\left(\left(\frac{8\sqrt{2}B_{\mathcal{F}_R}R_{\boldsymbol{\sigma}}}{\sigma^2_{\min}\delta} \right)^d, 1 \right)$$

□

B.3 Stable distributions

Proof of Lemma 3.3.8. For $\pi_{\boldsymbol{\theta}_1}, \pi_{\boldsymbol{\theta}_2} \in \mathfrak{T}$, denote $\boldsymbol{\Sigma}_j = \operatorname{diag}(\boldsymbol{\sigma}_j)$. Given (3.41) we have

$$\begin{aligned}\|\pi_{\boldsymbol{\theta}_1} - \pi_{\boldsymbol{\theta}_2}\|_{\mathcal{F}_R} &= \sup_{\boldsymbol{\omega}\in\mathbb{R}^d} \left| e^{\mathrm{i}\boldsymbol{\mu}_1^\top \boldsymbol{\omega}} e^{-\left(\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_1 \boldsymbol{\omega}\right)^{\alpha_1/2}} - e^{\mathrm{i}\boldsymbol{\mu}_2^\top \boldsymbol{\omega}} e^{-\left(\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_2 \boldsymbol{\omega}\right)^{\alpha_2/2}} \right| \\ &\leq \sup_{\boldsymbol{\omega}\in\mathbb{R}^d,\ \boldsymbol{\omega}\neq 0} e^{-a_1(\boldsymbol{\omega})} \left| e^{\mathrm{i}\boldsymbol{\mu}_1^\top \boldsymbol{\omega}} - e^{\mathrm{i}\boldsymbol{\mu}_2^\top \boldsymbol{\omega}} \right| + \sup_{\boldsymbol{\omega}\in\mathbb{R}^d,\ \boldsymbol{\omega}\neq 0} \left| e^{-a_1(\boldsymbol{\omega})} - e^{-a_2(\boldsymbol{\omega})} \right|\end{aligned}$$

where $a_j(\boldsymbol{\omega}) = \left(\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_j \boldsymbol{\omega}\right)^{\alpha_j/2}$ for $j = 1, 2$. Since $\psi_{\pi_{\boldsymbol{\theta}_1}}(0) - \psi_{\pi_{\boldsymbol{\theta}_2}}(0) = 0$ we consider only $\boldsymbol{\omega} \neq 0$ in the following for simplicity.

By a simple Taylor expansion the first term is bounded by

$$\sup_{\boldsymbol{\omega}} e^{-a_1(\boldsymbol{\omega})} \left| e^{\mathrm{i}\boldsymbol{\mu}_1^\top \boldsymbol{\omega}} - e^{\mathrm{i}\boldsymbol{\mu}_2^\top \boldsymbol{\omega}} \right| \leq \sup_{\boldsymbol{\omega}} e^{-\left(\frac{\sigma_{\min}^2}{2}\|\boldsymbol{\omega}\|_2^2\right)^{\alpha_1/2}} \|\boldsymbol{\omega}\|_2 \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2 \leq L_1 \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2 \tag{B.5}$$

where

$$L_1 = \sup_{x>0,\ \alpha\in[\alpha_{\min},2]} x e^{-\left(\frac{\sigma_{\min}^2}{2}x^2\right)^{\alpha/2}}$$

Without lost of generality, we suppose that $a_1 \geq a_2$ and bound the second term

$$\sup_{\boldsymbol{\omega}} \left| e^{-a_1} - e^{-a_2} \right| = \sup_{\boldsymbol{\omega}} e^{-a_2}(e^{-(a_1-a_2)} - 1) \leq \sup_{\boldsymbol{\omega}} e^{-\left(\frac{\sigma_{\min}^2}{2}\|\boldsymbol{\omega}\|_2^2\right)^{\alpha_1/2}} |a_1 - a_2|$$

We decompose

$$|a_1 - a_2| \leq \left| \left(\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_1 \boldsymbol{\omega}\right)^{\alpha_1/2} - \left(\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_2 \boldsymbol{\omega}\right)^{\alpha_1/2} \right| + \left| \left(\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_2 \boldsymbol{\omega}\right)^{\alpha_1/2} - \left(\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_2 \boldsymbol{\omega}\right)^{\alpha_2/2} \right|$$

and by Taylor expansion we have

$$\begin{aligned}\left| \left(\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_1 \boldsymbol{\omega}\right)^{\alpha_1/2} - \left(\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_2 \boldsymbol{\omega}\right)^{\alpha_1/2} \right| &\leq \frac{\alpha_1}{2} \left(\frac{\sigma_{\min}^2}{2}\|\boldsymbol{\omega}\|_2^2\right)^{\alpha_1/2-1} \left| \frac{1}{2}\boldsymbol{\omega}^\top \operatorname{diag}(\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2)\boldsymbol{\omega} \right| \\ &\leq \frac{\alpha_1}{2} \left(\frac{\sigma_{\min}^2}{2}\|\boldsymbol{\omega}\|_2^2\right)^{\alpha_1/2-1} \frac{1}{2}\|\boldsymbol{\omega}\|_4^2 \|\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2\|_2 \\ &\leq \frac{\alpha_1}{2(\sigma_{\min}^2)^{1-\alpha_1/2}} \left(\frac{\|\boldsymbol{\omega}\|_2}{\sqrt{2}}\right)^{\alpha_1} \|\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2\|_2\end{aligned}$$

since $\|\cdot\|_4 \leq \|\cdot\|_2$. Finally (recall that we consider only $\boldsymbol{\omega} \neq 0$), again by Taylor expansion

$$\begin{aligned}\left| \left(\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_2 \boldsymbol{\omega}\right)^{\alpha_1/2} - \left(\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_2 \boldsymbol{\omega}\right)^{\alpha_2/2} \right| &= \left| \log \frac{\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_2 \boldsymbol{\omega}}{2} \right| \left(\frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\Sigma}_2 \boldsymbol{\omega}\right)^{\alpha'/2} \frac{1}{2}|\alpha_1 - \alpha_2| \\ &\leq \left| \log\left(\frac{R_{\boldsymbol{\sigma}}}{2}\|\boldsymbol{\omega}\|_4^2\right) \right| \left(\frac{R_{\boldsymbol{\sigma}}}{2}\|\boldsymbol{\omega}\|_4^2\right)^{\alpha'/2} \frac{1}{2}|\alpha_1 - \alpha_2| \\ &= \left| \log\left(\sqrt{R_{\boldsymbol{\sigma}}}\frac{\|\boldsymbol{\omega}\|_2}{\sqrt{2}}\right) \right| \left(\sqrt{R_{\boldsymbol{\sigma}}}\frac{\|\boldsymbol{\omega}\|_2}{\sqrt{2}}\right)^{\alpha'} |\alpha_1 - \alpha_2|\end{aligned}$$

where α' is between α_1 and α_2.

Therefore we have

$$\sup_{\boldsymbol{\omega}} \left| e^{-a_1} - e^{-a_2} \right| \leq L_2 \|\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_2\|_2 + L_3 |\alpha_1 - \alpha_2| \tag{B.6}$$

where

$$L_2 = \sup_{x>0,\ \alpha\in[\alpha_{\min},2]} \frac{\alpha}{2(\sigma_{\min}^2)^{1-\alpha/2}} x^{\alpha} e^{-(x\sigma_{\min})^{\alpha}}$$

$$L_3 = \sup_{x>0,\ \alpha,\alpha'\in[\alpha_{\min},2]} \left|\log\left(\sqrt{R_{\boldsymbol{\sigma}}}x\right)\right| \left(\sqrt{R_{\boldsymbol{\sigma}}}x\right)^{\alpha'} e^{-(x\sigma_{\min})^{\alpha}}$$

which is the desired result. □

Proof of Lemma 3.3.9. Consider $\delta > 0$. Define $\mathcal{C}_{\boldsymbol{\mu}}$ a $\delta/(3L_1)$-covering of $\mathcal{D}_{\boldsymbol{\mu}} = \mathcal{B}_{\mathbb{R}^d,\|\cdot\|_2}(0, R_{\boldsymbol{\mu}})$ and $\mathcal{C}_{\boldsymbol{\sigma}}$ a $\delta/(3L_2)$-covering of $\mathcal{D}_{\boldsymbol{\sigma}} \subset \mathcal{B}_{\mathbb{R}^d,\|\cdot\|_2}(0, R_{\boldsymbol{\sigma}})$, both with respect to the Euclidean norm, and $\mathcal{C}_{\alpha}$ a $\delta/(3L_3)$-covering of $[\alpha_{\min}, 2]$ for $|\cdot|$. Given (B.4), for all $\boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\mu}} \times \mathcal{D}_{\boldsymbol{\sigma}} \times \mathcal{D}_{\alpha}$ there is $\bar{\boldsymbol{\theta}} \in \mathcal{C}_{\boldsymbol{\mu}} \times \mathcal{C}_{\boldsymbol{\sigma}} \times \mathcal{C}_{\alpha}$ such that $\|\pi_{\boldsymbol{\theta}} - \pi_{\bar{\boldsymbol{\theta}}}\|_{\mathcal{F}_R} \leq \delta$. Thus

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{T}, \delta\right) \leq |\mathcal{C}_{\boldsymbol{\mu}}| \cdot |\mathcal{C}_{\boldsymbol{\sigma}}| \cdot |\mathcal{C}_{\alpha}| \ .$$

Then, we conclude with the same arguments that Lemma 3.3.5:

$$|\mathcal{C}_{\boldsymbol{\mu}}| \leq \max\left(\left(\frac{12R_{\boldsymbol{\mu}}L_1}{\delta}\right)^d, 1\right), |\mathcal{C}_{\boldsymbol{\sigma}}| \leq \max\left(\left(\frac{12R_{\boldsymbol{\sigma}}L_2}{\delta}\right)^d, 1\right), |\mathcal{C}_{\alpha}| \leq \max\left(\frac{3L_3(2-\alpha_{\min})}{\delta}, 1\right) .$$

where the last inequality comes simply from the covering of a one-dimensional interval in $\mathbb{R}$. □

Appendix C

Generic Mixture Models

In this section, we give the proofs of Section 6.2 Chapter 6 on generic mixture models with ε-separation.

C.1 Proof of Lemma 6.2.4

To prove, Lemma 6.2.4, we will need the following intermediary results.

Lemma C.1.1. *Assume $h : \mathbb{R}_+ \to \mathbb{R}$ is differentiable and that $h'(x)$ is C-Lipschitz. Then for any $\delta, \nu \geq 0$:*

$$|h(0) - h(\delta) - h(\nu) + h(\delta + \nu)| \leq 2\delta\nu C\,.$$

Proof. Assume without loss of generality that $\delta = \min(\delta, \nu)$. Write $h(\delta) - h(0) = h'(c_1)\delta$ for some $c_1 \in [0, \delta]$ and $h(\delta + \nu) - h(\nu) = h'(c_2)\delta$ for some $c_2 \in [\nu, \delta + \nu]$, thus

$$|h(0) - h(\delta) - h(\nu) + h(\delta + \nu)| = |\delta(h'(c_2) - h'(c_1))| \leq C\delta\,|c_2 - c_1|\,,$$

bounded in absolute value by $2\delta\nu C$, since $|c_1 - c_2| \leq \delta + \nu \leq 2\nu$. □

Lemma C.1.2. *Let $\mu = \pi_{\boldsymbol{\theta}_1} - \pi_{\boldsymbol{\theta}_2}$ and $\mu' = \pi_{\boldsymbol{\theta}_3} - \pi_{\boldsymbol{\theta}_4}$ be two dipoles that are 1-separated, denote $d_{ij} = \varrho(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$. Consider $K \in \mathcal{E}(A, B, C, \gamma)$. Then we have:*

$$K(d_{13}) - f(d_{23}) - K(d_{14}) + f(d_{24}) \leq 2(B + C)d_{12}d_{34} \tag{C.1}$$

Proof. Assume without loss of generality that $d_{13} = \min(d_{13}, d_{23}, d_{14}, d_{24})$ and write

$$\begin{aligned} |K(d_{13}) - K(d_{23}) - K(d_{14}) + K(d_{24})| = {} & |K(d_{13}) - K(d_{23}) - K(d_{14}) + K(d_{23} + d_{14} - d_{13})| \\ & + |K(d_{24}) - K(d_{23} + d_{14} - d_{13})|\,. \end{aligned} \tag{C.2}$$

To bound the first term in (C.2) in absolute value, since we assumed without loss of generality that $d_{13} = \min(d_{13}, d_{23}, d_{14}, d_{24})$, we can apply Lemma C.1.2 with $h(x) := K(d_{13} + x)$, $\delta := d_{23} - d_{13} \geq 0$, $\nu := d_{14} - d_{13} \geq 0$, leading to

$$|K(d_{13}) - K(d_{23}) - K(d_{14}) + K(d_{23} + d_{14} - d_{13})| \leq 2C\,|(d_{23} - d_{13})(d_{14} - d_{13})| \leq 2Cd_{12}d_{34}\,.$$

To bound the second term in (C.2), put $g(u) := K(\sqrt{u})$, and note that $g'(u) = K'(\sqrt{u})/2\sqrt{u}$, implying since $K \in \mathcal{E}(A, B, C, \gamma)$ that $g'(u^2) \leq B/2$ for $u \geq 1$. Since $d_{23} + d_{14} - d_{13} \geq d_{23} \geq 1$ and $d_{24} \geq 1$, we first write

$$\begin{aligned} (K(d_{24}) - K(d_{23} + d_{14} - d_{13})) &= \left(g(d_{24}^2) - g((d_{23} + d_{14} - d_{13})^2)\right) \\ &\leq \tfrac{B}{2}\left|d_{24}^2 - (d_{23} + d_{14} - d_{13})^2\right|. \end{aligned}$$

Now, it holds

$$d_{24}^2 - (d_{23} + d_{14} - d_{13})^2 = d_{24}^2 - d_{23}^2 - d_{14}^2 + d_{13}^2 - 2(d_{13} - d_{23})(d_{13} - d_{14})\,,$$

the last product is bounded in absolute value by $2d_{12}d_{34}$, and it is easy to check by expanding the squared norms $d_{ij}^2 = \|\chi(\boldsymbol{\theta}_i) - \chi(\boldsymbol{\theta}_j)\|_2^2$ that

$$\left|d_{24}^2 - d_{23}^2 - d_{14}^2 + d_{13}^2\right| = 2\left|\langle \chi(\boldsymbol{\theta}_1) - \chi(\boldsymbol{\theta}_2), \chi(\boldsymbol{\theta}_3) - \chi(\boldsymbol{\theta}_4)\rangle\right| \leq 2d_{12}d_{34}\,.$$

Gathering everything we get the desired result. □

We can now prove Lemma 6.2.4.

Proof of Lemma 6.2.4. Denote $\mu = \xi_1\pi_{\boldsymbol{\theta}_1} - \xi_2\pi_{\boldsymbol{\theta}_2}$ and $\mu' = \xi_3\pi_{\boldsymbol{\theta}_3} - \xi_4\pi_{\boldsymbol{\theta}_4}$ two dipoles that are 1-separated, and without lost of generality suppose that $\xi_1 = \xi_3 = 1$, $\xi_2 = a \leq 1$, $\xi_4 = b \leq 1$. Our goal is to prove that $\frac{|\kappa(\mu,\mu')|}{\|\mu\|_\kappa \|\mu'\|_\kappa}$ is bounded.

Denote $d_{ij} = \varrho(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ and $K_{ij} = K(d_{ij}) = \kappa(\pi_{\boldsymbol{\theta}_i}, \pi_{\boldsymbol{\theta}_j})$. We have

$$\begin{aligned}\frac{|\kappa(\mu,\mu')|}{\|\mu\|_\kappa \|\mu'\|_\kappa} &= \frac{|K_{13} - aK_{23} - bK_{14} + abK_{24}|}{\sqrt{1 - 2aK_{12} + a^2}\sqrt{1 - 2bK_{34} + b^2}} \\ &\leq \frac{|K_{13} - K_{23} - K_{14} + K_{24}| + |(1-a)(K_{23} - K_{24})| + |(1-b)(K_{14} - K_{24})| + |(a-1)(b-1)K_{24}|}{\sqrt{(1-a)^2 + 2a(1-K_{12})}\sqrt{(1-b)^2 + 2b(1-K_{34})}}\end{aligned}$$

Applying Lemma C.1.2 we get:

$$|K_{13} - K_{23} - K_{14} + K_{24}| \leq 2(B+C)d_{12}d_{34}$$

as well as

$$\begin{aligned}|K_{23} - K_{24}| &\leq Bd_{34} \quad \text{(since } d_{23} \geq 1 \text{ and } d_{24} \geq 1\text{)} \\ |K_{14} - K_{24}| &\leq Bd_{12} \quad \text{(since } d_{14} \geq 1\text{)} \\ |K_{24}| &\leq A \\ 2(1 - K_{12}) &\geq \gamma d_{12}^2 \quad \text{(since } d_{12} \leq 1\text{)} \\ 2(1 - K_{34}) &\geq \gamma d_{34}^2 \quad \text{(since } d_{34} \leq 1\text{)}\end{aligned}$$

Therefore, if we denote $D = \max(2(B+C), A)$, we have

$$\begin{aligned}\frac{|\kappa(\mu,\mu')|}{\|\mu\|_\kappa \|\mu'\|_\kappa} &\leq D \cdot \frac{d_{12}d_{34} + (1-a)d_{34} + (1-b)d_{12} + (1-a)(1-b)}{\sqrt{(1-a)^2 + a\gamma d_{12}^2}\sqrt{(1-b)^2 + b\gamma d_{34}^2}} \\ &\leq D \cdot \frac{d_{12} + 1 - a}{\sqrt{(1-a)^2 + a\gamma d_{12}^2}} \cdot \frac{d_{34} + 1 - b}{\sqrt{(1-b)^2 + b\gamma d_{34}^2}} \\ &\leq \frac{D}{\gamma} g(1-a, d_{12}) g(1-b, d_{34})\end{aligned}$$

where g is defined on $[0,1]^2$ by $g(x,y) = \frac{x+y}{\sqrt{x^2 + (1-x)y^2}}$. We have

$$\begin{aligned}g(x,y) = \frac{x+y}{\sqrt{x^2 + (1-x)y^2}} &\leq \sqrt{2} \cdot \frac{x+y}{x + \sqrt{(1-x)y}} \leq \sqrt{2} \cdot \frac{x+y}{x + (1-x)y} \\ &\leq \sqrt{2}\left(1 + \frac{xy}{x + y - xy}\right) \leq \sqrt{2}\left(1 + \frac{1}{1/y + 1/x - 1}\right) \leq 2\sqrt{2}\end{aligned}$$

Gathering everything, we have

$$\frac{|\kappa(\mu,\mu')|}{\|\mu\|_\kappa \|\mu'\|_\kappa} \leq \frac{8D}{\gamma}$$

□

C.2 Admissibility, compatibility

Proof of Lemma 6.2.7. Let $\mu = \xi_1 \pi_{\boldsymbol{\theta}_1} - \xi_2 \pi_{\boldsymbol{\theta}_2}$ be a dipole, with $\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}') \leq 1$. Let $\ell \in \mathcal{Q}(D, L)$ be any function.

Writing $\mathbf{w} = [\xi, \xi']^T$, $\mathbf{K}$ a 2×2 matrix such that $K_{ij} = \kappa(\pi_{\boldsymbol{\theta}_1}, \pi_{\boldsymbol{\theta}_2}) = f(d_{12})$ where $d_{12} = \varrho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, and $\mathbf{L}$ a matrix with $L_{ij} = \ell_i \overline{\ell_j}$ with $\ell_i = \langle \ell, \pi_{\boldsymbol{\theta}_i} \rangle$. For any $W \geq 0$ we have

$$W^2 \|\mu\|_\kappa^2 - |\langle \ell, \mu \rangle|^2 = \mathbf{w}^T (W^2 \mathbf{K} - \mathbf{L}) \mathbf{w}.$$

Therefore it is sufficient to prove that there is W, that does not depend on the choice of function ℓ, such that $\mathbf{Q} = W^2 \mathbf{K} - \mathbf{L}$ is a positive semi-definite matrix. It is the case if its trace and determinant are non-negative. We have $\mathrm{tr}(\mathbf{Q}) = 2W^2 - |\ell_1|^2 - |\ell_2|^2 \geq 2(W^2 - D^2)$ since $\ell \in \mathcal{Q}(D, L)$. A sufficient condition for $\mathrm{tr}(\mathbf{Q}) \geq 0$ is therefore

$$W \geq D \tag{C.3}$$

Then, we have:

$$\begin{aligned}
\det(\mathbf{Q}) &= \left(W^2 - |\ell_1|^2\right)\left(W^2 - |\ell_2|^2\right) - \left|W^2 f(d_{12}) - \ell_1 \overline{\ell_2}\right|^2 \\
&= W^4 - W^2\left(|\ell_1|^2 + |\ell_2|^2\right) + |\ell_1|^2 |\ell_2|^2 - \left(W^2 f(d_{12}) - \mathrm{Re}\left(\ell_1 \overline{\ell_2}\right)\right)^2 - \left(\mathrm{Im}\left(\ell_1 \overline{\ell_2}\right)\right)^2.
\end{aligned}$$

Using $|\ell_1|^2 |\ell_2|^2 = \frac{1}{4}\left(\left(|\ell_1|^2 + |\ell_2|^2\right)^2 - \left(|\ell_1|^2 - |\ell_2|^2\right)^2\right)$, we get

$$\begin{aligned}
\det(\mathbf{Q}) = &\left(W^2 - \frac{1}{2}\left(|\ell_1|^2 + |\ell_2|^2\right)\right)^2 - \left(W^2 f(d_{12}) - \mathrm{Re}\left(\ell_1 \overline{\ell_2}\right)\right)^2 \\
&- \frac{1}{4}\left[\left(|\ell_1|^2 - |\ell_2|^2\right)^2 + 4\left(\mathrm{Im}\left(\ell_1 \overline{\ell_2}\right)\right)^2\right].
\end{aligned}$$

On the one hand, we have

$$\begin{aligned}
\left(W^2 - \frac{1}{2}\left(|\ell_1|^2 + |\ell_2|^2\right)\right)^2 - \left(W^2 f(d_{12}) - \mathrm{Re}\left(\ell_1 \overline{\ell_2}\right)\right)^2 &= \left(W^2 - \frac{1}{2}\left(|\ell_1|^2 + |\ell_2|^2\right) - W^2 f(d_{12}) + \mathrm{Re}\left(\ell_1 \overline{\ell_2}\right)\right) \\
&\quad \times \left(W^2 - \frac{1}{2}\left(|\ell_1|^2 + |\ell_2|^2\right) + W^2 f(d_{12}) - \mathrm{Re}\left(\ell_1 \overline{\ell_2}\right)\right) \\
&\geq \left(W^2\left(1 - f(d_{12})\right) - \frac{1}{2}|\ell_1 - \ell_2|^2\right)\left(2W^2 - 4D^2\right) \\
&\geq d_{12}^2\left(W^2\gamma - L^2\right)\left(W^2 - 2D^2\right)
\end{aligned}$$

given the hypotheses on f, ℓ and the fact that $d_{12} \leq 1$. On the other hand,

$$\begin{aligned}
\left(|\ell_1|^2 - |\ell_2|^2\right)^2 + 4\left(\mathrm{Im}\left(\ell_1 \overline{\ell_2}\right)\right)^2 &= \left(|\ell_1|^2 - |\ell_2|^2\right)^2 + 4|\ell_1|^2 |\ell_2|^2 - 4\left(\mathrm{Re}\left(\ell_1 \overline{\ell_2}\right)\right)^2 \\
&= \left(|\ell_1|^2 + |\ell_2|^2\right)^2 - 4\left(\mathrm{Re}\left(\ell_1 \overline{\ell_2}\right)\right)^2 \\
&= \left(|\ell_1|^2 + |\ell_2|^2 - 2\mathrm{Re}\left(\ell_1 \overline{\ell_2}\right)\right)\left(|\ell_1|^2 + |\ell_2|^2 + 2\mathrm{Re}\left(\ell_1 \overline{\ell_2}\right)\right) \\
&= |\ell_1 - \ell_2|^2 |\ell_1 + \ell_2|^2 \leq 4D^2 L^2 d_{12}^2.
\end{aligned}$$

Gathering everything, we have

$$\det(\mathbf{Q}) \geq d_{12}^2\left(\left(W^2\gamma - L^2\right)\left(W^2 - 2D^2\right) - D^2 L^2\right) \geq d_{12}^2 W^2 \gamma\left(W^2 - L^2/\gamma - 2D^2\right)$$

and therefore it is sufficient that

$$W \geq \left(L^2/\gamma + 2D^2\right)^{\frac{1}{2}} \tag{C.4}$$

□

C.3 Covering numbers of the secant set

Proof of Lemma 6.2.9. Recall that $\mathfrak{S} \subset \mathfrak{S}_{k,2,\varrho}(\mathfrak{T})$. Consider $\pi, \pi' \in \mathfrak{S}$. We can decompose

$$\pi - \pi' = \sum_{l=1}^{2k} \mu_l \tag{C.5}$$

where:

- for $l = 1, ..., k$, if there is an index $p \leq k$ such that $\varrho(\boldsymbol{\theta}_l, \boldsymbol{\theta}'_p) \leq 1$, we put $\mu_l = \xi_l \pi_{\boldsymbol{\theta}_l} - \xi'_p \pi_{\boldsymbol{\theta}'_p}$, if not we put $\mu_l = \xi_l \pi_{\boldsymbol{\theta}_l}$
- for $l = k+1, ..., 2k$, if $\pi_{\boldsymbol{\theta}'_{l-k}}$ is already selected as part of a dipole we put $\mu_l = 0$, if not we put $\mu_l = -\xi'_{l-k} \pi_{\boldsymbol{\theta}'_{l-k}}$.

Overall, the μ_l are 1-dipoles that are pairwise 1-separated.

Let $S \subset \{1, ..., 2k\}$ be the set of indexes such that $\mu_l \neq 0$. Since the kernel is characteristic, for $l \in S$ we have $\|\mu_l\|_\kappa \neq 0$. We have

$$\frac{\pi - \pi'}{\|\pi - \pi'\|_\kappa} = \frac{\sum_{l \in S} \mu_l}{\left\|\sum_{l \in S} \mu\right\|_\kappa} = \sum_{l \in S} \frac{\|\mu_l\|_\kappa}{\left\|\sum_{l \in S} \mu_l\right\|_\kappa} \cdot \frac{\mu_l}{\|\mu_l\|_\kappa} = \sum_{l=1}^{2k} \alpha_l \nu_l$$

where $\alpha_l = 0$ for all $l \notin S$ and $\alpha_l = \frac{\|\mu_l\|_\kappa}{\left\|\sum_{l \in S} \mu_l\right\|_\kappa}$ otherwise. The ν_l are defined:

- for $l = 1, ..., k$, if $l \in S$ then $\nu_l = \mu_l / \|\mu_l\|_\kappa \in \mathcal{D}(\xi_l \pi_{\boldsymbol{\theta}_l})$, else ν_l is any distribution in $\mathcal{D}(\xi_l \pi_{\boldsymbol{\theta}_l})$ (which has no influence since $\alpha_l = 0$ in that case);
- for $l = k+1, ..., 2k$, if $l \in S$ then $\nu_l = \xi'_{l-k} \pi_{\boldsymbol{\theta}'_{l-k}} / \left\|\xi'_{l-k} \pi_{\boldsymbol{\theta}'_{l-k}}\right\|_\kappa = \pi_{\boldsymbol{\theta}'_{l-k}}$ since $\|\pi_{\boldsymbol{\theta}}\|_\kappa = \sqrt{K(0)} = 1$, else ν_l is any distribution in $\mathfrak{T}$.

By Lemma 6.2.5 we have $\sum_{l=1}^{2k} \alpha_l^2 = \frac{\sum_{l \in S} \|\mu_l\|_\kappa^2}{\left\|\sum_{l \in S} \mu_l\right\|_\kappa^2} \leq 4$ from which $\|\boldsymbol{\alpha}\|_1 \leq 2\sqrt{2k} = r$ which proves the following inclusion:

$$\mathcal{S}^0(\pi, \mathfrak{S}) \subset \left\{ \sum_{l=1}^{k} \alpha_l \mu_l + \sum_{l=1}^{k} \alpha_{l+k} \pi_l \;\middle|\; \mu_l \in \mathcal{D}(\xi_l \pi_{\boldsymbol{\theta}_l}),\ \pi_l \in \mathfrak{T}, \|\boldsymbol{\alpha}\|_1 \leq r \right\} \tag{C.6}$$

We then apply Lemma A.3.6 to bound the covering numbers of this $2k$-mixture set, with $\mathcal{W} = \mathcal{B}_{\mathbb{R}^{2k}, \|\cdot\|_1}(0, r)$, using Lemma A.3.4 to express the covering numbers of $\mathcal{W}$ and Lemma 6.2.7 to bound $\mathrm{rad}_{\|\cdot\|_{\mathcal{F}_R}}(\mathcal{D}(\cdot)) \leq W_0 = (L^2/\gamma + 2D^2)^{\frac{1}{2}}$ and $\mathrm{rad}_{\|\cdot\|_{\mathcal{F}_R}}(\mathfrak{T}) \leq D \leq W_0$ (since the Lemma the maximum of these radii). Then using Lemma A.3.1 since the normalized secant set is included in the mixture set we get the result.

□

Proof of Lemma 6.2.10. We start by decomposing $\mathcal{D}(\xi \pi_{\boldsymbol{\theta}}) = S_1 \cup S_2$ where

$$\begin{aligned}
S_1 &= \left\{ \frac{\xi \pi_{\boldsymbol{\theta}} - \xi' \pi_{\boldsymbol{\theta}'}}{\|\xi \pi_{\boldsymbol{\theta}} - \xi' \pi_{\boldsymbol{\theta}'}\|_\kappa} \;\middle|\; \boldsymbol{\theta}' \in \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1),\ \xi' \in [0; 1],\ \|\xi \pi_{\boldsymbol{\theta}} - \xi' \pi_{\boldsymbol{\theta}'}\|_\kappa > 0,\ \xi' \leq \xi \right\} \\
&= \left\{ \frac{\pi_{\boldsymbol{\theta}} - a \pi_{\boldsymbol{\theta}'}}{\|\pi_{\boldsymbol{\theta}} - a \pi_{\boldsymbol{\theta}'}\|_\kappa} \;\middle|\; \boldsymbol{\theta}' \in \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1),\ \|\pi_{\boldsymbol{\theta}} - a \pi_{\boldsymbol{\theta}'}\|_\kappa > 0,\ a \in [0; 1] \right\} \\
S_2 &= \left\{ \frac{\xi \pi_{\boldsymbol{\theta}} - \xi' \pi_{\boldsymbol{\theta}'}}{\|\xi \pi_{\boldsymbol{\theta}} - \xi' \pi_{\boldsymbol{\theta}'}\|_\kappa} \;\middle|\; \boldsymbol{\theta}' \in \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1),\ \xi' \in [0; 1],\ \|\xi \pi_{\boldsymbol{\theta}} - \xi' \pi_{\boldsymbol{\theta}'}\|_\kappa > 0,\ \xi' \geq \xi \right\} \\
&= \left\{ \frac{a \pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}}{\|a \pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}\|_\kappa} \;\middle|\; \boldsymbol{\theta}' \in \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1),\ \|a \pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}\|_\kappa > 0,\ a \in [\xi; 1] \right\}.
\end{aligned}$$

Next, for a given $\eta > 0$ that will be defined later, we decompose each of these sets into an extruded and non-extruded part: $S_1 = S_1^{>\eta} \cup S_1^{\leq\eta}$ and $S_2 = S_2^{>\eta} \cup S_2^{\leq\eta}$, where

$$S_1^{>\eta} = \left\{ \frac{\pi_{\boldsymbol{\theta}} - a\pi_{\boldsymbol{\theta}'}}{\|\pi_{\boldsymbol{\theta}} - a\pi_{\boldsymbol{\theta}'}\|_\kappa} \in S_1 \;\middle|\; \|\pi_{\boldsymbol{\theta}} - a\pi_{\boldsymbol{\theta}'}\|_\kappa > \eta \right\}$$
$$S_1^{\leq\eta} = \left\{ \frac{\pi_{\boldsymbol{\theta}} - a\pi_{\boldsymbol{\theta}'}}{\|\pi_{\boldsymbol{\theta}} - a\pi_{\boldsymbol{\theta}'}\|_\kappa} \in S_1 \;\middle|\; \|\pi_{\boldsymbol{\theta}} - a\pi_{\boldsymbol{\theta}'}\|_\kappa \leq \eta \right\}$$
$$S_2^{>\eta} = \left\{ \frac{a\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}}{\|a\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}\|_\kappa} \in S_2 \;\middle|\; \|a\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}\|_\kappa > \eta \right\}$$
$$S_2^{\leq\eta} = \left\{ \frac{a\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}}{\|a\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}\|_\kappa} \in S_2 \;\middle|\; \|a\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}\|_\kappa \leq \eta \right\}$$

We then bound the covering numbers:

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{D}(\xi\pi_{\boldsymbol{\theta}}), \delta\right) \leq \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, S_1^{>\eta}, \delta\right) + \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, S_2^{>\eta}, \delta\right) + \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, S_1^{\leq\eta} \cup S_2^{\leq\eta}, \delta\right). \tag{C.7}$$

We have:

– applying Lemma A.3.5 with the sets $Y = \{\pi_{\boldsymbol{\theta}}\}$ a singleton and $Z = \left\{ a\pi_{\boldsymbol{\theta}'} \;\middle|\; \varrho(\boldsymbol{\theta}', \boldsymbol{\theta}) \leq 1,\ a \in [0;1] \right\}$, the norms $\|\cdot\|_a = \|\cdot\|_{\mathcal{F}_R}$ and $\|\cdot\|_b = \|\cdot\|_\kappa$ and the constants $A = 1$ and $B = W_0 \geq 1$, and using Lemma A.3.6 with $\mathrm{rad}_{\|\cdot\|_{\mathcal{F}_R}}(\mathfrak{T}) \leq D$ for the covering numbers of Z, and using Lemma A.3.2 with the fact that $\|\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}\|_{\mathcal{F}_R} \leq L\varrho(\boldsymbol{\theta}, \boldsymbol{\theta}')$, we get

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, S_1^{>\eta}, \delta\right) \leq \max\left(\frac{16W_0D}{\delta\eta}, 1\right) \mathcal{N}\left(\varrho, \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1), \frac{\delta\eta}{16W_0L}\right); \tag{C.8}$$

– again, applying Lemma A.3.5 with the sets $Y = \left\{ a\pi_{\boldsymbol{\theta}} \;\middle|\; a \in [\xi; 1] \right\}$ and $Z = \left\{ \pi_{\boldsymbol{\theta}'} \;\middle|\; \boldsymbol{\theta}' \in \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1) \right\}$, the same norms and constants, we get

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, S_2^{>\eta}, \delta\right) \leq \max\left(\frac{8W_0D(1-\xi)}{\delta\eta}, 1\right) \mathcal{N}\left(\varrho, \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1), \frac{\delta\eta}{8W_0L}\right) \leq \frac{1}{2}\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, S_1^{>\eta}, \delta\right) \tag{C.9}$$

where we bound the covering numbers of $\mathcal{S}_2^{>\eta}$ by that of $\mathcal{S}_1^{>\eta}$ for simplicity.

– finally, applying the hypotheses on the set $\mathcal{V}_{\boldsymbol{\theta}}$ with Lemma A.3.3 (placed in a set X of tampered distributions, assuming the random features are smooth), we get

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, S_1^{\leq\eta} \cup S_2^{\leq\eta}, 2(\delta + M\eta)\right) \leq \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{V}_{\boldsymbol{\theta}}, \delta\right),$$

meaning that for all $\delta \leq 4M\eta_{\max}$,

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, S_1^{\leq\delta/(4M)} \cup S_2^{\leq\delta/(4M)}, \delta\right) \leq \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{V}_{\boldsymbol{\theta}}, \frac{\delta}{4}\right). \tag{C.10}$$

Finally, combining (C.7), (C.8), (C.9) and (C.10) with $\eta := \frac{\delta}{4M}$ we get the result. □

C.4 Choice of kernel

Proof of Lemma 6.2.11. Let us search for σ_k^2 under the form $\sigma_k^2 = [2(a\ln k + b)]^{-1}$.

Define $K(x) = e^{-\frac{x^2}{2\sigma^2}}$, suppose that $\sigma^2 \leq 1/3$ (which will be indeed the case). An easy function study of $h(t) := (1 - t/2)\exp(\frac{t}{2\sigma^2})$ shows that, when $\sigma^2 \leq 1/2$, h is non-decreasing on $[0, 1]$ with $h(0) = 1$, implying that $1 - u^2/2 \geq f(u)$ for $0 \leq u \leq 1$. This verifies (i) in the definition of $\mathcal{E}(A, B, C, \gamma)$ (Def. 6.2.3), with $\gamma = 1$.

By an easy study of K'', K' is negative and increasing for $u^2 \geq \sigma^2$. Thus $|K'(u)|$ is decreasing for $u \geq 1$ and we can set $B = |K'(1)| = \exp(-\frac{1}{2\sigma^2})/\sigma^2$. Note that we have

$B > A = K(1)$ and therefore $\max(A, 2(C+B)) = 2(C+B)$.

Similarly, an easy study of $K^{(3)}$ shows that K'' is positive and decreasing for $u^2 \geq 3\sigma^2$. Since $\sigma^2 \leq 1/3$, K'' is positive decreasing for $u \geq 1$ and we can set $C = K''(1) = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2} - 1\right)\exp(-\frac{1}{2\sigma^2})$. As a result $(B+C) = \exp(-1/2\sigma^2)/\sigma^4$ and the condition $2(B+C) \leq 3/(64k)$ reads as:

$$\exp(-\tfrac{1}{2\sigma^2})/\sigma^4 \leq 3/(128k).$$

With the definition of σ_k^2, the desired property holds if

$$\begin{aligned} \sigma_k^4 e^{1/2\sigma_k^2} &\geq 128k/3 \\ \frac{k^a e^b}{4(a\ln k + b)^2} &\geq 128k/3 \\ \frac{k^{a-1}}{(a\ln k + b)^2} &\geq 512e^{-b}/3 \end{aligned}$$

Consider $K(k) := \ln\left(k^{a-1}/(a\ln k + b)^2\right) = (a-1)\ln k - 2\ln(a\ln k + b)$. A quick function study shows that its derivative is positive if $\ln k \geq 2/(a-1) - b/a$. As soon as $2/(a-1) - b/a \leq 0$, i.e.,

$$a \geq \frac{b}{b-2}, \tag{C.11}$$

the function K is therefore increasing for $k \geq 1$, its minimum is at $k = 1$, and the desired property holds if $1/b^2 \geq 512e^{-b}/3$, i.e.,

$$b - 2\ln b - \ln\frac{512}{3} \geq 0.$$

The latter holds true for, e.g., $b = 12$, (C.11) holds as soon as $a \geq b/(b-2) = 1.2$, which proves the result. □

Appendix D

Application to Mixtures of Diracs

D.1 Proof of Lemma 6.3.3

Proof of Lemma 6.3.3. Consider the embedding $\varphi : \mathcal{T} \to \mathfrak{T}$ defined as $\varphi(\boldsymbol{\theta}) = \delta_{\boldsymbol{\theta}}$, which is surjective by definition of $\mathfrak{T}$.

Consider $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{T}$. We have

$$\begin{aligned}\left\|\delta_{\boldsymbol{\theta}} - \delta_{\boldsymbol{\theta}'}\right\|_{\mathcal{F}_R} &= \sup_{\boldsymbol{\omega}} \frac{2}{c(\boldsymbol{\omega})} \left| e^{\mathrm{i}\boldsymbol{\omega}^\top \boldsymbol{\theta}} - e^{\mathrm{i}\boldsymbol{\omega}^\top \boldsymbol{\theta}'} \right| \\ &\leq \sup_{\boldsymbol{\omega}} \frac{2\sqrt{d}}{\lambda \|\boldsymbol{\omega}\|_2} \|\boldsymbol{\omega}\|_2 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 = L \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2\end{aligned}$$

where $L = 2\sqrt{d}/\lambda$. Hence φ is L-Lipschitz.

Finally, we have

$$\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathfrak{T}, \delta\right) \overset{\text{Lem. A.3.2}}{\leq} \mathcal{N}\left(\|\cdot\|_2, \mathcal{T}, \frac{\delta}{L}\right) \overset{\text{Lem. A.3.4}}{\leq} \max\left(\left(\frac{4LR_{\mathbf{c}}}{\delta}\right)^d, 1\right). \tag{D.1}$$

□

D.2 Proof of Lemma 6.3.4

Proof of Lemma 6.3.4. Consider $\boldsymbol{\theta} \in \mathcal{T}$, and $\eta \leq 1/2$. We must build a set $\mathcal{V}_{\boldsymbol{\theta}}$ that is close to two sets:

$$S_1 = \left\{ \frac{\pi_{\boldsymbol{\theta}} - a\pi_{\boldsymbol{\theta}'}}{\|\pi_{\boldsymbol{\theta}} - a\pi_{\boldsymbol{\theta}'}\|_\kappa} \;\middle|\; \boldsymbol{\theta}' \in \mathcal{T},\ \varrho(\boldsymbol{\theta}, \boldsymbol{\theta}') \leq 1,\ a \in [0;1],\ \|\pi_{\boldsymbol{\theta}} - a\pi_{\boldsymbol{\theta}'}\|_\kappa \leq \eta \right\} \tag{D.2}$$

$$S_2 = \left\{ \frac{a\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}}{\|a\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}\|_\kappa} \;\middle|\; \boldsymbol{\theta}' \in \mathcal{T},\ \varrho(\boldsymbol{\theta}, \boldsymbol{\theta}') \leq 1,\ a \in [0;1],\ \|a\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}\|_\kappa \leq \eta \right\} \tag{D.3}$$

Consider any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{T}$ and $a \in [0;1]$ such that $\varrho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \leq 1$ and $\|\pi_{\boldsymbol{\theta}_1} - a\pi_{\boldsymbol{\theta}_2}\|_\kappa \leq \eta$. We are going to approach $\mu = \frac{\pi_{\boldsymbol{\theta}_1} - a\pi_{\boldsymbol{\theta}_2}}{\|\pi_{\boldsymbol{\theta}_1} - a\pi_{\boldsymbol{\theta}_2}\|_\kappa}$ with some tempered distribution ν. We have

$$\|\mu - \nu\|_{\mathcal{F}_R} = \sup_{\boldsymbol{\omega}} \frac{2}{c(\boldsymbol{\omega})} \left| \frac{e^{\mathrm{i}\boldsymbol{\omega}^\top \boldsymbol{\theta}_1} - a e^{\mathrm{i}\boldsymbol{\omega}^\top \boldsymbol{\theta}_2}}{\|\pi_{\boldsymbol{\theta}_1} - a\pi_{\boldsymbol{\theta}_2}\|_\kappa} - \psi_\nu(\boldsymbol{\omega}) \right| \tag{D.4}$$

where ψ_ν is the characteristic function of ν.

Denote $\Delta_{\boldsymbol{\theta}} = \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1$, $b = 1 - a$ and $\alpha = \|\pi_{\boldsymbol{\theta}_1} - a\pi_{\boldsymbol{\theta}_2}\|_\kappa$. Also denote $\Delta_0 = \Delta_{\boldsymbol{\theta}}/\alpha$ and $b_0 = b/\alpha$, and finally $K_{12} = K(\varrho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)) \leq K(0) = 1$.

With these notations, the first term in (D.4) reads

$$\frac{e^{\mathrm{i}\boldsymbol{\omega}^\top \boldsymbol{\theta}_1} - a e^{\mathrm{i}\boldsymbol{\omega}^\top \boldsymbol{\theta}_2}}{\|\pi_{\boldsymbol{\theta}_1} - a\pi_{\boldsymbol{\theta}_2}\|_\kappa} = e^{\mathrm{i}\boldsymbol{\omega}^\top \boldsymbol{\theta}_1} \left(\frac{1 - e^{\mathrm{i}\alpha\boldsymbol{\omega}^\top \Delta_0}}{\alpha} \right) + b_0 e^{\mathrm{i}\boldsymbol{\omega}^\top \boldsymbol{\theta}_2}$$

The parameters Δ_0 and b_0 can be controlled: we have

$$\alpha^2 = 1 - 2(1-b)K_{12} + 1 + b^2 - 2b = 2(1-b)(1-K_{12}) + b^2$$

meaning that $b^2 \leq \alpha^2$ and thus $b_0 \leq 1$. Furthermore we have:

$$\|\Delta_0\|_2^2 = \frac{\lambda^2 \varrho^2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\alpha^2 \sigma_k^2} \overset{K \in \mathcal{E}(A,B,C,1)}{\leq} \frac{\lambda^2}{\sigma_k^2} \cdot \frac{2(1-K_{12})}{2(1-b)(1-K_{12}) + b^2}$$

hence either $1 - K_{12} = 0$ and $\|\Delta_0\| = 0$, either

$$\|\Delta_0\|_2^2 \leq \frac{\lambda^2}{(1-b)\sigma_k^2} \leq \frac{2\lambda^2}{\sigma_k^2} = r_\Delta^2. \tag{D.5}$$

using the fact that $-b \geq -\eta \geq -1/2$.

By Taylor expansion, we have

$$\left| \frac{1 - e^{\mathrm{i}\alpha \boldsymbol{\omega}^\top \Delta_0}}{\alpha} + \mathrm{i}\boldsymbol{\omega}^\top \Delta_0 \right| \leq \sup_{\alpha \in [0,\ \eta]} \left| \frac{\partial^2}{\partial^2 \alpha} (1 - e^{-\mathrm{i}\alpha \boldsymbol{\omega}^\top \Delta_0}) \right| \frac{\alpha}{2} = \left| \boldsymbol{\omega}^\top \Delta_0 \right|^2 \frac{\alpha}{2} \leq \|\boldsymbol{\omega}\|_2^2 \frac{\lambda^2 \eta}{\sigma_k^2} \tag{D.6}$$

since $\alpha \leq \eta$.

Hence, if we define $\nu = -\delta'_{\boldsymbol{\theta}_1, \Delta_0} + b_0 \delta_{\boldsymbol{\theta}_2}$ a tampered distribution where $\delta'_{\boldsymbol{\theta}_1, \Delta_0}$ is the derivative of the Dirac function at position $\boldsymbol{\theta}_1$ along direction Δ_0, *i.e.* such that $\psi_\nu(\boldsymbol{\omega}) = -\mathrm{i}\boldsymbol{\omega}^\top \Delta_0 e^{\mathrm{i}\boldsymbol{\omega}^\top \boldsymbol{\theta}_1} + b_0 e^{\mathrm{i}\boldsymbol{\omega}^\top \boldsymbol{\theta}_2}$, we have

$$\|\mu - \nu\|_{\mathcal{F}_R} = \sup_{\boldsymbol{\omega}} \frac{2}{c(\boldsymbol{\omega})} \left| \frac{1 - e^{\mathrm{i}\alpha \boldsymbol{\omega}^\top \Delta_0}}{\alpha} + \mathrm{i}\boldsymbol{\omega}^\top \Delta_0 \right| \leq \frac{2\sqrt{d(d+2)}}{\|\boldsymbol{\omega}\|_2^2 \lambda^2} \frac{\|\boldsymbol{\omega}\|_2^2 \lambda^2 \eta}{\sigma_k^2} = \frac{2\eta\sqrt{d(d+2)}}{\sigma_k^2} = M\eta.$$

Using this property with $\boldsymbol{\theta}_1 = \boldsymbol{\theta}$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}'$, for all $\mu \in S_1$ there is ν in

$$\mathcal{V}_1 = \left\{ -\delta'_{\boldsymbol{\theta}, \Delta} + b\delta_{\boldsymbol{\theta}'} \;\middle|\; \boldsymbol{\theta}' \in \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1),\ \Delta \in \mathcal{B}_{\mathbb{R}^d, \|\cdot\|_2}(0, r_\Delta),\ b \in [0;1] \right\}$$

such that $\|\mu - \nu\|_{\mathcal{F}_R} \leq M\eta$. Similarly, by taking $\boldsymbol{\theta}_1 = \boldsymbol{\theta}'$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}$, by symmetry for all $\mu \in S_2$ there is ν in

$$\mathcal{V}_2 = \left\{ \delta'_{\boldsymbol{\theta}', \Delta} - b\delta_{\boldsymbol{\theta}} \;\middle|\; \boldsymbol{\theta}' \in \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1),\ \Delta \in \mathcal{B}_{\mathbb{R}^d, \|\cdot\|_2}(0, r_\Delta),\ b \in [0;1] \right\}$$

such that $\|\mu - \nu\|_{\mathcal{F}_R} \leq M\eta$. Hence we define $\mathcal{V}_{\boldsymbol{\theta}} = \mathcal{V}_1 \cup \mathcal{V}_2$ the set of tempered distributions that satisfies the desired property, and must now bound its covering numbers.

Consider the product space $X = \mathcal{B}_{\mathbb{R}^d, \|\cdot\|_2}(0, r_\Delta) \times \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1) \times [0;1]$.

Let us begin with $\mathcal{V}_1$. Given $x = (\Delta, \boldsymbol{\theta}', b) \in X$, define the function $\varphi : X \mapsto \mathcal{V}_1$ by $\varphi(x) = -\delta'_{\Delta, \boldsymbol{\theta}} + b\delta_{\boldsymbol{\theta}'}$.

For $x_1 = (\Delta_1, \boldsymbol{\theta}_1, b_1)$ and $x_2 = (\Delta_2, \boldsymbol{\theta}_2, b_2)$ in X, we have

$$\begin{aligned} \|\varphi(x_1) - \varphi(x_2)\|_{\mathcal{F}_R} &= \left\| -\delta'_{\Delta_1, \boldsymbol{\theta}} + b_1 \delta_{\boldsymbol{\theta}_1} + \delta'_{\Delta_2, \boldsymbol{\theta}} - b_2 \delta_{\boldsymbol{\theta}_2} \right\|_{\mathcal{F}_R} \\ &\leq \left\| \delta'_{\Delta_1, \boldsymbol{\theta}} - \delta'_{\Delta_2, \boldsymbol{\theta}} \right\|_{\mathcal{F}_R} + \|b_1 \delta_{\boldsymbol{\theta}_1} - b_1 \delta_{\boldsymbol{\theta}_2}\|_{\mathcal{F}_R} + \|b_1 \delta_{\boldsymbol{\theta}_2} - b_2 \delta_{\boldsymbol{\theta}_2}\|_{\mathcal{F}_R} \end{aligned}$$

We bound each of those terms. First,

$$\begin{aligned} \left\| \delta'_{\Delta_1, \boldsymbol{\theta}} - \delta'_{\Delta_2, \boldsymbol{\theta}} \right\|_{\mathcal{F}_R} &= \sup_{\boldsymbol{\omega}} \frac{2}{c(\boldsymbol{\omega})} \left| \boldsymbol{\omega}^\top (\Delta_1 - \Delta_2) \right| \\ &\leq \sup_{\boldsymbol{\omega}} \frac{2\sqrt{d}\,\|\boldsymbol{\omega}\|_2}{\lambda \|\boldsymbol{\omega}\|_2} \|\Delta_1 - \Delta_2\|_2 = L_1 \|\Delta_1 - \Delta_2\|_2 \end{aligned}$$

where $L_1 = 2\sqrt{d}/\lambda$. Then we have

$$\begin{aligned}\|b_1\delta_{\boldsymbol{\theta}_1} - b_1\delta_{\boldsymbol{\theta}_2}\|_{\mathcal{F}_R} \leq \|\delta_{\boldsymbol{\theta}_1} - \delta_{\boldsymbol{\theta}_2}\|_{\mathcal{F}_R} &= \sup_{\boldsymbol{\omega}} \frac{2}{c(\boldsymbol{\omega})} \left|e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_1} - e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_2}\right| \\ &\leq \sup_{\boldsymbol{\omega}} \frac{2\sqrt{d}\,\|\boldsymbol{\omega}\|_2}{\lambda\,\|\boldsymbol{\omega}\|_2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 = L_2\varrho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2).\end{aligned}$$

where $L_2 = \frac{2\sqrt{d}}{\sigma_k}$. Finally, we have

$$\|b_1\delta_{\boldsymbol{\theta}_2} - b_2\delta_{\boldsymbol{\theta}_2}\|_{\mathcal{F}_R} \leq \|\delta_{\boldsymbol{\theta}_2}\|_{\mathcal{F}_R} |b_1 - b_2| \leq L_3 |b_1 - b_2|$$

with $L_3 = \sqrt{2}$.

Therefore if we denote $\mathcal{C}_1$ a $\frac{\delta}{3L_1}$-covering of $\mathcal{B}_{\mathbb{R}^d,\|\cdot\|_2}(0, r_\Delta)$, $\mathcal{C}_2$ a $\frac{\delta}{3L_2}$-covering of $\mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1)$ and $\mathcal{C}_3$ a $\frac{\delta}{3L_3}$-covering of $[0;1]$, for any $x \in X$ there exists an element $\bar{x} \in \mathcal{C}_1 \times \mathcal{C}_2 \times \mathcal{C}_3$ such that $\|\varphi(x) - \varphi(x')\|_{\mathcal{F}_R} \leq \delta$. Thus we have

$$\begin{aligned}\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{V}_1, \delta\right) &\leq |\mathcal{C}_1| \cdot |\mathcal{C}_2| \cdot |\mathcal{C}_3| \\ &\leq \mathcal{N}\left(\|\cdot\|_2, \mathcal{B}_{\mathbb{R}^d,\|\cdot\|_2}(0, r_\Delta), \frac{\delta}{3L_1}\right) \cdot \mathcal{N}\left(\varrho, \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1), \frac{\delta}{3L_2}\right) \cdot \mathcal{N}\left(\|\cdot\|_1, [0;1], \frac{\delta}{3L_3}\right) \\ &\leq \max\left(\left(\frac{12r_\Delta L_1}{\delta}\right)^d, 1\right) \cdot \max\left(\left(\frac{12L_2}{\delta}\right)^d, 1\right) \cdot \max\left(\frac{3L_3}{\delta}, 1\right)\end{aligned}$$

All the $\max(\cdot, 1)$ are resolved since $12r_\Delta L_1,\ 12L_2,\ 3L_3$ are all greater than 1, and we assumed $\delta \leq 1$ for simplicity.

We now turn to $\mathcal{V}_2$. Given $x = (\Delta, \boldsymbol{\theta}', b) \in X$, define the function $\varphi : X \mapsto \mathcal{V}_2$ by $\varphi(x) = \delta'_{\Delta,\boldsymbol{\theta}'} - b\delta_{\boldsymbol{\theta}}$.

For $x_1 = (\Delta_1, \boldsymbol{\theta}_1, b_1)$ and $x_2 = (\Delta_2, \boldsymbol{\theta}_2, b_2)$ in X, we have

$$\begin{aligned}\|\varphi(x_1) - \varphi(x_2)\|_{\mathcal{F}_R} &= \left\|\delta'_{\Delta_1,\boldsymbol{\theta}_1} - b_1\delta_{\boldsymbol{\theta}} - \delta'_{\Delta_2,\boldsymbol{\theta}_2} + b_2\delta_{\boldsymbol{\theta}}\right\|_{\mathcal{F}_R} \\ &\leq \left\|\delta'_{\Delta_1,\boldsymbol{\theta}_1} - \delta'_{\Delta_2,\boldsymbol{\theta}_1}\right\|_{\mathcal{F}_R} + \left\|\delta'_{\Delta_2,\boldsymbol{\theta}_1} - \delta'_{\Delta_2,\boldsymbol{\theta}_2}\right\|_{\mathcal{F}_R} + \|b_1\delta_{\boldsymbol{\theta}} - b_2\delta_{\boldsymbol{\theta}}\|_{\mathcal{F}_R}\end{aligned}$$

Similar to the previous case, we bound each of those terms:

$$\left\|\delta'_{\Delta_1,\boldsymbol{\theta}_1} - \delta'_{\Delta_2,\boldsymbol{\theta}_1}\right\|_{\mathcal{F}_R} \leq L_1 \|\Delta_1 - \Delta_2\|_2 .$$

Then

$$\begin{aligned}\left\|\delta'_{\Delta_2,\boldsymbol{\theta}_1} - \delta'_{\Delta_2,\boldsymbol{\theta}_2}\right\|_{\mathcal{F}_R} &= \sup_{\boldsymbol{\omega}} \frac{2}{c(\boldsymbol{\omega})} \left|\boldsymbol{\omega}^\top\Delta_2\right| \left|e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_1} - e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_2}\right| \\ &\leq \sup_{\boldsymbol{\omega}} \frac{2\sqrt{d(d+2)}}{\lambda^2\,\|\boldsymbol{\omega}\|_2^2} \|\boldsymbol{\omega}\|_2^2 \|\Delta\|_2 \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \\ &\leq L_4\varrho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\end{aligned}$$

where $L_4 = \frac{2r_\Delta\sqrt{d(d+2)}}{\lambda\sigma_k} = \frac{2\sqrt{2d(d+2)}}{\sigma_k^2}$. Finally,

$$\|b_1\delta_{\boldsymbol{\theta}} - b_2\delta_{\boldsymbol{\theta}}\|_{\mathcal{F}_R} \leq \|\delta_{\boldsymbol{\theta}}\|_{\mathcal{F}_R} |b_1 - b_2| \leq L_3 |b_1 - b_2| .$$

Therefore if we denote $\mathcal{C}_1$ a $\frac{\delta}{3L_1}$-covering of $\mathcal{B}_{\mathbb{R}^d,\|\cdot\|_2}(0,r_\Delta)$, $\mathcal{C}_2$ a $\frac{\delta}{3L_4}$-covering of $\mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta},1)$ and $\mathcal{C}_3$ a $\frac{\delta}{3L_3}$-covering of $[0;1]$, similar to the previous case we have

$$\begin{aligned}\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R},\mathcal{V}_1,\delta\right) &\leq |\mathcal{C}_1|\cdot|\mathcal{C}_2|\cdot|\mathcal{C}_3| \\ &\leq \mathcal{N}\left(\|\cdot\|_2,\mathcal{B}_{\mathbb{R}^d,\|\cdot\|_2}(0,r_\Delta),\frac{\delta}{3L_1}\right)\cdot\mathcal{N}\left(\varrho,\mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta},1),\frac{\delta}{3L_4}\right)\cdot\mathcal{N}\left(\|\cdot\|_1,[0;1],\frac{\delta}{3L_3}\right) \\ &\leq \max\left(\left(\frac{12r_\Delta L_1}{\delta}\right)^d,1\right)\cdot\max\left(\left(\frac{12L_4}{\delta}\right)^d,1\right)\cdot\max\left(\frac{3L_3}{\delta},1\right)\end{aligned}$$

Since $L_4 > L_2$, this bound is greater than the bound for the covering numbers of $\mathcal{V}_1$, and we bound the covering numbers of $\mathcal{V}_{\boldsymbol{\theta}}$ by twice that of $\mathcal{V}_2$ to obtain the final result.

□

D.3 Proof of Theorem 6.3.5

Proof. Applying all the Lemmas of Section 6.3.2, all the hypotheses necessary to apply Theorem 6.2.12 hold. We obtain the following constants:

- Compatibility constant: given in equation (6.45);
- Admissibility constant: we have

$$W_0 = 2\sqrt{\frac{d}{\sigma_k^2}+1} = \mathcal{O}\left(\sqrt{d\log k}\right)$$

and the admissibility constant

$$W_\Lambda = 4\sqrt{2k\left(\frac{d}{\sigma_k^2}+1\right)} = \mathcal{O}\left(\sqrt{kd\log k}\right)$$

- Finally, according to Lemma 6.3.4 there exist tangent sets $\mathcal{V}_{\boldsymbol{\theta}}$ such that their covering numbers have a common bound independent of $\boldsymbol{\theta}$. Combining this bound with Lemma 6.3.3 and equation (6.30), we get:

$$N = \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R},\mathcal{S}^0(\cdot,\mathfrak{S}),\frac{1}{4}\right) \leq \left[A_1A_2^d + A_3A_4^d\right]^k$$

with

$$\begin{aligned}A_1 &= 3\cdot 2^{32}\sqrt{2}k^2W_0^3\sqrt{d(d+2)}\sigma_k^{-2} \\ A_2 &= 2^{29}kd\sqrt{2k(d+2)}\sigma_k^{-2}\frac{R_{\mathbf{c}}}{\varepsilon_\lambda} \\ A_3 &= 3\cdot 2^{23}k\sqrt{k}W_0^2 \\ A_4 &= 3^2 2^{30}d\sqrt{2kd(d+2)}\sigma_k^{-2}\frac{R_{\mathbf{c}}}{\varepsilon_\lambda}\end{aligned}$$

and the sketch size (6.47) indeed scales as

$$m \geq cW_\Lambda^2\log\left(\frac{N}{\rho}\right) = \mathcal{O}\left(k^2d^2\texttt{polylog}\left(k,d,\frac{R_{\mathbf{c}}}{\varepsilon},\frac{1}{\rho}\right)\right)$$

□

D.4 Proof of Lemma 6.3.6

We will use the two following lemmas.

Lemma D.4.1. *Define $\mathcal{F}(L)$ the family of functions $f : \mathfrak{Z} \mapsto \mathbb{C}$ that are L-Lipschitz for some $L > 0$ (with respect to the norm $\|\cdot\|_2$), and $\mathcal{F} \subset \mathcal{F}(L)$. For any probability distribution $\pi^\star$ and any set of centroids $h = \{\mathbf{c}_1, \dots, \mathbf{c}_k\} \subset \mathbb{R}^d$, there exists a weight vector $\boldsymbol{\xi} \in \mathbb{S}^{k-1}$ such that*

$$\left\| \pi^\star - \sum_{l=1}^{k} \xi_l \delta_{\mathbf{c}_l} \right\|_{\mathcal{F}} \leq L \left(\mathcal{R}_{\pi^\star}(h) \right)^{\frac{1}{b}} \tag{D.7}$$

where $b = 2$ for k-means and $b = 1$ for k-medians.

Proof. It is well-known (see, e.g., [Sri+10]) that the Wasserstein distance between two distributions can be defined in terms of transport (so-called "earth mover's distance") but also equivalently as

$$\|\pi - \pi'\|_{\text{Wasserstein}} = \|\pi - \pi'\|_{\mathcal{F}(1)} = \sup_{f \in \mathcal{F}(1)} |\langle \pi - \pi', f \rangle| = \frac{1}{L} \|\pi - \pi'\|_{\mathcal{F}(L)} ,$$

Let $\pi^\star$ be a probability distribution and $h = \{\mathbf{c}_1, \dots, \mathbf{c}_k\} \subset \mathbb{R}^d$ be a set of centroids. Denote $V_l := \left\{ \mathbf{z} \in \mathfrak{Z} \mid l = \arg\min_p \|\mathbf{z} - \mathbf{c}_p\|_2 \right\}$ the Voronoi cell associated to $\mathbf{c}_i$ (ties are breaked arbitrarily to ensure this constitutes a partition of $\mathfrak{Z}$), define $\xi_l = \pi^\star(V_l)$. Then by using the transport characterization of the Wasserstein distance, considering the transport plan consisting in sending all points of V_l to $\mathbf{c}_l$, we conclude

$$\left\| \pi - \sum_{l=1}^{k} \xi_l \delta_{\mathbf{c}_l} \right\|_{\text{Wasserstein}} \leq \sum_{l=1}^{k} \xi_l \mathbb{E}_{\mathbf{z} \sim \pi^\star} \left[\|\mathbf{z} - \mathbf{c}_l\|_2 \,|\, \mathbf{z} \in V_l \right] = \mathbb{E}_{\mathbf{z} \sim \pi^\star} \min_l \|\mathbf{z} - \mathbf{c}_l\| = \mathcal{R}_{\pi^\star}^{k-\text{med.}}(h) .$$

This yields the result for k-medians. The result for k-means is an immediate consequence of Jensen's inequality:

$$\mathcal{R}_{\pi^\star}^{k-\text{med.}}(h) = \mathbb{E}_{\mathbf{z} \sim \pi^\star} \min_l \|\mathbf{z} - \mathbf{c}_l\|_2 \leq \left(\mathbb{E}_{\mathbf{z} \sim \pi^\star} \min_l \|\mathbf{z} - \mathbf{c}_l\|_2^2 \right)^{\frac{1}{2}} = \sqrt{\mathcal{R}_{\pi^\star}^{k-\text{means}}(h)}.$$

□

Lemma D.4.2. *Consider $\mathcal{F} \subset \mathcal{F}(L)$. Consider any mixture of Diracs $\pi_{\Theta,\xi} = \sum_{l=1}^{k} \xi_l \delta_{\boldsymbol{\theta}_l}$. Then, there exists a function $\sigma : (1:k) \mapsto (1:k)$ such that: for all $\sigma(l) \neq \sigma(p)$ we have $\left\|\boldsymbol{\theta}_{\sigma(l)} - \boldsymbol{\theta}_{\sigma(p)}\right\|_2 \geq \varepsilon$ and*

$$\left\| \sum_{l=1}^{k} \xi_l \delta_{\boldsymbol{\theta}_l} - \sum_{l=1}^{k} \xi_l \delta_{\boldsymbol{\theta}_{\sigma(l)}} \right\|_{\mathcal{F}} \leq L\varepsilon \tag{D.8}$$

In other words, we transform the original sum of Diracs by suppressing some of them and repeating others (which just results into Diracs with higher weight), such that all remaining Diracs are ε-separated and the resulting mixture is a good approximation of the first one.

Proof. Define $\Theta' = \{\boldsymbol{\theta}_{i_1}, \dots, \boldsymbol{\theta}_{i_s}\}$ a subset of Θ of maximal size $s \leq k$ such that the $\boldsymbol{\theta}_{i_l}$ are ε-separated. Then, for all $\boldsymbol{\theta}_l \in \Theta$, either it is included in Θ', either it is ε-close to a point in Θ', otherwise it could have been included in Θ' while retaining ε-separation, and Θ' would not be maximal. Then for all $l \in (1:k)$, if $\boldsymbol{\theta}_l$ is included in Θ' we define $\sigma(l) = l$, if not we pick one $\boldsymbol{\theta}_{i_l} \in \Theta'$ such that $\|\boldsymbol{\theta}_l - \boldsymbol{\theta}_{i_l}\| \leq \varepsilon$ and define $\sigma(l) = i_l$. At the end of the day, we have either $\boldsymbol{\theta}_l = \boldsymbol{\theta}_{\sigma(l)}$, either $\left\|\boldsymbol{\theta}_l - \boldsymbol{\theta}_{\sigma(l)}\right\|_2 \leq \varepsilon$, and therefore

$$\left\| \sum_{l=1}^{k} \xi_l \delta_{\boldsymbol{\theta}_l} - \sum_{l=1}^{k} \xi_l \delta_{\boldsymbol{\theta}_{\sigma(l)}} \right\|_{\mathcal{F}} \leq \sum_{l=1}^{k} \xi_l \sup_{f \in \mathcal{F}} \left| f(\boldsymbol{\theta}_l) - f(\boldsymbol{\theta}_{\sigma(l)}) \right| \leq L \sum_{l=1}^{k} \xi_l \left\|\boldsymbol{\theta}_l - \boldsymbol{\theta}_{\sigma(l)}\right\|_2 \leq L\varepsilon$$

□

Proof of Lemma 6.3.6. For either k-means or k-medians denote $h^\star = \{\mathbf{c}_1^\star, \dots, \mathbf{c}_k^\star\} = \arg\min_{h\in\mathcal{H}} \mathcal{R}_{\pi^\star}(h)$. Using Lemma D.4.1, there is a set of weights $\boldsymbol{\xi}$ such that for $\mathcal{F} \subset \mathcal{F}(L)$, denoting $\pi = \sum_{l=1}^k \xi_l \delta_{\mathbf{c}_l^\star}$:

$$\|\pi^\star - \pi\|_{\mathcal{F}} \leq L \left(\mathcal{R}_{\pi^\star}(h)\right)^{\frac{1}{b}}$$

Now, if the $\mathbf{c}_l^\star$ are ε-separated (*i.e.* π is in the model), we define $\pi = \pi_{\mathfrak{S}} \in \mathfrak{S}$ and $\varepsilon' = 0$. Otherwise, using Lemma D.4.2, we can define π' a k'-mixture (with $k' \leq k$) of ε-separated Diracs such that $\|\pi - \pi'\|_{\mathcal{F}} \leq L\varepsilon$. We fill π' with dummy Diracs ε-separated from the others with weight 0 to obtain $\pi_{\mathfrak{S}} = \pi' \in \mathfrak{S}$ (technically, assuming the hypothesis class is sufficiently large to contain such dummy Diracs, which we suppose for simplicity), and in that case $\varepsilon' = \varepsilon$. In both cases we have

$$\|\pi^\star - \pi_{\mathfrak{S}}\|_{\mathcal{F}} \leq \|\pi^\star - \pi\|_{\mathcal{F}} + \|\pi - \pi_{\mathfrak{S}}\|_{\mathcal{F}} \leq L\left(\left(\mathcal{R}_{\pi^\star}(h)\right)^{\frac{1}{b}} + \varepsilon'\right)$$

We now have to prove that $\|\cdot\|_{\mathcal{L}(\mathcal{H})} + 4W_{\mathcal{L}} \|\cdot\|_{\mathcal{F}_R} \leq \|\cdot\|_{\mathcal{F}(L)}$ for some L, *i.e.* that both $\mathcal{L}(\mathcal{H})$ and $\mathcal{F}_R$ are families of Lipschitz functions.

In the proof of Lemma 6.3.2 we showed that the features $\phi_{\boldsymbol{\omega}}(\mathbf{z})$ are $\frac{2\sqrt{d}}{\lambda}$-Lipschitz with respect to the Euclidean norm.

Finally, in the proof of Lemma 6.3.1 we showed that in the k-medians case the loss functions are 1-Lipschitz, while in the k-means case the loss functions are $4R_{\mathbf{c}}$-Lipschitz *on a $R_{\mathbf{c}}$-bounded domain*, which is why we have to restrain the sample space in this case. This concludes the proof. □

Appendix E

Application to Gaussian mixture models

E.1 Proof of Lemma 6.4.1

Proof of Lemma 6.4.1. We use a property from [Ahr05] on product of Gaussians:

$$\int \pi_1(\mathbf{z})\pi_2(\mathbf{z})d\mathbf{z} = \frac{1}{(2\pi)^{d/2}\left|\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\right\|^2_{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}\right) \tag{E.1}$$

We can write the kernel

$$\kappa(\mathbf{z},\mathbf{z}') = \exp\left(-\frac{1}{2}\left\|\mathbf{z}-\mathbf{z}'\right\|^2_{\boldsymbol{\Sigma}_\kappa}\right) = (2\pi)^{d/2}\left|\boldsymbol{\Sigma}_\kappa\right|^{\frac{1}{2}}\pi_\kappa(\mathbf{z}-\mathbf{z}')$$

where $\pi_\kappa = \mathcal{N}(0,\boldsymbol{\Sigma}_\kappa)$. Hence we have

$$\begin{aligned}\kappa(\pi_1,\pi_2) =& (2\pi)^{d/2}\left|\boldsymbol{\Sigma}_\kappa\right|^{\frac{1}{2}}\int_{\mathbf{z}}\pi_1(\mathbf{z})\left(\int_{\mathbf{z}'}\pi_2(\mathbf{z}')\pi_\kappa(\mathbf{z}-\mathbf{z}')d\mathbf{z}'\right)d\mathbf{z}\\ =& (2\pi)^{d/2}\left|\boldsymbol{\Sigma}_\kappa\right|^{\frac{1}{2}}\int_{\mathbf{z}}\pi_1(\mathbf{z})\pi_{2,\kappa}(\mathbf{z})d\mathbf{z},\end{aligned}$$

by convolution, where $\pi_{2,\kappa} = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2+\boldsymbol{\Sigma}_\kappa)$. Using (E.1) we get the desired result. □

E.2 Proof of Lemma 6.4.2

Proof of Lemma 6.4.2. Consider $h = ((\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k),\boldsymbol{\xi}) \in \mathcal{H}$, recall that $\ell(\mathbf{z},h) = -\log\pi_h(\mathbf{z})$ where $\pi_h = \sum_{l=1}^k \xi_l\mathcal{N}(\boldsymbol{\mu}_l,\boldsymbol{\Sigma})$. For all $\boldsymbol{\theta}\in\mathcal{T}$, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{z}\sim\pi_{\boldsymbol{\theta}}}\ell(\mathbf{z},h) = D_{\mathrm{KL}}\left(\pi_{\boldsymbol{\theta}}||\pi_h\right) + \mathrm{H}\left(\pi_{\boldsymbol{\theta}}\right) \geq \mathrm{H}\left(\pi_{\boldsymbol{\theta}}\right) = \frac{1}{2}\left[\log\left|2\pi\boldsymbol{\Sigma}\right| + \mathbb{E}_{\mathbf{z}\sim\pi_{\boldsymbol{\theta}}}\left\|\mathbf{z}-\boldsymbol{\theta}\right\|^2_{\boldsymbol{\Sigma}}\right]\\ = \frac{1}{2}\left[\log\left|2\pi\boldsymbol{\Sigma}\right| + \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(0,\mathbf{I})}\left\|\mathbf{z}\right\|_2^2\right] = \frac{1}{2}\left[\log\left|2\pi\boldsymbol{\Sigma}\right| + d\right] = \frac{1}{2}\log\left|2\pi e\boldsymbol{\Sigma}\right|\end{aligned}$$

where $e = e^1$. Then by concavity of the logarithm we have

$$\begin{aligned}\mathbb{E}_{\mathbf{z}\sim\pi_{\boldsymbol{\theta}}}\ell(\mathbf{z},h) = \mathbb{E}_{\mathbf{z}\sim\pi_{\boldsymbol{\theta}}}\left[-\log\sum_{l=1}^k\xi_l\pi_{\boldsymbol{\mu}_l}(\mathbf{z})\right] \leq \sum_{l=1}^k\xi_l\mathbb{E}_{\mathbf{z}\sim\pi_{\boldsymbol{\theta}}}\left[-\log\pi_{\boldsymbol{\mu}_l}(\mathbf{z})\right]\\ = \sum_{l=1}^k\xi_l\mathbb{E}_{\mathbf{z}\sim\pi_{\boldsymbol{\theta}}}\left[-\log\pi_{\boldsymbol{\mu}_l}(\mathbf{z})\right] = \sum_{l=1}^k\xi_l\left(D_{\mathrm{KL}}\left(\pi_{\boldsymbol{\theta}}||\pi_{\boldsymbol{\mu}_l}\right) + \mathrm{H}\left(\pi_{\boldsymbol{\theta}}\right)\right)\\ \overset{\text{Lem. 3.3.4}}{=} \sum_{l=1}^k\xi_l\frac{1}{2}\left\|\boldsymbol{\theta}-\boldsymbol{\mu}_l\right\|^2_{\boldsymbol{\Sigma}} + \mathrm{H}\left(\pi_{\boldsymbol{\theta}}\right) \leq 2R^2_{\boldsymbol{\mu}} + \frac{1}{2}\log\left|2\pi e\boldsymbol{\Sigma}\right|\end{aligned}$$

Hence

$$|\mathbb{E}_{\mathbf{z}\sim\pi_{\boldsymbol{\theta}}}\ell(\mathbf{z},h)| \leq 2R_{\boldsymbol{\mu}}^2 + \frac{1}{2}\left|\log\left|2\pi e\boldsymbol{\Sigma}\right|\right|$$

For the Lipschitz part, we write $f(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}\sim\pi_{\boldsymbol{\theta}}}\ell(\mathbf{z},h)$ and

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{z}\sim\pi_{\boldsymbol{\theta}}}\ell(\mathbf{z},h) = \nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mathbf{z}\sim\pi_0}\ell(\mathbf{z}+\boldsymbol{\theta},h) = \mathbb{E}_{\mathbf{z}\sim\pi_0}\nabla_{\boldsymbol{\theta}}\ell(\mathbf{z}+\boldsymbol{\theta},h)$$

We have

$$\begin{aligned}\nabla\ell(\mathbf{z},h) &= -\nabla\log\pi_h(\mathbf{z}) = -\frac{\nabla\pi_h(\mathbf{z})}{\pi_h(\mathbf{z})} = -\frac{\sum_{l=1}^k \xi_l\nabla\pi_{\boldsymbol{\mu}_l}(\mathbf{z})}{\pi_h(\mathbf{z})}\\ &= -\sum_{l=1}^k \frac{\xi_l\pi_{\boldsymbol{\mu}_l}(\mathbf{z})}{\pi_h(\mathbf{z})}\cdot\frac{\nabla\pi_{\boldsymbol{\mu}_l}(\mathbf{z})}{\pi_{\boldsymbol{\mu}_l}(\mathbf{z})} = \sum_{l=1}^k \beta_l(\mathbf{z})\nabla\left[-\log\pi_{\boldsymbol{\mu}_l}(\mathbf{z})\right]\end{aligned}$$

where $\beta_l(\mathbf{z}) = \frac{\xi_l\pi_{\boldsymbol{\mu}_l}(\mathbf{z})}{\pi_h(\mathbf{z})} \geq 0$, $\sum_{l=1}^k \beta_l(\mathbf{z}) = 1$. Since $\nabla\left[-\log\pi_{\boldsymbol{\mu}}(\mathbf{z})\right] = \boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu}_l)$ we have

$$\nabla\ell(\mathbf{z},h) = \boldsymbol{\Sigma}^{-1}\left(\mathbf{z} - \sum_{l=1}^k \beta_l(\mathbf{z})\boldsymbol{\mu}_l\right)$$

and thus

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}\sim\pi_0}\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}+\boldsymbol{\theta} - \sum_{l=1}^k \beta_l(\mathbf{z}+\boldsymbol{\theta})\boldsymbol{\mu}_l\right) = \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta} - \sum_{l=1}^k \gamma_l\boldsymbol{\mu}_l\right)$$

where $\gamma_l = \mathbb{E}_{\mathbf{z}\sim\pi_0}\beta_l(\mathbf{z}+\boldsymbol{\theta}) \geq 0$, $\sum_{l=1}^k \gamma_l = 1$. Given $\boldsymbol{\theta},\boldsymbol{\theta}' \in \mathcal{T}$ we have

$$\begin{aligned}|f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}')| &\leq \sup_{\boldsymbol{\theta}''\in\mathcal{T}} |\langle\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}''), \boldsymbol{\theta}-\boldsymbol{\theta}'\rangle| \leq \sup_{\boldsymbol{\theta}''\in\mathcal{T}} \|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}'')\|_{\boldsymbol{\Sigma}^{-1}} \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_{\boldsymbol{\Sigma}}\\ &\leq \left\|\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta} - \sum_{l=1}^k \gamma_l\boldsymbol{\mu}_l\right)\right\|_{\boldsymbol{\Sigma}^{-1}} \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_{\boldsymbol{\Sigma}}\\ &= \left\|\boldsymbol{\theta} - \sum_{l=1}^k \gamma_l\boldsymbol{\mu}_l\right\|_{\boldsymbol{\Sigma}} \varepsilon_\lambda\varrho(\boldsymbol{\theta},\boldsymbol{\theta}') \leq 2R_{\boldsymbol{\mu}}\varepsilon_\lambda\varrho(\boldsymbol{\theta},\boldsymbol{\theta}')\end{aligned}$$

□

E.3 Proof of Lemma 6.4.5

Proof of Lemma 6.4.5. The proof is extremely similar to the Dirac case, with different constants and bounds.

Consider $\boldsymbol{\theta} \in \mathcal{T}$, and $\eta \leq 1/2$. As in the Dirac case we must build a set $\mathcal{V}_{\boldsymbol{\theta}}$ that is close to two sets:

$$S_1 = \left\{\frac{\pi_{\boldsymbol{\theta}} - a\pi_{\boldsymbol{\theta}'}}{\|\pi_{\boldsymbol{\theta}} - a\pi_{\boldsymbol{\theta}'}\|_\kappa} \;\middle|\; \boldsymbol{\theta}' \in \mathcal{T},\ \varrho(\boldsymbol{\theta},\boldsymbol{\theta}') \leq 1,\ a \in [0;1],\ \|\pi_{\boldsymbol{\theta}} - a\pi_{\boldsymbol{\theta}'}\|_\kappa \leq \eta\right\} \tag{E.2}$$

$$S_2 = \left\{\frac{a\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}}{\|a\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}\|_\kappa} \;\middle|\; \boldsymbol{\theta}' \in \mathcal{T},\ \varrho(\boldsymbol{\theta},\boldsymbol{\theta}') \leq 1,\ a \in [0;1],\ \|a\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}\|_\kappa \leq \eta\right\} \tag{E.3}$$

Consider any $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2 \in \mathcal{T}$ and $a \in [0;1]$ such that $\varrho(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \leq 1$ and $\|\pi_{\boldsymbol{\theta}_1} - a\pi_{\boldsymbol{\theta}_2}\|_\kappa \leq \eta$. We are going to approach $\mu = \frac{\pi_{\boldsymbol{\theta}_1} - a\pi_{\boldsymbol{\theta}_2}}{\|\pi_{\boldsymbol{\theta}_1} - a\pi_{\boldsymbol{\theta}_2}\|_\kappa}$ with some tampered distribution ν. We have

$$\|\mu-\nu\|_{\mathcal{F}_R} = \sup_{\boldsymbol{\omega}} C_\lambda \left|\frac{e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_1} - ae^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_2}}{\|\pi_{\boldsymbol{\theta}_1} - a\pi_{\boldsymbol{\theta}_2}\|_\kappa}\cdot e^{-\frac{1}{2}\boldsymbol{\omega}^\top\boldsymbol{\Sigma}\boldsymbol{\omega}} - \psi_\nu(\boldsymbol{\omega})\right| \tag{E.4}$$

where ψ_ν is the characteristic function of ν.

Denote $\Delta_{\boldsymbol{\theta}} = \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1$, $b = 1 - a$ and $\alpha = \|\pi_{\boldsymbol{\theta}_1} - a\pi_{\boldsymbol{\theta}_2}\|_\kappa$. Also denote $\Delta_0 = \Delta_{\boldsymbol{\theta}}/\alpha$ and $b_0 = b/\alpha$, and finally $f_{12} = f(\varrho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2))$.

With these notations, the first term in (E.4) reads

$$\frac{e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_1} - ae^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_2}}{\|\pi_{\boldsymbol{\theta}_1} - a\pi_{\boldsymbol{\theta}_2}\|_\kappa} \cdot e^{-\frac{1}{2}\boldsymbol{\omega}^\top\boldsymbol{\Sigma}\boldsymbol{\omega}} = \left[e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_1}\left(\frac{1 - e^{\mathrm{i}\alpha\boldsymbol{\omega}^\top\Delta_0}}{\alpha}\right) + b_0 e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_2}\right] e^{-\frac{1}{2}\boldsymbol{\omega}^\top\boldsymbol{\Sigma}\boldsymbol{\omega}}$$

With the same computations than the proof of Lemma 6.3.4, it holds that $b_0 \leq 1$ and $\|\Delta_0\|_{\boldsymbol{\Sigma}} \leq r_\Delta$ with

$$r_\Delta = 2\sqrt{\frac{1 + \lambda^2/2}{\sigma_k^2}} \tag{E.5}$$

And by Taylor expansion

$$\left|\frac{1 - e^{\mathrm{i}\alpha\boldsymbol{\omega}^\top\Delta_0}}{\alpha} + \mathrm{i}\boldsymbol{\omega}^\top\Delta_0\right| \leq \left|\boldsymbol{\omega}^\top\Delta_0\right|^2 \frac{\alpha}{2} \leq \|\boldsymbol{\omega}\|_{\boldsymbol{\Sigma}^{-1}}^2 \|\Delta_0\|_{\boldsymbol{\Sigma}}^2 \frac{\alpha}{2} \leq \|\boldsymbol{\omega}\|_{\boldsymbol{\Sigma}^{-1}}^2 \eta \frac{2 + \lambda^2}{\sigma_k^2}.$$

Hence, if we define $\nu = -\pi'_{\boldsymbol{\theta}_1,\Delta_0} + b_0\pi_{\boldsymbol{\theta}_2}$ where $\pi'_{\boldsymbol{\theta}_1,\Delta_0}$ is the derivative of the Gaussian distribution with mean $\boldsymbol{\theta}_1$ and covariance $\boldsymbol{\Sigma}$ along direction Δ_0, *i.e.* such that $\psi_\nu(\boldsymbol{\omega}) = \left(-\mathrm{i}\boldsymbol{\omega}^\top\Delta_0 e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_1} + b_0 e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_2}\right) e^{-\frac{1}{2}\boldsymbol{\omega}\boldsymbol{\Sigma}\boldsymbol{\omega}}$, we have

$$\begin{aligned}\|\mu - \nu\|_{\mathcal{F}_R} &= \sup_{\boldsymbol{\omega}} C_\lambda \left|\frac{1 - e^{\mathrm{i}\alpha\boldsymbol{\omega}^\top\Delta_0}}{\alpha} + \mathrm{i}\boldsymbol{\omega}^\top\Delta_0\right| e^{-\frac{1}{2}\boldsymbol{\omega}^\top\boldsymbol{\Sigma}\boldsymbol{\omega}} \leq \frac{C_\lambda(2 + \lambda^2)\eta}{\sigma_k^2} \sup_{\boldsymbol{\omega}} \|\boldsymbol{\omega}\|_{\boldsymbol{\Sigma}^{-1}}^2 e^{-\frac{1}{2}\|\boldsymbol{\omega}\|_{\boldsymbol{\Sigma}^{-1}}^2} \\ &\leq \frac{C_\lambda(2 + \lambda^2)\eta}{\sigma_k^2} \sup_{R \in \mathbb{R}_+} Re^{-\frac{1}{2}R} = \frac{2C_\lambda(2 + \lambda^2)\eta}{e\sigma_k^2} = M\eta\,.\end{aligned}$$

by a quick study of the function $R \mapsto Re^{-\frac{1}{2}R}$, where $e = e^1$.

Using this property with $\boldsymbol{\theta}_1 = \boldsymbol{\theta}$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}$, for all $\mu \in S_1$ there is ν in

$$\mathcal{V}_1 = \left\{-\pi'_{\boldsymbol{\theta},\Delta} + b\pi_{\boldsymbol{\theta}'} \;\middle|\; \boldsymbol{\theta}' \in \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1),\ \Delta \in \mathcal{B}_{\mathbb{R}^d,\|\cdot\|_{\boldsymbol{\Sigma}}}(0, r_\Delta),\ b \in [0; 1]\right\}$$

such that $\|\mu - \nu\|_{\mathcal{F}_R} \leq M\eta$. Similarly, by symmetry for all $\mu \in S_2$ there is ν in

$$\mathcal{V}_2 = \left\{\pi'_{\boldsymbol{\theta}',\Delta} - b\pi_{\boldsymbol{\theta}} \;\middle|\; \boldsymbol{\theta}' \in \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1),\ \Delta \in \mathcal{B}_{\mathbb{R}^d,\|\cdot\|_{\boldsymbol{\Sigma}}}(0, r_\Delta),\ b \in [0; 1]\right\}$$

such that $\|\mu - \nu\|_{\mathcal{F}_R} \leq M\eta$. Hence we define $\mathcal{V}_{\boldsymbol{\theta}} = \mathcal{V}_1 \cup \mathcal{V}_2$ the set of tampered distributions satisfies the desired property, and must now bound its covering numbers.

Consider the product space $X = \mathcal{B}_{\mathbb{R}^d,\|\cdot\|_{\boldsymbol{\Sigma}}}(0, r_\Delta) \times \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta}, 1) \times [0; 1]$.

Let us begin with $\mathcal{V}_1$. Given $x = (\Delta, \boldsymbol{\theta}', b) \in X$, define the function $\varphi : X \mapsto \mathcal{V}_1$ by $\varphi(x) = -\pi'_{\Delta,\boldsymbol{\theta}} + b\pi_{\boldsymbol{\theta}'}$. For $x_1 = (\Delta_1, \boldsymbol{\theta}_1, b_1)$ and $x_2 = (\Delta_2, \boldsymbol{\theta}_2, b_2)$ in X, we have

$$\begin{aligned}\|\varphi(x_1) - \varphi(x_2)\|_{\mathcal{F}_R} &= \left\|-\pi'_{\Delta_1,\boldsymbol{\theta}} + b_1\pi_{\boldsymbol{\theta}_1} + \pi'_{\Delta_2,\boldsymbol{\theta}} - b_2\pi_{\boldsymbol{\theta}_2}\right\|_{\mathcal{F}_R} \\ &\leq \left\|\pi'_{\Delta_1,\boldsymbol{\theta}} - \pi'_{\Delta_2,\boldsymbol{\theta}}\right\|_{\mathcal{F}_R} + \|b_1\pi_{\boldsymbol{\theta}_1} - b_1\pi_{\boldsymbol{\theta}_2}\|_{\mathcal{F}_R} + \|b_1\pi_{\boldsymbol{\theta}_2} - b_2\pi_{\boldsymbol{\theta}_2}\|_{\mathcal{F}_R}\end{aligned}$$

We bound each of those terms. First,

$$\begin{aligned}\left\|\pi'_{\Delta_1,\boldsymbol{\theta}} - \pi'_{\Delta_2,\boldsymbol{\theta}}\right\|_{\mathcal{F}_R} &= \sup_{\boldsymbol{\omega}} C_\lambda \left|\boldsymbol{\omega}^\top(\Delta_1 - \Delta_2)\right| e^{-\frac{1}{2}\boldsymbol{\omega}^\top\boldsymbol{\Sigma}\boldsymbol{\omega}} \\ &\leq C_\lambda \|\Delta_1 - \Delta_2\|_{\boldsymbol{\Sigma}} \sup_{\boldsymbol{\omega}} \|\boldsymbol{\omega}\|_{\boldsymbol{\Sigma}^{-1}} e^{-\frac{1}{2}\|\boldsymbol{\omega}\|_{\boldsymbol{\Sigma}^{-1}}^2} = L_1 \|\Delta_1 - \Delta_2\|_{\boldsymbol{\Sigma}}\end{aligned}$$

where $L_1 = C_\lambda/\sqrt{e}$, since $\sup_{R\in\mathbb{R}_+} Re^{-\frac{1}{2}R^2} = e^{-\frac{1}{2}}$. Then we have

$$\begin{aligned}\|b_1\pi_{\boldsymbol{\theta}_1} - b_1\pi_{\boldsymbol{\theta}_2}\|_{\mathcal{F}_R} \leq \|\pi_{\boldsymbol{\theta}_1} - \pi_{\boldsymbol{\theta}_2}\|_{\mathcal{F}_R} &= \sup_{\boldsymbol{\omega}} C_\lambda \left|e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_1} - e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_2}\right| e^{-\frac{1}{2}\boldsymbol{\omega}^\top\boldsymbol{\Sigma}\boldsymbol{\omega}} \\ &\leq C_\lambda \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{\boldsymbol{\Sigma}} \sup_{\boldsymbol{\omega}} \|\boldsymbol{\omega}\|_{\boldsymbol{\Sigma}^{-1}} e^{-\frac{1}{2}\|\boldsymbol{\omega}\|^2_{\boldsymbol{\Sigma}^{-1}}} \leq L_2\varrho(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\end{aligned}$$

with $L_2 = C_\lambda\sigma_k^{-1}\sqrt{2+\lambda^2}$. Finally,

$$\|b_1\pi_{\boldsymbol{\theta}_2} - b_2\pi_{\boldsymbol{\theta}_2}\|_{\mathcal{F}_R} \leq \|\pi_{\boldsymbol{\theta}_2}\|_{\mathcal{F}_R} |b_1 - b_2| \leq C_\lambda |b_1 - b_2| \ .$$

Therefore if we denote $\mathcal{C}_1$ a $\frac{\delta}{3L_1}$-covering of $\mathcal{B}_{\mathbb{R}^d,\|\cdot\|_{\boldsymbol{\Sigma}}}(0,r_\Delta)$ for the norm $\|\cdot\|_{\boldsymbol{\Sigma}}$ and $\mathcal{C}_2$ a $\frac{\delta}{3L_2}$-covering of $\mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta},1)$ for the metric ϱ and $\mathcal{C}_3$ a $\frac{\delta}{3C_\lambda}$-covering of $[0;1]$, for any $x \in X$ there exists an element $\bar{x} \in \mathcal{C}_1\times\mathcal{C}_2\times\mathcal{C}_3$ such that $\|\varphi(x) - \varphi(x')\|_{\mathcal{F}_R} \leq \delta$. Thus we have

$$\begin{aligned}\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R}, \mathcal{V}_1, \delta\right) &\leq |\mathcal{C}_1|\cdot|\mathcal{C}_2|\cdot|\mathcal{C}_3| \\ &\leq \mathcal{N}\left(\|\cdot\|_{\boldsymbol{\Sigma}}, \mathcal{B}_{\mathbb{R}^d,\|\cdot\|_{\boldsymbol{\Sigma}}}(0,r_\Delta), \frac{\delta}{3L_1}\right)\cdot\mathcal{N}\left(\varrho, \mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta},1), \frac{\delta}{3L_2}\right)\cdot\mathcal{N}\left(\|\cdot\|_1, [0;1], \frac{\delta}{3C_\lambda}\right) \\ &\leq \max\left(\left(\frac{12r_\Delta L_1}{\delta}\right)^d, 1\right)\cdot\max\left(\left(\frac{12L_2}{\delta}\right)^d, 1\right)\cdot\max\left(\frac{3C_\lambda}{\delta}, 1\right)\end{aligned}$$

Then the $\max(\cdot,1)$ are resolved by using the fact that $\delta \leq 1$.

We now turn to $\mathcal{V}_2$. Given $x = (\Delta,\boldsymbol{\theta}',b) \in X$, define the function $\varphi : X \mapsto \mathcal{V}_2$ by $\varphi(x) = \pi'_{\Delta,\boldsymbol{\theta}'} - b\pi_{\boldsymbol{\theta}}$. For $x_1 = (\Delta_1,\boldsymbol{\theta}_1,b_1)$ and $x_2 = (\Delta_2,\boldsymbol{\theta}_2,b_2)$ in X, we have

$$\begin{aligned}\|\varphi(x_1) - \varphi(x_2)\|_{\mathcal{F}_R} &= \left\|\pi'_{\Delta_1,\boldsymbol{\theta}_1} - b_1\pi_{\boldsymbol{\theta}} - \pi'_{\Delta_2,\boldsymbol{\theta}_2} + b_2\pi_{\boldsymbol{\theta}}\right\|_{\mathcal{F}_R} \\ &\leq \left\|\pi'_{\Delta_1,\boldsymbol{\theta}_1} - \pi'_{\Delta_2,\boldsymbol{\theta}_1}\right\|_{\mathcal{F}_R} + \left\|\pi'_{\Delta_2,\boldsymbol{\theta}_1} - \pi'_{\Delta_2,\boldsymbol{\theta}_2}\right\|_{\mathcal{F}_R} + \|b_1\pi_{\boldsymbol{\theta}} - b_2\pi_{\boldsymbol{\theta}}\|_{\mathcal{F}_R}\end{aligned}$$

Again, we bound each of those terms. Similar to the previous case, we have

$$\left\|\pi'_{\Delta_1,\boldsymbol{\theta}_1} - \pi'_{\Delta_2,\boldsymbol{\theta}_1}\right\|_{\mathcal{F}_R} \leq L_1\|\Delta_1 - \Delta_2\|_{\boldsymbol{\Sigma}} .$$

Then

$$\begin{aligned}\left\|\pi'_{\Delta_2,\boldsymbol{\theta}_1} - \pi'_{\Delta_2,\boldsymbol{\theta}_2}\right\|_{\mathcal{F}_R} &= \sup_{\boldsymbol{\omega}} C_\lambda \left|\boldsymbol{\omega}^\top\Delta_2\right| \left|e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_1} - e^{\mathrm{i}\boldsymbol{\omega}^\top\boldsymbol{\theta}_2}\right| e^{-\frac{1}{2}\boldsymbol{\omega}^\top\boldsymbol{\Sigma}\boldsymbol{\omega}} \\ &\leq \sup_{\boldsymbol{\omega}} C_\lambda \left|\boldsymbol{\omega}^\top\Delta_2\right| \left|\boldsymbol{\omega}^\top(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\right| e^{-\frac{1}{2}\|\boldsymbol{\omega}\|^2_{\boldsymbol{\Sigma}^{-1}}} \\ &\leq \sup_{\boldsymbol{\omega}} C_\lambda \|\boldsymbol{\omega}\|^2_{\boldsymbol{\Sigma}^{-1}} \|\Delta_2\|_{\boldsymbol{\Sigma}} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{\boldsymbol{\Sigma}} e^{-\frac{1}{2}\|\boldsymbol{\omega}\|^2_{\boldsymbol{\Sigma}^{-1}}} \\ &\leq Cr_\Delta \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_1\|_{\boldsymbol{\Sigma}} \sup_{R\in\mathbb{R}_+} Re^{-\frac{1}{2}R} = L_3\varrho(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\end{aligned}$$

where $L_3 = \frac{2C_\lambda r_\Delta\sqrt{2+\lambda^2}}{e\sigma_k} = \frac{2\sqrt{2}C_\lambda(2+\lambda^2)}{e\sigma_k^2}$. And finally,

$$\|b_1\pi_{\boldsymbol{\theta}} - b_2\pi_{\boldsymbol{\theta}}\|_{\mathcal{F}_R} \leq \|\pi_{\boldsymbol{\theta}}\|_{\mathcal{F}_R} |b_1 - b_2| \leq C_\lambda |b_1 - b_2| .$$

Therefore if we denote $\mathcal{C}_1$ a $\frac{\delta}{3L_1}$-covering of $\mathcal{B}_{\mathbb{R}^d,\|\cdot\|_{\boldsymbol{\Sigma}}}(0,r_\Delta)$, $\mathcal{C}_2$ a $\frac{\delta}{3L_3}$-covering of $\mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta},1)$ and $\mathcal{C}_3$ a $\frac{\delta}{3C_\lambda}$-covering of $[0,1]$, similar to the previous case we have

$$\begin{aligned}\mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R},\mathcal{V}_1,\delta\right) &\leq |\mathcal{C}_1|\cdot|\mathcal{C}_2|\cdot|\mathcal{C}_3| \\ &\leq \mathcal{N}\left(\|\cdot\|_{\boldsymbol{\Sigma}},\mathcal{B}_{\mathbb{R}^d,\|\cdot\|_{\boldsymbol{\Sigma}}}(0,r_\Delta),\frac{\delta}{3L_1}\right)\cdot\mathcal{N}\left(\varrho,\mathcal{B}_{\mathcal{T},\varrho}(\boldsymbol{\theta},1),\frac{\delta}{3L_3}\right)\cdot\mathcal{N}\left(\|\cdot\|_1,[0,1],\frac{\delta}{3C_\lambda}\right) \\ &\leq \max\left(\left(\frac{12r_\Delta L_1}{\delta}\right)^d,1\right)\cdot\max\left(\left(\frac{12L_3}{\delta}\right)^d,1\right)\cdot\max\left(\frac{3C_\lambda}{\delta},1\right)\end{aligned}$$

Since we have $L_3 \geq L_2$, we bound the covering numbers of S_1 by those of S_2 and obtain the desired expression. □

E.4 Proof of Theorem 6.4.6

Proof of Theorem 6.4.6. Applying all the Lemmas of Section 6.4.2, all the hypotheses necessary to apply Theorem 6.2.12 hold. We obtain the following constants:

- Compatibility constant: given in equation (6.66);
- Admissibility constant: we have

$$W_0 = C_\lambda\left(2+\frac{2+\lambda^2}{\sigma_k^2}\right)^{\frac{1}{2}} = \mathcal{O}\left(e^{\frac{d}{2\lambda^2}}\sqrt{\log(k)(1+\lambda^2)}\right)$$

  and the admissibility constant

$$W_\Lambda = 4C_\lambda\sqrt{k}\left(1+\frac{1+\lambda^2/2}{\sigma_k^2}\right)^{\frac{1}{2}} = \mathcal{O}\left(e^{\frac{d}{2\lambda^2}}\sqrt{k\log(k)(1+\lambda^2)}\right)$$

- Finally, according to Lemma 6.4.5 there exist tangent sets $\mathcal{V}_{\boldsymbol{\theta}}$ such that their covering numbers have a common bound independent of $\boldsymbol{\theta}$. Combining this bound with Lemma 6.4.4 and equation (6.30), we get:

$$N = \mathcal{N}\left(\|\cdot\|_{\mathcal{F}_R},\mathcal{S}^0(\cdot,\mathfrak{S}),\frac{1}{4}\right) \leq \left[A_1A_2^d + A_3A_4^d\right]^k$$

  with

$$\begin{aligned}A_1 &= 3\cdot 2^{31}k^2W_0^3C_\lambda^2\frac{2+\lambda^2}{e\sigma_k^2} \\ A_2 &= 2^{27}e^{-1}k\sqrt{2k}C_\lambda^3W_0R_{\boldsymbol{\mu}}\left(\frac{2+\lambda^2}{\sigma_k^2}\right)^{\frac{3}{2}} \\ A_3 &= 3\cdot 2^{22}k\sqrt{2k}C_\lambda W_0^2 \\ A_4 &= 3^2 2^{32}\sqrt{2k}C_\lambda^3R_{\boldsymbol{\mu}}\left(\frac{2+\lambda^2}{e\sigma_k^2}\right)^{\frac{3}{2}}\end{aligned}$$

  and the sketch size (6.68) indeed scales as

$$\begin{aligned}m \geq cW_\Lambda^2\log\left(\frac{N}{\rho}\right) &= \mathcal{O}\left(e^{\frac{d}{\lambda^2}}(1+\lambda^2)k^2d\log(C_\lambda)\texttt{polylog}\left(k,d,R_{\boldsymbol{\mu}},\frac{1}{\rho}\right)\right) \\ &= \mathcal{O}\left(e^{\frac{d}{\lambda^2}}(1+\lambda^2)k^2d\frac{d}{\lambda^2}\texttt{polylog}\left(k,d,R_{\boldsymbol{\mu}},\frac{1}{\rho}\right)\right) \\ &= \mathcal{O}\left(e^{\frac{d}{\lambda^2}}k^2d^2\texttt{polylog}\left(k,d,R_{\boldsymbol{\mu}},\frac{1}{\rho}\right)\right)\end{aligned}$$

□

E.5 Kernel for rotated 2-dimensional Gaussians

In this last section we prove the expression of the mean kernel 7.1 given in the outlooks of the thesis in Chapter 7. in future investigations we will aim at proving that this mean kernel satisfies (or not) the conditions of Chapter 6.

Recall that we consider 2-dimensional gaussians with "flat" rotated covariance, defined as $\pi_\theta = \mathcal{N}\left(0, \mathbf{R}_\theta \text{diag}\left([\sigma_1^2, \sigma_2^2]\right)\mathbf{R}_\theta^\top\right)$ where

$$\mathbf{R}_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

such that

$$\mathbf{\Sigma}_\theta = R_\theta \mathbf{\Sigma} R_\theta^\top = \begin{bmatrix} \sigma_1^2 C_\theta^2 + \sigma_2^2 S_\theta^2 & (\sigma_1^2 - \sigma_2^2) C_\theta S_\theta \\ (\sigma_1^2 - \sigma_2^2) C_\theta S_\theta & \sigma_1^2 S_\theta^2 + \sigma_2^2 C_\theta^2 \end{bmatrix} \tag{E.6}$$

where $C_\theta = \cos(\theta)$ and $S_\theta = \sin(\theta)$. We consider a Gaussian kernel $\kappa(\mathbf{z}, \mathbf{z}') = e^{-\frac{1}{2}\|\mathbf{z}-\mathbf{z}'\|_2^2}$. In that case, using Lemma 6.4.1 the mean kernel expresses $\kappa(\pi_\theta, \pi_{\theta'}) = 1/\left|\mathbf{I} + \mathbf{\Sigma}_\theta + \mathbf{\Sigma}_\theta\right|^{\frac{1}{2}}$. Denote $C = \cos(\theta)$ and $C' = \cos(\theta')$ (resp. $S = \sin(\theta)$ and $S' = \sin(\theta')$), $C_+ = \cos(\theta + \theta')$ (resp. $S_+ = \sin(\theta + \theta')$) and $C_- = \cos(\theta - \theta')$. Simple trigonometry yields that $C^2 + (C')^2 = 1 + C_+ C_-$, $S^2 + (S')^2 = 1 - C_+ C_-$ and $SC + S'C' = S_+ C_-$. From here we can derive

$$\begin{aligned}
\left|\mathbf{I} + \mathbf{\Sigma}_\theta + \mathbf{\Sigma}_{\theta'}\right| =& \Big(1 + \sigma_1^2(C^2 + (C')^2) + \sigma_2^2(S^2 + (S')^2)\Big)\Big(1 + \sigma_1^2(S^2 + (S')^2) + \sigma_2^2(C^2 + (C')^2)\Big) \\
& - (\sigma_1^2 - \sigma_2^2)^2 (SC' + S'C)^2 \\
=& \Big(1 + \sigma_1^2 + \sigma_2^2 + (\sigma_1^2 - \sigma_2^2) C_+ C_-\Big)\Big(1 + \sigma_1^2 + \sigma_2^2 - (\sigma_1^2 - \sigma_2^2) C_+ C_-\Big) - (\sigma_1^2 - \sigma_2^2)^2 S_+^2 C_-^2 \\
=& (1 + \sigma_1^2 + \sigma_2^2)^2 - (\sigma_1^2 - \sigma_2^2)^2 (C_+^2 C_-^2 + S_+^2 C_-^2) \\
=& (1 + \sigma_1^2 + \sigma_2^2)^2 - (\sigma_1^2 - \sigma_2^2)^2 C_-^2 \\
=& (1 + \sigma_1^2 + \sigma_2^2)^2 - (\sigma_1^2 - \sigma_2^2)^2 + (\sigma_1^2 - \sigma_2^2)^2 S_-^2
\end{aligned}$$

which is the desired result.

Bibliography

[Ach01] Dimitris Achlioptas. "Database-friendly random projections". In: *Principles of database systems (PODS)*. 2001, pp. 274–281.

[AD03] Christophe Andrieu and Arnaud Doucet. "Online expectation-maximization type algorithms for parameter estimation in general state space models." In: *ICASSP (6)* 4 (2003).

[Ahr05] Peter Ahrendt. *The Multivariate Gaussian Probability Distribution*. Tech. rep. IMM, Technical University of Denmark, 2005.

[AJM08] Virtudes Alba Fernández, M. Dolores Jiménez Gamero, and J. Muñoz García. "A test for the two-sample problem based on empirical characteristic functions". In: *Computational Statistics and Data Analysis* 52.7 (2008), pp. 3730–3748.

[AJM09] Nir Ailon, Ragesh Jaiswal, and Claire Monteleoni. "Streaming k -means approximation". In: *Advances in Neural Information Processing Systems (NIPS)* (2009), pp. 10–18.

[Alo+09] Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. "NP-hardness of Euclidean sum-of-squares clustering". In: *Machine Learning* 75.2 (2009), pp. 245–248.

[AM05] Dimitris Achlioptas and Frank Mcsherry. "On spectral learning of mixtures of distributions". In: *Learning Theory* (2005), pp. 458–469.

[And+13] Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James Voss. "The more, the merrier: the blessing of dimensionality for learning large gaussian mixtures". In: *arXiv:1311.2891* 35 (2013), pp. 1–30.

[Aro50] Nachman Aronzajn. "Theory of reproducing kernels". In: *Transactions of the American Mathematical Society* 68 (1950), pp. 337–404.

[ATL16] Salem Alelyani, Jiliang Tang, and Huan Liu. "Feature Selection for Clustering : A Review". In: *Data Clustering: Algorithms and Applications* (2016), pp. 29–60.

[AV07] David Arthur and Sergei Vassilvitskii. "k-means++: The Advantages of Careful Seeding". In: *ACM-SIAM symposium on Discrete algorithms*. 2007, pp. 1027–1035.

[Bac15] Francis Bach. "On the Equivalence between Quadrature Rules and Random Features". In: (2015), pp. 1–25. arXiv: arXiv:1502.06800v1.

[Bar+08] Richard Baraniuk, Mark Davenport, Ronald A Devore, and Michael Wakin. "A simple proof of the restricted isometry property for random matrices". In: *Constructive Approximation* 28.3 (2008), pp. 253–263.

[Bar07] Richard Baraniuk. "Compressive sensing". In: *IEEE Signal Processing Magazine* 24.2 (2007), pp. 118–121.

[BD08a] Thomas Blumensath and Mike E. Davies. "Gradient pursuit for non-linear sparse signal modelling". In: *European Signal Processing Conference (EUSIPCO)*. 2008.

[BD08b] Thomas Blumensath and Mike E. Davies. "Iterative thresholding for sparse approximations". In: *Journal of Fourier Analysis and Applications* 14.5-6 (2008), pp. 629–654.

[BD09] Thomas Blumensath and Mike E. Davies. "Iterative hard thresholding for compressed sensing". In: *Applied and Computational Harmonic Analysis* 27.3 (2009), pp. 265–274. arXiv: arXiv:0805.0510v1.

[Bec13] Stephen Becker. "L-BFGSB-C". In: *https://github.com/stephenbeckr/L-BFGS-B-C* (2013).

[BGP13] Anthony Bourrier, Rémi Gribonval, and Patrick Pérez. "Compressive gaussian mixture estimation". In: *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. 2013, pp. 6024–6028.

[BGP15] Anthony Bourrier, Rémi Gribonval, and Patrick Pérez. "Compressive Gaussian Mixture estimation". In: *Compressed Sensing and its Applications - MATHEON Workshop 2013*. Ed. by Haulger Boche, Robert Calderbank, Gitta Kutyniok, and Jan Vybiral. Birkhäuser Basel, 2015, pp. 6024–6028.

[BKK13] Nikoletta Bassiou, Constantine Kotropoulos, and Evangelia Koliopoulou. "Symmetric alpha-Stable Sparse Linear Regression for Musical Audio Denoising". In: Ispa (2013), pp. 375–380.

[BLJ04] Francis Bach, Gert R.G. Lanckriet, and Michael I. Jordan. "Multiple kernel learning, conic duality, and the SMO algorithm". In: *International Conference on Machine Learning (ICML)*. 2004.

[Blu11] Thomas Blumensath. "Sampling and reconstructing signals from a union of linear subspaces". In: *IEEE Transactions on Information Theory* 57.7 (2011), pp. 4660–4671. arXiv: arXiv:0911.3514v2.

[Bor+06] Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans Peter Kriegel, Bernhard Schölkopf, and Alexander J. Smola. "Integrating structured biological data by Kernel Maximum Mean Discrepancy". In: *Bioinformatics* 22.14 (2006), pp. 49–57.

[Bou+14] Anthony Bourrier, Mike E. Davies, Tomer Peleg, and Rémi Gribonval. "Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems". In: *IEEE Transactions on Information Theory* 60.12 (2014), pp. 7928–7946.

[Bou14] Anthony Bourrier. "Compressed sensing and dimensionality reduction for unsupervised learning". PhD thesis. 2014.

[BP12] Kristian Bredies and Hanna Katriina Pikkarainen. "Inverse problems in spaces of measures". In: *ESAIM: Control, Optimisation and Calculus of Variations* 19.1 (2012), pp. 190–218.

[BRH08] Eran Bashan, Raviv Raich, and Alfred O. Hero. "Optimal two-stage search for sparse targets using convex criteria". In: *IEEE Transactions on Signal Processing* 56.11 (2008), pp. 5389–5402.

[BRM15] Petros T. Boufounos, Shantanu Rane, and Hassan Mansour. "Representation and Coding of Signal Geometry". In: *arXiv:1512.07636 [cs, math]* (2015), pp. 1–28. arXiv: 1512.07636.

[Bro+13] Simon A. Broda, Markus Haas, Jochen Krause, Marc S. Paolella, and Sven C. Steude. "Stable mixture GARCH models". In: *Journal of Econometrics* 172.2 (2013), pp. 292–306.

[Bro05] Mike Brooks. *VOICEBOX: Speech Processing Toolbox for MATLAB*. 2005.

[BS09] Liefeng Bo and Cristian Sminchisescu. "Efficient Match Kernels between Sets of Features for Visual Recognition". In: *Advances in Neural Information Processing System (NIPS)*. 2009.

[BS10a] Mikhail Belkin and Kaushik Sinha. "Polynomial learning of distribution families". In: *IEEE 51st Annual Symposium on Foundations of Computer Science*. Ieee, 2010. arXiv: arXiv:1004.4864v1.

[BS10b] Mikhail Belkin and Kaushik Sinha. "Toward Learning Gaussian Mixtures with Arbitrary Separation". In: *Conference On Learning Theory (COLT)*. 2010. arXiv: arXiv:0907.1054v2.

[BSR15] Nicholas Boyd, Geoffrey Schiebinger, and Benjamin Recht. "The Alternating Descent Conditional Gradient Method for Sparse Inverse Problems". In: (2015), pp. 1–21. arXiv: 1507.01562.

[BTA04] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Vol. 2004. C. Kluwer Academics Publisher, 2004.

[Bun+10] Florentina Bunea, Alexandre B. Tsybakov, Marten H. Wegkamp, and Adrian Barbu. "SPADES and mixture models". In: *The Annals of Statistics* 38.4 (2010), pp. 2525–2558. arXiv: arXiv:0901.2044v2.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[Byr+95] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. "A Limited Memory Algorithm for Bound Constrained Optimization". In: *SIAM Journal on Scientific Computing* 16.5 (1995), pp. 1190–1208. arXiv: arXiv:1011.1669v3.

[BZD10] Christos Boutsidis, A. Zouzias, and Petros Drineas. "Random Projections for k-means Clustering". In: *Advances in Neural Information and Processing Systems (NIPS)*. 2010, pp. 298–306.

[Can08] Emmanuel J. Candès. "The restricted isometry property and its implications for compressed sensing". In: *Comptes Rendus Mathematique* 346.9-10 (2008), pp. 589–592.

[Cas04] Roberto Casarin. "Bayesian Inference for Mixtures of Stable Distributions". In: *SSRN Electronic Journal* (2004), pp. 1–50.

[CDD09] Albert Cohen, Wolfgang Dahmen, and Ronald A Devore. "Compressed sensing and best k-term approximation". In: *Journal of the American mathematical Society* 22.1 (2009), pp. 211–231.

[CF00] Marine Carrasco and Jean-Pierre Florens. "Generalization of GMM to a continuum of moment conditions". In: *Econometric Theory* (2000).

[CF02] Marine Carrasco and Jean-Pierre Florens. "Efficient GMM estimation using the empirical characteristic function". In: *IDEI Working Paper* (2002).

[CF14] Marine Carrasco and Jean-Pierre Florens. "On The Asymptotic Efficiency Of Gmm". In: *Econometric Theory* 30.02 (2014), pp. 372–406.

[CFG12] Emmanuel J. Candès and Carlos Fernandez-Granda. "Super-Resolution from Noisy Data". In: October 2012 (2012), pp. 1–22. arXiv: 1211.0290.

[CGJ96] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. "Active Learning with Statistical Models". In: *Journal of Artificial Intelligence Research* 4 (1996), pp. 129–145. arXiv: 9603104 [cs].

[CH09] Graham Cormode and Marios Hadjieleftheriou. "Methods for finding frequent items in data streams". In: *The VLDB Journal* 19.1 (2009), pp. 3–20.

[Cha+12] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. "The convex geometry of linear inverse problems". In: *Foundations of Computational Mathematics (2012)* (2012). arXiv: arXiv:1012.0621v3.

[Cha+15] Olivier Chabiron, François Malgouyres, Jean Yves Tourneret, and Nicolas Dobigeon. "Toward Fast Transform Learning". In: *International Journal of Computer Vision* 114.2-3 (2015), pp. 195–216.

[Cha17] Antoine Chatalic. "Towards Scalable Sketched Learning". PhD thesis. 2017.

[Chw+15] Kacper Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. "Fast Two-Sample Testing with Analytic Representations of Probability Measures". In: *Advances in Neural Information Processing Systems (NIPS)*. 2015. arXiv: 1506.04725.

[CJJ12] Radha Chitta, Rong Jin, and Anil K Jain. "Efficient Kernel Clustering using Random Fourier Features". In: *IEEE 12th International Conference on Data Mining* (2012), pp. 161–170.

[CJS09] Robert Calderbank, Sina Jafarpour, and Robert Schapire. *Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain*. Tech. rep. 2009.

[CM05] Graham Cormode and S. Muthukrishnan. "An improved data stream summary: the count-min sketch and its applications". In: *Journal of Algorithms* 55.1 (2005), pp. 58–75.

[CM09] Olivier Cappé and Eric Moulines. "Online EM Algorithm for Latent Data Models". In: *Journal of the Royal Statistical Society* 71.3 (2009), pp. 593–613. arXiv: 0712.4273.

[Coh+15] Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. "Dimensionality Reduction for k-Means Clustering and Low Rank Approximation". In: *arXiv:1410.6801* (2015), p. 37. arXiv: 1410.6801.

[Cor+04] Graham Cormode, Flip Korn, S. Muthukrishnan, and Srivastava Divesh. "Diamond in the rough: Finding hierarchical heavy hitters in multi-dimensional data". In: *International Conference on Management of Data*. 2004, pp. 155–166.

[Cor+11] Graham Cormode, Minos Garofalakis, Peter J. Haas, and Chris Jermaine. "Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches". In: *Foundations and Trends in Databases* 4.xx (2011), pp. 1–294.

[CP11] Emmanuel J. Candès and Yaniv Plan. "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements". In: *IEEE Transactions on Information Theory* 57.4 (2011), pp. 2342–2359. arXiv: 1001.0339.

[CRT06a] Emmanuel J. Candès, Justin K. Romberg, and Terrence Tao. "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on Information Theory* 52.2 (2006), pp. 480–509.

[CRT06b] Emmanuel J. Candès, Justin K. Romberg, and Terrence Tao. "Stable signal recovery from incomplete and inaccurate measurements". In: *Communications on Pure and Applied Mathematics* 59.8 (2006), pp. 1207–1223.

[CRW17] Krzysztof Choromanski, Mark Rowland, and Adrian Weller. "The Unreasonable Effectiveness of Random Orthogonal Embeddings". In: (2017). arXiv: 1703.00864.

[CS02] Felipe Cucker and Steve Smale. "On the mathematical foundations of learning". In: *Bulletin of the American Mathematical Society* 39.1 (2002), pp. 1–49.

[CS16] B. Chandra and Rajesh K. Sharma. "Fast learning in Deep Neural Networks". In: *Neurocomputing* 171 (2016), pp. 1205–1215. arXiv: 1508.01887.

[CT04] Emmanuel J. Candès and Terrence Tao. "Decoding by linear programming". In: *IEEE Transactions on Information Theory* 51.12 (2004), pp. 4203–4215.

[CT06] Emmanuel J. Candès and Terrence Tao. "Near-optimal signal recovery from random projections: Universal encoding strategies?" In: *IEEE Transactions on Information Theory* 52.12 (2006), pp. 5406–5425.

[CT10] Emmanuel J. Candès and Terrence Tao. "The power of convex relaxation: Near-optimal matrix completion". In: *IEEE Transactions on Information Theory* 56.5 (2010), pp. 2053–2080. arXiv: 0903.1476.

[CV16] Krzysztof Choromanski and Sindhwani Vikas. "Recycling Randomness with Structure for Sublinear time Kernel Expansions". In: *arXiv:1605.09049v1* (2016). arXiv: arXiv:1605.09049v1.

[Das+05] Anirban Dasgupta, John Hopcroft, Jon Kleinberg, and Mark Sandler. "On learning mixtures of heavy-tailed distributions". In: *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS* 2005 (2005), pp. 491–500.

[Das99] Sanjoy Dasgupta. "Learning mixtures of Gaussians". In: *IEEE 51st Annual Symposium on Foundations of Computer Science*. May. 1999.

[DeC+15] Yohann DeCastro, Fabrice Gamboa, Didier Henrion, and Jean-Bernard Lasserre. "Exact solutions to Super Resolution on semi-algebraic domains in higher dimensions". In: *arXiv preprint arXiv:1502.02436* (2015), pp. 1–22. arXiv: 1502.02436.

[DFH14] Antoine Deleforge, Florence Forbes, and Radu Horaud. "High-dimensional regression with gaussian mixtures and partially-latent response variables". In: *Statistics and Computing* 25.5 (2014), pp. 893–911. arXiv: arXiv:1308.2302v3.

[DLR77] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B* 39.1 (1977), pp. 1–38.

[Don06] David L. Donoho. "Compressed sensing". In: *IEEE Transactions on Information Theory* 52.4 (2006), pp. 1289–1306.

[Don15] David L. Donoho. "50 years of Data Science". In: (2015), pp. 1–41.

[DP15] Vincent Duval and Gabriel Peyré. "Exact Support Recovery for Sparse Spikes Deconvolution". In: *Foundations of Computational Mathematics* 15.5 (2015), pp. 1315–1355. arXiv: 1306.6909.

[Duc07] John Duchi. *Derivations for Linear Algebra and Optimization*. Tech. rep. 2007.

[Fel+10] Dan Feldman, Morteza Monemizadeh, Christian Sohler, and David P Woodruff. "Coresets and Sketches for High Dimensional Subspace Approximation Problems". In: 1 (2010), pp. 630–649.

[FFK11] Dan Feldman, Matthew Faulkner, and Andreas Krause. "Scalable Training of Mixture Models via Coresets". In: *Proceedings of Neural Information Processing Systems* (2011), pp. 1–9.

[FHT03] Alexei A Fedotov, Peter Harremoës, and Flemming Topsøe. "Refinements of Pinsker's Inequality". In: *IEEE Transactions on Information Theory* 49.6 (2003), pp. 1491–1498.

[Fis12] Tom Fischer. "Existence, uniqueness, and minimality of the Jordan measure decomposition". In: *arXiv:1206.5449* (2012). arXiv: 1206.5449.

[FL11] Dan Feldman and Michael Langberg. "A unified framework for approximating and clustering data". In: *Proceedings of the forty-third annual ACM symposium on Theory of computing* 46109 (2011), pp. 569–578. arXiv: 1106.1379.

[FM03] Dmitriy Fradkin and David Madigan. "Experiments with random projections for machine learning". In: *International Conference on Knowledge discovery and data mining (KDD)*. 2003, pp. 517–522.

[FM77] Andrey Feuerverger and RA Mureika. "The empirical characteristic function and its applications". In: *The annals of Statistics* (1977).

[FM81] Andrey Feuerverger and Philip McDunnough. "On Some Fourier methods for Inference". In: *Journal of the American Statistical Association* 76.374 (1981), pp. 379–387.

[FR13] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. NY: Springer New York, 2013.

[FS05] Gereon Frahling and Christian Sohler. "A fast k -means implementation using coresets". In: *Proceedings of the twenty-second annual symposium on Computational geometry (SoCG)* 18.6 (2005), pp. 605–625.

[Fuk+07] K. Fukumizu, Arthur Gretton, X. Sun, and Bernhard Schölkopf. "Kernel Measures of Conditional Dependence". In: *Advances in Neural Information Processing System (NIPS)*. 2007.

[Gil+02] Anna C. Gilbert, Yannis Kotidis, S. Muthukrishnan, and Martin J. Strauss. "How to summarize the universe: Dynamic maintenance of quantiles". In: *International Conference on Very Large Data Bases (VLBD)*. 2002, pp. 454–465.

[GK99] Simon Godsill and Ercan E. Kuruoglu. "Bayesian inference for time series with heavy-tailed symmetric alpha -stable noise processes". In: (1999), pp. 1–28.

[GLA16] Mohammed Ghesmoune, Mustapha Lebbah, and Hanene Azzag. "State-of-the-art on clustering data streams". In: *Big Data Analytics* 1.1 (2016), p. 13.

[GPP16] Mina Ghashami, Daniel Perry, and Jeff M. Phillips. "Streaming Kernel Principal Component Analysis". In: *International Conference on Artificial Intelligence and Statistics* 41 (2016), pp. 1–16. arXiv: 1512.05059.

[Gre+06] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. "A Kernel Method for the Two-Sample Problem". In: *Advances in Neural Information Processing Systems (NIPS)*. 2006. arXiv: 0805.2368.

[Gre+12] Arthur Gretton, Bharath K. Sriperumbudur, Dino Sejdinovic, Heiko Strathmann, and Massimiliano Pontil. "Optimal kernel choice for large-scale two-sample tests". In: *Advances in Neural Information Processing Systems (NIPS)* (2012), pp. 1214–1222.

[Gri+17] Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, and Yann Traonmilin. "Compressive Statistical Learning with Random Feature Moments". In: *arXiv:1706.07180* (2017). arXiv: 1706.07180.

[Gri+96] Rémi Gribonval, Emmanuel Bacry, Stéphane Mallat, Philippe Depalle, and Xavier Rodet. "Analysis of sound signals with high resolution matching pursuit". In: *IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*. 1996, pp. 125–128.

[GSB15] Raja Giryes, Guillermo Sapiro, and Alex M Bronstein. "Deep Neural Networks with Random Gaussian Weights : A Universal Classification Strategy ?" In: *arXiv:1504.08291* (2015). arXiv: arXiv:1504.08291v1.

[Guh+00] Sudipto Guha, Nina Mishra, Rajeev Motwani, and Liadan OCallaghan. "Clustering Data Streams". In: *Proc. Ann. Symp. Foundations of Computer Science*. 2000.

[Hal05] Alastair R. Hall. *Generalized method of moments*. 2005.

[Hau+11] Jarvis Haupt, Rui Castro, Robert Nowak, and S T May. "Distilled Sensing : Adaptive Sampling for Sparse Detection and Estimation". In: *IEEE Transactions on Information Theory* 57.9 (2011), pp. 6222–6235. arXiv: arXiv:1001.5311v2.

[HK13] Daniel Hsu and Sham M. Kakade. "Learning mixtures of spherical gaussians: moment methods and spectral decompositions". In: *Conference on Innovations in Theoretical Computer Science*. 2013. arXiv: arXiv:1206.5766v4.

[HPM04] Sariel Har-Peled and Soham Mazumdar. "Coresets for k-Means and k-Median Clustering and their Applications". In: *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*. 2004, pp. 291–300.

[Jit+16] Wittawat Jitkrittum, Zoltán Szabó, Kacper Chwialkowski, and Arthur Gretton. "Interpretable Distribution Features with Maximum Testing Power". In: *Advances in Neural Information and Processing Systems (NIPS)* Nips (2016), pp. 1–21. arXiv: 1605.06796.

[Jol02] I T Jolliffe. "Principal Component Analysis, Second Edition". In: *Encyclopedia of Statistics in Behavioral Science* 30.3 (2002), p. 487. arXiv: arXiv:1011.1669v3.

[Jos+10] Sarang Joshi, Raj Varma Kommaraju, Jeff M. Phillips, and Suresh Venkatasubramanian. "Comparing Distributions and Shapes using the Kernel Distance". In: *Proceedings of the twenty-seventh annual symposium on Computational geometry (SoCG)*. 2010, pp. 47–56. arXiv: 1001.0591.

[JTD11] Prateek Jain, Ambuj Tewari, and Inderjit S. Dhillon. "Orthogonal matching pursuit with replacement". In: *Advances in Neural Information Processing Systems (NIPS)*. 2011, pp. 1215–1223.

[KCW16] Michael Kapralov, Comcast Cable, and David P Woodruff. "How to Fake Multiply by a Gaussian Matrix". In: *Icml* (2016), pp. 1–37. arXiv: 1606.05732.

[Ker+16] Nicolas Keriven, Anthony Bourrier, Rémi Gribonval, and Patrick Pérèz. "Sketching for Large-Scale Learning of Mixture Models". In: *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. 2016.

[Ker+17a] Nicolas Keriven, Nicolas Tremblay, Yann Traonmilin, and Rémi Gribonval. "Compressive K-means". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017. arXiv: 1610.08738.

[Ker+17b] Nicolas Keriven, Anthony Bourrier, Rémi Gribonval, and Patrick Pérèz. "Sketching for Large-Scale Learning of Mixture Models". In: *arXiv preprint arXiv:1606.02838* (2017), pp. 1–50. arXiv: 1606.02838.

[Ker16] Nicolas Keriven. "SketchMLbox : a Matlab toolbox for large-scale learning of mixture models". In: *http://sketchml.gforge.inria.fr* (2016).

[KL51] S. Kullback and R. A. Leibler. "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86. arXiv: 1511.00860.

[Kri+09] Sebastian Kring, Svetlozar T. Rachev, Markus Höchstötter, and Frank J. Fabozzi. "Estimation of alpha-stable sub-Gaussian distributions for asset returns". In: *Contributions to Economics* (2009), pp. 111–152.

[KW70] George S Kimeldorf and Grace Wahba. "A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines". In: *Ann. Math. Stat.* 41 (1970), pp. 495–502.

[KY02] J Knight and J Yu. *Empirical Characteristic Function in Time Series Estimation*. Vol. 18. 2002, pp. 691–721.

[Lan87] Henry J. Landau. *Moments in mathematics*. 1987.

[LCB07] Gaëlle Loosli, Stéphane Canu, and Léon Bottou. "Training Invariant Support Vector Machines using Selective Sampling". In: *Large Scale Kernel Machines* (2007), pp. 301–320.

[LCB98] Yann Lecun, Corinna Cortes, and Christopher JC Burges. *The MNIST database of handwritten digits*. 1998.

[LeM16] Luc LeMagoarou. "Matrices efficientes pour le traitement du signal et l'apprentissage automatique". PhD thesis. 2016.

[Les+04] Christina S. Leslie, Eleazar Eskin, Adiel Cohen, Jason Weston, and William Stafford Noble. "Mismatch string kernels for discriminative protein classification". In: *Bioinformatics* 20.4 (2004), pp. 467–476.

[LG16] Luc LeMagoarou and Rémi Gribonval. "Flexible Multi-layer Sparse Approximations of Matrices and Applications". In: *IEEE Journal of Selected Topics in Signal Processing* 10.4 (2016), pp. 688–700. arXiv: 1506.07300.

[LH95] Charles L. Lawson and Richard J. Hanson. "Solving least squares problems". In: *SIAM classics in applied mathematics* (1995).

[Llo82] Stuart P. Lloyd. "Least Squares Quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–137.

[Lod+02] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. "Text Classification using String Kernels". In: *Journal of Machine Learning Research* 2 (2002), pp. 419–444.

[LSS13] Quoc V Le, Tamás Sarlós, and Alexander J. Smola. "Fastfood — Approximating Kernel Expansions in Loglinear Time". In: *International Conference on Machine Learning (ICML)* 28.1 (2013). arXiv: arXiv:1408.3060v1.

[Luc+17] Mario Lucic, Matthew Faulkner, Andreas Krause, and Dan Feldman. "Training Mixture Models at Scale via Coresets". In: (2017). arXiv: 1703.08110.

[Mar63] Donald W. Marquardt. *An Algorithm for Least-Squares Estimation of Nonlinear Parameters*. 1963. arXiv: arXiv:1011.1669v3.

[MG09] Boris Mailhé and Rémi Gribonval. "LocOMP: algorithme localement orthogonal pour l'approximation parcimonieuse rapide de signaux longs sur des dictionnaires locaux". In: *Actes du XXIIe colloque GRETSI*. 2009.

[ML09] Marius Muja and David G Lowe. "Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration". In: *International Conference on Computer Vision Theory and Applications (VISAPP '09)* (2009), pp. 1–10. arXiv: 10.1.1.160.1721.

[MM09] Oldaric-Ambrym Maillard and Rémi Munos. "Compressed Least-Squares Regression". In: *Advances in Neural Information and Processing Systems (NIPS)*. 2009.

[MNT04] K Madsen, H B Nielsen, and O Tingleff. "Methods for non-linear least squares problems". In: *Infomatics and Mathematical Modeling* 2 (2004), pp. 1–30. arXiv: /dx.doi.org/10.1023{\%}2FA{\%}3A1010933404324 [http:].

[Mua+12] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. "Learning from distributions via support measure machines". In: *Advances in Neural Information Processing Systems (NIPS)* (2012).

[Mua+14] Krikamol Muandet, Kenji Fukumizu, Bharath K. Sriperumbudur, Arthur Gretton, and Bernhard Schölkopf. "Kernel Mean Estimation and Stein's Effect". In: *31st International Conference on Machine Learning* 32 (2014), pp. 10–18. arXiv: arXiv:1306.0842v2.

[Muk16] Yusuke Mukuta. "Kernel Approximation via Empirical Orthogonal Decomposition for Unsupervised Feature Learning". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 5222–5230.

[Mul97] A. Muller. "Integral Probability Metrics and Their Generating Classes of Functions". In: *Advances in Applied Probability* 29.2 (1997), pp. 429–443.

[MZ93] Stéphane Mallat and Zhifeng Zhang. "Matching Pursuit With Time-Frequency Dictionaries". In: *IEEE Transactions in Signal Processing* 41.12 (1993), pp. 3397–3415.

[Nam+13] Sangnam Nam, Mike E. Davies, Michael Elad, and Rémi Gribonval. "The cosparse analysis model and algorithms". In: *Applied and Computational Harmonic Analysis* 34.1 (2013), pp. 30–56. arXiv: 1106.4987.

[NJW01] Andrew Y Ng, Michael I. Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm". In: *Advances in Neural Information Processing Systems 14* (2001), pp. 849–856.

[NM94] Whitney K. Newey and Daniel McFadden. "Large sample estimation and hypothesis testing". In: *Handbook of Econometrics*. Vol. 4. 1994, pp. 2111–2245.

[Nol03] John P. Nolan. "Modeling financial data with stable distributions". In: *Handbook of Heavy Tailed Distributions in Finance: Handbooks in Finance* 1.May 2002 (2003), pp. 105–129.

[Nol13] John P. Nolan. "Multivariate elliptically contoured stable distributions: Theory and estimation". In: *Computational Statistics* 28.5 (2013), pp. 2067–2089.

[NPM01] John P. Nolan, a. K. Panorska, and J. H. McCulloch. "Estimation of stable spectral measures". In: *Mathematical and Computer Modelling* 34.9-11 (2001), pp. 1113–1122.

[NT09] Deanna Needell and Joel A. Tropp. "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples". In: *Applied and Computational Harmonic Analysis* 26.3 (2009), pp. 301–321. arXiv: 0803.2392.

[Oli+15] Junier B. Oliva, Avinava Dubey, Barnabas Poczos, Jeff Schneider, and Eric P. Xing. "Bayesian Nonparametric Kernel-Learning". In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2015. arXiv: 1506.08776.

[Ome15] Vadym Omelchenko. "Parameter estimation of sub-Gaussian stable distributions". In: *Kybernetika* June (2015), pp. 929–949.

[OSS15] Junier B. Oliva, Dougal J. Sutherland, and Jeff Schneider. "Deep Mean Maps". In: *arXiv:1511.04150* (2015). arXiv: arXiv:1511.04150v1.

[PAB16] Farhad Pourkamali-Anaraki and Stephen Becker. "Randomized Clustered Nystrom for Large-Scale Kernel Machines". In: (2016), pp. 1–31. arXiv: 1612.06470.

[PDG15] Gilles Puy, Mike E. Davies, and Rémi Gribonval. "Linear embeddings of low-dimensional subsets of a Hilbert space to R m". In: *European Signal Processing Conference (EUSIPCO)*. 2015, pp. 469–473.

[PGD13] Tomer Peleg, Rémi Gribonval, and Mike E. Davies. "Compressed Sensing and Best Approximation from Unions of Subspaces: Beyond Dictionaries". In: *21st European Signal Processing Conference (EUSIPCO 2013)* (2013).

[PRK93] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad. "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition". In: *Asilomar Conference on Signals, Systems and Computers*. 1993.

[PSW16] Brooks Paige, Dino Sejdinovic, and Frank Wood. "Super-Sampling with a Reservoir". In: *Uncertainty in Artificial Intelligence* (2016).

[PY15] Jeffrey Pennington and Felix X Yu. "Spherical Random Features for Polynomial Kernels". In: (2015).

[Ran71] William M. Rand. "Objective Criteria for the Evaluation of Clustering Methods". In: *Journal of American Statistical Association* 66.336 (1971), pp. 846–850.

[Rau08] Holger Rauhut. "On the impossibility of uniform sparse reconstruction using greedy methods". In: *Sampling Theory in Signal and Image* (2008), pp. 1–15.

[Reb+13] Hugo Reboredo, Francesco Renna, Robert Calderbank, and Miguel R. D. Rodrigues. "Compressive Classification". In: *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. 2013, pp. 1029–1032. arXiv: 1302.4660.

[Red+15] Sashank J. Reddi, Aaditya Ramdas, Barnabás Póczos, Aarti Singh, and Larry Wasserman. "On the Decreasing Power of Kernel and Distance based Nonparametric Hypothesis Tests in High Dimensions". In: *AAAI Conference on Artifical Intelligence*. 2015. arXiv: 1406.2083.

[RFP07] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization". In: *Optimization Online* 52.3 (2007), pp. 1–33. arXiv: 0706.4138.

[Rob11] J. C. Robinson. *Dimensions, embeddings, and attractors*. 2011.

[RQD00] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. "Speaker Verification Using Adapted Gaussian Mixture Models". In: *Digital Signal Processing* 10.1-3 (2000), pp. 19–41.

[RR07] Ali Rahimi and Benjamin Recht. "Random Features for Large Scale Kernel Machines". In: *Advances in Neural Information Processing Systems (NIPS)* (2007).

[RR09] Ali Rahimi and Benjamin Recht. "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning". In: *Advances in Neural Information Processing Systems (NIPS)* (2009).

[RR95] Douglas A. Reynolds and Richard C. Rose. "Robust text-independent speaker identification using Gaussian mixture speaker models". In: *IEEE Transactions on Speech and Audio Processing* 3.1 (1995), pp. 72–83.

[Rud62] Walter Rudin. *Fourier Analysis on Groups*. Interscience Publishers, 1962.

[Rud87] Walter Rudin. *Real and complex analysis*. McGraw-Hil. 1987.

[Rud91] Walter Rudin. *Functional Analysis*. 1991, pp. xv+424. arXiv: arXiv:1011.1669v3.

[Saa+16] Alaa Saade, Francesco Caltagirone, Igor Carron, Laurent Daudet, Angélique Dremeau, Sylvain Gigan, and Florent Krzakala. "Random projections through multiple optical scattering: Approximating Kernels at the speed of light". In: *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)* 2016-May (2016), pp. 6215–6219. arXiv: 1510.06664.

[Sch15] Bernhard Schölkopf. "Computing Functions of Random Variables via Reproducing Kernel Hilbert Space Representations". In: *Statistics and Computing* 25.4 (2015), pp. 755–766. arXiv: arXiv:1501.06794v1.

[Sch17] Vincent Schellekens. "Compressive Clustering of High-Dimensional Datasets by 1-bit Sketching". PhD thesis. 2017.

[SD16] Aman Sinha and John Duchi. "Learning Kernels with Random Features". In: *Advances in Neural Information and Processing Systems (NIPS)*. 2016.

[SGKR09] Diego Salas-Gonzalez, Ercan E. Kuruoglu, and Diego P. Ruiz. "Finite mixture of alpha-stable distributions". In: *Digital Signal Processing* 19.2 (2009), pp. 250–264.

[SGKR10] Diego Salas-Gonzalez, Ercan E. Kuruoglu, and Diego P. Ruiz. "Modelling with mixture of symmetric stable distributions using Gibbs sampling". In: *Signal Processing* 90.3 (2010), pp. 774–783.

[Sho+10] S. R Hosseini Shojaei, V. Nassiri, Gh R. Mohammadian, and A. Mohammadpour. "Mixture of skewed alpha-stable distributions". In: *AIP Conference Proceedings* 1305.March 2011 (2010), pp. 130–137.

[SHS01] Bernhard Schölkopf, Ralf Herbrich, and Alexander J. Smola. "A Generalized Representer Theorem". In: *COLT* (2001), pp. 416–426.

[Smo+07] Alexander J. Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. "A Hilbert Space Embedding for Distributions". In: *International Conference on Algorithmic Learning Theory*. 2007, pp. 13–31.

[Son+06] Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. "Large Scale Multiple Kernel Learning". In: *Journal of Machine Learning Research* 7 (2006), pp. 1531–1565.

[Son+08] Le Song, X Zhang, Alexander J. Smola, Arthur Gretton, and Bernhard Schölkopf. "Tailoring density estimation via reproducing kernel moment matching". In: *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)* (2008), pp. 992–9.

[Son08] Le Song. "Learning via Hilbert space embedding of distributions". PhD thesis. 2008.

[SQW15] Ju Sun, Qing Qu, and John Wright. "When Are Nonconvex Problems Not Scary?" In: *arXiv:1510.06096* (2015), pp. 1–6. arXiv: 1510.06096.

[Sri+09] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Gert R.G. Lanckriet, and Bernhard Schölkopf. "Kernel choice and classifiability for RKHS embeddings of probability distributions". In: *Advances in Neural Information and Processing Systems (NIPS)*. 2009.

[Sri+10] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. "Hilbert space embeddings and metrics on probability measures". In: *The Journal of Machine Learning Research* 11 (2010), pp. 1517–1561.

[Sri02] Karthik Sridharan. *A Gentle Introduction to Concentration Inequalities*. Tech. rep. 2002.

[Sri11] Bharath K. Sriperumbudur. "Mixture density estimation via hilbert space embedding of measures". In: *IEEE International Symposium on Information Theory*. 2011, pp. 1027–1030.

[SS01] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. 2001. arXiv: arXiv:1011.1669v3.

[SS07] Dougal J. Sutherland and Jeff Schneider. "On the Error of Random Fourier Features". In: (2007). arXiv: arXiv:1506.02785v1.

[SS15] Bharath K. Sriperumbudur and Zoltán Szabó. "Optimal rates for Random Fourier Features". In: *Advances in Neural Information Processing Systems (NIPS)* (2015), pp. 1144–1152. arXiv: 1506.0215.

[SS16] Ryan Spring and Anshumali Shrivastava. "Scalable and Sustainable Deep Learning via Randomized Hashing". In: (2016). arXiv: 1602.08194.

[SSH13] S.O. Sadjadi, M. Slaney, and L. Heck. "MSR Identity Toolbox v1.0: A MATLAB toolbox for speaker-recognition Research." In: *Speech and Language Processing Technical Committee Newsletter* (2013).

[Ste56] H. Steinhaus. "Sur la division des corps materiels en parties". In: *Bull. Acad. Polon. Sci. IV (C1.III)* IV.12 (1956), pp. 801–804.

[Sut+15] Dougal J. Sutherland, Junier B. Oliva, Poczos Barnabas, and Jeff Schneider. "Linear-time Learning on Distributions with Approximate Kernel Embeddings". In: *arXiv:1509.07553* (2015).

[TAL14] Jiliang Tang, Salem Alelyani, and Huan Liu. "Feature Selection for Classification: A Review". In: *Data Classification: Algorithms and Applications* (2014), pp. 37–64.

[TG15] Yann Traonmilin and Rémi Gribonval. "Stable recovery of low-dimensional cones in Hilbert spaces : One RIP to rule them all". In: (2015).

[Tha+02] Nitin Thaper, Sudipto Guha, Piotr Indyk, and Nick Koudas. "Dynamic multidimensional histograms". In: *International conference on Management of data (SIGMOD)*. 2002, pp. 428–439.

[Tib11] Robert Tibshirani. "Regression shrinkage and selection via the lasso: a retrospective". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.3 (2011), pp. 273–282. arXiv: 11/73273 [1369-7412].

[Tra98] Kien C. Tran. "Estimating mixtures of normal distributions via empirical characteristic function". In: *Econometric Reviews* 17.2 (1998), pp. 167–183.

[Tre+16a] Nicolas Tremblay, Gilles Puy, Pierre Borgnat, Rémi Gribonval, and Pierre Vandergheynst. "Accelerated Spectral Clustering Using Graph Filtering Of Random Signals". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016, pp. 4094–4098. arXiv: 1509.08863.

[Tre+16b] Nicolas Tremblay, Gilles Puy, Rémi Gribonval, and Pierre Vandergheynst. "Compressive Spectral Clustering". In: *Proceedings of The 33rd International Conference on Machine Learning* (2016), pp. 1–15. arXiv: 1602.02018.

[Tse01] P. Tseng. "Convergence of a block coordinate descent method for nondifferentiable minimization". In: *Journal of Optimization Theory and Applications* 109.3 (2001), pp. 475–494.

[Vap95] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., 1995.

[Vei12] Mark Veillette. "STBL: a MATLAB library for working with alpha stable distributions". In: *https://github.com/markveillette/stbl* (2012).

[VF10] Andrea Vedaldi and Brian Fulkerson. *VLFeat - An open and portable library of computer vision algorithms*. Tech. rep. 2010.

[Vis+10] S.V.N. Vishwanathan, Nicol Schraudolph, Risi Kondor, and Karsten M. Borgwardt. "Graph Kernels". In: *Journal of Machine Learning Research* 11 (2010), pp. 1201–1242. arXiv: 0807.0093.

[VW04] Santosh Vempala and Grant Wang. "A spectral algorithm for learning mixture models". In: *Journal of Computer and System Sciences* 68.4 (2004), pp. 841–860.

[VZ12] Andrea Vedaldi and Andrew Zisserman. "Efficient additive kernels via explicit feature maps". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.3 (2012), pp. 480–492.

[WA13] Andrew Gordon Wilson and Ryan P. Adams. "Gaussian process kernels for pattern discovery and extrapolation". In: *International Conference on Machine Learning (ICML)*. Vol. 28. 3. 2013, pp. 1067–1075. arXiv: 1302.4245.

[WS01] Christopher Williams and Matthias W. Seeger. "Using the Nystrom Method to Speed Up Kernel Machines". In: *NIPS Proceedings* 13 (2001), pp. 682–688. arXiv: arXiv:1011.1669v3.

[Xin+16] Felix Xinnan, Yu Ananda, Theertha Suresh, Krzysztof Choromanski, Daniel Holtmann-Rice, and Sanjiv Kumar. "Orthogonal Random Features". In: *Advances in Neural Information and Processing Systems (NIPS)*. 2016. arXiv: arXiv:1610.09072v1.

[XK10] Dinghai Xu and John Knight. "Continuous Empirical Characteristic Function Estimation of Mixtures of Normal Parameters". In: *Econometric Reviews* 30.1 (2010), pp. 25–50.

[XY13] Yangyang Xu and Wotao Yin. "A Block Coordinate Descent Method for Regularized Multiconvex Optimization with Applications to Nonnegative Tensor Factorization and Completion". In: *SIAM Journal on Imaging Sciences* 6.3 (2013), pp. 1758–1789.

[Yan+15] Zichao Yang, Alexander J. Smola, Le Song, and Andrew Gordon Wilson. "A la Carte — Learning Fast Kernels". In: *International Conference on Artificial Intelligence and Statistics (AISTATS)* 38 (2015), pp. 1098–1106. arXiv: 1412.6493.

[ZM15] Ji Zhao and Deyu Meng. "FastMMD: Ensemble of Circular Discrepancy for Efficient Two-Sample Test". In: *Journal Neural Computation* 27.6 (2015), pp. 1345–1372. arXiv: 1405.2664.