High dimensional multiple means estimation and testing with applications to machine learning

Jean-Baptiste Fermanian

université PARIS-SACLAY

High dimensional multiple means estimation and testing with applications to machine learning

Test et estimation multiple de moyennes en grande dimension avec applications à l'apprentissage automatique

Thèse de doctorat de l'Université Paris-Saclay

École doctorale de Mathématique Hadamard n° 574 (EDMH)
Spécialité de doctorat : Mathématiques appliquées
Graduate School : Mathématiques. Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Laboratoire de mathématiques d'Orsay (Université Paris-Saclay, CNRS)** sous la direction de **Gilles BLANCHARD**, professeur, et la co-direction de **Magalie FROMONT**, professeure.

Thèse soutenue à Paris-Saclay, le 4 Octobre 2024, par

Jean-Baptiste FERMANIAN

Composition du jury

Membres du jury avec voix délibérative

Elisabeth GASSIAT Professeure, Université Paris-Saclay	Présidente
Vladimir KOLTCHINSKII Professeur, Georgia Institute of Technology	Rapporteur & Examinateur
Samory KPOTUFE Professeur, Columbia University	Rapporteur & Examinateur
Alexandra CARPENTIER Professeure, Pötsdam University	Examinatrice
Etienne ROQUAIN Maître de conférence, Sorbonne Université	Examinateur
Alexandre TSYBAKOV Professeur, ENSAE Paris	Examinateur

THESE DE DOCTORAT

NNT : 2024UPASM035

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre: Test et estimation multiple de moyennes en grande dimension avec applications à l'apprentissage automatique.
Mots clés: Grande dimension (effective), estimation multiple de moyennes, test de proximité, vitesse minimax, kernel mean embedding, self-attention.

Résumé: Nous étudions dans cette thèse l'influence de la grande dimension dans des problèmes de test et d'estimation. Notre analyse porte sur la dépendance en la dimension de la vitesse de séparation d'un test de proximité et du risque quadratique de l'estimation multiples de vecteurs. Nous complétons les résultats existants en étudiant ces dépendances dans le cas de distributions non isotropes. Pour de telles distributions, le rôle de la dimension est alors joué par des notions de dimension effective définies à partir de la covariance des distributions. Ce cadre permet d'englober des données de dimension infinie comme le kernel mean embedding, outil de machine learning que nous chercherons à estimer. À l'aide de cette analyse, nous construisons des méthodes d'estimation simultanée de vecteurs moyennes de différentes distributions à partir d'échantillons indépendants de chacune. Ces estimateurs ont de meilleures performances théorique et pratique relativement aux moyennes empiriques, en particulier dans des situations défavorables où la dimension (effective) est grande. Ces méthodes utilisent explicitement ou implicitement la relative facilité du test par rapport à l'estimation. Elles reposent sur la construction d'estimateurs de distances et de moments de la covariance pour lesquels nous fournissons des bornes de concentration non asymptotiques. Un intérêt particulier est porté à l'étude de données bornées pour lesquels une analyse spécifique est nécessaire. Nos méthodes sont accompagnées d'une analyse minimax justifiant leur optimalité. Dans une dernière partie, nous proposons une interprétation du mécanisme d'attention utilisé dans les réseaux de neurones Transformers comme un problème d'estimation multiple de vecteurs. Dans un cadre simplifié, ce mécanisme partage des idées similaires avec nos approches et nous mettons en évidence son effet de débruitage en grande dimension.

Title: High dimensional multiple means estimation and testing with applications to machine learning.
Keywords: High (effective) dimension, multiple mean estimation, closeness testing, minimax rate, kernel mean embedding, self-attention.

Abstract: In this thesis, we study the influence of high dimension in testing and estimation problems. We analyze the dimension dependence of the separation rate of a closeness test and of the quadratic risk of multiple vector estimation. We complement existing results by studying these dependencies in the case of non-isotropic distributions. For such distributions, the role of dimension is played by notions of effective dimension defined from the covariance of the distributions. This framework covers infinite-dimensional data such as kernel mean embedding, a machine learning tool we will be seeking to estimate. Using this analysis, we construct methods for simultaneously estimating mean vectors of different distributions from independent samples of each. These estimators perform better theoretically and practically than the empirical mean in unfavorable situations where the (effective) dimension is large. These methods make explicit or implicit use of the relative ease of testing compared with estimation. They are based on the construction of estimators of distances and moments of covariance, for which we provide non-asymptotic concentration bounds. Particular interest is given to the study of bounded data, for which a specific analysis is required. Our methods are accompanied by a minimax analysis justifying their optimality. In a final section, we propose an interpretation of the attention mechanism used in Transformer neural networks as a multiple vector estimation problem. In a simplified framework, this mechanism shares similar ideas with our approaches, and we highlight its denoising effect in high dimension.


université
PARIS-SACLAY
FACULTÉ
DES SCIENCES
D'ORSAY


Laboratoire de
Mathématiques
d'Orsay


Fondation mathématique
FMJH
Jacques Hadamard

Remerciements

Je voudrais tout d'abord remercier mon directeur et ma directrice de thèse, Gilles et Magalie, pour m'avoir permis de mener cette thèse. Merci à toi Gilles pour ta passion, ta gentillesse et la régularité de ton accompagnement que tu sois à Orsay, au Mans, Oxford ou Berlin. Merci pour tes relectures malgré mon anglais parfois approximatif et mes démonstrations souvent obscures. Sache quand même que tu m'as converti à ton art du parenthésage, je ne me passe plus de tes macros. Magalie, merci à toi de m'avoir orienté vers ce projet. Tes dernières années ont été bien remplies, et je te remercie d'avoir gardé malgré tout le lien. Tes conseils et avis ont toujours été précieux.

I would like to thank my referees, Vladimir and Samory, to have taken the time to read and comment my thesis. You did me a great honor. Un grand merci aussi à Alexandra, Elisabeth, Etienne et Alexandre d'avoir accepté d'être membres de mon jury. Je vous en suis très reconnaissant.

Merci à l'équipe "doctorant de Gilles" qui avez souvent aussi été mes supports administratifs-trains, hôtels, démarches de soutenance, templates... Romain, tu as capté beaucoup trop vite que j'avais des vrais soucis avec ça, c'était presque vexant ! Bastien, j'appréciais beaucoup tes arrivées soudaines dans le bureau pour trainasser, y'avait pas de débat c'était l'heure de la pause. Ce fut très agréable de faire ces années ensemble. Olympio je suis vraiment très heureux de t'avoir rencontré dès ces premiers mois à Orsay. Tu m'as beaucoup fait rêvé avec tes woofing-hippie ou red-neck, vacances en stop et autres aventures. Nos conférences ensembles me requinquaient toujours efficacement !

J'aimerais remercier tout ceux que j'ai pu fréquenté de près ou de plus loin durant ces années au LMO. Elles ont été rythmées par de nombreux groupes de travail : réseaux de neurones, transformers et surtout la prédiction conforme. Pour ce dernier ce fut la belle occasion de te rencontrer Pierre et de travailler avec toi aux quatres coins de Paris, de Orsay à Bourse en passant par François Miterrand et Jussieu. On a plein de super projets et j'ai hâte de les mener à leur termes. Merci à Guillermo, Evgenii, Karl, Elisabeth, Zacharie, Sylvain, Margaux, Romain, Ulysse, Etienne R., Eric, Cyril, Etienne B. pour ces échanges durant ces différents GTs. Merci en particulier à toi Christophe pour ces discussions toujours très instructives sur mes différents sujets. Merci aux co-bureaux, Valentin et, bien sûr, Benjamin pour toujours avoir amené ta vitalité dans ce bureau. Merci aux nouveaux et anciens voisins de palier du 2A21 Bertrand, Pierre-André, Wojciech, Laure. Mention spéciale à Samy le spécialiste des langues extrème orientales. Pour finir dans la catégorie Orsay, j'aimerais remercier pêle mêle Ibrahim, Vincent, Perrine, Marc, El Mehdi, David, Thomas pour ces bavardages, au coin d'un couloir ou d'une conférence, toujours très sympathiques.

Many thanks to you Hannah, a good part of this work has been done in partnership with you. It was very rewarding to have your perspective as a computer scientist on our ideas sometimes a bit non-practicable.

Un grand merci à tout les copains pour votre présence et votre amitié durant ces dernières années. Je pense que c'était pour moi un ingrédient essentiel pour aller jusqu'au bout de ce projet. Merci à la team ciné renno-parisienne Dorian, Razvan et Xing Feng, à laquelle je m'excuse d'avoir raté autant de séances, à la team rennaise-décentralisée, Hermès, Julian, Paul, Paul, Mathias et Julien, nos rendez vous ponctuels à des barycentres sont toujours un plaisir. Je suis bien content de te rejoindre Hermès et un peu triste de te rater à Paris, Paul. D'ailleurs, Julien il va falloir que tu rates tes trains à Montpellier à partir de maintenant. Juju, Pierrot, Prisci, Tomtom, Lance et Bibi, je pense vous n'avez pas encore réalisé mais maintenant plus aucun d'entre nous n'est étudiant. J'espère que cela ne vous fera pas un trop gros coup de vieux. Merci d'être là depuis toutes ces années. Je voudrais particulièrement te remercier Bibi, ainsi que Hannah et Raph, qui êtes tous les trois tellement moteur dans nos relations. L'énergie que vous mettez à me relancer malgré mes temps de réponse parfois un peu longs est vraiment très touchante. Marine, Loulou, Gobi, vous êtes vraiment des dingos et ça fait très plaisir de vous avoir dans ma vie, ne changez pas. Je voudrais remercier tous ceux avec qui j'ai partagé des verres, des apéros, des soirées jeux, échecs, salsa, spectacles, vraies vacances ou vacances TT ces dernières années, en particulier Pierre, Maxime, Charlotte, Tristan, David, Gus, Séverine, Viviane, Armance, Ines, Clément, David le frère, Hugo, Mathilde, Margot,

Julie, Marie, Lucie, Matsime, Marion, Thomas, Marthe. Merci pour ces moments.

Marie-Christine, Hervé, cette période de thèse sera toujours liée dans mon esprit à mon entrée dans la famille K. En faisant le bilan de ces années, c'est dur de ne pas penser à votre accueil chaleureux et votre bienveillance. Je vous en remercie ainsi que Mathieu, Marie-Mars, Agnès, Johannes et Léo pour ces soirées parisiennes. Cécile et Alex hâte de bientôt passer plus de temps avec vous au soleil, cela nous changera de Rouen ou Dijon le 30 décembre !

Adeline, Christelle et Grégoire vous êtes de sacrés excités mais que j'aime fort. Je suis heureux qu'on soit une fratrie soudée.

Papa, Maman on a de la chance que vous nous bichonnez autant, c'est presque indécent. Je vous remercie pour votre zenitude et votre ouverture. Vous m'avez toujours laissé mener ma barque tranquillement, tout en étant là au besoin. C'est une chance de vous avoir comme parents.

Alice, merci pour ton soutien constant. Cette thèse te doit énormément. Tu peux dire au revoir aujourd'hui à cette colocataire parfois envahissante. Je suis impatient de la suite de nos aventures.

Contents

I Introduction (English)

The analysis of large-scale data is a major current challenge in statistical and machine learning research. Modern methods have access to and seek to process increasingly complex data (signals, texts, images, videos, etc.). Although in some fields, this increase in data complexity comes with an increase in the quantity of data, in others the number of data items is limited, and taking this complexity into account becomes essential. A small amount of data, relative to its size, generally causes a loss of performance known as the *curse of dimensionality*. This term, introduced by Bellman (1966), covers a wide range of phenomena, some of which we will present below (see also Giraud, 2021).

In parallel with this loss of performance, small-dimensional intuitions are no longer necessarily relevant in large dimensions, and some a priori more strange methods can become interesting. A fundamental and well-known example of this type of phenomenon is the inadmissibility of the empirical mean in high dimension: Stein (1956) shows that to estimate the mean vector of a Gaussian distribution, the empirical mean is not efficient and exhibits a better estimator (James and Stein, 1961). This estimator, while having the same rate of convergence in sample size as the empirical mean, outperforms it thanks to a weaker dependence of its error on dimension. This example illustrates the need to include in the analysis of a problem not only dependence on sample size, but also on the size or complexity of the space. To capture this dependence, a non-asymptotic problem analysis is required. On the minimax side, the analysis of Pinsker (1980) shows, for example, that the James-Stein estimator is optimal for a fixed sample size, and for the dimension of the space tending to infinity. For testing problems, we find this consideration of dimension in the minimax analysis of the signal detection problem of Baraud (2002) and Blanchard et al. (2018) for example. To control the error of our methods and capture the effects of high dimensionality, we will use in this manuscript concentration inequalities to obtain non-asymptotic bounds.

Even if the data originally belong to a high-dimensional ambient space, it has been observed that they often actually live in lower-dimensional subspaces (vector subspace, submanifold, small numbers of clusters...), which makes the methods work. Although a data distribution is supported throughout the space, some directions may be uninformative and just consist of noise. The difficulty of the problem is then no longer characterized by the ambient dimension, but by notions of *effective dimensions*. These quantities, justified by minimax analysis, take into account the covariance structure of the data and quantify the degrees of freedom of a distribution. For example, we would like to say that a variable distributed on a straight line immersed in a high-dimensional ambient space is only of dimension one. These notions are particularly useful for analyzing functional data which, although in an infinite-dimensional space, may have a small, finite effective dimension. For example, commonly used kernel methods inject data into a functional Hilbert space (more precisely, a RKHS). In this case, the error of a method depends on the effective dimension of the distribution. The notion of effective dimension makes it possible to analyze problems of finite and infinite dimension simultaneously, and thus blurs the difference between parametric and non-parametric.

In this thesis, we will seek to understand the effects of high dimensionality on testing and estimation problems, with a view to potentially improving these methods. This thesis consists of four works:

- the paper "High-Dimensional Multi-Task Averaging and Application to Kernel Mean Embedding" in collaboration with Hannah Marienwald and Gilles Blanchard published at *AISTATS*, (2021);
- the chapter "Nonasymptotic One- and Two-Sample Tests in High dimension with Unknown Covariance Structure" in collaboration with Gilles Blanchard of *Foundations of Modern Statistics: Festschrift in Honor of Vladimir Spokoiny*, which is a collection of original research paper, (2023);
- the preprint "Estimation of multiple mean vectors in high dimension" in collaboration with Hannah Marienwald and Gilles Blanchard, (2024);
- some first results on an analysis of the self-attention mechanism in high dimension in collaboration with Gilles Blanchard.

Main problems of the thesis: The primary issue addressed in this thesis is the simultaneous estimation of mean vectors in high dimensions. We aim to estimate mean vectors, denoted as μ_k, corresponding to various distributions $\mathbb{P}_k$, all defined over a common Hilbert space $\mathcal{H}$. This problem combines classical statistical questions related to high-dimensional mean estimation, dating back to Stein (1956), with more recent concerns in multi-task learning (MTL) (Bonilla et al., 2007; Micchelli and Pontil, 2004), where the goal is to simultaneously perform multiple independent tasks that share similarities. Our specific task here involves estimating mean vectors, which can be viewed as a simplified version of more complex problems such as regression, classification, covariance estimation, or distribution estimation. In our case, where the tasks focus on estimating mean vectors, the problem is referred to as multi-task averaging (MTA) (Feldman et al., 2014). This problem becomes more concrete when the vectors in question are the kernel mean embeddings (KME) of distributions. KMEs, known for their numerous properties, are central to kernel methods, and our goal is to improve their estimation within the MTL framework, where multiple embeddings are estimated simultaneously.

Our approach begins by considering the case of isotropic Gaussian distributions, $\mathbb{P}_k = \mathcal{N}(\mu_k, \sigma^2 I_d)$, and homogeneous samples, $N_k = N$. In this scenario, the minimax estimation error for the mean is achieved by the empirical mean and is of order $\sigma^2 d/N$ (in quadratic norm). It is well known that in high dimensions, this estimation error is significantly larger than the testing error—i.e., the smallest distance between two means at which they can be distinguished—which is of order $\sigma^2\sqrt{d}/N$. We thus propose to test pairwise equality between the vectors μ_k to detect closed means and then estimate each one using a shrinkage estimator towards the selected empirical means. The intuition is that shrinkage will reduce the variance of the estimation at the cost of introducing a bias. In our case, we expect this bias to be smaller than the variance in high dimensions ($\sqrt{d} \leq d$). Consequently, the improvement offered by this estimation will be notable in high-dimensional settings, where estimation is particularly challenging. This phenomenon recalls the paradox of Stein (1956), in which empirical means are shrunk toward a reference point to improve their quadratic risk.

A significant portion of this thesis is devoted to justifying this intuition, constructing a method, and extending it to non-Gaussian, non-isotropic data, and non-homogeneous samples. These extensions are necessary for dealing with the case of KMEs, where data reside in an infinite-dimensional functional space. For such distributions, the critical parameter of dimension is replaced by a notion of effective dimension, which will be constructed from the covariances of the distributions. This concept plays a key role in our study of the separation rate for proximity tests of means in non-isotropic distributions, allowing us to generalize the phenomenon observed in the isotropic Gaussian case to the KME setting and, more broadly, to bounded distributions.

In the remainder of this introduction, we present some of the key concepts related to our problem that were briefly mentioned above. In Section I.1.1, we introduce the estimator from James and Stein (1961), one of the most famous shrinkage estimators in high dimensions, and connect its construction to various ideas that will be used in our methods. Then, in Section I.1.2, we present the notion of separation rate for proximity tests and examine the influence of dimensionality on this rate in the case of isotropic Gaussian distributions. In Section I.2, we introduce the KME of a distribution and some of its applications. This high-dimensional object, widely used in machine learning, greatly motivates our problem and leads us to consider notions of effective dimension. These notions, along with several examples, are presented in Section I.3. Finally, in Section I.4, we provide a more detailed presentation of the various contributions of this thesis.

I.1 Estimation and testing in high dimensions

Estimation and testing problems, central to this thesis, have already been studied in terms of high dimensionality. Its effect on these problems has been highlighted in particular in the classical framework of

isotropic Gaussian distributions. In this section, we first present the Stein's paradox (Stein, 1956) related to vector estimation and its link with more modern problems and methods. We then present results on test separation rate, for which Baraud (2002) was the first to take into account the influence of dimension. These works are in fact the starting point for the various questions posed in this thesis.

I.1.1 Estimating a vector: Stein's paradox

Stein's paradox is a typical example of the counter-intuitive phenomena of high dimensionality. Consider a sample $(X_i)_{1\leq i\leq N}$ of random vectors in $\mathbb{R}^d$ with a Gaussian distribution denoted $\mathcal{N}(\mu,\sigma^2 I_d)$ where the variance σ^2 is known and the mean vector $\mu\in\mathbb{R}^d$ is unknown. The aim is to estimate this vector μ while minimizing the squared risk for the Euclidean distance. Let $\mathcal{G}_d=\left\{\mathcal{N}(\mu,\sigma^2 I_d)^{\otimes N}:\mu\in\mathbb{R}^d\right\}$ the set of distributions of Gaussian N-samples with fixed isotropic covariance, then the minimax estimation error of the vector μ is:

$$\inf_{\widehat{\mu}}\sup_{\mathbb{P}\in\mathcal{G}_d}\mathbb{E}\Big[\|\widehat{\mu}-\mu\|^2\Big]=d\,\frac{\sigma^2}{N}\,. \tag{I.1}$$

We can see the influence of dimension on the estimation error: the error increases with the dimension of the space. The empirical mean $\bar{X}_N=\frac{1}{N}\sum_{k=1}^N X_k$ achieves exactly this error and is therefore a minimax estimator of μ on this set of distributions. However, Stein (1956) shows that the empirical mean is inadmissible, more precisely that there are strictly better estimators in the sense of quadratic risk. A better estimator is, for example, the James-Stein estimator (James and Stein, 1961), which contracts the empirical mean to a reference point, traditionally 0. We will consider here its alternative version with a positive part, defined by

$$\mu^{\mathbf{JS}+}=\left(1-\frac{\sigma^2}{N}\frac{d-2}{\|\bar{X}_N\|^2}\right)_+\bar{X}_N, \tag{I.2}$$

and which has a squared error for the estimation of μ strictly better than the empirical mean (Baranchik, 1964). For a dimension $d\geq 2$:

$$\mathbb{E}\Big[\|\mu^{\mathbf{JS}+}-\mu\|^2\Big]\leq d\,\frac{\sigma^2}{N}\min\left(\frac{\tau}{1+\tau}+\frac{4}{d},1\right)\quad\text{where}\quad\tau=\tau(\mu)=\frac{N\|\mu\|^2}{d\sigma^2}\,, \tag{I.3}$$

(see e.g. Lemma 3.8 of Tsybakov, 2008 for this bound). The improvement over the empirical mean is greater in higher dimensions, i.e. when $d\gg N$. Indeed, in this case, the variance of the empirical mean (Eq.(I.1)) becomes very high and the shrinkage towards 0 becomes more efficient. The variance is reduced by adding a bias symbolized by the τ factor. At a fixed $\|\mu\|^2$, the τ factor decreases with dimension and increases with sample size. Inversely, the improvement is weaker as N grows: we then leave the high-dimensional framework and the shrinkage loses its interest. In all cases, however, the James-Stein estimator remains strictly better.

The James-Stein estimator is minimax on the class of $\mathcal{G}_d$ distributions, but also on the subset of Gaussian distributions with means close to 0. Let $\tau>0$, and define

$$\mathcal{P}_d(\tau)=\left\{\mathcal{N}(\mu,\sigma^2 I_d)^{\otimes N}:\|\mu\|^2\leq\tau d\,\sigma_N^2\right\},$$

then Pinsker (1980) shows that the minimax risk on this class verifies

$$\lim_{d\to\infty}\inf_{\widehat{\mu}}\sup_{\mathbb{P}\in\mathcal{P}_d(\tau)}\frac{\|\widehat{\mu}-\mu\|^2}{d\sigma_N^2}=\frac{\tau}{1+\tau},\quad\text{where}\quad\sigma_N^2=\frac{\sigma^2}{N}. \tag{I.4}$$

The estimator $\mu^{\mathbf{JS}+}$ reaches the minimax bound asymptotically in the dimension and obviously without knowing τ. In this model, the possible estimation error is still $O(N^{-1})$, but the gain is in the dimension dependence. The James-Stein estimator adapts to many contexts, such as a non-isotropic covariance (Bock,

1975), notions of risk different from quadratic risk (Berger, 1976) and can be constructed with a different shrinkage and variance estimation (Baranchik, 1970; Lehmann and Casella, 2006). More recently, Muandet et al. (2014) adapt it to estimate Kernel Mean Embeddings of distributions (see Section I.2).

Below we present different interpretations of the James-Stein estimator and how classical methods or ideas lead to consider it. Our aim is to introduce the reader to different ideas that are applicable to the estimation of the vector μ, but which we will later use in broader frameworks.

Oracle interpretation: The James-Stein estimator can be seen as the result of estimating the weight of the best shrinkage estimator towards 0. Consider the estimator $\widehat{\mu}_\omega = \omega \bar{X}_N$ where $\omega \in [0,1]$ and look for the estimator of this form minimizing the quadratic risk:

$$\min_{\omega\in[0,1]} \mathbb{E}\left[\|\widehat{\mu}_\omega - \mu\|^2\right] = \min_{\omega\in[0,1]} \left[\omega^2 d\sigma_N^2 + (1-\omega)^2\|\mu\|^2\right] = d\sigma_N^2 \frac{\tau(\mu)}{1+\tau(\mu)}$$

The optimal weight is $\omega^* = 1 - \frac{d\sigma_N^2}{\|\mu\|^2 + d\sigma_N^2}$. By knowing the norm of μ, we can improve its estimation, which seems quite natural. As this is unknown, the James-Stein estimator directly estimates $\|\mu\|^2 + d\sigma_N^2$ by $\|\bar{X}_N\|^2$, which leads us to consider (I.3) after injecting this estimator into the oracle weight formula. The James-Stein estimator will keep its performance close to that of the oracle estimator, thanks to the fact that estimating a distance (an one-dimensional quantity) in high dimension is much easier than estimating a vector. The induced error will be negligible relative to the gain.

Test interpretation: The choice of the shrinkage of the James-Stein estimator can be related to the testing problem:

$$(H_0) : \mu = 0\,, \qquad (H_1) : \mu \neq 0\,.$$

The statistic $P = d\sigma_N^2/\|\bar{X}_N\|^2$ is super-uniform under (H_0) (trivially by Markov's inequality as $\mathbb{E}\left[P^{-1}\right] = 1$) and can be used to test these hypotheses. As under (H_0), the vector μ should be estimated by 0, the James-Stein estimator uses the test statistic P to quantify the contraction towards 0:

$$\widehat{\mu}^{\mathbf{JS}+} = \left(1 - \frac{d-2}{d} P\right)_+ \widehat{\mu}^{\mathtt{NE}}.$$

The test is used to construct the estimator. This view can be found in Casella (1985), which considers a shrinkage of each coordinate towards the mean of the coordinates and relates this James-Stein type estimator to the problem of testing whether the coordinates of μ are all equal.

Regularization interpretation: Gruber (1998) links James-Stein type estimators and ridge type estimators, which are other shrinkage estimators. Thus, the James-Stein estimator $\widehat{\mu}^{\mathbf{JS}} = \left(1 - \sigma_N^2 \frac{d-2}{\|\bar{X}_N\|^2}\right)\bar{X}_N$, (Eq.(II.3) without the positive part), is the solution to the ridge regression problem

$$\widehat{\mu}^{\mathbf{JS}} = \operatorname*{Arg\,Min}_{\nu\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^N \|X_i - \nu\|^2 + \lambda\|\nu\|^2$$

for $\lambda = \frac{(d-2)\sigma_N^2}{\|\bar{X}_N\|^2 - (d-2)\sigma_N^2}$. To avoid choosing λ, we can also consider regularization by the norm. This produces a James-Stein-type estimator:

$$\left(1 - \frac{\widehat{\sigma}}{\sqrt{N}}\frac{\sqrt{d}}{\|\bar{X}_N\|}\right)_+ \bar{X}_N = \operatorname*{Arg\,Min}_{\nu\in\mathbb{R}^d} \sqrt{\frac{1}{N}\sum_{i=1}^N \|X_i - \nu\|^2} + \frac{1}{\sqrt{N}}\|\nu\|,$$

where $\widehat{\sigma}^2 = \frac{1}{d(N-1)}\sum_{i=1}^N \left\|X_i - \bar{X}_N\right\|^2$ is an unbiased estimator of σ^2. This estimator beats the empirical mean for means close to 0, $(\tau(\mu) < 1)$ but is not minimax. However, this example illustrates that with

a simple penalization by the norm, without any additional parameter, we find a shrinkage close to that of the James-Stein estimator, as well as the presence of the positive part. Note also that this regularization naturally estimates the variance by the empirical variance.

Bayesian interpretation: The works of Efron and Morris (1972, 1973, 1976) interpret the James-Stein estimator as an empirical Bayes problem. For example, assuming that each of the coordinates of μ is drawn independently according to the same normal distribution $\mu_i \sim \mathbb{Q} = \mathcal{N}(0, \tau\sigma^2)$, the Bayes estimator of the vector μ is then a shrinkage estimator $\omega \bar{X}_N$ with $\omega = \frac{1}{1+\tau}$. Marginally $X_i \sim \mathcal{N}(0, (1+\tau)\sigma^2 I_d)$, so we can estimate $1+\tau$ by $\|\bar{X}_N\|^2/(d\sigma_N^2)$ which again leads to the James-Stein estimator. More recently Brown and Greenshtein (2009), have reconsidered this approach for an arbitrary $\mathbb{Q}$ distribution. A more detailed discussion is given in Section V.7.1.

Multi-task interpretation: The estimation of each of the coordinates of the vector μ can be considered as a multi-task problem (Baxter, 1997; Caruana, 1997). Multi-task learning seeks to solve different problems simultaneously (regression, estimation, ...) for datasets with different distributions. To this end, the multiple task approach makes use of similarities between distributions (common structure, identical noise, ...). Here, our tasks would be to estimate each of the coordinates by minimizing the compound risk, i.e. the average of the errors of each estimator. This is equivalent to minimizing the squared error of the vector of estimators. The James-Stein estimator finally uses the fact that the data from all tasks have the same noise σ^2 to estimate it efficiently and build a shrinkage estimator for each coordinate. The problem of estimating different averages is known as *multi-task averaging*. The link between this problem and the James-Stein estimator is considered, for example, by Feldman et al. (2014) or Duan and Wang (2023).

These last two interpretations see the James-Stein estimator as the joint estimation of real quantities (the coordinates). A natural question to ask is whether it can be adapted to simultaneously estimate different vectors from noisy samples of each. This problem reduces to the James-Stein framework if we assume that the noises in each sample are isotropic Gaussian. However, if the noises are different, unknown, non-isotropic or even non-Gaussian, it is not obvious how to construct a James-Stein type estimator, and although partial answers can be found in the literature, the problem has not been considered as a whole. Among the questions raised is how the different degrees of freedom of the problem interact, i.e. the size of samples, their dimension and the number of vectors to be estimated. We will consider this problem in Sections III and Section V.

I.1.2 Vector separation distance

The influence of dimension on test performance was highlighted by Baraud (2002) in his non-asymptotic analysis of test separation rate in the Gaussian framework. The separation distance or separation rate of a test defined by Ingster (1982) in the asymptotic framework and adapted to the non-asymptotic framework by Baraud (2002) permits a minimax analysis of testing problems. In a general framework, for a distance γ between distributions and a set of hypotheses $\mathcal{H}_0$ and alternatives $\mathcal{H}_1$, we define the separation distance of hypothesis sets for $\alpha \in (0,1)$ as :

$$\delta_\alpha^* = \inf\left\{\delta \geq 0 \middle| \exists \text{ test } T \ : \sup_{\mathbb{P}\in\mathcal{H}_0} \mathbb{P}(T=1) + \sup_{\mathbb{P}\in\mathcal{H}_1 : \gamma(\mathbb{P},\mathcal{H}_0)\geq\delta} \mathbb{P}(T=0) \leq \alpha\right\}. \tag{I.5}$$

It is important to note that the separation distance depends strongly on the chosen distance γ between the distributions. Intuitively, the separation distance is the minimum distance between hypotheses and alternatives for which a test exists whose sum of type I and II errors is controlled by α. This notion is equivalent to the sample complexity of the problem. For a test, sample complexity is the minimum number

of data items for which the sum of type I and type II errors is controlled by α for alternatives at a fixed δ distance. These notions can be deduced from each other by elementary operations.

Remark I.1. Definition (I.5) is different from the original ones, Baraud (2002) and Ingster (1982) consider rather the minimal distance of alternatives to hypotheses for which there is a test of type I error exactly α and type II error controlled. However, the two definitions coexist in the literature.

Let us consider the problem of signal detection with Gaussian noise. Let $(X_i)_{1\leq i\leq n}$ be a sample of Gaussian vectors with distribution $\mathcal{N}(\mu, \sigma^2 I_d)$ and the test problem:

$$(H_0(\eta)) : \|\mu\| \leq \eta, \text{ against } (H_1(\eta)) : \|\mu\| > \eta, \tag{I.6}$$

where $\eta \geq 0$. In the classical case where $\eta = 0$ we refer to Baraud (2002) and in the case $\eta > 0$, known as relevant or precise hypothesis testing, to Blanchard et al. (2018). These works are distinguished by their non-asymptotic analyses of the role of the dimension. Indeed, in the non-parametric framework, the analyses focus instead on the dependence of rate on sample size and the influence of regularity. For a detailed analysis, readers may refer to Ingster and Suslina (1998).

In this model of Gaussian distributions with fixed variances, we can choose as distance γ between the distributions the Euclidean distance between the mean vectors: $\gamma\Big(\mathcal{N}(\mu, \sigma^2 Id), \mathcal{N}(\nu, \sigma^2 I_d)\Big) = \|\mu - \nu\|$. With this distance γ, the optimal separation distance for the test problem (I.6) is the minimum distance δ^* for which a test is able to differentiate between the distributions with means in the ball of radius η and those outside the ball of radius $\eta + \delta^*$. For $\eta = 0$, Baraud (2002) analyzes the dimension dependence of this separation rate and gives the following rate:

$$\delta^*_\alpha(\eta = 0) = \Theta_\alpha\Big(d^{1/4}\sigma_N\Big), \tag{I.7}$$

where $\sigma_N^2 = \sigma^2/N$ and Θ_α indicates lower and upper bounds depending only on α. This non-asymptotic analysis highlights the relative ease of the test compared with estimation. The minimum error of detection is $\sigma_N d^{1/4}$ for the test versus $\sigma_N\sqrt{d}$ (Eq.(I.1)) for the estimation. In other words, a test is able to ensure that the vector μ is $\sigma_N d^{1/4}$ close to a reference point (here 0), whereas an estimator of μ is only guaranteed to be $\sigma_N\sqrt{d}$ close to the true vector μ.

For non-zero η, Blanchard et al. (2018) have demonstrated the existence of two regimes. When η is small, the test error is still (II.7), whereas for η large the test becomes easier and the error loses its dimension dependence. More precisely:

$$\delta^*_\alpha(\eta) = \Theta_\alpha\left(\sigma_N \max\left(1, \min\left(d^{1/4}, \sqrt{d}\frac{\sigma_N}{\eta}\right)\right)\right). \tag{I.8}$$

Intuitively, as η grows, one direction becomes predominant and the problem becomes one-dimensional. More generally, testing whether $\mu \in \mathcal{C}$ where $\mathcal{C}$ is a convex set is always easier than estimating μ. Denoting $\delta^*(\mathcal{C})$ the test separation distance for the convex set $\mathcal{C}$, we have

$$\delta^*_\alpha(\mathcal{C}) = O_\alpha\left(\inf_{\widehat{\mu}} \sup_{\mathbb{P}\in\mathcal{H}_0\cup\mathcal{H}_1} \mathbb{E}[\|\widehat{\mu} - \mu\|]\right) = O_\alpha\Big(\sigma_N\sqrt{d}\Big). \tag{I.9}$$

The first inequality is true in all generality (using as test statistic the distance of an estimator $\widehat{\mu}$ to the convex $\mathcal{C}$) and the second in our framework of isotropic Gaussian distributions. This limiting case is reached for $\mathcal{C}$ an orthant ($\mathcal{C} = [-\infty, 0]^d$) and in this case the test error is the same as the estimation error (Theorem 3.6. of Blanchard et al., 2018 or Juditsky and Nemirovski, 2002 in a nonparametric framework).

The appearance of dimension in the separation rate or estimation error is closely linked to the finite-dimensional isotropic Gaussian model. However, this paradigm does not apply to some modern tools. For example, this is the case for the Kernel Mean Embedding (KME), a vector of a functional space that can be used to characterize distributions. In the next section, we present what a KME is, its interest in machine learning and its links with high-dimensional testing and estimation problems.

I.2 A high-dimensional tool: the Kernel Mean Embedding

The *Kernel Mean Embedding* (KME) is a machine learning tool introduced by Smola et al. (2007) and is intrinsically linked to kernel methods and Reproducing Kernel Hilbert Spaces (RKHS). The principle behind kernel methods (Aizerman, 1964) is to inject the data under study into a higher dimensional space and then apply a classical algorithm (typically a regularized linear method). The strength of these methods is that the change of space is only done by replacing the Euclidean scalar product by the scalar product of the Hilbert space defined by a kernel κ. A method that uses only scalar products on the data is thus very easy to adapt, as is the case with many methods such as support vector machine (Boser et al., 1992; Cortes and Vapnik, 1995), ridge regression (Cristianini and Shawe-Taylor, 2000; Hoerl and Kennard, 1970; Saunders et al., 1998) or principal component analysis (Hotelling, 1933; Pearson, 1901; Schölkopf et al., 1998).

To a distribution $\mathbb{Q}$ on the initial space, a distribution $\mathbb{P}$ on the RKHS is associated which is the pushforward measure of $\mathbb{Q}$ by the injection map into the RKHS. The KME of the distribution can simply be defined as the expectation of the $\mathbb{P}$ distribution and is, under weak assumptions, a vector of the RKHS. The KME of a distribution is therefore directly related to the RKHS and therefore to the choice of kernel κ. By choosing a right kernel, the KME can fully characterize the distribution and, among other things, induce an easy-to-calculate distance between distributions called *Maximum Mean Discrepancy* (MMD) (Borgwardt et al., 2006). This distance is simply the distance between the KMEs of the distributions in the RKHS. This property opens the way to numerous applications such as the two-sample test (Gretton et al., 2012), goodness-of-fit test (Chwialkowski et al., 2016), supervised (Muandet et al., 2012; Szabó et al., 2016) or unsupervised distributional learning (Jegelka et al., 2009). The MMD distance is also used to build generative models (Dziugaite et al., 2015; Li et al., 2017; Li et al., 2015).

In this section we present the construction of the KME of a distribution, followed by its estimation and its use in two-sample tests. A more complete overview can be found in Muandet et al. (2017).

I.2.1 Constructing the KME and the MMD distance

Kernel methods are built from kernel functions. Let us denote $\mathcal{X}$ the data space, a function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel if it is symmetrical ($\kappa(x, y) = \kappa(y, x)$) and if for any integer n, weights $a_1, \dots, a_n \in \mathbb{R}$ and points $x_1, \dots, x_n \in \mathcal{X}$:

$$\sum_{i,j=1}^{n} a_i a_j \kappa(x_i, x_j) \geq 0 \,.$$

In particular, this property implies that the diagonal of κ is positive ($\kappa(x, x) \geq 0$ for any point x). Kernel functions are stable by sum, dilatation, multiplication and passage to the limit, making them easy to construct. The following are the most classic kernels:

- $\kappa(x, y) = \mathbf{1}_{x=y}$ defines the trivial kernel;
- if $\mathcal{X} \subset \mathbb{R}^d$, then $\exp\left(-\frac{\|x-y\|_2^2}{h^2}\right)$ and $\exp\left(-\frac{\|x-y\|_2}{h}\right)$ where $h > 0$ define the Gaussian and Laplace kernels respectively;
- if $\phi : \mathcal{X} \to \mathcal{H}$ is an injection of $\mathcal{X}$ into a Hilbert space $\mathcal{H}$, the scalar product $\langle \phi(\cdot), \phi(\cdot) \rangle_{\mathcal{H}}$ defines a positive definite kernel. If $\mathcal{X}$ is a Hilbert space, then its scalar product is one.

Kernel functions are built for a wide range of data, from text (Joulin et al., 2017) to sequences, particularly in bioinformatics (Gusfield, 1997), which can be extended for trees and graphs (see Gärtner, 2003 or Shawe-Taylor and Cristianini, 2004 for an overview), for image analysis (Zhang et al., 2007) and also in topological data analysis for persistence diagrams (Carriere et al., 2017).

The purpose of kernels is to define a scalar product in a larger space: from a positive definite kernel, it is indeed possible to construct a Hilbert space for which this kernel defines a scalar product.

Proposition I.2. *Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a positive definite kernel. Consider the vector space*

$$\mathcal{H}_0 = \mathit{Vect}(\kappa(x, \cdot) : x \in \mathcal{X}) \subset \mathbb{R}^{\mathcal{X}},$$

with scalar product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ defined by

$$\langle \kappa(x, \cdot), \kappa(y, \cdot) \rangle_{\mathcal{H}} := \kappa(x, y) \quad \textit{for} \quad x, y \in \mathcal{X} \tag{I.10}$$

and extended by linearity. Then $\mathcal{H}$ the completed of $\mathcal{H}_0$ is a Hilbert space for the scalar product (I.10), extended by taking the limit. In particular, κ is a reproducing kernel for $\mathcal{H}$:

$$\langle h, \kappa(x, \cdot) \rangle_{\mathcal{H}} = h(x), \quad \forall\, h \in \mathcal{H},\ x \in \mathcal{X}, \quad \textit{(reproducing property).}$$

The kernel induces an injection ϕ between the data space $\mathcal{X}$ and the RKHS $\mathcal{H}$ defined for $x \in \mathcal{X}$ by

$$\phi(x) = \kappa(x, \cdot)\,, \quad \text{and for } y \in \mathcal{X} \quad \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = \kappa(x, y)\,.$$

As announced, the kernel κ enables the direct calculation of scalar products between the embeddings of two data items in the RKHS. Moreover, $\mathcal{H}$ is the only RKHS for which κ is a reproducing kernel (Moore-Aronszajn Theorem, Aronszajn, 1950). However, for a given kernel, it is possible to construct several Hilbert spaces for which this kernel does define a scalar product, but is not reproducing. For example, the kernel given by $\kappa(x, y) = xy$ for $x, y \in \mathbb{R}$ defines a scalar product on any line of $\mathbb{R}^d$ for an arbitrary d dimension ($\mathcal{H} = \{x\nu, x \in \mathbb{R}\}$ for some $\nu \in \mathbb{R}^d$).

The kernel associates to a point of $\mathcal{X}$ an image in the RKHS by the injection ϕ which can be generalized to the distributions on $\mathcal{X}$ by the KME.

Definition I.3 (KME)**.** Let $\mathbb{Q}$ be a distribution on $\mathcal{X}$ and κ a positive definite kernel, the *Kernel Mean Embedding* (KME) of the distribution $\mathbb{Q}$ in the RKHS $\mathcal{H}$ associated with κ is

$$\mu_{\mathbb{Q}} = \mathbb{E}_{X \sim \mathbb{Q}}[\kappa(X, \cdot)]\,. \tag{I.11}$$

If $\mathbb{E}_{X \sim \mathbb{Q}}\left[\sqrt{\kappa(X, X)}\right] < \infty$ then $\mathbb{Q}$ is an integrable measure in the Bochner sense in $\mathcal{H}$ and $\mu_{\mathbb{Q}} \in \mathcal{H}$ is well defined.

The KME existence condition is easily verified by considering a bounded kernel (e.g. the trivial, Gaussian, Laplace kernels). Thanks to the reproducing property of the kernel, for any function h of $\mathcal{H}$, we have $\mathbb{E}_{X \sim \mathbb{Q}}[h(X)] = \langle h, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}$. This property will be particularly useful for estimating distances between KMEs. An important question was for which kernels this distance between KMEs induces a distance between distributions, i.e. for which kernels the function $\mathbb{Q} \mapsto \mu_{\mathbb{Q}}$ is injective. Kernels verifying this property are called *characteristics*. Intuitively, the function class of the RKHS $\mathcal{H}$ must be rich enough for the KME to characterize the distribution. For example, for a compact $\mathcal{X}$ space, the kernel is characteristic if the RKHS $\mathcal{H}$ is dense in continuous bounded functions (Steinwart, 2001). Here are a few more examples of characteristic kernels for different spaces.

- The trivial kernel $\mathbf{1}_{x=y}$ is characteristic when $\mathcal{X}$ is finite (Borgwardt et al., 2006).
- The exponential kernel $\kappa(x, y) = \exp(\langle x, y \rangle)$ is characteristic when $\mathcal{X}$ is a compact set of $\mathbb{R}^d$. The KME is then the moment-generating function $\mu_{\mathbb{Q}}(x) = \mathbb{E}_{X \sim \mathbb{Q}}[\exp(\langle x, X \rangle)]\,.$
- The Gaussian and Laplace kernels are characteristic on $\mathbb{R}^d$ (Fukumizu et al., 2007).
- More generally for a translation-invariant kernel on $\mathbb{R}^d$ ($\kappa(x, y) = K(x - y)$), the KME relates to the characteristic function of the distribution. It is then characteristic if its Fourier transform support is equal to $\mathbb{R}^d$ (Sriperumbudur et al., 2011, 2008, 2010).

For a characteristic kernel, the distance between the KMEs of two distributions therefore induces a distance between the distributions. This distance is known as the *Maximum Mean Discrepancy* (MMD) (Gretton et al., 2012).

Definition I.4 (MMD)**.** Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a characteristic kernel and $\mathcal{H}$ its associated RKHS, the MMD distance between two distributions $\mathbb{P}$ and $\mathbb{Q}$ of $\mathcal{X}$ is defined by:

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} .$$

where $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ are the respective KMEs of $\mathbb{P}$ and $\mathbb{Q}$.

This distance can be seen as an integrable probability metric (Müller, 1997) on the RKHS $\mathcal{H}$. Indeed:

$$\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} = \sup_{h \in \mathcal{H}: \|h\|_{\mathcal{H}} \leq 1} \langle h, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle = \sup_{h \in \mathcal{H}: \|h\|_{\mathcal{H}} \leq 1} \left(\int_{\mathcal{X}} h d\mathbb{P} - \int_{\mathcal{X}} h d\mathbb{Q} \right).$$

So if $\mathcal{H}$ contains the bounded functions, the MMD distance dominates the distance in total variation, and for $\mathcal{X} = \mathbb{R}$, if it contains the indicator functions $\left\{ \mathbf{1}_{(-\infty, t)} \right\}_{t \in \mathbb{R}}$, it dominates the Kolmogorov distance. The MMD distance is equivalent to the energy distance (Sejdinovic et al., 2013) and relates to the optimal transport distances: it is the limit of the Sinkhorn divergence (Genevay et al., 2018). Compared with these distances, a strength of the MMD distance and the KME in general is the relative ease of their estimations.

I.2.2 Estimation

Let $\mathbb{Q}$ and $\mathbb{P}$ be two distributions on $\mathcal{X}$ known only via two samples $\{X_i\}_{1 \leq i \leq N}$ and $\{Y_j\}_{1 \leq j \leq M}$ of the distributions $\mathbb{Q}$ and $\mathbb{P}$ respectively. We present here classical estimators of the KME $\mu_{\mathbb{Q}}$ and of the MMD distance between $\mathbb{Q}$ and $\mathbb{P}$.

The KME of the $\mathbb{Q}$ distribution can be seen as the expectation of the random vector $Z_i = \kappa(X_i, \cdot)$ (where $X_i \sim \mathbb{Q}$) in the Hilbert space $\mathcal{H}$. So the KME $\mu_{\mathbb{Q}} = \mathbb{E}[Z_1]$ can be estimated by the classical empirical mean:

$$\widehat{\mu}_{\mathbb{Q}}(\cdot) = \frac{1}{N} \sum_{i=1}^{N} \kappa(X_i, \cdot) = \frac{1}{N} \sum_{i=1}^{N} Z_i .$$

When the kernel κ is bounded, then the random vector Z is also bounded in $\mathcal{H}$ and it is possible to use the concentration tools associated with bounded random variables to control deviations. Thus, using the inequality of McDiarmid et al. (1989), for all $u \geq 0$, with probability $1 - e^{-u}$:

$$\|\widehat{\mu}_{\mathbb{Q}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \leq \frac{L}{\sqrt{N}} \left(1 + \sqrt{2u} \right), \tag{I.12}$$

where $L^2 \geq \sup_{x \in \mathcal{X}} \kappa(x, x)$ is a bound on the diagonal of the kernel. Such an assumption is verified by the usual kernels (trivial, Gaussian, Laplace). Note that a reproducing kernel bounded on the diagonal is then bounded everywhere, since for any $x, y \in \mathcal{X}$

$$|\kappa(x, y)| = |\langle \kappa(x, \cdot), \kappa(y, \cdot) \rangle_{\mathcal{H}}| \leq \|\kappa(x, \cdot)\|_{\mathcal{H}} \|\kappa(y, \cdot)\|_{\mathcal{H}} = \sqrt{\kappa(x, x) \kappa(y, y)}, .$$

Estimating the KMEs of $\mathbb{Q}$ and $\mathbb{P}$ distributions can provide an estimator of their MMD distance by calculating the distance between them in the RKHS directly (although there are better estimators, see below). However, the use of KMEs is not summed up to define the MMD distance. Its estimation is needed, for example, for *distribution regression* problems where we seek to make a prediction from a sample (see for example Oliva et al., 2013 or Szabó et al., 2016). In causal inference, the KME of conditional laws is used as a proxy before regression (Mastouri et al., 2021; Singh et al., 2019). In these cases, full estimation of the KME is required.

The MMD distance between two distributions $\mathbb{Q}$ and $\mathbb{P}$ can be estimated without estimating their respective KMEs. Its most classic unbiased estimator (for the squared distance) is constructed using U-statistics:

$$\widehat{\mathrm{MMD}}^2(\mathbb{Q},\mathbb{P}) := \frac{1}{N(N-1)}\sum_{i\neq j=1}^{N}\kappa(X_i,X_j) - \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\kappa(X_i,Y_j) + \frac{1}{M(M-1)}\sum_{i\neq j=1}^{N}\kappa(Y_i,Y_j)\,. \tag{I.13}$$

This estimator can be considered naturally after noticing that $\|\mu_{\mathbb{Q}}\|^2_{\mathcal{H}} = \mathbb{E}[\langle Z, Z'\rangle_{\mathcal{H}}] = \mathbb{E}[\kappa(X,X')]$ where X, X' are independent and from $\mathbb{Q}$ distribution. Each of the terms of (I.13) estimates with no bias each of the terms of the development of the distance $\|\mu_{\mathbb{Q}} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}$. Deviations from this distance estimator can be controlled using concentration inequalities on U-statistics from Hoeffding (1963) (see also Gretton et al., 2012 in the KME framework). For all $u \geq 0$, with probability at least $1 - e^{-u}$:

$$\left|\widehat{\mathrm{MMD}}^2(\mathbb{Q},\mathbb{P}) - \|\mu_{\mathbb{Q}} - \mu_{\mathbb{P}}\|^2\right| \leq \frac{L^2}{\sqrt{\min(N,M)}}\sqrt{8u}\,. \tag{I.14}$$

These estimators of KME and MMD distance are optimal in sample size (Tolstikhin et al., 2017; Tolstikhin et al., 2016). In particular, the rate in the sample size does not depend on the dimension. However, in both cases, the rates given by the concentration bounds (I.12) and (I.14) do not take into account the covariance structure of the distribution or the dimension of the space. The terms related to these parameters are actually bounded by the bound L on the kernel. As the deviations are both in $O(L/\sqrt{N})$, the variance terms are actually "hidden" in the bound L on the kernel. To take their effects into account, we will need more precise concentration inequalities, of Bernstein type for example. More generally, the analysis of high-dimensional vector or distance estimation is used to be done under the assumption that the distributions are sub-Gaussian (e.g. Hsu et al., 2012; Koltchinskii and Lounici, 2017). However this framework, while including bounded distributions, does not capture the influence of these parameters on deviations by a direct application of existing results (see discussion in Section IV.2.4). To capture them, throughout the thesis, with the aim of building procedures adapted to KMEs, we will consider the framework of bounded data in a Hilbert space. Phenomena in estimation and testing comparable to those presented previously in Section II.1 will be observed.

I.2.3 Two-sample tests

The two-sample test consists in testing the equality of two distributions for only observed threw samples of each. Formally, for $\mathbb{P}$ and $\mathbb{Q}$, two distributions on the space $\mathcal{X}$, we seek to test

$$(H_0) : \mathbb{P} = \mathbb{Q}\,, \qquad \text{against} \qquad (H_1) : \mathbb{P} \neq \mathbb{Q}\,. \tag{I.15}$$

from two samples of each distribution. In one dimension, the historical tests for this problem are the Chi-squared test in the discrete framework (Pearson, 1900) and the Kolmogorov-Smirnov test in the continuous framework (Kolmogorov, 1933). These tests are based respectively on empirical estimators of the Chi-squared divergence and the Kolmogorov distance between $\mathbb{P}$ and $\mathbb{Q}$. The Kolmogorov-Smirnov test generalizes to high dimension (Bickel, 1969; Friedman and Rafsky, 1979), but has the disadvantage of a high algorithmic cost. Instead of considering these discrepancies, the kernel test proposed by Gretton et al. (2012) chooses to compare distributions using the MMD distance between $\mathbb{P}$ and $\mathbb{Q}$. In this framework, the test (I.15) can be rewritten as a vector equality test in the RKHS:

$$(H_0) : \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} = 0\,, \qquad \text{against} \qquad (H_1) : \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \neq 0\,. \tag{I.16}$$

Thanks to its flexibility (adaptable to the context by the choice of kernel) and the simplicity of MMD distance estimation (see Section I.2.2), this test has been widely broadcast. The question of its optimality

in the sense of the separation distance (I.5) is still being studied today. The original test, based on the U-statistic (I.13), is optimal for distributions with Hölder densities on $\mathbb{R}^d$ and for the distance L^2 between densities. However, the kernel chosen to construct such a minimax test must depend on the regularity parameter of these densities and then, in all generality, the test is sub-optimal (Balasubramanian et al., 2021). However, Schrab et al. (2023) construct a minimax and adaptive version of this test on $\mathbb{R}^d$ by aggregating a family of tests based on several kernels. To obtain a minimax test on spaces different of $\mathbb{R}^d$, Hagrass et al. (2022) use a regularization of the MMD distance by the covariance with a kernel adapted to the data. The separation distance of this test is then measured in terms of the Hellinger distance.

In the analyses cited, the test separation distance is evaluated using distances between distributions (L^2 distance between densities, Hellinger distance, Chi-squared divergence, ...). To relate these distances to the MMD distance, assumptions of regularity on the densities are required, and then the optimal rate in the sample size depends on the dimension. For example, Li and Yuan (2019) show that the separation distance for the L^2 norm between densities is $\Theta(N^{-4s/(4s+d)})$ where d is the dimension of the space, s the Sobolev regularity of the densities and N the sample sizes. The dependence on dimension is similar to that of nonparametric estimation rate. However, this dependence disappears when the test separation distance is considered directly in terms of the MMD distance between the distributions. In this case, the test separation distance is the separation distance of the vectors in the RKHS $\mathcal{H}$ and the same rate $\Theta(N^{-1/2})$ is found. Drawing a parallel with the Gaussian case, we might ask what influence the dimension of the space has on this separation distance, and whether it is possible to recover the form (I.8). The role of dimension for the separation distance, but also for the test error, will actually be played by a notion of effective dimension that we define in the next section.

I.3 Effective dimension

As already mentioned, analyzing and learning information from high-dimensional data is often possible because the data actually has a simpler structure. For example, the data may lie in a lower-dimensional subspace, in a manifold or can be clusters into a small number of balls. So the dimension of the ambient space of the distribution is not necessarily a critical quantity for quantifying the difficulty of a task. Instead, this role is played by notions of effective or intrinsic dimension of a distribution. These effective dimensions of a distribution $\mathbb{P}$ will depend on the considered problem and those we will consider will be constructed from its covariance operator (Baker, 1973). In what follows, we shall consider $\mathbb{P}$ to be a distribution on a Hilbert space $\mathcal{H}$. Examples will be given for $\mathcal{H} = \mathbb{R}^d$ and $\mathcal{H}$ a RKHS.

Definition I.5 (Covariance operator)**.** Let $\mathbb{P}$ be a distribution on a Hilbert space $\mathcal{H}$ such that $\mathbb{E}\left[\|X\|_{\mathcal{H}}^2\right] < \infty$, then its covariance operator is defined by

$$\Sigma(\mathbb{P}) : \begin{cases} \mathcal{H} & \to & \mathcal{H}\,, \\ y & \mapsto & \mathbb{E}\big[\langle y, X\rangle_{\mathcal{H}} X\big] - \big\langle y, \mathbb{E}\big[X\big]\big\rangle_{\mathcal{H}} \mathbb{E}\big[X\big]\,. \end{cases}$$

where X is a random variable of distribution $\mathbb{P}$.

On $\mathbb{R}^d$, for the canonical scalar product, the covariance operator is just the covariance matrix of the distribution: if $\mu = \mathbb{E}[X]$, then $\Sigma(\mathbb{P}) = \mathbb{E}\big[(X-\mu)(X-\mu)^T\big]$.

An infinite dimension distribution is for example the injection of a distribution into an RKHS $\mathcal{H}$. We can then consider the covariance operator of the push-forward distribution. Let $X \sim \mathbb{Q}$ be a random variable on $\mathcal{X}$ and $\mathbb{P}$ be the distribution of $\kappa(X, \cdot)$ on $\mathcal{H}$. The covariance operator is well defined when $\mathbb{E}[\kappa(X,X)]$ is finite (e.g. for bounded κ) and, in this case, for all $h \in \mathcal{H}$:

$$\Sigma(\mathbb{P})h = \mathbb{E}\big[\langle h, \kappa(X,\cdot)\rangle_{\mathcal{H}} \kappa(X,\cdot)\big] - \langle h, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}} \mu_{\mathbb{Q}}(\cdot)\,, \tag{I.17}$$

where $\mu_{\mathbb{Q}}$ is the KME (Definition I.3) of the distribution $\mathbb{Q}$. In particular, the covariance operator verifies for all $h, h' \in \mathcal{H}$:

$$\langle h', \Sigma(\mathbb{P})h \rangle_{\mathcal{H}} = \mathbb{E}[h(X)h'(X)] - \mathbb{E}[h(X)]\mathbb{E}[h(X')] = \mathrm{Cov}[h(X), h'(X)]$$

This identity is sometimes used as a definition of the operator. The operator $\Sigma(\mathbb{P})$ is bounded (hence continuous), positive and self-adjoint. Please note that we are considering here the centered version of the covariance operator. In fact, it is the moments of this operator that are used in our methods and that we sometimes have to estimate (see Section I.4.5). In kernel methods, its non-centered version is also used, in particular for tests of independence and conditional independence (Doran et al., 2014; Gretton et al., 2005). However, the centered version has recently been used for kernel principal component analysis (Sriperumbudur and Sterge, 2022) and two-sample tests (Hagrass et al., 2022; Li and Yuan, 2019). We will not, however, be interested in its estimation, but rather use it as a theoretical tool to define a notion of effective dimension. We will, however, need to estimate some of these moments (see Section I.4.5), for which we will provide estimators.

The notions of effective dimension that we will consider are expressed in terms of Schatten norms of the covariance operator of the distribution. The following definition introduces the three effective dimensions that will be considered in this manuscript.

Definition I.6 (Effective dimension)**.** Let $\mathbb{P}$ be a distribution on a Hilbert space $\mathcal{H}$ and $\Sigma := \Sigma(\mathbb{P})$ its covariance operator. We will call effective dimensions the following quantities:

$$d^{\mathrm{e}}(\mathbb{P}) = \frac{\operatorname{Tr}\Sigma}{\|\Sigma\|_{op}}, \quad d^{*}(\mathbb{P}) = \frac{\operatorname{Tr}\Sigma^2}{\|\Sigma\|_{op}^2}, \quad d^{\bullet}(\mathbb{P}) = \frac{(\operatorname{Tr}\Sigma)^2}{\operatorname{Tr}\Sigma^2}, \tag{I.18}$$

where Tr denotes the trace and $\|\cdot\|_{op}$ the operator norm.

Remark I.7. We sometimes express effective dimensions in terms of Schatten norms of the covariance: for $p \in \mathbb{N}^*$ the p-Schatten norm is defined by $\|\Sigma\|_p^p := \operatorname{Tr}(\Sigma^p)$, if this quantity exists. For $p \geq p'$, the Schatten norms satisfy $\|\Sigma\|_p \leq \|\Sigma\|_{p'}$.

According to the previous remark, the effective dimensions are well defined if the covariance operator Σ is of trace class, i.e. of finite trace. Indeed, for $(e_k)_{k\in\mathbb{N}}$ an orthonormal basis of $\mathcal{H}$, we have :

$$\operatorname{Tr}\Sigma = \sum_{k=0}^{+\infty} \langle e_k, \Sigma e_k \rangle_{\mathcal{H}} = \sum_{k=0}^{+\infty} \Big(\mathbb{E}\Big[\langle e_k, X\rangle_{\mathcal{H}}^2\Big] - \langle e_k, \mathbb{E}[X]\rangle_{\mathcal{H}}^2 \Big) = \mathbb{E}\Big[\|X\|_{\mathcal{H}}^2\Big] - \big\|\mathbb{E}[X]\big\|_{\mathcal{H}}^2 < \infty\,.$$

The last quantity is bounded by assumption and using Jensen's inequality.

In the random matrix literature, d^{e} is sometimes called the intrinsic dimension (Hsu et al., 2012; Tropp et al., 2015) or effective rank (Koltchinskii and Lounici, 2016), and d^* is known as the numerical rank or stable rank of Σ (Rudelson and Vershynin, 2007; Tropp et al., 2015).

These three notions of effective dimension give a quantification of the distribution's degrees of freedom. If the distribution $\mathbb{P}$ is isotropic, i.e. whose covariance is $\Sigma(\mathbb{P}) = \sigma^2 I_d$, where d is the dimension of the ambient space, then all these notions of effective dimension are equal to d. More generally, we have the following inequalities between effective dimensions:

$$d \geq d^{\bullet}(\mathbb{P}) \geq d^{\mathrm{e}}(\mathbb{P}) \geq d^{*}(\mathbb{P})\,. \tag{I.19}$$

The following problems give very simple situations in which each of the dimensions of Eq. (I.18) are involved.

- The dimension d^{e} characterizes estimation problems. For example, if we want to estimate the expectation of a distribution from an observation $X \sim \mathbb{P}$, the squared error of the estimation is :

  $$\mathbb{E}\Big[\|X - \mu\|^2\Big] = \sigma^2 d^{\mathrm{e}}\,,$$

  where $\mu = \mathbb{E}[X]$ and $\sigma^2 = \|\Sigma|_{op}$. We find the form (I.1).

- The dimension d^* is more relevant to test problems. For example, to detect that the expectation μ of a distribution is close to 0, the distance $\|\mu\|^2$ can be estimated by $\langle X, X' \rangle$ where X and X' are two independent observations. Then :

$$\mathbb{E}[\langle X, X' \rangle] = \|\mu\|^2, \quad \text{and} \quad \mathrm{Var}[\langle X, X' \rangle] \geq \sigma^4 d^*$$

where $\sigma^2 = \|\Sigma|_{op}$. Roughly speaking, we recover that the test error is $\sqrt{d}$ versus d for the estimation, but with different dimensions.

- The dimension $d^\bullet$ is less natural to interpret, but comes into play in Section V, where we will use tests to improve vector estimation. It can be seen as the ratio of the estimation error to the test error:

$$\sqrt{d^\bullet} = \frac{d^{\mathrm{e}}}{\sqrt{d^*}} \simeq \frac{\mathbb{E}\left[\|X - \mu\|^2\right]}{\sqrt{\mathrm{Var}[\langle X, X' \rangle]}}.$$

In the following examples, we calculate the effective dimensions of different distributions.

Example I.8 (Support of smaller dimension)**.** Suppose that the distribution $\mathbb{P}$ is supported in a vector subspace of dimension p. Then

$$d^{\mathrm{e}}(\mathbb{P}) \leq p.$$

Suppose now that $\mathcal{H}$ is of finite dimension d and consider $\widetilde{\mathbb{P}}$ the distribution $\mathbb{P}$ noised by addition of an independent and isotropic noise of covariance $\varepsilon^2 I_d$. Then

$$d^{\mathrm{e}}\big(\widetilde{\mathbb{P}}\big) \leq p + d\frac{\varepsilon^2}{\|\Sigma(\mathbb{P})\|_{op}}.$$

If the noise is low enough ($\varepsilon^2 \ll d^{-1}\|\Sigma(\mathbb{P})\|_{op}$), the effective dimension captures that the distribution $\mathbb{Q}$ is a noisy version of a lower-dimensional support distribution.

Example I.9 (Discrete RKHS)**.** Consider $\mathcal{H}$ the RKHS associated with the trivial kernel $\kappa(x, y) = \mathbf{1}\{x = y\}$ defined on the discrete space $\mathcal{X} = \{x_1, \ldots, x_m\}$. Let $\mathbb{Q} = \sum_{i=1}^m p_i \delta_{x_i}$ be a distribution on $\mathcal{X}$ and $\mathbb{P}$ the distribution of the pushfroward of $\mathbb{Q}$ into $\mathcal{H}$. The effective dimension d^{e} of $\mathbb{P}$ is then lower and upper bounded by :

$$\frac{1}{2}\frac{1 - \|p\|_2^2}{\|p\|_\infty(1 - \|p\|_\infty)} \leq d^{\mathrm{e}}(\mathbb{P}) \leq \frac{1 - \|p\|_2^2}{\|p\|_\infty(1 - \|p\|_\infty)}$$

where $p \in [0, 1]^m$ is the probability vector of $\mathbb{Q}$. In this case, the notion of effective dimension is interpreted as a normalized version of the Gini-Simpson index (Simpson, 1949). The Gini-Simpson index is $1 - |p\|_2^2$ and measures whether a population is diverse. In our framework, a diversified population (limiting case $p_i \simeq m^{-1}$) will have a large effective dimension ($d^{\mathrm{e}} \simeq m$).

Example I.10 (Translation kernel)**.** Consider $\mathcal{H}$ the RKHS associated with a translation kernel $\kappa_h(x, y) := K((x - y)/h)$ where $h > 0$, the *bandwidth* of the kernel, is a fixed parameter and $x, y \in \mathcal{X} = \mathbb{R}^d$. Let $\mathbb{Q}$ be a distribution on $\mathbb{R}^d$ with density f with respect to the Lebesgue measure and $\mathbb{P}_h$ the distribution of the pushforward of $\mathbb{Q}$ into $\mathcal{H}$. Assuming K and f are enough regular, for small bandwith, the effective dimension relates to the L^2 norm of the density. Indeed, as the bandwidth h tends to 0 :

$$d^\bullet(\mathbb{P}_h) \underset{h \to 0}{\sim} \frac{K(0)^2}{h^d \|K\|_{L^2}^2 \|f\|_{L^2}^2}.$$

A diffuse distribution ($\|f\|_{L^2}^2$ small) will have a larger effective dimension in the RKHS than a concentrated distribution. For example, if K is a Gaussian kernel and $\mathbb{Q} \sim \mathcal{N}(\mu, \sigma^2 I_d)$ a Gaussian distribution, then the effective dimension depends on the ratio between the variance and the bandwidth:

$$d^\bullet(\mathbb{P}_h) \underset{h\to 0}{\sim} \left(\frac{8\sigma^2}{h^2}\right)^{d/2}, \qquad \text{moreover} \qquad d^\bullet(\mathbb{P}_h) \underset{h\to\infty}{\longrightarrow} d\,. \tag{I.20}$$

In this Gaussian case, we explicitly find that the effective dimension will be greater for a distribution with high variance.

These effective dimensions also relate to the notion of local covariance dimension (Dasgupta and Freund, 2008) defined from the eigenvalues $\sigma_1^2 \geq \ldots \geq \sigma_d^2$ of the covariance Σ. For example, Verma et al. (2009) define that a distribution $\mathbb{P}$ is of covariance dimension (p, ε) if the p largest eigenvalues represent a proportion $(1-\varepsilon)$ of the covariance trace:

$$\sum_{i=1}^{p} \sigma_i^2 \geq (1-\varepsilon)\operatorname{Tr}\Sigma\,.$$

This notion can also be defined locally, by assuming that this condition is verified not for the covariance but for the covariances of the distributions restricted to each ball of a given radius. Note that for a fixed $\varepsilon > 0$, a distribution $\mathbb{P}$ has covariance dimension $(d^{\mathrm{e}}, \varepsilon)$ for effective dimensions of the order of the ambient dimension ($d^{\mathrm{e}} \gtrsim (1-\varepsilon)d$) or close to 1 ($d^{\mathrm{e}} \lesssim (1+\varepsilon)$). More generally, $\mathbb{P}$ is of covariance dimension $(d^{\mathrm{e}}, \varepsilon_{d^{\mathrm{e}}})$ for some $\varepsilon_{d^{\mathrm{e}}} \leq (1-\frac{1}{d^{\mathrm{e}}})(1-\frac{d^{\mathrm{e}}-1}{d-1})$.

On a more geometric side, for a probability measure, some notions of dimension will be defined from the evolution of a ball measure with its radius. Roughly speaking, the distribution $\mathbb{Q}$ will be of dimension d if the probability of any ball S_r of radius $r > 0$ is proportional to the volume of a d-dimensional ball, i.e. if we have $\mathbb{P}(S_r) < Cr^d$ for some constant C (for radius r bounded or going to 0, depending on the case). These notions are strongly linked to the Hausdorff (1918) and Assouad (1979) dimensions, generalized to distributions for example by the *pointwise dimension* (Young, 1982), the *information dimension* (Isham, 1993), the *doubling dimension* (see Heinonen, 2001) or by the notion of maximally homogeneous distribution (Kpotufe, 2011). If the distribution has a density on a sub-vector space, these dimensions will coincide with the dimension of this space, just like the dimensions we are considering (see Example II.8). However, these notions will diverge for discrete distributions. For such distributions and small radii, the ball measures will no longer evolve with the radius, giving a dimension equal to 0. Conversely, dimensions (I.18) will consider the structure of the values taken by the distribution in the ambient space. For example, for an uniform distribution on an orthonormal family of vectors (e_k):

$$d^{\mathrm{e}}\left(\frac{1}{n}\sum_{k=1}^{n}\delta_{e_k}\right) = n-1\,.$$

For a discrete distribution on a line, d^{e} will be equal to 1.

I.4 Contributions

A recurrent objective in this thesis is to make the bridge between high-dimensional phenomena studied for isotropic Gaussian data and current modern tools such as the KME presented above. The central problem we consider is the simultaneous estimation of mean vectors, which can be KMEs of different distributions. This problem can be considered as a multi-task learning instance, or even transfer learning where the user seeks to estimate this new vector with the help of estimators from other objects, or even federated learning where the distributed data may have variations in their distributions and only a vector can be transmitted.

We will consider it here in the most general way, seeing these vectors as elements of a Hilbert space whose observations are perturbed by Gaussian or bounded noise.

We therefore consider this problem in Section III and propose a method using a Stein-type effect where the improvement in estimation increases with dimension. Guarantees are given under homogeneity assumptions between the different distributions. For KMEs, high dimension then means high effective dimension. This method is based on tests and takes advantage of the relative ease of detection compared with estimation. In Section IV we generalize this testing phenomenon to non-isotropic data and in particular to KMEs. This work, although independent, also helps to extend our multiple vector estimation to heterogeneous data presented in Section V. In this section we propose two methods, one based on these tests and the other on an empirical risk minimization. The combination of the samples improves the estimation of each mean, particularly when the distributions have a common structure. We have observed that this phenomenon is indirectly present in the self-attention mechanism used in Transformers (Vaswani et al., 2017). In Section VI, we thus show that self-attention works as a denoising of high-dimensional data and that we find comparable behaviors in high dimension to those considered in the previous sections.

These contributions are detailed in this section.

I.4.1 Multiple estimation of mean vectors for homogeneous data

Section III is a work in collaboration with Gilles Blanchard and Hannah Marienwald (Marienwald et al., 2021) and focuses on the simultaneous estimation of mean vectors for different distributions. Consider the model:

$$\begin{cases} X_\bullet^{(k)} := (X_i^{(k)})_{1\leq i\leq N_k} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_k, \ k \in [\![B]\!]; \\ (X_\bullet^{(1)}, \dots, X_\bullet^{(B)}) \text{ independent.} \end{cases} \tag{I.21}$$

the objective is to estimate the mean vector $\mu_k = \mathbb{E}\left[X_1^{(k)}\right]$ of each distribution $\mathbb{P}_k$ for which a sample $X_\bullet^{(k)}$ is provided. This problem can be seen as a multi-task problem where each task is to estimate a vector μ_k. This vector can be, for example, the KME of a distributions $\mathbb{P}_k$ for which we have at our disposal a sample $X_\bullet^{(k)}$ already immersed in the RKHS. For a given sample, we construct a shrinkage estimator of its empirical mean towards a reference point. This reference point will not be chosen arbitrarily as for the James-Stein's estimator, but from the other samples. The aim is to use the relative ease in high dimension of the test compared with estimation to find a reference point close to the true mean.

For isotropic Gaussian data and homogeneous samples, $N_k = N$ and $\mathbb{P}_k = \mathcal{N}(\mu_k, \sigma^2 I_d)$, we estimate for each mean a set of τ-neighbors $\widehat{V}_i = \{j : T_{ij} = 1, j \in [\![B]\!]\}$ where T_{ij} is a test for the hypotheses:

$$(H_{0,ij}) : \|\mu_i - \mu_j\|^2 \leq \tau\sigma_N^2 d, \quad \text{against} \quad (H_{1,ij}) : \|\mu_i - \mu_j\|^2 > \tau\sigma_N^2 d. \tag{I.22}$$

where $\sigma_N^2 = \sigma^2/N$. The aim of these tests is to find samples whose means are close to the target mean with respect to the estimation error ($d\sigma_N^2$). The estimator considered for μ_i is a shrinkage estimator of the empirical mean to the average of the empirical means selected by the tests:

$$\widehat{\mu}_i = \gamma\widehat{\mu}_i^{\texttt{NE}} + \frac{1-\gamma}{|\widehat{V}_i|}\sum_{j\in\widehat{V}_i}\widehat{\mu}_j^{\texttt{NE}}, \tag{I.23}$$

where $\widehat{\mu}_j^{\texttt{NE}}$ is the empirical mean over sample j and $\gamma \in (0,1)$ a parameter to be fixed. The test error, in squared norm, for isotropic Gaussian distributions is of the order of $\sigma_N^2\sqrt{d}$, so we expect to be able to construct tests such that the selected neighbors are at a distance from the true mean μ_i of at most $\tau\sigma_N^2 d + \sigma_N^2\sqrt{d}$ (see Section I.1.2). Thus in Section III:

- We construct tests such that the bias added by this shrinkage is of order smaller than the estimation error of the empirical mean, and we show theoretically and experimentally that our method improves the estimation relatively to the empirical mean.

- The data are assumed to be isotropic Gaussian or bounded in order to apply our method to the estimation of KMEs. In both cases, we make an assumption of sample homogeneity: sample sizes and variances are assumed to be of the same order. This assumption justifies our symmetrical use of the means selected by the tests in (I.23).

This section can be seen as an introduction to Section V where the method is generalized to non-homogeneous samples. However, the method presented in Section III retains an independent interest by proposing a simple, low algorithmic cost approach with finer-grained theoretical guarantees. In particular, our bounds take into account the data dependency between test and estimate that we set aside in Section V.3 (see, for example, Theorem III.2).

I.4.2 One- or two-sample high-dimensional tests with unknown covariance structure

The method proposed in Section III is based on the analysis in the isotropic Gaussian framework of the separation distance (I.8). In order to generalize this method, we are interested in the two-sample means proximity test problem:

$$(H_0(\eta)) : \|\mu - \nu\| \leq \eta, \text{ against } (H_1(\eta, \delta)) : \|\mu - \nu\| > \eta + \delta. \tag{I.24}$$

where μ and ν are the respective mean vectors of $\mathbb{P}$ and $\mathbb{Q}$ distributions known via i.i.d. samples $\{X_i\}_{1\leq i\leq n}$ and $\{Y_i\}_{1\leq i\leq m}$. This problem is a generalization of the signal detection problem (I.6): it can be reduced by formally assuming that $m = +\infty$ or that $\mathbb{Q}$ is a Dirac distribution. When μ and ν are KMEs, this test can be used to test the proximity of two distributions in term of MMD distance. In this form, the separation distance is the smallest distance δ_α for which there is a test such that the sum of type I and II errors are controlled by a given $\alpha \in (0, 1)$. The important point in our analysis is that the covariances of each distribution are not assumed to be known and are potentially different. Our contributions are as follows:

- We perform a minimax analysis of the test (II.24) and give a lower bound on the separation distance for Gaussian data. We find the two regimes of Blanchard et al. (2018) (Eq.(I.8)) but where the dimension of the space is replaced by a notion of effective dimension of the problem.
- We construct tests reaching this lower bound for Gaussian and bounded data in a Hilbert space. Our tests are constructed from U-statistics of the form (I.13) and estimators of its quantiles. Concentration inequalities are given for all these estimators in the Gaussian and bounded framework.

This work has been done in collaboration with Gilles Blanchard (Blanchard and Fermanian, 2023).

I.4.3 Generalization of multiple mean estimation and minimaximality

Section V is based on these two previous parts and has been done in collaboration with Gilles Blanchard and Hannah Marienwald (Blanchard et al., 2024). We again consider the model (I.21), but no longer assume any homogeneity between the distributions. The sample sizes are different, as are the covariances of the distributions, which are assumed to be unknown. To decide whether to aggregate two estimators, we need to take into account both the proximity of their means and the ratio of their variances. Hence, even if all samples have the same mean, the estimators of those with a smaller variance should be preferred. For this purpose, we consider as estimator a convex combination of empirical means

$$\widehat{\mu}_{\boldsymbol{\omega}} = \sum_{k=1}^{B} \omega_k \widehat{\mu}_k^{\mathtt{NE}}, \tag{I.25}$$

where $\widehat{\mu}_k^{\mathtt{NE}}$ is the empirical mean of sample k and $\boldsymbol{\omega}$ is a weight vector in the simplex $\mathcal{S}_B$ (i.e. $\sum_{k=1}^B \omega_k = 1$ and $\omega_k \geq 0$). This form of estimator is more general than the one considered in Section III (the estimator (I.23) is reduced to (I.25) by taking $\gamma = \omega_1$ and $\omega_k = (1-\gamma)\mathbf{1}_{k\in\widehat{V}_1}|\widehat{V}_1|^{-1}$). Our aim is still to improve the estimation of each mean relatively to the naive estimation by the empirical mean. To this end, we propose two methods for estimating optimal $\boldsymbol{\omega}$ weights.

The first method presented in Section V.3 is based on tests and is an adaptation of the method in Section III in the heterogeneous case. To estimate the mean μ_1, our first step is to select (relatively) close means with (relatively) smaller variances. We will try to estimate:

$$V_\tau = \left\{k \in [\![B]\!] : \|\mu_k - \mu_1\|^2 \leq \tau\sigma_1^2 d^{\mathrm{e}}(\mathbb{P}_1)\right\}, \quad \text{et} \quad W_{(\varsigma)} = \left\{k : \sigma_k^2\big(d^*(\mathbb{P}_k)\big)^{1/2} \leq \varsigma\sigma_1^2\big(d^*(\mathbb{P}_1)\big)^{1/2}\right\} \tag{I.26}$$

where $\tau, \varsigma > 0$ are fixed parameters and $\sigma_k^2 = \|\Sigma_k\|_{\mathrm{op}}/N_k$. In words, V_τ contains the distributions whose means are close to μ_1 up to the estimation error and $W_{(\varsigma)}$ contains distributions whose test errors are not bigger than that of the $\mathbb{P}_1$ distribution. The set $W_{(\varsigma)}$ is used to exclude means that could be selected from the tests for V_τ but whose means could actually be too far from μ_1. These sets are estimated using tests. The second step consists in estimating weights $\boldsymbol{\omega}$ given by an oracle minimization of the theoretical risk. The weight ω_k assigned to the estimator $\widehat{\mu}_k^{\mathtt{NE}}$ will decrease with its estimation error $\sigma_k^2 d^{\mathrm{e}}(\mathbb{P}_k)$.

This approach differs from the first method presented by the non-symmetry of the relationship "being a τ-neighbor". The estimator of a sample with large variance and/or small size will have more neighbors and be more shrunk than the one of a large sample. The method will more improve the estimation for the samples with less initial information, but will not deteriorate the estimation for the others.

We give non-asymptotic bounds on the error of this method and on its various stages. We propose estimates of V_τ and $W_{(\varsigma)}$ that can be replaced by other estimators if required. This improves estimation in particular when the test error is small relative to the estimation error, i.e. when the effective dimension $\sqrt{d^\bullet} = d^{\mathrm{e}}/\sqrt{d^*}$ is large (for $\mathbb{P}_1$).

A weakness of the two test approaches (homogeneous and heterogeneous) is the need to choose the parameters τ and ς. We therefore propose a second method that uses ideas from the Q-aggregation of Lecué and Rigollet (2014). Instead of selecting neighbors using tests, we estimate some weights by minimizing an estimator of an upper bound of the quadratic risk $\mathcal{R}_1(\boldsymbol{\omega}) := \mathbb{E}\Big[\|\widehat{\mu}_{\boldsymbol{\omega}} - \mu_1\|^2\Big]$. This estimator is made up of two terms, the first $\widehat{L}_1(\boldsymbol{\omega})$ is an estimator of the risk $\mathcal{R}_1(\boldsymbol{\omega})$ and the second $\widehat{Q}_1(\boldsymbol{\omega})$ is an estimator of the deviations of $\Big|\widehat{L}_1(\boldsymbol{\omega}) - \mathcal{R}_1(\boldsymbol{\omega})\Big|$. Our estimator is then $\widehat{\mu}_{\widehat{\boldsymbol{\omega}}}$ where

$$\widehat{\boldsymbol{\omega}} \in \operatorname*{Arg\,Min}_{\boldsymbol{\omega}\in\mathcal{S}_B}\Big(\widehat{L}_1(\boldsymbol{\omega}) + \widehat{Q}_1(\boldsymbol{\omega})\Big).$$

The term $\widehat{Q}_1$ is derived from concentration inequalities and involves the covariances of the distributions. Weighting by $\widehat{Q}_1$ the minimization will impose a form of sparsity on the vector $\widehat{\boldsymbol{\omega}}$ in the same way as a ℓ_1 penalty, but taking into account the dimensionality of the different samples. We interpret these as tests implicitly performed by this regularization.

We construct such estimators and give bounds on the mean square error of the estimator $\widehat{\mu}_{\widehat{\boldsymbol{\omega}}}$. We find in these bounds the same rate of convergence $O((d^\bullet)^{-1/2})$ as in the test approach. The Q-aggregation has the advantage, however, of being adaptive in τ and ς. In practice, the two methods achieve comparable results.

Intuition. Our aim is to choose estimator that upper bounds the true risk, i.e. such that with high probability:

$$\mathcal{R}_1(\boldsymbol{\omega}) \leq \widehat{L}_1(\boldsymbol{\omega}) + \widehat{Q}_1(\boldsymbol{\omega}) + O(\sqrt{d^*}) \leq \mathcal{R}_1(\boldsymbol{\omega}) + O(\sqrt{d^*}), \quad \forall\boldsymbol{\omega} \in \mathcal{S}_B.$$

Indeed, we expect deviations in distance estimation to be of the order of $\sqrt{d^*}$. The parallel can be drawn with the deviation of a Gaussian vector $Z \sim \mathcal{N}(\mu, \Sigma)$. Suppose we want to estimate a "close" upper bound of $\mathcal{R} = \mathbb{E}\left[\|Z-\nu\|^2\right]$ for a vector $\nu \in \mathbb{R}^d$. Then for all $u \geq 0$, with probability $1 - e^{-u}$:

$$\mathbb{E}\left[\|Z-\nu\|^2\right] \leq \|Z-\nu\|^2 + 2\sigma^2\sqrt{(d^* + 2(\mu-\nu)^T\Sigma(\mu-\nu))u}$$

where $\sigma^2 = \|\Sigma\|_{\text{op}}$ and $d^* = d^*\big(\mathcal{N}(\mu,\Sigma)\big)$ is the effective dimension of the Gaussian distribution (see Lemma V.40). In this case, $\|Z-\nu\|^2$ would be our estimator $\widehat{L}_1$ and $\widehat{Q}_1$ would be an estimator of $\sqrt{(\mu-\nu)^T\Sigma(\mu-\nu)}$. The deviation of the order of $\sqrt{d^*}$ is clearly present.

The errors of these two methods are studied under a high (effective) dimensional point of view. As with the James-Stein estimator (see Section I.1.1), the improvement lies in the dimensional dependence of the rate of convergence. In Section V.5, we perform a minimax analysis of the problem and give lower bounds for the optimal improvement possible for one sample and in average over the samples. We then discuss the optimality of our two methods.

I.4.4 Effect of denoising of self-attention mechanism

Section VI presents some preliminary works in collaboration with Gilles Blanchard on the attention mechanism. This mechanism is used in the Transformer neural networks (Vaswani et al., 2017) widely used today, particularly for text or image data generation tasks. In these cases, the inputs of the neural network are no longer a single point, but a set of points that can be the words of a sentence or text, or the encodings of sub-parts of an image. To process this type of data, Transformers add additional layers to the neural network, known as attention layers, which seek to consider the points not individually, but as a whole. The intuition is very simple: to translate a word in a sentence, it is important to take its context into account. Formally, for points $X_1, \ldots, X_N$, the attention first constructs N new points:

$$a_{Q,K}(X_i) := \sum_{j=1}^{N} \omega_{ij} X_j \quad \text{where} \quad (\omega_{ij})_j = \texttt{Softmax}\Big(\big(\langle QX_i, KX_j\rangle\big)_j\Big) \in \mathcal{S}_N\,, \tag{I.27}$$

where $\mathcal{S}_N$ is still the simplex and Q and K are matrices learned during the neural network training. We assume here that these matrices are fixed, so we place ourselves after the training and seek to understand the action of attention on the data. By comparing this form with (II.25), we propose to interpret the attention mechanism as a form of multiple vector estimation. We thus assume that the points X_i are noisy observatoins of vectors μ_i, and we ask whether the new points $a_{Q,K}(X_i)$ would be less noisy compared to the original points X_i, i.e. whether for $1 \leq i \leq N$:

$$\mathbb{E}\left[\|a_{Q,K}(X_i) - \mu_i\|^2\right] < \mathbb{E}\left[\|X_i - \mu_i\|^2\right].$$

for the dimension tending to infinity. An improvement is expected when vectors μ_i have a simpler structure which can be learned through the other points. We analyze this question in a simplified framework where the random vectors X_i have a Gaussian isotropic distribution and where the matrices Q and K are proportional to the identity ($Q = K = I_d/\sqrt{h}$). In this case we study the values of the parameter h for which the self-attention mechanism is not degenerate (i.e. $a_{Q,K}(X_i) \neq X_i$ and $a_{Q,K}(X_i)$ different from the mean of the data). For such parameters, we exhibit certain structures of the points μ_i for which a denoising effect actually occurs (support of smaller dimension or small covering number for a certain radius). Based on this analysis of the effect of dimension, we propose a slightly modified version of the weights ω of (I.27) for which we obtain theoretically and on simulated data a denoising effect for a wider spectrum of parameter h ($h \simeq d$ for the original methods versus $\sqrt{d} \lesssim h \lesssim d$ for our modified versions).

An important assumption in our analysis is that the vectors μ_i have the same norm. For such vectors, the weights given by scalar products are the same as those given by the squared norm ($\texttt{Softmax}((\langle \mu_i, \mu_j \rangle)_j) = \texttt{Softmax}\Big((-\|\mu_i - \mu_j\|^2/2)_j\Big)$). Since we also assume that Q and K are proportional to the identity, we can interpret $\langle QX_i, KX_j \rangle$ as an estimator of the Euclidean distance between μ_i and μ_j (up to an additive and a factor $1/2$). As for $i = j$, this estimator is biased, our modified attention weights just remove this bias in two different ways (see Definition VI.7).

The assumption that the vectors μ_i are on the sphere is justified for us by the second stage of attention, which consists in combining the initial points with $a_{Q,K}(X_i)$ and constructing for $1 \leq i \leq N$:

$$A_{Q,K,V}(X_i) := \frac{X_i + V a_{Q,K}(X_i)}{\|X_i + V a_{Q,K}(X_i)\|}, . \tag{I.28}$$

for a matrix V also learned by the neural network. Thus the self attention associates to each point a normalized vector. Although the form (I.28) is frequently employed to study self-attention, the normalization step is often neglected. In our model, these vectors $A_{Q,K,V}(X_i)$ can be interpreted as (matrix) contraction estimators from X_i to $a_{Q,K}(X_i)$ and relate to empirical Bayes estimators (Brown and Greenshtein, 2009) for certain V matrices (see discussion in Section VI.1.1). For the moment, however, our analysis focuses on $a_{Q,K}(X_i)$ with the intuition that a denoising of $a_{Q,K}(X_i)$ would denoise $A_{Q,K,V}(X_i)$. These results, although restricted to simplified versions of the attention mechanism in an isotropic Gaussian framework, highlight a new phenomenon and raise many new questions.

I.4.5 Concentration inequalities

These various works repeatedly use estimators of distance between vectors and of moments of the covariance of distributions. These estimates are controlled by concentration bounds that may be of interest independently of the problems we consider. In particular, these bounds pay close attention to the role of (effective) dimension in deviations. In addition to Gaussian and bounded distributions, we also consider heavy-tailed distributions where only a finite fourth-order moment is assumed. For this type of data, "median of means" estimators are considered (see Lugosi and Mendelson, 2019a for example). Table 1 points to the various results disseminated in the thesis.

Target quantity	Gaussian setting	Bounded setting	Heavy-tailed setting
$\|\mu - \nu\|^2$	Proposition IV.6	Proposition IV.9	Proposition V.33
$\operatorname{Tr}\Sigma$	Proposition V.26	Proposition V.29	Proposition V.34
$\|\Sigma\|_{\text{op}}$	Proposition IV.10	Proposition IV.11	-
$\sqrt{\operatorname{Tr}\Sigma^2}$	Proposition IV.12	Proposition IV.13	Proposition V.34
$(\mu - \nu)^T \Sigma (\mu - \nu)$	Proposition V.37	Proposition V.38	-

Table 1: Survey of different concentration results: μ and Σ are respectively the mean and the covariance of a distribution and ν the possibly unknown mean of another distribution.

II Introduction (Français)

Un enjeu actuel et important de la recherche en statistique et machine learning est l'analyse de données de grande dimension. Les méthodes modernes ont accès et cherchent à traiter des données de plus en plus complexes (signals, textes, images, vidéos...). Bien que dans certains domaines cette complexification des données s'accompagne d'une augmentation de leur quantité, dans d'autres, le nombre de données est limité et la prise en compte de cette complexité devient essentielle. Ainsi un faible nombre de données relativement à leur dimension cause en général une perte de performance connue sous le terme général de *malédiction* ou *fléau de la dimension*. Cette appellation introduite par Bellman (1966) (*curse of dimensionality*) regroupe de nombreux phénomènes dont nous présenterons certains (voir aussi Giraud, 2021).

En parallèle de cette perte de performance, les intuitions de la petite dimension ne sont plus forcément pertinentes en grande dimension, et certaines méthodes a priori plus saugrenues peuvent devenir intéressantes. Un exemple fondamental et bien connu de ce type de phénomène est l'inadmissibilité de la moyenne empirique en grande dimension : Stein (1956) montre que pour estimer le vecteur moyenne d'une distribution gaussienne, la moyenne empirique n'est pas efficace et exhibe un meilleur estimateur (James et Stein, 1961). Cet estimateur, tout en ayant la même vitesse de convergence en la taille de l'échantillon que la moyenne empirique, le surpasse grâce à une dépendance plus faible de son erreur en la dimension. Cet exemple illustre la nécessité d'inclure dans l'analyse d'un problème non plus seulement la dépendance en la taille de l'échantillon mais aussi en la taille ou complexité de l'espace. Une possibilité pour capturer cette dépendance est de procéder à une analyse non asymptotique des problèmes. Du côté minimax, l'analyse de Pinsker (1980) montre par exemple que l'estimateur de James-Stein est optimal à taille d'échantillon fixé mais pour la dimension de l'espace tendant vers l'infini. Pour les problèmes de tests, on retrouve cette prise en compte de la dimension dans l'analyse minimax du problème de détection de signal de Baraud (2002) et Blanchard et al. (2018) par exemple. Pour contrôler l'erreur de nos méthodes et capturer les effets de la grande dimension, nous utiliserons dans ce manuscrit des inégalités de concentration pour obtenir des bornes non asymptotiques.

Bien que les données appartiennent à l'origine à un espace ambiant de grande dimension, il a été constaté qu'elles vivent souvent en réalité dans des sous-espaces de plus petite dimension (sous-espace vectoriel, sous-variété, petit nombre de clusters...), ce qui permet aux méthodes de fonctionner. La distribution des données peut ainsi être à support dans tout l'espace, mais avoir certaines directions non informatives et juste constituées de bruit. La difficulté du problème n'est alors plus caractérisée par la dimension ambiante mais par des notions de *dimensions effectives*. Ces quantités, justifiées par l'analyse minimax, prennent en compte la structure de covariance des données et quantifient les degrés de libertés d'une distribution. Par exemple on aimerait dire qu'une variable distribuée sur une droite plongée dans un espace ambiant de grande dimension n'est en réalité que de dimension 1. Ces notions sont particulièrement utiles pour analyser des données fonctionnelles qui bien que dans un espace de dimension infinie peuvent avoir une dimension effective finie et faible. Par exemple les méthodes à noyaux, couramment utilisées, injectent les données dans un espace de Hilbert fonctionnel (un RKHS plus précisément). Dans ce cas, l'erreur d'une méthode dépend de la dimension effective de la distribution des plongement des données dans le RKHS. Cette notion permet d'analyser simultanément des problèmes de dimension finie et infinie et floute ainsi la différence entre le paramétrique et le non paramétrique.

Dans cette thèse, nous chercherons à comprendre les effets de la grande dimension sur des problèmes de tests et d'estimation en vue de potentiellement améliorer ces méthodes. Cette thèse est constituée de quatre travaux :

- l'article "High-Dimensional Multi-Task Averaging and Application to Kernel Mean Embedding" en collaboration avec Hannah Marienwald et Gilles Blanchard publié à *AISTATS*, (2021) ;
- le chapitre "Nonasymptotic One- and Two-Sample Tests in High dimension with Unknown Covariance

Structure" en collaboration avec Gilles Blanchard de *Foundations of Modern Statistics : Festschrift in Honor of Vladimir Spokoiny*, (2023) ;

- le preprint "Estimation of multiple mean vectors in high dimension" en collaboration avec Hannah Marienwald et Gilles Blanchard, (2024) ;
- des premiers résultats sur une analyse du mécanisme de self-attention en grande dimension en collaboration avec Gilles Blanchard.

Aperçu des problématiques : Le problème principal considéré dans cette thèse est l'estimation simultanée de vecteurs moyennes en grande dimension. Nous cherchons à estimer des vecteurs moyennes notés μ_k de différentes distributions $\mathbb{P}_k$ définies sur un même espace de Hilbert $\mathcal{H}$. Ce problème mèle les questions statistiques classiques liées à l'estimation d'une moyenne en grande dimension datant de Stein (1956) et celles plus récentes de *multi task learning* (MTL) (Bonilla et al., 2007 ; Micchelli et Pontil, 2004) où l'objectif est d'effectuer simultanément plusieurs tâches indépendantes mais ayant des similarités. Notre tâche ici est l'estimation d'un vecteur moyenne qui peut être vue comme une version idéalisée de problèmes plus complexes comme de la régression, classification, estimation de covariance, de distribution... Dans notre cas où les tâches consistent à estimer un vecteur moyenne, on parle alors de *multi task averaging* (MTA) (Feldman et al., 2014). Ce problème devient plus concret lorque ces vecteurs sont les *kernel mean embeddings* (KME) de distributions. Cet outil aux nombreuses propriétés est central dans les méthodes à noyaux et nous chercherons à améliorer son estimation dans le cadre MTL où plusieurs d'entre eux sont estimés simultanément.

Le point de départ de nos approches vient cependant du cas de distributions gaussiennes isotropes, $\mathbb{P}_k = \mathcal{N}(\mu_k, \sigma^2 I_d)$, et d'échantillons homogènes, $N_k = N$. Dans ce cas, l'erreur minimax d'estimation d'une moyenne est atteinte par la moyenne empirique et est d'ordre $\sigma^2 d/N$ (en norme quadratique). Il est connu qu'en grande dimension cette erreur d'estimation est bien plus importante que l'erreur de test, la plus petite distance entre deux moyennes pour laquelle il est possible de les distinguer, qui est de l'ordre de $\sigma^2\sqrt{d}/N$. Nous proposons donc de tester l'égalité entre les vecteurs μ_k deux à deux pour détecter les moyennes proches et ensuiter estimer chacune d'entre elles par un estimateur de contraction vers les moyennes empiriques ainsi sélectionnées. L'intuition est que la contraction réduira la variance de l'estimation au prix de l'ajout d'un biais. Dans notre cas, on s'attend à un biais d'ordre inférieur à la variance en grande dimension ($\sqrt{d} \ll d$). L'amélioration donnée par cet estimateur sera ainsi notable en grande dimension où l'estimation est particulièrement difficile. Ce phénomène rappelle ainsi le paradoxe de Stein (1956) qui contracte la moyenne empirique vers un point de référence pour améliorer son risque quadratique.

Une importante partie de cette thèse est dédiée à la justification de cette intuition, la construction d'une méthode et sa généralisation à des données non gaussiennes, non isotropes et des échantillons non homogènes. Cette étude est nécessaire pour pouvoir traiter le cas des KMEs où les données sont dans un espace fonctionnel de dimension infinie. Pour de telles distributions, le paramètre critique de la dimension est remplacée par une notion de dimension effective qui sera construite à partir des covariance des distribution. Cette notion interviendra dans notre étude de la vitesse de séparation de tests de proximité de moyennes pour des distributions non isotropes, permettant de généraliser le phénomène connu du cas isotrope gaussien au cadre KME et, plus généralement, pour des distributions bornées.

Dans la suite de cette introduction nous présentons certaines des notions liées à notre problème que nous venons d'évoquer ci-dessus. En Section II.1.1, nous présentons l'estimateur de James et Stein (1961), un des plus célèbre estimateurs de contraction en grande dimension et relions sa construction à différentes idées qui seront utilisées dans nos méthodes. Nous présentons ensuite en Section II.1.2 la notion de vitesse de séparation de tests de proximité et l'influence de la dimension dans cette vitesse dans le cas de distributions gaussiennes isotropes. En Section II.2, nous introduisons le KME d'une distribution et certaines de ses applications. Cet objet de grande dimension, très utilisé en machine learning, motive grandement notre problème et nous amène à devoir considérer des notions de dimension effective. Ces notions sont présentées

en Section II.3 accompagnées de différents exemples. En Section II.4 nous présentons plus en détail les différentes contributions de la thèse.

II.1 Estimation et test en grande dimension

Les problèmes d'estimation et de test, centraux dans cette thèse, ont déjà été étudiés sous l'axe de la grande dimension. Son effet pour ces problèmes a été mis en valeur en particulier dans le cadre classique de distributions gaussiennes isotropes. Nous présentons donc dans cette section, dans un premier temps, le paradoxe de Stein (1956) lié à l'estimation d'un vecteur et son lien avec des problématiques et des méthodes plus modernes. Puis nous présenterons des résultats sur la vitesse de séparation de test pour lesquels Baraud (2002) est le premier à avoir pris en compte l'influence de la dimension. Ces travaux sont finalement le point de départ des différentes questions posées dans cette thèse.

II.1.1 Estimation d'un vecteur : le paradoxe de Stein

Le paradoxe de Stein est un exemple typique des phénomènes contre-intuitifs de la grande dimension. Considérons un échantillon $(X_i)_{1\leq i\leq N}$ de vecteurs aléatoires dans $\mathbb{R}^d$ de loi gaussienne notée $\mathcal{N}(\mu,\sigma^2 I_d)$ où la variance σ^2 est connue et le vecteur moyenne $\mu\in\mathbb{R}^d$ est inconnu. L'objectif est d'estimer ce vecteur μ en minimisant le risque quadratique pour la distance euclidienne. Soit $\mathcal{G}_d=\{\mathcal{N}(\mu,\sigma^2 I_d)^{\otimes N}:\mu\in\mathbb{R}^d\}$ l'ensemble des distributions de N-échantillons gaussiens à covariance isotrope fixée, l'erreur minimax d'estimation du vecteur μ est alors :

$$\inf_{\widehat{\mu}}\sup_{\mathbb{P}\in\mathcal{G}_d}\mathbb{E}\Big[\|\widehat{\mu}-\mu\|^2\Big]=d\,\frac{\sigma^2}{N}\,. \tag{II.1}$$

On constate l'influence de la dimension sur l'erreur d'estimation : l'erreur augmente avec la dimension de l'espace. La moyenne empirique $\bar{X}_N=\frac{1}{N}\sum_{k=1}^N X_k$ atteint exactement cette erreur et est donc un estimateur minimax de μ sur cet ensemble de distributions. Cependant Stein (1956) montre que la moyenne empirique est inadmissible, plus précisément qu'il existe des estimateurs strictement meilleurs au sens du risque quadratique. Un meilleur estimateur est par exemple l'estimateur de James-Stein (James et Stein, 1961) qui contracte la moyenne empirique vers un point de référence, traditionnellement 0. Nous considérerons plutôt ici sa version avec partie positive, définie par

$$\mu^{\mathbf{JS}+}=\left(1-\frac{\sigma^2}{N}\frac{d-2}{\|\bar{X}_N\|^2}\right)_+\bar{X}_N, \tag{II.2}$$

et qui a une erreur quadratique pour l'estimation de μ strictement meilleure que la moyenne empirique (Baranchik, 1964). Pour une dimension $d\geq 2$:

$$\mathbb{E}\Big[\|\mu^{\mathbf{JS}+}-\mu\|^2\Big]\leq d\,\frac{\sigma^2}{N}\min\left(\frac{\tau}{1+\tau}+\frac{4}{d},1\right)\quad\text{où}\quad\tau=\tau(\mu)=\frac{N\|\mu\|^2}{d\sigma^2}\,, \tag{II.3}$$

(voir par exemple Lemme 3.8 de Tsybakov, 2008 pour cette borne). L'amélioration par rapport à la moyenne empirique est plus importante en grande dimension, c'est-à-dire lorsque $d\gg N$. En effet, dans ce cas, la variance de la moyenne empirique (Eq.(II.1)) devient très mauvaise et la contraction vers 0 en devient plus efficace. La variance est réduite en ajoutant un biais symbolisé par le facteur τ. À $\|\mu\|^2$ fixé, le facteur τ décroît avec la dimension et augmente avec la taille de l'échantillon. À l'inverse, l'amélioration est plus faible lorsque N grandit : on sort alors du cadre de la grande dimension et la contraction perd de son intérêt. Cependant dans tous les cas, l'estimateur de James-Stein reste strictement meilleur.

L'estimateur de James-Stein est minimax sur la classe de distributions $\mathcal{G}_d$ mais l'est aussi sur le sous ensemble des distributions gaussiennes de moyennes proches de 0. Soit $\tau>0$, posons

$$\mathcal{P}_d(\tau)=\{\mathcal{N}(\mu,\sigma^2 I_d)^{\otimes N}:\|\mu\|^2\leq\tau d\,\sigma_N^2\},$$

alors Pinsker (1980) montre que le risque minimax sur cette classe vérifie

$$\lim_{d\to\infty} \inf_{\widehat{\mu}} \sup_{\mathbb{P}\in\mathcal{P}_d(\tau)} \frac{\mathbb{E}\left[\|\widehat{\mu}-\mu\|^2\right]}{d\sigma_N^2} = \frac{\tau}{1+\tau}\,, \quad \text{où} \quad \sigma_N^2 = \frac{\sigma^2}{N}. \tag{II.4}$$

L'estimateur $\mu^{\mathbf{JS}+}$ atteint la borne minimax asymptotiquement en la dimension et évidemment sans connaître τ. Dans ce modèle, l'erreur d'estimation possible est bien toujours en $O(N^{-1})$ mais le gain se fait au niveau de la dépendance en la dimension. L'estimateur de James-Stein s'adapte à de nombreux contextes par exemple à une covariance non isotrope (Bock, 1975), des notions de risque différentes du risque quadratique (Berger, 1976) et peut se construire avec une contraction différente et en estimant la variance (Baranchik, 1970 ; Lehmann et Casella, 2006). Plus récemment, Muandet et al. (2014) l'adaptent pour estimer des Kernel Mean Embeddings de distributions (voir Section II.2).

Nous présentons ci-dessous différentes interprétations de l'estimateur de James-Stein et comment des méthodes ou idées classiques mènent à le considérer. Notre objectif est de présenter au lecteur ou à la lectrice différentes idées applicables à l'estimation du vecteur μ mais que nous utiliserons par la suite dans des cadres plus larges.

Interprétation oracle : L'estimateur de James-Stein peut être vu comme issu de l'estimation du poids du meilleur estimateur de contraction vers 0. Considérons l'estimateur $\widehat{\mu}_\omega = \omega\bar{X}_N$ où $\omega \in [0,1]$ et cherchons l'estimateur de cette forme minimisant le risque quadratique :

$$\min_{\omega\in[0,1]} \mathbb{E}\left[\|\widehat{\mu}_\omega - \mu\|^2\right] = \min_{\omega\in[0,1]} \left[\omega^2 d\sigma_N^2 + (1-\omega)^2\|\mu\|^2\right] = d\sigma_N^2 \frac{\tau(\mu)}{1+\tau(\mu)}$$

Le poids optimal est $\omega^* = 1 - \frac{d\sigma_N^2}{\|\mu\|^2 + d\sigma_N^2}$. En connaissant la norme de μ on peut améliorer son estimation ce qui semble assez naturel. Celle-ci étant inconnue l'estimateur de James-Stein estime directement $\|\mu\|^2 + d\sigma_N^2$ par $\|\bar{X}_N\|^2$ ce qui amène à considérer (II.3) après avoir injecté cet estimateur dans la formule du poids oracle. L'estimateur de James-Stein va garder des performances proches de celles de l'estimateur oracle grâce au fait que l'estimation d'une distance (quantité unidimensionnelle) en grande dimension est bien plus facile que celle d'un vecteur. L'erreur induite va être négligeable relativement au gain.

Interprétation test : Le choix de la contraction de l'estimateur de James-Stein peut être relié au problème de test :

$$(H_0) : \mu = 0\,, \qquad (H_1) : \mu \neq 0\,.$$

La statistique $P = d\sigma_N^2/\|\bar{X}_N\|^2$ est super-uniforme sous (H_0) (trivialement par l'inégalité de Markov car $\mathbb{E}\left[P^{-1}\right] = 1$) et peut être utilisée pour tester ces deux hypothèses. Comme sous (H_0), le vecteur μ devrait être estimé par 0, l'estimateur de James-Stein utilise la statistique de test P pour quantifier la contraction vers 0 :

$$\widehat{\mu}^{\mathbf{JS}+} = \left(1 - \frac{d-2}{d}P\right)_+ \widehat{\mu}^{\mathtt{NE}}.$$

Le test permet de construire l'estimateur. Cette vision se retrouve dans Casella (1985) qui considère une contraction de chaque coordonnée vers la moyenne des coordonnées et relie cet estimateur de type James-Stein au problème de tester si les coordonnées de μ sont toutes égales.

Interprétation régularisation : Gruber (1998) lie les estimateurs de type James-Stein et les estimateurs de type ridge qui sont d'autres estimateurs de contraction. Ainsi l'estimateur de James-Stein $\widehat{\mu}^{\mathbf{JS}} = \left(1 - \sigma_N^2 \frac{d-2}{\|\bar{X}_N\|^2}\right)\bar{X}_N$, (Eq.(II.3) sans la partie positive), est solution du problème de régression

ridge

$$\widehat{\mu}^{\mathbf{JS}} = \underset{\nu\in\mathbb{R}^d}{\operatorname{Arg\,Min}} \frac{1}{N}\sum_{i=1}^{N}\|X_i-\nu\|^2 + \lambda\|\nu\|^2$$

pour $\lambda = \frac{(d-2)\sigma_N^2}{\|\bar{X}_N\|^2-(d-2)\sigma_N^2}$. Pour éviter le choix de λ, on peut aussi envisager une régularisation par la norme. On obtient alors un estimateur de type James-Stein :

$$\left(1-\frac{\widehat{\sigma}}{\sqrt{N}}\frac{\sqrt{d}}{\|\bar{X}_N\|}\right)_+ \bar{X}_N = \underset{\nu\in\mathbb{R}^d}{\operatorname{Arg\,Min}} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|X_i-\nu\|^2} + \frac{1}{\sqrt{N}}\|\nu\|,$$

où $\widehat{\sigma}^2 = \frac{1}{d(N-1)}\sum_{i=1}^{N}\left\|X_i-\bar{X}_N\right\|^2$ est un estimateur de σ^2. Cet estimateur bat la moyenne empirique pour des moyennes proche de 0, $(\tau(\mu)<1)$ mais n'est pas minimax. Cependant cet exemple illustre qu'avec une simple pénalisation par la norme, sans paramètre supplémentaire, on retrouve une contraction proche de celle de l'estimateur de James-Stein ainsi que la présence de la partie positive. On peut aussi remarquer que cette régularisation estime naturellement la variance par la variance empirique.

Interprétation bayésienne : Les travaux de Efron et Morris (1972, 1973, 1976) interprètent l'estimateur de James-Stein comme un problème de Bayes empirique. Par exemple en supposant que chacune des coordonnées de μ est tirée indépendamment selon une même loi normale $\mu_i \sim \mathbb{Q} = \mathcal{N}(0,\tau\sigma^2)$, l'estimateur de Bayes du vecteur μ est alors un estimateur de contraction $\omega\bar{X}_N$ avec $\omega = \frac{1}{1+\tau}$. Marginalement $X_i \sim \mathcal{N}(0,(1+\tau)\sigma^2 I_d)$, donc on peut ainsi estimer $1+\tau$ par $\|\bar{X}_N\|^2/(d\sigma_N^2)$ ce qui mène encore à l'estimateur de James-Stein. Plus récemment Brown et Greenshtein (2009), ont reconsidéré cette approche pour une loi $\mathbb{Q}$ arbitraire. Une plus ample discussion est donnée en Section V.7.1.

Interprétation tâches multiples : L'estimation de chacune des coordonnées du vecteur μ peut être considérée comme un problème de tâches multiples (Baxter, 1997 ; Caruana, 1997). L'apprentissage de multiples tâches cherche à résoudre différents problèmes simultanément (régression, estimation, ...) pour des jeux de données de différentes distributions. Pour cela, l'approche tâche multiple utilise des similarités entre les distributions (structure commune, bruit identique,...). Ici nos tâches seraient d'estimer chacune des coordonnées en minimisant le risque composé (ou *compound risk*), c'est-à-dire la moyenne des erreurs de chaque estimateur. Ceci est bien équivalent à minimiser l'erreur quadratique du vecteur des estimateurs. L'estimateur de James-Stein utilise finalement que les données de toutes les tâches ont le même bruit σ^2 pour l'estimer efficacement et construire un estimateur de contraction pour chaque coordonnée. Le problème d'estimer différentes moyennes est connu sous le nom de *multi-task averaging*. Le lien de ce problème avec l'estimateur de James-Stein est considéré par exemple par Feldman et al. (2014) ou Duan et Wang (2023).

Ces deux dernières interprétations voient l'estimateur de James-Stein comme l'estimation jointe de quantités réelles (les coordonnées). Une question naturelle à se poser est s'il est possible de l'adapter pour estimer simultanément différents vecteurs à partir d'échantillons bruités de chacun. Ce problème se ramène au cadre de James-Stein si l'on suppose que les bruits de chaque échantillon sont gaussiens isotropes. Cependant si les bruits sont différents, inconnus, non isotropes ou même non gaussiens, il n'est pas évident de construire un estimateur de type James-Stein et bien que des réponses partielles se trouvent dans la littérature, le problème n'a pas été considéré dans son ensemble. Parmi les questions soulevées, on peut se demander comment vont interagir les différents degrés de libertés du problème i.e. le nombre de données, leur dimension et le nombre de vecteurs à estimer. Nous considèrerons ce problème dans les Sections III et Section V.

II.1.2 Distance de séparation de vecteurs

L'influence de la dimension dans les performances d'un test a été mise en valeur par Baraud (2002) dans son analyse non asymptotique de la vitesse de séparation de tests dans le cadre gaussien. La distance ou vitesse de séparation d'un test définie par Ingster (1982) dans le cadre asymptotique et adaptée au cadre non asymptotique par Baraud (2002) permet une analyse minimax des problèmes de test. Dans un cadre général, pour une distance γ entre les distributions et un ensemble d'hypothèses $\mathcal{H}_0$ et d'alternatives $\mathcal{H}_1$, nous définissons la distance de séparation des ensembles d'hypothèses pour $\alpha \in (0,1)$ comme :

$$\delta^*_\alpha = \inf\left\{ \delta \geq 0 \middle| \exists \text{ test } T \ : \sup_{\mathbb{P}\in\mathcal{H}_0} \mathbb{P}(T=1) + \sup_{\mathbb{P}\in\mathcal{H}_1 : \gamma(\mathbb{P},\mathcal{H}_0)\geq\delta} \mathbb{P}(T=0) \leq \alpha \right\}. \tag{II.5}$$

Il est important de remarquer que la distance de séparation dépend fortement de la distance γ choisie entre les distributions. Intuitivement la distance de séparation optimale est la distance minimale entre les hypothèses et les alternatives pour laquelle un test existe dont la somme des erreurs de type I et II est contrôlée par α. Cette notion est équivalente à la complexité d'échantillon (sample complexity) du problème. Pour un test la complexité d'échantillon est le nombre minimal de données pour lesquelles la somme des erreurs de type I et II sont contrôlées par α pour des alternatives à une distance δ fixée. Ces notions se déduisent l'une de l'autre par des opérations élémentaires.

Remarque II.1. La définition (II.5) est différente des originales, Baraud (2002) et Ingster (1982) considèrent plutôt la distance minimale des alternatives aux hypothèses pour lesquelles il existe un test d'erreur de type I exactement α et d'erreur de type II contrôlée. Les deux définitions cohabitent cependant dans la littérature.

Considérons le problème de détection de signal avec bruit gaussien. Soit $(X_i)_{1\leq i\leq n}$ un échantillon de vecteurs gaussiens de loi $\mathcal{N}(\mu, \sigma^2 I_d)$ et le problème de test :

$$(H_0(\eta)) : \|\mu\| \leq \eta, \text{ contre } (H_1(\eta)) : \|\mu\| > \eta, \tag{II.6}$$

où $\eta \geq 0$. Dans le cas classique où $\eta = 0$ nous nous référons à Baraud (2002) et dans le cas $\eta > 0$, connu sous le nom de test d'hypothèses pertinentes ou précises, à Blanchard et al. (2018). Ces travaux se distinguent par leurs analyses non asymptotiques du rôle de la dimension. En effet dans le cadre non paramétrique, les analyses se concentrent plutôt sur la dépendance de la vitesse en la taille de l'échantillon et l'influence de la régularité. Pour une analyse détaillée, les lecteurs peuvent se référer à Ingster et Suslina (1998).

Dans ce modèle de distributions gaussiennes à variances fixées, nous pouvons choisir comme distance γ entre les distributions la distance euclidienne entre les vecteurs moyennes $\gamma\Big(\mathcal{N}(\mu,\sigma^2 Id), \mathcal{N}(\nu,\sigma^2 I_d)\Big) = \|\mu - \nu\|$. Avec cette distance γ, la distance de séparation optimale pour le problème de test (II.6) est la distance minimale δ^* pour laquelle un test est capable de différencier les distributions de moyennes dans la boule de rayon η et celles en dehors de la boule de rayon $\eta + \delta^*$. Pour $\eta = 0$, Baraud (2002) analyse la dépendance en la dimension de cette vitesse de séparation et donne la vitesse :

$$\delta^*_\alpha(\eta = 0) = \Theta_\alpha\Big(d^{1/4}\sigma_N\Big), \tag{II.7}$$

où $\sigma_N^2 = \sigma^2/N$ et Θ_α indique des bornes inférieures et supérieures dépendant seulement de α. Cette analyse non asymptotique met en lumière la relative facilité du test par rapport à l'estimation. L'erreur minimale de détection est de $\sigma_N d^{1/4}$ pour le test contre $\sigma_N\sqrt{d}$ (Eq.(II.1)) pour l'estimation. Autrement dit un test est capable d'assurer que le vecteur μ est proche à $\sigma_N d^{1/4}$ d'un point de référence (ici 0) alors qu'un estimateur de μ n'est assuré que d'être à une distance $\sigma_N\sqrt{d}$ du vrai vecteur μ.

Pour η non nul, Blanchard et al. (2018) ont mis en évidence l'existence de deux régimes. Lorsque η est petit, l'erreur de test est bien (II.7), alors que pour η grand le test devient plus facile et l'erreur perd sa dépendance en la dimension. Plus précisément :

$$\delta_\alpha^*(\eta) = \Theta_\alpha\left(\sigma_N \max\left(1, \min\left(d^{1/4}, \sqrt{d}\frac{\sigma_N}{\eta}\right)\right)\right). \tag{II.8}$$

Intuitivement, lorsque η grandit, une direction devient prépondérante et le problème se ramène à un problème unidimensionnel. Plus généralement, tester si $\mu \in \mathcal{C}$ où $\mathcal{C}$ est un convexe est toujours plus facile que d'estimer μ. En notant $\delta^*(\mathcal{C})$ la distance de séparation du test pour le convexe $\mathcal{C}$, on a

$$\delta_\alpha^*(\mathcal{C}) = O_\alpha\left(\inf_{\widehat{\mu}} \sup_{\mathbb{P}\in\mathcal{H}_0\cup\mathcal{H}_1} \mathbb{E}[\|\widehat{\mu} - \mu\|]\right) = O_\alpha\left(\sigma_N\sqrt{d}\right). \tag{II.9}$$

La première inégalité étant vraie en toute généralité (en utilisant comme statistique de test la distance d'un estimateur $\widehat{\mu}$ au convexe $\mathcal{C}$) et la seconde dans notre cadre de distributions gaussiennes isotropes. Ce cas limite est atteint pour $\mathcal{C}$ un orthant ($\mathcal{C} = [-\infty, 0]^d$) et dans ce cas l'erreur de test est la même que l'erreur d'estimation (Théorème 3.6. de Blanchard et al., 2018 ou Juditsky et Nemirovski, 2002 dans un cadre non paramétrique).

L'apparition de la dimension dans la vitesse de séparation ou l'erreur d'estimation est très liée au modèle gaussien isotrope en dimension finie. Cependant certains outils modernes sortent de ce paradigme. C'est le cas par exemple du Kernel Mean Embedding (KME), vecteur d'un espace fonctionnel qui va permettre de caractériser des distributions. Nous présentons donc dans la prochaine section ce qu'est un KME, son intérêt en machine learning et ses liens avec les problèmes de tests et d'estimation en grande dimension.

II.2 Un outil de grande dimension : le Kernel Mean Embedding

Le *Kernel Mean Embedding* (KME) est un outil de machine learning introduit par Smola et al. (2007) et intrinsèquement lié aux méthodes à noyaux et aux espaces de Hilbert à noyaux reproduisants (Reproducing Kernel Hilbert Space - RKHS). Le principe des méthodes à noyaux (Aizerman, 1964) est d'injecter les données étudiées dans un espace de plus grande dimension et d'appliquer ensuite un algorithme classique (typiquement une méthode linéaire régularisée). La force de ces méthodes est que le changement d'espace se traduit seulement par le remplacement du produit scalaire euclidien par le produit scalaire de l'espace de Hilbert défini par un noyau κ. Une méthode qui n'utilise que des produits scalaires sur les données s'adapte ainsi très facilement, c'est le cas de nombreuses méthodes comme le support vector machine (Boser et al., 1992 ; Cortes et Vapnik, 1995), la régression ridge (Cristianini et Shawe-Taylor, 2000 ; Hoerl et Kennard, 1970 ; Saunders et al., 1998) ou l'analyse par composantes principales (Hotelling, 1933 ; Pearson, 1901 ; Schölkopf et al., 1998).

À une distribution $\mathbb{Q}$ sur l'espace initial on associe une distribution $\mathbb{P}$ sur le RKHS qui est la mesure image, ou push-forward, de $\mathbb{Q}$ via l'injection dans le RKHS. Le KME de la distribution peut simplement se définir comme l'espérance de la distribution $\mathbb{P}$ et, sous de faibles hypothèses, est un vecteur du RKHS. Si le noyau est bien choisi, le KME caractérise totalement la distribution et entre autre induit une distance facile à calculer entre des distributions appelée *Maximum Mean Discrepancy* (MMD) (Borgwardt et al., 2006). Cette distance est simplement la distance entre les KMEs des distributions. Cette propriété ouvre la voie à de nombreuses applications comme par exemple du two sample test (Gretton et al., 2012), goodness-of-fit test (Chwialkowski et al., 2016), de l'apprentissage multi-instance ou de distributions à la fois supervisé (Muandet et al., 2012 ; Szabó et al., 2016) aussi bien que non supervisé (Jegelka et al., 2009). On peut aussi relever l'utilisation du MMD pour la construction de modèles génératifs (Dziugaite et al., 2015 ; Li et al., 2017 ; Li et al., 2015).

Dans cette section nous présenterons la construction du KME d'une distribution puis son estimation et son utilisation dans les tests à deux échantillons. Un panorama plus complet peut être trouvé dans Muandet et al. (2017).

II.2.1 Construction du KME et de la distance MMD

Les méthodes à noyaux sont construites à partir de fonctions noyaux. Notons $\mathcal{X}$ l'espace de nos données, une fonction $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est un noyau défini positif si elle est symétrique ($\kappa(x, y) = \kappa(y, x)$) et si pour tout entier n, tous poids $a_1, \dots, a_n \in \mathbb{R}$ et tous points $x_1, \dots, x_n \in \mathcal{X}$:

$$\sum_{i,j=1}^{n} a_i a_j \kappa(x_i, x_j) \geq 0 \,.$$

Cette propriété induit en particulier que la diagonale de κ est positive ($\kappa(x, x) \geq 0$ pour tout point x). Les fonctions noyaux sont stables par somme, dilatation, multiplication et passage à la limite ce qui les rend assez faciles à construire. Les noyaux suivants sont les plus classiques :

- $\kappa(x, y) = \mathbf{1}_{x=y}$ définit le noyau trivial ;
- si $\mathcal{X} \subset \mathbb{R}^d$, alors $\exp\left(-\frac{\|x-y\|_2^2}{h^2}\right)$ et $\exp\left(-\frac{\|x-y\|_2}{h}\right)$ où $h > 0$ définissent respectivement les noyaux gaussien et de Laplace ;
- si $\phi : \mathcal{X} \to \mathcal{H}$ est une injection de $\mathcal{X}$ dans un espace de Hilbert $\mathcal{H}$, le produit scalaire $\langle \phi(\cdot), \phi(\cdot) \rangle_{\mathcal{H}}$ définit un noyau défini positif. Si $\mathcal{X}$ est un espace de Hilbert, son produit scalaire en est donc un.

Des fonctions noyaux sont construites pour des données très variées comme des textes (Joulin et al., 2017), des suites, arbres ou graphes (voir Gärtner, 2003 ou Shawe-Taylor et Cristianini, 2004 pour un panorama) utilisés dans ces cas en particulier en bioinformatique (Gusfield, 1997), pour des images (Zhang et al., 2007) et aussi pour des diagrammes de persistance en analyse topologique de données (Carriere et al., 2017).

L'objectif des noyaux est de définir un produit scalaire dans un espace plus grand : à partir d'un noyau défini positif, il est en effet possible de construire un espace de Hilbert pour lequel ce noyau définit un produit scalaire.

Proposition II.2. *Soit $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ un noyau défini positif. Soit l'espace vectoriel*

$$\mathcal{H}_0 = \mathit{Vect}(\kappa(x, \cdot) : x \in \mathcal{X}) \subset \mathbb{R}^{\mathcal{X}} \,,$$

muni du produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ défini par

$$\langle \kappa(x, \cdot), \kappa(y, \cdot) \rangle_{\mathcal{H}} := \kappa(x, y) \quad \textit{pour} \quad x, y \in \mathcal{X} \tag{II.10}$$

et étendu par linéarité. Alors $\mathcal{H}$ le complété de $\mathcal{H}_0$ est un espace de Hilbert pour le produit scalaire (II.10), étendu par passage à la limite. En particulier κ est un noyau reproduisant pour $\mathcal{H}$:

$$\langle h, \kappa(x, \cdot) \rangle_{\mathcal{H}} = h(x), \quad \forall\, h \in \mathcal{H},\ x \in \mathcal{X}, \quad \textit{(propriété reproduisante).}$$

Le noyau induit une injection ϕ entre l'espace des données $\mathcal{X}$ et le RKHS $\mathcal{H}$ définie pour $x \in \mathcal{X}$ par

$$\phi(x) = \kappa(x, \cdot) \,, \quad \text{et pour } y \in \mathcal{X} \quad \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = \kappa(x, y) \,.$$

Comme annoncé, le noyau κ permet de calculer directement les produits scalaires entre les représentants de deux données dans le RKHS. D'autre part, $\mathcal{H}$ est l'unique RKHS dont κ est le noyau reproduisant (Théorème de Moore-Aronszajn, Aronszajn, 1950). Cependant pour un noyau donné, il est possible de construire plusieurs espaces de Hilbert pour lesquels ce noyau définit bien un produit scalaire mais n'est pas

reproduisant. Par exemple, le noyau donné par $\kappa(x, y) = xy$ pour $x, y \in \mathbb{R}$ définit un produit scalaire sur n'importe quelle droite de $\mathbb{R}^d$ pour une dimension d arbitraire ($\mathcal{H} = \{x\nu, x \in \mathbb{R}\}$ pour un certain $\nu \in \mathbb{R}^d$).

Le noyau associe aux points de $\mathcal{X}$ des images dans le RKHS par l'injection ϕ qui se généralise aux distributions sur $\mathcal{X}$ par le KME.

Définition II.3 (KME)**.** Soit $\mathbb{Q}$ une distribution sur $\mathcal{X}$ et κ un noyau défini positif, le *Kernel Mean Embedding* (KME) de la distribution $\mathbb{Q}$ dans le RKHS $\mathcal{H}$ associé à κ est

$$\mu_{\mathbb{Q}} = \mathbb{E}_{X\sim\mathbb{Q}}[\kappa(X, \cdot)]\,. \tag{II.11}$$

Si $\mathbb{E}_{X\sim\mathbb{Q}}\left[\sqrt{\kappa(X, X)}\right] < \infty$ alors $\mathbb{Q}$ est une mesure intégrable au sens de Bochner dans $\mathcal{H}$ et $\mu_{\mathbb{Q}} \in \mathcal{H}$ est bien défini.

La condition d'existence du KME est facilement vérifiée en considérant un noyau borné (par exemple le noyau trivial, gaussien, de Laplace ...). Grâce à la propriété reproduisante du noyau, pour toute fonction h de $\mathcal{H}$, on a $\mathbb{E}_{X\sim\mathbb{Q}}[h(X)] = \langle h, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}}$. Cette propriété va être en particulier très utile pour estimer des distances entre KME. Une question importante a été de savoir pour quels noyaux cette distance entre KME induit bien une distance entre les distributions, c'est à dire pour quels noyaux la fonction $\mathbb{Q} \mapsto \mu_{\mathbb{Q}}$ est bien injective. Les noyaux vérifiant cette propriété sont dits *caractéristiques*. Intuitivement il faut que la classe de fonction du RKHS $\mathcal{H}$ soit assez riche pour que le KME caractérise la distribution. Par exemple, pour un espace $\mathcal{X}$ compact, le noyau est caractéristique si le RKHS $\mathcal{H}$ est dense dans les fonctions continues bornées (Steinwart, 2001). Voici quelques autres exemples de noyaux caractéristiques pour différents espaces.

- Le noyau trivial $\mathbf{1}_{x=y}$ est caractéristique lorsque $\mathcal{X}$ est fini (Borgwardt et al., 2006).
- Le noyau exponentiel $\kappa(x, y) = \exp(\langle x, y\rangle)$ est caractéristique pour $\mathcal{X}$ compact de $\mathbb{R}^d$. Le KME est alors la fonction génératrice des moments $\mu_{\mathbb{Q}}(x) = \mathbb{E}_{X\sim\mathbb{Q}}[\exp(\langle x, X\rangle)]\,.$
- Les noyaux gaussien et de Laplace sont caractéristiques sur $\mathbb{R}^d$ (Fukumizu et al., 2007).
- Plus généralement pour un noyau invariant par translation sur $\mathbb{R}^d$ ($\kappa(x, y) = K(x - y)$), le KME se relie à la fonction caractéristique de la distribution. Il est alors caractéristique si sa transformée de Fourier est à support égal à tout $\mathbb{R}^d$ (Sriperumbudur et al., 2011, 2008, 2010).

Pour un noyau caractéristique, la distance entre les KME de deux distributions induit donc une distance entre les distributions. Cette distance est connue sous le nom de *maximum mean discrepancy* (MMD) (Gretton et al., 2012).

Définition II.4 (MMD)**.** Soit $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ un noyau caractéristique et $\mathcal{H}$ son RKHS associé, la distance MMD entre $\mathbb{P}$ et $\mathbb{Q}$ distributions de $\mathcal{X}$ est définie par :

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}\,.$$

où $\mu_{\mathbb{P}}$ et $\mu_{\mathbb{Q}}$ sont les KMEs de $\mathbb{P}$ et $\mathbb{Q}$.

Cette distance peut se voir comme une métrique intégrable de probabilités (Müller, 1997) sur le RKHS $\mathcal{H}$. En effet :

$$\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} = \sup_{h\in\mathcal{H}:\|h\|_{\mathcal{H}}\leq 1} \langle h, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\rangle = \sup_{h\in\mathcal{H}:\|h\|_{\mathcal{H}}\leq 1} \left(\int_{\mathcal{X}} h d\mathbb{P} - \int_{\mathcal{X}} h d\mathbb{Q}\right).$$

Ainsi si $\mathcal{H}$ contient les fonctions bornées, elle domine la distance en variation totale, et pour $\mathcal{X} = \mathbb{R}$, si elle contient les fonctions indicatrices $\left\{\mathbf{1}_{(-\infty,t)}\right\}_{t\in\mathbb{R}}$ elle domine la distance de Kolmogorov. La distance MMD est équivalente à la distance énergie (Sejdinovic et al., 2013) et se relie aux distances de transport optimal : elle est la limite de la divergence Sinkhorn (Genevay et al., 2018). Par rapport à ces distances, une force de la distance MMD et du KME en général est la relative facilité de leurs estimations.

II.2.2 Estimation

Soient $\mathbb{Q}$ et $\mathbb{P}$ deux distributions sur $\mathcal{X}$ connues seulement via deux échantillons $\{X_i\}_{1\leq i\leq N}$ et $\{Y_j\}_{1\leq j\leq M}$ des distributions $\mathbb{Q}$ et $\mathbb{P}$ respectivement. Nous présentons ici des estimateurs classiques du KME $\mu_{\mathbb{Q}}$ et de la distance MMD entre $\mathbb{Q}$ et $\mathbb{P}$.

Le KME de la distribution $\mathbb{Q}$ peut se voir comme l'espérance du vecteur aléatoire $Z_i = \kappa(X_i, \cdot)$ (où $X_i \sim \mathbb{Q}$) dans l'espace de Hilbert $\mathcal{H}$. Ainsi le KME $\mu_{\mathbb{Q}} = \mathbb{E}[Z_1]$ peut s'estimer par la moyenne empirique classique :

$$\widehat{\mu}_{\mathbb{Q}}(\cdot) = \frac{1}{N}\sum_{i=1}^{N} \kappa(X_i, \cdot) = \frac{1}{N}\sum_{i=1}^{N} Z_i \,.$$

Lorsque le noyau κ est borné, le vecteur aléatoire Z est alors lui aussi borné dans $\mathcal{H}$ et il est possible d'utiliser les outils de concentration liés aux variables aléatoires bornées pour contrôler les déviations. Ainsi en utilisant l'inégalité de McDiarmid et al. (1989), pour tout $u \geq 0$, avec probabilité $1 - e^{-u}$:

$$\|\widehat{\mu}_{\mathbb{Q}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \leq \frac{L}{\sqrt{N}}\Big(1 + \sqrt{2u}\Big), \tag{II.12}$$

où $L^2 \geq \sup_{x\in\mathcal{X}} \kappa(x, x)$ est une borne sur la diagonale du noyau. Une telle hypothèse est vérifiée par les noyaux usuels (trivial-Gaussien-de Laplace). On peut remarquer d'ailleurs qu'un noyau reproduisant borné sur la diagonale est alors borné partout, en effet pour tout $x, y \in \mathcal{X}$

$$|\kappa(x, y)| = |\langle \kappa(x, \cdot), \kappa(y, \cdot)\rangle_{\mathcal{H}}| \leq \|\kappa(x, \cdot)\|_{\mathcal{H}}\|\kappa(y, \cdot)\|_{\mathcal{H}} = \sqrt{\kappa(x, x)\kappa(y, y)} \,.$$

L'estimation de KMEs de distributions $\mathbb{Q}$ et $\mathbb{P}$ peut permettre d'estimer leur distance MMD en calculant directement leur distance dans le RKHS (bien qu'il y ait mieux comme estimateur, voir ci-dessous). Cependant l'utilisation du KME ne se résume pas à la définition de la distance MMD. Son estimation est par exemple nécessaire pour des problèmes de *distribution régression* où l'on cherche à faire une prédiction à partir d'un échantillon (voir par exemple Oliva et al., 2013 ou Szabó et al., 2016). En inférence causale, le KME de lois conditionnelles est utilisé comme proxy avant de procéder à une régression (Mastouri et al., 2021 ; Singh et al., 2019). Dans ces cas l'estimation complète du KME est nécessaire.

La distance MMD entre deux distributions $\mathbb{Q}$ et $\mathbb{P}$ peut s'estimer sans directement estimer les KMEs respectifs. Un estimateur non biaisé (de la distance au carré) le plus classique se construit à l'aide de U-statistiques :

$$\widehat{\mathrm{MMD}}^2(\mathbb{Q}, \mathbb{P}) := \frac{1}{N(N-1)}\sum_{i\neq j=1}^{N} \kappa(X_i, X_j) - \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \kappa(X_i, Y_j) + \frac{1}{M(M-1)}\sum_{i\neq j=1}^{N} \kappa(Y_i, Y_j) \,. \tag{II.13}$$

Cet estimateur se considère naturellement après avoir remarqué que $\|\mu_{\mathbb{Q}}\|^2_{\mathcal{H}} = \mathbb{E}[\langle Z, Z'\rangle_{\mathcal{H}}] = \mathbb{E}[\kappa(X, X')]$ où X, X' sont indépendants de loi $\mathbb{Q}$. Chacun des termes de (II.13) estime sans biais chacun des termes du développement de la distance $\|\mu_{\mathbb{Q}} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}$. Les déviations de cet estimateur de la distance peuvent se contrôler à l'aide des inégalités de concentration sur les U-statistiques de Hoeffding (1963) (voir aussi Gretton et al., 2012 dans le cadre KME). Pour tout $u \geq 0$, avec probabilité $1 - e^{-u}$:

$$\left|\widehat{\mathrm{MMD}}^2(\mathbb{Q}, \mathbb{P}) - \|\mu_{\mathbb{Q}} - \mu_{\mathbb{P}}\|^2\right| \leq \frac{L^2}{\sqrt{\min(N, M)}}\sqrt{8u} \,. \tag{II.14}$$

Ces estimateurs du KME et de la distance MMD sont optimaux en la taille de l'échantillon (Tolstikhin et al., 2017 ; Tolstikhin et al., 2016). Cependant, dans les deux cas, les vitesses données par les bornes de concentration (II.12) et (II.14) ne prennent pas en compte la structure de covariance de la distribution ou

la dimension de l'espace. Les termes liés à ces paramètres sont en réalité bornés par la borne L sur le noyau. Pour pouvoir prendre en compte leurs effets nous aurons besoin d'inégalités de concentration plus précises, de type Bernstein par exemple. Plus généralement, l'analyse de l'estimation de vecteurs ou de distance en grande dimension se fait plutôt sous l'hypothèse que les distributions sont sous-gaussiennes (Hsu et al., 2012 ; Koltchinskii et Lounici, 2017 par exemple). Ce cadre, bien qu'englobant des distributions bornées, ne permet cependant pas de capturer l'influence de ces paramètres sur les déviations par une application directe de résultats existants (voir discussion en Section IV.2.4). Pour les capturer, tout au long de la thèse, en ayant comme objectif de construire des procédures adaptées à des KMEs, nous considèrerons le cadre de données bornées dans un espace de Hilbert. Des phénomènes en estimation et en test comparables à ceux présentés précédemment en Section II.1 seront constatés.

II.2.3 Tests à deux échantillons

Le test à deux échantillons consiste à tester l'égalité de deux distributions à partir d'échantillons de chacune d'entre elles. Formellement pour $\mathbb{P}$ et $\mathbb{Q}$, deux distributions sur l'espace $\mathcal{X}$, on cherche à tester

$$(H_0) : \mathbb{P} = \mathbb{Q}\,, \qquad \text{contre} \qquad (H_1) : \mathbb{P} \neq \mathbb{Q}\,. \tag{II.15}$$

à partir de deux échantillons de chacune des lois. En une dimension, les tests historiques pour ce problème sont les tests du Khi-deux dans le cadre discret (Pearson, 1900) et le test de Kolmogorov-Smirnov dans le cadre continu (Kolmogorov, 1933). Ces tests sont construits respectivement sur des estimateurs empiriques de la divergence du Khi-deux et de la distance de Kolmogorov entre $\mathbb{P}$ et $\mathbb{Q}$. En grande dimension, le test de Kolmogorov-Smirnov se généralise (Bickel, 1969 ; Friedman et Rafsky, 1979) mais a le désavantage d'avoir un important coût algorithmique. Au lieu de considérer ces divergences, le test à noyau proposé par Gretton et al. (2012) choisit de comparer les distributions à l'aide la distance MMD entre $\mathbb{P}$ et $\mathbb{Q}$. Dans ce cadre le test (II.15) se réécrit comme un test d'égalité de vecteurs dans le RKHS :

$$(H_0) : \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} = 0\,, \qquad \text{contre} \qquad (H_1) : \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \neq 0\,. \tag{II.16}$$

Grâce à sa flexibilité (adaptable au contexte par le choix du noyau) et la simplicité de l'estimation de la distance MMD (voir Section II.2.2), ce test a été largement diffusé. La question de son optimalité au sens de la distance de séparation (II.5) est encore étudiée aujourd'hui. Le test original, construit à partir de la U-statistique (II.13), est optimal pour des distributions à densités höldériennes sur $\mathbb{R}^d$ et pour la distance L^2 entre les densités. Cependant, le noyau choisi pour construire un tel test minimax doit dépendre du paramètre de régularité de ces densités et en toute généralité le test est sous optimal (Balasubramanian et al., 2021). Schrab et al. (2023) construisent cependant une version minimax et adaptative de ce test sur $\mathbb{R}^d$ en aggrégeant une famille de tests basés sur plusieurs noyaux. Pour avoir un test minimax sur des espaces différents de $\mathbb{R}^d$, Hagrass et al. (2022) utilisent une régularisation de la distance MMD par la covariance avec un noyau adapté aux données. La distance de séparation de ce test est alors mesurée en terme de distance de Hellinger.

Dans les analyses citées, la distance de séparation du test est évaluée à l'aide de distances entre les distributions (distance L^2 entre les densités, Hellinger, divergence du Khi-deux, ...). Pour relier ces distances à la distance MMD, des hypothèses de régularité sur les densités sont nécessaires et les vitesses optimales en la taille de l'échantillon dépendent alors de la dimension. Par exemple Li et Yuan (2019) montrent que la distance de séparation pour la norme L^2 entre les densités est en $\Theta(N^{-4s/(4s+d)})$ où d est la dimension de l'espace, s la régularité de Sobolev des densités et N la taille des échantillons. On retrouve une dépendance en la dimension semblable à celle de la vitesse d'estimation non paramétrique. Cette dépendance disparaît cependant lorsque la distance de séparation du test est considérée directement en terme de distance MMD entre les distributions. Dans ce cas, la distance de séparation du test est la distance de séparation des vecteurs dans le RKHS $\mathcal{H}$ et on retrouve d'ailleurs la même vitesse en $\Theta(N^{-1/2})$. En faisant le parallèle

avec le cas gaussien on peut se demander quelle est l'influence de la dimension de l'espace sur cette distance de séparation et s'il est possible de retrouver une forme (II.8). Le rôle de la dimension pour la distance de séparation, mais aussi pour l'erreur de test, va être en réalité joué par une notion de dimension effective que nous définissons dans la section suivante.

II.3 Dimension effective

Comme nous l'avons déjà évoqué, analyser et apprendre de l'information de données de grande dimension est souvent possible car les données possèdent en réalité une structure plus simple. Ainsi la dimension de l'espace dans lequel vit une distribution n'est pas forcément une quantité critique pour quantifier la difficulté d'une tâche. Ce rôle est plutôt joué par des notions de dimension effective ou intrinsèque d'une distribution. Ces dimensions intrinsèques d'une distribution $\mathbb{P}$ dépendent du problème considéré et celles que nous considèrerons seront construites à partir de son opérateur de covariance (Baker, 1973). Dans toute la suite on considèrera que $\mathbb{P}$ est une distribution sur un espace de Hilbert $\mathcal{H}$. Des exemples seront donnés pour $\mathcal{H} = \mathbb{R}^d$ et $\mathcal{H}$ un RKHS.

Définition II.5 (Opérateur de covariance)**.** Soit $\mathbb{P}$ une distribution sur un espace de Hilbert $\mathcal{H}$ telle que $\mathbb{E}\Big[\|X\|_{\mathcal{H}}^2\Big] < \infty$, son opérateur de covariance est alors défini par

$$\Sigma(\mathbb{P}) : \left\{ \begin{array}{ccc} \mathcal{H} & \to & \mathcal{H}\,, \\ y & \mapsto & \mathbb{E}\big[\langle y, X\rangle_{\mathcal{H}} X\big] - \big\langle y, \mathbb{E}\big[X\big]\big\rangle_{\mathcal{H}} \mathbb{E}\big[X\big]\,. \end{array} \right.$$

où X est une variable aléatoire de loi $\mathbb{P}$.

Sur $\mathbb{R}^d$, pour le produit scalaire canonique, l'opérateur de covariance est seulement la matrice de covariance de la distribution : si $\mu = \mathbb{E}[X]$, alors $\Sigma(\mathbb{P}) = \mathbb{E}\big[(X-\mu)(X-\mu)^T\big]$.

En dimension infinie, on peut considérer l'opérateur de covariance de l'injection d'une distribution dans un RKHS $\mathcal{H}$. Soit $X \sim \mathbb{Q}$ une variable aléatoire sur $\mathcal{X}$ et $\mathbb{P}$ la distribution de $\kappa(X, \cdot)$ sur $\mathcal{H}$. L'opérateur de covariance est bien défini lorsque $\mathbb{E}[\kappa(X,X)]$ est finie (par exemple pour κ borné) et, dans ce cas, pour tout $h \in \mathcal{H}$:

$$\Sigma(\mathbb{P})h = \mathbb{E}\big[\langle h, \kappa(X,\cdot)\rangle_{\mathcal{H}} \kappa(X,\cdot)\big] - \langle h, \mu_{\mathbb{Q}}\rangle_{\mathcal{H}} \mu_{\mathbb{Q}}(\cdot)\,, \tag{II.17}$$

où $\mu_{\mathbb{Q}}$ est le KME (Definition II.3) de la distribution $\mathbb{Q}$. En particulier l'opérateur de covariance vérifie pour tout $h, h' \in \mathcal{H}$:

$$\langle h', \Sigma(\mathbb{P})h\rangle_{\mathcal{H}} = \mathbb{E}[h(X)h'(X)] - \mathbb{E}[h(X)]\mathbb{E}[h(X')] = \mathrm{Cov}[h(X), h'(X)]$$

Cette identité est parfois utilisée comme définition de l'opérateur. L'opérateur $\Sigma(\mathbb{P})$ est borné (donc continu), positif et auto-adjoint. Nous faisons remarquer au lecteur que nous considérons ici la version centrée de l'opérateur de covariance. En effet ce sont les moments de cet opérateur qui interviennent dans nos méthodes et que nous aurons parfois à estimer (voir Section II.4.5). Dans les méthodes noyaux sa version non centrée est aussi utilisée, en particulier pour des tests d'indépendances et d'indépendance conditionnelle (Doran et al., 2014 ; Gretton et al., 2005). On peut cependant relever l'utilisation récente de la version centrée pour de l'analyse à composante principale à noyaux (Sriperumbudur et Sterge, 2022) et pour des two samples tests (Hagrass et al., 2022 ; Li et Yuan, 2019). Nous ne nous intéresserons cependant pas à son estimation mais l'utiliserons plutôt comme un outil théorique pour définir justement une notion de dimension effective. Nous aurons cependant besoin d'estimer certains de ces moments (voir Section II.4.5) pour lesquels nous fournirons des estimateurs.

Les notions de dimension effective que nous considèrerons s'expriment en fonction de normes de Schatten de l'opérateur de covariance de la distribution. La définition suivante introduit les trois dimensions effectives qui seront considérées dans ce manuscrit.

Définition II.6 (Dimension effective)**.** Soit $\mathbb{P}$ une distribution sur un espace de Hilbert $\mathcal{H}$ et $\Sigma := \Sigma(\mathbb{P})$ son opérateur de covariance. On appellera dimensions effectives les quantités suivantes :

$$d^{\mathrm{e}}(\mathbb{P}) = \frac{\operatorname{Tr}\Sigma}{\|\Sigma\|_{op}}, \quad d^{*}(\mathbb{P}) = \frac{\operatorname{Tr}\Sigma^2}{\|\Sigma\|_{op}^2}, \quad d^{\bullet}(\mathbb{P}) = \frac{(\operatorname{Tr}\Sigma)^2}{\operatorname{Tr}\Sigma^2}, \tag{II.18}$$

où Tr désigne la trace et $\|\cdot\|_{op}$ la norme opérateur.

Remarque II.7. Nous exprimerons parfois les dimensions effectives en fonction des normes de Schatten de la covariance, pour $p \in \mathbb{N}^*$ la norme p-Schatten est définie par $\|\Sigma\|_p^p := \operatorname{Tr}(\Sigma^p)$, si cette quantité existe. Pour $p \geq p'$, les normes de Schatten vérifient $\|\Sigma\|_p \leq \|\Sigma\|_{p'}$.

D'après la remarque précédente, les dimension effectives sont bien définies si l'opérateur de covariance Σ est de classe trace, c'est-à-dire de trace finie. En effet pour $(e_k)_{k\in\mathbb{N}}$ une base orthonormée de $\mathcal{H}$, on a :

$$\operatorname{Tr}\Sigma = \sum_{k=0}^{+\infty} \langle e_k, \Sigma e_k\rangle_{\mathcal{H}} = \sum_{k=0}^{+\infty} \Big(\mathbb{E}\Big[\langle e_k, X\rangle_{\mathcal{H}}^2\Big] - \langle e_k, \mathbb{E}[X]\rangle_{\mathcal{H}}^2\Big) = \mathbb{E}\Big[\|X\|_{\mathcal{H}}^2\Big] - \big\|\mathbb{E}[X]\big\|_{\mathcal{H}}^2 < \infty\,.$$

La dernière quantité est bien bornée par hypothèse et en utilisant l'inégalité de Jensen.

Dans la littérature des matrices aléatoires, d^{e} est parfois appelée la dimension intrinsèque (Hsu et al., 2012 ; Tropp et al., 2015) ou rang effectif (Koltchinskii et Lounici, 2016), et d^* est connue sous le nom de rang numérique ou rang stable de Σ (Rudelson et Vershynin, 2007 ; Tropp et al., 2015).

Ces trois notions de dimension effective donnent une quantification des degrés de libertés de la distribution. Si la distribution $\mathbb{P}$ est isotrope, c'est à dire dont la covariance est $\Sigma(\mathbb{P}) = \sigma^2 I_d$, où d est la dimension de l'espace ambiant, alors toutes ces notions de dimension effective sont égales à d. Plus généralement on a les inégalités suivantes entre les dimensions effectives :

$$d \geq d^{\bullet}(\mathbb{P}) \geq d^{\mathrm{e}}(\mathbb{P}) \geq d^{*}(\mathbb{P})\,. \tag{II.19}$$

Les problèmes suivants donnent des situations très simples dans lesquelles chacune des dimensions de Eq.(II.18) interviennent.

- La dimension d^{e} caractérise plutôt les problèmes d'estimation. Par exemple si on cherche à estimer l'espérance d'une distribution à partir d'une observation $X \sim \mathbb{P}$, l'erreur quadratique d'estimation est alors :
$$\mathbb{E}\Big[\|X-\mu\|^2\Big] = \sigma^2 d^{\mathrm{e}}\,,$$
où $\mu = \mathbb{E}[X]$ et $\sigma^2 = \|\Sigma\|_{op}$. On retrouve la forme (II.1).

- La dimension d^* intervient plutôt dans les problèmes de test. Par exemple pour détecter l'espérance μ d'une distribution est proche de 0, la distance $\|\mu\|^2$ peut être estimée par $\langle X, X'\rangle$ où X et X' sont deux observations indépendantes. Alors :
$$\mathbb{E}[\langle X, X'\rangle] = \|\mu\|^2\,, \quad \text{et} \quad \operatorname{Var}[\langle X, X'\rangle] \geq \sigma^4 d^*$$
où $\sigma^2 = \|\Sigma\|_{op}$. On retrouve grossièrement que l'erreur de test est en $\sqrt{d}$ contre d pour l'estimation mais avec des dimensions différentes.

- La dimension $d^{\bullet}$ est moins naturelle à interpréter mais intervient dans le Chapitre V dans lequel nous utiliserons des tests pour améliorer l'estimation de vecteurs. Elle peut être vue comme le rapport de l'erreur d'estimation sur l'erreur de test :
$$\sqrt{d^{\bullet}} = \frac{d^{\mathrm{e}}}{\sqrt{d^*}} \simeq \frac{\mathbb{E}\Big[\|X-\mu\|^2\Big]}{\sqrt{\operatorname{Var}[\langle X, X'\rangle]}}\,.$$

Dans les exemples suivants nous calculons des dimensions effective de différentes distributions.

Exemple II.8 (Support de plus petite dimension)**.** Supposons que $\mathbb{P}$ est à support dans un sous espace vectoriel de dimension p. Alors

$$d^{\mathrm{e}}(\mathbb{P}) \leq p.$$

Supposons maintenant que $\mathcal{H}$ est de dimension finie d et considérons $\widetilde{\mathbb{P}}$ la distribution $\mathbb{P}$ bruitée par addition d'un bruit indépendant et isotrope de covariance $\varepsilon^2 I_d$. Alors

$$d^{\mathrm{e}}\big(\widetilde{\mathbb{P}}\big) \leq p + d\frac{\varepsilon^2}{\|\Sigma(\mathbb{P})\|_{op}}.$$

Si le bruit est assez faible ($\varepsilon^2 \ll d^{-1}\|\Sigma(\mathbb{P})\|_{op}$), la dimension effective capture que la distribution $\mathbb{Q}$ est une version bruitée d'une distribution de support de plus petite dimension.

Exemple II.9 (RKHS discret)**.** Considérons $\mathcal{H}$ le RKHS associé au noyau trivial $\kappa(x,y) = \mathbf{1}\{x=y\}$ défini sur l'espace discret $\mathcal{X} = \{x_1, \ldots, x_m\}$. Soit $\mathbb{Q} = \sum_{i=1}^m p_i \delta_{x_i}$ une distribution sur $\mathcal{X}$ et $\mathbb{P}$ la distribution du plongement de $\mathbb{Q}$ dans $\mathcal{H}$. La dimension effective d^{e} de $\mathbb{P}$ est alors encadrée par :

$$\frac{1}{2}\frac{1-\|p\|_2^2}{\|p\|_\infty(1-\|p\|_\infty)} \leq d^{\mathrm{e}}(\mathbb{P}) \leq \frac{1-\|p\|_2^2}{\|p\|_\infty(1-\|p\|_\infty)}$$

où $p \in [0,1]^m$ est le vecteur de probabilités de $\mathbb{Q}$. Dans ce cas la notion de dimension effective s'interprète comme une version renormalisé de l'indice de Gini-Simpson (Simpson, 1949). L'indice de Gini-Simpson est $1-\|p\|_2^2$ et mesure si une population est diversifiée. Dans notre cadre une population diversifiée (cas limite $p_i \simeq m^{-1}$) aura une grande dimension effective ($d^{\mathrm{e}} \simeq m$).

Exemple II.10 (Noyau de translation)**.** Considérons $\mathcal{H}$ le RKHS associé à un noyau de translation $\kappa_h(x,y) := K((x-y)/h)$ où $h > 0$, la fenêtre ou *bandwidth* du noyau, est un paramètre fixé et $x, y \in \mathcal{X} = \mathbb{R}^d$. Soit $\mathbb{Q}$ une distribution sur $\mathbb{R}^d$ à densité f par rapport à la mesure de Lebesgue et $\mathbb{P}_h$ la distribution du plongement de $\mathbb{Q}$ dans $\mathcal{H}$. En supposant K et f assez régulières, pour des petites fenêtres, la dimension effective se relie à la norme L^2 de la densité. En effet lorsque le bandwidth h tend vers 0 :

$$d^{\bullet}(\mathbb{P}_h) \underset{h\to 0}{\sim} \frac{K(0)^2}{h^d \|K\|_{L^2}^2 \|f\|_{L^2}^2}\,.$$

Une distribution diffuse ($\|f\|_{L^2}^2$ petite) aura une dimension effective plus grande dans le RKHS qu'une distribution concentrée. Par exemple si K est un noyau gaussien et $\mathbb{Q} \sim \mathcal{N}(\mu, \sigma^2 I_d)$ une loi gaussienne, la dimension effective dépend alors du ratio entre la variance et la fenêtre :

$$d^{\bullet}(\mathbb{P}_h) \underset{h\to 0}{\sim} \left(\frac{8\sigma^2}{h^2}\right)^{d/2}, \qquad \text{de plus} \qquad d^{\bullet}(\mathbb{P}_h) \underset{h\to\infty}{\longrightarrow} d\,. \tag{II.20}$$

Dans ce cas gaussien, on retrouve explicitement que la dimension effective sera plus grande pour une distribution de grande variance.

Ces dimensions effectives se relient aussi à la notion de dimension de covariance locale (Dasgupta et Freund, 2008) définie à partir des valeurs propres $\sigma_1^2 \geq \ldots \geq \sigma_d^2$ de la covariance Σ. Par exemple, Verma et al. (2009) définissent qu'une distribution $\mathbb{P}$ est de dimension de covariance (p,ε) si les p plus grandes valeurs propres représentent une proportion $(1-\varepsilon)$ de la trace de la covariance :

$$\sum_{i=1}^{p} \sigma_i^2 \geq (1-\varepsilon)\,\mathrm{Tr}\,\Sigma\,.$$

Cette notion peut aussi se définir localement, en supposant cette condition vérifiée non plus pour la covariance mais pour les covariance des distributions restreintes à chaque boule d'un certain rayon. On peut remarquer que pour $\varepsilon > 0$ fixé, une distribution $\mathbb{P}$ est de dimension de covariance $(d^{\mathrm{e}}, \varepsilon)$ pour des dimension effectives de l'ordre de la dimension ambiante ($d^{\mathrm{e}} \gtrsim (1-\varepsilon)d$) ou proches de 1 ($d^{\mathrm{e}} \lesssim (1+\varepsilon)$). Plus généralement, $\mathbb{P}$ est de dimension de covariance $(d^{\mathrm{e}}, \varepsilon_{d^{\mathrm{e}}})$ pour $\varepsilon_{d^{\mathrm{e}}} \leq (1-\frac{1}{d^{\mathrm{e}}})(1-\frac{d^{\mathrm{e}}-1}{d-1})$.

D'un côté plus géométrique, pour une mesure de probabilité, certaines notions de dimension vont se définir en fonction de l'évolution avec le rayon de la mesure des boules. Grossièrement, une distribution $\mathbb{Q}$ sera de dimension d si la probabilité de toute boule S_r de rayon $r > 0$ est dominée par le volume par le volume d'une boulde d-dimensionnelle, c'est-à-dire si on a $\mathbb{P}(B_r) < Cr^d$ pour une certaine constante C (selon les cas pour r borné ou tendant vers 0). Ces notions sont fortement liées aux dimensions de Hausdorff (1918) et de Assouad (1979), généralisées aux distributions par exemple par la *pointwise dimension* (Young, 1982), l'*information dimension* (Isham, 1993), la *doubling dimension* (voir Heinonen, 2001) ou par la notion de distribution maximalement homogène (Kpotufe, 2011). Si la distribution est à densité sur un sous espace vectoriel, ces dimensions vont coïncider avec la dimension de cet espace tout comme les dimensions que nous considérons (voir Exemple II.8). Cependant, ces notions vont diverger pour des distributions discrètes. Pour de telles distributions et des rayons assez petits, les mesures des boules ne vont plus évoluer avec le rayon ce qui donne une dimension égale à 0. À l'inverse, les dimensions (II.18) vont bien considérer la structure des valeurs prises par la distribution dans l'espace. Par exemple pour une distribution uniforme sur une famille orthonormée de vecteurs $(e_k)_{1\leq k\leq n}$:

$$d^{\mathrm{e}}\left(\frac{1}{n}\sum_{k=1}^{n}\delta_{e_k}\right) = n-1\,.$$

Pour une distribution discrète sur une droite, les dimensions seront de 1.

II.4 Contributions

Un objectif récurrent dans cette thèse est de faire le pont entre les phénomènes de grande dimension étudié pour des données gaussiennes isotropiques avec les outils modernes actuels comme le KME présenté précédemment. Le problème central que nous considérons est l'estimation simultanée de vecteurs moyennes, pouvant être justement des KMEs de différentes distributions. Ce problème peut être considéré comme une question d'apprentissage multi-tâche, voire aussi d'apprentissage par transfert où l'utilisateur cherche à estimer ce nouveau vecteur en s'aidant d'estimations d'autres objets ou même d'apprentissage fédéré où les données distribuées peuvent avoir des variations dans leurs distributions et que seul un vecteur peut être transmis. Nous le considérerons ici de manière la plus générale en voyant ces vecteurs comme des éléments d'un espace de Hilbert dont les observations sont perturbées par un bruit Gaussien ou borné.

Nous considérons ce problème en Section III et proposons une méthode utilisant un effet de type Stein où l'amélioration de l'estimation augmente avec la dimension. Les garanties sont données sous des hypothèses d'homogénéité entre les différentes distributions. Pour les KMEs, la grande dimension signifie alors une grande dimension effective. Cette méthode est construite à partir de tests et utilise la relative facilité de détection relativement à l'estimation. En Section IV nous généralisons ce phénomène de test à des données non isotropiques et en particulier aux KMEs. Ce travail, bien qu'indépendant, permet aussi d'étendre notre estimation multiple de vecteurs à des données hétérogènes en Section V. Dans cette section nous proposons deux méthodes, l'une reposant sur des tests et l'autre sur la minimisation du risque empirique. Fondamentalement la combinaison des échantillons permet de débruiter l'estimation de chaque moyenne en particulier lorsque les distributions ont une structure commune. On constate que ce phénomène se retrouve indirectement dans le mécanisme de self-attention utilisé dans les Transformers (Vaswani et al., 2017). En Section VI, nous montrons ainsi que la self attention fonctionne comme un débruitage de données de grande

dimension et qu'on retrouve des comportements comparables en grande dimension à ceux considérés dans les précédentes sections.

Ces différentes contributions sont détaillées dans cette section.

II.4.1 Estimation multiple de vecteurs moyennes pour des données homogènes

La Section III est un travail en collaboration avec Gilles Blanchard et Hannah Marienwald (Marienwald et al., 2021) et s'intéresse à l'estimation simultanée des vecteurs moyennes de différentes distributions. Considérons le modèle :

$$\begin{cases} X_\bullet^{(k)} := (X_i^{(k)})_{1\le i\le N_k} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_k,\ k \in [\![B]\!]; \\ (X_\bullet^{(1)}, \dots, X_\bullet^{(B)}) \text{ indépendants.} \end{cases} \tag{II.21}$$

l'objectif est d'estimer le vecteur moyenne $\mu_k = \mathbb{E}\left[X_1^{(k)}\right]$ de chacune des distributions $\mathbb{P}_k$ pour lesquelles un échantillon $X_\bullet^{(k)}$ est à disposition. Pour un échantillon donné, nous construisons un estimateur de contraction de sa moyenne empirique vers un point de référence. Ce point de référence ne va cependant pas être choisi arbitrairement mais à partir des autres échantillons. L'objectif va être d'utiliser la relative facilité du test par rapport à l'estimation en grande dimension pour trouver un point de référence proche de la vraie moyenne.

Pour des données Gaussiennes isotropes et des échantillons homogènes, $N_k = N$ et $\mathbb{P}_k = \mathcal{N}(\mu_k, \sigma^2 I_d)$, nous estimons pour chaque moyenne un ensemble de τ-voisins $\widehat{V}_i = \{j : T_{ij} = 1, j \in [\![B]\!]\}$ où T_{ij} est un test pour les hypothèses :

$$(H_{0,ij}) : \|\mu_i - \mu_j\|^2 \le \tau\sigma_N^2 d, \quad \text{contre} \quad (H_{1,ij}) : \|\mu_i - \mu_j\|^2 > \tau\sigma_N^2 d. \tag{II.22}$$

où $\sigma_N^2 = \sigma^2/N$. Le but de ces tests est de trouver des échantillons dont les moyennes sont proches de la moyenne cible relativement à l'erreur d'estimation ($d\sigma_N^2$). L'estimateur considéré pour μ_i est une contraction de la moyenne empirique vers la moyenne des moyennes empiriques sélectionnées par les tests

$$\widehat{\mu}_i = \gamma\widehat{\mu}_i^{\texttt{NE}} + \frac{1-\gamma}{|\widehat{V}_i|}\sum_{j\in\widehat{V}_i}\widehat{\mu}_j^{\texttt{NE}}, \tag{II.23}$$

où $\widehat{\mu}_j^{\texttt{NE}}$ est la moyenne empirique sur l'échantillon j et $\gamma \in (0,1)$ un paramètre à fixer. L'erreur de test, en norme au carré, pour des distributions Gaussiennes isotropes, est de l'ordre de $\sigma_N^2\sqrt{d}$, on s'attend donc à pouvoir construire des tests tels que les voisins sélectionnés soient à une distance de la vraie moyenne μ_i d'au plus $\tau\sigma_N^2 d + \sigma_N^2\sqrt{d}$ (voir Section II.1.2). Ainsi en Section III :

- Nous construisons des tests tel que le biais ajouté par cette contraction est d'ordre plus petit que l'erreur d'estimation de la moyenne empirique et nous montrons théoriquement et expérimentalement que notre méthode améliore l'estimation relativement à la moyenne empirique.
- Les données sont supposées Gaussiennes isotropes ou bornées dans le but d'appliquer notre méthode à l'estimation de KMEs. Dans ces deux cadres, nous faisons une hypothèse d'homogénéité des échantillons : les tailles et variances des échantillons sont supposées du même ordre. Cette hypothèse justifie notre utilisation symétrique des moyennes sélectionnées par les tests dans (II.23).

Cette section peut être considérée comme une introduction à la Section V où la méthode est généralisée à des échantillons non homogènes. Cependant la méthode présentée en Section III garde un intérêt indépendant en proposant une approche simple, peu coûteuse algorithmiquement et dont les garanties théoriques sont plus fines. En particulier nos bornes prennent en compte la dépendance des données entre test et estimation qui nous mettons de côté en Section V.3 (voir par exemple Théorème III.2).

II.4.2 Tests à un ou deux échantillons en grande dimension avec une structure de covariance inconnue

La méthode proposée en Section III repose sur l'analyse dans le cadre isotrope gaussien de la distance de séparation (II.8). Dans le but de généraliser cette méthode, nous nous sommes intéressés au problème de test à deux échantillons de proximité de moyennes

$$(H_0(\eta)) : \|\mu - \nu\| \leq \eta, \text{ contre } (H_1(\eta, \delta)) : \|\mu - \nu\| > \eta + \delta. \tag{II.24}$$

où μ et ν sont les vecteurs moyennes respectifs de distributions $\mathbb{P}$ et $\mathbb{Q}$ connues via des échantillons i.i.d. $\{X_i\}_{1\leq i\leq n}$ et $\{Y_i\}_{1\leq i\leq m}$. Ce problème est une généralisation du problème de détection de signal (II.6) : on peut s'y ramener en supposant formellement que $m = +\infty$ ou que $\mathbb{Q}$ est une distribution Dirac. Lorsque μ et ν sont des KMEs ce test permet de tester la proximité de deux distributions. Sous cette forme, la distance de séparation est la plus petite distance δ_α pour laquelle il existe un test tel que la somme des erreurs de type I et II soient contrôlées par $\alpha \in (0, 1)$ donné. Le point important dans notre analyse est que les covariances de chaque distribution ne sont pas supposées connues et qu'elles sont potentiellement différentes. Nos contributions sont les suivantes :

- Nous procédons à une analyse minimax du test (II.24) et donnons une borne inférieure à la distance de séparation pour des données gaussiennes. Nous retrouvons les deux régimes de Blanchard et al. (2018) (Eq.(II.8)) mais où la dimension de l'espace est remplacée par une notion de dimension effective du problème.

- Nous construisons des tests atteignant cette borne inférieure pour des données gaussiennes et bornées dans un espace de Hilbert. Nos tests sont construits à partir de la U-statistique de la forme (II.13) et d'estimateurs de ses quantiles. Des inégalités de concentration sont données pour tous ces estimateurs dans le cadre gaussien et borné.

Ce travail a été fait en collaboration avec Gilles Blanchard (Blanchard et Fermanian, 2023).

II.4.3 Généralisation de l'estimation multiple de moyennes et minimaximalité

La Section V s'appuie sur ces deux précédentes parties et a été faite en collaboration avec Gilles Blanchard et Hannah Marienwald (Blanchard et al., 2024). Nous considérons à nouveau le modèle (II.21) mais ne supposons plus aucune homogénéité entre les distributions. Les tailles d'échantillons sont différentes ainsi que les covariances des distributions supposées inconnues. Pour choisir d'aggréger deux estimateurs, il faudra prendre en compte à la fois la proximité des moyennes mais aussi le rapport des variances. Même si tous les échantillons ont la même moyenne, ceux avec une petite variance auront tout intérêt à être privilégiés. Pour cela nous nous considérons donc comme estimateur des combinaisons convexe des moyennes empiriques

$$\widehat{\mu}_{\boldsymbol{\omega}} = \sum_{k=1}^{B} \omega_k \widehat{\mu}_k^{\mathtt{NE}}, , \tag{II.25}$$

où $\widehat{\mu}_k^{\mathtt{NE}}$ est la moyenne empirique de l'échantillon k et $\boldsymbol{\omega}$ est un vecteur de poids dans le simplexe $\mathcal{S}_B$ (i.e. $\sum_{k=1}^{B} \omega_k = 1$ et $\omega_k \geq 0$). Cette forme d'estimateur est plus générale que celle considérée en Section III (l'estimateur (II.23) se ramène à (II.25) en prenant $\gamma = \omega_1$ et $\omega_k = (1-\gamma)\mathbf{1}_{k\in\widehat{V}_1}|\widehat{V}_1|^{-1}$). Notre but est toujours d'améliorer l'estimation de chaque moyenne relativement à l'estimation naïve par la moyenne empirique. Pour cela nous proposons deux méthodes pour estimer des poids $\boldsymbol{\omega}$ optimaux.

La première méthode présentée en Section V.3 repose sur des tests et est une adaptation de la méthode de la Section III dans le cas hétérogène. Pour estimer la moyenne μ_1, notre première étape consiste à

sélectionner des moyennes (relativement) proches et de variances (relativement) plus faible. Nous chercherons à estimer :

$$V_\tau = \Big\{k \in [\![B]\!] : \|\mu_k - \mu_1\|^2 \le \tau\sigma_1^2 d^{\mathrm{e}}(\mathbb{P}_1)\Big\}, \quad \text{et} \quad W_{(\varsigma)} = \Big\{k : \sigma_k^2\big(d^*(\mathbb{P}_k)\big)^{1/2} \le \varsigma\sigma_1^2\big(d^*(\mathbb{P}_1)\big)^{1/2}\Big\} \tag{II.26}$$

où $\tau, \varsigma > 0$ sont des paramètres fixes et $\sigma_k^2 = \|\Sigma_k\|_{\mathrm{op}}/N_k$. En mots, V_τ contient les distribution de moyennes proches relativement à l'erreur d'estimation de μ_1 et $W_{(\varsigma)}$ contient les distributions d'erreur de tests plus faibles que celle de la distribution $\mathbb{P}_1$. L'ensemble $W_{(\varsigma)}$ permet d'exclure des moyennes qui pourraient être sélectionnées des tests pour V_τ mais dont les moyennes seraient en réalité trop loin de μ_1. Ces ensembles sont estimés à l'aide de tests. La seconde étape consiste à estimer des poids $\boldsymbol{\omega}$ donnés par une minimisation oracle du risque théorique. Le poids ω_k attribué à l'estimateur $\widehat{\mu}_k^{\mathtt{NE}}$ va décroître avec son erreur d'estimation $\sigma_k^2 d^{\mathrm{e}}(\mathbb{P}_k)$.

Cette approche diffère de la première méthode présentée par la non symétrie de la relation "être un τ-voisin". L'estimateur d'un échantillon de grande variance et/ou de petite taille aura plus de voisins et sera plus contracté qu'un échantillon de grande taille. La méthode améliorera plus les échantillons ayant peu d'informations initialement mais ne détériorera pas l'estimation pour les autres.

Nous donnons des bornes non asymptotiques sur l'erreur de cette méthode et sur ses différents étapes. Nous proposons des estimations de V_τ et $W_{(\varsigma)}$ qui peuvent être remplacés par d'autres estimateurs si besoin. Cette améliore l'estimation en particulier lorsque l'erreur de test est faible relativement à l'erreur d'estimation, c'est-à-dire lorsque la dimension effective $\sqrt{d^\bullet} = d^{\mathrm{e}}/\sqrt{d^*}$ est grande (pour $\mathbb{P}_1$).

Un point faible des deux approches tests (homogènes et hétérogènes) est la nécessité de choisir les paramètres τ et ς. Nous proposons donc une seconde méthode qui utilise des idées de la Q-aggrégation de Lecué et Rigollet (2014). Au lieu de sélectionner par des tests les voisins, nous estimons les poids en minimisant un estimateur d'une borne supérieure du risque quadratique $\mathcal{R}_1(\boldsymbol{\omega}) := \mathbb{E}\Big[\|\widehat{\mu}_{\boldsymbol{\omega}} - \mu_1\|^2\Big]$. Cet estimateur est constitué de deux termes, le premier terme $\widehat{L}_1(\boldsymbol{\omega})$ est un estimateur de $\mathcal{R}_1(\boldsymbol{\omega})$ et le second $\widehat{Q}_1(\boldsymbol{\omega})$ un estimateur des déviations de $\Big|\widehat{L}_1(\boldsymbol{\omega}) - \mathcal{R}_1(\boldsymbol{\omega})\Big|$. Notre estimateur est alors $\widehat{\mu}_{\widehat{\boldsymbol{\omega}}}$ où

$$\widehat{\boldsymbol{\omega}} \in \operatorname*{Arg\,Min}_{\boldsymbol{\omega}\in\mathcal{S}_B}\Big(\widehat{L}_1(\boldsymbol{\omega}) + \widehat{Q}_1(\boldsymbol{\omega})\Big).$$

Le terme $\widehat{Q}_1$ est tiré d'inégalités de concentration et fait intervenir les covariances des distributions. La pondération par $\widehat{Q}_1$ va imposer une forme de sparsité au vecteur $\widehat{\boldsymbol{\omega}}$ comme pourrait le faire une pénalisation ℓ_1 mais qui prend en compte la dimensionalité des différents échantillons. Nous l'interprétons comme des tests implicitement effectué par cette régularisation.

Nous construisons de tels estimateurs et donnons des bornes sur l'erreur quadratique moyenne de l'estimateur $\widehat{\mu}_{\widehat{\boldsymbol{\omega}}}$. Nous retrouvons d'ailleurs dans ces bornes la même vitesse de convergence que l'approche test en $O((d^\bullet)^{-1/2})$. La Q-aggrégation a l'avantage cependant d'être adaptative en τ et ς. En pratique, les deux méthodes obtiennent des résultats comparables.

Intuition. Notre objectif est de choisir des estimateurs pour avoir une borne sur le vrai risque, c'est-à-dire que le risque soit borné avec grande probabilité :

$$\mathcal{R}_1(\boldsymbol{\omega}) \le \widehat{L}_1(\boldsymbol{\omega}) + \widehat{Q}_1(\boldsymbol{\omega}) + O(\sqrt{d^*}) \le \mathcal{R}_1(\boldsymbol{\omega}) + O(\sqrt{d^*}), \quad \forall \boldsymbol{\omega} \in \mathcal{S}_B.$$

On s'attend en effet à ce que les déviations de l'estimation d'une distance soit de l'ordre de $\sqrt{d^*}$. On peut faire le parallèle avec la déviation d'un vecteur Gaussien $Z \sim \mathcal{N}(\mu, \Sigma)$. Supposons qu'on veuille estimer $\mathcal{R} = \mathbb{E}\Big[\|Z - \nu\|^2\Big]$ pour un vecteur $\nu \in \mathbb{R}^d$. Alors pour tout $u \ge 0$, avec probabilité $1 - e^{-u}$:

$$\mathbb{E}\Big[\|Z - \nu\|^2\Big] \le \|Z - \nu\|^2 + 2\sigma^2\sqrt{(d^* + 2(\mu-\nu)^T\Sigma(\mu-\nu))u}$$

où $\sigma^2 = \|\Sigma\|_{\text{op}}$ et $d^* = d^*\big(\mathcal{N}(\mu, \Sigma)\big)$ est la dimension effective de la loi gaussienne (voir Lemme V.40). Dans ce cas $\|Z - \nu\|^2$ serait notre estimateur $\widehat{L}_1$ et $\widehat{Q}_1$ serait un estimateur de $\sqrt{(\mu-\nu)^T\Sigma(\mu-\nu)}$. On retrouve bien la déviation de l'ordre de $\sqrt{d^*}$.

Les erreurs de ces deux méthodes sont étudiées sous un l'axe de grande dimension (effective). Tout comme l'estimateur de James-Stein (voir Section II.1.1), l'amélioration se fait au niveau de la dépendance en la dimension de la vitesse de convergence. En Section V.5, nous faisons une analyse minimax du problème et donnons des bornes inférieures pour l'amélioration optimale possible pour un échantillon et en moyenne sur les tous les échantillons. Nous discutons de l'optimalité de nos deux méthodes.

II.4.4 Effet de débruitage de la self-attention

La Section VI présente des travaux préliminaires en collaboration avec Gilles Blanchard sur le mécanisme d'attention. Ce mécanisme est utilisé dans les réseaux de neurones Transformers (Vaswani et al., 2017) très utilisés aujourd'hui en particulier pour des tâches de génération de données de texte ou d'image. Dans ces cas, les entrées du réseau de neurones ne sont plus un point mais un ensemble de points pouvant être les mots d'une phrase ou d'un texte ou bien les encodement de sous parties d'une image. Pour traiter ce type de données, les Transformers ajoutent des couches supplémentaires dans le réseau de neurone dites d'attention qui vont chercher à considérer non plus les points individuellement mais dans leur globalité. L'intuition est très simple : pour traduire un mot d'une phrase il est important de prendre en compte son contexte. Formellement, pour des points $X_1, \dots, X_N$, l'attention construit dans une première étape N nouveaux points :

$$a_{Q,K}(X_i) := \sum_{j=1}^{N} \omega_{ij} X_j \quad \text{où} \quad (\omega_{ij})_j = \texttt{Softmax}\Big(\big(\langle QX_i, KX_j\rangle\big)_j\Big) \in \mathcal{S}_N\,, \tag{II.27}$$

où $\mathcal{S}_N$ est le toujours le simplexe et Q et K sont des matrices apprises durant l'entraînement du réseau de neurone. Nous supposons ici que ces matrices sont fixes, nous plaçons donc après l'apprentissage et cherchons à comprendre l'action de l'attention sur les données. En comparant cette forme avec (II.25), nous proposons d'interpréter le mécanisme d'attention comme une forme d'estimation multiple de vecteurs. Nous supposons ainsi que les points X_i sont des versions bruitées de vecteurs μ_i ayant une structure plus simple et nous nous demandons si les nouveaux points $a_{Q,K}(X_i)$ seraient moins bruités par rapport aux points originaux X_i, c'est-à-dire si pour $1 \le i \le N$:

$$\mathbb{E}\Big[\|a_{Q,K}(X_i) - \mu_i\|^2\Big] < \mathbb{E}\Big[\|X_i - \mu_i\|^2\Big]\,.$$

pour la dimension tendant vers l'infini. Nous analysons cette question dans un cadre simplifié où les vecteurs X_i sont gaussiens isotropes et où les matrices Q et K sont proportionnelles à l'identité ($Q = K = I_d/\sqrt{h}$) et dans ce cas nous étudions les valeurs du paramètres h pour lesquelles le mécanisme de self-attention n'est pas dégénéré (i.e. $a_{Q,K}(X_i) \neq X_i$ et $a_{Q,K}(X_i)$ différent de la moyenne des données). Pour un tel paramètre nous exhibons certaines structures des moyennes μ_i pour lesquelles il y a effectivement débruitage (support de plus petite dimension ou petit nombre de couverture pour un certain rayon). À partir de cette analyse de l'effet de la dimension, nous proposons une version légèrement modifée des poids ω de (II.27) pour lesquels nous obtenons théoriquement et sur des données simulées un effet de débruitage pour un plus large spectre de paramètre h ($h \simeq d$ pour les méthodes originales contre $\sqrt{d} \lesssim h \lesssim d$ pour notre version modifiée).

Une hypothèse important de notre analyse est que les vecteurs μ_i sont de même norme. Dans ce cas, on peut remarquer que les poids donnés par des produits scalaires sont les mêmes que ceux donnés par la norme carré ($\texttt{Softmax}((\langle \mu_i, \mu_j\rangle)_j) = \texttt{Softmax}\Big((-\|\mu_i - \mu_j\|^2/2)_j\Big)$). Comme nous supposons de plus que les matrices Q et K sont proportionnelles à l'identité, nous pouvons interpréter $\langle QX_i, KX_j\rangle$ comme un

estimateur de la distance euclidienne entre μ_i et μ_j (à constante additive près et facteur $1/2$). Comme pour $i = j$, cet estimateur est biaisé, notre modification des poids d'attention va juste enlever ce biais.

L'hypothèse que les vecteurs μ_i soient sur la phère est justifée pour nous par la seconde étape de l'attention qui consiste à combiner les points initiaux avec $a_{Q,K}(X_i)$ et construire pour $1 \leq i \leq N$:

$$A_{Q,K,V}(X_i) := \frac{X_i + V a_{Q,K}(X_i)}{\|X_i + V a_{Q,K}(X_i)\|} .$$

pour une matrice V apprise aussi par le réseau de neurone. Ainsi à la sortie de la self attention, à chaque point est bien associé un vecteur normalisé. C'est d'ailleurs sous cette forme que la self attention est plutôt étudiée bien que la normalisation soit parfois négligée. Dans notre modèle, les vecteurs $A_{Q,K,V}(X_i)$ peuvent s'interpréter comme des estimateurs de contraction (matricielle) de X_i vers $a_{Q,K}(X_i)$ et se relient aux estimateurs de type empirical Bayes (Brown et Greenshtein, 2009) pour certaines matrices V (voir discussion en Section VI.1.1). Notre analyse se concentre cependant pour l'instant sur $a_{Q,K}(X_i)$. Ces résultats bien que restreints à des versions simplifiées du mécanisme d'attention dans un cadre gaussien isotrope mettent en lumière un phénomène, à notre connaissance, méconnu et amènent de nombreuses nouvelles questions.

II.4.5 Inégalités de concentration

Ces différents travaux utilisent de façon répétée des estimateurs de distance entre les vecteurs et de moments de la covariance des distributions. Ces estimations sont contrôlées par des bornes de concentration pouvant être d'un intérêt indépendamment des problèmes que nous considérons. En particulier ces bornes portent une grande attention au rôle de la dimension (effective) dans les déviations. En dehors des données gaussiennes et bornées, nous considérons aussi des distributions à queues lourdes où seul un moment d'ordre quatre fini est supposé. Pour ce type de données des estimateurs de type "median of means" sont considérés (voir Lugosi et Mendelson, 2019a par exemple). La Table 2 pointe vers les différents résultats disséminés dans la thèse.

Quantité estimée	Cadre Gaussien	Cadre borné	Cadre queues-lourdes
$\|\mu - \nu\|^2$	Proposition IV.6	Proposition IV.9	Proposition V.33
$\operatorname{Tr}\Sigma$	Proposition V.26	Proposition V.29	Proposition V.34
$\|\Sigma\|_{\mathrm{op}}$	Proposition IV.10	Proposition IV.11	-
$\sqrt{\operatorname{Tr}\Sigma^2}$	Proposition IV.12	Proposition IV.13	Proposition V.34
$(\mu - \nu)^T \Sigma (\mu - \nu)$	Proposition V.37	Proposition V.38	-

TABLE 2 : Recensement des différents résultats de concentration : μ et Σ sont respectivement la moyenne et covariance d'une distribution et ν la moyenne potentiellement inconnue d'une autre.

III High dimensional multi-task averaging

In this section, we propose an improved estimator for the multi-task averaging problem, whose goal is the joint estimation of the means of multiple distributions using separate, independent data sets. The naive approach is to take the empirical mean of each data set individually, whereas the proposed method exploits similarities between tasks, without any related information being known in advance. First, for each data set, similar or neighboring means are determined from the data by multiple testing. Then each naive estimator is shrunk towards the local average of its neighbors. We prove theoretically that this approach provides a reduction in mean squared error. This improvement can be significant when the dimension of the input space is large, demonstrating a "blessing of dimensionality" phenomenon. An application of this approach is the estimation of multiple kernel mean embeddings, which plays an important role in many modern applications. The theoretical results are verified on artificial and real world data.

Contents

III.1 Introduction

The estimation of means from i.i.d. data is arguably one of the oldest and most classical problems in statistics. In this work we consider the problem of estimating *multiple* means $\mu_1, \dots, \mu_B$ of probability distributions $\mathbb{P}_1, \dots, \mathbb{P}_B$, over a common space $\mathcal{X} = \mathbb{R}^d$ (or possibly a real Hilbert space $\mathcal{H}$). We assume that for each individual distribution $\mathbb{P}_i$, we observe an i.i.d. data set $X_\bullet^{(i)}$ of size N_i, and that these data sets have been collected independently from each other.

In the rest of the section, we will call each such data set $X_\bullet^{(i)}$ a *bag*. Mathematically, our model is thus

$$\begin{cases} X_\bullet^{(i)} := (X_k^{(i)})_{1\le k\le N_i} \overset{i.i.d.}{\sim} \mathbb{P}_i,\ 1 \le i \le B; \\ (X_\bullet^{(1)}, \dots, X_\bullet^{(B)}) \text{ independent,} \end{cases} \tag{III.1}$$

where $\mathbb{P}_1, \dots, \mathbb{P}_B$ are square integrable distributions on $\mathbb{R}^d$ which we call *tasks*, and our goal is the estimation of their means

$$\mu_i := \mathbb{E}_{X\sim\mathbb{P}_i}[X] \in \mathbb{R}^d,\ 1 \le i \le B. \tag{III.2}$$

Given an estimate $\widehat{\mu}_i$ of μ_i, we will be interested in its squared error $\|\widehat{\mu}_i - \mu_i\|^2$, and aim at controlling it either with high probability or in average (mean squared error, MSE):

$$\mathrm{MSE}(i, \widehat{\mu}_i) := \mathbb{E}\big[\|\widehat{\mu}_i - \mu_i\|^2\big];$$

this error can be considered either individually for each task $\mathbb{P}_i$ or averaged over all tasks.

This problem is also known as multi-task averaging (MTA) (Feldman et al., 2014), an instance of the multi-task learning (MTL) problem. Prior work on MTL showed that learning multiple tasks jointly yields better performance compared to individual single task solutions (Caruana, 1997; Evgeniou et al., 2005; Feldman et al., 2014).

In this chapter we adapt the idea of joint estimation to the multi-task averaging problem and will show that we can take advantage of some unknown *structure* in the set of tasks to improve the estimation. Here, by individual estimation we mean that our natural baseline is the naive estimator (NE) given by the simple empirical mean:

$$\widehat{\mu}_i^{\mathrm{NE}} := \frac{1}{N_i}\sum_{k=1}^{N_i} X_k^{(i)}; \quad \mathrm{MSE}(i, \widehat{\mu}_i^{\mathrm{NE}}) = \frac{1}{N_i}\operatorname{Tr}\Sigma_i, \tag{III.3}$$

where Σ_i is the covariance matrix of $\mathbb{P}_i$.

Our motivation for considering this setting is the growing number of large databases taking the above form, where independent bags corresponding to different but conceptually similar distributions are available; for example, one can think of i as an index for a large number of individuals, for each of which a number of observations (assumed to be sampled from an individual-specific distribution) are available, say medical records, or online activity information collected by some governmental or corporate mass spying device.

While estimating means in such databases is of interest of its own, a particularly important motivation to consider this setting is that of Kernel Mean Embedding (KME), a technique enjoying sustained attention in the statistical and machine learning community since its introduction in the seminal paper of Smola et al. (2007); see Section II.2.

The core principle of KME is to represent the distribution $\mathbb{P}_Z$ of a random variable Z via the mean of $X = \phi(Z)$, where ϕ is a rich enough feature mapping from the input space $\mathcal{Z}$ to a (reproducing kernel) Hilbert space $\mathcal{H}$. In practice, it is assumed that we have an i.i.d. bag $(Z_k)_{1\le k\le N}$ from $\mathbb{P}$, which is used to estimate its KME. Here we are interested again in the situation where a large number of independent data sets from different distributions are available, and we want to estimate their KMEs jointly. This is, therefore, an instance of the model (III.1), once we set $\mathcal{X} := \mathcal{H}$ and $X_k^{(i)} := \phi(Z_k^{(i)})$.

III.1.1 Relation to Previous Work

The fact that the naive estimator (III.3) can be improved upon when multiple, real-valued means are to be estimated simultaneously, has a long history in mathematical statistics. More precisely, let us introduce the following isotropic Gaussian setting:

$$\mathbb{P}_i = \mathcal{N}(\mu_i, I_d);\ N_i = N, \qquad 1 \le i \le B, \tag{GI}$$

on which we will come back in the sequel.

As shown in Stein (1956), for $B = 1$ with $d \geq 3$ the naive estimator is inadmissible, i.e. there exists a strictly better estimator, with a lower MSE for any true mean vector μ_1. An explicit example of a better estimator is given by the celebrated *James-Stein* estimator (JSE) (James and Stein, 1961), which shrinks adaptively the naive estimator towards $\mathbf{0}$, or more generally, towards an a priori fixed vector ν_0.

The MTA problem was introduced by Feldman et al. (2014), who proposed an approach which regularizes the estimation such that similar tasks shall have similar means as well. However, they assumed the pairwise task similarity to be given, which is unfeasible in most practical applications. In addition to our own approach, we will also introduce a variation of theirs, suitable for the KME framework, that *estimates* the task similarity instead of assuming it to be known. Martínez-Rego and Pontil (2013) proposed a method based on spectral clustering of the tasks and applying Feldman et al. (2014)'s method separately on each cluster, but without theoretical analysis.

Variations of the JSE can be shown to yield possible improvements over the NE in more general situations as well (see Fathi et al., 2020 for recent results in non-Gaussian settings). This has also been exploited for KME in Muandet et al. (2016), where a Stein-type estimator in kernel space was shown to generally improve over naive KME estimation. To the best of our knowledge, no shrinkage estimator for KME explicitly designed for or taking advantage of the MTA setting exists.

In the remainder of this work we will proceed as follows. Section III.2 introduces the basic idea of the approach and starts with a general discussion. We will expose in Section III.3 a theoretical analysis proving that the presented method improves upon the naive estimation in terms of squared error, possibly by a large factor. The general theoretical results will be discussed explicitly for the Gaussian setting (Section III.3.3) and in the KME framework (Section III.3.4). The approach is then tested for the KME setting on artificial and real world data in Section III.4. All proofs are found in the Sections III.6.1 to III.6.6, Section III.6.7 gives a detailed description of the estimators compared in the experiments, and Section III.6.8 presents additional numerical results in the Gaussian setting.

III.2 Method

The basic idea of our approach is to improve the estimation of a mean of a task by basing its estimation not on its own bag alone, but concatenating the samples from all bags it is *sufficiently similar* to. Since in most practical applications task similarity is not known, we will propose a statistical test that assesses task relatedness based on the given data.

III.2.1 Overview of the Approach

In the remainder of the section we will use the notation $[\![n]\!] := \{1, \dots, n\}$. For convenience of exposition, assume the (GI) setting. In this case, the naive estimators all have the same MSE, $\overline{\sigma}^2 := d/N$. Fix a particular task (reindexed $i = 0$) with mean μ_0 that we wish to estimate, and assume for now we are given the *side information* that for some constant $\tau > 0$, it holds $\Delta_{0i}^2 := \|\mu_0 - \mu_i\|^2 \leq \tau\overline{\sigma}^2$ for some "neighbor tasks" $i \in [\![V]\!]$ (a subset of the larger set of B tasks within range $\tau\overline{\sigma}^2$ to μ_0, reindexed for convenience). Consider the estimator $\widetilde{\mu}_0$ obtained by a simple average of neighbor naive estimators, $\widetilde{\mu}_0 = \frac{1}{V+1}\sum_{i=0}^{V} \widehat{\mu}_i^{\mathrm{NE}}$. We can bound via usual bias-variance decomposition, independence of the bags and convexity of the squared norm:

$$\mathrm{MSE}(0, \widetilde{\mu}_0) \;=\; \left\| \frac{1}{V+1} \sum_{i=1}^{V} (\mu_0 - \mu_i) \right\|^2 + \frac{\overline{\sigma}^2}{V+1} \;\leq\; \overline{\sigma}^2 \frac{(1 + V\tau)}{V+1}. \tag{III.4}$$

Thus, the above bound guarantees that $\widetilde{\mu}_0$ improves over $\widehat{\mu}_0^{\mathrm{NE}}$ whenever $\tau < 1$, and leads to a relative improvement of order $\max(\tau, V^{-1})$.

In practice, we *don't* have *any* a priori side information on the configuration of the means. A simple idea is, therefore, to estimate the quantities Δ_{0i}^2 from the data by an estimator $\widehat{\Delta}_{0i}^2$ and select only those bags for which $\widehat{\Delta}_{0i}^2 \leq \widetilde{\tau}\overline{\sigma}^2$. This is in a nutshell the principle of our proposed method.

The deceptive simplicity of the above idea might be met with some deserved skepticism. One might expect that the typical estimation error of $\widehat{\Delta}_{0i}^2$ would be of the same order as the MSE of the naive estimators. Consequently, we could at best guarantee with high probability a bound of $\Delta_{0i}^2 \lesssim \overline{\sigma}^2$ for the estimated neighbor tasks, i.e. $\tau \approx 1$, which does not lead to any substantial theoretical improvement when using (III.4). The reason why the above criticism is pessimistic, even in the worst case, is the role of the dimension d. From high-dimensional statistics, it is known that the rate of *testing* for $\Delta_{0i}^2 = 0$, i.e. the minimum ρ^2 such that a statistical test can detect $\Delta_{0i}^2 \geq \rho^2$ with probability close to 1, is faster than the rate of *estimation*, $\rho^2 \simeq \sqrt{d}/N = \overline{\sigma}^2/\sqrt{d}$ (see e.g. Baraud, 2002; Blanchard et al., 2018). Thus, we can reliably determine neighbor tasks with $\tau \approx 1/\sqrt{d}$. Based on (III.4), we can hope again for an improvement of order up to $\mathcal{O}(1/\sqrt{d})$ over NE, which is significant even for a moderately large dimension. In the rest of the section, we develop the idea sketched here more precisely and illustrate its consequences on KME by numerical experiments. The message we want to convey is that the *curse* of higher dimensional data with its effect on MSE can be to a limit mitigated by a *relative blessing* because we can take advantage of neighboring tasks more efficiently.

III.2.2 Proposed Approach

Denote $\overline{\sigma}_i^2 = \mathrm{MSE}(i, \widehat{\mu}_i^{\mathrm{NE}}), i \in [\![B]\!]$. Introduce the following notation: $\Delta_{ij} := \|\mu_i - \mu_j\|$. In general, our approach assumes that we have at hand a family of tests $(T_{ij})_{1 \leq i,j \leq B}$ for the null hypotheses $H_{ij}^0 : \Delta_{ij}^2 > \tau\overline{\sigma}_i^2$ against the alternatives $H_{ij}^1 : \Delta_{ij}^2 \leq \tau'\overline{\sigma}_i^2$, for $0 \leq \tau' < \tau$. The exact form of the tests will be discussed later for specific settings.

We denote the set of detected neighbors of task $i \in [\![B]\!]$ as $V_i := \{j : T_{ij} = 1, j \in [\![B]\!]\}$; we can safely assume $T_{ii} = 1$ so that that $i \in V_i$ always holds and $|V_i| \geq 1$. We will also denote $V_i^* = V_i \setminus \{i\}$. For $\gamma \in [0, 1]$, define the modified estimator

$$\widetilde{\mu}_i := \gamma\widehat{\mu}_i^{\mathrm{NE}} + \frac{(1-\gamma)}{|V_i|}\sum_{j \in V_i}\widehat{\mu}_j^{\mathrm{NE}}, \tag{III.5}$$

which can be interpreted as a local shrinkage estimator pulling the naive estimator towards the simple average of its neighbors.

III.3 Theoretical results

We will assume that the naive estimators defined by (III.3) satisfy

$$\max_{i \in [\![B]\!]} \mathrm{MSE}(i, \widehat{\mu}_i^{\mathrm{NE}}) \leq \overline{\sigma}^2. \tag{III.6}$$

Define the notation

$$G(\tau) := \left\{(i,j) \in [\![B]\!]^2 : \Delta_{ij}^2 \leq \tau\overline{\sigma}^2\right\}; \qquad \overline{G}(\tau) := \left\{(i,j) \in [\![B]\!]^2 : \Delta_{ij}^2 \geq \tau\overline{\sigma}^2\right\},$$

and two following events:

$$A(\tau) := \left\{\max_{(i,j) \in \overline{G}(\tau)} T_{ij} = 1\right\}; \qquad B(\tau') := \left\{\min_{(i,j) \in G(\tau')} T_{ij} = 0\right\};$$

so $\mathbb{P}[A(\tau)]$ is the collective false positive rate of the tests (or family-wise error rate) while $\mathbb{P}[B(\tau')]$ is the collective false negative rate to detect $\Delta_{ij}^2 \leq \tau'\overline{\sigma}^2$ (family-wise Type II error rate).

III.3.1 A General Result under Independence of Estimators and Tests

We start with a result assuming that the tests $(T_{ij})_{(i,j)\in[\![B]\!]^2}$ and the estimators $(\widehat{\mu}_i^{\mathrm{NE}})_{i\in[\![B]\!]}$ are independent. This can be achieved for instance by splitting the original bags into two.

Theorem III.1. *Assume model (III.1) holds as well as (III.2), and that (III.6) holds. Furthermore, assume that there exists a family of tests $(T_{ij})_{(i,j)\in[\![B]\!]^2}$ that is independent of $(X_\bullet^{(i)})_{i\in[\![B]\!]}$. For a fixed constant $\tau > 0$, consider the family of estimators $(\widetilde{\mu}_i)_{i\in[\![B]\!]}$ defined by (III.5) with respective parameters*

$$\gamma_i := \frac{\tau|V_i^*|}{(1+\tau)|V_i^*|+1}. \tag{III.7}$$

Then, conditionally to the event $A^c(\tau)$, it holds

$$\forall i \in [\![B]\!] : \mathrm{MSE}(i,\widetilde{\mu}_i) \leq \left(\frac{\tau|V_i^*|+1}{(1+\tau)|V_i^*|+1}\right)\overline{\sigma}^2. \tag{III.8}$$

Let $\mathcal{N}$ denote the covering number of the set of means $\{\mu_j, j\in[\![B]\!]\}$ by balls of radius $\sqrt{\tau'}\overline{\sigma}/2$. Then, conditionally to the events $A^c(\tau)$ and $B^c(\tau')$ (for $\tau' < \tau$), it holds

$$\frac{1}{B}\sum_{i=1}^{B}\mathrm{MSE}(i,\widetilde{\mu}_i) \leq \left(\frac{\tau}{1+\tau} + \frac{\mathcal{N}}{B}\frac{1}{1+\tau}\right)\overline{\sigma}^2. \tag{III.9}$$

The proof can be found in Section III.6. In a nutshell, conditional to the favorable event $A^c(\tau)$, and because the tests are independent of the estimators, we can use the argument leading to (III.4), extended to take into account the shrinkage factor γ, and optimize the value of γ to obtain (III.7), (III.8). If $B^c(\tau')$ is satisfied as well, we can deduce (III.9) directly from (III.8).

Discussion.

- The factor in the individual MSE bound (III.8) is strictly less than 1 as soon as $|V_i| > 1$. As the number of neighbors $|V_i|$ grows, the factor is larger than but approaches $\tau/(1+\tau)$. Therefore, there is a general trade-off between τ and the number of neighbors in a neighborhood of radius $\sqrt{\tau}\overline{\sigma}$. Nevertheless, in order to aim at possibly significant improvement over naive estimation, a small value of τ should be taken.
- The factor in the averaged MSE bound (III.9) is also always smaller than 1 (as expected from the individual MSE bound). It has a nice interpretation in terms of the ratio $\mathcal{N}/B$: if $\mathcal{N} \ll B$, the improvement factor will be very close to $\tau/(1+\tau)$. Thus, we collectively can improve over the naive estimation wrt MSE as soon as the set of means has a small covering number (at scale $\sqrt{\tau'}\overline{\sigma}/2$) in comparison to its cardinality. This condition can be met in different structural low complexity situations, e.g. clustered means, means being sparse vectors, set of means on a low-dimensional manifold. Note that the method does not need information about said structure in advance and is in this sense adaptive to it.

III.3.2 Using the Same Data for Tests and Estimation

We now present a general result in the case where the estimators and tests are not assumed to be independent (e.g. computed from the same data.) To this end we introduce the following additional events:

$$C(\tau) : \left\{\max_{i\neq j}|\langle\widehat{\mu}_i^{\mathrm{NE}} - \mu_i, \widehat{\mu}_j^{\mathrm{NE}} - \mu_j\rangle| > \tau\overline{\sigma}^2\right\}; \qquad C'(\tau) : \left\{\max_i\|\widehat{\mu}_i^{\mathrm{NE}} - \mu_i\|^2 > \overline{\sigma}^2 + \tau\overline{\sigma}^2\right\}.$$

Theorem III.2. *Assume that there exists a family of tests $(T_{ij})_{(i,j)\in[\![B]\!]^2}$. For a given $\tau > 0$ consider the family of estimators $(\widetilde{\mu}_i)_{i\in[\![B]\!]}$ defined by (III.5) with respective parameters*

$$\gamma_i := \frac{\tau}{1+\tau}. \tag{III.10}$$

Then, for $\tau' \geq \tau$, with probability greater than $1 - \mathbb{P}[A(\tau) \cup B(\tau') \cup C(\tau) \cup C'(\tau)]$, it holds

$$\forall i \in [\![B]\!] : \|\widetilde{\mu}_i - \mu_i\|^2 \leq 2\overline{\sigma}^2\left(\tau + \frac{\tau + |V_i|^{-1}}{1+\tau}\right). \tag{III.11}$$

Let $\mathcal{N}$ denote the covering number of the set of means $\{\mu_j, j \in [\![B]\!]\}$ by balls of radius $\sqrt{\tau'}\overline{\sigma}/2$. Then, with the same probability as above, it holds

$$\frac{1}{B}\sum_{i=1}^{B}\|\widetilde{\mu}_i - \mu_i\|^2 \leq 2\overline{\sigma}^2\left(\tau + \frac{\tau}{1+\tau} + \frac{\mathcal{N}}{B}\frac{1}{1+\tau}\right). \tag{III.12}$$

The interpretation of the above result is similar to that of Theorem III.1, with the caveat that the factor in the MSE bound is not always bounded by 1 as earlier; but the qualitative behaviour when τ is small, which is the relevant regime, is the same as previously described.

III.3.3 The Gaussian Setting

In view of the previous results, the crucial point is whether there exists a family of tests such that the events $A(\tau), B(\tau'), C(\tau), C'(\tau)$ have small probability, for a value of τ significantly smaller than 1, and τ' of the same order as τ (up to an absolute numerical constant). This is what we establish now in the Gaussian setting.

Proposition III.3. *Assume (GI) is satisfied. For a fixed $\alpha \in (0,1)$, define the tests*

$$T_{ij} = \mathbf{1}\Big\{\big\|\widehat{\mu}_i^{\mathrm{NE}} - \widehat{\mu}_j^{\mathrm{NE}}\big\|^2 \leq \zeta d/N\Big\}, \tag{III.13}$$

with $\zeta := \left(\sqrt{2+\tau} - 4\sqrt{\delta}\right)^2$, where we put $\delta := (2\log B + \log\alpha^{-1})/d$.

Then, provided $\tau \geq \max(C\delta, \sqrt{C\delta})$ (with $C = 10^3$), it holds $\mathbb{P}[A(\tau)] \leq \alpha$, $\mathbb{P}[B(\tau')] \leq \alpha$ with $\tau' = \tau/3$, $\mathbb{P}[C(\tau)] \leq 2\alpha$ and $\mathbb{P}[C'(\tau)] \leq \alpha$.

The above result is significant in combination with Theorems III.1 and III.2 when δ is small, which is the case if $\log(B)/d$ is small. The message is the following: in a high-dimensional setting, provided $B \ll e^d$, we can reach a large improvement compared to the naive estimators, if the set of means exhibits structure, as witnessed by a small covering number at scale $d^{\frac{1}{4}}\sqrt{(\log B)/N}$. The best-case scenario is when all the means are tightly clustered around a few values, so that $\mathcal{N}$ is small but B is large, then the improvement in the MSE is by a factor of order $\sqrt{(\log B)/d}$.

III.3.4 Methodology and Theory in the Kernel Mean Embedding Framework

We recall that the principle of KME posits a reproducing kernel k on an input space $\mathcal{Z}$, corresponding to a feature mapping $\Phi : \mathcal{Z} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space, with $k(z,z') = \langle\phi(z), \phi(z')\rangle$. The feature mapping ϕ can be extended to *probability distributions* $\mathbb{P}$ on $\mathcal{Z}$, via $\phi(\mathbb{P}) := \mathbb{E}_{Z\sim\mathbb{P}}[\phi(Z)]$, provided this expectation exists, which can be guaranteed for instance if ϕ is bounded. This gives rise to an extended kernel on probability distributions via $k(\mathbb{P},\mathbb{Q}) := \langle\phi(\mathbb{P}), \phi(\mathbb{Q})\rangle = \mathbb{E}_{(Z,Z')\sim\mathbb{P}\otimes\mathbb{Q}}[k(Z,Z')]$.

As explained in the introduction, if we have a large number of distributions $(\mathbb{P}_i)_{i\in[\![B]\!]}$ for each of which an independent bag $(Z_k^{(i)})_{1\leq k\leq N_i}$ is available, and we wish to collectively estimate their KMEs, this is an

instance of the model (III.1)-(III.2) under the transformation $X_k^{(i)} := \phi(Z_k^{(i)})$. The distributions $\mathbb{P}_i$ are replaced by their image distribution through ϕ s.t. $\mu_i = \phi(\mathbb{P}_i)$ and the naive estimators are $\widehat{\mu}_i^{\mathrm{NE}} = \phi(\widehat{\mathbb{P}}_i)$, where $\widehat{\mathbb{P}}_i$ is the empirical measure associated to bag $Z_\bullet^{(i)}$. We will make the assumption that the kernel is bounded, $\sup_{z\in\mathcal{Z}} k(z,z) = \sum_{z\in\mathcal{Z}} \|\phi(z)\|^2 \leq L^2$, resulting in the following "bounded setting":

$$\forall i \in [\![B]\!] : N_i = N \text{ and } \left\|X_k^{(i)}\right\| \leq L, \mathbb{P}_i - \text{a.s.}, k \in [\![N]\!]. \tag{BS}$$

(note in particular that we still assume that all bags have the same size for the theoretical results.)

As always for kernel-based methods, elements of the Hilbert space $\mathcal{H}$ are an abstraction which are never explicitly represented in practice; instead, norms and scalar products between elements, that can be written as linear combinations of sample points, can be computed by straightforward formulas using the kernel. In this perspective, a central object is the *inter-task Gram matrix* K defined as $K_{ij} := k(\mathbb{P}_i, \mathbb{P}_j) = \langle \mu_i, \mu_j \rangle, (i,j) \in [\![B]\!]^2$. In the framework of *inference on distributions*, the distributions $\mathbb{P}_i$ act as (latent) training points and the matrix K as the usual kernel Gram matrix for kernel inference. In contrast to what is assumed in standard kernel inference, K is not directly observed but approximated by $\widehat{K}$ s.t. $\widehat{K}_{ij} := \langle \widehat{\mu}_i, \widehat{\mu}_j \rangle$, for some estimators $(\widehat{\mu}_i)_{i\in[\![B]\!]}$ of the true KMEs. The following elementary proposition links the quality of approximation of the means with the corresponding inter-task Gram matrix:

Proposition III.4. *Assume the model (III.1)-(III.2) under the assumption $\left\|X_k^{(i)}\right\| \leq L$ for all k, i. Let $\widehat{\mu}_i$ be estimators of μ_i bounded by L, and the matrices K and $\widehat{K}$ defined as the Gram matrices of $(\mu_i)_{i\in[\![B]\!]}$ and $(\widehat{\mu}_i)_{i\in[\![B]\!]}$, respectively. Then*

$$\left\| \frac{1}{B}(K - \widehat{K}) \right\|_{\mathrm{Fr.}}^2 \leq \frac{4L^2}{B} \sum_{i\in[\![B]\!]} \|\mu_i - \widehat{\mu}_i\|^2, \tag{III.14}$$

where $\|K\|_{\mathrm{Fr.}} := \mathrm{Tr}(KK^T)^{\frac{1}{2}}$ is the Frobenius norm.

This result further illustrates the interest of improving the task-averaged squared error.

In order to apply our general results Theorems III.1 and III.2, we must again find suitable values of τ (as small as possible) and τ' (as close to τ as possible) so that the probability of the events $A(\tau), B(\tau'), C(\tau), C'(\tau)$ is small, in the setting (BS). In that context, the role of the dimension d will be played by the *effective dimension* $\mathrm{Tr}\,\Sigma / \|\Sigma\|_{op}$, where Σ is the covariance operator for the variable X. More precisely, since this quantity can change from one source distribution to the the other, we will make the following assumption: there exists $d^{\mathrm{e}} > 0$ such that

$$\forall i \in [\![B]\!] : \qquad d^{\mathrm{e}} \|\Sigma_i\|_{\mathrm{op}} \leq \mathrm{Tr}\,\Sigma_i \leq N\overline{\sigma}^2. \tag{III.15}$$

Observe that in view of (III.3), the upper bound above is merely a reformulation of (III.6) and, therefore, not a new assumption; the lower bound is.

We consider tests based on the unbiased estimate of the maximum mean discrepancy (MMD; note that the MMD between tasks i and j is exactly Δ_{ij}^2):

$$U_{ij} = \frac{1}{N(N-1)} \sum_{\substack{k,\ell=1 \\ k\neq\ell}}^{N} \left(\langle X_k^{(i)}, X_\ell^{(i)} \rangle + \langle X_k^{(j)}, X_\ell^{(j)} \rangle \right) - \frac{2}{N^2} \sum_{k,\ell=1}^{N} \langle X_k^{(i)}, X_\ell^{(j)} \rangle.$$

Proposition III.5. *Consider model (III.1), the bounded setting (BS) and assume (III.15) holds. Define*

$$r(t) := 5\left(\sqrt{\left(\frac{1}{d^{\mathrm{e}}} + \frac{L}{N\overline{\sigma}} \right) t} + \frac{Lt}{N\overline{\sigma}} \right), \tag{III.16}$$

and

$$\tau_{\min}(t) := r(t)\max\Big(\sqrt{2}, r(t)\Big). \tag{III.17}$$

For a fixed $t \geq 1$, define the tests T_{ij} for i, j in $[\![B]\!]^2$

$$T_{ij} := \mathbf{1}\big\{U_{ij} < \tau\overline{\sigma}^2/2\big\}. \tag{III.18}$$

Then, provided $\tau \geq 144\tau_{min}(t)$, it holds

$$\mathbb{P}[A(\tau) \cup B(\tau/4) \cup C(\tau/7) \cup C'(\tau/48)] \leq 14B^2e^{-t}\,.$$

The quantity $r(t)$ above (taking $t = \log(14B^2\alpha^{-1})$, where $1-\alpha$ is the target probability) plays a role analogous to δ in the Gaussian setting (Proposition III.3). As the bag size N becomes sufficiently large, we expect $\overline{\sigma} = \mathcal{O}(N^{-\frac{1}{2}})$ and, therefore, $\overline{\sigma}N = \mathcal{O}(N^{\frac{1}{2}})$. Hence, provided N is large enough, the quantity $r(t)$ is mainly of the order $\sqrt{\log(B)/d^{\mathrm{e}}}$. Like in the Gaussian case, this factor determines the potential improvement with respect to the naive estimator, which can be very significant if the effective data dimensionality d^{e} is large.

From a technical point of view, capturing precisely the role of the effective dimension required us to establish concentration inequalities for deviations of sums of bounded vector-valued variables improving over the classical vectorial Bernstein's inequality of Pinelis and Sakhanenko (1986). We believe this result (see Corollary III.10) to be of interest of its own and to have potential other applications.

III.4 Experiments and evaluation

We validate our theoretical results in the KME setting[1] on both synthetic as well as real world data. The neighboring kernel means are determined from the tests as described in Eq. (III.18). More specifically, in practice we use the modification that (i) we adapt the formula for possibly unequal bag sizes, and (ii) in each test T_{ij} we replace $\overline{\sigma}^2$ by the task-dependent unbiased estimate

$$\widehat{\mathrm{MSE}}(i, \widehat{\mu}_i^{\mathrm{NE}}) := \frac{1}{2N_i^2(N_i-1)} \cdot \sum_{k\neq\ell}^{N_i} k(Z_k^{(i)}, Z_k^{(i)}) - 2k(Z_k^{(i)}, Z_\ell^{(i)}) + k(Z_\ell^{(i)}, Z_\ell^{(i)}). \tag{III.19}$$

We analyze three different variations of our method which we call similarity test based (STB) approaches. `STB-0` corresponds to Eq. (III.5) with $\gamma = 0$. `STB weight` uses model optimization to find a suitable value for γ, whereas `STB theory` sets γ as defined in Eq. (III.7). However, here we replaced τ with $c \cdot \tau$, where $c > 0$ is a multiplicative constant, to allow for more flexibility.

We compare their performances to the naive estimation, `NE`, and the regularized shrinkage estimator, `R-KMSE`, (Muandet et al., 2016) which also estimates the KME of each bag separately but shrinks it towards zero. Furthermore, we modified the multi-task averaging approach presented in Feldman et al. (2014) such that it can be used for the estimation of kernel mean embeddings. Similar to our idea, this method shrinks the estimation towards related tasks. However, they require the task similarity to be known. Therefore, we test two options: `MTA const` assumes constant similarity for each bag; `MTA stb` uses the proposed test from Eq. (III.18) to assess the bags for their similarity. See Section III.6.7 for a detailed description of the tested methods.

In the presented results, each considered method has up to two tuning parameters that, in our experiments, are picked in order to optimize averaged test error. Therefore, the reported results can be understood as close to "oracle" performance – the best potential of each method when parameters are close to optimal tuning. While this can be considered unrealistic for practice, a closely related situation can occur in the setting where the user wishes to use the method on test bags of size N, and has at hand a limited number

[1] In the Gaussian setting, we report numerical results in the Section III.6.8.

of training bags of much larger size $N' \gg N$. From each such training bag, one can subsample N points, use the method for estimation of the means of all bags of size N (incl. subsampled bags), and monitor the error with respect to the means of the full training bags (of size N', used as a ground truth proxy). This allows a reasonable calibration of the tuning parameters.

III.4.1 Synthetic Data

The toy data consists of multiple, two-dimensional Gaussian distributed bags $Z_\bullet^{(i)}$ with fixed means but randomly rotated covariance matrices, i.e.

$$Z_\bullet^{(i)} \sim \mathcal{N}\Big(\mathbf{0}, R(\theta_i)\Sigma R(\theta_i)^T\Big) = \mathbb{P}_i \ , \qquad \theta_i \sim \mathcal{U}(-\pi/4, \pi/4),$$

where the covariance matrix $\Sigma = \mathrm{diag}(1, 10)$ is rotated using rotation matrix $R(\theta_i)$ according to angle θ_i. The different estimators are evaluated using the unbiased, squared MMD between the estimation $\widetilde{\mu}_i$ and μ_i as loss. Since μ_i is unknown, it must be approximated by another (naive) estimation $\widehat{\mu}_i^{\mathrm{NE}}(Y_\bullet^{(i)})$ based on independent test bags $Y_\bullet^{(i)}$ from the same distribution as $Z_\bullet^{(i)}$, with $|Y_\bullet^{(i)}| = 1000$. The test bag $Y_\bullet^{(i)}$ has much larger size than the training bag $Z_\bullet^{(i)}$, as a consequence the estimator $\widehat{\mu}_i^{\mathrm{NE}}(Y_\bullet^{(i)})$ has a lower MSE than all considered estimators based on $Z_\bullet^{(i)}$, and can be used as a proxy for the true μ_i.[2] In order to guarantee comparability, all methods use a Gaussian RBF with the kernel width fixed to the average feature-wise standard deviation of the data. Optimal values for the model parameter, e.g. ζ and γ for `STB weight`, are selected such that they minimize the estimation error averaged over 100 trials. Once the values for the parameters are fixed, another 200 trials of data are generated to estimate the final generalization error. Different experimental setups were tested:

(a) **Different Bag Sizes** $B = 50$ and $N_i \in [10, 300]$ for all $i \in [\![B]\!]$,
(b) **Different Number of Bags** $B \in [10, 300]$ and $N_i = 50$ for all $i \in [\![B]\!]$,
(c) **Imbalanced Bags** $B = 50$ and $N_1 = 10, \ldots, N_{50} = 300$,
(d) **Clustered Bags** $N_i, B = 50$ for all $i \in [\![B]\!]$ but the Gaussian distributions are no longer centered around $\mathbf{0}$. Instead, each ten bags form a cluster with the cluster centers equally spaced on a circle. The radius of the circle is varied between 0 and 5, to model different degrees of overlap between clusters.

The results for the experiments on the synthetic data can be found in Figure 1(a) to (d). The estimation of the KME becomes more accurate as the bag size per bag increases. Nevertheless, all of the tested methods provide an increase in estimation performance over the naive estimation, although, the improvement for larger bag sizes decreases for `R-KMSE` and `MTA const`. As expected, methods that use the local neighborhood of the KME yield lower estimation error when the number of available bags increases. Interestingly, this decrease seems to converge towards a capping value, which might reflect the intrinsic dimensionality of the data as indicated by Theorems III.1 and III.2 combined with Proposition III.5. Although we assumed equal bag sizes in the theoretical results, the proposed approaches provide accurate estimations also for the imbalanced setting. Figure 1(c) shows that the improvement is most significant for bags with few samples, which is consistent with results on other multi-task learning problems (see e.g. Feldman et al., 2014). However, when the KME of a bag with many samples is shrunk towards a neighbor with few samples, the estimation can be deteriorated (compare results on (a) with those on (c) for large bag sizes). A similar effect can be seen in the results on the clustered setting. When the bags overlap, a bag from a different cluster might be considered as neighbor which leads to a stronger estimation bias. When the tasks have similar centers or are strictly separated, the methods show similar performance to what is shown in Figure 1(b).

[2] Additionally, the estimation of the squared loss is unbiased if the diagonal entries of the Gram matrix will be included for $Z_\bullet^{(i)}$ but excluded for $Y_\bullet^{(i)}$.

To summarize, NE and R-KMSE give worst performances because they estimate the kernel means separately. Even though MTA const assumes all tasks to be related, it improves the estimation performance even when the bags are not similar. However, the methods that derive the task similarity from the local neighborhood achieve most accurate KME estimations in all of the tested scenarios, especially STB weight and STB theory.

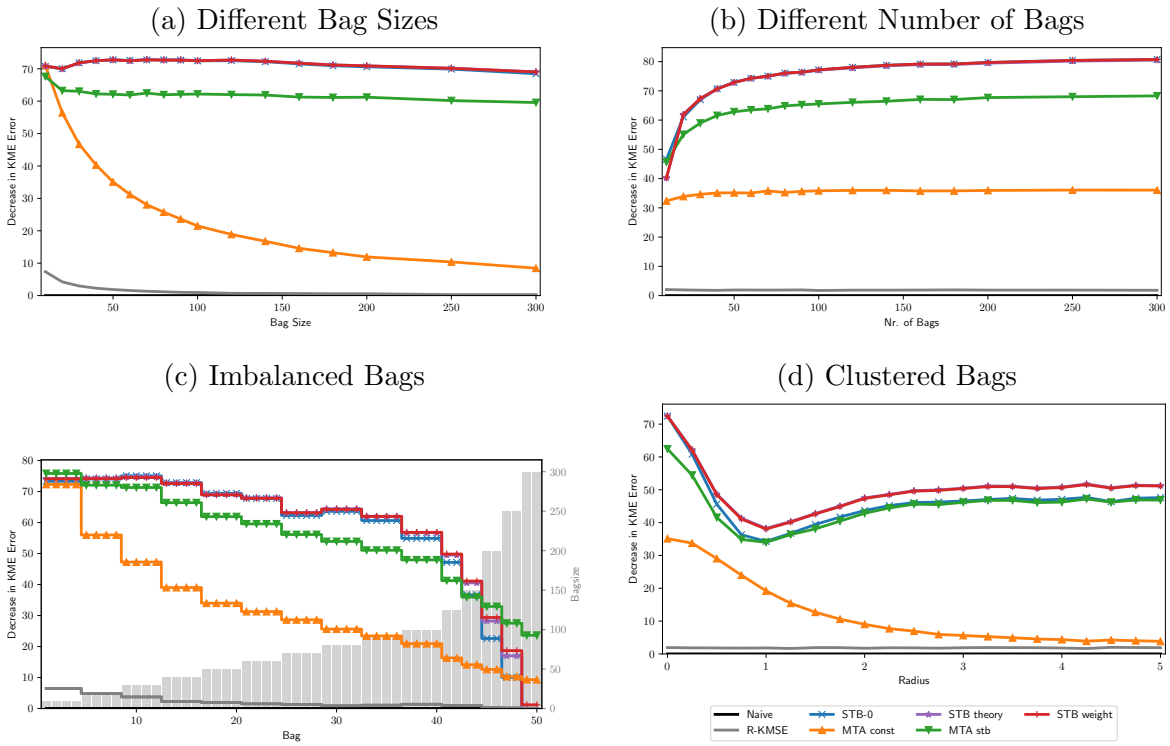

Figure 1: Decrease in KME estimation error compared to NE in percent on experimental setups (a) to (d). Higher is better. STB-0, STB weight and STB theory give similar results so that their results might be printed on top of each other.

III.4.2 Real World Data

We test our methods on a remote sensing data set. The AOD-MISR1 data set is a collection of 800 bags with each 100 samples. The samples correspond to randomly selected pixels from a MISR satellite, where each instance is formed by 12 reflectances from three MISR cameras.[3] It can be used to predict the aerosol optical depth (AOD) which poses an important problem in climate research (Wang et al., 2011).

The data is standardized such that each of the features has unit standard deviation and is centered around zero. In each out of the 100 trials, we randomly subsample 20 samples from each bag, on which the KME estimation is based. This estimation is then compared to the naive estimation on the complete bag. Cross-validation, with 400 bags for training and testing, is used to optimize for the model parameters of each approach and then estimate its error. Again, all methods use a Gaussian RBF with the kernel width fixed to one. The results are shown in Table 3.

Again, all of the methods provide a more accurate estimation of the KME than the naive approach. The estimations given by STB-0 are similar to those of NE, because STB-0 considers very few bags as neighbors. This lets us conclude that the bags are rather isolated than overlapping. MTA stb, STB weight and STB

[3] We only use 12 out of 16 features because the remaining four are constant per bag.

Table 3: Decrease in KME estimation error compared to `NE` in percent on the AOD-MISR1 data.

METHOD	%	METHOD	%	METHOD	%
`R-KMSE`	8.83	`MTA const`	13.92	`STB theory`	21.83
`STB-0`	1.43	`MTA stb`	17.17	`STB weight`	22.73

`theory` might give better estimations because they allow for more flexible shrinkage. Again, `STB weight` and `STB theory` are outperforming the remaining methods.

III.5 Conclusion

In this section we proposed an improved estimator for the multi-task averaging problem. The estimation is improved by shrinking the naive estimation towards the average of its neighboring means. The neighbors of a task are found by multiple testing so that task similarities must not be known a priori. Provided that appropriate tests exist, we proved that the introduced shrinkage approach yields a lower mean squared error for each task individually and also on average. We show that there exists a family of statistical tests suitable for isotropic Gaussian distributed data or for means that lie in a reproducing kernel Hilbert space. Theoretical analysis shows that this improvement can be especially significant when the (effective) dimension of the data is large, using the property that the typical detection radius of the tests is much better than the standard estimation error in high dimension. This property is particularly important for the estimation of multiple kernel mean embeddings (KME) which is an interesting application relevant for the statistical and machine learning community. The proposed estimator and the theoretical results can naturally be translated to the KME framework.

We tested different variations of the presented approach on synthetic and real world data and compared its performance to other state-of-the-art methods. In all of the conducted experiments, the proposed shrinkage estimators yield the most accurate estimations.

Since the estimation of a KME is often only an intermediate step for solving a final task, as for example in distributional regression (Szabó et al., 2016), further effort must be made to assess whether the improved estimation of the KME also leads to a better final prediction performance. Furthermore, the results on the imbalanced toy data sets have shown that the shrinkage estimator particularly improves the estimation of small bags. However, when the KME of a bag with many samples is shrunk towards a neighbor with low bag size, its estimation might be distorted. Therefore, we develop in following sections a similarity test and a weighting scheme that take the bag size into account. From a theoretical perspective, we also investigate in Section V if the improvement factor with respect to the naive estimates is optimal in a suitable minimax sense.

III.6 Appendix of Section III

III.6.1 Proof of Theorem III.1

We argue conditional to the tests, below expectations are taken with respect to the samples $(X_\bullet^{(b)})_{b\in[\![B]\!]}$ only. Assume the event $A^c(\tau)$ holds, implying for all i:

$$j \in V_i \Rightarrow \Delta_{ij}^2 \leq \tau\sigma^2. \tag{III.20}$$

Take $i = 1$ without loss of generality, and denote $V = V_1, V^* = V_1 \setminus \{1\}$, and $v = |V_1|$. We also put $\eta = 1-\gamma$. We use an argument similar to that leading to (III.4) using independence of the bags, triangular

inequality and (III.20):

$$\begin{aligned}
\mathrm{MSE}(1,\widetilde{\mu}_1) &= \mathbb{E}\left[\left\|(1-\eta)(\widehat{\mu}_1^{\mathrm{NE}}-\mu_1)+\frac{\eta}{v}\sum_{j\in V}(\widehat{\mu}_i^{\mathrm{NE}}-\mu_1)\right\|^2\right] \\
&= \frac{\eta^2}{v^2}\left(\left\|\sum_{i\in V^*}(\mu_i-\mu_1)\right\|^2+\sum_{i\in V^*}\mathbb{E}\left[\left\|\mu_i-\widehat{\mu}_i^{\mathrm{NE}}\right\|^2\right]\right) \\
&\quad + (1-\eta(1-v^{-1}))^2\mathbb{E}\left[\left\|\widehat{\mu}_1^{\mathrm{NE}}-\mu_1\right\|^2\right] \\
&\le \sigma^2\Big(\frac{\eta^2}{v^2}\big((v-1)^2\tau+(v-1)\big)+(1-\eta(1-v^{-1}))^2\Big) \\
&= \sigma^2\big(\eta^2(1-v^{-1})\big((1-v^{-1})\tau+1\big)-2\eta(1-v^{-1})+1\big).
\end{aligned}$$

The optimal value of $\gamma = 1-\eta$ is given by (III.7) and gives rise to (III.8).

Assume additionally that $B^c(\tau')$ holds. Let $\varepsilon := \sqrt{\tau'}\sigma/2$ and let $\mathcal{C} := \{x_1,\dots,x_{\mathcal{N}}\}$ be an ε-covering of the set of means. Let $\pi(i)$ be the index of the element of $\mathcal{C}$ closest to μ_i, and $N_k := \{b\in[\![B]\!] : \pi(b)=k\}$, $i\in[\![\mathcal{N}]\!]$. By the triangular inequality, for any $i\in[\![\mathcal{N}]\!]$, $b\in N_i$ one has $|V_b|\ge|N_{\pi(i)}|$. Hence averaging (III.8) over i we get

$$\begin{aligned}
\frac{1}{B}\sum_{b=1}^{B}\mathrm{MSE}(b,\widetilde{\mu}_b) &\le \frac{\sigma^2}{B}\sum_{i\in[\![B]\!]}\frac{\tau(|N_{\pi(i)}|-1)+1}{1+(1+\tau)(|N_{\pi(i)}|-1)} \\
&= \frac{\sigma^2}{B}\sum_{k\in[\![\mathcal{N}]\!]}\frac{|N_k|(\tau(|N_k|-1)+1)}{1+(1+\tau)(|N_k|-1)}.
\end{aligned}$$

The above take the form $\sum_k f(|N_k|)$, and it is straightforward to check that f is convex. Since it holds $1\le|N_k|\le B-\mathcal{N}+1$ for all k, and $\sum_{k\in[\![\mathcal{N}]\!]}|N_k|=B$, the maximum of the above expression is attained for an extremal point of this convex domain, i.e., by symmetry, $N_1 = B-\mathcal{N}+1$ and $N_k=1$ for $k\ge 2$. Therefore

$$\begin{aligned}
\frac{1}{B}\sum_{b=1}^{B}\mathrm{MSE}(b,\widetilde{\mu}_b) &\le \frac{\sigma^2}{B}\left((\mathcal{N}-1)+\frac{(B-\mathcal{N}+1)((B-\mathcal{N})\tau+1)}{(B-\mathcal{N})(1+\tau)+1}\right) \\
&= \frac{\sigma^2}{B}\left(\mathcal{N}+\frac{(B-\mathcal{N})^2\tau}{(B-\mathcal{N})(1+\tau)+1}\right) \\
&\le \sigma^2\left(\frac{\tau}{\tau+1}+\frac{\mathcal{N}}{B}\frac{1}{\tau+1}\right).
\end{aligned}$$

□

III.6.2 Proof of Theorem III.2

We follow the same general line as in theorem III.1. Assume the event $A^c(\tau)\cap B^c(\tau')\cap C^c(\tau)\cap C'^c(\tau)$ holds. Take $i=1$ without loss of generality, and denote $V = V_1, V^* = V_1\setminus\{1\}$, and $v=|V_1|$. We still put $\eta = 1-\gamma$. Then

$$\begin{aligned}
\|\widetilde{\mu}_1-\mu_1\|^2 &= \left\|(1-\eta)(\widehat{\mu}_1^{\mathrm{NE}}-\mu_1)+\frac{\eta}{v}\sum_{j\in V}(\widehat{\mu}_j^{\mathrm{NE}}-\mu_1)\right\|^2 \\
&\le 2\left(\left\|(1-\eta(1-v^{-1}))(\widehat{\mu}_1^{\mathrm{NE}}-\mu_1)+\frac{\eta}{v}\sum_{j\in V^*}(\widehat{\mu}_j^{\mathrm{NE}}-\mu_j)\right\|^2+\frac{\eta^2}{v^2}\left\|\sum_{j\in V^*}\mu_j-\mu_1\right\|^2\right).
\end{aligned}$$

Let us upper bound the different terms. Because $j \in V$, we know that $\Delta_{j1} \leq \tau\overline{\sigma}^2$, so by the triangular inequality

$$\frac{\eta}{v}\left\| \sum_{j\in V^*} \mu_j - \mu_1 \right\| \leq \frac{\eta}{v} \sum_{j\in V^*} \|\Delta_{ij}\| \leq \eta(1-v^{-1})\sqrt{\tau}\overline{\sigma}.$$

Let us develop the other term :

$$\begin{aligned}
&\left\|(1-\eta(1-v^{-1}))(\widehat{\mu}_1^{\mathrm{NE}} - \mu_1) + \frac{\eta}{v}\sum_{j\in V^*}(\widehat{\mu}_j^{\mathrm{NE}} - \mu_j)\right\|^2 \\
&\quad = (1-\eta(1-v^{-1}))^2\|\widehat{\mu}_1^{\mathrm{NE}} - \mu_1\|^2 + \frac{2\eta(1-\eta(1-v^{-1}))}{v}\sum_{j\in V^*}\langle \widehat{\mu}_1^{\mathrm{NE}} - \mu_1, \widehat{\mu}_j^{\mathrm{NE}} - \mu_j\rangle \\
&\qquad + \frac{\eta^2}{v^2}\sum_{j\neq k\in V^*}\langle \widehat{\mu}_j^{\mathrm{NE}} - \mu_j, \widehat{\mu}_k^{\mathrm{NE}} - \mu_k\rangle + \frac{\eta^2}{v^2}\sum_{j\in V^*}\|\widehat{\mu}_j^{\mathrm{NE}} - \mu_j\|^2 \\
&\quad \leq \overline{\sigma}^2\Big[(1-\eta(1-v^{-1}))^2(1+\tau) + 2\eta(1-\eta(1-v^{-1}))(1-v^{-1})\tau \\
&\qquad\qquad + \eta^2(1-v^{-1})^2\tau + \eta^2 v^{-1}(1-v^{-1})(1+\tau)\Big].
\end{aligned}$$

Let us associate the two expressions, we obtain that :

$$\|\widetilde{\mu}_1 - \mu_1\|^2 \leq 2\overline{\sigma}^2\Big[\tau + 1 - 2(1-v^{-1})\eta + (1-v^{-1})(1+\tau)\eta^2\Big].$$

The expression is minimal when $\eta = (1+\tau)^{-1}$. By the same arguments about using covering numbers as in the proof of Theorem III.1, we obtain that with probability greater than $1-\mathbb{P}[A(\tau)\cup B(\tau')\cup C(\tau)\cup C'(\tau)]$:

$$\begin{aligned}
\frac{1}{B}\sum_{i\in[\![B]\!]}\|\widetilde{\mu}_i - \mu_i\|^2 &\leq \frac{2\overline{\sigma}^2}{B}\sum_{i\in[\![B]\!]}\tau + \frac{\tau + |V_i|^{-1}}{1+\tau} \\
&\leq 2\overline{\sigma}^2\left(\tau + \frac{\tau}{1+\tau} + \frac{\mathcal{N}}{B}\frac{1}{1+\tau}\right).
\end{aligned}$$

□

III.6.3 Proof of Proposition III.3

Recall that we assume the (GI) model. We first consider the behavior of a single test

$$T_{ij} = \mathbf{1}\Big\{\big\|\widehat{\mu}_i^{\mathrm{NE}} - \widehat{\mu}_j^{\mathrm{NE}}\big\|^2 \leq \zeta\overline{\sigma}^2\Big\},$$

where $\overline{\sigma}^2 := d/N$, we also put $\Delta^2 = \Delta_{ij}^2$ for short. The random variable $Z := \widehat{\mu}_i^{\mathrm{NE}} - \widehat{\mu}_j^{\mathrm{NE}}$ is distributed as $\mathcal{N}(\mu_i - \mu_j, 2n^{-1}I_d)$ by independence of the bags. From classical concentration results for chi-squared variables recalled as Proposition III.6 in Section III.6.5, for any $\alpha \in (0,1)$ either of the inequalities below hold with probability $1-\alpha$:

$$\sqrt{\Delta^2 + 2\overline{\sigma}^2} - 4\overline{\sigma}\sqrt{\frac{\log\alpha^{-1}}{d}} \leq \|Z\| \leq \sqrt{\Delta^2 + 2\overline{\sigma}^2} + 2\overline{\sigma}\sqrt{\frac{\log\alpha^{-1}}{d}}. \tag{III.21}$$

Put $\delta := (\log\alpha^{-1})/d$ for short.

We start with analyzing Type I error: if $\Delta^2 \geq \tau\overline{\sigma}^2$, then the above lower bound implies $\|Z\|^2 \geq \overline{\sigma}^2\big(\sqrt{2+\tau} - 4\sqrt{\delta}\big)^2$, so $T_{ij} = 0$ if we choose $\zeta := \big(\sqrt{2+\tau} - 4\sqrt{\delta}\big)^2$. By union bound over $(i,j) \in [\![B]\!]^2$, with this choice we guarantee that $\mathbb{P}[A(\tau)] \leq \alpha$ if we replace α by αB^2 (i.e. take $\delta = (2\log B + \log\alpha^{-1})/d$). This establishes the bound on family-wise type I error.

We now analyze type II error: assume now that we have picked $\zeta := \big(\sqrt{2+\tau} - 4\sqrt{\delta}\big)^2$, with $\tau \geq \max(C\delta, \sqrt{C\delta})$, $C = 1000$, and assume $\Delta^2 \leq \tau'\overline{\sigma}^2$. Then assuming the upper bound in (III.21) is satisfied, we ensure $T_{ij} = 1$ provided

$$\sqrt{\tau' + 2} \leq \sqrt{\tau + 2} - 6\sqrt{\delta}.$$

Note that the condition on τ ensures that the above right-hand-side is positive. Taking squares and further bounding, a sufficient condition for the above is $\tau' \leq \tau - 12\sqrt{(2+\tau)\delta}$. Using the condition on τ, it holds

$$12\sqrt{(2+\tau)\delta} \leq 12\sqrt{3C^{-1}}\tau \leq \frac{2}{3}\tau,$$

hence $\tau' \leq \tau/3$ is a sufficient condition. This ensures, by the union bound, that $\mathbb{P}[B(\tau')] \leq \alpha$ when replacing δ by $\delta' = (2\log B + \log \alpha^{-1})/d$ as above.

We now turn to controlling the probability of the events $C(\tau)$ and $C'(\tau)$. For fixed i, j put $X_1 = \widehat{\mu}_i^{\mathrm{NE}} - \mu_i$, $X_2 = \widehat{\mu}_j^{\mathrm{NE}} - \mu_j$. Under the (GI) model, X_1, X_2 are independent $\mathcal{N}(0, N^{-1}I_d)$. Applying the result of Proposition III.7, we obtain that for $\alpha \in (0,1)$, we have probability at least $1 - 2\alpha$:

$$|\langle X_i, X_j\rangle| \leq \overline{\sigma}^2\big(\sqrt{2\delta} + \delta\big),$$

where we have put $\delta := (\log \alpha^{-1})/d$ as previously. As soon as $\tau \geq \max(C\delta, \sqrt{C\delta})$, $(C \geq 1)$ we obtain $|\langle X_i, X_j\rangle| \leq 3\tau\overline{\sigma}^2/\sqrt{C}$ on the above event, implying that the event $C(\tau)$ is a fortiori satisfied for $C = 10^3$.

From estimate (III.24) in Proposition III.6, we have with probability at least $1 - \alpha$:

$$\|X_1\| \leq \overline{\sigma}^2\big(1 + \sqrt{2\delta}\big) \leq \overline{\sigma}^2\big(1 + 2\tau C^{-1}\big),$$

under the same condition on τ as above. As previously, by the union bound the above estimates are true simultaneously for all i, j with the indicated probabilities if we replace δ by $\delta' = (2\log B + \log \alpha^{-1})/d$, and $C'(\tau)$ is satisfied when taking $C = 10^3$. □

III.6.4 Results in the Bounded Setting (for KME Estimation)

Proof of Proposition III.4

$$\begin{aligned}
\big\|(K - \widehat{K})\big\|_{\mathrm{Fr.}}^2 &= \sum_{(i,j)\in[\![B]\!]^2} \big(\langle \mu_i, \mu_j\rangle - \langle \widehat{\mu}_i, \widehat{\mu}_j\rangle\big)^2 \\
&= \sum_{(i,j)\in[\![B]\!]^2} \big(\langle \mu_i - \widehat{\mu}_i, \mu_j\rangle + \langle \widehat{\mu}_i, \mu_j - \widehat{\mu}_j\rangle\big)^2 \\
&\leq 2 \sum_{(i,j)\in[\![B]\!]^2} \big(\langle \mu_i - \widehat{\mu}_i, \mu_j\rangle^2 + \langle \widehat{\mu}_i, \mu_j - \widehat{\mu}_j\rangle^2\big) \\
&\leq 2L^2 \sum_{(i,j)\in[\![B]\!]^2} \big(\|\mu_i - \widehat{\mu}_j\|^2 + \|\mu_j - \widehat{\mu}_j\|^2\big) \\
&\leq 4L^2 B \sum_{i\in[\![B]\!]} \|\mu_i - \widehat{\mu}_j\|^2.
\end{aligned}$$

□

Proof of Proposition III.5 Recall the notation

$$r(t) = 5\left(\sqrt{\left(\frac{1}{d^{\mathrm{e}}} + \frac{L}{N\overline{\sigma}}\right)t} + \frac{Lt}{N\overline{\sigma}}\right), \tag{III.22}$$

and

$$\tau_{\min}(t) := r(t)\max\Big(\sqrt{2}, r(t)\Big). \tag{III.23}$$

Introduce the notation $q(t) := \overline{\sigma} r(t)$; $\xi(t) := \overline{\sigma}^2 \tau_{\min}(t) = q(t)\max(\sqrt{2}\overline{\sigma}, q(t))$. Let $i, j \in [\![B]\!]^2$ be fixed and $t \geq 1$. We put $\tau = \lambda^2 \tau_{\min}(t)$ with $\lambda \geq 12$.

Suppose that $\|\Delta_{ij}\|^2 > \tau\overline{\sigma}^2 = \lambda^2 \tau_{\min}\overline{\sigma}^2 = \lambda^2 \xi(t)$. We use the concentration inequality (III.40) for bounded variables, proved in Section III.6.6, and obtain that with probability greater than $1 - 8e^{-t}$, and using the definition of $\xi(t)$:

$$U_{ij} \geq \|\Delta_{ij}\|^2 - 2\|\Delta_{ij}\| q(t) - 8\sqrt{2\overline{\sigma}^2} q(t) - 32 q^2(t) \geq \|\Delta_{ij}\|\Big(\|\Delta_{ij}\| - 2q(t)\Big) - 40\xi(t)\,.$$

(To be more precise, (III.40) proves the above estimate for the value of $q(t)$ defined by (III.37), the value of $q(t)$ defined in the present proof is an upper bound for it, so the above also holds.)

Observe $\|\Delta_{ij}\| \geq \lambda\sqrt{\xi(t)} \geq 12\sqrt{\xi(t)} \geq 2q(t)$. By monotonicity in $\|\Delta_{ij}\|$ under that condition, it holds $\|\Delta_{ij}\|\Big(\|\Delta_{ij}\| - 2q(t)\Big) \geq \sqrt{\lambda\xi(t)}(\lambda\sqrt{\xi(t)} - 2q(t)) \geq \lambda(\lambda-2)\xi(t)$. That leads to

$$U_{ij} \geq (\lambda^2 - 2\lambda - 40)\xi(t) \geq (\lambda^2/2)\xi(t) = (\tau/2)\overline{\sigma}^2,$$

where we have used that $\lambda^2 - 2\lambda - 40 \geq \lambda^2/2$ for $\lambda \geq 12$. So

$$\mathbb{P}\big[\|\Delta_{ij}\|^2 > \tau\overline{\sigma}^2 \quad \text{and} \quad T_i = 1\big] \leq 8e^{-t}.$$

Suppose now $\|\Delta_{ij}\|^2 < (\tau/4)\overline{\sigma}^2 = (\lambda^2/4)\xi(t)$. Then, according to the concentration ineqality (III.39), with probability greater than $1 - 8e^{-t}$, it holds

$$\begin{aligned} U_{ij} &\leq \|\Delta_{ij}\|^2 + 2\|\Delta\| q(t) + 2\sqrt{2\overline{\sigma}^2} q(t) + 11 q^2(t) \\ &\leq \Big(\lambda^2/4 + \lambda + 13\Big)\xi(t) \\ &\leq (\lambda^2/2)\xi(t) = (\tau/2)\overline{\sigma}^2. \end{aligned}$$

We have used that $\lambda^2/4 + \lambda + 13 \leq \lambda^2/2$ for $\lambda \geq 12$. So

$$\mathbb{P}\big[\|\Delta_{ij}\|^2 < \tau\overline{\sigma}^2/4 \quad \text{and} \quad T_i = 0\big] \leq 2e^{-t}.$$

An union bound over $(i,j) \in [\![B]\!]^2$ gives that

$$\mathbb{P}[A(\tau) \cup B(\tau/4)] \leq 8B^2 e^{-t}\,.$$

Remarking that

$$\overline{\sigma}^2\tau/7 \geq 20 q(t)\max(q(t), \sqrt{2\overline{\sigma}^2}) \quad \text{and} \quad \overline{\sigma}^2\tau/48 \geq 3q(t)\max(q(t), \sqrt{2\overline{\sigma}^2}) \geq 2q(t)\sqrt{2\overline{\sigma}^2} + q^2(t)$$

and using the concentration inequalities (III.30) and (III.36) gives

$$\mathbb{P}[C(\tau/7)] \leq 6(B^2 - B)e^{-t}\,, \quad \text{and} \quad \mathbb{P}[C'(\tau/48)] \leq Be^{-t}\,.$$

□

III.6.5 Concentration Results in the Gaussian Setting

Proposition III.6. *Let Z be a normal $\mathcal{N}(\mu, \sigma^2 I_d)$ random variable in $\mathbb{R}^d$. Then for any $t \geq 0$:*

$$\mathbb{P}\left[\|Z\| \geq \sqrt{\|\mu\|^2 + \sigma^2 d} + \sigma\sqrt{2t}\right] \leq e^{-t}\,, \tag{III.24}$$

and

$$\mathbb{P}\left[\|Z\| \leq \sqrt{\|\mu\|^2 + \sigma^2 d} - 2\sigma\sqrt{2t}\right] \leq e^{-t}. \tag{III.25}$$

Proof. The stated inequalities are direct consequences of classical deviation inequalities for (noncentral) χ^2 variables. Put $\lambda := \|\mu\|^2$, then for the upper deviation bound, Lemma 8.1 of Birgé (2001) states that

$$\mathbb{P}\Big[\|Z\|^2 \geq \lambda + d\sigma^2 + 2\sqrt{(2\lambda + d\sigma^2)\sigma^2 t} + 2\sigma^2 t\Big] \leq e^{-t}\,,$$

and we have

$$\lambda + d\sigma^2 + 2\sqrt{(2\lambda + d\sigma^2)\sigma^2 t} + 2\sigma^2 t \leq \Big(\sqrt{\lambda + d\sigma^2} + \sigma\sqrt{2t}\Big)^2\,,$$

implying (III.24). For the lower deviation bound, Lemma 8.1 of Birgé (2001) states that

$$\mathbb{P}\Big[\|Z\|^2 \leq \lambda + d\sigma^2 - 2\sqrt{(2\lambda + d\sigma^2)\sigma^2 t}\Big] \leq e^{-t}\,,$$

and we have

$$\Big(\lambda + d\sigma^2 - 2\sqrt{(2\lambda + d\sigma^2)\sigma^2 t}\Big)_+ \geq \sqrt{\lambda + d\sigma^2}\Big(\sqrt{\lambda + d\sigma^2} - 2\sigma\sqrt{2t}\Big)_+ \geq \Big(\sqrt{\lambda + d\sigma^2} - 2\sigma\sqrt{2t}\Big)_+^2,$$

leading to (III.25). □

Proposition III.7. *Let X_1, X_2 be independent $\mathcal{N}(0, \sigma^2 I_d)$ variables in dimension d. Then for any $t \geq 0$:*

$$\mathbb{P}\Big[\langle X_1, X_2\rangle \geq \sigma^2\Big(\sqrt{2dt} + t\Big)\Big] \leq e^{-t}. \tag{III.26}$$

Proof. Without loss of generality assume $\sigma^2 = 1$. For two independent one-dimensional Gaussian variables G_1, G_2, one has for any $\lambda \in [0, 1]$:

$$\mathbb{E}[\exp \lambda G_1 G_2] = \mathbb{E}[\mathbb{E}[\exp \lambda G_1 G_2 | G_2]] = \mathbb{E}\left[\exp \frac{\lambda^2}{2} G_2^2\right] = \frac{1}{\sqrt{1 - \lambda^2}}\,,$$

so that

$$\log \mathbb{E}[\exp \lambda \langle X_1, X_2\rangle] = \frac{d}{2}(-\log(1 - \lambda^2)) \leq \frac{d}{2}\frac{\lambda^2}{(1 - \lambda)}\,.$$

Applying Lemma 8.2 of Birgé (2001) gives (III.26). □

III.6.6 Concentration Results in the Bounded Setting

Studying concentration in the kernel setting means having concentration results of bounded variables taking values in a separable Hilbert space. Recall that $k(x, y) = \langle \phi(z), \phi(z')\rangle_{\mathcal{H}}$ for all z,z' in $\mathcal{H}$, so that if k is bounded by L^2, then the map ϕ is bounded by L. To obtain concentration results, we will use Talagrand's inequality.

Theorem III.8 (Talagrand's inequality)**.** *Let $X_1^s, ..., X_N^s$ be iid real random variables indexed by $s \in S$ where S is a countable index set, and L be a positive constant such that:*

$$\mathbb{E}[X_k^s] = 0\,, \quad \textit{and} \quad |X_k^s| \leq L \ a.s. \quad \forall k \in [\![N]\!],\, s \in S\,.$$

Let us note $Z = \sup_{s\in S} \sum_{k=1}^N X_k^s$, then for all $t \geq 0$:

$$\mathbb{P}\Big[Z - \mathbb{E}[Z] \geq 2\sqrt{(2v + 16L\mathbb{E}[Z])t} + 2Lt\Big] \leq e^{-t};$$
$$\mathbb{P}\Big[-Z + \mathbb{E}[Z] \geq 2\sqrt{(4v + 32L\mathbb{E}[Z])t} + 4Lt\Big] \leq e^{-t},$$

where $v = \sup_{s\in S} \sum_{k=1}^N \mathbb{E}\big[(X_k^s)^2\big]$.

Talagrand's inequality appeared originally in Talagrand, 1996, with the above form (using additional symmetrization and contraction arguments from Ledoux and Talagrand, 1991) appearing in Massart, 2000. The constants in the upper deviation bound have been improved by Rio, 2002 and Bousquet, 2002, however no such improvement is available for lower devations as far as we know. The above version is taken from Massart, 2007 p. 169–170, (5.45) and (5.46) combined with (5.47) there.

Because a Hilbertian norm can be viewed as a supremum, we can use Talagrand's inequality to obtain a concentration inequality for the norm of the sum of bounded Hilbert-valued random variables.

Proposition III.9. *Let $(Z_k)_{1\leq k\leq N}$ be i.i.d. random variables taking values in a separable Hilbert space $\mathcal{H}$, whose norm is bounded by L a.s. Let μ and Σ denote their common mean and covariance operator. Let*

$$V = \left\| \frac{1}{N}\sum_{k=1}^{N} Z_k \right\|, \quad \text{and} \quad V_c = \left\| \frac{1}{N}\sum_{k=1}^{N} Z_k - \mu \right\|.$$

Then for any $t \geq 0$:

$$\mathbb{P}\Big[V^2 \geq \|\mu\|^2 + \big(\mathbb{E}[V_c] + q_\Sigma(t)\big)^2 + 2\|\mu\| q_\Sigma(t)\Big] \leq 2e^{-t}\,, \tag{III.27}$$

and

$$\mathbb{P}\Big[V^2 \leq \|\mu\|^2 + \big(\mathbb{E}[V_c] - 2q_\Sigma(t)\big)_+^2 - 2\|\mu\| q_\Sigma(t)\Big] \leq 2e^{-t}\,, \tag{III.28}$$

where

$$q_\Sigma(t) = 2\sqrt{\left(\frac{2\|\Sigma\|_{\mathrm{op}}}{N} + 16L\frac{\sqrt{\operatorname{Tr}\Sigma}}{N^{3/2}}\right)t} + \frac{2L}{N}t. \tag{III.29}$$

Proof. Let us denote $q(t)$ for $q_\Sigma(t)$ for this proof. We start with bounding the deviations of V_c. Observe that

$$V_c = \sup_{\|u\|_{\mathcal{H}}=1} \frac{1}{N}\sum_{k=1}^{N} \langle u, Z_k - \mu\rangle,$$

where the supremum can be restricted to u in a dense countable subset $\mathcal{S}$ of the unit sphere, since $\mathcal{H}$ is separable. We can therefore apply Talagrand's inequality with $X_k^u := N^{-1}\langle u, Z_k - \mu\rangle$; it holds $|X_k^u| \leq L/N$, and note that since $\Sigma = \mathbb{E}[(Z-\mu)\otimes(Z-\mu)^*]$, it holds

$$\mathbb{E}\big[(X_k^u)^2\big] = N^{-2}\mathbb{E}\Big[\langle u, Z_k - \mu\rangle^2\Big] = N^{-2}\langle u, \Sigma u\rangle,$$

so that $\sup_{u\in\mathcal{S}}\sum_{k=1}^{N}\mathbb{E}\big[(X_k^u)^2\big] = N^{-1}\|\Sigma\|_{\mathrm{op}}$. Furthermore, $\mathbb{E}[V_c] \leq N^{-\frac{1}{2}}\sqrt{\operatorname{Tr}\Sigma}$ by Jensen's inequality, which we use to further bound the deviation term by $q(t)$.

By Theorem III.8, with probability greater than $1-e^{-t}$ for $t \geq 0$, it holds

$$V_c \leq \mathbb{E}[V_c] + q(t)\,, \tag{III.30}$$

and with probability greater than $1-e^{-t}$,

$$V_c \geq \mathbb{E}[V_c] - 2q(t)\,. \tag{III.31}$$

We turn to bounding the deviations of $V^2 - \|\mu\|^2$. Observe

$$V^2 - \|\mu\|^2 = V_c^2 + \frac{2}{N}\sum_{k=1}^{N}\langle Z_k - \mu, \mu\rangle. \tag{III.32}$$

Using Bernstein's inequality for the variables $W_i = \langle Z_i - \mu, \mu \rangle$, satifying $\mathbb{E}[W_i] = 0$, $\mathbb{E}\big[W_i^2\big] = \langle \mu, \Sigma\mu \rangle \leq \|\Sigma\|_{\mathrm{op}} \|\mu\|^2$, and $|W_i| \leq L\|\mu\|$, we have that with probability greater than $1 - e^{-t}$, for $t \geq 0$:

$$\frac{1}{N}\sum_{i=1}^{N} \langle Z_i - \mu, \mu \rangle \leq \|\mu\| \left[\sqrt{\frac{2\|\Sigma\|_{\mathrm{op}} t}{N}} + \frac{4Lt}{3N} \right] \leq \|\mu\| q(t)\,. \tag{III.33}$$

Combining inequality (III.33) with (III.32) and (III.30) gives that with probability greater than $1-2e^{-t}$:

$$V^2 - \|\mu\|^2 \leq \left(\mathbb{E}[V_c] + q(t)\right)^2 + 2\|\mu\| q(t)\,,$$

and, combining (III.33), (III.32) and (III.31), we have with probability greater than $1 - 2e^{-t}$:

$$V^2 - \|\mu\|^2 \geq \left(\mathbb{E}[V_c] - 2q(t)\right)_+^2 - 2\|\mu\| q(t)\,.$$

□

Corollary III.10. *Using the setting and notation of Proposition III.9, we have*

$$-2q_\Sigma(1) + \sqrt{\frac{\mathrm{Tr}\,\Sigma}{N}} \leq \mathbb{E}[V_c] \leq \sqrt{\frac{\mathrm{Tr}\,\Sigma}{N}}\,.$$

As a consequence, for any $t > 0$,

$$\mathbb{P}\left[V^2 \geq \|\mu\|^2 + \left(\sqrt{\frac{\mathrm{Tr}\,\Sigma}{N}} + q_\Sigma(t) \right)^2 + 2\|\mu\| q_\Sigma(t) \right] \leq 2e^{-t}\,, \tag{III.34}$$

and for any $t \geq 1$,

$$\mathbb{P}\left[V^2 \leq \|\mu\|^2 + \left(\sqrt{\frac{\mathrm{Tr}\,\Sigma}{N}} - 4q_\Sigma(t) \right)_+^2 - 2\|\mu\| q_\Sigma(t) \right] \leq 2e^{-t}\,, \tag{III.35}$$

Remark. To the expert reader, we want to point out that the above concentration estimates are sharper than the Bernstein's concentration inequality for vector random variables due to Pinelis and Sakhanenko (1986) (Corollary 1 there) and which has found many uses in the recent literature on kernel methods. The reason is that in Pinelis and Sakhanenko's result, which concerns deviations of the centered process V_c, the deviation term (in factor of t) for V_c is proportional to $\sqrt{\mathrm{Tr}\,\Sigma/N}$. The inequality of Pinelis and Sakhanenko also only bounds upper deviations.

In contrast, in the above result, the term $\sqrt{\mathrm{Tr}\,\Sigma/N} = \mathbb{E}\big[\|V_c\|^2\big]^{\frac{1}{2}}$ appears with constant 1, and the main deviation term (in factor of t) only involves $\sqrt{\|\Sigma\|_{\mathrm{op}}/N}$, which is better by a factor of $1/\sqrt{d^{\mathrm{e}}}$. We also obtain the informative lower deviation bound (III.35).

To summarize, Pinelis and Sakhanenko (1986)'s inequality controls the upper deviations of V_c from zero in terms of a factor of its expectation, while the above concentration inequalities control the two-sided deviations of V_c^2 from its *expectation*, which is $\mathrm{Tr}\,\Sigma/N$, in terms of a factor of its typical deviation, which is $\|\Sigma\|_{\mathrm{op}}/N$.

This improvement makes the above bound first-order correct and mimic more closely the Gaussian chi-squared deviation phenomenon of Proposition III.6. This sharpness (and the fact that we get a control for two-sided deviations) is crucial in order to be able to capture the behavior of the effective dimension, see in particular Proposition III.12 below for the analysis of the MMD U-statistic, for which the exact cancellation of the first order terms is paramount.

Proof. The upper bound of the mean of V_c is given directly by Jensen's inequality. For the lower bound, we can rewrite Talagrand's inequality (III.30) equivalently under the following form: there exists ξ, an exponential random variable of parameter 1, such that almost surely

$$V_c \leq \mathbb{E}[V_c] + q_\Sigma(\xi) = \mathbb{E}[V_c] + \alpha\sqrt{\xi} + \beta\xi,$$

where α and β are given by (III.29). Taking the square and then the mean gives :

$$\begin{aligned}\mathbb{E}\left[V_c^2\right] &\leq \mathbb{E}\left[\left(\mathbb{E}[V_c] + \alpha\sqrt{\xi} + \beta\xi\right)^2\right] \\ &\leq \mathbb{E}\left[\left(\mathbb{E}[V_c] + (\alpha+\beta)\sqrt{\xi}\right)^2 + 2(\alpha+\beta)\mathbb{E}[V_c]\xi + (\alpha+\beta)^2\xi^2\right].\end{aligned}$$

We can use now the concavity of the function $\xi \mapsto \left(\mathbb{E}[V_c] + (\alpha+\beta)\sqrt{\xi}\right)^2$ and Jensen's inequality, obtaining

$$\mathbb{E}\left[V_c^2\right] \leq \left(\mathbb{E}[V_c] + (\alpha+\beta)\right)^2 + 2(\alpha+\beta)\mathbb{E}[V_c] + 2(\alpha+\beta)^2 \leq \left(\mathbb{E}[V_c] + 2(\alpha+\beta)\right)^2.$$

Because $\mathbb{E}\left[V_c^2\right] = \operatorname{Tr}\Sigma/N$, and $(\alpha+\beta) = q_\Sigma(1)$ by definition, we obtain that

$$\mathbb{E}[V_c] \geq \sqrt{\frac{\operatorname{Tr}\Sigma}{N}} - 2q_\Sigma(1).$$

If $t \geq 1$, it holds $q(t) \geq q(1)$ and we can plug in the above estimates for $\mathbb{E}[V_c]$ into (III.27) and (III.28) to obtain (III.34) and (III.35), respectively (note that the condition $t \geq 1$ is only needed for the lower devation bound). □

Proposition III.11. *Let $(X_k)_{1\leq k\leq N} \overset{i.i.d.}{\sim} X$ and $(Y_k)_{1\leq k\leq N} \overset{i.i.d.}{\sim} Y$ be independent families of centered random variables bounded by L in a separable Hilbert space $\mathcal{H}$. Let Σ_X and Σ_Y be their respective covariance operators, $\overline{\sigma}^2$ and d^{e} such that*

$$\max(\operatorname{Tr}\Sigma_X, \operatorname{Tr}\Sigma_Y)/N \leq \overline{\sigma}^2\,; \quad \min\left(\frac{\operatorname{Tr}\Sigma_X}{\|\Sigma_X\|_{op}}, \frac{\operatorname{Tr}\Sigma_Y}{\|\Sigma_Y\|_{op}}\right) \geq d^{\mathrm{e}}.$$

Then for any $t \geq 0$:

$$\mathbb{P}\left[\left\langle \frac{1}{N}\sum_{k=1}^N X_k, \frac{1}{N}\sum_{k=1}^N Y_k \right\rangle \geq 20q(t)\max(\overline{\sigma}, q(t))\right] \leq 6e^{-t}, \tag{III.36}$$

where

$$q(t) = 2\sqrt{\left(\frac{4\overline{\sigma}^2}{d^{\mathrm{e}}} + 16L\frac{\sqrt{2\overline{\sigma}^2}}{N}\right)t} + \frac{2L}{N}t. \tag{III.37}$$

Proof. Let us remark that

$$\left\langle \frac{1}{N}\sum_{k=1}^N X_k, \frac{1}{N}\sum_{k=1}^N Y_k \right\rangle = \frac{1}{2N^2}\left[\left\|\sum_{k=1}^N X_k + Y_k\right\|^2 - \left\|\sum_{k=1}^N X_k\right\|^2 - \left\|\sum_{k=1}^N Y_k\right\|^2\right].$$

So, by Corollary III.10, with probability greater than $1-6e^{-t}$, for $t \geq 1$, and using $(a-b)_+^2 \geq a^2 - 2ab$:

$$\begin{aligned}2\left\langle \frac{1}{N}\sum_{k=1}^N X_k, \frac{1}{N}\sum_{k=1}^N Y_k \right\rangle &\leq \left(\sqrt{\frac{\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y}{N}} + q(t)\right)^2 - \left(\sqrt{\frac{\operatorname{Tr}\Sigma_X}{N}} - 4q(t)\right)_+^2 - \left(\sqrt{\frac{\operatorname{Tr}\Sigma_Y}{N}} - 4q(t)\right)_+^2 \\ &\leq q(t)(19\overline{\sigma} + q(t)) \leq 20q(t)\max(\overline{\sigma}, q(t)).\end{aligned}$$

□

Proposition III.12. *Let $(X_k)_{1\le i\le N} \overset{i.i.d.}{\sim} X$ and $(Y_k)_{1\le i\le N} \overset{i.i.d.}{\sim} Y$ be independent families of random variables bounded by L in $\mathcal{H}$. Let μ_x, Σ_X and μ_Y, Σ_Y denote their respective means and covariance operators. Let U the statistic defined as*

$$U = \frac{1}{N(N-1)} \sum_{\substack{k,\ell=1 \\ k\neq\ell}}^{N} \langle X_k, X_\ell\rangle_{\mathcal{H}} - \frac{2}{N^2} \sum_{k,\ell=1}^{N} \langle X_k, Y_\ell\rangle_{\mathcal{H}} + \frac{1}{N(N-1)} \sum_{\substack{k,\ell=1 \\ k\neq\ell}}^{N} \langle Y_k, Y_\ell\rangle_{\mathcal{H}}. \tag{III.38}$$

Then for any $t \ge 1$, $N \ge 2$:

$$\mathbb{P}\Big[U \ge \|\mu_X - \mu_Y\|^2 + 2\|\mu_X - \mu_Y\| q(t) + 2\sqrt{2\overline{\sigma^2}} q(t) + 11 q^2(t)\Big] \le 8e^{-t}, \tag{III.39}$$

and

$$\mathbb{P}\Big[U \le \|\mu_X - \mu_Y\|^2 - 2\|\mu_X - \mu_Y\| q(t) - 8\sqrt{2\overline{\sigma^2}} q(t) - 32 q^2(t)\Big] \le 8e^{-t}, \tag{III.40}$$

where $q(t)$ is given by (III.37).

Proof. Observe that

$$\begin{aligned} U &= \left\| \frac{1}{N}\sum_{k=1}^{N} X_k - \frac{1}{N}\sum_{k=1}^{N} Y_k \right\|^2 + \frac{1}{N-1}\left(\left\| \frac{1}{N}\sum_{k=1}^{N} X_k \right\|^2 + \left\| \frac{1}{N}\sum_{k=1}^{N} Y_k \right\|^2 - \frac{1}{N}\sum_{k=1}^{N} \|X_k\|^2 - \frac{1}{N}\sum_{k=1}^{N} \|Y_k\|^2 \right) \\ &=: \left\| \frac{1}{N}\sum_{k=1}^{N} X_k - \frac{1}{N}\sum_{k=1}^{N} Y_k \right\|^2 + \frac{1}{N-1} H\,. \end{aligned}$$

Using now the upper bound of Bernstein's inequality, since $\mathbb{E}\big[\|X\|^2\big] = \|\mu_X\|^2 + \operatorname{Tr}\Sigma_X$, with probability greater than $1 - e^{-t}$ it holds:

$$\frac{1}{N}\sum_{k=1}^{N} \|X_k\|^2 \ge \operatorname{Tr}\Sigma_X + \|\mu_X\|^2 - \sqrt{2L^2\overline{\sigma^2} t} - \frac{2L^2 t}{3N}\,.$$

So using (III.34) (twice), with probability greater than $1 - 6e^{-t}$:

$$\begin{aligned} H \le& \|\mu_X\|^2 + 2\|\mu_X\| q(t) + \left(\sqrt{\frac{\operatorname{Tr}\Sigma_X}{N}} + q(t)\right)^2 + \|\mu_Y\|^2 + 2\|\mu_Y\| q(t) + \left(\sqrt{\frac{\operatorname{Tr}\Sigma_Y}{N}} + q(t)\right)^2 \\ &- \operatorname{Tr}\Sigma_X - \|\mu_X\|^2 + \sqrt{2L^2\overline{\sigma^2} t} + \frac{2L^2 t}{3N} - \operatorname{Tr}\Sigma_Y - \|\mu_Y\|^2 + \sqrt{2L^2\overline{\sigma^2} t} + \frac{2L^2 t}{3N} \\ \le& -(N-1)/N\Big(\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y\Big) + 4Lq(t) + 4\sqrt{\overline{\sigma^2}} q(t) + 2q^2(t) + 2\sqrt{2L^2\overline{\sigma^2} t} + \frac{4L^2 t}{3N} \\ \le& -(N-1)/N\Big(\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y\Big) + (2+4N) q^2(t)\,. \end{aligned}$$

Using again (III.34), and $N \ge 2$, with probability greater than $1 - 8e^{-t}$:

$$\begin{aligned} U &\le \|\mu_X - \mu_Y\|^2 + 2\|\mu_X - \mu_Y\| q(t) + \left(\sqrt{\frac{\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y}{N}} + q(t)\right)^2 - \frac{\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y}{N} + 10 q^2(t) \\ &\le \|\mu_X - \mu_Y\|^2 + 2\|\mu_X - \mu_Y\| q(t) + 2\sqrt{2\overline{\sigma^2}} q(t) + 11 q^2(t)\,, \end{aligned}$$

which is (III.39).

We proceed similarly for lower deviations of U: using again Bernstein's inequality and (III.35), with probability greater than $1 - 6e^{-t}$, and using $(a-b)_+^2 \ge a^2 - 2ab$:

$$\begin{aligned} H \ge& \|\mu_X\|^2 - 2\|\mu_X\| q(t) + \left(\sqrt{\frac{\operatorname{Tr}\Sigma_X}{N}} - 4q(t)\right)_+^2 + \|\mu_Y\|^2 - 2\|\mu_Y\| q(t) + \left(\sqrt{\frac{\operatorname{Tr}\Sigma_Y}{N}} - 4q(t)\right)_+^2 \\ &- \operatorname{Tr}\Sigma_X - \|\mu_X\|^2 - \sqrt{2L^2\overline{\sigma^2} t} - \frac{2L^2 t}{3N} - \operatorname{Tr}\Sigma_Y - \|\mu_Y\|^2 - \sqrt{2L^2\overline{\sigma^2} t} - \frac{2L^2 t}{3N} \\ \ge& -(N-1)/N(\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y) - 16N q^2(t)\,, \end{aligned}$$

which implies, using again (III.35), and $N \geq 2$, that with probability greater than $1 - 8e^{-t}$ it holds:

$$\begin{aligned} U &\geq \|\mu_X - \mu_Y\|^2 - 2\|\mu_X - \mu_Y\| q(t) + \left(\sqrt{\frac{\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y}{N}} - 4q(t)\right)_+^2 - \frac{\operatorname{Tr}\Sigma_X + \operatorname{Tr}\Sigma_Y}{N} - 16q^2(t) \\ &\geq \|\mu_X - \mu_Y\|^2 - 2\|\mu_X - \mu_Y\| q(t) - 8\sqrt{2\overline{\sigma}^2} q(t) - 32q^2(t)\,, \end{aligned}$$

which is (III.40).

□

III.6.7 Details on the Tested Methods in the Numerical Experiments

In the following, the methods that are tested in the experiments are described in more detail. Recall, that $V_i := \{j : T_{ij} = 1, j \in [\![B]\!]\}$ and let T_{ij} be defined as in Eq (III.18), i.e. V_i holds the neighboring kernel means of bag i. All of the methods give KME estimations of the form

$$\tilde{\mu}_i := \sum_{j \in [\![B]\!]} \omega_{ij} \cdot \widehat{\mu}_j^{\mathrm{NE}},$$

where the definition of the weighting w_{ij} depends on the applied method.

1. `NE` considers each bag individually. Therefore, the weighting is simply

$$\omega_{ij} = \begin{cases} 1, & \text{for } i = j \\ 0, & \text{otherwise.} \end{cases}$$

2. `R-KMSE` was proposed by Muandet et al. (2016). It estimates each KME individually but shrinks it towards 0. The amount of shrinkage depends on the data and is defined as

$$\omega_{ij} = \begin{cases} 1 - \frac{\lambda}{1+\lambda}, & \text{for } i = j \\ 0, & \text{otherwise} \end{cases}$$

where

$$\lambda = \frac{\varrho - \rho}{(1/N_b - 1)\varrho + (N_b - 1)\rho}$$

with $\varrho = 1/N_i \sum_{k=1}^{N_i} k(Z_k^{(i)}, Z_k^{(i)})$ and $\rho = 1/N_i^2 \sum_{k,\ell=1}^{N_i} k(Z_k^{(i)}, Z_\ell^{(i)})$.

3. `STB-0` is described in Eq. (III.5) with γ set to 0, i.e.

$$\omega_{ij} = \begin{cases} \frac{1}{|V_i|}, & \text{for } j \in V_i \\ 0, & \text{otherwise.} \end{cases}$$

4. `STB theory` is defined by Eq. (III.5). It uses the optimal value for γ as described in Eq. (III.7) that was proven to be optimal. Here, τ is replaced by its empirical counterpart ζ and another multiplicative constant $c > 0$ was added to allow for more flexibility. Its specific value must be found using model optimization.

$$\omega_{ij} = \begin{cases} \gamma + \frac{1-\gamma_i}{|V_i|}, & \text{for } i = j \\ \frac{1-\gamma_i}{|V_i|}, & \text{for } i \neq j, j \in V_i \\ 0, & \text{otherwise} \end{cases}$$

with

$$\gamma_i = \frac{c \cdot \zeta \cdot (|V_i| - 1)}{(1 + c \cdot \zeta) \cdot (|V_i| - 1) + 1}.$$

5. `STB weight` is also described by Eq. (III.5) but the optimal value of γ is found by model optimization

$$\omega_{ij} = \begin{cases} \gamma + \frac{1-\gamma}{|V_i|}, & \text{for } i = j \\ \frac{1-\gamma}{|V_i|}, & \text{for } i \neq j, j \in V_i \\ 0, & \text{otherwise.} \end{cases}$$

6. `MTA const` is based on a multi-task averaging approach described in Feldman et al. (2014) which we translated to the KME framework as

$$\omega_{ij} = \left(\left(I + \frac{\gamma}{B} D \cdot L(A) \right)^{-1} \right)_{ij}. \tag{III.41}$$

   Here, $D = \text{diag}\big((E_i)_{i \in [\![B]\!]}\big)$ as defined in Eq. (III.19) and $L(A)$ denotes the graph Laplacian of task-similarity matrix A. For `MTA const` the similarity is assumed to be constant, i.e. $A = a \cdot (\mathbf{1}\mathbf{1}^T)$ with $a = \frac{1}{B(B-1)} \sum_{i,j \in [\![B]\!]} \big\| \widehat{\mu}_i^{\text{NE}} - \widehat{\mu}_j^{\text{NE}} \big\|_{\mathcal{H}}^2$. Again, the optimal value for γ must be found using model optimization.

7. `MTA stb` is defined as in Eq. (III.41). In contrast to `MTA const`, the similarity matrix A is defined as

$$A_{ij} = \begin{cases} 1, & \text{for } j \in V_i \\ 0, & \text{otherwise.} \end{cases}$$

The methods `STB-0`, `STB weight`, `STB theory` and `MTA stb` use all the similarity test defined by T_{ij} which depends on ζ. Nevertheless, the optimal value for ζ is found by model optimization for each method individually.

III.6.8 Numerical Results in the Gaussian Setting

In this section we report numerical comparisons of the proposed approaches in the idealized Gaussian setting (GI). In that setting, since the tests and proposed estimates only depend on the naive estimators, we can reduce each bag to its naive estimator, in other words we can assume $N = 1$ (only one observation per bag). We consider the following models for the means $(\mu_i)_{i \in [\![B]\!]}$ (in each case the number of bags is $B = 2000$):

- Model **UNIF**: ambient dimension $d = 1000$, the means $(\mu_i)_{i \in [\![B]\!]}$ are distributed uniformly over the lower-dimensional cube $[-20, 20]^{d'}$, $d' = 10$ (the remaining coordinates are set to 0).
- Model **CLUSTER**: ambient dimension $d = 1000$, the means are clustered in 20 clusters of centers $(m_i)_{i \in [\![10]\!]}$, drawn as $\mathcal{N}(0, I_d)$, in each cluster the means are drawn as Gaussians $\mathcal{N}(m_i, 0.1 * I_d)$,
- Model **SPHERE**: ambient dimension $d = 1000$, the 6 first coordinates of the means are distributed uniformly on the sphere of radius 50 in $\mathbb{R}^6$, the rest are set to 0.
- Model **SPARSE**: ambient dimension $d = 50$, the means are 2-sparse vectors with two random coordinates distributed as $\text{Unif}[0, 20]$.

In each case, we first select the parameter for the tests (parameter ζ in (III.13)) from the oracle `STB-0` performance. This value is held fixed and the shrinkage parameter in methods `MTA stb`, `STB theory`, `STB weight` is again determined as its "oracle" value by minimization over the squared error, as done in the KME experiments.

For comparison, we also display the results of the classical positive-part James-Stein estimator (`JS+`, Baranchik, 1970), which is a shrinkage estimator applied separately on each bag. It has no tuning parameter.

Table 4: Decrease in averaged squared estimation error compared to `NE` in percent on the Gaussian data (higher is better). Averaged results over 20 trials. Standard error of one given trial is of order 5.10^{-3}.

	JS+	MTA const	MTA stb	STB-0	STB theory	STB weight
UNIF	0.439	0.427	0.653	0.796	0.813	0.813
CLUSTER	0.495	0.508	0.979	0.980	0.980	0.980
SPHERE	0.285	0.285	0.745	0.894	0.898	0.898
SPARSE	0.224	0.162	0.367	0.402	0.441	0.443

IV Nonasymptotic one-and two-sample tests in high dimension with unknown covariance structure

Let $\mathbb{X} = (X_i)_{1\leq i\leq n}$ be an i.i.d. sample of square-integrable variables in $\mathbb{R}^d$, with common expectation μ and covariance matrix Σ, both unknown. We consider the problem of testing if μ is η-close to zero, i.e. $\|\mu\| \leq \eta$ against $\|\mu\| \geq (\eta + \delta)$; we also tackle the more general two-sample mean closeness (also known as *relevant difference*) testing problem. The aim of this work is to obtain nonasymptotic upper and lower bounds on the minimal separation distance δ such that we can control both the Type I and Type II errors at a given level. The main technical tools are concentration inequalities, first for a suitable estimator of $\|\mu\|^2$ used a test statistic, and secondly for estimating the operator and Frobenius norms of Σ coming into the quantiles of said test statistic. These properties are obtained for Gaussian and bounded distributions. A particular attention is given to the dependence in the pseudo-dimension d_* of the distribution, defined as $d_* := \|\Sigma\|_2^2/\|\Sigma\|_\infty^2$. In particular, for $\eta = 0$, the minimum separation distance is $\Theta(d_*^{1/4}\sqrt{\|\Sigma\|_\infty/n})$, in contrast with the minimax estimation distance for μ, which is $\Theta(d_e^{1/2}\sqrt{\|\Sigma\|_\infty/n})$ (where $d_e := \|\Sigma\|_1/\|\Sigma\|_\infty$). This generalizes a phenomenon spelled out in particular by Baraud (2002).

Contents

IV.1 Introduction

We consider the following fundamental signal detection problem: given an i.i.d. sample $\mathbb{X} = (X_i)_{1\leq i\leq n}$ from a square integrable distribution $\mathbb{P}_X$ on $\mathbb{R}^d$ (or possibly a separable Hilbert space, under some conditions which will be discussed later) with $\mu = \mathbb{E}[X_1]$, test the hypothesis of "η-closeness to zero" of the mean:

$$(H_0(\eta)) : \|\mu\| \leq \eta, \text{ against } (H_1(\eta,\delta)) : \|\mu\| > \eta + \delta. \tag{IV.1}$$

In fact, we consider the following more general two-sample mean closeness testing problem: for $\mathbb{X} = (X_i)_{1\leq i\leq n}$ and $\mathbb{Y} = (Y_i)_{1\leq i\leq m}$ two independent samples of i.i.d. variables with distributions $\mathbb{P}_X, \mathbb{P}_Y$ on

$\mathbb{R}^d$ with respective means μ and ν, test the hypothesis of η-closeness (or similarity) of the two means,

$$(H_0(\eta)) : \|\mu - \nu\| \leq \eta, \text{ against } (H_1(\eta, \delta)) : \|\mu - \nu\| > \eta + \delta. \tag{IV.2}$$

Observe that we can always formally subsume setting (IV.1) into setting (IV.2), by letting m go to infinity and/or assuming (if needed) that the covariance of Y_1 is zero. Therefore, in the contribution section we will concentrate mainly on setting (IV.2).

The problem (IV.1) (and numerous extensions thereof) has been a long-time subject of attention in mathematical statistics. For the zero mean test problem, i.e. $\eta = 0$, the celebrated works of Ingster (1982, 1993) in the Gaussian white noise model are seminal. In the case $\eta > 0$, the problems (IV.1)-(IV.2) are known as testing for *precise hypotheses*, *relevant hypotheses* or *relevant differences* (Berger and Delampady, 1987); this setting has found applications in particular in biostatistics for bioequivalence testing (see e.g. Wellek, 2002). (See next sections for a more detailed discussion of related literature.) In this work, we will consider the situation where the involved distributions are either Gaussian or of bounded norm (and hence sub-Gaussian), but with unknown covariance matrix acting as a nuisance parameter.

We are interested in finding bounds on the separation distance δ, i.e. a bound on the minimum value of δ such that there exists a test with both Type I and Type II error rates bounded by a "small" prescribed quantity. Our interest here is more on the constructive side, so that we will concentrate on feasible procedures that are in particular adaptive to the covariances of the involved distributions. A matching lower bound (for any fixed covariance structure) will be provided in the Gaussian setting. We emphasize that our focus is on *finite sample* (i.e. nonasymptotic) results, as will be discussed below.

IV.1.1 Relation to white noise model in nonparametric statistics

In the isotropic Gaussian case (white noise) with known variance , and for $\eta = 0$, the signal detection problem (IV.1) has been studied in much generality, in particular in the infinite-dimensional setting where $\mathbb{R}^d$ is replaced by a separable Hilbert space. In this situation, due to the fact that the white noise model on an infinite-dimensional Hilbert space cannot be represented by a random variable taking values in that space, the canonical model which is considered instead is the Gaussian sequence model for the coordinates of each of the observations in an orthonormal basis (in fact the Gaussian sequence model with known variance is usually considered with a single observation of the sequence):

$$X^{(i)} = \mu^{(i)} + \sigma\varepsilon^{(i)}, \qquad i \in \mathbb{N}_{>0}, \tag{IV.3}$$

where $(\varepsilon^{(i)})_{i\geq 1}$ is an i.i.d. standard normal sequence. This fundamental model in nonparametric statistics allows to represent in a clean way many functional spaces of interest for the signal μ through geometrical properties of its expansion coefficients $(\mu^{(i)})_{i\geq 1}$ in a suitable basis. Since in that infinite-dimensional setting the alternative $\|\mu\|^2 > \delta^2$ is "too big" and gives rise to trivial separation rates, the usual focus is on considering restricted alternatives of the form $\left\{\mu \in \mathcal{F}; \|\mu\| \geq \delta^2\right\}$, for a given nonparametric set $\mathcal{F}$. Classical alternatives of interest include in particular ℓ_2 ellipsoids (corresponding to Hilbert norms of different strengths), ℓ_p bodies, and Besov bodies. Interpreted in functional spaces, these alternatives correspond respectively to balls in Sobolev spaces (typically when considering Fourier basis coefficient expansions) or in Besov spaces (for suitable wavelet basis coefficient expansions).

The literature on these topics is profound and extensive, see e.g. Ingster and Suslina (2012) for a comprehensive overview. The case of certain classes of ℓ_2-ellipsoids appears to have been studied first by Ingster (1982) and Ermakov (1991), then a remarkable series of works of Ingster (1993) and Ingster and Suslina (1998) established minimax testing rates for general ℓ_2 ellipsoids as well as other alternatives. V. Spokoiny's contribution is prominent in this body of literature, in particular for dealing with the case of Besov bodies (Lepski and Spokoiny, 1999) as well as considering the problem of statistical adaptivity over a family of alternatives (Spokoiny, 1996).

This very limited overview of the topic of testing in the white noise model is meant to contrast with the setting considered here. On the one hand, we will not consider a particular form of alternative; on the other hand, we assume that the observations can truly be represented as elements in a possibly infinite-dimensional separable Hilbert space. Under the Gaussian assumption, this means that the covariance operator Σ of the noise process is assumed to have a finite trace, which also prevents the triviality problem mentioned above for the white Gaussian noise setting. If we represent the observation coordinates in a diagonalizing basis of Σ, our setting in the Gaussian setting amounts to the Gaussian sequence model (IV.3) wherein the constant parameter σ is replaced by a square integrable sequence $(\sigma^{(i)})_{i\geq 1}$. Note that formally normalizing the i-th observation coordinate by $\sigma^{(i)}$ would give rise again to model (IV.3), however the separation distance would then be measured in the weak norm $\left\|\Sigma^{1/2}\mu\right\|$.

IV.1.2 Relation to "modern" and high-dimensional statistics

Since we only consider test separation distance without a specific alternative, the setup we consider can be considered as less elaborate, at least in the sense of asymptotic theory, than the settings with various non-parametric alternatives discussed above. On the other hand, our focus is specifically on the following points:

1. Finite-sample analysis;
2. Non-Gaussian data (we will only consider bounded data here);
3. Robustness to misspecification (here under the form of the relaxed composite null $\|\mu\|^2 \leq \eta^2$, also called *relevant hypothesis testing*).

These features have been rightly identified by V. Spokoiny as the defining features of "modern" approach to statistics (Spokoiny, 2012; Spokoiny and Dickhaus, 2015). The problem of testing a null hypothesis defined as a neighborhood rather than an exact match has been tackled under different settings in the statistics literature, especially for the two-sample testing case. For example, motivated by bioequivalence testing between populations, Munk and Czado (1998) consider the problem of testing closeness of two real distributions as measured in Mallows distance, Dette and Munk (1998) that of closeness in L^2 distance of two nonparametric (Hölder regular) regression functions; Dette et al. (2020a), the closeness in supremum norm distance of two mean functions in a Banach functional data setting; Dette et al. (2020b), the closeness in L^2 distance of the functional mean of time series. In all cases, the underlying principle is to estimate the target distance — as will be also case in this section — and the data is not always assumed to be Gaussian, but the corresponding analysis based on Gaussian asymptotic theory. To estimate the quantiles of the test statistic, Dette and Munk (1998) choose to estimate the variance, Dette et al. (2020b) use a self-normalized procedure and give asymptotic bounds; Dette et al. (2020a) propose a bootstrap approach and obtain an asymptotic convergence of the test statistic. Our approach is a direct estimation of the variance with nonasymptotic guarantees.

Taking the above aspects into account in the theory, in particular non-asymptotic analysis, is motivated by a large number of high-dimensional applications, where it appears that relying on traditional asymptotic of Gaussian parametric or non-parametric theory can possibly be problematic if done without care. Finite sample theory allows to delineate more precisely in which situations traditional approximations still can be relied upon, and to study non-standard asymptotics, in particular when key parameters, such as dimensionality, can themselves depend on the sample size n. It is also of use when considering multiple testing scenarios, where multiplicity has to be taken into account precisely.

Another fruitful modern insight is that high-dimensional statistical models tend to blur the line between parametric and non-parametric point of views. Precise non-asymptotic results in a finite-dimensional setting, but where the role of key model parameters (in particular, dimensionality or effective dimensionality) is precisely analyzed, can provide key theoretical components for analyzing non-parametric settings. In the

signal testing framework considered here, this way of thinking has in particular been pioneered by Baraud (2002), who obtained sharp non-asymptotic results for the problem (IV.1) in the case $\eta = 0$, and for the finite-dimensional counterpart of the white noise model (IV.3), i.e. the isotropic setting $\Sigma = \sigma^2 I_d$ in dimension d. Baraud further demonstrated that this result provided a valuable and versatile tool to analyze models of typical interest in high-dimensional statistics (such as sparse alternatives) as well as non-parametric alternatives (such as those mentioned in the previous section). A key insight from Baraud's work is that the minimum separation distance in that setting is $\mathcal{O}(d^{1/4}\sigma/\sqrt{n})$, in contrast with minimax estimation distance for μ, which is $\Theta(d^{1/2}\sigma/\sqrt{n})$: the testing separation distance is smaller than the minimax estimation error by a factor $d^{1/4}$.

Analyzing precisely the role of dimensionality (ambient or effective) in minimax testing separation rates and the difference with minimax estimation rates has been a subject of interest in recent literature in various settings, highlighting similar related phenomena. For instance, Lam-Weil et al. (2022) consider the problem of testing equality of two high-dimensional multinomial distributions and study the minimum ℓ_1 separation distance in a vicinity of a reference distribution π (which implicitly determines a notion of local effective dimensionality). Since this model has bounded data, our analysis could be applied in that setting, however it concerns separation in ℓ_2 distance (the separation in ℓ_1 distance exhibits considerably more involved behavior). Ostrovskii et al. (2020) consider a different type of two-sample testing problem, in a regression context, where the goal is to determine which one of the two distributions has a given (known to the user) regression vector. They give a sharp bound on the minimum separation distance between the two regression vectors including the role of the dimension, also exhibiting a difference with estimation rates.

Coming back to our model, the results of Baraud (2002) provide a sharp answer, but only in the case $\eta = 0$ and for isotropic Gaussian (white noise) data with known variance. Still in the Gaussian isotropic case, the minimum separation rates for any value of $\eta \geq 0$ were precisely characterized by Blanchard et al. (2018). We also consider the Gaussian setting in the present work, but analyze the generalized situation where the covariance matrix Σ can be arbitrary (and unknown). In this situation, the role of the dimensionality d is played by proxy quantities depending on Σ, sometimes called effective dimensionality or effective rank. For the signal testing problem however, it turns out that the proxy dimensionalities for testing and estimation differ. Namely, for $\eta = 0$, we find that the minimax separation distance is $\mathcal{O}(d_*^{1/4}\sqrt{\|\Sigma\|_\infty/n})$, where $d_* := \|\Sigma\|_2^2/\|\Sigma\|_\infty^2$, while the minimax estimation distance for μ is $\Theta(d_e^{1/2}\sqrt{\|\Sigma\|_\infty/n})$, where $d_e := \|\Sigma\|_1/\|\Sigma\|_\infty$. (Notice that $d_* \leq d_e \leq d$ in general, while these quantities are all equal in the isotropic setting.) Furthermore, we also study the estimation of key quantities $\|\Sigma\|_\infty^{1/2}$ and $\|\Sigma\|_2$ determining the proxy dimensionality and the testing threshold[4].

A crucial mathematical tool in high-dimensional statistics is to obtain sharp concentration inequalities for quadratic forms of random vectors. These are closely related to technical tools used in the present work. An important point in such inequalities is to quantify as precisely as possible up to which point quadratic forms of non-Gaussian vectors can mimic the Gaussian behavior (i.e. that of central and non-central weighted chi-squared statistics). This topic has received a good deal of attention in the recent years and V. Spokoiny also made substantial contributions to that area (Spokoiny and Dickhaus, 2015; Spokoiny and Zhilova, 2013). In the present work, we derive from scratch the needed concentration inequalities; we discuss in more detail the relation to V. Spokoiny's own work and to related literature in Section IV.2.4.

IV.1.3 Relation to machine learning and kernel mean embeddings of distributions

An application setting which motivated us to consider in detail the case of bounded data is that of testing of the data distribution via kernel mean embedding (KME) methods, a principle which has garnered a lot

[4] With the notation $\|\Sigma\|_p$ we mean p-Schatten norm. We will freely use in this section the equivalent notation $\|\Sigma\|_\infty = \|\Sigma\|_{\mathrm{op}}$, $\|\Sigma\|_1 = \mathrm{Tr}(\Sigma)$, $\|\Sigma\|_2^2 = \mathrm{Tr}(\Sigma^2)$.

of attention in the machine literature since the seminal paper of Smola et al. (2007). It has been advocated in particular for two-sample (Gretton et al., 2012) and goodness-of-fit (Chwialkowski et al., 2016) testing; see Section II.2.

We describe the KME principle briefly. Assume Z is a random variable with distribution $\mathbb{P}_Z$ taking values in the measurable space $\mathcal{Z}$, and that one has at hand a fixed mapping $\Phi : \mathcal{Z} \to \mathcal{H}$, where $\mathcal{H}$ is a separable Hilbert space. To this mapping is associated a reproducing kernel Hilbert space (RKHS) $\mathcal{H}'$ with kernel $k(z, z') := \langle \Phi(z), \Phi(z') \rangle$.

Assuming the variable $X = \Phi(Z)$ is Bochner integrable[5] (which is the case in particular when the mapping Φ is bounded), the kernel mean embedding of $\mathbb{P}_Z$ is defined as $\Phi(\mathbb{P}_Z) := \mathbb{E}[\Phi(Z)] \in \mathcal{H}$ (using a rather natural overload of notation for Φ). The *maximum mean discrepancy* (MMD) between distributions $\mathbb{P}, \mathbb{Q}$ in the domain of definition of Φ is defined as the semimetric

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) := \|\Phi(\mathbb{P}) - \Phi(\mathbb{Q})\|.$$

Since $\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) > 0$ implies $\mathbb{P} \neq \mathbb{Q}$, this principle can be used for simple goodness-of-fit testing (testing for $\mathbb{P}_Z = \mathbb{P}_0$ for some known distribution $\mathbb{P}_0$, given an i.i.d. sample from $\mathbb{P}_Z$) and two-sample testing (testing for $\mathbb{P}_Z = \mathbb{P}_{Z'}$, given two independent i.i.d. samples from $\mathbb{P}_Z$ and $\mathbb{P}_{Z'}$); in each case, the test statistic is a suitable estimator of $\mathrm{MMD}_k(\mathbb{P}_Z, \mathbb{P}_0)$, resp. $\mathrm{MMD}_k(\mathbb{P}_Z, \mathbb{P}_{Z'})$ from the observed data. More generally one may want to test the relaxed null hypothesis $\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) \leq \eta$ and analyze the power of the test in terms of the MMD separation itself. This is indeed a particular case of (IV.1)-(IV.2), when considering the Hilbert-valued variable $X = \Phi(Z)$ and, for two-sample testing, $Y = \Phi(Z')$.

A common situation is when Φ is bounded in norm by some constant L, or equivalently in terms of the kernel, $\sup_{z \in \mathcal{Z}} k(z, z) \leq L^2$. This ensures in particular that Φ is defined on all distributions. Analyzing our original setting with norm-bounded but potentially infinite-dimensional data is therefore suited to this case.

Gretton et al. (2012) derive the asymptotic distribution of the (suitably renormalized) MMD test statistic, which is identical to the one we use below (once interpreted in the KME setting). Unsurprisingly, a Gaussian limiting behavior is identified. Our study analyzes this behavior from a non-asymptotic point of view; this can be particularly of interest for situation where the mapping Φ (or equivalently the associated kernel) is to depend on the sample size, or when performing a large number of such tests in parallel: in this case uniformly valid nonasymptotic bounds are a a valuable tool for further analysis. See Section III for such a multiple test scenario in the context of multiple task averaging. Multiple tests can also be aggregated to test a global hypothesis, see Fromont et al. (2012) in the context of two-sample testing based on the KME approach.

In our study, the power of the test is investigated for alternatives of the form (IV.1)-(IV.2), which, interpreted in the KME setting, correspond to $\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) \geq \eta + \delta$. The power of KME-based tests (in the goodness-of-fit case) was also investigated by Balasubramanian et al. (2021), but for alternatives measured in a χ^2 distance separation, more precisely, of the form $\{\mathbb{Q} \in \mathcal{F}; \chi^2(\mathbb{P}_0, \mathbb{Q}) \geq \delta\}$, where $\mathcal{F}$ is a nonparametric set of distributions whose density with respect to $\mathbb{P}_0$ is approximated at a given rate by functions in the RKHS $\mathcal{H}'$ associated to k, in the sense of interpolation with $L^2(\mathbb{P}_0)$. This is close in spirit to nonparametric points of view discussed in Section IV.1.1, in the sense that χ^2-separation alone is too weak to get nontrivial separation rates and one has to additionally consider intersection with nonparametric sets of interest. Again, because we choose to analyze alternatives measured in MMD_k-separation itself, the results we obtain in this setting have a different nature.

[5] that is, the real random variable $\|\Phi(Z)\|$ is integrable, which guarantees that the integral of $\Phi(Z)$ is well-defined in a strong sense as an element of the Hilbert space; see e.g. Cohn (1980).

IV.1.4 Overview of contributions

The main contribution of this section is to give upper bounds on the optimal (minmax) testing separation distance for problems (IV.1) and (IV.2) over classes of probability distributions with fixed covariance matrix Σ for sample $\mathbb{X}$, as well as S for sample $\mathbb{Y}$ in the two-sample case. The covariance structures are considered as nuisance parameters and we investigate precisely how they influence the testing separation distance. Let $\mathcal{P}$ be a family of distributions for the two samples (we consider the Gaussian setting and the bounded setting), and $\mathcal{P}_{\Sigma,S}$ the subsets of distributions of $\mathcal{P}$ with $\mathrm{Cov}[X_1] = \Sigma$, and $\mathrm{Cov}[Y_1] = S$ (in the two-sample case). Consider the sets of distributions

$$\mathcal{H}_0(\eta, \Sigma, S) := \{\mathbb{P} \in \mathcal{P}_{\Sigma,S} | \mathbb{P} \text{ satisfies } H_0(\eta)\},$$
$$\mathcal{A}_\delta(\eta, \Sigma, S) := \{\mathbb{P} \in \mathcal{P}_{\Sigma,S} | \mathbb{P} \text{ satisfies } H_1(\eta, \delta)\},$$

then the optimal separation distance is, for $\alpha \in (0, 1)$:

$$\delta^*(\alpha, \Sigma, S, \eta) = \inf\left\{\delta \geq 0 \middle| \exists \text{ test } T \ : \sup_{\mathbb{P}\in\mathcal{H}_0} \mathbb{P}(T = 1) + \sup_{\mathbb{P}\in\mathcal{A}_\delta} \mathbb{P}(T = 0) \leq \alpha\right\}. \tag{IV.4}$$

In the Gaussian setting, we establish that δ^* is upper bounded up to a constant factor via

$$\delta^*(\alpha, \Sigma, S, \eta) \lesssim \sigma\kappa_\alpha \max\left(1, \min\left(d_*^{\frac{1}{4}}, d_*^{\frac{1}{2}}\frac{\sigma\kappa_\alpha}{\eta}\right)\right), \tag{IV.5}$$

(Theorem IV.7) where $\kappa_\alpha := \sqrt{-\log(\alpha)}$, and, in the one-sample case, $\sigma^2 := \|\Sigma\|_{\mathrm{op}}/n$ is a scalar variance factor and $d_* := \mathrm{Tr}\,\Sigma^2/\|\Sigma\|_{\mathrm{op}}^2$ a notion of effective dimension. In the two-sample case, we obtain also (IV.5), with $\sigma^2 := \|M_{m,n}\|_{\mathrm{op}}$, and $d_* := \mathrm{Tr}\,M_{m,n}^2/\sigma^4$, where $M_{m,n} := (\Sigma/n + S/m)$ (Theorem IV.8). In the one-sample case, this result can be formulated equivalently in terms of *sample complexity* n^* needed to detect at given error level α and separation distance δ for problem (IV.1):

$$n^*(\alpha, \Sigma, S, \eta) \lesssim \|\Sigma\|_{\mathrm{op}}\kappa_\alpha\delta^{-1}\max\left(\delta^{-1}, d_*^{\frac{1}{2}}(\max(\delta, \eta))^{-1}\right). \tag{IV.6}$$

This result is established first when assuming that Σ, S are known, then we show that it holds as well when they are unknown (under some mild assumptions on the sample size, see Corollary IV.14 for an explicit statement in the one-sample case and condition (IV.27) there). Matching minimax lower bounds are given for one and two-sample problems in the Gaussian setting. In the bounded setting, we derive upper bounds only, which take the same flavor as (IV.5) under some mild assumptions on the sample sizes.

IV.1.5 Organization of the section

We present in Section IV.2 our main results. In order to cover both the Gaussian and bounded settings under the same umbrella, we start in Section IV.2.1 by a generic result: assuming some suitable concentration for an estimate U of the squared signal norm $\|\mu\|^2$ holds (Assumption IV.1), as well as for estimators of its quantiles (Assumption IV.2), for the problems (IV.1) and (IV.2) we propose in Theorem IV.3 sufficient conditions on δ such that we can control the Type I and Type II errors of a test T based on U. In the following sections, the Gaussian setting and the bounded setting are considered separately. In Section IV.2.2, we give concentration results for U to fulfill Assumption IV.1. In Section IV.2.3 we give results to fulfill Assumption IV.2, which are related to the estimation of $\|\Sigma\|_\infty^{1/2}$ and $\|\Sigma\|_2$. The proofs of the corresponding results are found in Sections IV.3.1 to IV.3.6, respectively.

IV.2 Main results

We will build a test for the model (IV.2) based on an estimator U of the distance $\|\mu - \nu\|^2$, typically a modified U-statistic as defined below. We will first consider a general point of view to deduce bounds

on the separation rate when U satisfies certain concentration properties; this will then apply both to the Gaussian and bounded settings.

IV.2.1 A general result to upper bound separation rates

As mentioned earlier, from now on we concentrate primarily on the two-sample setting, being understood that upper bounds for the one-sample setting can be deduced readily. In order to define a general framework encompassing as particular cases the more specific settings considered below, in this section we will assume a generic statistical model $\mathcal{P}$ for the distribution of the samples $\mathbb{X}$ and $\mathbb{Y}$, which we recall we always assume to be independent and i.i.d. with respective squared integrable marginal distributions $\mathbb{P}_X, \mathbb{P}_Y$. We will thus use without comment the fact that a distribution $\mathbb{P} \in \mathcal{P}$ equivalently specifies the marginal distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$ of the samples. We will consider the covariance matrices Σ, S of $\mathbb{P}_X, \mathbb{P}_Y$ as nuisance parameters influencing the optimal separation distance, and define the sub-models

$$\mathcal{P}_{\Sigma,S} = \{\mathbb{P} \in \mathcal{P} : \mathrm{Cov}[\mathbb{P}_X] = \Sigma, \mathrm{Cov}[\mathbb{P}_Y] = S\};$$

$\mathcal{P}_\Sigma$ is defined in an analogous way for the one-sample setting.

The first property we require is a form of 2-sided concentration of U around the target quantity:

Assumption IV.1. *For any (Σ, S) and distribution $\mathbb{P} \in \mathcal{P}_{\Sigma,S}$; for any given $\alpha \in (0,1)$ there exist $q_1 = q_1(\Sigma, S, \alpha), q_2 = q_2(\Sigma, S, \alpha)$ in $\mathbb{R}_+$ such that:*

$$\mathbb{P}\big[\big|U - \|\mu - \nu\|^2\big| \geq \|\mu - \nu\| q_1 + q_2\big] \leq \alpha\,. \tag{IV.7}$$

Additionally, we will consider the situation where the quantities q_1, q_2 (which are necessary to find a suitable testing threshold) are not known but must also be estimated from the data; this is the case if the covariance matrices (Σ, S) are unknown. This leads us to our second assumption:

Assumption IV.2. *Suppose Assumption IV.1 holds, with the notation introduced therein. For any $\alpha \in (0,1)$ there exist two estimators $\widehat{Q}_1 = \widehat{Q}_1(\alpha)$ and $\widehat{Q}_2 = \widehat{Q}_2(\alpha)$ in $\mathbb{R}_+$ such that, for any (Σ, S) and distribution $\mathbb{P} \in \mathcal{P}_{\Sigma,S}$:*

$$\mathbb{P}\left[\left|q_1(\Sigma, S, \alpha) - \widehat{Q}_1(\alpha)\right| \geq \frac{1}{2} q_1(\Sigma, S, \alpha)\right] \leq \alpha\,, \tag{IV.8}$$

$$\mathbb{P}\left[\left|q_2(\Sigma, S, \alpha) - \widehat{Q}_2(\alpha)\right| \geq \frac{1}{2} q_2(\Sigma, S, \alpha)\right] \leq \alpha\,. \tag{IV.9}$$

(In the "oracle" case where the covariances Σ, S are assumed to be known, of course Assumption IV.2 is trivially satisfied taking $\widehat{Q}_1 = q_1$, $\widehat{Q}_2 = q_2$.) The following generic result transforms the above assumptions into an estimate of the separation distance for setting (IV.2).

Theorem IV.3. *Let $\mathcal{P}$ be a statistical model for setting (IV.2), and U be a statistic. Let Assumptions IV.1 and IV.2 be granted. Given $\eta \geq 0$ and $\alpha \in (0,1)$, let T be the test defined by*

$$T = 1\Big\{U - \eta^2 > 2\eta \widehat{Q}_1(\alpha) + 2\widehat{Q}_2(\alpha)\Big\}\,. \tag{IV.10}$$

Then for any (Σ, S), provided

$$\delta > 2q_1 + \min\big(2\sqrt{q_2}, 2\eta^{-1} q_2\big)\,, \tag{IV.11}$$

it holds, for any distribution $\mathbb{P} \in \mathcal{P}_{\Sigma,S}$:

$$\mathbb{P}[T = 1] \leq 3\alpha\,, \ \text{if } \mathbb{P} \text{ satisfies } (H_0(\eta));$$
$$\mathbb{P}[T = 0] \leq 3\alpha\,, \ \text{if } \mathbb{P} \text{ satisfies } (H_1(\eta, \delta)).$$

IV.2.2 Concentration properties of the test statistic

The rest of this section is dedicated to establishing the validity of Assumptions IV.1 and IV.2 for the following statistic $U(\mathbb{X}, \mathbb{Y})$:

$$U(\mathbb{X}, \mathbb{Y}) := \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \langle X_i, X_j \rangle + \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \langle Y_i, Y_j \rangle - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \langle X_i, Y_j \rangle \,. \tag{IV.12}$$

Observe that provided expectations μ, ν exist, $U(\mathbb{X}, \mathbb{Y})$ is an unbiased estimator of $\|\mu - \nu\|^2$. In the KME setting as described in Section IV.1.3, inner products are replaced by kernel evaluations and the above statistic is the standard unbiased estimate of the squared MMD between $\mathbb{P}_X$ and $\mathbb{P}_Y$. As announced previously, we will concentrate on the following two settings:

Definition IV.4 (Gaussian setting)**.** The samples $\mathbb{X}$ and $\mathbb{Y}$ are i.i.d. Gaussian in $\mathbb{R}^d$ of marginal distributions $\mathbb{P}_X = \mathcal{N}(\mu, \Sigma)$ and $\mathbb{P}_Y = \mathcal{N}(\nu, S)$, respectively.

In the Gaussian setting, we will assume a finite ambient dimension d for technical reasons: our proofs rely on the Gauss-Lipschitz concentration inequality, which applies in finite dimension. As will appear clearly however, all our results to come are dimension-free in the sense that d never enters the picture, instead only norms of Σ, S come into play. We surmise that our results would apply as well in the same form in the Hilbert-valued setting provided $\mathrm{Tr}(\Sigma)$ and $\mathrm{Tr}(S)$ are finite, but did not try to write down a precise approximation argument to this end.

Definition IV.5 (Bounded setting)**.** The samples $\mathbb{X}$ and $\mathbb{Y}$ are i.i.d. in a separable Hilbert space $\mathcal{H}$ with norm bounded by $L > 0$. The covariance operators for the marginal sample distributions are denoted Σ and S, respectively; observe that they have finite trace by the boundedness assumption.

Propositions IV.6 and IV.9 give concentration bounds for the statistic U, ensuring Assumption IV.1 in the two above settings.

Proposition IV.6. *Assume the Gaussian setting holds and $n, m \geq 2$. Then with probability at least $1 - \alpha$,*

$$\left| U - \|\mu - \nu\|^2 \right| \leq \|\mu - \nu\| q_1 + q_2 \,, \tag{IV.13}$$

where U is defined in (IV.12) and

$$q_1(\Sigma, S, \alpha) = \sqrt{2 \left(\frac{\|\Sigma\|_{\mathrm{op}}}{n} + \frac{\|S\|_{\mathrm{op}}}{m} \right) u(\alpha)} \,, \tag{IV.14}$$

$$q_2(\Sigma, S, \alpha) = 32 \left(\frac{\sqrt{\mathrm{Tr}\, \Sigma^2}}{n} + \frac{\sqrt{\mathrm{Tr}\, S^2}}{m} \right) u(\alpha) \,. \tag{IV.15}$$

where $u(\alpha) := -\log \alpha + \log 8$.

Let us simplify somewhat the above expression when plugged into Theorem IV.3 in the case of signal detection (IV.1). We also give a matching lower bound (up to constant factor) for the optimal separation distance.

Theorem IV.7. *Consider the signal detection problem (IV.1) and assume the Gaussian setting with covariance matrix Σ. Then the minimum separation distance δ^* given by (IV.4) so that the type I and II errors for problem (IV.1) are less that $\alpha \in (0, 1)$ for all $\mathbb{P} \in \mathcal{P}_\Sigma$ is upper bounded by*

$$\delta^*(\alpha, \Sigma, \eta) \lesssim \sigma_n \sqrt{u} \max\left(1, \min\left(d_*^{\frac{1}{4}}, \sqrt{d_* u} \cdot \frac{\sigma_n}{\eta} \right) \right), \tag{IV.16}$$

where $u(\alpha) := -\log\alpha + \log 60$. *If* $d_* \geq 3$, *then it is lower bounded by*

$$\delta^*(\alpha, \Sigma, \eta) \geq \sigma_n \sqrt{\frac{1-\alpha}{12}} \max\left(1, \min\left(d_*^{\frac{1}{4}}, \sqrt{d_*(1-\alpha)} \cdot \frac{\sigma_n}{\eta}\right)\right), \tag{IV.17}$$

where $\sigma_n^2 := \|\Sigma\|_{\mathrm{op}}/n$, *and* $d_* := \operatorname{Tr}\Sigma^2/\|\Sigma\|_{\mathrm{op}}^2$. *(The symbol* $\lesssim$ *indicates inequality up to a numerical factor).*

Observe that it holds $d_* \leq d_e$, where $d_e = \operatorname{Tr}\Sigma/\sigma^2$ is the "effective dimensionality" coming into play for signal estimation rates (namely $\mathbb{E}\left[\|\overline{X} - \mu\|^2\right]^{1/2} = \sigma\sqrt{d_e/n}$, where $\overline{X}$ is the empirical mean). In the finite d-dimensional case with $\Sigma = I_d$, it holds $d = d_e = d_*$, and the separation (IV.16) has been shown to be optimal in the Gaussian setting for $\eta = 0$ by Baraud (2002) and for any $\eta \geq 0$ by Blanchard et al. (2018). It exhibits a continuous transition between the signal detection setting ($\eta = 0$, $\delta^* \simeq d^{1/4}\sigma/\sqrt{n}$) and the hyperplane testing setting (which is equivalent to the 1-dimensional setting by rotational invariance; $\eta \to \infty$, $\delta^* \simeq \sigma/\sqrt{n}$). In that particular situation, we observe that the signal separation distance is smaller by a factor $d^{1/4}$ than the signal estimation error, a phenomenon typical of high-dimensional statistics. In the more general setting studied here where Σ can be arbitrary, this difference between rates can be all the more marked, since in addition d_* can be much smaller than d_e.

We obtain a similar result for the two-sample problem:

Theorem IV.8. *Consider the two-sample mean problem* (IV.2) *and assume the Gaussian setting with covariance matrices* Σ, S. *Then the minimum separation distance* δ^* *so that the type I and II errors for problem* (IV.2) *is less than* $\alpha \in (0,1)$ *for all* $\mathbb{P} \in \mathcal{P}_{\Sigma,S}$ *is upper bounded by*

$$\delta^*(\alpha, \Sigma, S, \eta) \lesssim \sigma_{n,m}\sqrt{u} \max\left(1, \min\left(d_*^{\frac{1}{4}}, \sqrt{d_* u} \cdot \frac{\sigma_{n,m}}{\eta}\right)\right), \tag{IV.18}$$

where $u := -\log\alpha + \log 60$. *If* $d_* \geq 3$, *then it is lower bounded by*

$$\delta^*(\alpha, \Sigma, S, \eta) \geq \sigma_{n,m}\sqrt{\frac{1-\alpha}{48}} \max\left(1, \min\left(d_*^{\frac{1}{4}}, \sqrt{d_*(1-\alpha)} \cdot \frac{\sigma_{n,m}}{\eta}\right)\right), \tag{IV.19}$$

where $\sigma_{n,m}^2 := \|M_{n,m}\|_{\mathrm{op}}$, *and* $d_* := \operatorname{Tr} M_{n,m}^2/\sigma_{n,m}^4$, *for* $M_{n,m} := \Sigma/n + S/m$. *(The symbol* $\lesssim$ *indicates inequality up to a numerical factor).*

Here the effective dimension d^* depends on the two covariance matrices Σ and S, weighted by the size of the samples.

Remark. As mentioned in the introduction, by letting m go to infinity in the two-sample case, we recover the bounds of the one sample case (up to a constant factor). It is worth examining if the converse holds, i.e. if there is an argument to reduce the two-sample problem to the simpler one-sample case (this would simplify some technical aspects of the proofs, somewhat). For the *upper* bounds on the minimum separation distance, this is the case in some specific situations: for equal sample sizes $n = m$, the two-sample case can be reduced to the one-sample problem setting by pairing the samples and considering the single sample $(X_i - Y_i)_{1\leq i\leq n}$, and one can recover this way in essence the two-sample result. If $\Sigma = S$, and for general sample sizes, we can also reduce to the single sample with size $\min(m,n)$, $(X_i - Y_i)_{1\leq i\leq \min(m,n)}$, and recover again the two-sample result up to a numerical factor. However a reduction argument in the general case has eluded us. Concerning the *lower* bound, the argument for the two-sample case indeed hinges on a reduction the one-sample case, by considering the sub-models where one of the two sample means is known, see Section IV.3.4.

We now turn to the bounded setting.

Proposition IV.9. *Assume the bounded setting holds and $n, m \geq 2$. Then with probability at least $1-\alpha$,*

$$\left|U - \|\mu-\nu\|^2\right| \leq \|\mu-\nu\| q_1 + q_2 \,, \tag{IV.20}$$

where U is defined in (IV.12) and

$$q_1(\Sigma, S, \alpha) = 2\sqrt{2\left(\frac{\|\Sigma\|_{\mathrm{op}}}{n} + \frac{\|S\|_{\mathrm{op}}}{m}\right)u} + \frac{4Lu}{3(n \wedge m)} \,,$$
$$q_2(\Sigma, S, \alpha) = 614\left(\frac{\sqrt{\operatorname{Tr}\Sigma^2}}{n} + \frac{\sqrt{\operatorname{Tr} S^2}}{m}\right)u + 3708\frac{L^2u^2}{(n\wedge m)^2} \,,$$

with $u(\alpha) = -\log\alpha + \log 2$.

Thus, in the bounded setting we can guarantee that the behavior of the test is qualitatively the same as in the Gaussian setting (see e.g. Theorem IV.7) — and this from a non-asymptotic point view, provided $n \wedge m \geq uL^2/\sigma^2$, where $\sigma^2 = \|\Sigma\|_{\mathrm{op}}$.

A special case of interest is when the data lies on the sphere of radius L, i.e. $\|X_i\| = \|Y_j\| = L$ a.s. In this case $L^2 = \operatorname{Tr}\Sigma$ and the above condition can be rewritten $n \wedge m \geq u d_e$. This situation is met in particular in the KME setting, see Section IV.1.3, when using a translation-invariant kernel $k(z, z') = k_\circ(z - z')$, in which case $L^2 = k_\circ(0)$.

IV.2.3 Quantile estimation

Since we are considering the case where Σ, S can be arbitrary in this work, it is natural to assume that these are not known in advance. We study next the estimation of the quantities q_1 and q_2, in both settings (bounded and Gaussian), in order to check Assumption IV.2 for our generic theorem. If we can grant that assumption, Theorem IV.3 guarantees that the separation distance remains qualitatively the same as in the "oracle" situation where they are known. To simplify the exposition, in this section we will present results for the one-sample problem only; similar results, although slightly more technical, can be obtained for the two-sample problem. Thus, we need to have estimators of $\|\Sigma\|_{\mathrm{op}}$ and $\operatorname{Tr}\Sigma^2$ — more precisely, of their square root.

For q_1, we will use the empirical covariance operator $\widehat{\Sigma} := \widehat{\Sigma}(\mathbb{X})$:

$$\widehat{\Sigma}(\mathbb{X}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})(X_i - \widehat{\mu})^T \,, \tag{IV.21}$$

where $\widehat{\mu} := \widehat{\mu}(\mathbb{X})$ is the empirical mean of the sample $\mathbb{X}$.

Proposition IV.10 (Gaussian setting)**.** *Assume $\mathbb{X} = (X_i)_{1\leq i\leq n}$ are i.i.d. Gaussian vectors of covariance Σ. For $u \geq 0$, with probability at least $1 - 3e^{-u}$:*

$$\left|\left\|\widehat{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\right| \leq 3\sqrt{2}\|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\left(\sqrt{\frac{d_e}{n}} + \sqrt{\frac{u}{n}}\right) , \tag{IV.22}$$

where $\widehat{\Sigma}$ is defined in (IV.21) and $d_e = \operatorname{Tr}\Sigma/\|\Sigma\|_{\mathrm{op}}$.

Proposition IV.11 (Bounded setting)**.** *Assume that $\mathbb{X} = (X_i)_{1\leq i\leq n}$ are i.i.d. bounded in norm by L and with covariance Σ. For $u \geq 0$, with probability at least $1 - 2e^{-u}$:*

$$\left|\left\|\widehat{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\right| \leq 4L\left(2\sqrt{\frac{d_e}{n}} + \sqrt{\frac{2u}{n}} + \frac{u}{3n}\right) \tag{IV.23}$$

where $\widehat{\Sigma}$ is defined in (IV.21) and $d_e = \operatorname{Tr}\Sigma/\|\Sigma\|_{\mathrm{op}}$.

These concentration bounds are not sharp in an asymptotic sense, where the main term for the scaling of the deviations is expected to follow that of asymptotic normality for eigenvalues of the empirical covariance operators, as in the classical results of Anderson (2003), but they are largely sufficient for our purposes (see Corollary IV.14 below). Some refined related nonasymptotic bounds can be found in the recent literature. In particular, Koltchinskii and Lounici (2017) derive nonasymptotic results for controlling $\left\|\widehat{\Sigma} - \Sigma\right\|$ in the Gaussian setting, and in the centered case where $\mu = 0$ is known. In fact, in essence the result of our technical Proposition IV.23 in the proof section (which is like Proposition IV.11 but in the centered case) can be deduced from the results of Koltchinskii and Lounici (2017) by elementary arguments. We decided to include a standalone proof here; while we do rely on the estimates of Koltchinskii and Lounici (2017) (or rather on the improved version of van Handel, 2017) for the expectation of the difference, we derive an upper bound on the deviation by a rather direct application of the Gauss-Lipschitz concentration. While Koltchinskii and Lounici (2017) also rely on such arguments, their proofs are much more involved, for the reason that they study the norm or the difference while we only are interested in the difference of the (root) norms here. Finally, we also mention very recent results of Jirak and Wahl (2018) for sharp nonasymptotic control of spectral quantities related to Σ, which could also potentially be applied here, though it seems at first glance that a logarithmic dependence in the dimension could enter into play.

For the bounded setting (Proposition IV.11), the bound (IV.23) could presumably be improved to have $\sqrt{\|\Sigma\|_{\mathrm{op}}}$ instead of L for the main terms. The results of Theorem 9 of Koltchinskii and Lounici (2017) under a sub-Gaussian assumption do not seem to be able to imply Proposition IV.11, see the more detailed discussion below in Section IV.2.4.

Turning now to q_2, we will estimate $\sqrt{\operatorname{Tr}\Sigma^2}$ using the following statistic $\widehat{T} := \widehat{T}(\mathbb{X})$, which is an unbiased estimator of $\operatorname{Tr}\Sigma^2$:

$$\widehat{T}(\mathbb{X}) := \frac{1}{4n(n-1)(n-2)(n-3)} \sum_{i\neq j\neq k\neq l} \langle X_i - X_k, X_j - X_l\rangle^2 . \tag{IV.24}$$

Proposition IV.12 (Gaussian setting)**.** *Assume $\mathbb{X} = (X_i)_{1\leq i\leq n}$ are i.i.d. Gaussian vectors of covariance Σ and $n \geq 4$. Then for all $u \geq 0$:*

$$\mathbb{P}\left[\left|\sqrt{\widehat{T}} - \sqrt{\operatorname{Tr}\Sigma^2}\right| \geq 30\sqrt{\frac{\operatorname{Tr}\Sigma^2}{n}u^2}\right] \leq e^4 e^{-u} , \tag{IV.25}$$

where $\widehat{T}$ is defined in (IV.24).

Proposition IV.13 (Bounded setting)**.** *Assume that $\mathbb{X} = (X_i)_{1\leq i\leq n}$ are i.i.d. bounded in norm by L and with covariance Σ and $n \geq 4$. Then for all $u \geq 0$:*

$$\mathbb{P}\left[\left|\sqrt{\widehat{T}} - \sqrt{\operatorname{Tr}\Sigma^2}\right| \geq 12L^2\sqrt{\frac{u}{n}}\right] \leq 2e^{-u} . \tag{IV.26}$$

where $\widehat{T}$ is defined in (IV.24).

Thanks to these concentration results, we can construct estimators of $q_1(\Sigma,\alpha)$ and $q_2(\Sigma,\alpha)$ satisfying Assumption IV.2. In the Gaussian setting, we give the following explicit corollary of Propositions IV.10 and IV.12; the proof is straightforward and omitted.

Corollary IV.14 (Gaussian setting)**.** *Consider the signal detection problem* (IV.1) *and assume the Gaussian setting holds. Let $\alpha \in (0,1)$, $u = u(\alpha) = -\log\alpha + \log 8$, and $\widehat{Q}_1(\alpha)$ and $\widehat{Q}_2(\alpha)$ be the statistics defined by*

$$\widehat{Q}_1(\alpha) = \sqrt{\frac{2\left\|\widehat{\Sigma}(\mathbb{X})\right\|_{\mathrm{op}}}{n}u} , \quad \widehat{Q}_2(u) = 32\frac{\sqrt{\widehat{T}(\mathbb{X})}}{n}u ,$$

where $\widehat{\Sigma}$ is defined in (IV.21) and $\widehat{T}$ in (IV.24). Then for any Σ, provided

$$n \gtrsim \max\big(d_e(\Sigma), u, u^4\big)\,, \tag{IV.27}$$

(we recall $d_e(\Sigma) = \operatorname{Tr}\Sigma/\|\Sigma\|_{\mathrm{op}}$), then it holds, for any distribution $\mathbb{P} \in \mathcal{P}_\Sigma$:

$$\mathbb{P}\Big[\Big|\widehat{Q}_1(\alpha) - q_1(\Sigma,\alpha)\Big| \le q_1(\Sigma,\alpha)/2\Big] \le \alpha,$$
$$\mathbb{P}\Big[\Big|\widehat{Q}_2(\alpha) - q_2(\Sigma,\alpha)\Big| \le q_2(\Sigma,\alpha)/2\Big] \le \alpha,$$

where q_1, q_2 are as defined in (IV.14),(IV.15) (with $m = \infty$).

The condition (IV.27) for n is needed to grant Assumption IV.2: it ensures that the deviations of the estimators $\|\widehat{\Sigma}\|_{\mathrm{op}}^{1/2}$ and $\widehat{T}$ coming from Proposition IV.10 and IV.12 are smaller than their target quantities $\|\Sigma\|_{\mathrm{op}}^{1/2}/2$ and $(\operatorname{Tr}\Sigma^2)^{1/2}/2$, respectively. The requirement that the size of the sample is larger than the effective dimension d_e appears mild.

For the bounded setting and the signal detection problem (IV.1), estimators $\widehat{Q}_1$ and $\widehat{Q}_2$ satisfying Assumption IV.2 can also be constructed in a similar way from Propositions IV.11 and IV.13 (details omitted). In the bounded setting, the quantiles q_1 and q_2 of U are composed of two terms, the first (and larger) one gives the dependence in the covariance of the distribution, the second depends on the bound L. This additional term will have to be taken into account, and the condition on n analogous to (IV.27) will involve L. In general this will not be a problem since L or an upper bound on L is supposed to be known, as is the case for instance in the kernel setting (see the concluding discusssion in the previous section). Finally, for the two-sample test problem (IV.2), comparable results can be obtained using the estimators $\widehat{\Sigma}(\mathbb{Y})$ and $\widehat{T}(\mathbb{Y})$; we omit the details.

IV.2.4 Concluding remarks

A technical discussion point: Gaussian, sub-Gaussian, and bounded vectors. The utility of our systematic distinction between the Gaussian and bounded case can be disputed in the light of recent concentration literature (see e.g. Hsu et al., 2012; Koltchinskii and Lounici, 2017 and further references therein) deriving results holding for sub-Gaussian random vectors, a seemingly more general setting emcompassing both the Gaussian and bounded settings as particular cases (since bounded variables are sub-Gaussian by Hoeffding's inequality).

This point deserves a specific discussion. The sub-Gaussianity assumption for a vector variable X (assumed centered for simplicity here) often takes the following form: for any unit vector u, denoting $X_u = \langle X, u\rangle$, it is assumed that $\|X_u\|_{\psi_2} \le C\sqrt{\operatorname{Var}[X_u]}$ (where $\|.\|_{\psi_2}$ is the Orlicz ψ_2-norm); or equivalently in terms of Laplace transform,

$$\log(\mathbb{E}[\exp\lambda(X_u)]) \le (C')^2\lambda^2\operatorname{Var}[X_u]/2 \text{ for all } \lambda \ge 0. \tag{IV.28}$$

A key point is that the factors C or C' in those definitions should be independent of u, and they generally appear as global factors in the derived deviation inequalities. If the only information we have is that $\|X\|$ is bounded a.s. by L, we see that the factors C or C' should be taken of the order of $\sup_{\|u\|=1}(L/\sqrt{\operatorname{Var}[X_u]}) = L\|\Sigma^{-1}\|_{\mathrm{op}}^{1/2}$, which is not acceptable in a high-dimensional setting, and in particular for the application to KME described in Section IV.1.3, where one might expect that $\|\Sigma^{-1}\|_{\mathrm{op}}$ can get arbitrarily large or even infinite.

Some works (such as Spokoiny and Zhilova, 2013 and the appendix of Spokoiny and Dickhaus, 2015) consider settings going beyond sub-Gaussianity, i.e. when (IV.28) is only required to hold for $\lambda \le M^{-1}$. This allows in principle for more general variables, e.g. chi-squared type statistics or variables admitting Berstein- or Bennett-type control of their Laplace transform, while making the constant C' in (IV.28)

controlled by a fixed numerical constant. Under this assumption the "first-order" terms are of the correct order, i.e. typically only depend on the variance Σ. Unfortunately, the value of M comes up into additional terms, and since its value has to be independent of u, in the bounded setting the uniformity with respect to u means that M should be again taken of the order of $\|\Sigma^{-1}\|_{\mathrm{op}}^{1/2}$.

To summarize, despite our best efforts we were not able to derive from existing general results, working under the (possibly extended) sub-Gaussian assumption, a concentration in the bounded setting that would not involve $\|\Sigma^{-1}\|_{\mathrm{op}}$, and this is the reason why we treated it separately with tools specific to bounded variables such as the Bousquet-Talagrand inequality. It would be of course of notable interest to obtain results under a general sub-Gaussian assumption $\sup_{\|u\|=1}\|X_u\|_{\psi_2} \leq L$, and control deviations only involving various norms of Σ for the main terms, possibly L for smaller-order terms, but not depending on $\|\Sigma^{-1}\|_{\mathrm{op}}$.

Perspectives. We finally list a few items for future developments.

- It would be interesting to obtain a version of Proposition (IV.11) where the main term does not involve the bound L.
- A recent trend of research developed "robust" exponential concentration bounds for estimators of scalars and vectors with minimal moments assumptions (see e.g. Lugosi and Mendelson, 2019a for a survey of recent advances). It seems a very interesting question to study if such robust procedures can be pushed to the testing setting and enjoy similar nonasymptotic controls to the Gaussian and bounded settings under much relaxed distributional assumptions. Preliminary calculations seem to indicate that the "median-of-means" (MoM) approach can be applied to U-statistics without particular problems and that Assumption IV.1 can be granted for MoM versions of U-statistics under the assumption of existing moments of order 4, and presumably Assumption IV.2 under moments of order 8.
- We have analyzed here quantile estimation by direct estimation of unknown quantities coming into the quantile bounds. In practice, quantile estimation by some form of resampling procedure would be often sharper and preferred. V. Spokoiny also made notable recent contributions to this topic (Naumov et al., 2019; Spokoiny and Zhilova, 2015). In the setting of two-sample testing where the null hypothesis is strict equality, it is possible to obtain tests with exact nonasymptotic level based on permutation tests and variations thereof; see Fromont et al. (2012) for such approaches for testing equality of distributions based on the KME methodology, and Kim et al. (2020) for recent broad results on minimax optimality for the power of permutation-based tests. Estimating quantiles via bootstrap procedures is also an interesting direction to pursue in setting, in the case where the null hypothesis is based on closeness rather than equality of signals, so that exact permutation tests do not apply; Dette et al. (2020a) recently proposed nonstandard bootstrap procedures to tackle this issue.
- Lower bounds establishing the optimality of the separation rates appearing have been established in the Gaussian case in Theorem IV.7. It would be nice find such a lower bound in the bounded case.

IV.3 Proofs for Section IV

The proofs of some of the technical results, first stated without justification along the text, can be found in Section IV.3.7. We first state a standard technical lemma which we will use several times in the following proofs.

Lemma IV.15. *Let $a \in \mathbb{R}_+$ and $b \in \mathbb{R}$, then*

$$-\min\left(\sqrt{b}, \frac{|b|}{a}\right) \leq \sqrt{(a^2+b)_+} - a \leq \min\left(\sqrt{|b|}, \frac{|b|}{2a}\right). \tag{IV.29}$$

IV.3.1 Proof of Theorem IV.3

Let us denote $D := \|\mu - \nu\|$. Under (H_0) we have $D \leq \eta$ and thus:

$$\begin{aligned}\mathbb{E}_{H_0}[T] &= \mathbb{P}_{H_0}\left[U - \eta^2 > 2\eta\widehat{Q}_1 + 2\widehat{Q}_2\right] \\ &\leq \mathbb{P}_{H_0}\left[U > D^2 + Dq_1 + q_2\right] \\ &\qquad + \mathbb{P}_{H_0}\left[\left|q_1 - \widehat{Q}_1\right| > q_1/2\right] + \mathbb{P}_{H_0}\left[\left|q_2 - \widehat{Q}_2\right| > q_2/2\right] \\ &\leq 3\alpha\,,\end{aligned}$$

where we have used Assumptions IV.1 and IV.2.

We will prove below that under (H_1), we have

$$\mathbb{P}_{H_1}\left[D^2 - Dq_1(u) - q_2(u) \leq \eta^2 + \eta 2\widehat{Q}_1 + 2\widehat{Q}_2\right] \leq 2\alpha\,, \tag{IV.30}$$

which entails:

$$\begin{aligned}\mathbb{P}_{H_1}[T = 0] &= \mathbb{P}_{H_1}\left[U - \eta^2 \leq 2\eta\widehat{Q}_1 + 2\widehat{Q}_2\right] \\ &\leq \mathbb{P}_{H_1}\left[U \leq D^2 - Dq_1 - q_2\right] + \mathbb{P}_{H_1}\left[D^2 - Dq_1 - q_2 \leq \eta^2 + \eta 2\widehat{Q}_1 + 2\widehat{Q}_2\right] \\ &\leq 3\alpha\,,\end{aligned}$$

and the proof is complete. We now prove inequality (IV.30). Let us first solve the following quadratic inequality in $Z \geq 0$:

$$Z^2 - Zq_1 - q_2 \geq \eta^2 + 3\eta q_1 + 3q_2\,. \tag{IV.31}$$

The equation is satisfied when

$$Z \geq \frac{q_1 + \sqrt{(2\eta + 3q_1)^2 + 16q_2}}{2}\,;$$

furthermore, by Lemma IV.15 and the assumed inequality (IV.11), we have that

$$\frac{q_1 + \sqrt{(2\eta + 3q_1)^2 + 16q_2}}{2} \leq \eta + 2q_1 + \min\left(2\sqrt{q_2}, \frac{2q_2}{\eta}\right) \leq \eta + \delta\,.$$

Under (H_1), $D \geq \eta + \delta$, so D satisfies equation (IV.31). We conclude by remarking that, using Assumption IV.2:

$$\begin{aligned}\mathbb{P}_{H_1}\left[D^2 - Dq_1(u) - q_2(u) \leq \eta^2 + \eta 2\widehat{Q}_1 + 2\widehat{Q}_2\right] &\leq \mathbb{P}[H_1]\eta^2 + \eta 2\widehat{Q}_1 + 2\widehat{Q}_2 \geq \eta^2 + 3\eta q_1 + 3q_2 \\ &\leq 2\alpha\,.\end{aligned}$$

□

IV.3.2 Proof of Propositions IV.6 and IV.9

As much for the Gaussian case as for the bounded case, we will give concentration bounds for the statistic U defined in (IV.12), by decomposing the statistic in four parts. Let us define:

$$U_{\mathbb{X}} := \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \langle X_i - \mu, X_j - \mu \rangle\,, \qquad U_{\mathbb{Y}} := \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \langle Y_i - \nu, Y_j - \nu \rangle\,,$$

$$U_{\mathbb{X},\mathbb{Y}} := \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \langle X_i - \mu, Y_j - \nu \rangle, \qquad U_* := \left\langle \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu) - \frac{1}{m} \sum_{j=1}^{m} (Y_j - \nu), \mu - \nu \right\rangle.$$

We have that

$$U = \|\mu - \nu\|^2 - 2U_* + U_{\mathbb{X}} + U_{\mathbb{Y}} - 2U_{\mathbb{X},\mathbb{Y}}\,. \tag{IV.32}$$

Gaussian setting. We first need some results on Gaussian variables. The first result is a decoupling theorem of Vershynin (2018).

Proposition IV.16 (Vershynin, 2018, Theorem 6.1.1)**.** *Let $X_1, \dots, X_n$ be independent centered and weakly (i.e. Pettis) integrable vectors in a Hilbert space, $(a_{ij})_{1\le i,j\le n}$ a family of real numbers and $F : \mathbb{R} \mapsto \mathbb{R}$ a convex function. Then*

$$\mathbb{E}\left[F\left(\sum_{i\neq j} a_{ij}\langle X_i, X_j\rangle\right)\right] \le \mathbb{E}\left[F\left(4\sum_{i,j} a_{ij}\langle X_i, X'_j\rangle\right)\right],$$

where (X'_i) is an independent copy of (X_i).

The following lemma is standard; see e.g. Birgé (2001), Lemma 8.2.

Lemma IV.17. *Let X a real random variable such that for all $0 < t < b^{-1}$:*

$$\log\big(\mathbb{E}\big[e^{tX}\big]\big) \le \frac{(at)^2}{1-bt},$$

where a and b are two positive constants. Then, for all $t \ge 0$:

$$\mathbb{P}\Big[X \ge 2a\sqrt{t} + bt\Big] \le e^{-t}.$$

Proposition IV.18. *Let X and Y be two independent Gaussian vectors following the distributions $\mathcal{N}(0,\Sigma)$ and $\mathcal{N}(0,S)$ respectively. Then for $t < (\|S\|_{\mathrm{op}}\|\Sigma\|_{\mathrm{op}})^{-1/2}$:*

$$\log \mathbb{E}[\exp(t\langle X, Y\rangle)] \le \frac{t^2 \operatorname{Tr}(S\Sigma)}{2(1 - t\sqrt{\|S\|_{\mathrm{op}}\|\Sigma\|_{\mathrm{op}}})}.$$

Using Lemma IV.17, for all $u \ge 0$:

$$\mathbb{P}\left[\langle X, Y\rangle \ge \sqrt{2\operatorname{Tr}(S\Sigma)u} + \sqrt{\|S\|_{\mathrm{op}}\|\Sigma\|_{\mathrm{op}}}u\right] \le e^{-u}.$$

We can now prove Proposition IV.6. The samples $\mathbb{X}$ and $\mathbb{Y}$ have respective distributions $\mathcal{N}(\mu,\Sigma)$ and $\mathcal{N}(\nu,S)$. We will obtain a concentration inequality for U using its decomposition (IV.32).

Let us first find concentration inequalities for $U_{\mathbb{X}}$ and $U_{\mathbb{Y}}$. Using decoupling (see Proposition IV.16) we have for all $t < (4\|\Sigma\|_{\mathrm{op}})^{-1}$:

$$\mathbb{E}[\exp(tn(n-1)U_{\mathbb{X}})] \le \mathbb{E}\left[\exp\left(4t\left\langle \sum_{i=1}^n X_i - \mu, \sum_{i=1}^n X'_i - \mu\right\rangle\right)\right],$$

where X'_i are independent copies of the X_is. Then using Proposition IV.18, it holds with probability at least $1 - 2e^{-u}$:

$$n(n-1)|U_{\mathbb{X}}| \le 4n\left(\sqrt{2\operatorname{Tr}\Sigma^2 u} + \|\Sigma\|_{\mathrm{op}}u\right). \tag{IV.33}$$

The same method works for $U_{\mathbb{Y}}$. The concentration of $U_{\mathbb{X},\mathbb{Y}}$ is directly obtained using Proposition IV.18. Finally U_* is a centered 1-dimensional Gaussian with variance $(\mu-\nu)^T\big(\frac{\Sigma}{n} + \frac{S}{m}\big)(\mu-\nu)$ and we use the classical bound $\mathbb{P}\big[|N| \ge \sigma\sqrt{2t}\big] \le 2e^{-t}$ for $N \sim \mathcal{N}(0,\sigma^2)$. Thus we obtain that with probability at least $1 - 8e^{-u}$:

$$\begin{aligned}\big|U - \|\mu-\nu\|^2\big| \le& \frac{4}{n-1}\left(\sqrt{2\operatorname{Tr}\Sigma^2 u} + \|\Sigma\|_{\mathrm{op}}u\right) + \frac{4}{m-1}\left(\sqrt{2\operatorname{Tr} S^2 u} + \|S\|_{\mathrm{op}}u\right)\\ &+ \frac{4}{\sqrt{nm}}\left(\sqrt{2\operatorname{Tr}\Sigma S u} + (\|\Sigma\|_{\mathrm{op}}\|S\|_{\mathrm{op}})^{\frac12}u\right)\\ &+ \sqrt{2(\mu-\nu)^T\left(\frac{\Sigma}{n} + \frac{S}{m}\right)(\mu-\nu)u}.\end{aligned}$$

We conclude by upper bounding the operator norms $\|\Sigma\|_{\mathrm{op}}$ and $\|S\|_{\mathrm{op}}$ by $\sqrt{\operatorname{Tr}\Sigma^2}$ and $\sqrt{\operatorname{Tr} S^2}$ and for the third term we use that

$$(2\operatorname{Tr}(\Sigma S))^{\frac{1}{2}} \leq (4\operatorname{Tr}\Sigma^2 \operatorname{Tr} S^2)^{\frac{1}{4}} \leq (\operatorname{Tr}\Sigma^2)^{\frac{1}{2}} + (\operatorname{Tr} S^2)^{\frac{1}{2}}.$$

We finally use $(n-1)^{-1} \leq 2n^{-1}$ for $n \geq 2$ and similarly for m. It is easy to check that the fourth term is upper bounded by q_1 defined in (IV.14). It just remains to use that $u \geq 1$ to get $u \geq \sqrt{u}$ and (IV.13).

Bounded setting. The concentration of U is obtained in the bounded setting using a concentration inequality for degenerate U-statistics of Houdré and Reynaud-Bouret (2003). We present here a somewhat simplified version suited for our purpose[6].

Theorem IV.19 (Houdré and Reynaud-Bouret, 2003, Theorem 3.4)**.** *Let $T_1, \dots, T_N$ be independent random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a Borel space $(\mathcal{T}, \mathcal{G})$. Let*

$$U_N = \sum_{i=2}^{N} \sum_{j=1}^{i-1} g_{i,j}(T_i, T_j)\,,$$

where $g_{i,j} : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ are measurable Borelian functions satisfying

$$\mathbb{E}[g_{i,j}(T_i, T_j)|T_i] = \mathbb{E}[g_{i,j}(T_i, T_j)|T_j] = 0\,.$$

Let us suppose that the following quantities are finite

$$A := \sup_{t,t',i,j} |g_{i,j}(t,t')|\,,$$

$$B^2 := \max\left\{ \sup_{t,i}\left(\sum_{j=1}^{i-1} \mathbb{E}[g_{i,j}(t,T_j)^2]\right), \sup_{t,j}\left(\sum_{i=j+1}^{n} \mathbb{E}[g_{i,j}(T_i,t)^2]\right)\right\},$$

$$C^2 := \sum_{i=2}^{N}\sum_{j=1}^{i-1} \mathbb{E}[g_{i,j}(T_i,T_j)^2]\,.$$

Then for all $u > 0$:

$$\mathbb{P}\left[U_N \geq 4C(\sqrt{2u} + 2\sqrt{2}u) + 202Bu^{3/2} + 196Au^2\right] \leq 2.77e^{-u}\,. \tag{IV.34}$$

Let us prove Proposition IV.9. We recall that we suppose here that the samples $\mathbb{X}$ and $\mathbb{Y}$ are both bounded by L. To obtain a deviation inequality for the statistic U, we consider separately the statistics $U_{\mathbb{X}} + U_{\mathbb{Y}} - 2U_{\mathbb{X},\mathbb{Y}}$ and then U_*.

Using Theorem IV.19 with $N = n+m$, $T_i := X_i - \mu$ for $1 \leq i \leq n$ and $T_i = Y_i - \nu$ for $n+1 \leq i \leq n+m$, $\mathcal{T} = \{u : \|u\| \leq 4L^2\}$ and

$$g_{ij}(\cdot,\cdot) = \begin{cases} \frac{1}{n(n-1)}\langle\cdot,\cdot\rangle, & \text{if } 1 \leq i,j \leq n,\\ \frac{1}{m(m-1)}\langle\cdot,\cdot\rangle, & \text{if } n+1 \leq i,j \leq n+m,\\ -\frac{1}{nm}\langle\cdot,\cdot\rangle, & \text{otherwise,}\end{cases}$$

we get that with probability greater than $1 - 5.54e^{-u}$:

$$|U_{\mathbb{X}} + U_{\mathbb{Y}} - 2U_{\mathbb{X},\mathbb{Y}}|/2 \leq 307\left(\frac{\sqrt{\operatorname{Tr}\Sigma^2}}{n} + \frac{\sqrt{\operatorname{Tr} S^2}}{m}\right)u + 1854L^2u^2\,. \tag{IV.35}$$

[6] In the original result the u deviation term involves an additional constant D and we simply use $D \leq C$ here.

To obtain the above, we have upper bounded A, B, C by:

$$A \leq \frac{8L^2}{(n \wedge m)^2}, \quad B^2 \leq \frac{8L^2}{(n \wedge m)^2}\left(\frac{\|\Sigma\|_{\mathrm{op}}}{n} + \frac{\|S\|_{\mathrm{op}}}{m}\right), \quad C^2 = \frac{3}{2}\left(\frac{\operatorname{Tr}\Sigma^2}{n} + \frac{\operatorname{Tr} S^2}{m}\right);$$

then, using that $2\sqrt{ab} \leq a + b$ and that $\|\Sigma\|_{\mathrm{op}} \leq \sqrt{\operatorname{Tr}\Sigma^2}$, we get (IV.35).

For U_*, we use Bernstein's inequality (i.e. combining Lemmas IV.32 and IV.17) to get that with probability at least $1 - 2e^{-u}$, it holds:

$$|U_*| \leq \|\mu - \nu\|\left(\sqrt{2\left(\frac{\|\Sigma\|_{\mathrm{op}}}{n} + \frac{\|S\|_{\mathrm{op}}}{m}\right)u} + \frac{2Lu}{3n \wedge m}\right). \tag{IV.36}$$

Combining (IV.35) and (IV.36), we obtain the claim of Proposition IV.9. □

IV.3.3 Proof of Theorem IV.7

The upper bound is directly obtained using Theorem IV.3. Assumption IV.1 is satisfied as a consequence of Proposition IV.6. We do not consider estimation of nuisance parameters related to the covariance matrix Σ which is assumed to be fixed and known for this result; thus Assumption IV.2 is trivially satisfied by taking $\widehat{Q}_1 = q_1(\Sigma, \alpha)$, $\widehat{Q}_2 = q_2(\Sigma, \alpha)$.

Let us now prove the lower bound (IV.17). The following proof is an adaptation to the non-isotropic Gaussian setting of the proof of Theorem 5.1 in Blanchard et al. (2018). Let $\alpha \in (0, 1)$, ans Σ be a positive semidefinite matrix. Without loss of generality, we can assume that Σ is diagonal: $\Sigma = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_1 \geq \ldots \geq \lambda_d > 0$. Let us denote $\mathbb{P}_{\mu,\Sigma}$ the distribution of $\mathcal{N}(\mu, \Sigma)$ for $\mu \in \mathbb{R}^d$ and introduce the Gaussian mixture distribution:

$$\mathbb{Q}^n_\Sigma := \frac{1}{2^{d-1}} \sum_{m \in \mathcal{M}} \mathbb{P}^{\otimes n}_{m,\Sigma}, \tag{IV.37}$$

where

$$\mathcal{M} = \left\{(\lambda_1 v_1 h, \ldots, \lambda_{d-1} v_{d-1} h, \eta) \,|\, v \in \{-1, 1\}^{d-1}\right\}.$$

We take $h^2 := \frac{(\eta+\delta)^2 - \eta^2}{\operatorname{Tr}\Sigma^2 - \lambda_d^2}$. Then, for all $m \in \mathcal{M}$,

$$\|m\|_d = \sqrt{\eta^2 + (\operatorname{Tr}\Sigma^2 - \lambda_d^2)h^2} = \eta + \delta\,.$$

Let $\nu = (0, \ldots, \eta)$, it holds

$$\begin{aligned}
\sup_{\mathbb{P} \in \mathcal{H}_0} \mathbb{P}^{\otimes n}(\phi = 1) + \sup_{\mathbb{P} \in \mathcal{A}_\delta} \mathbb{P}^{\otimes n}(\phi = 0) &\geq \mathbb{P}^{\otimes n}_{\nu,\Sigma}(\phi = 1) + \mathbb{Q}^n_\Sigma(\phi = 0) \\
&\geq 1 - \frac{1}{2}\left\|\mathbb{P}^{\otimes n}_{\nu,\Sigma} - \mathbb{Q}^n_\Sigma\right\|_{\mathrm{TV}} \\
&\geq 1 - \frac{1}{2}\left(\int_{\mathbb{R}^{d\times n}} \left(\frac{d\mathbb{Q}^n_\Sigma}{d\mathbb{P}^{\otimes n}_{\nu,\Sigma}}\right)^2 d\mathbb{P}^{\otimes n}_{\nu,\Sigma} - 1\right)^{\frac{1}{2}}.
\end{aligned} \tag{IV.38}$$

see for instance Baraud (2002). For a tensor product of Gaussian distributions with fixed, equal covariance, the empirical mean is a sufficient statistic because the Radon-Nikodym derivative of a tensor product of Gaussian measures w.r.t. the Lebesgue measure can be written for $x_1, \ldots, x_n \in \mathbb{R}^d$ as

$$\frac{d\mathbb{P}^{\otimes n}_{m,\Sigma}}{d\lambda^{\otimes n}}(x_1, \ldots, x_n) = \phi_{m,\Sigma/n}(\bar{x}) F_\Sigma(x_1, \ldots, x_n)\,,$$

where $\bar{x}$ is the mean of the x_is, $\phi_{m,\Sigma/n}$ is the p.d.f. of a normal $\mathcal{N}(m, \Sigma/n)$ variable, and F_Σ is a function of $(x_1, \dots, x_n)$ which only depends on Σ. Therefore

$$\frac{d\mathbb{Q}_\Sigma^n}{d\mathbb{P}_{\nu,\Sigma}^{\otimes n}}(x_1, \dots, x_n) = \frac{d\mathbb{Q}_{\Sigma/n}^1}{d\mathbb{P}_{\nu,\Sigma/n}}(\bar{x}),$$

and thus

$$\int_{\mathbb{R}^{d\times n}} \left(\frac{d\mathbb{Q}_\Sigma^n}{d\mathbb{P}_{\nu,\Sigma}^{\otimes n}}\right)^2 d\mathbb{P}_{\nu,\Sigma}^{\otimes n} = \int_{\mathbb{R}^d} \left(\frac{d\mathbb{Q}_{\Sigma/n}^1}{d\mathbb{P}_{\nu,\Sigma/n}}\right)^2 d\mathbb{P}_{\nu,\Sigma/n}\,.$$

Thus the problem boils down to studying a single Gaussian vector of covariance Σ/n; for the following we will assume $n = 1$ and replace at the end Σ by Σ/n. Let us compute the densities F_ν and Q of these two distributions. For $x \in \mathbb{R}^d$:

$$F_\nu(x) = \left(\det \Sigma(2\pi)^d\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\lambda_d}(x_d - \eta)^2\right) \prod_{i=1}^{d-1} \exp\left(-\frac{x_i^2}{2\lambda_i}\right),$$

and

$$\begin{aligned} Q(x) &= \left(\det \Sigma(2\pi)^d\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\lambda_d}(x_d - \eta)^2\right) \times \frac{1}{2^{d-1}} \sum_{\substack{v_i\in\{-1,1\}\\ 1\le i\le d-1}} \prod_{i=1}^{d-1} \exp\left(-\frac{1}{2\lambda_i}(x_i - h\lambda_i v_i)^2\right) \\ &= \left(\det \Sigma(2\pi)^d\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\lambda_d}(x_d - \eta)^2 - \frac{h^2}{2}\sum_{i=1}^{d-1}\lambda_i\right) \times \prod_{i=1}^{d-1} \exp\left(-\frac{x_i^2}{2\lambda_i}\right)\cosh(hx_i)\,. \end{aligned}$$

Using that $\mathbb{E}\left[\cosh^2(aZ)\right] = \exp(a^2\sigma^2)\cosh(a^2\sigma^2)$ when $Z \sim \mathcal{N}(0, \sigma^2)$, we have that

$$\begin{aligned} \int_{\mathbb{R}^d} \frac{Q(x)^2}{F_\nu(x)}dx &= \left(\det \Sigma(2\pi)^d\right)^{-\frac{1}{2}} \exp\left(-h^2\sum_{i=1}^{d-1}\lambda_i\right) \int_{\mathbb{R}} \exp\left(-\frac{1}{2\lambda_d}(x_d - \eta)^2\right)dx_d \\ &\qquad \times \prod_{i=1}^{d-1} \int_{\mathbb{R}} \cosh^2(hx_i)\exp\left(-\frac{x_i^2}{2\lambda_i}\right)dx_i \\ &= \exp\left(-h^2\sum_{i=1}^{d-1}\lambda_i\right) \prod_{i=1}^{d-1} \exp(h^2\lambda_i)\cosh(h^2\lambda_i) \\ &= \prod_{i=1}^{d-1} \cosh(h^2\lambda_i). \end{aligned}$$

By Taylor expansion, we obtain the bound

$$h^2\lambda_i \le 1 \Rightarrow \cosh(h^2\lambda_i) \le 1 + \frac{e}{2}\lambda_i^2 h^4\,.$$

From this and the definition of h we deduce:

$$\log \prod_{i=1}^{d-1} \cosh(h^2\lambda_i) \le \frac{e}{2}(\operatorname{Tr}\Sigma^2 - \lambda_d^2)h^4 = \frac{e}{2(\operatorname{Tr}\Sigma^2 - \lambda_d^2)}\left((\eta+\delta)^2 - \eta^2\right)^2.$$

The end of the proof follows the same steps as the proof of Theorem 5.1 of Blanchard et al., 2018. That leads us to the final result: if

$$\delta \le \sqrt{\|\Sigma\|_{\mathrm{op}}\sqrt{d_* - 1}s + \eta^2} - \eta \quad \text{where} \quad s := \sqrt{\frac{2}{e}\log(1 + 4(1-\alpha)^2)}\,,$$

and

$$d_* \geq 1 + \frac{2}{e}\ln(5), \quad \text{i.e.} \quad d_* \geq 3\,,$$

then using (IV.38)

$$\sup_{\mathbb{P}\in\mathcal{H}_0} \mathbb{P}[\phi = 1] + \sup_{\mathbb{P}\in\mathcal{A}_\delta} \mathbb{P}[\phi = 0] > \alpha\,.$$

It follows

$$\begin{aligned}\delta^* &\geq \sqrt{\|\Sigma\|_{\mathrm{op}}\sqrt{d_*-1}s + \eta^2} - \eta \\ &\geq 2^{-\frac{3}{2}}\min\left(\sqrt{s\|\Sigma\|_{\mathrm{op}}}(d_*-1)^{\frac{1}{4}}, s\|\Sigma\|_{\mathrm{op}}\frac{(d_*-1)^{\frac{1}{2}}}{\eta}\right),\end{aligned}$$

and we obtain the inequality corresponding to the second part of the maximum in the right-hand side of (IV.17) by using that $s \geq (1-\alpha)$ and that $d_* - 1 \geq 2d_*/3$ because $d_* \geq 3$.

Let us prove now that $\delta^* \gtrsim \sqrt{\|\Sigma\|_{\mathrm{op}}}$. Let us consider the eigenvector e_1 associated to the maximum eigenvalue $\|\Sigma\|_{\mathrm{op}}$. Then $\mathbb{P}_{\eta e_1,\Sigma} \in \mathcal{H}_0$ and $\mathbb{P}_{(\eta+\delta)e_1,\Sigma} \in \mathcal{A}_\delta$. Let us denote $\lambda_1 = \|\Sigma\|_{\mathrm{op}}/n$, we have:

$$\begin{aligned}\int_{\mathbb{R}^d}\left(\frac{d\mathbb{P}^{\otimes n}_{(\eta+\delta)e_1,\Sigma}}{d\mathbb{P}^{\otimes n}_{\eta e_1,\Sigma}}\right)^2 d\mathbb{P}^{\otimes n}_{\eta e_1,\Sigma} &= \int_{\mathbb{R}^d}\left(\frac{d\mathbb{P}_{(\eta+\delta)e_1,\Sigma/n}}{d\mathbb{P}_{\eta e_1,\Sigma/n}}\right)^2 d\mathbb{P}_{\eta e_1,\Sigma/n} \\ &= \frac{e^{-\delta^2/\lambda_1}}{\sqrt{\lambda_1 2\pi}}\int_{\mathbb{R}}\exp\left(-\frac{(x-\eta)^2}{2\lambda_1}\right)\exp\left(\frac{2\delta(x-\eta)}{\lambda_1}\right)dx \\ &= \exp\left(\frac{3\delta^2}{\lambda_1}\right).\end{aligned}$$

If $\delta \leq \sqrt{\frac{\lambda_1}{3}\log(1+4(1-\alpha)^2)}$, then using (IV.38)

$$\sup_{\mathbb{P}\in\mathcal{H}_0} \mathbb{P}[\phi = 1] + \sup_{\mathbb{P}\in\mathcal{A}_\delta} \mathbb{P}[\phi = 0] > \alpha\,.$$

It follows that:

$$\delta^* \geq \sqrt{\|\Sigma/n\|_{\mathrm{op}}(1-\alpha)}.$$

IV.3.4 Proof of Theorem IV.8

This proof is similar to the proof of Theorem IV.7, so some details will be skipped. As in the one-sample case the upper bound is directly obtained using Theorem IV.3 and Proposition IV.6. We just additionally use the following upper bounds:

$$\begin{aligned}\frac{\sqrt{\mathrm{Tr}\,\Sigma^2}}{n} + \frac{\sqrt{\mathrm{Tr}\,S^2}}{m} &\leq \sqrt{2}\sqrt{\frac{\mathrm{Tr}\,\Sigma^2}{n^2} + \frac{\mathrm{Tr}\,S^2}{m^2}} \leq \sqrt{2}\sqrt{\mathrm{Tr}\left(\frac{\Sigma}{n} + \frac{S}{m}\right)^2}\,; \\ \frac{\|\Sigma\|_{\mathrm{op}}}{n} + \frac{\|S\|_{\mathrm{op}}}{m} &\leq 2\max\left(\frac{\|\Sigma\|_{\mathrm{op}}}{n}, \frac{\|S\|_{\mathrm{op}}}{m}\right) \leq 2\left\|\frac{\Sigma}{n} + \frac{S}{m}\right\|_{\mathrm{op}},\end{aligned}$$

where the last inequality holds because Σ, S are both positive semidefinite.

The lower bound in the two-sample case is a direct consequence of the one-sample case, by reduction to the case where one of the two sample means is known, say equal to zero. More specifically, let Σ and S be two symmetric positive semidefinite matrices, we consider again the distribution $\mathbb{Q}^n_\Sigma$ defined in (IV.37). Then

$$\int_{\mathbb{R}^{d\times(n+m)}}\left(\frac{d\mathbb{Q}^n_\Sigma \otimes \mathbb{P}^{\otimes m}_{0,S}}{d\mathbb{P}^{\otimes n}_{\nu,\Sigma}\otimes\mathbb{P}^{\otimes m}_{0,S}}\right)^2 d\mathbb{P}^{\otimes n}_{\nu,\Sigma}\otimes\mathbb{P}^{\otimes m}_{0,S} = \int_{\mathbb{R}^{d\times n}}\left(\frac{d\mathbb{Q}^n_\Sigma}{d\mathbb{P}^{\otimes n}_{\nu,\Sigma}}\right)^2 d\mathbb{P}^{\otimes n}_{\nu,\Sigma}\,.$$

Then using the previous results of the proof of Theorem IV.7 we obtain that

$$\delta^*(\alpha) \geq \left(n^{-1} \sqrt{\operatorname{Tr} \Sigma^2 - \lambda_d^2} s + \eta^2 \right)^{\frac{1}{2}} - \eta, \tag{IV.39}$$

with $s = \sqrt{\frac{2}{e} \log(1 + 4(1-\alpha)^2)}$. By the same token we obtain that

$$\delta^*(\alpha) \geq \left(m^{-1} \sqrt{\operatorname{Tr} S^2 - \ell_d^2} s + \eta^2 \right)^{\frac{1}{2}} - \eta, \tag{IV.40}$$

where ℓ_d is the smallest eigenvalue of the matrix S. Because $d_* \geq 3$, it holds

$$\begin{aligned} \max\left(n^{-2}\left(\operatorname{Tr}\Sigma^2 - \lambda_d^2\right), m^{-2}\left(\operatorname{Tr} S^2 - \ell_d^2\right)\right) &\geq \frac{2}{3} \max\left(n^{-2} \operatorname{Tr}\Sigma^2, m^{-2} \operatorname{Tr} S^2\right)^{\frac{1}{2}} \\ &\geq \frac{1}{6} \operatorname{Tr}\left(\frac{\Sigma}{n} + \frac{S}{m}\right)^2, \end{aligned}$$

and by combining (IV.39) and (IV.40), we obtain that

$$\delta^*(\alpha) \geq (2\sqrt{12})^{-1} \sigma \min\left(\sqrt{s} d_*^{\frac{1}{4}}, s \frac{\sigma d_*^{\frac{1}{2}}}{\eta} \right),$$

where $\sigma = \|\Sigma/n + S/m\|_{\mathrm{op}}$. We obtain (IV.19) using again that $s \geq 1 - \alpha$.

The last part of the lower bound is obtained as in the one-sample case using first the distributions $\mathbb{P}^{\otimes n}_{(\eta+\delta)e_1,\Sigma} \otimes \mathbb{P}^{\otimes m}_{0,S}$ and $\mathbb{P}^{\otimes n}_{\eta e_1,\Sigma} \otimes \mathbb{P}^{\otimes m}_{0,S}$ where e_1 is still the eigenvector associated to the biggest eigenvalue of Σ. We obtain that $\delta^*(\alpha) \gtrsim \|\Sigma/n\|_{\mathrm{op}}^{1/2}$. By the same token, we obtain that $\delta^*(\alpha) \gtrsim \|S/m\|_{\mathrm{op}}^{1/2}$ and conclude the proof using that $2\max(\|\Sigma/n\|_{\mathrm{op}}, \|S/m\|_{\mathrm{op}}) \geq \|\Sigma/n + S/m\|_{\mathrm{op}}$.

IV.3.5 Proof of Propositions IV.10 and IV.11

We want to obtain a concentration inequality for the estimator $\sqrt{\|\widehat{\Sigma}\|_{\mathrm{op}}}$. To this end, we will first study the following:

$$\widetilde{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)(X_i - \mu)^T, \tag{IV.41}$$

where μ is the true mean of the sample $\mathbb{X}$. Then we have:

$$\left\|\widehat{\Sigma} - \widetilde{\Sigma}\right\|_{\mathrm{op}} = \left\| -(\mu - \widehat{\mu})(\mu - \widehat{\mu})^T \right\|_{\mathrm{op}} = \|\mu - \widehat{\mu}\|^2. \tag{IV.42}$$

Gaussian setting. The concentration of $\|\widetilde{\Sigma}\|_{\mathrm{op}}^{1/2}$ is a consequence of the classical Lipschitz Gaussian concentration property (see e.g. Theorem 3.4 in Massart, 2003).

Theorem IV.20 (Gaussian Lipschitz concentration)**.** *Let $X = (x_1, \ldots, x_d)$ be a vector of i.i.d. standard Gaussian variables, and $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a L-Lipschitz function with respect to the Euclidean norm. Then for all $t \geq 0$:*

$$\mathbb{P}[f(X) - \mathbb{E}[f(X)] \geq t] \leq e^{-\frac{t^2}{2L^2}}. \tag{IV.43}$$

The following corollary is a direct consequence of that theorem (we provide a proof in Section IV.3.7), which will be used to control the term in (IV.42).

Corollary IV.21. *Let X a random Gaussian vector of distribution $\mathcal{N}(\mu, \Sigma)$. Then for all $u \geq 0$:*

$$\mathbb{P}\left[\|X\| \geq \sqrt{\|\mu\|^2 + \operatorname{Tr}\Sigma} + \sqrt{2\|\Sigma\|_{\mathrm{op}} u} \right] \leq e^{-u}. \tag{IV.44}$$

We will use the results of Koltchinskii and Lounici (2017) giving an upper bound of the expectation of the operator norm of the deviations of $\widetilde{\Sigma}$ from its expectation. The constants come from the improved version given by van Handel (2017).

Theorem IV.22 (van Handel, 2017)**.** *Let $\mathbb{X} = (X_i)_{1\le i\le n}$ a sample of independent Gaussian vectors of distribution $\mathcal{N}(0,\Sigma)$, then*

$$\mathbb{E}\Big[\big\|\widetilde{\Sigma}-\Sigma\big\|_{\mathrm{op}}\Big] \le \|\Sigma\|_{\mathrm{op}}\left((2+\sqrt{2})\sqrt{\frac{d_e}{n}}+2\frac{d_e}{n}\right), \tag{IV.45}$$

where $d_e = \operatorname{Tr}\Sigma/\|\Sigma\|_{\mathrm{op}}$ and $\widetilde{\Sigma}$ is defined in equation (IV.41).

We can now prove a concentration inequality for $\|\widetilde{\Sigma}\|_{\mathrm{op}}^{1/2}$.

Proposition IV.23. *Let $\mathbb{X} = (X_i)_{1\le i\le n}$ a sample of independent $\mathcal{N}(\mu,\Sigma)$ Gaussian vectors, then for $u \ge 0$, with probability at least $1-2e^{-u}$:*

$$\left|\big\|\widetilde{\Sigma}\big\|_{\mathrm{op}}^{\frac12} - \|\Sigma\|_{\mathrm{op}}^{\frac12}\right| \le 2\sqrt{\frac{2\operatorname{Tr}\Sigma}{n}}+\sqrt{\frac{2u\|\Sigma\|_{\mathrm{op}}}{n}}, \tag{IV.46}$$

where $\widetilde{\Sigma}$ is defined in (IV.41).

Remark IV.24. In (IV.46), the lower and upper bounds have been brought together, but the lower bound is in fact slightly better than the upper bound. This is due to the lower bound of the expectation where $\operatorname{Tr}\Sigma$ can be replaced by $\|\Sigma\|_{\mathrm{op}}$, see (IV.49) below.

Proof. We remark that

$$\begin{aligned}
\big\|\widetilde{\Sigma}\big\|_{\mathrm{op}}^{\frac12} &= \sup_{\|u\|_d=1}\sqrt{u^t\widetilde{\Sigma}u}\\
&= \sup_{\|u\|_d=1}\frac{1}{\sqrt{n}}\left(\sum_{i=1}^n\langle u, X_i-\mu\rangle^2\right)^{\frac12}\\
&= \sup_{\|u\|_d=1}\sup_{\|v\|_n=1}\frac{1}{\sqrt{n}}\sum_{i=1}^n\langle u, X_i-\mu\rangle v_i\\
&\overset{\text{dist}}{\sim} \sup_{\|u\|_d=1}\sup_{\|v\|_n=1}\frac{1}{\sqrt{n}}\sum_{i=1}^n\big\langle u, \Sigma^{\frac12}g_i\big\rangle v_i,
\end{aligned}$$

where $(g_i)_{i=1\dots n}$ are i.i.d. standard Gaussian vectors and $\|\cdot\|_p$ for $p\in\mathbb{N}$ is defined as the Euclidean norm in $\mathbb{R}^p$. Let u and v be unit vectors in $\mathbb{R}^d$ and $\mathbb{R}^n$ respectively and $f_{u,v}:\mathbb{R}^{d\times n}\to\mathbb{R}$:

$$f_{u,v}(y) := \frac{1}{\sqrt{n}}\sum_{i=1}^n\big\langle u, \Sigma^{\frac12}y_i\big\rangle v_i, \quad y\in\mathbb{R}^{d\times n}.$$

These functions are Lipschitz: indeed for all $z, y\in\mathbb{R}^{d\times n}$ we have:

$$\begin{aligned}
f_{u,v}(y)-f_{u,v}(z) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\big\langle u, \Sigma^{\frac12}(y_i-z_i)\big\rangle v_i \le \frac{1}{\sqrt{n}}\sum_{i=1}^n\|\Sigma\|_{\mathrm{op}}^{\frac12}\|y_i-z_i\|_d|v_i|\\
&\le \frac{\|\Sigma\|_{\mathrm{op}}^{\frac12}}{\sqrt{n}}\sqrt{\sum_{i=1}^n\|y_i-z_i\|_d^2} = \frac{\|\Sigma\|_{\mathrm{op}}^{\frac12}}{\sqrt{n}}\|y-z\|_{d\times n}.
\end{aligned} \tag{IV.47}$$

A supremum of Lipschitz functions is Lipschitz, thus we can use the Gaussian Lipschitz concentration (Theorem IV.20), and get for all $x \geq 0$:

$$\mathbb{P}\left[\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \mathbb{E}\left[\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}}\right] \geq \sqrt{\frac{2x\|\Sigma\|_{\mathrm{op}}}{n}}\right] \leq e^{-x}\,, \tag{IV.48}$$

with the same control for lower deviations.

It remains to upper bound $\left|\mathbb{E}\left[\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{1/2}\right] - \|\Sigma\|_{\mathrm{op}}^{1/2}\right|$. For one direction, using Jensen's and triangular inequalities and inequality (IV.29), we get:

$$\begin{aligned}
\mathbb{E}\left[\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}}\right] - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}} &\leq \sqrt{\|\Sigma\|_{\mathrm{op}} + \mathbb{E}\left[\left\|\widetilde{\Sigma} - \Sigma\right\|_{\mathrm{op}}\right]} - \sqrt{\|\Sigma\|_{\mathrm{op}}} \\
&\leq \min\left(\sqrt{\mathbb{E}\left[\left\|\widetilde{\Sigma} - \Sigma\right\|_{\mathrm{op}}\right]}, \frac{\mathbb{E}\left[\left\|\widetilde{\Sigma} - \Sigma\right\|_{\mathrm{op}}\right]}{2\sqrt{\|\Sigma\|_{\mathrm{op}}}}\right) \\
&\leq 2\sqrt{\frac{2\operatorname{Tr}\Sigma}{n}}\,.
\end{aligned}$$

For the last inequality, we have used Theorem IV.22 for the expectation and then the fact that

$$\min\Big((a\sqrt{x} + bx)^{1/2}, (a\sqrt{x} + bx)/2\Big) \leq \max(\sqrt{a+b}, (a+b)/2)\sqrt{x}$$

where $a = 2 + \sqrt{2}$, $b = 2$ and $x = d_e/n$. This is achieved by treating cases $x \leq 1$ and $x \geq 1$ separately.

For the other direction, a reformulation of (IV.48) is that there exists a random variable $g \sim \mathrm{Exp}(1)$ such that:

$$\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} \leq \mathbb{E}\left[\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}}\right] + \sqrt{\frac{2g\|\Sigma\|_{\mathrm{op}}}{n}}\,.$$

Taking the square then the expectation and then applying Jensen's inequality to the concave function $x \mapsto (a + b\sqrt{x})^2$ $(a, b \geq 0)$, we obtain:

$$\begin{aligned}
\|\Sigma\|_{\mathrm{op}} \leq \mathbb{E}\left[\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}\right] &\leq \mathbb{E}_g\left[\left(\mathbb{E}\left[\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}}\right] + \sqrt{\frac{2g\|\Sigma\|_{\mathrm{op}}}{n}}\right)^2\right] \\
&\leq \left(\mathbb{E}\left[\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}}\right] + \sqrt{\frac{2\|\Sigma\|_{\mathrm{op}}}{n}}\right)^2,
\end{aligned}$$

and thus

$$\mathbb{E}\left[\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}}\right] - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}} \geq -\sqrt{\frac{2\|\Sigma\|_{\mathrm{op}}}{n}} \geq -2\sqrt{\frac{2\operatorname{Tr}\Sigma}{n}}\,. \tag{IV.49}$$

□

Proof of Proposition IV.10. It holds

$$\left|\left\|\widehat{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\right| \leq \left|\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\right| + \left\|\widehat{\Sigma} - \widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}}.$$

Then, from (IV.42):

$$\|\widehat{\Sigma} - \widetilde{\Sigma}\|_{\mathrm{op}}^{\frac{1}{2}} \leq \|\mu - \widehat{\mu}\|\,.$$

According to Proposition IV.23 and Corollary IV.21, we obtain that for $u \geq 0$, with probability at least $1 - 3e^{-u}$:

$$\left|\left\|\widehat{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\right| \leq 2\sqrt{\frac{2\operatorname{Tr}\Sigma}{n}} + \sqrt{\frac{2u\|\Sigma\|_{\mathrm{op}}}{n}} + \sqrt{\frac{\operatorname{Tr}\Sigma}{n}} + \sqrt{\frac{2u\|\Sigma\|_{\mathrm{op}}}{n}}\,.$$

So, for $u \geq 0$, with probability at least $1 - 3e^{-u}$:

$$\left| \|\widehat{\Sigma}\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}} \right| \leq 3\sqrt{\frac{2\operatorname{Tr}\Sigma}{n}} + 2\sqrt{\frac{2u\|\Sigma\|_{\mathrm{op}}}{n}}\,.$$

Bounded setting. We first recall the following concentration result for bounded random vectors in the formulation of Bousquet (2002).

Theorem IV.25 (Talagrand-Bousquet inequality)**.** *Assume $(X_i)_{1\leq i\leq n}$ are i.i.d. with marginal distribution $\mathbb{P}$. Let $\mathcal{F}$ be a countable set of functions from $\mathcal{X}$ to $\mathbb{R}$ and assume that all functions f in $\mathcal{F}$ are $\mathbb{P}$-measurable, square-integrable, bounded by M and satisfy $\mathbb{E}[f] = 0$. Then we denote*

$$Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(X_i).$$

Let σ be a positive real number such that $\sigma^2 \geq \sup_{f\in\mathcal{F}} \operatorname{Var}[f(X_1)]$. Then for all $u \geq 0$, $\varepsilon > 0$ we have:

$$\mathbb{P}\left[Z \geq \mathbb{E}[Z](1+\varepsilon) + \sqrt{2un\sigma^2} + \frac{Mu}{3}(1+\varepsilon^{-1}) \right] \leq e^{-u}.$$

The following corollary is a direct consequence of Theorem IV.25. Some refinement of this result in the same vein (including two-sided deviation control in the uncentered case) have been presented in Section III.6.6 (Proposition III.9 and Corollary III.10).

Corollary IV.26. *Let X_i for $i = 1, \dots, n$ i.i.d. random vectors bounded by L with expectation μ, covariance Σ in a separable Hilbert space $\mathcal{H}$. Then for $u \geq 0$, with probability at least $1 - e^{-u}$:*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} X_i - \mu \right\| \leq 2\sqrt{\frac{\operatorname{Tr}\Sigma}{n}} + \sqrt{\frac{2\|\Sigma\|_{\mathrm{op}}u}{n}} + \frac{4Lu}{3n}\,.$$

Lemma IV.27. *Let X_i for $i = 1, \dots, n$ i.i.d. random vectors bounded by L with expectation μ, covariance Σ in a separable Hilbert space $\mathcal{H}$. Then*

$$\mathbb{E}\left[\left\| \widetilde{\Sigma} - \Sigma \right\|_{\mathrm{op}} \right] \leq \sqrt{\frac{\mathbb{E}[\|X_1 - \mu\|^4]}{n}}\,, \tag{IV.50}$$

where $\widetilde{\Sigma}$ is defined in (IV.41).

Remark IV.28. Using the boundedness of the variables we can upper bound this variance:

$$\mathbb{E}\left[\|X_1 - \mu\|^4\right] \leq 4L^2 \operatorname{Tr}\Sigma.$$

Proposition IV.29. *Let $(X_i)_{1\leq i\leq n}$ be i.i.d. random vectors in a separable Hilbert space $\mathcal{H}$, with norm bounded by L and covariance Σ, then for any for $u \geq 1$, with probability at least $1 - e^{-u}$:*

$$\left\| \widetilde{\Sigma} - \Sigma \right\|_{\mathrm{op}} \leq 2\sqrt{\frac{\mathbb{E}[\|X_1 - \mu\|^4]}{n}} + L\sqrt{\frac{2\|\Sigma\|_{\mathrm{op}}u}{n}} + \frac{8L^2u}{3n}\,, \tag{IV.51}$$

where $\widetilde{\Sigma}$ is defined in (IV.41).

Proof. We denote in this proof $Z_i := X_i - \mu$ for $1 \leq i \leq n$. Let us first remark that if B_1 is the unit ball of $\mathcal{H}$, then:

$$\left\| \widetilde{\Sigma} - \Sigma \right\|_{\mathrm{op}} = \sup_{u,v\in B_1} \frac{1}{n}\sum_{i=1}^{n} \langle v, \left(Z_i Z_i^T - \Sigma\right) u \rangle =: \sup_{u,v\in B_1} \frac{1}{n}\sum_{i=1}^{n} f_{u,v}(X_i)\,.$$

Since the variables X_i have norm bounded by L, it can be assumed equivalently that they take their values in $B_L = LB_1$, and it holds $\sup_{x\in B_L}\sup_{u,v\in B_1} f_{u,v}(x) \leq 8L^2$. Furthermore, since $(u,v)\mapsto f_{u,v}(x)$ is continuous, and the Hilbert space $\mathcal{H}$ is separable, the uncountable set B_1 can be replaced by a countable dense subset. Thus we can apply Theorem IV.25, and obtain that with probability at least $1-e^{-x}$:

$$\left\|\widetilde{\Sigma}-\Sigma\right\|_{\mathrm{op}} \leq 2\mathbb{E}\left[\|\widetilde{\Sigma}-\Sigma\|_{\mathrm{op}}\right] + L\sqrt{\frac{2\|\Sigma\|_{\mathrm{op}}x}{n}} + \frac{16L^2x}{3n}\,,$$

where we have used for the variance term:

$$\begin{aligned}\sup_{u,v\in B_1}\mathbb{E}\left[\left\langle v,\left(Z_iZ_i^T-\Sigma\right)u\right\rangle^2\right] &\leq \sup_{u,v\in B_1}\mathbb{E}\left[\langle v,Z_i\rangle^2\langle Z_i,u\rangle^2\right]\\ &\leq 4nL^2\|\Sigma\|_{\mathrm{op}}\,.\end{aligned}$$

We conclude using the upper bound of the expectation from Lemma IV.27. □

Proof of Proposition IV.11. As in the Gaussian case, we have:

$$\left|\left\|\widehat{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\right| \leq \left|\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\right| + \left\|\widehat{\Sigma}-\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}}.$$

From Lemma IV.15 and Proposition IV.29, we have with probability at least $1-e^{-u}$:

$$\left|\left\|\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\right| \leq 4L\sqrt{\frac{\operatorname{Tr}\Sigma}{n\|\Sigma\|_{\mathrm{op}}}} + \sqrt{\frac{16}{3}\frac{L^2u}{n}}\,,$$

where we have used that:

$$\sqrt{\frac{\operatorname{Var}[\|Z_1\|^2]}{n}} \leq \frac{2L\sqrt{\operatorname{Tr}\Sigma}}{\sqrt{n}}\,.$$

Using

$$\left\|\widehat{\Sigma}-\widetilde{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} \leq \|\mu-\widehat{\mu}\|\,,$$

and according to Corollary IV.26, we obtain that for $u\geq 0$, with probability at least $1-2e^{-u}$:

$$\begin{aligned}\left|\left\|\widehat{\Sigma}\right\|_{\mathrm{op}}^{\frac{1}{2}} - \|\Sigma\|_{\mathrm{op}}^{\frac{1}{2}}\right| &\leq \left(4L\sqrt{\frac{\operatorname{Tr}\Sigma}{n\|\Sigma\|_{\mathrm{op}}}} + \sqrt{\frac{16}{3}\frac{L^2u}{n}}\right)\\ &\qquad + \left(2\sqrt{\frac{\operatorname{Tr}\Sigma}{n}} + \sqrt{\frac{2\|\Sigma\|_{\mathrm{op}}u}{n}} + \frac{4Lu}{3n}\right)\\ &\leq 8L\sqrt{\frac{\operatorname{Tr}\Sigma}{n\|\Sigma\|_{\mathrm{op}}}} + 4L\left(\sqrt{\frac{2u}{n}} + \frac{u}{3n}\right),\end{aligned}$$

where we have used for the last inequality that $\|\Sigma\|_{\mathrm{op}} \leq 4L^2$. □

IV.3.6 Proof of Propositions IV.12 and IV.13

From a sample $\mathbb{X} = (X_i)_{1\leq i\leq n}$ of i.i.d. random vectors, we want to estimate $\operatorname{Tr}\Sigma^2$ where Σ is their common covariance matrix. The statistic $\widehat{T}$ defined in (IV.24) is an unbiased estimator of $\operatorname{Tr}\Sigma^2$. This statistic is also invariant by translation.

If we denote $\mathfrak{S}_n$ the set of permutations of $\{1,\dots,n\}$, $\widehat{T}$ can be rewritten as:

$$\widehat{T} = \frac{1}{n!}\sum_{\sigma\in\mathfrak{S}_n}\frac{1}{\lfloor n/4\rfloor}\sum_{i=1}^{\lfloor n/4\rfloor}\frac{1}{4}\left\langle X_{\sigma(4i)}-X_{\sigma(4i-2)}, X_{\sigma(4i-1)}-X_{\sigma(4i-3)}\right\rangle^2; \tag{IV.52}$$

namely by symmetry, all the 4-tuples appear the same number of times in the right-hand side, so we just need to divide by the number of terms to obtain the identity (IV.52). We will use this decomposition to obtain a concentration of the statistic $\widehat{T}$ for the Gaussian case and the bounded case, since the inner sum for each fixed permutation is a sum of $\lfloor n/4 \rfloor$ i.i.d. terms.

Gaussian setting. Because the statistic is invariant by translation we can assume without loss of generality that $\mu = 0$. To obtain a deviation inequality for $\widehat{T}^{1/2}$, we will first find a concentration inequality for $\widehat{T}$ and then use Lemma IV.15. We obtain concentration via control of moments of $\widehat{T}$, so we first need some upper bounds on Gaussian moments. The following lemma is proved in Section IV.3.7.

Lemma IV.30. *Let $Z_i := \langle X_i^1 - X_i^3, X_i^2 - X_i^4 \rangle^2/4$, where X_i^j for $i = 1, \dots, m$ and $1 \le j \le 4$ are i.i.d. Gaussian random vectors $\mathcal{N}(0, \Sigma)$. Then for all $q \in \mathbb{N}$:*

$$\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m} Z_i - \operatorname{Tr}\Sigma^2\right)^{2q}\right] \le \left(4\sqrt{2}\phi q^2 \frac{\operatorname{Tr}\Sigma^2}{\sqrt{m}}\right)^{2q}, \tag{IV.53}$$

where $\phi = (1+\sqrt{5})/2$ is the golden ratio.

We deduce from this lemma a concentration inequality for $\widehat{T}$.

Proposition IV.31. *Let $(X_i)_{1 \le i \le n}$, $n \ge 4$ be i.i.d. random vectors with distribution $\mathcal{N}(\mu, \Sigma)$. Then for all $u \ge 0$:*

$$\mathbb{P}\left[\left|\widehat{T} - \operatorname{Tr}\Sigma^2\right| \ge 30\frac{u^2 \operatorname{Tr}\Sigma^2}{\sqrt{n}}\right] \le e^4 e^{-u}, \tag{IV.54}$$

where $\widehat{T}$ is defined in (IV.24).

Proof. Using Lemma IV.30, (IV.52) and the convexity of the function $x \mapsto x^{2q}$, we can upper bound the moments of $\widehat{T}$:

$$\mathbb{E}\left[\left(\widehat{T} - \operatorname{Tr}\Sigma^2\right)^{2q}\right] \le \left(4\sqrt{2}\phi q^2 \frac{\operatorname{Tr}\Sigma^2}{\sqrt{\lfloor n/4 \rfloor}}\right)^{2q}. \tag{IV.55}$$

Let $t \ge 0$ and $q \in \mathbb{N}$, then by Markov's inequality

$$\mathbb{P}\left[\left|\widehat{T} - \operatorname{Tr}\Sigma^2\right| \ge t\right] \le t^{-2q}\mathbb{E}\left[\left(\widehat{T} - \operatorname{Tr}\Sigma^2\right)^{2q}\right]. \tag{IV.56}$$

Let us choose q as:

$$q = \left\lfloor \frac{e^{-1}}{2\sqrt{\phi}2^{\frac{1}{4}}} t^{\frac{1}{2}} \left(\frac{\operatorname{Tr}\Sigma^2}{\sqrt{\lfloor n/4 \rfloor}}\right)^{-\frac{1}{2}} \right\rfloor,$$

so that (IV.55), (IV.56) entail

$$\mathbb{P}\left[\left|\widehat{T} - \operatorname{Tr}\Sigma^2\right| \ge t\right] \le e^{-4q}.$$

Let us now take

$$t = \frac{e^2\sqrt{2}\phi}{4} u^2 \frac{\operatorname{Tr}\Sigma^2}{\sqrt{\lfloor n/4 \rfloor}} \le 30\frac{u^2 \operatorname{Tr}\Sigma^2}{\sqrt{n}},$$

where we have used that $\lfloor n/4 \rfloor \ge n/7$ for $n \ge 4$; we obtain that for all $u \ge 0$:

$$\mathbb{P}\left[\left|\widehat{T} - \operatorname{Tr}\Sigma^2\right| \ge 30\frac{u^2 \operatorname{Tr}\Sigma^2}{\sqrt{n}}\right] \le e^4 e^{-u}.$$

□

Proposition IV.12 directly follows from Proposition IV.31 and Lemma IV.15.

Bounded setting. As in the Gaussian case, we first obtain a concentration inequality for $\widehat{T}$ and then using Lemma IV.15, we obtain one for $\widehat{T}^{1/2}$. We will need the following classical Bernstein's inequality (see for instance Vershynin, 2018, Exercise 2.8.5 for the version below) which gives an upper bound on the Laplace transform of the sum of bounded random variables.

Lemma IV.32 (Bernstein's inequality)**.** *Let $(X_i)_{1\le i\le m}$ be i.i.d. real centered random variables bounded by B such that*

$$\mathbb{E}\big[X_1^2\big] \le \sigma^2 .$$

Then for all $t < 3/B$:

$$\log\Big(\mathbb{E}\big[e^{t\sum X_i}\big]\Big) \le \frac{1}{2}\frac{m\sigma^2 t^2}{1 - Bt/3} .$$

Via Bernstein's inequality we obtain the following result.

Proposition IV.33. *Let $(X_i)_{1\le i\le n}$, $n \ge 4$ be i.i.d. Hilbert-valued random variables with norm bounded by L and covariance Σ, and $\widehat{T}$ defined by (IV.24). Then for all $t \ge 0$:*

$$\mathbb{P}\left[\left|\widehat{T} - \operatorname{Tr}\Sigma^2\right| \ge 8L^2\sqrt{\frac{\operatorname{Tr}\Sigma^2 t}{n}} + \frac{10L^4 t}{n}\right] \le 2e^{-t} . \tag{IV.57}$$

where $\widehat{T}$ is defined in (IV.24).

Proof. Let X, X', Y, Y' be i.i.d. Hilbert-valued random vectors of expectation μ, covariance Σ and with norm bounded by L, and $Z := \langle X - Y, X' - Y'\rangle^2/4$. Then it holds $0 \le Z \le 4L^4$, $\mathbb{E}[Z] = \operatorname{Tr}\Sigma^2$ and

$$\begin{aligned} |Z - \mathbb{E}[Z]| &\le 4L^4; \\ \operatorname{Var}[Z] &\le 4L^4\mathbb{E}[Z] = 4L^4 \operatorname{Tr}\Sigma^2. \end{aligned}$$

Now using the convexity of the exponential function, (IV.52) and then Lemma IV.32, we can upper bound the Laplace transform of $\widehat{T}$ as follows:

$$\log\Big(\mathbb{E}\big[e^{t\widehat{T}}\big]\Big) \le \frac{1}{2\lfloor n/4\rfloor}\frac{4L^4 \operatorname{Tr}\Sigma^2 t^2}{1 - 4L^4 t/(3\lfloor n/4\rfloor)} ,$$

for all t such that the right-hand sise is well defined, i.e. the denominator is strictly positive. Now using Lemma IV.17, and $\lfloor n/4\rfloor \ge n/7$ for $n \ge 4$, for all $t \ge 0$ it holds

$$\mathbb{P}\left[\left|\widehat{T} - \operatorname{Tr}\Sigma^2\right| \ge 8L^2\sqrt{\frac{\operatorname{Tr}\Sigma^2 t}{n}} + \frac{10L^4 t}{n}\right] \le 2e^{-t} . \tag{IV.58}$$

□

Proof of Proposition IV.13. Assuming the event entering into (IV.58) holds, we will use the inequalities of Lemma IV.15:

$$\begin{aligned} \sqrt{\widehat{T}} - \sqrt{\operatorname{Tr}\Sigma^2} &\le \sqrt{\operatorname{Tr}\Sigma^2 + 8L^2\sqrt{\frac{\operatorname{Tr}\Sigma^2 t}{n}}} - \sqrt{\operatorname{Tr}\Sigma^2} + \sqrt{\frac{10L^4 t}{n}} \\ &\le 4L^2\sqrt{\frac{t}{n}} + L^2\sqrt{\frac{10t}{n}} \le 8L^2\sqrt{\frac{t}{n}} . \end{aligned}$$

For the other side, we proceed analogously:

$$\sqrt{\widehat{T}} - \sqrt{\operatorname{Tr}\Sigma^2} \geq \sqrt{\left(\operatorname{Tr}\Sigma^2 - 8L^2\sqrt{\frac{\operatorname{Tr}\Sigma^2 t}{n}}\right)_+} - \sqrt{\operatorname{Tr}\Sigma^2} - \sqrt{\frac{10L^4 t}{n}}$$
$$\geq -8L^2\sqrt{\frac{t}{n}} - L^2\sqrt{\frac{10t}{n}} \geq -12L^2\sqrt{\frac{t}{n}}\,.$$

□

IV.3.7 Additional proofs

Proof of Lemma IV.15. This Lemma completes the Lemma 6.1.3 of Blanchard et al. (2018). This is its complete proof.

Let a in $\mathbb{R}_+$, it is well known that for $b \geq -a^2$:

$$a - \sqrt{|b|} \leq \sqrt{a^2 + b} \leq a + \sqrt{|b|}\,.$$

On the other hand, suppose that $b \geq 0$, the Taylor expansion of the function $b \mapsto \sqrt{a^2+b} - a$ gives that there exists $c \in (0, b)$ such that:

$$\sqrt{a^2+b} - a = \frac{b}{2\sqrt{a^2+c}} \leq \frac{b}{2a}\,.$$

Suppose now that $0 \geq b \geq -a^2$, then

$$\sqrt{a^2+b} \geq a + \frac{b}{a} \Leftrightarrow b \geq 2b + \frac{b^2}{a^2} \Leftrightarrow b \geq -a^2\,.$$

The equation (IV.29) is still true when $b < -a^2$ because then:

$$-a \geq -\sqrt{|b|} \geq -\frac{|b|}{a}\,.$$

□

Proof of Proposition IV.18. Let g be a standard Gaussian random vector in $\mathbb{R}^d$, and $U^T D U$ be the singular value decomposition of the matrix $S^{1/2}\Sigma S^{1/2}$ where $D = \operatorname{diag}(\lambda, \ldots, \lambda_d)$. Then we have the following equalities in distribution

$$Y^T \Sigma Y \overset{\text{dist}}{\sim} g^T S^{\frac{1}{2}} \Sigma S^{\frac{1}{2}} g \overset{\text{dist}}{\sim} g^T U^T D U g \overset{\text{dist}}{\sim} g^T D g\,.$$

The last equality is a consequence of the invariance by rotation of Gaussian vectors. Then for $t < 1/\sqrt{\|\Sigma\|_{\text{op}}\|S\|_{\text{op}}}$:

$$\mathbb{E}\left[e^{t\langle X,Y\rangle}\right] = \mathbb{E}\left[e^{\frac{t^2\|\Sigma^{\frac{1}{2}}Y\|^2}{2}}\right] = \mathbb{E}\left[e^{\frac{t^2 g^T D g}{2}}\right] = \mathbb{E}\left[\exp\left(\frac{t^2}{2}\sum_{i=1}^d \lambda_i g_i^2\right)\right].$$

Using the independence of the coordinates and that $-\log(1-x) = \log\left(1 + \frac{x}{1-x}\right) \leq \frac{x}{1-x} \leq \frac{x}{1-\sqrt{x}}$ for $x < 1$, we obtain:

$$\log\left(\mathbb{E}\left[e^{t\langle X,Y\rangle}\right]\right) = \sum_{i=1}^d -\frac{1}{2}\log(1-t^2\lambda_i) \leq \sum_{i=1}^d \frac{1}{2}\frac{t^2\lambda_i}{1-t^2\lambda_i} \leq \frac{1}{2}\frac{t^2\operatorname{Tr}(S^{\frac{1}{2}}\Sigma S^{\frac{1}{2}})}{1 - t\|S^{\frac{1}{2}}\Sigma S^{\frac{1}{2}}\|_{\text{op}}^{\frac{1}{2}}}\,.$$

We conclude using that $\operatorname{Tr}(S^{\frac{1}{2}}\Sigma S^{\frac{1}{2}}) = \operatorname{Tr}(\Sigma S)$ and that $\|S^{\frac{1}{2}}\Sigma S^{\frac{1}{2}}\|_{\text{op}} \leq \|S\|_{\text{op}}\|\Sigma\|_{\text{op}}$. □

Proof of Corollary IV.21. We use the representation $X \overset{\text{dist}}{\sim} (\Sigma^{\frac{1}{2}} g + \mu)$, where g is a standard Gaussian random variable. We then have

$$\|X\| \overset{\text{dist}}{\sim} \|\Sigma^{\frac{1}{2}} g + \mu\| = f(g)\,,$$

where for $y \in \mathbb{R}^d$:

$$f(y) = \left\|\Sigma^{\frac{1}{2}} y + \mu\right\|.$$

This function f is Lipschitz with constant $\|\Sigma^{\frac{1}{2}}\|_{\text{op}}$. We conclude using Theorem IV.20 and Jensen's inequality:

$$\mathbb{E}\big[\|X\|\big] \leq \sqrt{\|\mu\|^2 + \operatorname{Tr}\Sigma}\,.$$

□

Proof of Corollary IV.26. We apply Theorem IV.25, with $\varepsilon = 1$ and the set of functions $\mathcal{F} = \{f_u\}_{\|u\|_{\mathcal{H}}=1}$ where $f_u : x \in \mathcal{H} \mapsto \langle x, u\rangle_{\mathcal{H}}$ for $u \in \mathcal{H}$. We can find a countable subset of the unit sphere because $\mathcal{H}$ is separable. Then

$$Z = \sup_{\|u\|_{\mathcal{H}}=1} \sum_{i=1}^{n} \langle X_i - \mu, u\rangle_{\mathcal{H}} = \left\| \sum_{i=1}^{n} X_i - \mu \right\|_{\mathcal{H}}.$$

We conclude using that for all u in the unit sphere of $\mathcal{H}$, $\operatorname{Var}[\langle X_i - \mu, u\rangle_{\mathcal{H}}] \leq \|\Sigma\|_{\text{op}}$ and $|\langle X_i - \mu, u\rangle_{\mathcal{H}}| \leq 2L$ a.s. We use Jensen's inequality to upper bound the expectation: $\mathbb{E}[Z] \leq (n \operatorname{Tr}\Sigma)^{\frac{1}{2}}$. □

Proof of Lemma IV.27. We upper bound the operator norm with the Frobenius norm. We denote in this proof $Z_i := X_i - \mu$. It holds:

$$\begin{aligned}
\mathbb{E}\Big[\big\|\Sigma - \widetilde{\Sigma}\big\|_{\text{op}}\Big] &\leq \mathbb{E}\left[\sqrt{\operatorname{Tr}\left(\Sigma - \widetilde{\Sigma}\right)^2}\right] \\
&\leq \left(\mathbb{E}\left[\operatorname{Tr}\left(\frac{1}{n^2}\Big(\sum_i (Z_i Z_i^T)^2 + \sum_{i\neq j} Z_i Z_i^T Z_j Z_j^T\Big) - \widetilde{\Sigma}\Sigma - \Sigma\widetilde{\Sigma} + \Sigma^2\right)\right]\right)^{\frac{1}{2}} \\
&= \left(\frac{\mathbb{E}\big[\|Z\|^4\big]}{n} - \frac{\operatorname{Tr}\Sigma^2}{n}\right)^{\frac{1}{2}} \leq \sqrt{\frac{\mathbb{E}[\|Z\|^4]}{n}} \leq \frac{2L\sqrt{\operatorname{Tr}\Sigma}}{\sqrt{n}}\,.
\end{aligned}$$

□

Proof Lemma IV.30. First let us remark that if X and X' are independent $\mathcal{N}(0,\Sigma)$ Gaussian vectors, then

$$\langle X, X'\rangle \overset{\text{dist}}{\sim} \sum_{i=1}^{d} \lambda_i g_i g_i'\,,$$

where g_i and g_i' are independent standard Gaussian random variables and the λ_is are the eigenvalues of Σ. Then for $q \in \mathbb{N}$, recalling $\mathbb{E}\big[g_i^{2q}\big] = (2q!)/(2^q q!)$,

$$\begin{aligned}
\mathbb{E}\Big[\langle X, X'\rangle^{2q}\Big] &= \sum_{p_1+\ldots+p_d=q} \binom{2q}{2p_1,\ldots,2p_d} \prod_{i=1}^{d} (\lambda_i)^{2p_i} \left(\frac{(2p_i)!}{2^{p_i} p_i!}\right)^2 \\
&\leq (2q)! \sum_{p_1+\ldots+p_d=q} \prod_{i=1}^{d} (\lambda_i^2)^{p_i} \\
&\leq (2q)! (\operatorname{Tr}\Sigma^2)^q,
\end{aligned}$$

where we have used $(2p)! \leq 2^{2p} p!^2$. Using this bound, we upper bound the moments of the Z_i-s:

$$\left|\mathbb{E}[Z_i^q]\right| = 2^{-2q}\mathbb{E}\Big[\left\langle X_i^1 - X_i^3, X_i^2 - X_i^4\right\rangle^{2q}\Big] \leq (2q)!(\operatorname{Tr}\Sigma^2)^q\,.$$

We now upper bound the moments of $Z_i - \operatorname{Tr}\Sigma$. Let Z_i' be an independent copy of Z_i, then since $\mathbb{E}[Z_i'] = \operatorname{Tr}\Sigma^2$, by Jensen's inequality

$$\mathbb{E}\Big[\big(Z_i - \operatorname{Tr}\Sigma^2\big)^{2q}\Big] \leq \mathbb{E}\Big[\big(Z_i - Z_i'\big)^{2q}\Big] \leq 2^{2q}\mathbb{E}\Big[Z_i^{2q}\Big] \leq (4q)!(2\operatorname{Tr}\Sigma^2)^{2q}\,.$$

For the odd moments we use that the function $(\cdot)^{2q+1}$ is increasing:

$$-(\operatorname{Tr}\Sigma^2)^{2q+1} \leq \mathbb{E}\Big[\big(Z_i - \operatorname{Tr}\Sigma^2\big)^{2q+1}\Big] \leq \mathbb{E}\Big[Z_i^{2q+1}\Big] \leq (4q+2)!(\operatorname{Tr}\Sigma^2)^{2q+1}\,,$$

so for all $q \geq 0$:

$$\left|\mathbb{E}\Big[\big(Z_i - \operatorname{Tr}\Sigma^2\big)^{q}\Big]\right| \leq (2q)!(2\operatorname{Tr}\Sigma^2)^q\,. \tag{IV.59}$$

It remains to upper bound the moments of the sum:

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^m Z_i^2 - \operatorname{Tr}\Sigma^2\right)^{2q}\right] &= \frac{1}{m^{2q}} \sum_{\substack{p_1+\ldots+p_m=2q\\ p_i\neq 1}} \binom{2q}{p_1,\ldots,p_m} \prod_{i=1}^m \mathbb{E}\Big[\big(Z_i - \operatorname{Tr}\Sigma^2\big)^{p_i}\Big] \\
&\leq \frac{1}{m^{2q}} \sum_{\substack{p_1+\ldots+p_m=2q\\ p_i\neq 1}} \frac{(2q)!}{p_1!\ldots p_m!} \prod_{i=1}^m (2p_i)!(2\operatorname{Tr}\Sigma^2)^{p_i} \\
&\leq (2q)!\left(\frac{2\operatorname{Tr}\Sigma^2}{m}\right)^{2q} (2q)^{2q} \sum_{\substack{p_1+\ldots+p_m=2q\\ p_i\neq 1}} 1.
\end{aligned}$$

Let us count the number of terms in this last sum. Consider first that we have k non-null terms $(p_{i_1},\ldots,p_{i_k})$. Their sum is equal to $2q$ but because these terms are strictly greater than 1, we also have that $(p_{i_1}-2) + \ldots + (p_{i_k}-2) = 2q-2k$, where all terms of this sum are nonnegative. The number of k-partitions of $2q-2k$ is $\binom{(2q-2k)+(k-1)}{k-1} = \binom{2q-k-1}{k-1}$ and then the number of terms in the sum is equal to:

$$\begin{aligned}
\sum_{k=0}^m \binom{m}{k}\binom{2q-k-1}{k-1} &= \sum_{k=0}^{m\wedge q} \binom{m}{k}\binom{2q-k-1}{k-1} \\
&\leq m^q \sum_{k=0}^q \binom{2q-k-1}{k-1} = m^q F(2q-1) \leq m^q \phi^{2q}\,,
\end{aligned}$$

where $F(\cdot)$ is the Fibonacci sequence and $\phi = (1+\sqrt{5})/2$ is the golden ratio. So using that $(2q)! \leq (2q)^q q^q$ we obtain that

$$\mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^m Z_i - \operatorname{Tr}\Sigma^2\right)^{2q}\right] \leq (2\phi^2)^q \left(\frac{\operatorname{Tr}\Sigma^2}{\sqrt{m}}\right)^{2q} (2q)^{4q}. \tag{IV.60}$$

□

V Estimation of multiple mean vectors in high dimension with full heterogeneity

In Section III, to estimate numerous multi-dimensional means of various probability distributions, we assume that the distributions are homogeneous. Formally we have assumed that the risks of each empirical mean are upper bounded by a known quantity (Eq.(III.6)) and that their effective dimension are of same order (Eq.(III.15)). However, this assumption prevents us from considering cases where some bags have high size or small variance relative to the others and could really improve their estimation.

For this purpose, in this section, the estimators are formed through convex combinations of the empirical means with weights depending of their covariance structure and their sample size. We introduce two strategies to find appropriate data-dependent convex combination weights: a first one employing a testing procedure to identify neighbouring means with low variance, which results in a closed-form plug-in formula for the weights, and a second one determining weights via minimization of an upper confidence bound on the quadratic risk. We evaluate the improvement in quadratic risk offered by our methods compared to the empirical means. Our analysis focuses on a dimensional asymptotics perspective, showing that our methods asymptotically approach an oracle (minimax) improvement as the effective dimension of the data increases. We demonstrate the efficacy of our methods in estimating multiple kernel mean embeddings through experiments on both simulated and real-world datasets.

Contents

V.1 Introduction

We study the problem of jointly estimating multiple vector means $\mu_1, \dots, \mu_B$ of distinct probability distributions $\mathbb{P}_1, \dots, \mathbb{P}_B$ over $\mathbb{R}^d$ (an extension to Hilbert spaces is also discussed). The estimation of the means is based on a family of independent sample sets, $X_\bullet^{(1)}, \dots, X_\bullet^{(B)}$, where each $X_\bullet^{(k)}$ with $k \in [\![B]\!] := \{1, \dots, B\}$ comprises of N_k samples drawn i.i.d. from $\mathbb{P}_k$. Formally, the joint model is

$$\begin{cases} X_\bullet^{(k)} := (X_i^{(k)})_{1 \le i \le N_k} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_k, \ k \in [\![B]\!]; \\ (X_\bullet^{(1)}, \dots, X_\bullet^{(B)}) \text{ independent.} \end{cases} \tag{V.1}$$

The distributions are assumed to be at least square-integrable. We refer to a set of samples $X_\bullet^{(k)}$ as *bag* and to $\mathbb{P}_k$ as *task*, in line with the domain of multi-task learning. Our aim is to define estimators $\widehat{\mu}_k$ and analyse the risk given by the expected squared distance to the true means μ_k.

Evident candidates are empirical means taken separately for each bag, which we call *naive* estimators. The question we want to tackle is whether it is possible to improve over these individual naive estimators by exploiting similarities between tasks. We propose and study particular estimators $\widehat{\mu}_k$ formed by a convex combination of naive estimators of "related" tasks. We insist that absolutely no information about the underlying similarity or task structure is assumed to be known *a priori*. Roughly speaking, we measure relatedness between tasks by estimating the distance between their means.

The goal is to analyse the *relative* risk of the proposed estimators, i.e., the ratio of their risk to that of the corresponding naive estimator. The following questions will guide our estimator construction and analysis:

(a) what would be the ideal "oracle" convex combination estimator, if some additional *a priori* information about task relatedness were known?

(b) can an empirical estimator approach the oracle relative risk from the data only, in a suitable asymptotical sense?

(c) is the oracle relative risk minimax optimal in a suitable asymptotical sense?

Because we focus on the relative risk, the usual asymptotics of the sample size going to infinity is not the most relevant one (though we will assume that the sample sizes are "large enough"). Rather, we will focus on *high-dimensional asymptotics* where the dimension grows large. More precisely, we mean a notion of *effective* dimension rather than ambient space dimension: the effective dimension of a task will be defined from spectral quantities related to its covariance matrix, as is common in high-dimensional statistics.

Motivations for this work. The framework under examination is motivated by scenarios involving large volumes of high-dimensional data. These scenarios typically involve the categorization of independent samples into homogeneous units that may exhibit differences but also varying degrees of similarity. Examples include medical or educational records sourced from different institutions, or purchase histories organised by individual clients on an internet platform. This framework also intersects with the concepts of federated and personalised machine learning (McAuley, 2022; Tan et al., 2022). An application of particular interest within this framework is that of kernel mean embeddings of distributions (Muandet et al., 2017). This involves estimating means of distributions after a formal mapping of the data into a Hilbert space. Notably, in this context, one anticipates that the effective dimensionality of the mapped data will be high.

Relation to previous work. The problem of estimating multiple means has a long and rich history in statistics, starting in particular with the seminal work of Stein on the eponymous paradox and the James-Stein estimator (James and Stein, 1961), continued with the empirical Bayes point of view on the latter (Efron and Morris, 1972), up to modern considerations on the topic (Brown and Greenshtein, 2009; Jiang and Zhang, 2009). The topic of "multitask learning" also provides a more recent angle on the problem (Duan and Wang, 2023; Feldman et al., 2014). We defer a detailed discussion to Section V.7, but stress that most previous works analysed the *compound* (or cumulated) risk over all tasks and its behaviour in the asymptotics $B \to \infty$, in a one- or fixed-dimensional setting. By contrast, we will be interested in analyzing the individual risk separately for each task, and in "high dimensional" asymptotics.

We start with a description of the considered setting in Section V.2. Sections V.3 and V.4 introduce two approaches to form convex combination estimators of the means, provide bounds on their relative risks, and a comparison of the two. A minimax analysis for suitable distribution classes is conducted in Section V.5. Finally, experiments on artificial and true data are presented in Section V.6. All proofs are provided in Section V.9, wherein Section V.9.1 contains a list of the used notation for the reader's convenience.

V.2 Setting and notation

V.2.1 Loss and risk

We consider the squared norm loss and expected risk

$$L_k(\widehat{\mu}_k) := \|\widehat{\mu}_k - \mu_k\|^2 ; \qquad R_k(\widehat{\mu}_k) := \mathbb{E}[L_k(\widehat{\mu}_k)]. \tag{V.2}$$

of an estimator $\widehat{\mu}_k$ for μ_k. The empirical mean $\widehat{\mu}_k^{\mathtt{NE}} := \frac{1}{N_i}\sum_{i=1}^{N_i} X_k^{(i)}$, called the *naive estimator*, serves as a reference. Due to the unbiasedness of the naive estimator, its variance is equal to its risk. More specifically, let the *naive risk* be denoted by

$$s_k^2 := R_k(\widehat{\mu}_k^{\mathtt{NE}}) = \frac{\operatorname{Tr}\Sigma_k}{N_k}, \tag{V.3}$$

where Σ_k is the covariance of task k. Then any estimator $\widehat{\mu}_k$ is analysed in terms of its *relative risk* to the naive — lower is better :

$$\frac{R_k(\widehat{\mu}_k)}{s_k^2}. \tag{V.4}$$

In contrast to the compound decision setting, our goal is to analyse the relative risk for each task separately. For this reason, the focus is on a specific task, say $k = 1$ and $R_1(\widehat{\mu}_1)/s_1^2$ without loss of generality. In Section V.5.2 the relative risk averaged over tasks $\frac{1}{B}\sum_{k=1}^{B} R_k(\widehat{\mu}_k)/s_k^2$ is considered.

V.2.2 High-dimensional asymptotics

Observe from (V.3) that the naive risk s_1^2 decreases at the parametric rate $\mathcal{O}(N_1^{-1})$. We expect the risk of a competing estimator $\widehat{\mu}_1$ to follow the same trend. As a consequence, the role of the sample size will

cancel out in the relative risk. In order to state meaningful results, it is necessary to obtain sharp estimates of the other factors in the rate.

To this end, we shift the perspective from a standard asymptotic view point, $N_1 \to \infty$, to high-dimensional asymptotics, emphasizing the behaviour of the risks as the dimensionality grows. There are different possible definitions of *effective dimensionality* of a distribution, generally linked to the spectral decay of the covariance matrix and ratios of its Schatten norms. The following ones will be relevant to our analysis:

$$d_k^\bullet := \frac{(\operatorname{Tr}\Sigma_k)^2}{\operatorname{Tr}\Sigma_k^2}, \qquad d_k^{\mathrm{e}} := \frac{\operatorname{Tr}\Sigma_k}{\|\Sigma_k\|_\infty}. \tag{V.5}$$

Observe that in the isotropic setting $\Sigma_k \propto I_d$, the effective dimensions $d_k^\bullet$ and d_k^{e} coincide with the ambient dimension d, as one would expect. In all cases it holds $1 \leq \sqrt{d_k^\bullet} \leq d_k^{\mathrm{e}} \leq d_k^\bullet \leq d$. In random matrix literature, d^{e} is sometimes called intrinsic dimension (Hsu et al., 2012; Tropp et al., 2015) or effective rank (Koltchinskii and Lounici, 2016), and $(d^{\mathrm{e}})^2/d^\bullet$ is known as the numerical or stable rank of Σ (Rudelson and Vershynin, 2007; Tropp et al., 2015). Most notably, we uncover a "blessing of dimensionality" phenomenon: in a nutshell, we will show that the relative risks of our estimators asymptotically approach a suitable notion of oracle relative risk as the (effective) dimensionality increases.

V.2.3 Distributional assumptions

For our theoretical analysis, we consider the following different possible distributional assumptions:

Assumption V.1 (**GS**, Gaussian setting)**.** *For all $k \in [\![B]\!]$, the distribution $\mathbb{P}_k$ is $\mathcal{N}(\mu_k, \Sigma_k)$.*

Assumption V.2 (**BS**, Bounded setting)**.** *For all $k \in [\![B]\!]$, $\mathbb{P}_k$ has support in the ball of radius M centred at 0.*

The (**BS**) setting is of particular interest for the application to kernel mean embeddings, for which the assumption of a bounded kernel is very common. All results for (**BS**) are presented in $\mathbb{R}^d$ but can be extended to a separable Hilbert space (up to adequate adaptation of notation).

Section V.9.4 covers another distributional assumption: heavy-tailed distributions with finite fourth moment. These results only hold for some of the proposed estimators (the testing approach, introduced in Section V.3) and are, thus, not discussed further elsewhere.

V.2.4 Simplifying settings

At times we will discuss unrealistic but simplifying settings to help with the exposition or to illuminate our theoretical findings.

Setting (**ECSS**, Equal Covariance and Sample Sizes). For all $k \in [\![B]\!]$, $\Sigma_k = \Sigma$ and $N_k = N$, which implies that $s_k^2, d_k^\bullet, d_k^{\mathrm{e}}$ do not depend on k.

Setting (**KC**, Known Covariances). For all $k \in [\![B]\!]$, Σ_k is known. Consequently, all derived quantities $\operatorname{Tr}\Sigma_k, \operatorname{Tr}\Sigma_k^2, d_k^\bullet, d_k^{\mathrm{e}}, s_k^2$ are also known.

We will first derive the estimators assuming known covariances (**KC**) but later provide estimates for covariance-related quantities if those are unknown. If the covariances and sample sizes are homogeneous (**ECSS**) the risks are more transparent and interpretable which will help to illuminate our theoretical findings. We insist that the final algorithms neither assume (**KC**) nor (**ECSS**).

V.2.5 Naive estimator aggregation

As announced earlier, without loss of generality we focus on estimating task $k = 1$. Furthermore, we focus on estimators which can be written as convex combinations (aggregation) of naive estimators. Let $\mathcal{S}_B$ denote the $(B-1)$-dimensional simplex, and $\boldsymbol{\omega} = (\omega_1, \dots, \omega_B) \in \mathcal{S}_B$ be a weight vector, then

$$\widehat{\mu}_{\boldsymbol{\omega}} := \sum_{k \in [\![B]\!]} \omega_k \widehat{\mu}_k^{\mathtt{NE}} \qquad \text{s.t.} \qquad \sum_{k \in [\![B]\!]} \omega_k = 1 \text{ and } \forall k \in [\![B]\!] : \omega_k \geq 0, \tag{V.6}$$

whose loss and risk will be abbreviated as $L_1(\boldsymbol{\omega})$ and $R_1(\boldsymbol{\omega})$, respectively. While the weight vector $\boldsymbol{\omega}$ may be data-dependent later, for the present considerations we assume that the weights are *deterministic*. In this case, using independence of the naive estimators and the notation $\Delta_k := \mu_k - \mu_1$, we restate the risk $R_1(\boldsymbol{\omega})$ by its bias-variance decomposition for a fixed $\boldsymbol{\omega}$ as

$$R_1(\boldsymbol{\omega}) = \Big\| \sum_{k \in [\![B]\!]} \omega_k (\mu_k - \mu_1) \Big\|^2 + \sum_{k \in [\![B]\!]} \omega_k^2 s_k^2 = \sum_{k,k' \in [\![B]\!]} \omega_k \omega_{k'} \langle \Delta_k, \Delta_{k'} \rangle + \sum_{k \in [\![B]\!]} \omega_k^2 s_k^2 \,, \tag{V.7}$$

where the first term corresponds to the (squared) bias and the second to the variance. Intuitively, we want to give higher weights to tasks that are close (small task bias $\|\Delta_k\|$) and can be accurately estimated (small naive risk s_k^2). At a first glance, we could set as a goal to find suitable weights $\boldsymbol{\omega}$ that minimise (V.7); this, however, would require full knowledge of the Gram matrix $(\langle \Delta_k, \Delta_{k'} \rangle)_{k,k' \in [\![B]\!]}$, in addition to the naive risks s_k^2. Estimation of the full Gram matrix, accurate enough to approach exact minimization of (V.7), appears unattainable if the number of tasks B is large and the Gram matrix becomes high-dimensional, which is the scenario we are interested in. For this reason, we will consider optimizing the risk given more limited information, which includes a subset of neighbouring tasks close to the target in relative sense but not their exact position. We define the oracle risk as the minimiser of the worst-case risk of (V.7) as if this partial information was known to the oracle.

We will consider two strategies to approach that oracle programme from data. In Section V.3 we aggregate only means close to the target which are identified by a test procedure. Minimization of an upper bound of the risk yields their weights. In Section V.4 we minimise directly an upper confidence bound of the aggregate risk (V.7) but have to take into account that the means that are further away induce a large uncertainty on the bias term. In both cases, we compare the obtained relative risk to that of the oracle. Additionally, we study the minimax risk under the oracle information in Section V.5 and whether the proposed estimators match it.

V.3 A testing approach

A low-risk aggregation estimator (V.6) combines naive estimations that — at best — provide a reduction in variance but add only a small bias, cf. (V.7). Our first approach explicitly controls the bias. We aim at identifying a subset of *neighbour* tasks whose means are sufficiently close to the target task. We then restrict the support of the weights to that subset and form a convex combination of neighbouring naive estimations. This approach and its analysis generalise ideas introduced in Marienwald et al. (2021). Let us first introduce some additional notation.

Definition V.3 (τ-neighbouring tasks)**.** Recall the notation $\Delta_k = \mu_k - \mu_1$. For a fixed $\tau > 0$, let $V_\tau \subseteq [\![B]\!]$ denote the set of all *τ-neighbouring tasks* (of task 1) as:

$$V_\tau := \Big\{ k \in [\![B]\!] : \|\Delta_k\|^2 \leq \tau s_1^2 \Big\}. \tag{V.8}$$

For $\tau = 0$, for the sake of later notational coherence we define $V_0 := \{1\}$ which deviates from (V.8) as V_0 does not contain any other tasks $k \neq 1$ even if $\Delta_k = 0$.

Note that this notion of τ-neighbourhood is relative to the naive risk of task 1, and that $1 \in V_\tau$ always holds.

Definition V.4 (Relative aggregated variance ν)**.** For a subset $U \subseteq [\![B]\!]$ of tasks, define their relative aggregated variance (to that of task 1) as:

$$\nu(U) := \frac{s^2(U)}{s_1^2}, \text{ with } s^2(U) := \left(\sum_{k\in U} \frac{1}{s_k^2}\right)^{-1}. \tag{V.9}$$

Observe that $s^2(U)$ is the variance of the optimal convex combination of unbiased, independent estimators that have different variances s_k^2 — a classical problem of statistics. The quantity $\nu(U)$ is, again, relative to the naive risk of task 1.

The quantity τ can be seen as the worst-case relative bias of a convex combination of their naive estimators for the goal of estimating μ_1, while $\nu(V_\tau)$ is a best-case relative variance (i.e., all the tasks in V_τ would in fact have mean μ_1). We introduce the following auxiliary function, which will capture an optimal trade-off between these two quantities. It provides a common reference value for the relative risks of our estimators and is of fundamental importance for the remainder of this manuscript.

Definition V.5. Define the function $\mathcal{B} : \mathbb{R} \times [0,1] \to [0,1]$ as

$$\mathcal{B}(\tau,\nu) := \left(\frac{\tau}{1+\tau}\right) + \left(\frac{1}{1+\tau}\right)\left(\frac{\nu}{1+\tau(1-\nu)}\right). \tag{V.10}$$

Observe that $\mathcal{B}(0,\nu) = \nu$, $\mathcal{B}(\tau,0) = \frac{\tau}{1+\tau}$, and $\mathcal{B}$ is increasing in both of its variables.

In the next section, we derive a form of optimal or "oracle" weights for combining naive estimators of tasks belonging to any given subset $V \subseteq V_\tau$, and identify $\mathcal{B}$ as a bound on its relative risk. The following sections (V.3.2 to V.3.4) are concerned with approximating the oracle bound by estimating unknown quantities and using a plug-in principle.

V.3.1 Oracle procedure

For a fixed $\tau > 0$, assume an oracle provides a set of neighbours V with the guarantee that $V \subseteq V_\tau$ holds. We restrict our attention to convex combinations of naive estimators only in set V, i.e., estimators $\widehat{\mu}_{\boldsymbol{\omega}}$ as in (V.6) with $\boldsymbol{\omega} \in \mathcal{S}_V$, the set of convex weights of support included in V. Using the Cauchy-Schwartz inequality in (V.7) (with $\Delta_1 = 0$), for such aggregated estimators we obtain the risk bound

$$R_1(\boldsymbol{\omega}) \leq \tau s_1^2(1-\omega_1)^2 + \sum_{k\in V} \omega_k^2 s_k^2, \quad \text{for all} \quad \boldsymbol{\omega} \in \mathcal{S}_V, \tag{V.11}$$

which can be optimised for $\boldsymbol{\omega}$. A bound on the oracle relative risk is presented next.

Lemma V.6. *Let $\tau > 0$ be fixed. For all $V \subseteq V_\tau$, the weights $\boldsymbol{\omega}_V^* \in \mathcal{S}_V$ that minimise (V.11) yield the bound*

$$\frac{R_1(\boldsymbol{\omega}_V^*)}{s_1^2} \leq \mathcal{B}(\tau, \nu(V)). \tag{V.12}$$

The oracle weights $\boldsymbol{\omega}_V^*$ are given by:

$$\omega_{V,k}^*(\tau, \boldsymbol{s}) = (1-\lambda)\mathbf{1}\{k=1\} + \lambda\frac{s^2(V)}{s_k^2}, \text{ where } \lambda := \frac{1}{1+\tau(1-\nu(V))}. \tag{V.13}$$

It holds $\mathcal{B}(\tau, \nu(V)) \in [\frac{\tau}{1+\tau}, 1]$, i.e., this bound cannot be better than $\frac{\tau}{1+\tau}$. We will call $\frac{\tau}{1+\tau}$ *best potential improvement* (that can be guaranteed by the oracle bound). The bound on the relative risk depends on the relative neighbourhood size τ and the relative aggregated variance $\nu(V)$. Because $\mathcal{B}$ increases in both variables, small τ and $\nu(V)$ are beneficial. This coincides with what we noted from the bias-variance decomposition (V.7). If τ is fixed, it is of advantage to consider as many τ-neighbours as possible so that $\nu(V)$ decreases, i.e., to take $V = V_\tau$. On the other hand, reducing the neighbourhood size τ reduces the bias but also leads to a smaller set of neighbours, ergo, a larger relative aggregated variance $\nu(V_\tau)$. Thus, there is a trade-off between both quantities. We may aim at a relative risk close to $\min_{\tau>0} \mathcal{B}(\tau, V_\tau)$ but for the remainder of this section we assume $\tau > 0$ fixed beforehand.

The following observations enable additional insight into the involved quantities:

(a) $\mathcal{B}(0, \nu(V)) = \nu(V)$, i.e., when $\tau \searrow 0$, which implies that all tasks in V have the same mean, the bound is given by the relative aggregated variance, as should be expected from the remark following Definition V.4.

(b) $\mathcal{B}(\tau, 0) = \frac{\tau}{1+\tau}$, the best potential improvement is reached when $s^2(V) \searrow 0$. This happens if at least one of the τ-neighbouring means is known with perfect precision and it becomes a "reference point". This scenario is comparable to the classical James-Stein setting, for which the origin is such a reference point and the James-Stein estimate improves most if the target is close to the origin (see Section I.1.1 for a detailed discussion). However, $s^2(V) \searrow 0$ also happens when τ-neighbours have a non-zero variance, but their number grows large.

(c) $\mathcal{B}(\tau, \nu(V))$ remains unchanged if we replace a group of neighbours $V \setminus \{1\}$ by a single τ-neighbour with variance $s^2(V \setminus \{1\})$.

In (**ECSS**) setting, the oracle bound (V.12) is similar to the one obtained with the method of Section III (see Eq.(III.8)).

V.3.2 From an oracle to an empirical procedure

In practice, the oracle information about the relative neighbours is unavailable. However, we can hope to approach the oracle setting by estimating the set of τ-neighbours V_τ and their risks s_k^2. We will assume that such estimates are independent of the samples used to compute $(\widehat{\mu}_k^{\mathtt{NE}})_{k\in[\![B]\!]}$. (To this end, one might resort to sample splitting.) The independence assumption of estimates is emphasised by a tilde notation: $(\widetilde{V}, \widetilde{\boldsymbol{s}}^2)$.

The simplest is to plug in such estimates into the oracle formula (V.13). The next proposition quantifies how the relative risk of the plug-in procedure can be bounded, provided the estimation error is.

Proposition V.7. *Let $\tau > 0$ be fixed. Assume $\widetilde{V} \subseteq [\![B]\!], \widetilde{\boldsymbol{s}}^2 = (\widetilde{s}_k^2)_{k\in[\![B]\!]} \in \mathbb{R}_+^B$ are possibly random but independent of the samples in model (V.1). Let V^* be some deterministic reference set, such that $1 \in V^*$. Let $(\widetilde{V}, \widetilde{\boldsymbol{s}}^2)$ be plugged in for $(V, \boldsymbol{s}^2)$ into (V.13), giving rise to weight vector $\widetilde{\boldsymbol{\omega}}$. Conditionally to the event*

$$\begin{cases} V^* \subseteq \widetilde{V} \subseteq V_\tau, \\ \left|\widetilde{s}_k^2 - s_k^2\right| \leq \eta s_k^2, \text{ for all } k \in \widetilde{V}, \text{ and some } \eta \in [0, 1), \end{cases} \tag{V.14}$$

it holds

$$\frac{R_1(\widetilde{\boldsymbol{\omega}})}{s_1^2} \leq \left(\frac{1+\eta}{1-\eta}\right)\mathcal{B}\big(\tau, \nu(\widetilde{V})\big) \leq \left(\frac{1+\eta}{1-\eta}\right)\mathcal{B}\big(\tau, \nu(V^*)\big). \tag{V.15}$$

Comparing the oracle relative risk bound (V.12) with that of the empirical procedure (V.15), note first the requirement that all estimated neighbours are τ-neighbours ($\widetilde{V} \subseteq V_\tau$); secondly, the oracle risk is deteriorated by two factors: the excess factor $(1+\eta)/(1-\eta) \geq 1$ which quantifies what we lose due to

estimation of the neighbours' risks; and the replacement of the set of true neighbours by the smaller set V^*, under the requirement that $V^* \subseteq \widetilde{V}$ holds. To summarise, we expect the risk of the empirical procedure to be close to the oracle risk if (1) the relative estimation error η for naive risks is small, and (2) we can guarantee the "sandwiching" property $V^* \subseteq \widetilde{V} \subseteq V_\tau$, with V^* as large as possible; typically we would be satisfied with $V^* = V_{(1-\varepsilon)\tau}$ for a small ε.

The next sections will introduce such estimates and the fulfillment of event (V.14) under certain conditions, starting with the estimation of neighbour tasks.

V.3.3 Finding neighbours (known covariances)

For now let us assume (**KC**); we will generalise to unknown covariances in the next section. Accordingly, the naive risks s_k^2 are known, so that $\eta = 0$ in the context of (V.15), and we focus on the estimation of the set of neighbours. We assume that we are doing so using independent "tilde" data $(\widetilde{X}_\bullet^{(k)})_{k\in[\![B]\!]}$ which are drawn from (V.1) but independent of $(X_\bullet^{(k)})_{k\in[\![B]\!]}$ (e.g., using sample splitting). For clarity $X_\bullet^{(k)}$ and $\widetilde{X}_\bullet^{(k)}$ are assumed to be of the same size N_k. Given the first requirement $\widetilde{V} \subseteq V_\tau$, it is natural to think of $\widetilde{V}$ as the output of a multiple test procedure (for which the null hypothesis for task k is *not* being a τ-neighbour, i.e., $\|\Delta_k\| > \tau s_1^2$).

Our approach is based on results for two-sample mean vector testing of Section IV. Assume $N_k \geq 2$ for all $k \in [\![B]\!]$. For $k \in [\![B]\!] \setminus \{1\}$, we form an unbiased estimator for $\|\Delta_k\|^2$ based on the U-statistics

$$\widetilde{U}_k := \sum_{\ell\in\{1,k\}} \sum_{\substack{i,j=1\\ i\neq j}}^{N_\ell} \frac{\langle \widetilde{X}_i^{(\ell)}, \widetilde{X}_j^{(\ell)} \rangle}{N_\ell(N_\ell-1)} - 2\sum_{i=1}^{N_1}\sum_{j=1}^{N_k} \frac{\langle \widetilde{X}_i^{(1)}, \widetilde{X}_j^{(k)} \rangle}{N_1 N_k}\,. \tag{V.16}$$

The following proposition is a direct consequence of Proposition IV.6:

Proposition V.8. *Assume (**GS**), (**KC**) hold and let $\alpha \in (0,1)$, $\tau > 0$ be fixed. Let $\widetilde{T}_k^{(\tau)}$ be given by*

$$\widetilde{T}_k^{(\tau)} := \mathbf{1}\Big\{\widetilde{U}_k \leq \tau s_1^2\Big\}. \tag{V.17}$$

Put for $k \in [\![B]\!]$

$$\tau_{\min}^k := 32\left(\frac{1}{\sqrt{d_1^\bullet}} + \frac{s_k^2/s_1^2}{\sqrt{d_k^\bullet}}\right)\log(8\alpha^{-1})\,, \tag{V.18}$$

then it holds:

$$\text{if } \|\mu_1-\mu_k\|^2 > \tau_k^+ s_1^2: \qquad \mathbb{P}\Big[\widetilde{T}_k^{(\tau)} = 1\Big] \leq \alpha; \tag{V.19}$$

$$\text{if } \|\mu_1-\mu_k\|^2 \leq \tau_k^- s_1^2: \qquad \mathbb{P}\Big[\widetilde{T}_k^{(\tau)} = 0\Big] \leq \alpha. \tag{V.20}$$

where $\tau_k^\pm = \Big(\sqrt{\tau} \pm \sqrt{\tau_{\min}^k}\Big)_+^2$.

Equations (V.19)-(V.20) can be understood as controls of the type I/II error level for the test of $\|\Delta_k\|^2 > \tau_k^+ s_1^2$ versus the alternative $\|\Delta_k\|^2 \leq \tau_k^- s_1^2$. It is possible to make the original null hypothesis $\|\Delta_k\|^2 > \tau s_1^2$ appear through notation translation ($\sqrt{\tau} \leftarrow \sqrt{\tau_k^-}$, $\sqrt{\tau_k^+} \leftarrow \sqrt{\tau}$, if we assume additionally $\tau \geq \tau_{\min}^k$). We prefer to keep the above more symmetric form, also because the rejection set (V.17) has a simple form, used in practice.

The test is able to identify mean differences very accurately relative to the target threshold τs_1^2 if $\tau \gg \tau_{\min}^k$. Formula (V.18) highlights the crucial role of the effective dimensionality for this minimal threshold of reliable detection. In the simplified (**ECSS**) setting, this threshold is simply of order $1/\sqrt{d_1^\bullet}$.

This reflects the known phenomenon that testing is more reliable than estimation in high dimensions; distances that can be detected might be of smaller order than the typical estimation error. For fixed τ and increasing dimension, the inconclusive gap between the null and the alternative vanishes with increasing dimension — a desirable property given the sandwiching property that we aim for (see (V.14)).

In general non-(**ECSS**) configurations, we still want to keep $\tau^k_{\min}$ small of order $1/\sqrt{d^\bullet_1}$. In view of the second term in (V.18), this suggests to only consider tasks with $s_k^2/\sqrt{d^\bullet_k} \le \varsigma s_1^2/\sqrt{d^\bullet_1}$ for some constant $\varsigma \ge 1$. To this aim, denote the set of tasks satisfying this criterion as

$$W_{(\varsigma)} := \left\{ k \in [\![B]\!] : \frac{s_k^2}{\sqrt{d^\bullet_k}} \le \varsigma \frac{s_1^2}{\sqrt{d^\bullet_1}} \right\} = \left\{ k \in [\![B]\!] : \frac{\|\Sigma_k\|_2}{N_k} \le \varsigma \frac{\|\Sigma_1\|_2}{N_1} \right\}, \tag{V.21}$$

and correspondingly the set of whittled down neighbours as

$$V_{\tau,\varsigma} := V_\tau \cap W_{(\varsigma)}. \tag{V.22}$$

Note that since we are under (**KC**), the set $W_{(\varsigma)}$ is assumed to be fully known for now. (We will consider estimating it in the next section.) Then the following corollary makes the obtained sandwiching property explicit:

Corollary V.9. *Let $\varsigma \ge 1$ be fixed. Assume (**GS**) and (**KC**) hold and let $\alpha \in (0,1)$. Then, defining*

$$\widetilde{V}_{\tau,\varsigma} := \left\{ k \in [\![B]\!] : \widetilde{T}_k^{(\tau)} = 1 \right\} \cap W_{(\varsigma)}$$

(where $\widetilde{T}_k^{(\tau)}$ is as in (V.17)), with probability at least $1-\alpha$ it holds

$$V_{\tau^-,\varsigma} \subseteq \widetilde{V}_{\tau,\varsigma} \subseteq V_{\tau^+}, \tag{V.23}$$
$$\text{where } \tau^\pm := \left(\sqrt{\tau} \pm \sqrt{\varsigma \tau^\circ_{\min}}\right)_+^2, \quad \tau^\circ_{\min} := 64 \log(8B\alpha^{-1})/\sqrt{d^\bullet_1}.$$

The sandwiching property (V.23) provides a direct link to Proposition V.7. More specifically, Corollary V.9 together with Proposition V.7 guarantee with high probability that the bound on the relative risk of the plug-in estimate $\widehat{\mu}_{\widetilde{\omega}}$ of (V.13) using the estimated set of neighbours $\widetilde{V}_{\tau,\varsigma}$ is bounded by $\mathcal{B}\big(\tau^+, \nu(V_{\tau^-,\varsigma})\big)$ (recall $\eta = 0$ for now because of (**KC**), and $\widetilde{V}_0 := \{1\}$). Furthermore, for fixed τ, if $d^\bullet_1/(\log B)^2 \to \infty$ then $\tau^\circ_{\min}$ vanishes and it holds $\tau^- \approx \tau \approx \tau^+$. Under (**ECSS**), we can simply take $\varsigma = 1$ and have $V_{\tau^-,\varsigma} = V_{\tau^-}$, ensuring a relative risk very close to the oracle $\mathcal{B}\big(\tau, \nu(V_\tau)\big)$. In a general context, there is an additional trade-off through the choice of the constant ς. In both cases, closeness to the oracle relative risk *improves* with increasing effective dimensionality.

V.3.4 Unknown covariances

In a realistic setting the covariances are unknown, especially in high dimensions. In this section, we estimate all quantities relevant for the fulfilment of Proposition V.7, using the same independent "tilde" data $(\widetilde{X}^{(k)}_\bullet)_{k\in[\![B]\!]}$ as in the previous section. For simplicity we assume that the sizes N_k of the "tilde" samples are the same as that of the main sample, as we would get by equal-size splitting. Observe that it is not necessary to estimate the full covariance matrices Σ_k, but only scalar quantities related to their Schatten norms. In particular, in the Gaussian setting we have the following result for the natural unbiased estimators of s_k^2:

Proposition V.10. *Let $\widetilde{s}_k^2 := \frac{1}{N_k(N_k-1)} \sum_{i=1}^{N_k} \big\|\widetilde{X}_i^{(k)} - \widetilde{\mu}_k^{\text{NE}}\big\|^2$, where $\widetilde{\mu}_k^{\text{NE}} := N_k^{-1} \sum_{i=1}^{N_k} \widetilde{X}_i^{(k)}$, and let $\alpha \in (0,1)$. Assume (**GS**) holds. Then with probability at least $1-\alpha$:*

$$\forall k \in [\![B]\!] : \qquad \big|\widetilde{s}_k^2 - s_k^2\big| \le \left(4\sqrt{2}\frac{\log(2B\alpha^{-1})}{\sqrt{d^\bullet_k N_k}}\right) s_k^2. \tag{V.24}$$

When $N_k \gtrsim \log^2(2B\alpha^{-1})$ for all k, the estimation of s_k^2 has relative accuracy of order $1/\sqrt{d_k^\bullet}$ with probability $(1-\alpha)$. This finding can be used for the fulfillment of the second requirement of condition (V.14). It also allows to preserve the qualitative results of Proposition V.8 (up to numerical factors) for test (V.17) wherein $\widetilde{s}_1^2$ is plugged in for s_1^2. Finally, we also replace $\|\Sigma_k\|_2$ in the definition (V.21) of set $W_{(\varsigma)}$ by suitable estimators; Proposition V.28 gives the details. It provides a quantitatively precise version of the sandwiching property analogous to (V.23) with all unknown quantities are replaced by their proposed estimators.

We combine the obtained results in an illustrative example. It shows a fully empirical algorithm that approximates the (whittled down) oracle $B(\tau, V_{\tau,\varsigma})$ (numerical constants are made explicit for concreteness but not meant to be sharp):

Proposition V.11. *Assume (**GS**) holds. Let $\alpha \in (0, 1/3)$. Consider the following plug-in versions of the quantities appearing in (V.17), (V.21):*

$$\widetilde{W}_{(\varsigma)} := \left\{ k \in [\![B]\!] : \frac{\widetilde{Z}_k^{(2)}}{N_k} \le \varsigma \frac{\widetilde{Z}_1^{(2)}}{N_1} \right\}, \qquad \widetilde{\widetilde{T}}_k^{(\tau)} := \mathbf{1}\Big\{ \widetilde{U}_k \le \tau \widetilde{s}_k^2 \Big\}, \tag{V.25}$$

where $\widetilde{s}_k^2$ as in Prop. V.10, and $\widetilde{Z}_k^{(2)}$ estimates $\|\Sigma_k\|_2$ as defined in (V.56). Define the set of estimated τ-neighbours

$$\widetilde{\widetilde{V}}_{\tau,\varsigma} := \Big\{ k \in \widetilde{W}_{(\varsigma)} : \widetilde{\widetilde{T}}_k^{(\tau)} = 1 \Big\}. \tag{V.26}$$

Assume $N_k \ge a(4 + \log(2B\alpha^{-1}))^4$ for all $k \in [\![B]\!]$, for a big enough numerical constant a ($a = 4400$ works). For fixed $\tau > 0, \varsigma \ge 1$, consider the weights $\widetilde{\boldsymbol{\omega}}^\sharp$ obtained by the modified *plug-in $\big(\widetilde{\widetilde{V}}_{\widetilde{\tau},3\varsigma}, \widetilde{\boldsymbol{s}}^2\big)$ for $(V, \boldsymbol{s^2})$ in (V.13), where*

$$\widetilde{\tau} := \left(1 + \frac{1}{60\sqrt{\widetilde{d}_1^\bullet}}\right)\Big(\sqrt{\tau} + \sqrt{6\varsigma\widetilde{\tau}_{\min}^\circ}\Big)^2; \qquad \widetilde{\tau}_{\min}^\circ := \frac{32\big(\log(8B\alpha^{-1})\big)}{\sqrt{\widetilde{d}_1^\bullet}}; \qquad \sqrt{\widetilde{d}_1^\bullet} := \frac{N_1 \widetilde{s}_1^2}{\widetilde{Z}_1^{(2)}}. \tag{V.27}$$

Then with probability at least $1 - 3\alpha$ over the draw of the "tilde" sample $(\widetilde{X}_\bullet^{(k)})_{k\in[\![B]\!]}$, it holds

$$\frac{R_1(\widetilde{\boldsymbol{\omega}}^\sharp)}{s_1^2} \le \left(1 + \frac{1}{10\sqrt{\min_k d_k^\bullet}}\right)\left(1 + \frac{30\sqrt{\varsigma \log(8B\alpha^{-1})}}{(d_1^\bullet)^{\frac{1}{4}}\sqrt{\tau}}\right)^2 \mathcal{B}\big(\tau, \nu(V_{\tau,\varsigma})\big),$$

where the expected risk is with respect to the main sample $(X_\bullet^{(k)})_{k\in[\![B]\!]}$.

V.3.5 Discussion

To summarise, for fixed values of $\tau, \varsigma, B, (N_k)_{k\in[\![B]\!]}$, the bound on the relative risk of $\widetilde{\boldsymbol{\omega}}^\sharp$ becomes arbitrarily close to the oracle bound in the high-dimensional asymptotics $d_1^\bullet \to \infty$. We stress that this applies for *fixed* sample sizes N_k, provided $N_k \gtrsim \log^4 B$. Consequently, the fully empirical procedure is (with high probability) not worse than the naive estimator up to a risk factor very close to 1 (since the oracle bound $\mathcal{B}$ is always less than 1), and potentially performs much better if there are many true τ-neighbouring tasks (again, as reflected by the oracle factor). The conclusion still holds true if $\tau, \varsigma, B, (N_k)$ vary with $d_1^\bullet$ ($\tau \to 0$ and/or $B \to \infty$ being the most interesting situations) provided $\varsigma \log(B)/\sqrt{d_1^\bullet} = o(\tau)$ holds and as $N_k \gtrsim \log^4 B$ as before.

Beyond the Gaussian setting. The results presented above hold under the Gaussian distributional assumptions (**GS**). However, the required components — specifically, concentration of estimators for distances between two means and for Schatten norms of the covariances — can be extended with appropriate modifications to the bounded (**BS**) and heavy-tailed (**HT**) distributional settings. Detailed results are presented in in Section V.9.4 and show the qualitative robustness of our approach beyond the Gaussian setting.

Beyond the testing approach. The testing approach has two flaws: first, the theoretical necessity to partition the data entails a certain loss of efficiency, such as a reduction by a factor of $1/2$ when the data is equally split. This consideration has been disregarded in the preceding discussion, where the oracle risk was restricted to the main sample. Second, the issue of parameter selection of τ and ς persists. As previously elucidated, the oracle relative risk $\mathcal{B}$ exhibits a bias-variance trade-off: the aggregated variance decreases with an increase in the number of τ-neighbours, consequently, with the worst-case relative bias τ. Ideally, parameters should be adaptively chosen to strive for optimal oracle improvement $\min_{\tau\geq 0,\varsigma\geq 1}\mathcal{B}\big(\tau,\nu(V_{\tau,\varsigma})\big)$. The next section introduces an alternative approach pursuing this objective. Additionally, Section V.5 analyses whether the derived bounds are optimal.

V.4 A "Q-aggregation" approach

In this section, we propose an alternative approach for forming the weights of the convex combination estimator (V.6). The weights are found by direct minimization of an upper confidence bound of the risk $R_1(\boldsymbol{\omega})$, i.e.,

$$\widehat{\boldsymbol{\omega}} \in \arg\min_{\boldsymbol{\omega}\in\mathcal{S}_B}\Big(\widehat{L}_1(\boldsymbol{\omega}) + u\widehat{Q}_1(\boldsymbol{\omega})\Big)\,, \tag{V.28}$$

where $\widehat{L}_1(\boldsymbol{\omega})$ is an unbiased estimate of the risk. The idea of this scheme bears resemblance to Q-aggregation (Lecué and Rigollet, 2014), because the objective function will be a quadratic function of $\boldsymbol{\omega}$. The objective aims at taking into account all individual distances between the bags, rather than selecting those less than a fixed threshold as in the testing approach. The penalization term $\widehat{Q}_1(\boldsymbol{\omega})$ shall be a high probability upper bound on the difference between estimated and true *loss* $(\widehat{L}_1(\boldsymbol{\omega}) - L_1(\boldsymbol{\omega}))$. Observe that the penalization term also depends on the weight vector, since giving more weight to tasks that are further away from the target (large $\|\Delta_k\|$) will result in a larger variability of the risk estimate $\widehat{L}_1(\boldsymbol{\omega})$. The parameter u is a calibration constant. Compared to the testing approach, one advantage is that it is not necessary to choose the parameters τ and ς. Furthermore no sample splitting is needed. On the other hand, the procedure is more computationally demanding since there is no closed form solution to (V.28). Instead, a solution $\widehat{\boldsymbol{\omega}}$ can be obtained by exponentiated gradient descent on the simplex (Kivinen and Warmuth, 1997).

We present specific choices for $\widehat{L}_1(\boldsymbol{\omega}), \widehat{Q}_1(\boldsymbol{\omega})$ and an analysis of the relative risk of the resulting Q-aggregation estimator for (**GS**) in Section V.4.1 and for (**BS**) thereafter. In contrast to Lecué and Rigollet (2014), we focus on the effect of the dimension rather than that of the sample size which provides a novel analysis.

V.4.1 Gaussian setting

Under assumption (**GS**) we propose to use the following estimates to form the Q-aggregation estimator:

$$\widehat{L}_1(\boldsymbol{\omega}) = \Big\|\sum_{k=2}^{B}\omega_k(\widehat{\mu}_i^{\texttt{NE}} - \widehat{\mu}_1^{\texttt{NE}})\Big\|^2 + (2\omega_1 - 1)\widehat{s}_1^2\,, \tag{V.29}$$

$$\widehat{s}_1^2 := \frac{1}{N_1(N_1-1)}\sum_{i=1}^{N_1}\Big\|X_i^{(1)} - \widehat{\mu}_1^{\texttt{NE}}\Big\|^2\,, \tag{V.30}$$

$$\widehat{Q}_1(\boldsymbol{\omega}) := \sum_{k=2}^{B}\omega_k\sqrt{\frac{\widehat{q}_k}{N_1}}, \quad \text{where} \quad \widehat{q}_k := \frac{1}{N_1-1}\sum_{i=1}^{N_1}\Big\langle\widehat{\mu}_1^{\texttt{NE}} - \widehat{\mu}_k^{\texttt{NE}}, X_i^{(1)} - \widehat{\mu}_1^{\texttt{NE}}\Big\rangle^2\,. \tag{V.31}$$

It can be checked easily that $\widehat{s}_1^2$ is an unbiased estimator of the naive risk s_1^2, and that the estimator $\widehat{L}_1(\boldsymbol{\omega})$ is an unbiased estimate of the *conditional* risk $\mathbb{E}\big[\widehat{L}_1(\boldsymbol{\omega}) - L_1(\boldsymbol{\omega})\big|X_\bullet^{(k)}, k\geq 2\big] = 0$. With these choices we establish the following result for the average risk of the Q-aggregation estimator:

Proposition V.12. *Assume (**GS**) holds, and let $u_0 \in \mathbb{R}_+$ be fixed such that $\log(17B) \le u_0 \le (N_1-1)/2$. With $\widehat{L}_1(\boldsymbol{\omega})$ and $\widehat{Q}_1(\boldsymbol{\omega})$ as defined in (V.29),(V.31), let*

$$\widehat{\boldsymbol{\omega}} \in \arg\min_{\boldsymbol{\omega}\in\mathcal{S}_B} \Big(\widehat{L}_1(\boldsymbol{\omega}) + 16\sqrt{u_0}\,\widehat{Q}_1(\boldsymbol{\omega})\Big). \tag{V.32}$$

Then it holds:

$$\frac{R_1(\widehat{\boldsymbol{\omega}})}{s_1^2} \le \frac{1}{s_1^2}\min_{\boldsymbol{\omega}\in\mathcal{S}_B}\Big[R_1(\boldsymbol{\omega})(1+CBe^{-u_0/2}) + CQ_1(\boldsymbol{\omega})\sqrt{u_0}\Big] + C\frac{u_0}{\sqrt{d_1^\bullet}}, \tag{V.33}$$

where $C>0$ is an absolute constant, and (recalling $\Delta_k = \mu_k - \mu_1$)

$$Q_1(\boldsymbol{\omega}) := \sum_{k=2}^{B}\omega_k\sqrt{\frac{q_k}{N_1}}, \quad \textit{with} \quad q_k := \Delta_k^T\Sigma_1\Delta_k + \frac{\operatorname{Tr}\Sigma_1\Sigma_k}{N_k}\,. \tag{V.34}$$

The above bound (V.33) has the form of an "oracle inequality" relating the relative risk of the Q-aggregation approach to the minimum of the attainable relative risk of any aggregation estimator with fixed weight $\boldsymbol{\omega}$ but with a penalization term $Q_1(\boldsymbol{\omega})$. The extra additive term (outside the minimum) vanishes in high effective dimension, but indicates that the relative risk bound cannot be better than $O(\log B/\sqrt{d_1^\bullet})$. We also emphasise the requirement $\log B \lesssim N_1$ implicit in the condition on the calibration parameter u_0. The effect of the penalization term $Q_1(\boldsymbol{\omega})$ on the oracle bound (V.33) might appear obscure: depending on the weights $\boldsymbol{\omega}$, the penalization might outweigh the main risk term $R_1(\boldsymbol{\omega})$. It is noteworthy that this term penalises tasks with distant means (term $\Delta_k^T\Sigma_1\Delta_k$) or with high variance (term $\operatorname{Tr}\Sigma_1\Sigma_k/N_k$). To provide further clarification, we present the following corollary which bounds the relative risk of the Q-aggregation method in terms of the relative risk of the oracle testing approach $\mathcal{B}(\tau,\nu)$:

Corollary V.13. *Assume (**GS**) holds. Let $u_0 \in \mathbb{R}_+$ be fixed, such that $\log 17B \le u_0 \le (N_1-1)/2$, and $\widehat{\boldsymbol{\omega}}$ as defined in (V.32). Then it holds:*

$$\frac{R_1(\widehat{\boldsymbol{\omega}})}{s_1^2} \le \Big(1+CBe^{-u_0/2}\Big)\inf_{\substack{\tau\ge 0\\ \varsigma\ge 1}}\Big[\mathcal{B}\big(\tau,\nu(V_{\tau,\varsigma})\big) + C\varsigma\sqrt{\frac{u_0}{d_1^{\mathrm{e}}}}\Big]. \tag{V.35}$$

where $C>0$ is an absolute constant, $\mathcal{B}(\cdot,\cdot)$, $\nu(\cdot)$ are as defined in (V.10), (V.9) and $V_{\tau,\varsigma}$ as in (V.21)-(V.22).

As a simple illustration, assume the tasks satisfy (**ECSS**) and have equal means ($\mu_k = \mu_1$ for $k \in [\![B]\!]$), but the estimator does not have this information. The oracle merges all tasks and has relative risk $\inf_{\tau,\varsigma}\mathcal{B}\big(\tau,\nu(V_{\tau,\varsigma})\big) = B^{-1}$ for $\tau\to 0, \varsigma = 1$. For $u_0 = \log 17B$, the relative risk of the Q-aggregation method (V.35) becomes

$$\frac{R_1(\widehat{\omega})}{s_1^2} \le C\max\left\{\frac{1}{B}, \sqrt{\frac{\log B}{d_1^{\mathrm{e}}}}\right\},$$

where $C\approx 1$ if d_1^{e} and B are large. We observe again a blessing of dimensionality; the best improvement is obtained when d_1^{e} is high ($d_1^{\mathrm{e}} \ge B^2\log B$ ensures a relative risk bound of order $1/B$, which is the best improvement even if the information of equal means had been known).

The standard James-Stein problem (Section I.1.1) can be cast as a particular limiting case of our general setting (V.1) with only two bags, the second of which with known mean equal to 0 and serving as a reference point:

Assumption V.14 (**JS**, James-Stein setting)**.** *$B=2$, $\mu_2 = 0$, formally $N_2 = \infty$ and $s_2^2 = 0$.*

In the (**JS**) setting, the Q-aggregation method exhibits the same asymptotic behaviour as the James-Stein estimator $\widehat{\mu}^{\mathbf{JS}+}$ without knowing the covariance Σ. Corollary V.15 is deduced from Proposition V.12.

Corollary V.15. *Assume (**JS**) and (**GS**), let $N_1 \geq 7$, $(N_1-1)/2 \geq u_0 \geq 3$, and $\widehat{\boldsymbol{\omega}}$ as defined in (V.32). Then:*

$$\frac{R_1(\widehat{\boldsymbol{\omega}})}{s^2} \leq \frac{\|\mu_1\|^2}{s_1^2+\|\mu_1\|^2}\Big(1+Ce^{-u_0/2}\Big) + C\sqrt{\frac{u_0}{d^{\mathrm{e}}}}\,, \tag{V.36}$$

where $C>0$ is some absolute constant.

The first term is, up to the multiplicative factor, Stein's error $\tau/(1+\tau)$ with $\tau = \|\mu^2\|/s^2$. In the dimensional asymptotic $d^{\mathrm{e}} \to \infty$, assume $u_0 \to \infty$ such that $u_0 = o(d^{\mathrm{e}})$ and suppose the mean satisfies $\|\mu\|^2 \leq \tau s^2$, then the estimator attains the Pinsker bound (II.4):

$$\lim_{d^{\mathrm{e}}\to\infty} \sup_{\substack{\mu_1,s_1:\\ \|\mu_1\|^2\leq\tau s_1^2}} \frac{R_1(\widehat{\boldsymbol{\omega}})}{s_1^2} \leq \frac{\tau}{1+\tau}\,.$$

V.4.2 Comparison with the testing approach

Let us compare the bounds obtained for the test method (Proposition V.11) to that for the Q-aggregation approach (Corollary V.13), in high-dimensional asymptotics $d_1^\bullet, d_1^{\mathrm{e}} \to \infty$. We start with an analysis of the conditions on the other parameters $\{\tau, \varsigma, B, (N_k)_{k\in[\![B]\!]}\}$ under which the obtained bounds guarantee that the relative risk of either method is bounded by the oracle bound $\mathcal{B}\big(\tau, \nu(V_{\tau,\varsigma})\big)$ up to a factor asymptotically converging to 1, a property which we call "oracle-consistency" for short.

Recall from Section V.3.5 that the relative risk of the test method is oracle-consistent (as $d_1^\bullet \to \infty$), provided $\varsigma\log(B)/\sqrt{d_1^\bullet} = o(\tau)$ and $N_k \gtrsim \log^4 B$ hold. Aside from these conditions the parameters $\tau, \varsigma, B, (N_k)$ can vary with $d_1^\bullet$. On the other hand, (V.35) shows that the Q-aggregation method is oracle-consistent (as $d_1^{\mathrm{e}} \to \infty$) with respect to *any* (τ,ς) provided that $N_1 \gtrsim \log(Bd_1^{\mathrm{e}})$, and $\varsigma\sqrt{\log(Bd_1^{\mathrm{e}})/d_1^{\mathrm{e}}} = o(\tau)$ (taking $u_0 = 2\log Bd_1^{\mathrm{e}}$). The additive terms in (V.35) are then negligible compared to $\mathcal{B}(\tau,\nu)$, due to $\mathcal{B}(\tau,.) \geq \tau/(1+\tau)$. Note also that it does not require any condition on N_k for $k \neq 1$.

If $d_1^\bullet$ and d_1^{e} are of the same order (e.g. in the isotropic setting), the above parameter conditions for consistency of either method are very similar with only minor differences. One such difference is that the test method is guaranteed to be oracle-consistent even if $B, \tau, \varsigma, (N_k)$ are fixed, i.e., must not change as $d_1^\bullet \to \infty$; while we require $N_1 \to \infty$ (though only at a logarithmic rate in $B, d_1^\bullet$) to warrant oracle consistency of the aggregation estimator. If d_1^{e} is of order $\sqrt{d_1^\bullet}$ (for example, for a slow power decrease of the eigenvalues λ_i of the covariance, $\lambda_i = i^{-\alpha}$ for $1 \leq i \leq d$ and $\alpha \in (1/2, 1)$), then the oracle consistency conditions for the Q-aggregation method are narrower.

Still, one has to keep in mind that oracle-consistency for the testing approach only holds for the specific parameters (τ,ς) that must be provided by the user, while the Q-aggregation method is oracle consistent with respect to any choice (τ,ς) satisfying the delineated conditions. In other words, the relative risk of the Q-aggregation method qualitatively enjoys the same asymptotic guarantees as the testing approach with *optimally selected* τ and ς subject to the above conditions. This and the fact that the Q-aggregation does not use data splitting is a strong argument in its favour. On the other hand, the testing method has the advantage of being more flexible and easily adapts to non-Gaussian distributions, e.g., bounded or heavy-tailed distributions (see Section V.9.4). With a modification of the penalization term, the Q-aggregation method can also be applied to bounded distributions, as shown next, but it currently does not accommodate heavy-tailed data distributions.

V.4.3 Bounded setting

Our results for the Q-aggregation estimator can be extended to the bounded setting (**BS**) where the data lie in a ball of radius M centred in 0. A precise value for M is often known. For example, if the data lies in a reproducing kernel Hilbert space associated with a bounded kernel, M^2 will be the bound on the

kernel. The methodology we propose for (**BS**) closely resembles the one outlined for the Gaussian setting. It utilises the same estimates, (V.29)-(V.30)-(V.31), for the risk estimation and its deviations. In order to compensate the lack of regularity of bounded compared to Gaussian data, an additional penalization term $\widehat{Q}^{\mathbf{BS}}(\omega)$ is introduced, which depends on M.

Proposition V.16. *Assume (**BS**). Let $u_0 \in \mathbb{R}_+$ with $2\log N_1 + \log(B) \leq u_0 \leq N_1$, and*

$$\widehat{\boldsymbol{\omega}} \in \arg\min_{\boldsymbol{\omega}\in\mathcal{S}_B} \Big(\widehat{L}_1(\boldsymbol{\omega}) + 4\sqrt{2u_0}\widehat{Q}_1(\boldsymbol{\omega}) + C_0 u_0 \widehat{Q}_1^{\mathbf{BS}}(\boldsymbol{\omega})\Big), \tag{V.37}$$

where $\widehat{L}_1, \widehat{Q}_1$ are defined in (V.29), (V.31) resp., $C_0 > 1424$ works, and

$$\widehat{Q}_1^{\mathbf{BS}}(\boldsymbol{\omega}) = \frac{M}{N_1}\sum_{i=2}^{B} \omega_i \big\|\widehat{\mu}_i^{NE} - \widehat{\mu}_1^{NE}\big\|. \tag{V.38}$$

Assume $N_k \geq (d_k^\bullet)^\beta$ for some $\beta > 0$ and all $k \in [\![B]\!]$, then:

$$\frac{R_1(\widehat{\boldsymbol{\omega}})}{s_1^2} \leq \min_{\tau>0,\varsigma\geq 1} \big(\mathcal{B}\big(\tau, \nu(V_{\tau,\varsigma})\big) + C\varsigma\varepsilon\big) + C\phi_1\varepsilon\,, \qquad \varepsilon := \max\left\{\sqrt{\frac{u_0}{d_1^{\mathrm{e}}}}\,,\, \frac{u_0}{(d_1^\bullet)^{\beta/2}}\right\}, \tag{V.39}$$

where $\mathcal{B}(\cdot,\cdot)$, $\nu(\cdot)$ are as defined in (V.10), (V.9), $V_{\tau,\varsigma}$ as in (V.21)-(V.22), C an absolute constant, and $\phi_1 := M^2/\operatorname{Tr}\Sigma_1$.

The quantity β reflects the trade-off between the requirement on the number of samples and the rate of convergence to the oracle bound. A bound similar to that in the Gaussian case will only be obtained if a stricter condition on the bag sizes is met ($N_k \gtrsim d_k^\bullet$ instead of $N_1 \gtrsim \log B$ as in Corollary V.13). In contrast to (V.35), there is no multiplicative constant in front of the bound, however, the additive term now involves the quantity ϕ_1 (see Section V.10 for a discussion of this quantity in the framework of kernel mean embedding (KME) estimation with a bounded kernel, which is our primary motivation for analyzing the bounded setting).

V.5 Minimax results

This section explores if the oracle relative risk upper bound $\mathcal{B}(\tau, \nu(V_\tau))$ as defined in (V.12), which has been utilised as benchmark in previous sections, is optimal in a minimax sense. As before, we will first examine the estimation of a single mean. Subsequently, we extend the analysis to the compound relative risks averaged over tasks.

Our aim is to establish minimax bounds matching the upper bounds over distribution classes that are as restrictive as possible. Since a minimax lower bound on a distribution class also applies to every superclass containing it, bounds on restrictive classes are more insightful. To achieve this, we narrow down the distribution classes by fixing as many parameters as possible to arbitrary values. As employed throughout this manuscript, we will adopt a high-dimensional asymptotics viewpoint and focus on minimax statements as the effective dimension grows large.

V.5.1 Single task relative risk

We derive a lower minimax bound for a class of distributions that closely match the assumptions proposed to introduce the oracle bound (V.12): a known subset of τ-neighbours V in arbitrary position, all other parameters (sample sizes, covariances, ...) being fixed. We additionally assume that all task covariance matrices are proportional to each other ("aligned"), which appears to be the least favourable setting.

Definition V.17. Let $\tau \in \mathbb{R}_+; B, V \in \mathbb{N}_{>0}$ with $B \geq V$, $\boldsymbol{s^2} = (s_1^2, \ldots, s_B^2) \in \mathbb{R}_+^B$, $(N_k)_{k\in[\![B]\!]} \in \mathbb{N}_{>0}^B$, and Σ a symmetric positive definite matrix of size d with $\operatorname{Tr}\Sigma = 1$ be fixed. We denote by $\mathcal{P}_{\text{single}}(\tau, V, \Sigma, \boldsymbol{s^2})$ the set of joint distributions for tasks following model (V.1) such that:

(i) The total number of bags is B and the number of samples per bag is given by $(N_k)_{k\in[\![B]\!]}$. (Omitted from the distribution class notation for simplicity.)

(ii) (**GS**) holds.

(iii) The task covariances are given by $\Sigma_k = N_k s_k^2 \Sigma$ (i.e., all tasks have covariances proportional to Σ and the naive risks are specified by the vector $\boldsymbol{s^2}$).

(iv) The mean vectors $(\mu_k)_{k\in[\![B]\!]}$ can vary freely subject to:

$$\|\mu_1 - \mu_k\|^2 \leq \tau s_1^2, k \in [\![V]\!].$$

A minimax lower bound, as by Proposition V.18 below, over that model holds over any larger model; for instance, the model where Σ_1 is arbitrarily fixed and the other covariances may vary freely provided that the naive risks still match the prescribed $\boldsymbol{s^2}$.

Proposition V.18. *It holds*

$$\inf_{\widehat{\mu}_1} \sup_{\mathbb{Q}\in\mathcal{P}_{\text{single}}(\tau,V,\Sigma,\boldsymbol{s^2})} \frac{R_1(\mathbb{Q},\widehat{\mu}_1)}{s_1^2} \geq \mathcal{B}\big(\tau,\nu([\![V]\!])\big) - \varepsilon\big(d^{\text{e}}(\Sigma)\big),$$

where $\mathcal{B}$ is defined in (V.12), ν in (V.9), the infimum is over all estimators $\widehat{\mu}_1$ for μ_1, and $R_1(\mathbb{Q},\widehat{\mu}_1)$ indicates its risk (V.2) under distribution $\mathbb{Q}$. The function $\varepsilon(t)$ is independent of any parameters *and satisfies $\varepsilon(t) = O((\log t)/t)$ as $t \to \infty$.*

This minimax lower bound can be compared with the upper bounds obtained for the testing and Q-aggregation methods, Proposition V.11 and Corollary V.13, resp. In the case of (**ECSS**) (so that $V_{\tau,\varsigma} = V_\tau$ for any $\varsigma \geq 1$ and we can ignore the role of ς), the lower and upper bounds match. This shows that the oracle relative risk $\mathcal{B}(\tau,\nu(V_\tau))$ is indeed minimax in the sense of high-dimensional asymptotics, provided that $\log(B) = o(d_1^{\text{e}})$. Furthermore, the Q-aggregation method is *asymptotically minimax adaptive* over the parameter $\tau > 0$. This can be seen as a generalization of classical results on the James-Stein estimator (see Section I.1.1). Observe also that for the upper and lower bounds the dimension-dependent remainder terms do not depend on other parameters, which makes the dimensional asymptotics uniform with respect to those parameters.

If (**ECSS**) does not hold, there can be a discrepancy between the minimax lower bound and the obtained upper bounds due to the exclusion of high variance tasks in the latter (V_τ against $V_{\tau,\varsigma}$). An unfavourable regime illustrating this gap is the following: suppose there are many tasks that are τ-neighbours of the target (τ being fixed independently of the dimension but $V \approx d_1^\bullet$) with significantly higher variances though ($\varsigma = s_k^2/s_1^2 \approx V^{1/2}$ for all $2 \leq k \leq V$). In that scenario, the upper bounds of Proposition V.11 and Corollary V.13 do not guarantee convergence to $\mathcal{B}(\tau,\nu(V_\tau)) \approx \tau/(1+\tau)$, since the remainder terms $\varsigma/\sqrt{d_1^\bullet}$ (resp. $\varsigma/\sqrt{d_1^{\text{e}}}$) do not converge to zero for high-dimensional asymptotics. This gap can amount to an arbitrary large factor since τ can be arbitrarily small. However, the scenario where a target task is surrounded by numerous neighbours with significantly higher variance can only arise for a small proportion of the tasks. This implies that this concern is alleviated when evaluating the relative risk averaged across all tasks, as shown next.

V.5.2 Compound relative risk

We define the compound relative risk as the relative risk averaged over all tasks. As we only studied upper bounds for a single task so far, we first derive new upper bounds for the compound relative risk. We then proceed to derive minimax bounds on restrictive distribution classes under which the task means exhibit a certain clustering or covering structure.

Definition V.19. Let $\boldsymbol{\mu} = (\mu_k)_{k\in[\![B]\!]}$ be a collection of vectors of $\mathbb{R}^d$, $J \in \mathbb{N}_{>0}$, and $\boldsymbol{\mathcal{C}}$ a J-partition of $[\![B]\!]$ (i.e., $\boldsymbol{\mathcal{C}} = (\mathcal{C}_j)_{j\in[\![J]\!]}$ with $\mathcal{C}_1 \sqcup \ldots \sqcup \mathcal{C}_J = [\![B]\!]$). The diameters of the partition $\boldsymbol{\mathcal{C}}$ applied to $\boldsymbol{\mu}$ are defined as:

$$\texttt{diam}(\boldsymbol{\mathcal{C}}, \boldsymbol{\mu}) = \left(\max_{k,\ell\in\mathcal{C}_j} \|\mu_k - \mu_\ell\|\right)_{j\in[\![J]\!]} \in \mathbb{R}_+^J. \tag{V.40}$$

We shall refer to parts as "groups" rather than clusters, because the partitioning can in principle be arbitrary. However, the intuition is that the set of vectors $\boldsymbol{\mu}$ exhibits more structure if it can be partitioned into a limited number of groups with small diameter. For instance, if it is strongly clustered, or supported on a set of small metric entropy such as a low-dimensional manifold. The compound relative risk of the Q-aggregation approach can then be upper bounded as follows:

Proposition V.20. *Assume (**GS**) holds, and let $u_0 \in \mathbb{R}_+$ such that $\log 17B \le u_0 \le (\min_k N_k - 1)/2$. For $k \in [\![B]\!]$, define $\widehat{L}_k(\boldsymbol{\omega}), \widehat{Q}_k(\boldsymbol{\omega})$ analogously to (V.29),(V.31) and*

$$\widehat{\boldsymbol{\omega}}_k \in \arg\min_{\boldsymbol{\omega}\in\mathcal{S}_B} \left(\widehat{L}_k(\boldsymbol{\omega}) + 16\sqrt{u_0}\,\widehat{Q}_k(\boldsymbol{\omega})\right). \tag{V.41}$$

Then it holds:

$$\frac{1}{B}\sum_{k=1}^{B} \frac{R_k(\widehat{\boldsymbol{\omega}}_k)}{s_k^2} \le \left(1 + CBe^{-u_0/2}\right) \min_{\boldsymbol{\mathcal{C}}} \left(\mathcal{L}^*\big(\boldsymbol{s}, \boldsymbol{\mathcal{C}}, \texttt{diam}(\boldsymbol{\mathcal{C}}, \boldsymbol{\mu})\big) + C\frac{u_0}{\min_{k\in[\![B]\!]} (d_k^{\mathrm{e}})^{1/2}}\right), \tag{V.42}$$

where the minimum is taken over all partitions $\boldsymbol{\mathcal{C}}$ of $[\![B]\!]$, C is an absolute constant, and for $\boldsymbol{\zeta} \in \mathbb{R}_+^J$:

$$\mathcal{L}^*(\boldsymbol{s}, \boldsymbol{\mathcal{C}}, \boldsymbol{\zeta}) := \frac{1}{B}\sum_{j=1}^{J}\sum_{k\in C_j} \mathcal{B}(\tau_{j,k}, \nu_{j,k}), \qquad \tau_{j,k} := \frac{\zeta_j^2}{s_k^2}, \quad \nu_{j,k} := \frac{s^2(\mathcal{C}_j)}{s_k^2}, \tag{V.43}$$

and $\mathcal{B}$ is defined in (V.12).

Similarly to the estimation of a single mean, the bound on the compound relative risk depends on the maximum distance between tasks of the same group relative to the naive risk of each task, and on the relative aggregated variances (V.9) in each group. Remarkably, the compound relative risk bound does not involve any "whittling down" of high-variance tasks as in the single task bound (V.22), and holds under arbitrary inhomogeneity of the tasks and sample sizes.

The quantity $\mathcal{L}^*$ equates to an oracle compound relative risk and is minimax under high-dimensional asymptotics. To show this, we extend the single task model V.21 to a joint distribution class such that the tasks are divided into inhomogeneous groups.

Definition V.21. Let $B \in \mathbb{N}_{>0}, \boldsymbol{s^2} = (s_1^2, \ldots, s_B^2) \in \mathbb{R}_+^B$, $(N_k)_{k\in[\![B]\!]} \in \mathbb{N}_{>0}^B$, and Σ a symmetric positive definite matrix of size d with $\operatorname{Tr}\Sigma = 1$ be fixed.
Let $J \in \mathbb{N}_{>0}$, $\boldsymbol{\mathcal{C}}$ be a J-partition of $[\![B]\!]$ and $\boldsymbol{\zeta} \in \mathbb{R}_+^J$. We define $\mathcal{P}_{\mathrm{mult}}(\boldsymbol{\mathcal{C}}, \boldsymbol{\zeta}, \Sigma, \boldsymbol{s})$ as the set of tasks according to model (V.1) with:

(i)-(iii) as in Definition V.17;

(iv) The mean vectors $\boldsymbol{\mu} = (\mu_k)_{k\in[\![B]\!]}$ can vary freely subject to

$$\boldsymbol{\mu} \in \left\{\boldsymbol{\mu} \in \mathbb{R}^{d\times B} : \texttt{diam}(\boldsymbol{\mathcal{C}}, \boldsymbol{\mu}) \le \boldsymbol{\zeta} \text{ (coordinate-wise inequality) } \right\}.$$

In words, $\mathcal{P}_{\mathrm{mult}}(\boldsymbol{\mathcal{C}}, \boldsymbol{\zeta}, \Sigma, \boldsymbol{s})$ is the set of Gaussian tasks with fixed, aligned covariances and naive risks prescribed by the vector $\boldsymbol{s}$, such that the groups of mean vectors given by partition $\boldsymbol{\mathcal{C}}$ have diameters bounded by the respective entries of vector $\boldsymbol{\zeta}$.

Proposition V.22. *Let $\boldsymbol{s} \in \mathbb{R}_+^B$, $J \in \mathbb{N}_{>0}$, $\boldsymbol{\mathcal{C}}$ a J-partition of $[\![B]\!]$ and $\boldsymbol{\zeta} \in \mathbb{R}_+^J$ be fixed. It holds*

$$\lim_{d^{\mathrm{e}} \to \infty} \sup_{\substack{\Sigma: \\ d^{\mathrm{e}}(\Sigma) = d^{\mathrm{e}}}} \inf_{\widehat{\boldsymbol{\mu}}} \sup_{\mathbb{Q} \in \mathcal{P}_{\mathrm{mult}}(\boldsymbol{\mathcal{C}}, \boldsymbol{\zeta}, \Sigma, \boldsymbol{s})} \frac{1}{B} \sum_{k=1}^{B} \frac{R_k(\mathbb{Q}, \widehat{\mu}_k)}{s_k^2} \geq \mathcal{L}^*(\boldsymbol{s}, \boldsymbol{\mathcal{C}}, \boldsymbol{\zeta}/2), \tag{V.44}$$

where the infimum is over all joint estimators $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_B)$.

In particular, since it holds $\mathcal{L}^*(\boldsymbol{s}, \boldsymbol{\mathcal{C}}, \boldsymbol{\zeta}/2) \geq \mathcal{L}^*(\boldsymbol{s}, \boldsymbol{\mathcal{C}}, \boldsymbol{\zeta})/4$, the upper bound matches the lower minimax bound up to a fixed constant factor in a dimensional asymptotics sense (by choosing $u_0 = \log 17B$ and provided that $\log B/(\min_k(d_k^{\mathrm{e}})^{-1/2}) = o(\mathcal{L}^*)$). Moreover, (V.42) shows that the Q-aggregation estimator is (up to that constant factor) asymptotically minimax adaptive with respect to the choice of grouping $\boldsymbol{\mathcal{C}}$ of the task means, the corresponding group diameters, and the bag variances.

As in the single task case, the minimax bound $\mathcal{L}^*$ only depends on the bag sizes through the naive risks $\boldsymbol{s}$: bags with large variance and many samples are statistically equivalent to bags with low variance and few samples. Similarly, the improvement only depends on the relative aggregated variance of each group, not on the number of bags. Proposition V.23 gives an interpretable upper bound for $\mathcal{L}^*$:

Proposition V.23. *Let $\boldsymbol{s} \in \mathbb{R}_+^B$, $J \in \mathbb{N}_{>0}$, $\boldsymbol{\mathcal{C}}$ a J-partition of $[\![B]\!]$ and $\boldsymbol{\zeta} \in \mathbb{R}_+^J$, it holds:*

$$\mathcal{L}^*(\boldsymbol{s}, \boldsymbol{\mathcal{C}}, \boldsymbol{\zeta}) \leq \sum_{j=1}^{J} \frac{|\mathcal{C}_j|}{B} \cdot \frac{\bar{\tau}_j + |\mathcal{C}_j|^{-1}}{\bar{\tau}_j + 1}, \quad \bar{\tau}_j := \frac{\zeta_j^2}{\bar{s}^2(\mathcal{C}_j)}, \quad \bar{s}^2(\mathcal{C}_j) := \left(\frac{1}{|\mathcal{C}_j|} \sum_{k \in \mathcal{C}_j} s_k^{-2} \right)^{-1}, \tag{V.45}$$

implying in particular:

$$\mathcal{L}^*(\boldsymbol{s}, \boldsymbol{\mathcal{C}}, \boldsymbol{\zeta}) \leq \min\left(1, \frac{\bar{\tau}_*}{1 + \bar{\tau}_*} + \frac{J}{B}\right), \qquad \bar{\tau}_* := \sum_{j=1}^{J} \frac{|\mathcal{C}_j|}{B} \bar{\tau}_j. \tag{V.46}$$

If all risks and diameters are equal, $s_k^2 = s^2$ and $\zeta_j^2 = \zeta^2$ for all $k \in [\![B]\!]$ and $j \in [\![J]\!]$, then the bound of (V.46) is sharp up to a factor at most 2.7.

Bound (V.46) elucidates that the compound oracle relative risk $\mathcal{L}^*$ is small when (i) there are few groups relative to the number of bags (i.e., J/B small); and (ii) groups have on average a small squared diameter relative to the harmonic mean of the naive risks of its constituent tasks.

Eq. (V.42) implies that the compound risk is upper bounded by $\mathcal{L}^*$ for any valid partitioning. As an illustrative example we consider the **(ECSS)** setting and $\boldsymbol{\mathcal{C}}$ as a $\sqrt{\tau}s$-covering of $\boldsymbol{\mu}$ for a given τ. Then $\bar{\tau}_* = \tau$ and the number of groups J is the covering number $N(\boldsymbol{\mu}, \sqrt{\tau}s)$. This highlights that the Q-aggregation strategy will be very effective to reduce the compound risk if the set of true means can be covered by a relatively small number of balls, in comparison to the total number of tasks, with a radius significantly smaller than the standard deviation of the naive estimates.

This bound takes a form akin to the findings presented in Section III (Theorem III.1 and III.2)where only the **(ECSS)** setting is examined and where is used a testing strategy comparable to that of the previous section. The parameter of these tests, though, has to be fixed by the user. In contrast, the Q-aggregation approach attains the oracle trade-off between the "bias" term $\tau/(1+\tau)$ and the "variance" term $N(\boldsymbol{\mu}, \sqrt{\tau}s)/B$ without the need to specify τ.

Finally, observe that the first term $\tau/(1+\tau)$ resembles the best potential improvement and is reminiscent of the oracle improvement factor of the James-Stein estimator, which can be conceived as a special case; see Section I.1.1 for additional details.

V.6 Application: estimation of multiple Kernel Mean Embeddings

We emphasise that our discussion and theoretical results include the case when $\mathcal{X}$ is a reproducing kernel Hilbert space (RKHS), in which case the mean corresponds to a kernel mean embedding (KME) (Muandet et al., 2017; Smola et al., 2007). Let $\mathcal{Z}$ be a measurable space enriched with a reproducing kernel $\kappa : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ and its corresponding RKHS $\mathcal{H}$. The kernel mean embedding $\mu_{\mathbb{P}_Z} \in \mathcal{H}$ of distribution $\mathbb{P}_Z$ on $\mathcal{Z}$ and its empirical (naive) estimation $\widehat{\mu}_{\mathbb{P}_Z}$, which is based on the samples $(Z_n)_{1 \leq n \leq N_Z} \sim \mathbb{P}_Z$, are defined as

$$\mu_{\mathbb{P}_Z} = \int_{\mathcal{Z}} \kappa(z, \cdot) \, \mathrm{d}\mathbb{P}_Z(z) \ , \ \widehat{\mu}_{\mathbb{P}_Z} = \frac{1}{N_Z} \sum_{n=1}^{N_Z} \kappa(Z_n, \cdot). \tag{V.47}$$

The estimation of multiple KMEs is an instance of model (V.1) once we identify $\mathcal{X} = \mathcal{H}$ and $X_k^{(i)} = \kappa(Z_k^{(i)}, \cdot)$ for a bounded reproducing kernel κ; this allows a direct application of our theoretical results for the bounded setting.

For characteristic kernels the map from $\mathbb{P}$ to $\mu_{\mathbb{P}}$ is injective and contains information about all moments of $\mathbb{P}$, so that $\mu_{\mathbb{P}}$ provides a unique representation of $\mathbb{P}$. Thus, KMEs can naturally be used to define a metric on probability distributions. Let $\mathbb{P}, \mathbb{Q}$ denote distributions and their KMEs $\mu_{\mathbb{P}}, \mu_{\mathbb{Q}}$ respectively. The maximum mean discrepancy (MMD) expresses the distance between $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ in $\mathcal{H}$

$$\begin{aligned} \mathrm{MMD}^2(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}^2 \ , \\ \widehat{\mathrm{MMD}}^2(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}}) &= \sum_{n \neq n'=1}^{N} \frac{\kappa(Z_n, Z_{n'})}{N(N-1)} - 2 \sum_{n=1}^{N} \sum_{m=1}^{M} \frac{\kappa(Z_n, Y_m)}{NM} + \sum_{m \neq m'}^{M} \frac{\kappa(Y_m, Y_{m'})}{M(M-1)}, \end{aligned}$$

where $\widehat{\mathrm{MMD}}^2$ denotes an unbiased estimate based on the samples $(Z_n)_{1 \leq n \leq N} \sim \mathbb{P}$ and $(Y_m)_{1 \leq m \leq M} \sim \mathbb{Q}$. For characteristic kernels it holds that $\mathrm{MMD}^2(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}}) = 0$ iff $\mathbb{P} = \mathbb{Q}$ (Gretton et al., 2012), which enables a large range of possible applications.

V.6.1 Motivation and Related Work

KMEs are employed for a variety of statistical tests, e.g., two-sample tests (Gretton et al., 2012), goodness-of-fit tests (Chwialkowski et al., 2016), and tests on statistical independence based on the Hilbert Schmidt independence criterion (Gretton et al., 2007). It also finds application in machine learning, e.g., for unsupervised (Jegelka et al., 2009) or supervised distributional learning (Muandet et al., 2012; Szabó et al., 2016), density estimation (Muandet et al., 2014), as part of the optimization criterion of the learning (Brehmer and Cranmer, 2020; Fakoor et al., 2020), and so on. Due to the wide variety of kernel functions, kernel mean embeddings can in general be used on various data types and for structured data. See Muandet et al. (2017) for an in-depth overview on KMEs and their applications.

The success of applying the KME or the MMD resp. relies heavily on the ability to accurately estimate the kernel mean based on sample data. The naive empirical estimator (V.47) was recently superseded by a James-Stein-like estimator (Muandet et al., 2014). They showed that this estimator is admissible and consistent for a suitable choice of shrinkage. Other single KME estimation strategies were proposed since then, e.g., non-linear shrinkage (Muandet et al., 2016), an empirical Bayesian approach (Filippi et al., 2016), and more robust estimations based on marginalised corrupted data (Xia et al., 2022), or a MOM approach (Lerasle et al., 2019). To the best of our knowledge, there is no prior work on the improved estimation of *multiple* kernel mean embeddings except for Marienwald et al. (2021).

V.6.2 Description of the Experiments

We evaluate the estimation of multiple kernel mean embeddings on artificial and real-world data.

Methods: We only sketch the best performing methods here. A complete list and detailed description of the tested methods can be found in Suection V.11, where we also provide pseudocode that demonstrates how the methods can be implemented in practice. More specifically, we found that methods based on Q-aggregation benefit from restricting the support of the weights from $\boldsymbol{\omega} \in \mathcal{S}_B$ to $\boldsymbol{\omega} \in \mathcal{S}_V$. However, the neighbouring test merely functions as a safeguard here with a much larger value for τ (cf. (V.8)) than that used for the testing approaches. We referred to methods, that are based on the testing procedure which finds neighbours for the construction of the convex combination as in Cor. V.9, as similarity test-based (`STB`). The approaches differ in their weighting schemes for these neighbours. `STB opt` calculates the oracle weights (V.13) where the aggregated variances are replaced by their empirical estimations. `STB orth` performs constrained risk minimization and posits an orthogonality assumption, $\langle \widehat{\mu}_j^{\mathtt{NE}} - \widehat{\mu}_i^{\mathtt{NE}}, \widehat{\mu}_{j'}^{\mathtt{NE}} - \widehat{\mu}_i^{\mathtt{NE}} \rangle = 0$ for all $j \neq j'$, which might be unrealistic in practice but yields a closed-form solution for the weights. Finally, `STB egd` minimises the Q-aggregation objective (V.37) and applies exponentiated gradient descent on the simplex (Kivinen and Warmuth, 1997) to approximate the solution.

We compare their performances to the naive estimation (`NE`), and we modify the multitask-averaging approach from Feldman et al. (2014) (`MTA const`) so that it is applicable to the estimation of KMEs. It assumes a constant similarity across tasks. Some more results of our previously proposed approach, `STB weight` (Section III) which was not designed to handle inhomogenous data, and of the regularised kernel mean shrinkage estimator `R-KMSE`, proposed in Muandet et al. (2016), that shrinks the estimation towards the origin and is performed separately on each bag can be found in Blanchard et al. (2024) with a discussion on the computational complexity of all approaches and on the choice of default parameter which have been found by cross-validation.

Experimental Metric: In the kernel case, the true KME μ is unknown even for synthesised data. We use a (naive) estimation based on an independent sample of the same distribution as approximation. Because this proxy is computed on a very large sample, it can be assumed to have low risk and to be more accurate than the estimation performed by any method on much smaller bags. The squared MMD between the (proxy) true KME μ_i of bag $i \in [\![B]\!]$ and its estimation $\widehat{\mu}_i^{\mathtt{m}}$, of form (V.6), performed by method $\mathtt{m}$ with weights $\omega_{i\cdot}^{\mathtt{m}}$ is then used as error measure

$$\widehat{\mathrm{MMD}}^2(\mu_i, \widehat{\mu}_i^{\mathtt{m}}) = \sum_{j,j' \in [\![B]\!]} \omega_{ij}^{\mathtt{m}} \omega_{ij'}^{\mathtt{m}} \sum_{n=1}^{N_j} \sum_{n'=1}^{N_{j'}} \frac{\kappa(Z_n^{(j)}, Z_{n'}^{(j')})}{N_j N_{j'}} - \sum_{j \in [\![B]\!]} 2\,\omega_{ij}^{\mathtt{m}} \sum_{n=1}^{N_j} \sum_{m=1}^{M_i} \frac{\kappa(Z_n^{(j)}, Y_m^{(i)})}{N_j M_i} + \sum_{m \neq m'}^{M_i} \frac{\kappa(Y_m^{(i)}, Y_{m'}^{(i)})}{M_i(M_i - 1)}, \tag{V.48}$$

where $Y_i, Z_i \sim \mathbb{P}_i$ independent with $|Y_i| = M_i \gg N_i = |Z_i|$ for all $i \in [\![B]\!]$, so that Y_i can be used to calculate the proxy and Z_i for the estimation. Each method is validated on the same data to guarantee comparability. This estimation error is averaged over multiple trials $\overline{\mathrm{MMD}}^2(\mu_i, \widehat{\mu}_i^{\mathtt{m}})$ and its decrease compared to the naive estimation $\widehat{\mu}^{\mathtt{NE}}$ is reported for all experiments

$$\Big(\big(\overline{\mathrm{MMD}}^2(\mu_i, \widehat{\mu}_i^{\mathtt{NE}}) - \overline{\mathrm{MMD}}^2(\mu_i, \widehat{\mu}_i^{\mathtt{m}}) \big) \,/\, \overline{\mathrm{MMD}}^2(\mu_i, \widehat{\mu}_i^{\mathtt{NE}}) \Big) \cdot 100 \;\; [\%].$$

Artificial Gaussian Data: The toy data sets are Gaussian distributed in $\mathbb{R}^2$ with fixed means and randomly rotated covariance matrices. For $i \in [\![B]\!]$ and $B = 50$

$$Z_\bullet^{(i)}, Y_\bullet^{(i)} \sim \mathcal{N}\Big(m_i, R(\theta_i) \Sigma R(\theta_i)^T\Big) = \mathbb{P}_i\,, \quad \theta_i \sim \mathcal{U}\Big(-\frac{\pi}{4}, \frac{\pi}{4}\Big),$$

where the rotation matrix $R(\theta_i)$ rotates the matrix $\Sigma = \mathrm{diag}(1, 10)$ according to angle θ_i. We generate $|Y_\bullet^{(i)}| = 1000$ data for the "proxy truth". A Gaussian RBF kernel, with a kernel width set to the average

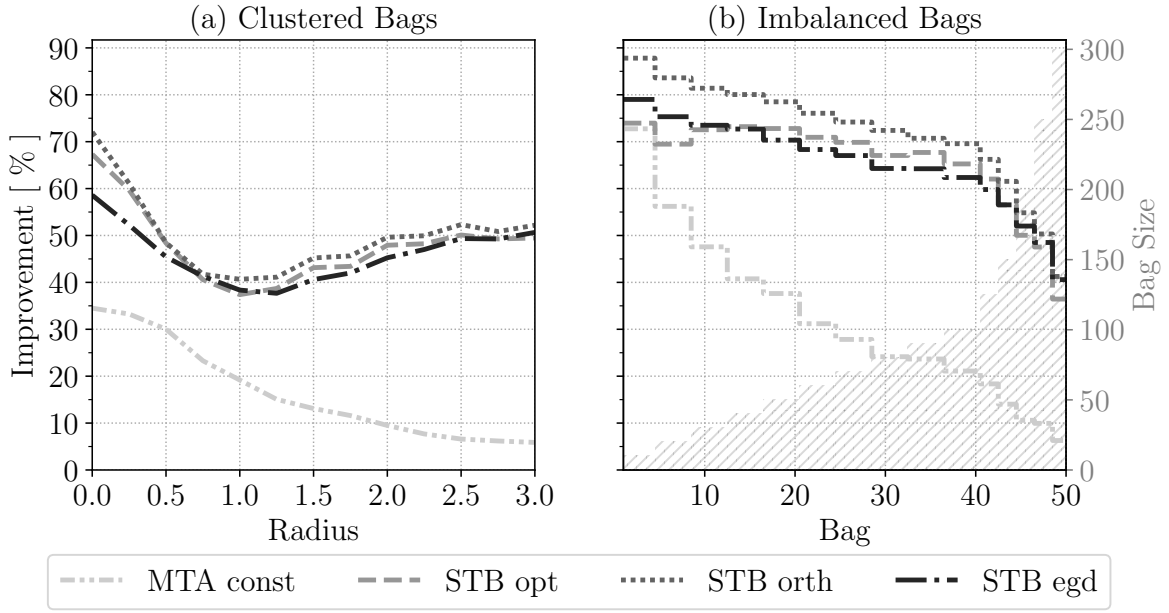

Figure 2: Decrease in average quadratic estimation error compared to `NE` in percent on Gaussian data settings (a) and (b) resp. Higher is better. The hashed histogram bars in (b) show the bag sizes for the bags 1 to 50, which vary between 10 and 300 (right axis).

feature-wise standard deviation of the data, maps the data from the two-dimensional input space to the infinite dimensional RKHS. Two setups are tested:

(a) Clustered Bags: $N_i = 50$ for all $i \in [\![B]\!]$. In the input space, each ten bags form a cluster where the cluster centres ($= m_i$) lie equally spaced on a circle. The radius of that circle varies between 0 and 3, which creates different amount of overlap between the clusters.

(b) Imbalanced Bags: $m_i = 0$ for all $i \in [\![B]\!]$. The bags $Z_\bullet^{(i)}$ are highly imbalanced, i.e. $N_i \in [10, 300]$. Because the tasks only vary in the rotation of their covariance matrices, we know that their KMEs lie on a low dimensional manifold in the RKHS. Because of the different bag sizes, the individual KMEs have different estimation accuracies.

The experiments are repeated for 100 trials; the results of the methods with default parameter choices are shown in Fig. 2.

All methods provide an improvement over `NE`, which is most significant for bags with few samples. This was already observed in other multi-task learning problems, e.g., see Marienwald et al. (2021) or Feldman et al. (2014). The constant similarity assumption of `MTA const` leads to an inadequate estimation for large radii or large bags. Namely, a KME with large bag size is shrunk to the grand empirical mean of all bags even though it includes high-variance (low sample size) or distant bags. This impairs the improvement. This effect is alleviated by the proposed `STB` approaches, that define the shrinkage according to the variance of and the distances between the KMEs. They show high performance for the tested settings. For $0.5 <$ radius < 2, the similarity test might mistake a bag of another cluster for a neighbour due to the strong overlap between the clusters, which explains the slight performance dip. All the proposed methods provide similarly accurate results. Despite its unrealistic orthogonality assumption, `STB orth` performs best on the artificial data.

Flow Cytometry Data: Flow cytometry is fundamental to biomedical research and clinical practice. It provides a multiparametric, single-cell analysis of a suspension or sample. The flow cytometer analyses

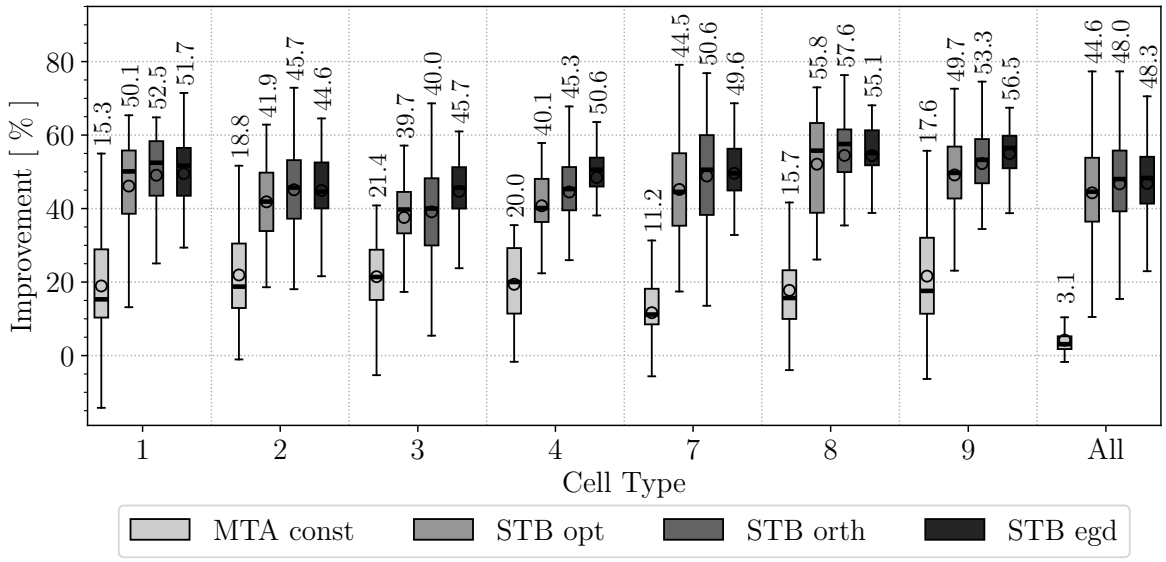

Figure 3: Decrease in estimation error compared to `NE` in percent on the flow cytometry data. Higher is better. The number next to the boxplot quantifies the median, which is also depicted as a line. The mean is visualised as a circle. From left to right: results on individual cell types $1, 2, 3, 4, 7, 8, 9$, and all cell types taken jointly.

the size, shape and internal complexity of cells and can detect the presence and amount of different fluorochromes (which in turn reveal insights about the presence of proteins or structures within the cell). These characteristics might then be used to classify the cells into different populations. Applications are vast, but well-known examples are differential blood count, or immunophenotyping of leukemia or in HIV infections (Adan et al., 2017; McKinnon, 2018).

The data set we use corresponds to the T-cell panel of the Human ImmunoPhenotyping Consortium (Finak et al., 2016). Seven laboratories were asked to perform a flow cytometry analysis of three replicates of blood samples of three patients. All laboratories were asked to follow the same experimental protocol and used the same seven markers to characterise the cells ($d = 7$). Based on the observed characteristics the cells were then classified into ten different populations or cell types. We use this structure (laboratory, replicate, patient, cell type) to divide the data into bags. We excluded bags with less than 1000 data points, which leads to 424 bags in total. Each data point $Z_n^{(i)} \in \mathbb{R}^7$ in a bag i corresponds to one cell. As the number of cells varies, the bags are highly imbalanced. We use a Gaussian RBF kernel with kernel width of 950 to map the cell features to a RKHS. The kernel choice and width are in accordance with Dussap et al. (2023). The (proxy) true KME is approximated by a naive estimation based on $Y_\bullet^{(i)}$ with $|Y_\bullet^{(i)}| = 1000$ (bags with more samples are capped). The sizes of the bags that are used for the estimation are chosen proportional to the bag sizes of the original input data, $N_i \in [7, 125]$, to mimic a realistic setting. In each one of the 100 trials, a subset of samples $Z_\bullet^{(i)}$ with $|Z_\bullet^{(i)}| = N_i$ is drawn randomly from $Y_\bullet^{(i)}$, on which the methods perform their estimation. We conducted experiments on each cell type separately so that $B \in [43, 62]$, and on all cell types jointly ($B = 424$). Cell types $5, 6$ and 10, for which $B < 7$, are excluded for the separate but included in the joint analysis.

The results are depicted in Fig. 3. On average, all methods provide an improvement over `NE`. For some trials, `MTA const` gives worse estimations than `NE` (negative improvement), see e.g., cell type 1. When all cell types are considered jointly, its performance drops significantly. The `STB` approaches give more accurate estimations than `MTA const` and provide an improvement of $\approx 50\%$ for all cell types. `STB egd` gives the most accurate and stable estimations across the different settings but also has high computational

complexity.

In summary, our presented methods provide an improvement over the naive estimation and over other state-of-the-art methods. Although `R-KMSE` or `MTA const` give more accurate estimations than the sample average, the provided improvements vary whether a shrinkage towards a common reference point or the grand mean resp. complies with the underlying data. In contrast, our proposed methods identify inhomogeneous task similarities and are applicable to imbalanced data sets (which, therefore, surpass our previously introduced method `STB weight`). While `STB egd` provides in most cases the highest improvement with least variance, it also requires the most computational complexity. `STB orth` provides a good trade-off.

V.7 Relation and comparison to previous work

We review related literature grouped along two axes: the first is rooted in statistics, compound decision rules and the empirical Bayes point of view, and secondly a more recent one related to multitask learning. We first emphasise again the seminal importance of the James-Stein (JS) estimator (James and Stein, 1961) for a single vector mean, which can be seen as a particular setting of model (V.1). Historically important is the realization that the sample average $\widehat{\mu}_i^{\texttt{NE}} := \frac{1}{N_i}\sum_{k=1}^{N_i} X_k^{(i)}$, despite being MLE (in the Gaussian model) and BLUE, is inadmissible and dominated by the shrinkage-based JS estimator. Pinsker (1980) should be credited for an early "dimensional asymptotics" point of view, analysing the minimax risk if the mean vector belongs to a ball of $\mathbb{R}^d$ as a by-product of his celebrated minimax analysis of estimators in Sobolev ball models (see, e.g., Nussbaum, 1996 for a discussion). The risk of the JS estimator is asymptotically close to that minimax in the isotropic Gaussian model if $d \to \infty$, as well as adaptive to the radius of the ball (Beran, 1996); see more details in Section I.1.1.

V.7.1 Empirical Bayes and compound decision point of view

The celebrated series of works by Efron and Morris (1972, 1973, 1976) advocated for an interpretation of the JS estimator as a compound decision problem and an empirical Bayes point of view (Robbins, 1951, 1964; Zhang, 2003): the problem of estimating a single mean vector in $\mathbb{R}^B$ with standard Gaussian noise is better seen as B-many estimations of one-dimensional means observed with independent observation noise (which in model (V.1) corresponds to $B > 1$ means in dimension $d = 1$). The authors compare the performance of the JS estimator to that of a Bayesian model, i.e., the means are themselves drawn from a centred Gaussian prior. The Bayes rule under the fully Gaussian model (prior and observations) is solely determined by the prior variance, which is usually unknown, hence, called "oracle" in the present discussion. The JS estimator can then be interpreted as being empirically Bayes as it replaces the oracle (prior) variance with an empirically estimated counterpart. The compound risk is shown to converge to the oracle Bayes risk, as B grows.

Efron and Morris (1976) generalised this analysis to the multidimensional case which is an instance of model (V.1) for arbitrary d and Gaussian task distributions with identical covariances. They proposed a multidimensional version of the JS estimator. Similarly to the one-dimensional case, this is interpreted as an empirical Bayes procedure with a multidimensional Gaussian prior, whose unknown covariance is replaced by an empirically estimated counterpart. If $(d+2)/B \to 0$, then the risk of the multidimensional JS estimator approaches that of the oracle Bayes rule.

The nonparametric empirical Bayes estimator developed by Brown and Greenshtein (2009) (see also Jiang and Zhang, 2009 for a closely related, independent work) is in the same line of thought, but considers a completely arbitrary prior on the means (in dimension $d = 1$). In that situation, the oracle Bayes procedure can be expressed in terms of the marginal, nonparametric mixture density of the observations across tasks and of its derivative (to establish this, Gaussian partial integration is used, thus, relying heavily on the assumption of isotropic Gaussian tasks). The proposed estimator replaces the true density with a kernel density estimate (while Jiang and Zhang, 2009 adopt a Generalised Maximum Likelihood Empirical Bayes

estimator to estimate the prior). For a Gaussian kernel and as $B \to \infty$, this estimator approaches the oracle Bayes rule.

Similar to our approach, George (1986) proposed a weighted combination of shrinkage estimators, e.g., multiple JS estimators. The weights are assumed to be known but can adapt to the data to some extent. He showed that an aggregation of Bayes rules is again Bayes on a mixture prior where the weights naturally translate to prior probabilities.

We emphasise the following key differences of this important line of work to the present one:

(a) The above approaches focus on the *compound* risk, while we analyse the risk of each individual task. The compound relative risk, analysed in Section V.5.2, is a different quantity from the ratio between compound risk and oracle Bayes risk.

(b) In the empirical Bayes framework, the focus lies on asymptotics as the number of independent tasks B grows large, while ours is on the growing (effective) dimension. Consequently, the choice of "oracle" reference for analyzing risk ratios differs between the two perspectives. Within the empirical Bayes paradigm, the compound oracle Bayes risk serves as the reference. We adopt a task-specific oracle improvement relative to the naive estimator. Thus, the theoretical outcomes derived from these divergent approaches are not readily comparable.

Concerning the role of the dimension, consistency with the oracle Bayes reference requires $d/B \to 0$ for the parametric approach of Efron and Morris (1976) and presumably an even more stringent condition for the nonparametric approaches of Brown and Greenshtein (2009) or Jiang and Zhang (2009). In fact they only considered the case $d = 1$, but since both works rely on metric entropy estimates on appropriate function spaces, one would expect those to suffer of the curse of dimensionality.

Consistency with the oracle, as considered here, requires roughly $\mathrm{polylog}(B)/d \to 0$, thereby accommodating a broader spectrum of regimes. For instance, when $B = \Theta(d^\alpha)$ for arbitrary $\alpha > 0$, our approach ensures consistency with our oracle improvement, yet fails to achieve consistency with the oracle Bayes with a Gaussian prior if $\alpha \leq 1$. Conversely, the regime where $B \to \infty$ while d remains fixed, which is pertinent to empirical Bayes analyses, does not yield meaningful results in our framework (though, allowing the dimension to increase at an arbitrary small power of B remains viable).

In summary, our perspective is tailored towards high-dimensional scenarios, with possibly non-isotropic covariance structures, whereas the empirical Bayes methodology is not inherently designed for such settings. Moreover, we emphasise the minimax property of our oracle improvement across suitable models as the dimension grows.

(c) We allow non-Gaussian data.

(d) We allow strong task heterogeneity (e.g., the covariances are not shared across tasks).

V.7.2 Multitask learning point of view

Feldman et al. (2014) viewed the many means estimation problem (V.1) as a multi-task learning problem (Caruana, 1997; Zhang and Yang, 2021), which gave rise to the term multi-task averaging. Also inspired by the JS estimator, the proposed approach extends the empirical compound risk minimization with a regularization term that favours the alignment of mean estimations for "related" tasks. The notion of "task relatedness" is encoded as a similarity matrix considered as *a priori* information. In absence of specific information, the similarities are taken constant across tasks and the method reduces to shrinkage towards the grand mean. The theoretical analysis focused mainly on the low-dimensional setting and the oracle weights when $B = 2$. Their data-driven similarity estimation yielded inconclusive results. Martínez-Rego and Pontil (2013) mitigated the default constant similarities in the absence of information by first clustering

the tasks into different groups and then applying the approach of Feldman et al. (2014) on each cluster separately; but a theoretical analysis of this approach was not conducted. In our work, we also propose to assimilate estimators of related tasks and thereby define an appropriate shrinkage direction. We eliminate the disadvantage of both approaches, i.e., constant or known similarities, by estimating them solely based on the available data. We also extend significantly our preliminary work (Section III) which was limited to the testing approach unfit for heterogeneous tasks, and with less precise theoretical results.

Recent work of Duan and Wang (2023) considers a general multi-task learning setting which includes the multiple mean estimation problem as a special case. Comparable to that of Feldman et al. (2014), their estimators are determined by compound empirical risk minimization with a regularization term measuring alignment to a predetermined model of task relatedness, e.g., the means form K clusters or are close to a linear subspace of dimension K. The proposed estimators depend on the considered task relatedness and on K. Once interpreted in terms of relative squared risk, the theoretical bounds obtained by Duan and Wang (2023) are not bounded by a constant but can grow as $\mathcal{O}(K^2)$ in the worst case where the fit to the posited task relatedness is poor. For the relative risk to be significantly less than $\mathcal{O}(1)$, the bounds require the condition $\delta \lesssim s_1/K$, where δ represents closeness to the model (cluster radius resp. distance to linear subspace). By contrast, in our analysis we do not posit a particular task relatedness or value of K to define the estimators; those are adaptive to the most advantageous grouping model, including cluster number and size, describing the structure of the true parameters (see Section V.5). Our relative risk bounds are worst-case bounded (and even bounded close to 1), and show a significant improvement in favourable cases even for the number of groups K growing with the number of tasks B. On the other hand, our approach won't result in a significant risk improvement if the task means belong to a low-dimensional subspace but are very far apart from each other. Still, using the covering complexity point of view discussed in Section V.5, an improvement can be shown if the tasks increase in number and are drawn, say, from a fixed a priori distribution having a low-dimensional support while the ambient dimension grows.

V.8 Conclusion

Considering the estimation of multiple mean vectors in high dimensions from independent samples, we focused on estimators formed as convex combinations of empirical averages of each sample. We proposed a test-then-aggregate method generalizing the approach of Marienwald et al. (2021), and a direct Q-aggregation approach where the weights are found by minimization of an adequate objective. From a theoretical perspective, we established asymptotic convergence to an oracle risk in an appropriate "dimensional asymptotics" sense, as the effective dimensionality grows. This oracle risk was proved to be exactly minimax under certain homogeneity conditions for the single-task risk, and minimax up to a fixed factor for the compound relative risk (without homogeneity conditions). One advantage of the Q-aggregation method is its theoretical adaptivity with respect to parameters that have to be user-provided for the testing approach. We demonstrated the efficacy of the proposed methods on showcase experiments for estimating multiple kernel mean embeddings on controlled artificial datasets and real-world flow cytometry data.

Future investigations will aim to address the discrepancy between the lower and upper bounds for the single mean estimation in extremely inhomogeneous cases (we suspect the minimax lower bound could be too conservative in such a case because it does not take into account the problem of neighbour detection). Another important open direction is the integration in the multiple-mean estimation setting of recent advances on single-mean estimation in high dimension, achieving sub-Gaussian performance even under heavy-tailed distributions or samples that were adversarially corrupted, e.g., the median of means estimator (Lugosi and Mendelson, 2019b, 2020, see Fathi et al., 2020; Lugosi and Mendelson, 2019a for an overview), or efficiently computable estimators (e.g., Cheng et al., 2019; Depersin and Lecué, 2022). Finally, a significant future avenue is to extend our approach from mean estimation to more general high-dimensional multi-task learning problems such as those considered by Duan and Wang (2023).

V.9 Proofs for Section V

V.9.1 Nomenclature

B	number of tasks, Sec. V.1
$\mathcal{B}(\tau, \nu)$	oracle risk, (V.10)
(BS)	bounded assumption, Sec. V.2.3
$\boldsymbol{\mathcal{C}}$	J-partition of $[\![B]\!]$, Def. V.19
d	ambient dimension, Sec. V.1
$d_k^\bullet$	effective dimension, (V.5)
d_k^{e}	effective dimension, (V.5)
$\mathtt{diam}(\boldsymbol{\mathcal{C}}, \boldsymbol{\mu})$	diameter of partition $\boldsymbol{\mathcal{C}}$ of $\boldsymbol{\mu}$, (V.40)
Δ_k	difference between μ_k and μ_1, Sec. V.2.5
(ECSS)	equal covariances and sample sizes, Sec. V.2.3
η	relative estimation error of s_k^2, (V.14)
(GS)	Gaussian assumption, Sec. V.2.3
(HT)	heavy-tailed assumption, Sup. V.9.4
J	nr. of parts of the partition, Def. V.19
k	index of task, Sec. V.1
(KC)	known covariances, Sec. V.2.3
$L_k(\widehat{\mu})$	loss of estimator $\widehat{\mu}$, (V.2)
$L_k(\boldsymbol{\omega})$	loss of aggregation estimator $\widehat{\mu}_{\boldsymbol{\omega}}$, (V.2.5)
$\widehat{L}_k(\boldsymbol{\omega})$	estimator for cond. risk, (V.29), Sec. V.1
$\mathcal{L}^*(\boldsymbol{s}, \boldsymbol{\mathcal{C}}, \boldsymbol{\zeta})$	compound oracle risk, (V.43)
M	radius of ball in which the bounded data lies, Sec. V.4.3
μ_k	expectation of distribution k
$\widehat{\mu}_k$	estimator of μ_k
$\widehat{\mu}_k^{\mathtt{NE}}$	naive estimation (empirical average) of μ_k, Sec. V.1
$\widehat{\mu}_{\boldsymbol{\omega}}$	aggregation estimator, (V.6)
$[\![n]\!]$	integers 1 to n, Sec. V.1
N_k	number of samples (bag size) of task k, Sec. V.1
$\|a\|$	canonical norm of vector a, Sec. V.2
$\|\Sigma\|_p$	Schatten norm of matrix Σ, Sec. V.2.2
$\|\Sigma\|_\infty$	operator norm of matrix Σ, Sec. V.2.2
$\nu(U)$	relative aggregated variance, (V.9)
$\boldsymbol{\omega}$	aggregation weights, (V.6)
$\mathbb{P}_k$	k-th task (probability distribution), Sec. V.1
$\mathcal{P}_{\mathrm{mult}}(\boldsymbol{\mathcal{C}}, \boldsymbol{\zeta}, \Sigma, \boldsymbol{s})$	class of distributions, Def. V.21
$\mathcal{P}_{\mathrm{single}}(\tau, V, \Sigma, \boldsymbol{s^2})$	class of distributions, Def. V.17
$\widehat{Q}_1(\boldsymbol{\omega})$	prob. upper bound on $(\widehat{L}_1(\boldsymbol{\omega}) - L_1(\boldsymbol{\omega}))$, (V.31)
$\widehat{Q}_1^{\mathbf{BS}}(\boldsymbol{\omega})$	additional penalization for **(BS)**, (V.38)
$R_k(\widehat{\mu})$	risk of estimator $\widehat{\mu}$, (V.2)
$R_k(\boldsymbol{\omega})$	risk of aggregation estimator $\widehat{\mu}_{\boldsymbol{\omega}}$, (V.2.5)
$s^2(U)$	harmonic mean of the risks of the tasks in U(V.9)
$\mathcal{S}_B$	$(B-1)$-dimensional simplex, Sec. V.2.5
s_k^2	naive risk, (V.3)
$\mathcal{S}_V$	set of convex weights of support incl. in V, Sec. V.2.5
ς	threshold for $W_{(\varsigma)}$, (V.21)
Σ_k	covariance matrix of k-th task, Sec. V.2
$\widetilde{T}_k^{(\tau)}$, $\widetilde{\widetilde{T}}_k^{(\tau)}$	empirical similarity test on independent copy data, (V.17)-(V.25)
$\tau, \tau_{\min}^k, \tau_{\min}^\circ, \tau^\pm$	thresholds for similarity test, (V.8)-(V.18)-(V.23)-(V.23)
$\tau/1+\tau$	best potential improvement
$\widetilde{U}_k$	unbiased estimator for $\|\Delta_k\|^2$, (V.16)

$\widetilde{V}$	estimation of V_τ, (V.14)
V_τ	τ-neighbouring tasks, (V.8)
$V_{\tau,\varsigma}$	trimmed V_τ, (V.22)
V^*	subset of $\widetilde{V}$, (V.14)
$W_{(\varsigma)}$	set of tasks with bounded variance, (V.21)
ϕ_k	ratio of the radius M with the covariance trace $\operatorname{Tr}\Sigma_k$,
$X_\bullet^{(k)}$	k-th bag, (V.1)
$\widetilde{X}_\bullet^{(k)}$	independent copy of k-th bag, Sec. V.3.2
ζ	bound on the diameter of the J-partition, Def. V.21

V.9.2 Proofs for Section V.3.1 and Section V.3.2

Proof of Lemma V.6 The weights $\boldsymbol{\omega}^*$ are obtained by minimizing the upper bound (V.11) using KKT conditions, for instance. However, to verify the bound (V.12), it suffices to substitute the weights (V.13) into (V.11). Let us denote $\nu = \frac{s^2(V)}{s_1^2}$, from (V.12):

$$\begin{aligned}\frac{R_1(\boldsymbol{\omega}^*)}{s_1^2} &\le \tau(1-\omega_1^*)^2 + \sum_{k\in V}(\omega_k^*)^2\frac{s_k^2}{s_1^2}\\ &= \tau\lambda^2(1-\nu)^2 + (1-\lambda)^2 + 2\lambda(1-\lambda)\nu + \lambda^2\nu.\end{aligned}$$

By substituting λ with its value from Equation (V.13), we obtain:

$$\begin{aligned}\frac{R_1(\boldsymbol{\omega}^*)}{s_1^2} &\le \frac{\tau(1-\nu)^2 + \tau^2(1-\nu)^2 + 2\tau(1-\nu)\nu + \nu}{(1+\tau(1-\nu))^2}\\ &= \frac{\tau(1-\nu)((1-\nu)+\tau(1-\nu)+\nu) + \nu(\tau(1-\nu)+1)}{(1+\tau(1-\nu))^2}\\ &= \frac{\tau(1-\nu)+\nu}{1+\tau(1-\nu)} = \mathcal{B}(\tau,\nu) = \mathcal{B}\left(\tau, \frac{s^2(V)}{s_1^2}\right).\end{aligned}$$

Thus, the inequality holds as claimed. □

Proof of Proposition V.7 Recall that we assume the following event holds:

$$\begin{cases} 1 \in V^* \subseteq \widetilde{V} \subseteq V_\tau, \\ \left|\widetilde{s}_k^2 - s_k^2\right| \le \eta s_k^2, \text{ for all } k \in \widetilde{V}, \end{cases} \tag{V.14}$$

for quantities $\widetilde{V}, \widetilde{s}$ which are considered as nonrandom for this proof (e.g., they are computed from an independent sample and we argue conditionally to that sample). Denote

$$\overline{R}_1(\widetilde{V}, \boldsymbol{\omega}) := \tau s_1^2(1-\omega_1)^2 + \sum_{k\in\widetilde{V}}\omega_k^2 s_k^2$$

the risk upper bound from (V.11) wherein we used the index set $\widetilde{V}$. Due to the first $\widetilde{V} \subset V_\tau$ the same argument leading up to (V.11), it holds $R_1(\boldsymbol{\omega}) \le \overline{R}_1(\widetilde{V}, \boldsymbol{\omega})$ for all $\boldsymbol{\omega} \in \mathcal{S}_{\widetilde{V}}$. Denoting now

$$\widetilde{R}_1(\widetilde{V}, \boldsymbol{\omega}) := \tau \widetilde{s}_1^2(1-\omega_1)^2 + \sum_{k\in\widetilde{V}}\omega_k^2 \widetilde{s}_k^2$$

the plug-in version of $\overline{R}_1(\widetilde{V}, \boldsymbol{\omega})$, we have, putting $\varepsilon_k := \left|s_k^2 - \widetilde{s}_k^2\right|$:

$$\forall \boldsymbol{\omega} \in \mathcal{S}_{\widetilde{V}} : \left|\overline{R}_1(\widetilde{V}, \boldsymbol{\omega}) - \widetilde{R}_1(\widetilde{V}, \boldsymbol{\omega})\right| \le \tau\varepsilon_1(1-\omega_1)^2 + \sum_{k\in\widetilde{V}}\omega_k^2\varepsilon_k \le \left(\max_{k\in\widetilde{V}}\frac{\varepsilon_k}{s_k^2}\right)\overline{R}_1(\widetilde{V}, \boldsymbol{\omega}),$$

which entails, from the second part of event (V.14):

$$(1-\eta)\overline{R}_1(\widetilde{V},\boldsymbol{\omega}) \le \widetilde{R}_1(\widetilde{V},\boldsymbol{\omega}) \le (1+\eta)\overline{R}_1(\widetilde{V},\boldsymbol{\omega})$$

Since $\widetilde{\boldsymbol{\omega}}$ is a minimiser of $\widetilde{R}_1(\widetilde{V},\boldsymbol{\omega})$, it holds for any other $\boldsymbol{\omega} \in \mathcal{S}_{\widetilde{V}}$:

$$R_1(\widetilde{\boldsymbol{\omega}}) \le \overline{R}_1(\widetilde{V},\widetilde{\boldsymbol{\omega}}) \le (1-\eta)^{-1}\widetilde{R}_1(\widetilde{V},\widetilde{\boldsymbol{\omega}}) \le (1-\eta)^{-1}\widetilde{R}_1(\widetilde{V},\boldsymbol{\omega}) \le \left(\frac{1+\eta}{1-\eta}\right)\overline{R}_1(\widetilde{V},\boldsymbol{\omega}).$$

Minimizing the latter inequality over $\boldsymbol{\omega}$ yields (from Lemma V.6):

$$\frac{R_1(\widetilde{\boldsymbol{\omega}})}{s_1^2} \le \left(\frac{1+\eta}{1-\eta}\right)\mathcal{B}\big(\tau,\nu(\widetilde{V})\big) \le \left(\frac{1+\eta}{1-\eta}\right)\mathcal{B}\big(\tau,\nu(V^*)\big),$$

due to $V^* \subseteq \widetilde{V}$ and the monotonicity properties of ν, $\mathcal{B}$. □

V.9.3 Proofs for Section V.3.3

We start with a generic result linking concentration of the test statistic to the properties of the associated test. It will allow to handle different distributional settings as particular cases.

We recall that $\widetilde{U}_k$ is the test U-statistic given by (V.16) using independent "tilde" data.

Assumption V.24 (**TSC**, Test Statistic Concentration)**.** *Assume that for all $k \in [\![B]\!]$ and $\alpha \in (0,1)$, there exists $q_k(\alpha)$:*

$$\mathbb{P}\Big[\Big|\widetilde{U}_k - \|\Delta_k\|^2\Big| \ge \|\Delta_k\| q_k(\alpha) + c_0^2 q_k^2(\alpha)\Big] \le \alpha\,. \tag{V.49}$$

where $c_0 \ge 2$ is a numerical constant.

Put $u_\alpha := \log(8/\alpha)$, it is established that:

- The assumption is satisfied under (**GS**) for $q_k^2(\alpha) = 2\Big(\frac{\|\Sigma_1\|_2}{N_1} + \frac{\|\Sigma_k\|_2}{N_k}\Big)u_\alpha$ and $c_0 = 4$. (Proposition IV.6)
- The assumption is satisfied under (**BS**) for $q_k^2(\alpha) = 16\Big(\frac{\|\Sigma_1\|_2}{N_1} + \frac{\|\Sigma_k\|_2}{N_k}\Big)u_\alpha + 4\frac{M^2u_\alpha^2}{N_1^2 \wedge N_k^2}$ and $c_0 = 31$. (Proposition IV.9)
- The assumption is satisfied under (**HT**) for $q_k^2(\alpha) = 16\Big(\frac{\|\Sigma_1\|_2}{N_1} + \frac{\|\Sigma_i\|_2}{N_i}\Big)u_\alpha$ and $c_0 = 2$ but for $\alpha \ge 8e^{-N_1 \wedge N_i}$. (Proposition V.33).

Proposition V.25. *Grant assumption (**TSC**) and let $\alpha \in (0,1)$, $\tau > 0$ be fixed. Let $\widetilde{T}_k$ be given by*

$$\widetilde{T}_k := \mathbf{1}\Big\{\widetilde{U}_k \le \tau s_1^2\Big\}. \tag{V.50}$$

Define $\tau_{\min}^k := 2c_0^2 s_1^{-2} q_k^2(\alpha)$, then it holds:

$$\text{if } \|\mu_1 - \mu_k\| > (\sqrt{\tau} + \sqrt{\tau_{\min}^k})s_1 : \qquad \mathbb{P}\Big[\widetilde{T}_k = 1\Big] \le \alpha; \tag{V.51}$$

$$\text{if } \|\mu_1 - \mu_k\| \le (\sqrt{\tau} - \sqrt{\tau_{\min}^k})s_1 : \qquad \mathbb{P}\Big[\widetilde{T}_k = 0\Big] \le \alpha. \tag{V.52}$$

Proof of Prop. V.25 We assume for the rest of the proof that

$$\Big|\widetilde{U}_k - \|\Delta_k\|^2\Big| \le \|\Delta_k\| q_k(\alpha) + c_0^2 q_k^2(\alpha)$$

holds, which according to Assumption (**TSC**) is the case with probability at least $1-\alpha$. Using $q_k^2(\alpha)s_1^{-2} = \tau_{\min}c_0^{-2}/2$ and putting $x := \frac{\|\Delta_k\|}{\sqrt{\tau}s_1}$, the above inequality entails

$$\left|\frac{\widetilde{U}_k}{\tau s_1^2} - x^2\right| \leq x\sqrt{\frac{\tau_{\min}}{2\tau}}c_0^{-1} + \frac{\tau_{\min}}{2\tau} \leq x\frac{\varepsilon_\tau}{2\sqrt{2}} + \frac{\varepsilon_\tau^2}{2}, \tag{V.53}$$

where we have used $c_0 \geq 2$ and where $\varepsilon_\tau := \sqrt{\tau_{\min}/\tau}$. This entails

$$\tau^{-1}s_1^{-2}(\widetilde{U}_k - \tau s_1^2) \leq x^2 + x\frac{\varepsilon_\tau}{2\sqrt{2}} + \frac{\varepsilon_\tau^2}{2} - 1.$$

Assuming $\varepsilon_\tau \leq 1$, the largest root of the quadratic polynomial on the right-hand-side above is lower bounded as

$$x_+ = -\frac{\varepsilon_\tau}{4\sqrt{2}} + \sqrt{1 - \frac{15}{32}\varepsilon_\tau^2} \geq 1 - \varepsilon_\tau,$$

using $\sqrt{1-a} \geq 1 - \sqrt{a}$ for $a \in [0,1]$. Thus, $0 \leq x \leq 1 - \varepsilon_\tau$ is a sufficient condition ensuring $\widetilde{T}_k = 1$, implying (V.52) since $(1-\varepsilon_\tau)^2\tau = \left(\sqrt{\tau} - \sqrt{\tau_{\min}}\right)^2$. (The case $\varepsilon_\tau > 1$ is trivial since the statement is void in that configuration.)

Similarly, (V.53) entails

$$\tau^{-1}s_1^{-2}(\widetilde{U}_k - \tau s_1^2) \geq x^2 - x\frac{\varepsilon_\tau}{2\sqrt{2}} - \frac{\varepsilon_\tau^2}{2} - 1;$$

the largest root of the quadratic polynomial on the right-hand-side above is upper bounded as

$$x'_+ = \frac{\varepsilon_\tau}{4\sqrt{2}} + \sqrt{1 + \frac{17}{32}\varepsilon_\tau^2} \leq 1 + \varepsilon_\tau,$$

using $\sqrt{1+a} \leq 1 + \sqrt{a}$. Thus, $x > 1 + \varepsilon_\tau$ is a sufficient condition ensuring $\widetilde{T}_k = 0$, implying (V.51) since $(1+\varepsilon_\tau)^2\tau = \left(\sqrt{\tau} + \sqrt{\tau_{\min}}\right)^2$.

Proof of Prop. V.8. Proposition IV.6 states that under (**GS**) it holds with probability at least $1-\alpha$ that

$$\left|\widetilde{U}_k - \|\Delta_k\|^2\right| \leq \|\Delta_k\| q'_k \sqrt{u_\alpha} + 16 q_k^2 u_\alpha, \tag{V.54}$$

where

$$q_k^2 = 2\left(\frac{\|\Sigma_1\|_2}{N_1} + \frac{\|\Sigma_k\|_2}{N_k}\right) = 2s_1^2\left(\frac{1}{\sqrt{d_1^\bullet}} + \frac{s_k^2/s_1^2}{\sqrt{d_k^\bullet}}\right),$$

and

$$(q'_k)^2 := 2\left(\frac{\|\Sigma_1\|_\infty}{N_1} + \frac{\|\Sigma_k\|_\infty}{N_k}\right);$$

since $\|\Sigma\|_\infty \leq \|\Sigma\|_2$, we have $q'_k \leq q_k$, so that assumption (V.49) is satisfied with $c_0 = 4$. The claim is then a consequence of Proposition V.25.

Proof of Cor. V.9. For any $k \in \widetilde{V}$, we have $k \in W_{(c)}$, and since $\varsigma \geq 1$, it holds (with the notation used in Proposition V.8, but using α/B in place of α)

$$\tau_{\min}^{(k)} = 64(u_\alpha + \log B)\left(\frac{1}{\sqrt{d_1^\bullet}} + \frac{s_k^2}{s_1^2\sqrt{d_k^\bullet}}\right) \leq \frac{\tau_{\min}^\circ}{2} + \frac{\varsigma\tau_{\min}^\circ}{2} \leq \varsigma\tau_{\min}^\circ,$$

and the result is a direct consequence of Proposition V.8 (combined with a union bound over $k \in [\![B]\!]$).

V.9.4 Proofs for Section V.3.4: estimating Schatten norms and plug-in estimates

We will be concentrating one one bag at a time and for this reason omit the task index k in the next results. Thus, we assume $\widetilde{X}_1, \dots, \widetilde{X}_N$ (with $N \geq 4$) are i.i.d. data points in $\mathbb{R}^d$ with expectation μ and known covariance matrix Σ. We start with estimators for the Schatten norms $\|\Sigma\|_p$, $p = 1, 2$.

We can use the natural unbiased estimator for any fixed $\|\Sigma\|_1 = \operatorname{Tr}\Sigma$,

$$\widetilde{Z}^{(1)} := \frac{1}{N-1}\sum_{i=1}^{N} \|\widetilde{X}_i - \widetilde{\mu}\|^2 = \frac{1}{2N(N-1)}\sum_{i\neq j}\left\|\widetilde{X}_i - \widetilde{X}_j\right\|^2, \tag{V.55}$$

where $\widetilde{\mu} = N^{-1}\sum_{i=1}^{N}\widetilde{X}_i$ is the empirical mean of the (sub-)sample.

Gaussian setting We have the following error control in the Gaussian setting:

Proposition V.26. *Assume (**GS**) holds. For $u \geq 1$, if $N \geq 2$*

$$\mathbb{P}\left[\left|\widetilde{Z}^{(1)} - \operatorname{Tr}\Sigma\right| \geq 4\sqrt{\frac{2\operatorname{Tr}\Sigma^2}{N}u}\right] \leq 2e^{-u}\,.$$

Proof of Proposition V.26 Let $\boldsymbol{X} = (\widetilde{X}_1 - \widetilde{\mu}, \dots, \widetilde{X}_N - \widetilde{\mu}) \in \mathbb{R}^{dN}$. Then $\boldsymbol{X}$ is a centred Gaussian vector with covariance matrix $\boldsymbol{\Sigma} := \Gamma \otimes \Sigma$ where $\Gamma = I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T \in \mathbb{R}^{N\times N}$, $\mathbf{1}_N = (1, \dots, 1) \in \mathbb{R}^N$ and $\otimes$ denotes the Kronecker product. Note that it holds $\operatorname{Tr}\Gamma = (N-1), \boldsymbol{\Sigma}^2 = \Gamma^2 \otimes \Sigma^2 = \Gamma \otimes \Sigma^2$, $\operatorname{Tr}\boldsymbol{\Sigma} = \operatorname{Tr}\Gamma\operatorname{Tr}\Sigma = (N-1)\operatorname{Tr}\Sigma$, and $\operatorname{Tr}(\boldsymbol{\Sigma}^2) = (N-1)\operatorname{Tr}\Sigma^2$. Then, according to Lemma V.40, for $u \geq 1$, with probability greater than $1 - 2e^{-u}$:

$$\begin{aligned}
\|\boldsymbol{X}\|_2^2 &\leq \operatorname{Tr}\boldsymbol{\Sigma} + 2\sqrt{\operatorname{Tr}\boldsymbol{\Sigma}^2 u} + 2\|\boldsymbol{\Sigma}\|_\infty u \leq (N-1)\operatorname{Tr}\Sigma + 4\sqrt{(N-1)\operatorname{Tr}\Sigma^2 u}\,,\\
\|\boldsymbol{X}\|_2^2 &\geq \operatorname{Tr}\boldsymbol{\Sigma} - 2\sqrt{\operatorname{Tr}\boldsymbol{\Sigma}^2 u} \geq (N-1)\operatorname{Tr}\Sigma - 2\sqrt{(N-1)\operatorname{Tr}\Sigma^2 u}\,.
\end{aligned}$$

We have used that $\sqrt{u} \leq u$ for $u \geq 1$. We conclude by remarking that $\|\boldsymbol{X}\|_2^2 = (N-1)\widetilde{Z}^{(1)}$. □

As in Section IV, we can estimate $\|\Sigma\|_2 = \sqrt{\operatorname{Tr}\Sigma^2}$ using the following U-statistic, which is an unbiased estimator of $\operatorname{Tr}\Sigma^2$:

$$(\widetilde{Z}^{(2)})^2 := \frac{1}{4N(N-1)(N-2)(N-3)}\sum_{i\neq j\neq k\neq l}\langle X_i - X_k, X_j - X_l\rangle^2\,. \tag{V.56}$$

Proposition V.27 restate the concentration bound of $\widetilde{Z}^{(2)}$ given in Proposition IV.12.

Proposition V.27. *Assume (**GS**) holds and $N \geq 4$. Then for all $u \geq 0$:*

$$\mathbb{P}\left[\left|\widetilde{Z}^{(2)} - \sqrt{\operatorname{Tr}\Sigma^2}\right| \geq 30\sqrt{\frac{\operatorname{Tr}\Sigma^2}{N}u^2}\right] \leq e^4 e^{-u}\,. \tag{V.57}$$

Proof of Proposition V.10 Proposition V.10 is a consequence of the above Proposition V.26, using the union bound over $k \in [\![B]\!]$. □

Propositions V.26 and V.27 can now be used to handle the plug-in versions of the quantities considered in Section V.3.3 when covariances are unknown:

Proposition V.28. *Assume (**GS**) holds, let $c \geq 1$ be a fixed number and let $\alpha \in (0,1)$. Assume that we have estimates $\widetilde{Z}_1^{(1)}$ for $\|\Sigma_1\|_1$ and $\widetilde{Z}_k^{(2)}$ for $\|\Sigma_k\|_2$, $k \in [\![B]\!]$, (depending on the independent "tilde" data only) such that with probability $1-\alpha$ it holds simultaneously for some constants $\eta_1, \eta_2 \in (0,1)$:*

$$\left|\widetilde{Z}_1^{(1)} - \|\Sigma_1\|_1\right| \leq \eta_1\|\Sigma_1\|_1; \tag{V.58}$$

$$\left|\widetilde{Z}_k^{(2)} - \|\Sigma_k\|_2\right| \leq \eta_2\|\Sigma_k\|_2, \text{ for all } k \in [\![B]\!]. \tag{V.59}$$

Consider the following plug-in versions of the quantities appearing in (V.17), (V.21)*:*

$$\widetilde{W}_{(\varsigma)} := \left\{ k \in [\![B]\!] : \frac{\widetilde{Z}_k^{(2)}}{N_k} \le \varsigma \frac{\widetilde{Z}_1^{(2)}}{N_1} \right\}, \qquad \widetilde{\widetilde{T}}_k^{(\tau)} := \mathbf{1}\left\{ \widetilde{U}_k \le \tau \frac{\widetilde{Z}_1^{(1)}}{N_1} \right\}. \tag{V.60}$$

Then, defining

$$\widetilde{\widetilde{V}}_{\tau,\varsigma} := \left\{ k \in \widetilde{W}_{(\varsigma)} : \widetilde{\widetilde{T}}_k^{(\tau)} = 1 \right\},$$

with probability at least $1 - 3\alpha$ *(with respect to the "tilde" data) it holds*

$$V_{\tau_-,\varsigma/\beta} \subseteq \widetilde{\widetilde{V}}_{\tau,\varsigma} \subseteq V_{\tau_+,\varsigma\beta}, \tag{V.61}$$

where $\tau_\pm := (1 \pm \eta_1)\big(\sqrt{\tau} \pm \sqrt{\tau^\circ_{\min}}\big)_+^2$*, (with* $\tau^\circ_{\min} = 64(\log 8B\alpha^{-1})/\sqrt{d_1^\bullet}$*),*
and $\beta := (1+\eta_2)/(1-\eta_2)$*.*

Proof. Assume that (V.58)-(V.59) are satisfied. Then $W_{(\varsigma/\beta)} \subseteq \widetilde{W}_{(\varsigma)} \subseteq W_{(\beta\varsigma)}$, with $\beta := (1+\eta_2)/(1-\eta_2)$. Furthermore, recalling $\widetilde{T}_k^{(\tau)} := \mathbf{1}\{\widetilde{U}_k \le \tau s_1^2\}$, then we have $\widetilde{T}_k^{((1-\eta_1)\tau)} \le \widetilde{\widetilde{T}}_k^{(\tau)} \le \widetilde{T}_k^{((1+\eta_1)\tau)}$; therefore

$$\left\{ k \in W_{(\varsigma/\beta)} : \widetilde{T}_k^{((1-\eta_1)\tau)} = 1 \right\} =: \widetilde{V}_- \subseteq \widetilde{\widetilde{V}} \subseteq \widetilde{V}_+ := \left\{ k \in W_{(\beta\varsigma)} : \widetilde{T}_k^{((1+\eta_1)\tau)} = 1 \right\}.$$

We can apply Corollary V.9 separately to $\widetilde{V}_-, \widetilde{V}_+$ and get that with probability $1-3\alpha$ (accounting for the union bound together with event (V.58)-(V.59)), (V.61) holds.. □

Proof of Prop. V.11. From Proposition V.26 with $u = \log(4B\alpha^{-1})$ and a union bound over tasks, with probability at least $1 - \alpha/2$ it holds

$$\forall k \in [\![B]\!] : \qquad \left| \widetilde{Z}_k^{(1)} - \|\Sigma_k\|_1 \right| \le \|\Sigma_k\|_2 \frac{\sqrt{32}\log(4B\alpha^{-1})}{\sqrt{N_k}} \le \frac{1}{\sqrt{a}} \|\Sigma_k\|_2, \tag{V.62}$$

where for the last inequality we used the assumption $N_k \ge a(4 + \log(2B\alpha^{-1}))^4 \ge a(4+\log 6)^2(4 + \log(2B\alpha^{-1}))^2 \ge 32a(\log(4B\alpha^{-1}))^2$ (also using $\alpha \le 1/3$ in that estimate).

Similarly, from Proposition V.27 with $u = (4 + \log(2B\alpha^{-1}))$, with probability at least $1 - \alpha/2$ it holds

$$\forall k \in [\![B]\!] : \qquad \left| \widetilde{Z}_k^{(2)} - \|\Sigma_k\|_2 \right| \le 30\|\Sigma\|_2 \frac{(4 + \log(2B\alpha^{-1}))^2}{\sqrt{N_k}} \le \frac{30}{\sqrt{a}} \|\Sigma_k\|_2. \tag{V.63}$$

Therefore, conditions (V.58)-(V.59) are satisfied simultaneously with probability $1-\alpha$, with $\eta_1 = \frac{1}{\sqrt{a}} \frac{1}{\sqrt{d_1^\bullet}}$ and $\eta_2 = \frac{30}{\sqrt{a}}$ (with $a \ge 4400$).

We apply Proposition V.28, but using the the values $(\widetilde{\tau}, 3c)$ given by (V.27) in place of (τ, ς). As a result we get with high probability the sandwiching property (V.61),

$$V_{\widetilde{\tau}_-,\varsigma} \subseteq \widetilde{\widetilde{V}}_{\widetilde{\tau},3\varsigma} \subseteq V_{\widetilde{\tau}_+}, \tag{V.64}$$

denoting $\widetilde{\tau}_\pm$ the formula for $\tau_\pm$ of Proposition V.28 where we replace (τ, ς) by $(\widetilde{\tau}, 3\varsigma)$. We proceed to get bounds for $\widetilde{\tau}_\pm = (1 \pm \eta_1)(\sqrt{\widetilde{\tau}} \pm \sqrt{3\varsigma\tau^\circ_{\min}})^2$.

Let us start with bounding the estimation error of $d_1^\bullet$ by $\widetilde{d}_1^\bullet$: it holds

$$\sqrt{\widetilde{d}_1^\bullet} = \frac{N_1 \widetilde{s}_1^2}{\widetilde{Z}_1^{(2)}} = \frac{\widetilde{Z}_1^{(1)}}{\widetilde{Z}_1^{(2)}} \le \frac{1+\eta_1}{1-\eta_2} \sqrt{d_1^\bullet} \le 2\sqrt{d_1^\bullet},$$

where the last inequality holds if $a \geq 4400$. We deduce

$$\widetilde{\tau}^{\circ}_{\min} = \frac{32\log(8B\alpha^{-1})}{\sqrt{\widetilde{d}^{\bullet}_1}} \geq \frac{1}{2}\cdot\frac{32\log(8B\alpha^{-1})}{\sqrt{d^{\bullet}_1}} = \tau^{\circ}_{\min}/2,$$

as defined in Proposition V.28. Furthermore, we have for $\eta_1 = \frac{1}{\sqrt{ad^{\bullet}_1}} \leq \frac{1}{\sqrt{a}}$ and $a \geq 4400$:

$$\frac{1}{1-\eta_1} = 1 + \frac{\eta_1}{1-\eta_1} \leq 1 + \frac{1/\sqrt{a}}{1-1/\sqrt{a}}\frac{1}{\sqrt{d^{\bullet}_1}} \leq 1 + \frac{1}{60\sqrt{d^{\bullet}_1}}.$$

Using the previous estimates we obtain

$$\widetilde{\tau} := \left(1 + \frac{1}{60\sqrt{\widetilde{d}^{\bullet}_1}}\right)\left(\sqrt{\tau} + \sqrt{6\varsigma\widetilde{\tau}^{\circ}_{\min}}\right)^2 \geq \frac{1}{1-\eta_1}\left(\sqrt{\tau} + \sqrt{3\varsigma\tau^{\circ}_{\min}}\right)^2.$$

It follows :

$$\widetilde{\tau}_- = (1-\eta_1)\left(\sqrt{\widetilde{\tau}} - \sqrt{3\varsigma\tau^{\circ}_{\min}}\right)^2 \geq \tau.$$

Now to get an upper bound on $\widetilde{\tau}_+$, similarly to above we have

$$\sqrt{\widetilde{d}^{\bullet}_1} \geq \frac{1-\eta_1}{1+\eta_2}d^{\bullet}_1 \geq \frac{\sqrt{d^{\bullet}_1}}{2},$$

and thus $\widetilde{\tau}^{\circ}_{\min} \leq 2\tau^{\circ}_{\min}$; it follows

$$\begin{aligned}\widetilde{\tau}_+ = (1+\eta_1)\left(\sqrt{\widetilde{\tau}} + \sqrt{3\varsigma\tau^{\circ}_{\min}}\right)^2 &\leq \left(1 + \frac{1}{66\sqrt{d^{\bullet}_1}}\right)\left(1 + \frac{1}{30\sqrt{d^{\bullet}_1}}\right)(\sqrt{\tau} + 3\sqrt{3\varsigma\tau^{\circ}_{\min}})^2 \\ &= \xi\tau,\end{aligned}$$

where $\xi := (1 + 1/(30\sqrt{d^{\bullet}_1}))(1 + 1/(66\sqrt{d^{\bullet}_1}))(1 + 3\sqrt{3\varsigma\tau^{\circ}_{\min}/\tau})^2$.

With these estimates in hand the sandwiching property (V.64) implies

$$V_{\tau,\varsigma} \subseteq \widetilde{\widetilde{V}}_{\widetilde{\tau},3\varsigma} \subseteq V_{\xi\tau}.$$

We use this property to apply Proposition (V.7) as earlier, and obtain

$$\frac{R_1(\widetilde{\boldsymbol{\omega}})}{s_1^2} \leq \left(\frac{1+\eta}{1-\eta}\right)\mathcal{B}(\xi\tau, \nu(V_{\tau,\varsigma})) \leq \left(1 + \frac{1}{25\sqrt{\min_k d^{\bullet}_k}}\right)\xi\mathcal{B}(\tau, \nu(V_{\tau,\varsigma})).$$

Elementary estimates lead to

$$\left(1 + \frac{1}{25\sqrt{\min_k d^{\bullet}_k}}\right)\xi \leq \left(1 + \frac{1}{10\sqrt{\min_k d^{\bullet}_k}}\right)\left(1 + \frac{30\sqrt{\varsigma\log(8B\alpha^{-1})}}{(d^{\bullet}_1)^{\frac{1}{4}}\sqrt{\tau}}\right)^2.$$

□

Bounded setting Proposition V.29 and Proposition V.30 give concentration bounds for $\widetilde{Z}^{(1)}$ and $\widetilde{Z}^{(2)}$ in bounded setting.

Proposition V.29. *Assume (**BS**) holds. For $u \geq 1$, if $N \geq 2$*

$$\mathbb{P}\left[\left|\widetilde{Z}^{(1)} - \operatorname{Tr}\Sigma\right| \geq 2\sqrt{2\frac{\operatorname{Var}[\|X_1-\mu\|^2]}{N}u} + 32\frac{M^2u}{N}\right] \leq 4e^{-u}.$$

Proof. Let us first remark that:

$$\widetilde{Z}^{(1)} = \frac{1}{N-1}\sum_{i=1}^{N}\left\|\widetilde{X}_i - \mu\right\|^2 - \frac{N\|\widetilde{\mu}-\mu\|^2}{N-1}$$

Using Bernstein's inequality (Lemma V.42), with probability greater than $1-2e^{-u}$:

$$\left|\sum_{i=1}^{N}\left\|\widetilde{X}_i - \mu\right\|^2 - N\operatorname{Tr}\Sigma\right| \leq \sqrt{2N\mathrm{Var}[\|X_1-\mu\|^2]u} + 8M^2u\,.$$

Using McDiarmid's inequality (Boucheron et al., 2004; McDiarmid, 1998), for $f(x_1,\dots,x_N) = \|N^{-1}\sum_{i=1}^{N}(x_i-\mu)\|$, with probability greater than $1-2e^{-u}$:

$$\begin{aligned}-\frac{4M^2}{N} \leq \|\widetilde{\mu}-\mu\|^2 - \frac{\operatorname{Tr}\Sigma}{N} &\leq \left(\mathbb{E}[\|\widetilde{\mu}-\mu\|] + \sqrt{\frac{2M^2u}{N}}\right)^2 - \frac{\operatorname{Tr}\Sigma}{N} \\ &\leq \left(\mathbb{E}[\|\widetilde{\mu}-\mu\|]^2 - \frac{\operatorname{Tr}\Sigma}{N}\right) + 2\mathbb{E}[\|\widetilde{\mu}-\mu\|]\sqrt{\frac{2M^2u}{N}} + \frac{2M^2u}{N} \leq 8\frac{M^2u}{N}\,,\end{aligned}$$

where we have used successively Jensen's inequality, that $\operatorname{Tr}\Sigma \leq 4M^2$ and $u \geq 1$. It only stays to use that $(N-1)^{-1} \leq 2N^{-1}$ for $N \geq 2$ and a triangle inequality to conclude the proof, with probability at least $1-4e^{-u}$:

$$\begin{aligned}\left|\widetilde{Z}^{(1)} - \operatorname{Tr}\Sigma\right| &\leq \frac{\sqrt{2N\mathrm{Var}[\|X_1-\mu\|^2]u}}{N-1} + \frac{8M^2u}{N-1} + \frac{8M^2u}{N-1} \\ &\leq 2\sqrt{2\frac{\mathrm{Var}[\|X_1-\mu\|^2]}{N}u} + 32\frac{M^2u}{N}\,.\end{aligned}$$

□

Similarly as in the Gaussian setting, we can estimate $\|\Sigma\|_2$ using the U-statistic (V.56). Proposition V.30 is a restatement of Proposition IV.13.

Proposition V.30 (Blanchard and Fermanian, 2023, Prop. 13)**.** *Assume (***BS***) holds and $N \geq 4$. Then for all $u \geq 0$:*

$$\mathbb{P}\left[\left|\widetilde{Z}^{(2)} - \sqrt{\operatorname{Tr}\Sigma^2}\right| \geq 12M^2\sqrt{\frac{u}{N}}\right] \leq 2e^{-u}\,. \tag{V.65}$$

Thanks to these concentration results, we are able to give a bound on the estimation error of the test method for bounded data on the model of Proposition V.11.

Proposition V.31. *Assume (***BS***) holds. Let $\alpha \in (0,1/3)$. Consider the set of estimated τ-neighbours $\widetilde{\widetilde{V}}_{\tau,\varsigma}$ defined in (V.26), assume $N_k \geq a\phi_k^2 d_k^\bullet \log(8B\alpha^{-1})$ for all $k \in [\![B]\!]$, for a big enough constant a ($a = 576$ works), and where $\phi_k := M^2/(\operatorname{Tr}\Sigma_k)$.*

For fixed $\tau > 0$, $\varsigma \geq 1$, consider the weights $\widetilde{\boldsymbol{\omega}}^\sharp$ obtained by the modified *plug-in $\left(\widetilde{\widetilde{V}}_{\widetilde{\tau},3\varsigma}, \widetilde{\boldsymbol{s}}^2\right)$ for $(V, \boldsymbol{s}^2)$ in (V.13), where*

$$\widetilde{\tau} := \left(1 + \frac{1}{2\sqrt{\widetilde{d}_1^\bullet}}\right)\left(\sqrt{\tau} + \sqrt{6\widetilde{\tau}^\circ_{\min}}\right)^2; \qquad \widetilde{\tau}^\circ_{\min} := \frac{80c_0^2\varsigma\left(\log(8B\alpha^{-1})\right)}{\sqrt{\widetilde{d}_1^\bullet}}; \qquad \sqrt{\widetilde{d}_1^\bullet} := \frac{N_1\widetilde{s}_1^2}{\widetilde{Z}_1^{(2)}}. \tag{V.66}$$

and $c_0 = 31$. Then with probability at least $1-3\alpha$ over the draw of the "tilde" sample $(\widetilde{X}_\bullet^{(k)})_{k\in[\![B]\!]}$, it holds

$$\frac{R_1(\widetilde{\boldsymbol{\omega}}^\sharp)}{s_1^2} \leq \left(1 + \frac{4}{\sqrt{\min_k d_k^\bullet}}\right)\left(1 + \frac{900\sqrt{\varsigma\log(8B\alpha^{-1})}}{(d_1^\bullet)^{\frac{1}{4}}\sqrt{\tau}}\right)^2 \mathcal{B}\big(\tau, \nu(V_{\tau,\varsigma})\big),$$

where the expected risk is with respect to the main sample $(X_\bullet^{(k)})_{k\in[\![B]\!]}$.

Proof of Prop. V.31. From Proposition V.29 with $u = \log(8B\alpha^{-1})$ and a union bound over tasks, with probability at least $1 - \alpha/2$ it holds

$$\forall k \in [\![B]\!] : \qquad \left|\widetilde{Z}_k^{(1)} - \|\Sigma_k\|_1\right| \leq 2\sqrt{2\frac{\|\Sigma_k\|_1 M^2 u}{N_k}} + 32\frac{M^2 u}{N_k} \leq \frac{1}{3}\|\Sigma_k\|_2 \tag{V.67}$$

where for the last inequality we used the assumption $N_k \geq 64a\phi_k^2 d_k^\bullet \log(8B\alpha^{-1})$. Similarly, from Proposition V.30 with $u = \log(4B\alpha^{-1})$, with probability at least $1 - \alpha/2$ it holds

$$\forall k \in [\![B]\!] : \qquad \left|\widetilde{Z}_k^{(2)} - \|\Sigma_k\|_2\right| \leq 12M^2\sqrt{\frac{u}{N}} \leq \frac{1}{6}\|\Sigma_k\|_2. \tag{V.68}$$

Therefore, as in the Gaussian case (see proof of Proposition V.11), with $\eta_2 = 1/6$ and $\beta = (1+\eta_2)/(1-\eta_2) \leq 3$:

$$\widetilde{W}_\varsigma \subseteq W_{3\varsigma}$$

Let $\tau_{\min}^\circ = 80c_0^2\varsigma u(d_1^\bullet)^{-1/2}$, one can check that $\tau_{\min}^\circ \geq \tau_{\min}^k$ for all $k \in V_{\tau,\varsigma}$. Indeed, in bounded setting:

$$\tau_{\min}^k \leq 2c_0^2\left(16u\frac{1+\varsigma}{\sqrt{d_1^\bullet}} + 4u^2 s_1^{-2}\left(\frac{\phi_1 s_1^2}{N_1} + \frac{\phi_k s_k^2}{N_k}\right)\right) \leq 2c_0^2\left(\frac{32\varsigma u}{\sqrt{d_1^\bullet}} + 4u\frac{1+\varsigma}{d_1^\bullet}\right) \leq \frac{80c_0^2\varsigma u}{\sqrt{d_1^\bullet}}$$

where we have used that $\varsigma \geq 1$, the assumption on N_k and the expression of $\tau_{\min}^k$ given by Proposition V.25. We apply Proposition V.25 to $\widetilde{\tau}$ defined in (V.66), then with high probability:

$$V_{\widetilde{\tau}_-,\varsigma} \subseteq \widetilde{\widetilde{V}}_{\widetilde{\tau},3\varsigma} \subseteq V_{\widetilde{\tau}_+}, \tag{V.69}$$

where $\widetilde{\tau}_\pm = (1 \pm \eta_1)\left(\sqrt{\widetilde{\tau}} \pm \sqrt{\tau_{\min}^\circ}\right)$. We proceed to get bounds for $\widetilde{\tau}_\pm$.

Let us start with bounding the estimation error of $d_1^\bullet$ by $\widetilde{d}_1^\bullet$: it holds

$$\sqrt{\widetilde{d}_1^\bullet} = \frac{N_1\widetilde{s}_1^2}{\widetilde{Z}_1^{(2)}} = \frac{\widetilde{Z}_1^{(1)}}{\widetilde{Z}_1^{(2)}} \leq \frac{1+\eta_1}{1-\eta_2}\sqrt{d_1^\bullet} \leq 2\sqrt{d_1^\bullet},$$

where $\eta_1 = (d_1^\bullet)^{-1/2}/3 \leq 1/3$. We deduce

$$\widetilde{\tau}_{\min}^\circ = \frac{80c_0^2\varsigma u}{\sqrt{\widetilde{d}_1^\bullet}} \geq \frac{1}{2}\cdot\frac{80c_0^2\varsigma u}{\sqrt{d_1^\bullet}} = \tau_{\min}^\circ/2,$$

Furthermore, as $\eta_1 \leq 1/3$:

$$\frac{1}{1-\eta_1} = 1 + \frac{\eta_1}{1-\eta_1} \leq 1 + \frac{1}{2\sqrt{d_1^\bullet}}.$$

Using the previous estimates we obtain

$$\widetilde{\tau} := \left(1 + \frac{1}{2\sqrt{\widetilde{d}_1^\bullet}}\right)\left(\sqrt{\tau} + \sqrt{6\varsigma\widetilde{\tau}_{\min}^\circ}\right)^2 \geq \frac{1}{1-\eta_1}\left(\sqrt{\tau} + \sqrt{3\varsigma\tau_{\min}^\circ}\right)^2.$$

It follows :

$$\widetilde{\tau}_- = (1-\eta_1)\left(\sqrt{\widetilde{\tau}} - \sqrt{3\varsigma\tau_{\min}^\circ}\right)^2 \geq \tau.$$

Now to get an upper bound on $\widetilde{\tau}_+$, similarly to above we have

$$\sqrt{\widetilde{d}_1^\bullet} \geq \frac{1-\eta_1}{1+\eta_2}d_1^\bullet \geq \frac{\sqrt{d_1^\bullet}}{2},$$

and thus $\widetilde{\tau}^{\circ}_{\min} \le 2\tau^{\circ}_{\min}$; it follows

$$\widetilde{\tau}_+ = (1+\eta_1)\big(\sqrt{\widetilde{\tau}} + \sqrt{3\varsigma\tau^{\circ}_{\min}}\big)^2 \le \left(1+\frac{1}{3\sqrt{d_1^\bullet}}\right)\left(1+\frac{1}{\sqrt{d_1^\bullet}}\right)(\sqrt{\tau}+3\sqrt{3\varsigma\tau^{\circ}_{\min}})^2 = \xi\tau,$$

where $\xi := (1+1/(3\sqrt{d_1^\bullet}))(1+1/\sqrt{d_1^\bullet})(1+3\sqrt{3\varsigma\tau^{\circ}_{\min}/\tau})^2$.

With these estimates in hand the sandwiching property (V.69) implies

$$V_{\tau,\varsigma} \subseteq \widetilde{\widetilde{V}}_{\widetilde{\tau},3\varsigma} \subseteq V_{\xi\tau}.$$

We use this property to apply Proposition (V.7) as earlier, and obtain

$$\frac{R_1(\widetilde{\boldsymbol{\omega}})}{s_1^2} \le \left(\frac{1+\eta}{1-\eta}\right)\mathcal{B}(\xi\tau, \nu(V_{\tau,\varsigma})) \le \left(1+\frac{1}{2\sqrt{\min_k d_k^\bullet}}\right)\xi\mathcal{B}(\tau,\nu(V_{\tau,\varsigma})).$$

Elementary estimates lead to

$$\left(1+\frac{1}{2\sqrt{\min_k d_k^\bullet}}\right)\xi \le \left(1+\frac{4}{\sqrt{\min_k d_k^\bullet}}\right)\left(1+\frac{900\sqrt{\varsigma\log(8B\alpha^{-1})}}{(d_1^\bullet)^{\frac{1}{4}}\sqrt{\tau}}\right)^2.$$

□

Heavy-tailed setting Similarly as in Sup. V.9.4 and V.9.4, we provide in this section estimators of $\|\Delta_k\|^2$, $\|\Sigma_k\|_1$ and $\|\Sigma_k\|_2$ but for heavy-tailed data. These estimators can be directly used to estimate the neighbours $V_{\tau,\varsigma}$ and the oracle weights to then apply the testing approach in this setting.

Assumption V.32 (**HT**, Heavy-tailed setting)**.** *For all $k \in [\![B]\!]$, $\mathbb{P}_k$ has a finite fourth moment.*

Consider a statistic $T(N; x_1, \dots x_N)$ in $\mathbb{R}$, the Median of Blocks statistics $\mathrm{MOB}_b(T)$ for b a divisor of N is defined by the median of the statistics T^a, $1 \le a \le b$ built from a b-partition of $x_1, \dots x_N$:

$$\mathrm{MOB}_k(T) := \mathrm{Median}(T^a, 1 \le a \le b)$$

where $T^a = T(N/b; x_{aN/b+1}, \dots x_{(a+1)N/b})$. If b does not divide N, it suffices to partition the sample into sub-samples of size $\lfloor N/b \rfloor$ and $\lceil N/b \rceil$. If the original estimator is constructed from different samples (e.g., (V.16)), each sample is partitioned into b subsamples.

Proposition V.33. *Assume (**HT**) holds, let $0 \le u \le N$ and $b = \lceil u \rceil$, let $U(X_\bullet^{(1)}, X_\bullet^{(k)})$ the estimator of $\|\Delta_k\|^2$ defined in (V.16), then, with probability greater than $1 - e^{-u/8}$:*

$$\left|\mathrm{MOB}_b(U(X_\bullet^{(1)}, X_\bullet^{(k)})) - \|\Delta_k\|^2\right| \le 4\sqrt{\Delta_k^T\left(\frac{\Sigma_1}{N_1}+\frac{\Sigma_k}{N_k}\right)\Delta_k u} + 4\left(\frac{\|\Sigma_1\|_2}{N_1}+\frac{\|\Sigma_k\|_2}{N_k}\right)u\,. \tag{V.70}$$

In the kernel setting, the statistic $U(X_\bullet^{(1)}, X_\bullet^{(k)})$ is an estimator of the MMD distance between $\mathbb{P}_1$ and $\mathbb{P}_k$. (Lerasle et al., 2019 proposed a different robust estimator of this quantity called MONK, but we focus here on the MOB estimator, which has the advantage to be easier to compute and to study.)

Proposition V.34. *Assume (**HT**) holds, let $0 \le u \le N/4$ and $b = \lceil u \rceil$:*

$$\mathbb{P}\left[\left|\mathrm{MOB}_b(Z^{(1)}) - \mathrm{Tr}\,\Sigma\right| \ge C\sqrt{\frac{\mathrm{Var}[\|X_1-\mu\|^2]u}{N}} + C\frac{\sqrt{\mathrm{Tr}\,\Sigma^2}u}{N}\right] \le e^{-u/8}\,,$$

$$\mathbb{P}\left[\left|\sqrt{\mathrm{MOB}_b(Z^{(2)})} - \sqrt{\mathrm{Tr}\,\Sigma^2}\right| \ge C\sqrt{\frac{M_X u}{N}}\right] \le e^{-u/8}\,,$$

where $Z^{(1)}$ is defined in (V.55), $Z^{(2)}$ in (V.56), $C > 0$ is an absolute constant and $M_X = \mathbb{E}\big[\|X_1-\mu\|^4\big]$.

Proposition V.33 and Proposition V.34 are different consequences of Lemma V.35 below. Some more refined concentration bounds can be derived for MOB-type statistics (see, e.g., Devroye et al., 2016; Minsker, 2019), but the present results are sufficient to show that in the (**HT**) setting suitable statistics satisfy Assumption (**TSC**) and (V.58)-(V.59).

Proof of Proposition V.33. According to Lemma V.35, we only need compute the variances of the statistics $\widetilde{U}_a$,

$$\begin{aligned}\mathrm{Var}\Big[\widetilde{U}_k\Big] &= 4\sqrt{\Delta_k^T\left(\frac{\Sigma_1}{N_1}+\frac{\Sigma_k}{N_k}\right)\Delta_k} + 2\,\mathrm{Tr}\left(\frac{\Sigma_1}{N_1}+\frac{\Sigma_k}{N_k}\right)^2 + 2\left(\frac{\|\Sigma_1\|_2}{N_1^2(N_1-1)}+\frac{\|\Sigma_i\|_2}{N_i^2(N_i-1)}\right)\\ &\le 4\sqrt{\Delta_k^T\left(\frac{\Sigma_1}{N_1}+\frac{\Sigma_i}{N_i}\right)\Delta_k} + 4\left(\frac{\|\Sigma_1\|_2}{N_1}+\frac{\|\Sigma_k\|_2}{N_k}\right) =: \widetilde{v}(N_1,N_i)\end{aligned}$$

We apply Lemma V.35 with $N = N_1 + N_i$ and $v(N/u) := \widetilde{v}(N_1/u, N_k/u)$. □

Proof of Proposition V.34.

For $Z^{(1)}$ the concentration bound is deduced directly from the variance:

$$\mathrm{Var}\Big[Z^{(1)}\Big] = \frac{\mathrm{Var}\big[\|X-\mu\|^2\big]}{N} + \frac{2\|\Sigma\|_2^2}{N(N-1)}\,.$$

For $Z^{(2)}$ we can first assume w.l.g. than X is centred. Then $Z^{(2)}$ can be developed as:

$$(Z^{(2)})^2 = \frac{1}{N^{(2)}}\sum_{i\neq j}\langle X_i,X_j\rangle^2 - \frac{2}{N^{(3)}}\sum_{i\neq j\neq k}\langle X_i,X_j\rangle\langle X_i,X_k\rangle - \frac{1}{N^{(4)}}\sum_{i\neq j\neq k\neq q}\langle X_i,X_j\rangle\langle X_k,X_q\rangle\,.$$

where $n^{(p)} = n(n-1)\dots(n-p+1)$ for $n \ge p \in \mathbb{N}$. Let us first compute $\mathrm{Var}\big[(Z^{(2)})^2\big]$:

$$\begin{aligned}\mathrm{Var}\Big[(Z^{(2)})^2\Big] &\le \frac{2}{N^{(2)}}\mathbb{E}\Big[\langle X,X'\rangle^4\Big] + \frac{4(N-2)}{N^{(2)}}\mathbb{E}\Big[\big(X^T\Sigma X\big)^2\Big]\\ &\quad + \frac{4}{N^{(3)}}\big((3!)M_X^2 + 2(N-3)\,\mathrm{Tr}\,\Sigma^4\big) + \frac{4!}{N^{(4)}}M_X^2\\ &\le C\frac{\|\Sigma\|_\infty^2 M_X}{N} + C\frac{M_X^2}{N^2}\end{aligned}$$

where $C > 0$ is some absolute constant. Then according to Lemma V.35, for $u \le N/4$, with probability grater than $1 - e^{-u/8}$:

$$\Big|\mathrm{MOB}_b((Z^{(2)})^2) - \mathrm{Tr}\,\Sigma^2\Big| \le C\|\Sigma\|_\infty\sqrt{\frac{M_X u}{N}} + C\frac{M_X u}{N}\,, \tag{V.71}$$

Using that $\Big|\sqrt{(a^2+b)_+} - a\Big| \le \min\Big(\sqrt{|b|}, \frac{b}{a}\Big)$ for $a \in \mathbb{R}_+$ and $b \in \mathbb{R}$, (see, e.g., Lemma 15 of Blanchard and Fermanian, 2023), assuming (V.71), then

$$\begin{aligned}\Big|\mathrm{MOB}_b(Z^{(2)}) - \sqrt{\mathrm{Tr}\,\Sigma^2}\Big| &\le \max_{\varepsilon\in\{-1,1\}}\left|\sqrt{\mathrm{Tr}\,\Sigma^2 + \varepsilon C\|\Sigma\|_\infty\sqrt{\frac{M_X u}{N}}} - \sqrt{\mathrm{Tr}\,\Sigma^2}\right| + C\sqrt{\frac{M_X u}{N}}\\ &\le C\frac{\|\Sigma\|_\infty}{\sqrt{\mathrm{Tr}\,\Sigma^2}}\sqrt{\frac{M_X u}{N}} + C\sqrt{\frac{M_X u}{N}} \le C\sqrt{\frac{M_X u}{N}}\,.\end{aligned}$$

□

Lemma V.35. *Let $T(N; x_1, \dots x_N)$ a statistic build from N i.i.d. random variables such that for all $N \ge N_0$:*

$$\mathbb{E}[T(N; X_1, \dots, X_N)] = \mathbb{E}[T], \qquad \mathrm{Var}[T(N; X_1, \dots, X_N)] \le v(N),$$

where $v : \mathbb{R}_+ \to \mathbb{R}_+$ is nonincreasing. Let $1 \le u \le N/(N_0+1)$ and $b = \lceil u \rceil$, then

$$\mathbb{P}\left[|\mathrm{MOB}_b(T) - \mathbb{E}[T]| \ge \sqrt{4v\left(\frac{N}{4u}\right)}\right] \le e^{-u/8}\,.$$

Proof of Lemma V.35.
First assume that $b|N$. Let us denote $T_a := T(N/b; x_{(a-1)N/b+1}, \dots x_{aN/b})$ for $a \in [\![b]\!]$. Then for all $a \in [\![b]\!]$, by Markov's inequality:

$$\mathbb{P}\Big[|T_a - \mathbb{E}[T]| \ge \sqrt{4v(N/k)}\Big] \le \frac{1}{4}\,. \tag{V.72}$$

Then, $|\mathrm{MOB}_b(T) - \mathbb{E}[T]| \ge \sqrt{4v(N/b)}$ implies that at least $b/2$ of T_a satisfies

$$|T_a - \mathbb{E}[T]| \ge \sqrt{4v(N/b)}\,.$$

By independence of the T_a and Hoeffding's inequality:

$$\mathbb{P}\Big[|\mathrm{MOB}_b(T) - \mathbb{E}[T]| > \sqrt{4v(N/b)}\Big] \le \mathbb{P}\left[\mathrm{Bin}\left(b, \frac{1}{4}\right) \ge \frac{b}{2}\right] \le e^{-b/8}\,,$$

where Bin denotes the Binomial distribution. Because $u \le b \le u+1$ and v is a noninccreasing function, we can conclude:

$$e^{-b/8} \le e^{-u/8}\,, \qquad v\left(\frac{N}{b}\right) \le v\left(\frac{N}{u+1}\right) \le v\left(\frac{N}{4u}\right).$$

If $b \nmid N$, equation (V.72) is still verified with $v\big(\lfloor \frac{N}{b} \rfloor\big)$ instead of $v\big(\frac{N}{b}\big)$ and:

$$\left\lfloor \frac{N}{\lceil u \rceil} \right\rfloor \ge \frac{N}{\lceil u \rceil} - 1 \ge \frac{N}{2\lceil u \rceil} \qquad \text{if } \lceil u \rceil \le N/2$$
$$\left\lfloor \frac{N}{\lceil u \rceil} \right\rfloor = 1 \ge \frac{N}{2\lceil u \rceil} \qquad \text{if } N \ge \lceil u \rceil > N/2\,.$$

We conclude using that $\lceil u \rceil \le (u+1) \le 2u$ for $u \ge 1$. □

V.9.5 Proofs for Section V.4

Proof of Proposition V.12 Let $\widehat{\boldsymbol{\omega}} \in \arg\min_{\boldsymbol{\omega} \in \mathcal{S}_B} \big(\widehat{L}_1(\boldsymbol{\omega}) + 16\sqrt{u_0}\widehat{Q}_1(\boldsymbol{\omega})\big)$. Denote $\mathcal{X}^{-1} = (X_\bullet^{(k)})_{k \ne 1}$ the observed bag data except for the first bag, which corresponds to the target task.

First step : bound in conditional probability. As a first step, we obtain a high-probability bound for $L_1(\widehat{\boldsymbol{\omega}})$. For $x \ge 1$, define the event $A(x)$:

$$A(x) := \left\{ \begin{array}{lll} \sqrt{\mathfrak{q}_k} \le c_1(x)\sqrt{\widehat{q}_k} + C\frac{s_1^2}{d_1^e}\sqrt{N_1 x}, & 2 \le k \le B, & \text{(a)} \\ \sqrt{\widehat{q}_k} \le \left(1 + \sqrt{\frac{2x}{N_1 - 1}}\right)\left(\sqrt{\mathfrak{q}_k + \frac{s_1^4}{d_1^\bullet} N_1} + \frac{s_1^2}{d_1^e}\sqrt{2N_1 x}\right), & 2 \le k \le B, & \text{(b)} \\ \left|\widehat{s}_1^2 - s_1^2\right| \le C\frac{s_1^2}{\sqrt{d_1^\bullet N_1}}x\,, & & \text{(c)} \\ \|\widehat{\mu}_1^{\mathtt{NE}} - \mu_1\|^2 \le s_1^2 + C\frac{s_1^2}{\sqrt{d_1^\bullet}}x\,, & & \text{(d)} \\ |\langle \widehat{\mu}_k^{\mathtt{NE}} - \mu_1, \widehat{\mu}_1^{\mathtt{NE}} - \mu_1 \rangle| \le \sqrt{2\frac{\mathfrak{q}_k}{N_1}x}, & 2 \le k \le B, & \text{(e)} \end{array} \right\}$$

where $\mathfrak{q}_k = (\widehat{\mu}_k^{\mathtt{NE}} - \mu_1)^T \Sigma_1 (\widehat{\mu}_k^{\mathtt{NE}} - \mu_1)$ and $c_1(x) = \sqrt{e}\exp(x/(N_1 - 1))$. For the whole proof, the notation C will denote an absolute numeric constant whose value can change between lines. The probability of the event A conditionally to $\mathcal{X}^{(-1)}$ is bounded as:

$$\mathbb{P}\Big[A^c(x,y)|\mathcal{X}^{(-1)}\Big] \le (6B+4)e^{-x}. \tag{V.73}$$

We combine a union bound with estimates for each individual bound: bounds (a) and (b) are consequences of Proposition V.37 with $\nu = \widehat{\mu}_k^{\mathtt{NE}}$. For (a), we have used that $\sqrt{\mathfrak{q}_k} \le \sqrt{\mathfrak{q}_k + \operatorname{Tr}\Sigma_1^2/N_1}$. Bound (c) is a rewriting of Proposition V.26. Bound (d) is a consequence of Lemma V.40 with $X = \widehat{\mu}_1^{\mathtt{NE}} - \mu_1$, $\mu = 0$, $\Sigma = \Sigma_1/N_1$; bounding $\sqrt{x}$ by x, and $\|\Sigma_1\|_\infty$ by $\sqrt{\operatorname{Tr}\Sigma_1^2}$. Finally (e) is deduced from Lemma V.39 with $X = \langle \widehat{\mu}_k^{\mathtt{NE}} - \mu_1, \widehat{\mu}_1^{\mathtt{NE}} - \mu_1\rangle$, $m = 0$ and $\sigma^2 = \mathfrak{q}_k$.

From now on, assume that event $A(x)$ holds. Then,

$$\begin{aligned}
L_1(\widehat{\boldsymbol{\omega}}) &= \Big\| \sum_{k=1}^B \widehat{\omega}_k(\widehat{\mu}_k^{\mathtt{NE}} - \widehat{\mu}_1^{\mathtt{NE}}) + (\widehat{\mu}_1^{\mathtt{NE}} - \mu_1)\Big\|^2 \\
&= \Big\| \sum_{k=2}^B \widehat{\omega}_k(\widehat{\mu}_k^{\mathtt{NE}} - \widehat{\mu}_1^{\mathtt{NE}})\Big\|^2 + 2\sum_{k=2}^B \widehat{\omega}_k\langle \widehat{\mu}_k^{\mathtt{NE}} - \widehat{\mu}_1^{\mathtt{NE}}, \widehat{\mu}_1^{\mathtt{NE}} - \mu_1\rangle + \big\|\widehat{\mu}_1^{\mathtt{NE}} - \mu_1\big\|^2 \\
&= \Big\| \sum_{k=2}^B \widehat{\omega}_k(\widehat{\mu}_k^{\mathtt{NE}} - \widehat{\mu}_1^{\mathtt{NE}})\Big\|^2 + 2\sum_{k=2}^B \widehat{\omega}_k\langle \widehat{\mu}_k^{\mathtt{NE}} - \mu_1, \widehat{\mu}_1^{\mathtt{NE}} - \mu_1\rangle + (2\widehat{\omega}_1 - 1)\|\widehat{\mu}_1^{\mathtt{NE}} - \mu_1\|^2 \\
&= \widehat{L}_1(\widehat{\boldsymbol{\omega}}) + 2\sum_{k=2}^B \widehat{\omega}_k\langle \widehat{\mu}_k^{\mathtt{NE}} - \mu_1, \widehat{\mu}_1^{\mathtt{NE}} - \mu_1\rangle + (2\widehat{\omega}_1 - 1)\big(\|\widehat{\mu}_1^{\mathtt{NE}} - \mu_1\|^2 - s_1^2\big) \\
&\quad + (2\widehat{\omega}_1 - 1)\big(s_1^2 - \widehat{s}_1^2\big)
\end{aligned}$$

Using (e) and then (a) for the second term, (d) for the third and (c) for the last, we get:

$$\begin{aligned}
L_1(\widehat{\boldsymbol{\omega}}) &\le \widehat{L}_1(\widehat{\boldsymbol{\omega}}) + 2c_1(x)\sqrt{2x}\sum_{k=2}^B \widehat{\omega}_k\sqrt{\frac{\widehat{q}_k}{N_1}} + Cs_1^2\left(\frac{x}{\sqrt{d_1^\bullet}} + \frac{x}{d_1^{\mathrm{e}}}\right) \\
&\le \left(1 \vee \frac{c_1(x)\sqrt{2x}}{8\sqrt{u_0}}\right)\min_{\boldsymbol{\omega}\in\mathcal{S}_B}\left(\widehat{L}_1(\boldsymbol{\omega}) + 16\sqrt{u_0}\sum_{k=2}^B \omega_k\sqrt{\frac{\widehat{q}_k}{N_1}}\right) + Cs_1^2\frac{x}{\sqrt{d_1^\bullet}}\,. \qquad \text{(V.74)}
\end{aligned}$$

The appearance of the minimum is a consequence of the definition of $\widehat{\boldsymbol{\omega}}$.

Second step : conditional bound in expectation. We can now deduce, from the previous step, a bound in expectation conditionally to all samples expect the first one. For any fixed $\omega \in \mathcal{S}_B$, we first want to compare $\widehat{L}_1(\boldsymbol{\omega})$ to its conditional expectation $\mathbb{E}\big[\widehat{L}_1(\boldsymbol{\omega})\big|\mathcal{X}^{(-1)}\big]$ which is equal to the conditional expectation of the loss L_1:

$$\mathbb{E}\big[\widehat{L}_1(\boldsymbol{\omega})\big|\mathcal{X}^{(-1)}\big] = \Big\|\sum_{k=2}^B \omega_k(\widehat{\mu}_k^{\mathtt{NE}} - \mu_1)\Big\|^2 + \omega_1^2 s_1^2 = \mathbb{E}\big[L_1(\boldsymbol{\omega})\big|\mathcal{X}^{(-1)}\big].$$

For any fixed $\boldsymbol{\omega} \in \mathcal{S}_B$, as $x \ge 1$:

$$\begin{aligned}
\widehat{L}_1(\boldsymbol{\omega}) &= \Big\|\sum_{k=2}^B \omega_k(\widehat{\mu}_k^{\mathtt{NE}} - \widehat{\mu}_1^{\mathtt{NE}})\Big\|^2 + (2\omega_1 - 1)\hat{s}_1^2 \\
&= \Big\|\sum_{k=2}^B \omega_k(\widehat{\mu}_k^{\mathtt{NE}} - \mu_1) + (1-\omega_1)(\mu_1 - \widehat{\mu}_1^{\mathtt{NE}})\Big\|^2 + (2\omega_1 - 1)\hat{s}_1^2 \\
&= \mathbb{E}\Big[L_1(\boldsymbol{\omega})|\mathcal{X}^{(-1)}\Big] + 2(1-\omega_1)\sum_{k=2}^B \omega_k\langle \widehat{\mu}_k^{\mathtt{NE}} - \mu_1, \mu_1 - \widehat{\mu}_1^{\mathtt{NE}}\rangle \\
&\quad + (1-\omega_1)^2\big(\|\widehat{\mu}_1^{\mathtt{NE}} - \mu_1\|^2 - s_1^2\big) + (2\omega_1 - 1)(\hat{s}_1^2 - s_1^2) \\
&\le \mathbb{E}\Big[L_1(\boldsymbol{\omega})|\mathcal{X}^{(-1)}\Big] + 2\sqrt{2x}\sum_{k=2}^B \omega_k\sqrt{\frac{\mathfrak{q}_k}{N_1}} + C\frac{s_1^2 x}{\sqrt{d_1^\bullet}}, \qquad \text{(V.75)}
\end{aligned}$$

using (c), (d), (e) again. From (b), for all $k \in [\![B]\!]$, and using again $x \geq 1$:

$$\begin{aligned}\sqrt{\widehat{q}_k} &\leq \left(1+\sqrt{\frac{2x}{N_1-1}}\right)\left(\sqrt{\mathfrak{q}_k + \frac{s_1^4}{d_1^\bullet}N_1} + \frac{s_1^2}{d_1^{\mathrm{e}}}\sqrt{2N_1 x}\right) \\ &\leq \left(1+\sqrt{\frac{2x}{N_1-1}}\right)\sqrt{\mathfrak{q}_k} + C\left(\sqrt{x}+\frac{x}{\sqrt{N_1-1}}\right)\sqrt{\frac{N_1}{d_1^\bullet}}s_1^2. \end{aligned} \tag{V.76}$$

Then, plugging (V.75) and (V.76) into (V.74) , for all $\boldsymbol{\omega} \in \mathcal{S}_B$, as $x \geq 1$:

$$\begin{aligned}L_1(\widehat{\boldsymbol{\omega}}) \leq \left(1 \vee \frac{c_1(x)\sqrt{2x}}{8\sqrt{u_0}}\right)\Bigg[&\mathbb{E}\Big[L_1(\boldsymbol{\omega})|\mathcal{X}^{(-1)}\Big] \\ &+ \left(2\sqrt{2x} + 16\sqrt{u_0}\left(1+\sqrt{\frac{2x}{N_1-1}}\right)\right)\sum_{k=2}^{B}\omega_k\sqrt{\frac{\mathfrak{q}_k}{N_1}} \\ &+ \frac{Cs_1^2}{\sqrt{d_1^\bullet}}\left(x + C\sqrt{u_0}\left(\sqrt{x}+\frac{x}{\sqrt{N_1-1}}\right)\right)\Bigg].\end{aligned}$$

By rearranging the terms and using that $u_0 \leq N_1 - 1$ and $x \geq 1$:

$$\begin{aligned}L_1(\widehat{\boldsymbol{\omega}}) \leq \left(1 \vee \frac{c_1(x)\sqrt{2x}}{8\sqrt{u_0}}\right)\Bigg[&\mathbb{E}\Big[L_1(\boldsymbol{\omega})|\mathcal{X}^{(-1)}\Big] \\ &+ C\big(\sqrt{u_0}+\sqrt{x}\big)\sum_{k=2}^{B}\omega_k\sqrt{\frac{\mathfrak{q}_k}{N_1}} + \frac{Cs_1^2}{\sqrt{d_1^\bullet}}(\sqrt{u_0 x}+x)\Bigg] =: \psi(x)P(x)\end{aligned}$$

where $\psi(x) := 1 \vee \frac{c_1(x)\sqrt{2x}}{8\sqrt{u_0}}$ and P is a degree 2 polynomial in $\sqrt{x}$ with coefficients that are constant conditionally to $\mathcal{X}^{(-1)}$. We will denote the shifted version of ψ and P by ψ_s and P_s, for $v \geq 0$:

$$\psi_s(v) := \psi(v + \log(6B+4))\,, \quad P_s(v) = P(v+\log(6B+4))\,. \tag{V.77}$$

Both notations will be used depending on the case for the sake of readability. So for all $v \geq 0$

$$\mathbb{P}\Big[L_1(\widehat{\boldsymbol{\omega}}) \geq \psi_s(v)P_s(v)|\mathcal{X}^{(-1)}\Big] \leq e^{-v}.$$

thanks to (V.73) after taking $x = v + \log(6B+4) \geq 1$. Then there exists a random variable ξ following an exponential distribution of parameter 1 conditionally to $\mathcal{X}^{(-1)}$, such that $L_1(\widehat{\omega}) \leq \psi_s(\xi)P_s(\xi)$ almost surely. Let us first simplify the expression of ψ, recalling that by assumption $(N_1-1)/2 \geq u_0 \geq \log(17B) \geq 1/2 + \log(6B+4) \geq 1/2 + \log(10) \geq 5/2$, then for $x \leq u_0$:

$$\sqrt{2}c_1(x) \leq \sqrt{2}\exp\left(\frac{1}{2}+\frac{u_0}{N_1-1}\right) \leq \sqrt{2}\exp\left(\frac{1}{2}+\frac{1}{2}\right) \leq \sqrt{2}e \leq 8.$$

Thus, for $x \leq u_0$, $\psi(x) = 1$. For $x \geq u_0$, it holds $c_1(x) \geq \sqrt{e} \geq 1$, so that:

$$\begin{aligned}\psi(x) \leq \frac{c_1(x)\sqrt{x}}{\sqrt{u_0}} &\leq \exp\left(\frac{1}{2}+\frac{x-u_0}{N_1-1}+\frac{u_0}{N_1-1}\right)\sqrt{\frac{x}{u_0}} \\ &\leq e\exp\left(\frac{x-\log(6B+4)}{5}\right)\sqrt{\frac{x}{u_0}}.\end{aligned} \tag{V.78}$$

We can now bound the conditional expectation $\mathbb{E}\big[L_1(\widehat{\boldsymbol{\omega}})|\mathcal{X}^{(-1)}\big]$ separating the values before and after u_0:

$$\begin{aligned}\mathbb{E}\Big[L_1(\widehat{\boldsymbol{\omega}})|\mathcal{X}^{(-1)}\Big] &\leq \mathbb{E}\Big[\psi_s(\xi)P_s(\xi)|\mathcal{X}^{(-1)}\Big]\\ &= \mathbb{E}\Big[\psi_s(\xi)P_s(\xi)(1_{\xi+\log(6B+4)\leq u_0}+\mathbf{1}_{\xi+\log(6B+4)>u_0})|\mathcal{X}^{(-1)}\Big]\\ &\leq P_s(u_0-\log(6B+4))+\mathbb{E}\Big[\psi_s(\xi)P_s(\xi)\mathbf{1}_{\xi+\log(6B+4)>u_0}|\mathcal{X}^{(-1)}\Big]\\ &\leq P(u_0)+\mathbb{E}\left[e\exp(\xi/5)\sqrt{\frac{\xi+\log(6B+4)}{u_0}}P_s(\xi)\mathbf{1}_{\xi+\log(6B+4)>u_0}|\mathcal{X}^{(-1)}\right]. \quad \text{(V.79)}\end{aligned}$$

We have used that P (and P_s) is increasing on $\mathbb{R}_+$ (P is a polynomial with positive coefficients) and the bound (V.78). The second term in (V.79) can be upper bounded using Lemma V.36, as $\sqrt{\xi+\log(6B+4)}P_s(\xi)$ can be seen as a polynomial of degree 3 evaluated in $\sqrt{\xi+\log(6B+4)}$. We apply (V.82) to this polynomial with $a=\log(6B+4)$, $\delta=u_0-\log(6B+4)$, $\rho=1/5$, $d=3$ and $\gamma=1/2$. As $a\geq\log(10)\geq 2$ and $\delta\geq 1/2$, the condition required by Lemma V.82 is satisfied: $(\delta+a)(1-\rho)\geq 2\geq 3/2=\gamma d$. Then it holds:

$$\begin{aligned}&\mathbb{E}\left[\exp(\xi/5)\sqrt{\frac{\xi+\log(6B+4)}{u_0}}P_s(\xi)1_{\xi+\log(6B+4)>u_0}|\mathcal{X}^{(-1)}\right]\\ &\qquad\leq C\sqrt{\frac{u_0}{u_0}}P_s(u_0-\log(6B+4))e^{-(4/5)(u_0-\log(6B+4))}\leq CP(u_0)Be^{-u_0/2}. \quad \text{(V.80)}\end{aligned}$$

Combining (V.79) and (V.80) and replacing $P(u_0)$ by its value, we obtain:

$$\mathbb{E}\Big[L_1(\widehat{\boldsymbol{\omega}})|\mathcal{X}^{(-1)}\Big] \quad\leq\quad \mathbb{E}\Big[L_1(\boldsymbol{\omega})|\mathcal{X}^{(-1)}\Big](1+CBe^{-u_0/2})+C\sqrt{u_0}\sum_{k=2}^{B}\omega_k\sqrt{\frac{\mathfrak{q}_k}{N_1}}+Cs_1^2\frac{u_0}{\sqrt{d_1^\bullet}}.$$

Third step : unconditional bound. We now simply take the expectation with respect to $\mathcal{X}^{(-1)}$. From the previous bound, using Jensen's inequality, for all $\boldsymbol{\omega}\in\mathcal{S}_B$:

$$\mathbb{E}[L_1(\widehat{\boldsymbol{\omega}})]\leq\mathbb{E}[L_1(\boldsymbol{\omega})](1+CBe^{-u_0/2})+C\sqrt{u_0}\sum_{k=2}^{B}\omega_k\sqrt{\frac{\mathbb{E}[\mathfrak{q}_k]}{N_1}}+Cs_1^2\frac{u_0}{\sqrt{d_1^\bullet}}.$$

We obtain (V.33) as $\mathbb{E}[\mathfrak{q}_k]=q_k$. □

Lemma V.36. *Let $\xi\sim\mathcal{E}(1)$ be an exponential random variable, and ρ,a,δ be positive real numbers. Then for all $p\geq 0$ such that $p<(\delta+a)(1-\rho)$, it holds:*

$$\mathbb{E}\big[(\xi+a)^p e^{\rho\xi}1_{\xi\geq\delta}\big]\leq\left(1-\rho-\frac{p}{a+\delta}\right)^{-1}(\delta+a)^p e^{-\delta(1-\rho)}. \quad \text{(V.81)}$$

Let P a polynomial of degree d and $\gamma>0$ such that $\gamma d<(\delta+a)(1-\rho)$, then:

$$\mathbb{E}\big[P((\xi+a)^\gamma)e^{\rho\xi}1_{\xi\geq\delta}\big]\leq\left(1-\rho-\frac{d\gamma}{a+\delta}\right)^{-1}P((\delta+a)^\gamma)e^{-\delta(1-\rho)}. \quad \text{(V.82)}$$

Proof. As $p<(\delta+a)(1-\rho)$, then $p<(\delta+a)(1-\rho-\varepsilon)$ for all $\varepsilon<1-\rho-p/(a+\delta)$. The function $x\mapsto F(x):=(x+a)^p e^{(\rho-(1-\varepsilon))x}$ on $\mathbb{R}_+$ attains its maximum in $x_*:=p(1-\rho-\varepsilon)^{-1}-a$ and then decreases to 0. As $x_*<\delta$, we have $F(x)\leq F(\delta)$ for all $x\geq\delta$, thus:

$$\mathbb{E}\big[(\xi+a)^p e^{\rho\xi}1_{\xi\geq\delta}\big]=\mathbb{E}\Big[F(\xi)e^{(1-\varepsilon)\xi}1_{\xi\geq\delta}\Big]\leq F(\delta)\mathbb{E}\Big[e^{(1-\varepsilon)\xi}1_{\xi\geq\delta}\Big]=(\delta+a)^p e^{-(1-\rho)\delta}\varepsilon^{-1}.$$

As the inequality is true for all $\varepsilon<1-\rho-p/(a+\delta)$ we get (V.81). Equation (V.82) is obtained by applying (V.81) to each of the monomials of degree $k\leq d$ as $k\gamma\leq d\gamma<(\delta+a)(1-\rho)$, upper bounding the first factor and summing. □

Proofs of Corollary V.15 and Corollary V.13

Proof of Corollary V.15 According to Proposition V.12, for $B = 2$, $\mu_2 = 0$ and $\Sigma_2 = 0$; for all $\omega_1 \in (0,1)$:

$$R_1(\widehat{\omega}) \leq \big((1-\omega_1)^2\|\mu_1\|^2 + \omega_1 s_1^2 + 2(1-\omega_1)\eta\big)(1 + Ce^{-u_0/2}) + Cs_1^2\sqrt{\frac{u_0}{d_1^{\mathrm{e}}}}, ,$$

where $\eta = C\frac{\|\mu_1\|s_1}{\sqrt{d_1^{\mathrm{e}}}}\sqrt{u_0}$. Let us choose $\omega_1 = \min\Big(\frac{\|\mu_1\|^2+\eta}{\|\mu_1\|^2+s_1^2}, 1\Big)$. Then if $\eta \leq s_1^2$:

$$\begin{aligned}
R_1(\widehat{\omega}) &\leq (1 + Ce^{-u_0/2})\frac{\|\mu_1\|^2 s_1^2 + 2s_1^2\eta - \eta^2}{\|\mu_1\|^2 + s_1^2} + Cs_1^2\sqrt{\frac{u_0}{d_1^{\mathrm{e}}}} \\
&\leq (1 + Ce^{-u_0/2})\frac{\|\mu_1\|^2 s_1^2}{\|\mu_1\|^2 + s_1^2} + Cs_1^2\sqrt{\frac{u_0}{d_1^{\mathrm{e}}}}\frac{2\|\mu_1\|s_1}{\|\mu_1\|^2 + s_1^2} + Cs_1^2\sqrt{\frac{u_0}{d_1^{\mathrm{e}}}} \\
&\leq (1 + Ce^{-u_0/2})\frac{\|\mu_1\|^2 s_1^2}{\|\mu_1\|^2 + s_1^2} + Cs_1^2\sqrt{\frac{u_0}{d_1^{\mathrm{e}}}}\,,
\end{aligned}$$

where we have used that $2ab \leq a^2 + b^2$. Otherwise, if $\eta \geq s_1^2$:

$$\begin{aligned}
R_1(\widehat{\omega}) &\leq s_1^2(1 + Ce^{-u_0/2}) + Cs_1^2\sqrt{\frac{u_0}{d_1^{\mathrm{e}}}} \\
&\leq (1 + Ce^{-u_0/2})\frac{\|\mu_1\|^2 s_1^2}{\|\mu_1\|^2 + s_1^2} + (1 + Ce^{-u_0/2})\frac{s_1^4}{\|\mu_1\|^2 + s_1^2} + Cs_1^2\sqrt{\frac{u_0}{d_1^{\mathrm{e}}}}\,.
\end{aligned}$$

We conclude using that $s_1^2 \leq C\frac{\|\mu_1\|^2}{d_1^{\mathrm{e}}}u_0$ in this case.

Proof of Corollary V.13. Let $\tau \geq 0, \varsigma \geq 1$ be fixed. Let k be an element of $V_{\tau,\varsigma} = W_{(\varsigma)} \cap V_\tau$ with $k \neq 1$. We start by upper bounding q_k, with q_k defined in (V.34). Since $k \in W_{(\varsigma)}$, it holds $\operatorname{Tr}\Sigma_k^2 \leq \varsigma^2\frac{N_k^2}{N_1^2}\operatorname{Tr}\Sigma_1^2$, so that

$$\begin{aligned}
\operatorname{Tr}\Sigma_1\Sigma_k \leq \frac{1}{2}\left(\frac{N_k}{N_1}\operatorname{Tr}\Sigma_1^2 + \frac{N_1}{N_k}\operatorname{Tr}\Sigma_k^2\right) &\leq \frac{1+\varsigma^2}{2}\frac{N_k}{N_1}\operatorname{Tr}\Sigma_1^2 \\
&\leq \frac{N_k}{N_1}\frac{(1+\varsigma^2)(\operatorname{Tr}\Sigma_1)^2}{2d_1^\bullet} \\
&= N_k N_1\frac{\varsigma^2 s_1^4}{d_1^\bullet}.
\end{aligned}$$

Since $k \in V_\tau$, it holds

$$\frac{\Delta_k^T\Sigma_1\Delta_k}{N_1} \leq \frac{\|\Sigma_1\|_\infty}{N_1}\|\Delta_k\|^2 \leq \frac{\operatorname{Tr}\Sigma_1}{N_1}\frac{1}{d_1^{\mathrm{e}}}\tau s_1^2 = \frac{\tau s_1^4}{d_1^{\mathrm{e}}}.$$

Joining these estimates, we get

$$\frac{q_k}{N_1} \leq \frac{\Delta_k^T\Sigma_1\Delta_k}{N_1} + \frac{\operatorname{Tr}\Sigma_1\Sigma_k}{N_1 N_k} \leq s_1^4\left(\frac{\tau}{d_1^{\mathrm{e}}} + \frac{\varsigma^2}{d_1^\bullet}\right).$$

Therefore, for $\boldsymbol{\omega}$ a vector of the simplex $\mathcal{S}_B$ having support in $W^{(\varsigma)} \cap V_\tau$, using $d_1^{\mathrm{e}} \leq d_1^\bullet$ it holds

$$Q_1(\boldsymbol{\omega}) = \sum_{k\geq 2}\omega_k\sqrt{\frac{q_k}{N_1}} \leq (1-\omega_1)\sqrt{\tau+\varsigma^2}\frac{s_1^2}{\sqrt{d_1^{\mathrm{e}}}}. \tag{V.83}$$

We now choose the weight vector $\boldsymbol{\omega}^* = \boldsymbol{\omega}^*_{V_{\tau,c}}$ given by the oracle weights of (V.13), for the set $V = V_{\tau,\varsigma}$. From Lemma V.6, this gives rise to $R_1(\boldsymbol{\omega}^*) \le \mathcal{B}(\tau,\nu)$, where $\nu = \nu(V_{\tau,\varsigma})$; furthermore we have the explicit expression

$$(1-\omega_1^*) = \lambda(1-\nu), \qquad \text{where } \lambda = \frac{1}{1+\tau(1-\nu)},$$

so that it holds (since $\nu \in [0,1]$)

$$(1-\omega_1^*)\sqrt{\tau} = \frac{(1-\nu)\sqrt{\tau}}{1+\tau(1-\nu)} \le \max\left(\frac{\tau(1-\nu)}{1+\tau(1-\nu)}, \frac{\sqrt{\tau}(1-\nu)}{1+\sqrt{\tau}(1-\nu)}\right) \le 1.$$

Plugging this into (V.83), we get $Q_1(\boldsymbol{\omega}^*) \le 2\varsigma s_1^2/\sqrt{d_1^{\mathrm{e}}}$, then (V.35) since the obtained estimate holds for any $\tau \ge 0, \varsigma \ge 1$.

Proof of Proposition V.16 We follow the same general canvas as in the proof of Proposition V.12.

First step : bound in conditional probability. Let us recall the definitions of $Q^{\mathbf{BS}}(\boldsymbol{\omega})$ and $\widehat{q}_k$:

$$\widehat{Q}^{\mathbf{BS}}(\boldsymbol{\omega}) := \frac{M}{N_1}\sum_{k=2}^{B}\omega_k\|\widehat{\mu}_k^{\mathtt{NE}} - \widehat{\mu}_1^{\mathtt{NE}}\|, \quad \widehat{q}_k = \frac{1}{N_1-1}\sum_{p=1}^{N_1}\left\langle \widehat{\mu}_k^{\mathtt{NE}} - \widehat{\mu}_1^{\mathtt{NE}}, X_p^{(1)} - \widehat{\mu}_1^{\mathtt{NE}}\right\rangle^2.$$

We will need the following quantity $\widehat{q}'_k$ which is close to $\widehat{q}_k$ but easier to control:

$$\widehat{q}'_k = \frac{1}{N_1-1}\sum_{p=1}^{N_1}\left\langle \widehat{\mu}_k^{\mathtt{NE}} - \mu_1, X_p^{(1)} - \widehat{\mu}_1^{\mathtt{NE}}\right\rangle^2.$$

The estimated weight vector $\widehat{\boldsymbol{\omega}}$ for the estimation of μ_1 is chosen as

$$\widehat{\boldsymbol{\omega}} \in \operatorname*{Arg\,Min}_{\boldsymbol{\omega}\in\mathcal{S}_B}\left(\widehat{L}_1(\boldsymbol{\omega}) + 4\sqrt{2u_0}\widehat{Q}_1(\boldsymbol{\omega}) + 1424u_0\widehat{Q}(\boldsymbol{\omega})^{\mathbf{BS}}\right).$$

Let $u := u_0 - \log B$, and define the events:

$$A_1 = \left\{\|\widehat{\mu}_k^{\mathtt{NE}} - \mu_1\|_{\Sigma_1} \le 2\sqrt{\widehat{q}'_k} + 711\frac{\|\widehat{\mu}_k^{\mathtt{NE}} - \mu_1\|M}{\sqrt{N_1}}(u+\log B), 2 \le k \le B\right\},$$

$$A_2 = \left\{\left|\|\widehat{\mu}_1^{\mathtt{NE}} - \mu_1\|^2 - \widehat{s}_1^2\right| \le C\frac{s_1^2}{\sqrt{d_1^{\bullet}}}u + C\frac{M^2}{N_1^2}u^2\right\},$$

and

$$A_3 = \left\{\langle \widehat{\mu}_k^{\mathtt{NE}} - \mu_1, \widehat{\mu}_1^{\mathtt{NE}} - \mu_1\rangle \le \sqrt{2\frac{u+\log B}{N_1}}\|\widehat{\mu}_k^{\mathtt{NE}} - \mu_1\|_{\Sigma_1} + \frac{2\|\widehat{\mu}_k^{\mathtt{NE}} - \mu_1\|M}{3N_1}(u+\log B), 2 \le k \le B\right\},$$

where we recall that for ν a vector and Σ an operator, $\|\nu\|_\Sigma^2 := \langle \nu, \Sigma\nu\rangle$. For $i \in \{1,3\}$, $\mathbb{P}\big[A_i|\mathcal{X}^{(-1)}\big] \ge 1 - e^{-u}$ and $\mathbb{P}\big[A_2|\mathcal{X}^{(-1)}\big] \ge 1 - 2e^{-u}$ because of Proposition V.38 for A_1, Lemma V.42 for A_3 and for A_2, because $\|\widehat{\mu}_1^{\mathtt{NE}} - \mu_1\|^2 - \widehat{s}_1^2$ is a U-statistic:

$$\|\widehat{\mu}_1^{\mathtt{NE}} - \mu_1\|^2 - \widehat{s}_1^2 = \frac{1}{N_1(N_1-1)}\sum_{\ell\neq p=1}^{N_1}\left\langle X_\ell^{(1)} - \mu_1, X_p^{(1)} - \mu_1\right\rangle, \tag{V.84}$$

the concentration is a direct consequence of Houdré and Reynaud-Bouret (2003) (or see Proposition IV.9). Then the event $A = A_1 \cap A_2 \cap A_3$ conditionally to $\mathcal{X}^{(-1)}$ is of probability greater than $1 - 4e^{-u}$.
The differences between respectively $\widehat{q}_k$ and $\widehat{q}'_k$ for $k \in [\![B]\!]$ can be bounded independently of k :

$$\left|\sqrt{\widehat{q}_k} - \sqrt{\widehat{q}'_k}\right| \le \sqrt{\frac{1}{N_1(N_1-1)}\sum_{p=1}^{N_1}\langle \widehat{\mu}_1^{\mathtt{NE}} - \mu_1, X_p - \widehat{\mu}_1^{\mathtt{NE}}\rangle^2} =: \Delta q. \tag{V.85}$$

Assume A, then:

$$L_1(\widehat{\boldsymbol{\omega}}) = \widehat{L}_1(\widehat{\boldsymbol{\omega}}) + 2\sum_{k=2}^{B}\widehat{\omega}_k\langle\widehat{\mu}_k^{\mathtt{NE}} - \mu_1, \widehat{\mu}_1^{\mathtt{NE}} - \mu_1\rangle + (2\widehat{\omega}_1 - 1)\big(\|\widehat{\mu}_1^{\mathtt{NE}} - \mu_1\|^2 - \widehat{s}_1^2\big)$$
$$\leq \widehat{L}_1(\widehat{\boldsymbol{\omega}}) + 2\sum_{k=2}^{B}\widehat{\omega}_k\left(\sqrt{2\frac{u+\log B}{N_1}}\|\widehat{\mu}_k^{\mathtt{NE}} - \mu_1\|_{\Sigma_1} + \frac{2\|\widehat{\mu}_k^{\mathtt{NE}} - \mu_1\|M}{3N_1}(u+\log B)\right)$$
$$+ \frac{Cs_1^2}{\sqrt{d_1^\bullet}}u + \frac{CM^2}{N_1^2}u^2,$$

where we have used the events A_2 and A_3. Then using the event A_1, the bound (V.85) and a triangle inequality we get:

$$L_1(\widehat{\boldsymbol{\omega}}) \leq \widehat{L}_1(\widehat{\boldsymbol{\omega}}) + 4\sqrt{\log B + u}\sum_{k=2}^{B}\widehat{\omega}_k\sqrt{\frac{2\widehat{q}_k}{N_1}} + 1424(\log B + u)\sum_{k=2}^{B}\widehat{\omega}_k\frac{M\|\widehat{\mu}_k^{\mathtt{NE}} - \widehat{\mu}_1^{\mathtt{NE}}\|}{N_1}$$
$$+ C\frac{\Delta q}{\sqrt{N_1}}\sqrt{\log B + u} + C\frac{\|\widehat{\mu}_1^{\mathtt{NE}} - \mu_1\|M}{N_1}(\log B + u) + \frac{Cs_1^2}{\sqrt{d_1^\bullet}}u + \frac{CM^2}{N_1^2}u^2.$$

Using the choice of $\widehat{\boldsymbol{\omega}}$, conditionally to A:

$$L_1(\widehat{\boldsymbol{\omega}}) \leq \min_{\boldsymbol{\omega}\in\mathcal{S}_B}\left(\widehat{L}_1(\boldsymbol{\omega}) + 4\sqrt{2u_0}\widehat{Q}(\boldsymbol{\omega}) + 1424u_0\widehat{Q}^{\mathbf{BS}}(\boldsymbol{\omega})\right)$$
$$+ C\frac{\Delta q}{\sqrt{N_1}}\sqrt{\log B + u} + C\frac{\|\widehat{\mu}_1^{\mathtt{NE}} - \mu_1\|M}{N_1}(\log B + u) + \frac{Cs_1^2}{\sqrt{d_1^\bullet}}u + \frac{CM^2}{N_1^2}u^2.$$

Second and third steps: bound in expectation. Let us bound some expectation using Jensen's inequality:

$$\mathbb{E}\left[\sqrt{\widehat{q}_k'}\right] \leq \sqrt{\|\mu_k - \mu_1\|_{\Sigma_1}^2 + \frac{\mathrm{Tr}(\Sigma_1\Sigma_k)}{N_k}}\,, \quad \mathbb{E}[\Delta q] \leq \frac{M\sqrt{\mathrm{Tr}\,\Sigma_1}}{N_1} + \frac{\sqrt{\mathrm{Tr}\,\Sigma_1^2}}{\sqrt{N_1}}. \tag{V.86}$$

The expectation of $\sqrt{\widehat{q}_k}$ can be bounded using that $\sqrt{\widehat{q}_k} \leq \sqrt{\widehat{q}_k'} + \Delta q$. We can now bound the risk. Let $\boldsymbol{\omega} \in \mathcal{S}_B$:

$$R_1(\widehat{\boldsymbol{\omega}}) \leq \mathbb{E}[L_1(\widehat{\boldsymbol{\omega}})1_A] + M^2\mathbb{P}[A^c]$$
$$\leq L_1(\boldsymbol{\omega}) + 4\sqrt{2u_0}\sum_{k=2}^{B}\omega_k\frac{\mathbb{E}\left[\sqrt{\widehat{q}_k}\right]}{\sqrt{N_1}} + 1424u_0\sum_{k=2}^{B}\omega_k\frac{M(\|\mu_k - \mu_1\| + s_1 + s_k)}{N_1}$$
$$+ C\frac{\mathbb{E}[\Delta q]}{\sqrt{N_1}}\sqrt{\log B + u} + C\frac{s_1 M}{N_1}(\log B + u) + C\frac{s_1^2}{\sqrt{d_1^\bullet}}u + C\frac{M^2}{N_1^2}u^2 + 3M^2e^{-u}$$

Because $u \geq 2\log N_1$, the last term is upper bounded by the previous one. Using (V.86) and by bringing together the terms:

$$R_1(\widehat{\boldsymbol{\omega}}) \leq R_1(\boldsymbol{\omega}) + 4\sqrt{2(\log B + u)}Q(\boldsymbol{\omega}) + 1424(\log B + u)\sum_{k=2}^{B}\omega_k\frac{M(\|\mu_k - \mu_1\| + s_k)}{N_1}$$
$$+ C\frac{s_1^2}{\sqrt{d_1^\bullet}}(u + \sqrt{\log B + u}) + C\frac{Ms_1}{N_1}(\log B + u) + C\frac{M^2}{N_1^2}u^2\,, \tag{V.87}$$

where Q is defined in (V.34). Let $\tau, \varsigma > 0$ and $\boldsymbol{\omega}^* = \boldsymbol{\omega}^*_{V_{\tau,\varsigma}}$ be defined as in (V.13). Then as in the proof of Corollary V.13:

$$R_1(\boldsymbol{\omega}^*) = s_1^2\mathcal{B}(\tau, \nu(V_{\tau,\varsigma}))\,, \quad Q(\boldsymbol{\omega}^*) \leq C\sqrt{\frac{1+\varsigma^2}{d_1^{\mathrm{e}}}}s_1^2. \tag{V.88}$$

Up to bound the third term in the upper bound (V.87), let us bound s_k^2 for $k \in V_{\tau,\varsigma}$. On the one hand:

$$s_k^2 = \frac{\operatorname{Tr}\Sigma_k}{N_k} \leq \frac{4M^2}{N_k} = 4\operatorname{Tr}\Sigma_1 \frac{\phi_1}{N_k} = 4s_1^2 \frac{\phi_1 N_1}{N_k}\,.$$

On the other hand, as $k \in V_{\tau,\varsigma} \subset W_{(\varsigma)}$:

$$s_k^2 = \frac{\operatorname{Tr}\Sigma_k}{N_k} = \sqrt{d_k^\bullet}\frac{\sqrt{\operatorname{Tr}\Sigma_k^2}}{N_k} \leq \sqrt{d_k^\bullet}\varsigma\frac{\sqrt{\operatorname{Tr}\Sigma_1^2}}{N_1} = s_1^2\varsigma\sqrt{\frac{d_k^\bullet}{d_1^\bullet}}\,.$$

Combining these two bounds:

$$s_k^2 \leq 4s_1^2 \min\left(\frac{\phi_1 N_1}{N_k}, \varsigma\sqrt{\frac{d_k^\bullet}{d_1^\bullet}}\right).$$

As we assume $N_k \geq (d_k^\bullet)^\beta$, for $k \in V_{\tau,\varsigma}$:

$$s_k^2 \leq 4s_1^2 \min\left(\frac{\phi_1 N_1}{(d_k^\bullet)^\beta}, \varsigma\sqrt{\frac{d_k^\bullet}{d_1^\bullet}}\right) \leq 4s_1^2 \max_{d\geq 1} \min\left(\frac{\phi_1 N_1}{d^\beta}, \varsigma\sqrt{\frac{d}{d_1^\bullet}}\right) = 4s_1^2(\phi_1 N_1)^{\frac{1}{1+2\beta}}\left(\frac{\varsigma}{\sqrt{d_1^\bullet}}\right)^{\frac{2\beta}{1+2\beta}}.$$

We can now bound the third term in (V.87). As $\omega_k^* = 0$ for $k \notin V_{\tau,\varsigma}$:

$$\begin{aligned}\sum_{k=2}^B \omega_k^* \frac{M(\|\mu_k - \mu_1\| + s_k)}{N_1} &\leq \frac{M}{N_1}(1-\omega_1^*)\left(\sqrt{\tau}s_1 + 2s_1(\phi_1 N_1)^{\frac{1}{2(1+2\beta)}}\left(\frac{\varsigma}{\sqrt{d_1^\bullet}}\right)^{\frac{\beta}{1+2\beta}}\right)\\ &\leq s_1^2\left((1-\omega_1^*)\sqrt{\frac{\tau\phi_1}{N_1}} + 2\phi_1^{\frac{1+\beta}{1+2\beta}}\left(\frac{\varsigma}{N_1\sqrt{d_1^\bullet}}\right)^{\frac{\beta}{1+2\beta}}\right).\end{aligned}$$

As $N_1 \geq (d_1^\bullet)^\beta$ and $(1-\omega_1^*)\sqrt{\tau} \leq 1$ (by definition of ω_1^*), we get:

$$\sum_{k=2}^B \omega_k^* \frac{M(\|\mu_k - \mu_1\| + s_k)}{N_1} \leq 2s_1^2\left(\frac{\sqrt{\phi_1}}{(d_1^\bullet)^{\beta/2}} + \frac{\phi_1^{\frac{1+\beta}{1+2\beta}}\varsigma^{\frac{\beta}{1+2\beta}}}{(d_1^\bullet)^{\beta/2}}\right). \tag{V.89}$$

Injecting the bounds (V.88) and (V.89) into (V.87) leads to:

$$\begin{aligned}\frac{R_1(\widehat{\boldsymbol{\omega}})}{s_1^2} &\leq \min_{\tau>0,\varsigma>0}\left(\mathcal{B}(\tau,\nu(V_{\tau,\varsigma})) + C\varsigma\sqrt{\frac{u_0}{d_1^{\mathrm{e}}}} + Cu_0\frac{\phi_1^{\frac{1+\beta}{1+2\beta}}\varsigma^{\frac{\beta}{1+2\beta}}}{(d_1^\bullet)^{\beta/2}}\right)\\ &\quad + Cu_0\frac{\sqrt{\phi_1}}{(d_1^\bullet)^{\beta/2}} + C\sqrt{\frac{u_0}{d_1^{\mathrm{e}}}} + C\frac{u_0}{\sqrt{d_1^\bullet}} + C\frac{u_0\sqrt{\phi_1}}{\sqrt{N_1}} + C\frac{\phi_1 u^2}{N_1}\\ &\leq \min_{\tau>0,\varsigma>0}\left(\mathcal{B}(\tau,\nu(V_{\tau,\varsigma})) + C\varsigma\sqrt{\frac{u_0}{d_1^{\mathrm{e}}}} + Cu_0\frac{\phi_1^{\frac{1+\beta}{1+2\beta}}\varsigma^{\frac{\beta}{1+2\beta}}}{(d_1^\bullet)^{\beta/2}}\right) + C\sqrt{\frac{u_0}{d_1^{\mathrm{e}}}} + C\frac{u_0\phi_1}{(d_1^\bullet)^{\beta/2}}.\end{aligned}$$

As $\phi_1^{\frac{1+\beta}{1+2\beta}}\varsigma^{\frac{\beta}{1+2\beta}} \leq \max(\phi_1,\varsigma) \leq \phi_1 + \varsigma$, we obtain:

$$\frac{R_1(\widehat{\boldsymbol{\omega}})}{s_1^2} \leq \min_{\tau>0,\varsigma>0}\left(\mathcal{B}(\tau,\nu(V_{\tau,\varsigma})) + C\varsigma\max\left(\sqrt{\frac{u_0}{d_1^{\mathrm{e}}}}, \frac{u_0}{(d_1^\bullet)^{\beta/2}}\right)\right) + C\sqrt{\frac{u_0}{d_1^{\mathrm{e}}}} + C\frac{u_0\phi_1}{(d_1^\bullet)^{\beta/2}}.$$

□

V.9.6 Concentration inequalities

Concentration for $\widehat{q}$. Consider first the Gaussian setting (**GS**).

Proposition V.37. *Let $X_1, \dots, X_N$ i.i.d. Gaussian random vectors of distribution $\mathcal{N}(\mu_1, \Sigma_1)$ and $\nu \in \mathbb{R}^d$. Let $\widehat{q} = \frac{1}{N-1}\sum_{k=1}^N \langle \widehat{\mu}_1^{\mathtt{NE}} - \nu, X_k - \widehat{\mu}_1^{\mathtt{NE}}\rangle^2$, then for all $x \geq 0$:*

$$\mathbb{P}\left[\sqrt{\widehat{q}} \geq \left(1 + \sqrt{\frac{2x}{N-1}}\right)\left(\sqrt{\|\mu_1 - \nu\|_{\Sigma_1}^2 + \frac{\operatorname{Tr}\Sigma_1^2}{N}} + \|\Sigma_1\|_\infty \sqrt{\frac{2x}{N}}\right)\right] \leq 2e^{-x}\,, \tag{V.90}$$

and

$$\mathbb{P}\left[\sqrt{\widehat{q}} \leq e^{-1/2-x/(N-1)}\left(\sqrt{\|\mu_1 - \nu\|_{\Sigma_1}^2 + \frac{\operatorname{Tr}\Sigma_1^2}{N}} - 2\|\Sigma_1\|_\infty \sqrt{\frac{2x}{N}}\right)\right] \leq 2e^{-x}\,, \tag{V.91}$$

where $\|\mu_1 - \nu\|_{\Sigma_1}^2 = (\mu_1 - \nu)^T \Sigma_1 (\mu_1 - \nu)$.

Proof. Let us consider the random vector $Z \in \mathbb{R}^N$ with $Z_k = \langle \widehat{\mu}_1^{\mathtt{NE}} - \nu, X_k - \widehat{\mu}_1^{\mathtt{NE}}\rangle$, then $\widehat{q} = \|Z\|_N^2/(N-1)$, where $\|\cdot\|_N$ is the Euclidian norm in $\mathbb{R}^N$. Conditionally to $\widehat{\mu}_1^{\mathtt{NE}}$, Z is a Gaussian vector of distribution $\mathcal{N}(0, e(\widehat{\mu}_1^{\mathtt{NE}})\Gamma)$, where $e(\widehat{\mu}_1^{\mathtt{NE}}) = (\widehat{\mu}_1^{\mathtt{NE}} - \nu)^T \Sigma_1 (\widehat{\mu}_1^{\mathtt{NE}} - \nu)$ and $\Gamma = I_N - \mathbf{1}_N \mathbf{1}_N^T/N$ with $\mathbf{1}_N = (1, \dots, 1) \in \mathbb{R}^N$. The eigenvalues of Γ are 1 with multiplicity $N-1$ and 0. So $\|Z\|^2/e(\widehat{\mu}_1^{\mathtt{NE}})$ has a $\chi^2(N-1)$ distribution. Then conditionally to $\widehat{\mu}_1^{\mathtt{NE}}$:

$$\widehat{q} = \frac{\|Z\|^2}{N-1} \sim \frac{e(\widehat{\mu}_1^{\mathtt{NE}})}{N-1}\chi^2(N-1)\,.$$

Then according to Lemma V.40 and Lemma V.41, for all $x \geq 0$:

$$\mathbb{P}\left[\sqrt{\frac{\widehat{q}}{e(\widehat{\mu}_1^{\mathtt{NE}})}} \geq 1 + \sqrt{\frac{2x}{N-1}}\middle|\widehat{\mu}_1^{\mathtt{NE}}\right] \leq e^{-x}\,, \quad \mathbb{P}\left[\sqrt{\frac{\widehat{q}}{e(\widehat{\mu}_1^{\mathtt{NE}})}} \leq e^{-1/2}e^{-x/(N-1)}\middle|\widehat{\mu}_1^{\mathtt{NE}}\right] \leq e^{-x}.$$

Let $g = \Sigma_1^{1/2}(\widehat{\mu}^{\mathtt{NE}} - \nu) \sim \mathcal{N}(\Sigma_1^{1/2}(\mu_1 - \nu), \Sigma_1^2/N)$, as $\|g\|^2 = e(\widehat{\mu}_1^{\mathtt{NE}})$, from Lemma V.40 with $\Sigma_1^{1/2}(\mu_1 - \nu) \to \mu$ and $\Sigma_1^2/N \to \Sigma$, we get that for all $x \geq 0$:

$$\mathbb{P}\left[\sqrt{e(\widehat{\mu}_1^{\mathtt{NE}})} \geq \sqrt{(\mu_1 - \nu)^T\Sigma_1(\mu_1 - \nu) + \frac{\operatorname{Tr}\Sigma_1^2}{N}} + \|\Sigma_1\|_\infty\sqrt{\frac{2x}{N}}\right] \leq e^{-x}\,,$$
$$\mathbb{P}\left[\sqrt{e(\widehat{\mu}_1^{\mathtt{NE}})} \leq \sqrt{(\mu_1 - \nu)^T\Sigma_1(\mu_1 - \nu) + \frac{\operatorname{Tr}\Sigma_1^2}{N}} - 2\|\Sigma_1\|_\infty\sqrt{\frac{2x}{N}}\right] \leq e^{-x}\,.$$

We have used that for all $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$ and $x \geq 0$:

$$\left(\sqrt{\|\mu\|^2 + \operatorname{Tr}\Sigma} + \sqrt{2\|\Sigma\|_\infty x}\right)^2 \geq \left(\|\mu\|^2 + \operatorname{Tr}\Sigma\right) + 2\sqrt{(\operatorname{Tr}\Sigma^2 + 2\mu^T\Sigma\mu)x} + 2\|\Sigma\|_\infty x\,,$$
$$\left(\sqrt{\|\mu\|^2 + \operatorname{Tr}\Sigma} - 2\sqrt{2\|\Sigma\|_\infty x}\right)_+^2 \leq \left(\left(\|\mu\|^2 + \operatorname{Tr}\Sigma\right) - 2\sqrt{(\operatorname{Tr}\Sigma^2 + 2\mu^T\Sigma\mu)x}\right)_+\,,$$

as $(a-b)_+^2 \leq (a^2 - ab)_+$ for $a, b > 0$.
Equations (V.90) and (V.91) are obtained by combining these concentration inequalities. □

In the bounded setting (**BS**), Proposition V.38 gives a concentration bound for $\widehat{q}'$, which is a slightly different statistic from $\widehat{q}$ because we consider $\mu_1 - \nu$ known for $\widehat{q}'$.

Proposition V.38. *Assume (**BS**), let $\nu \in \mathbb{R}^d$ and $\widehat{q}' = \frac{1}{N-1}\sum_{k=1}^N \langle \mu_1 - \nu, X_k - \widehat{\mu}_1^{\mathtt{NE}}\rangle^2$. Then for all $u \geq 1$:*

$$\mathbb{P}\left[2\sqrt{\widehat{q}'} \leq \sqrt{(\mu_1 - \nu)\Sigma_1(\mu_1 - \nu)} - 711\frac{\|\mu_1 - \nu\|M}{\sqrt{N-1}}u\right] \leq e^{-u}.$$

Proof. Let us first denote $\delta := \mu_1 - \nu$ and $Z' := \sqrt{\widehat{q}}$. We are going to use Talagrand's inequality (Theorem V.43). So let us first rewrite Z':

$$\begin{aligned} Z' &= \sup_{\|v\|_N=1} \frac{1}{\sqrt{N-1}} \sum_{k=1}^N v_k \langle \delta, X_k - \widehat{\mu}_1^{\mathtt{NE}} \rangle \\ &= \sup_{\|v\|_N=1} \frac{1}{\sqrt{N-1}} \sum_{k=1}^N \langle \delta, X_k - \mu_1 \rangle \left(v_k - \frac{1}{N} \sum_{q=1}^N v_q \right). \end{aligned}$$

Let $T = \{v \in \mathbb{R}^N, \|v\|_N = 1\}$ (or a countable dense subset) and define for $v \in T$:

$$X_k^v := \frac{1}{\sqrt{N-1}} \langle \delta, X_k - \mu_1 \rangle \left(v_k - \frac{1}{N} \sum_{q=1}^N v_q \right),$$

then:

$$|X_k^v| \le \frac{2\|\delta\| M}{\sqrt{N-1}}, \quad \sup_{v \in T} \sum_{k=1}^N \mathbb{E}\big[(X_k^v)^2\big] \le \frac{\delta^T \Sigma \delta}{N-1} \le \frac{4\|\delta\|^2 M^2}{N-1}.$$

Using Theorem V.43, with probability greater than $1 - e^{-u}$, $u \ge 1$:

$$Z' \ge \mathbb{E}[Z'](1-\varepsilon) - C(\varepsilon) \frac{\|\delta\| M}{\sqrt{N-1}} u,$$

where $C(\varepsilon) = 8(2 + \varepsilon^{-1})$ for some $\varepsilon > 0$. We just need to lower bound $\mathbb{E}[Z']$ by $\sqrt{\mathbb{E}[(Z')^2]} = \sqrt{\delta^T \Sigma_1 \delta}$. For that, using again Talagrand's inequality, it exists an exponential random variable $\xi \sim \mathcal{E}(1)$ such that:

$$Z' \le \mathbb{E}[Z'](1+\varepsilon) + C(\varepsilon) \frac{\|\delta\| M}{\sqrt{N-1}} \xi$$

Then:

$$\begin{aligned} \mathbb{E}\big[(Z')^2\big] &\le \mathbb{E}\left[\left(\mathbb{E}[Z'](1+\varepsilon) + C(\varepsilon) \frac{\|\delta\| M}{\sqrt{N-1}} \xi \right)^2 \right] \\ &\le \left(\mathbb{E}[Z'](1+\varepsilon) + \sqrt{2} C(\varepsilon) \frac{\|\delta\| M}{\sqrt{N-1}} \right)^2, \end{aligned}$$

and we get that $(1+\varepsilon)\mathbb{E}[Z'] \ge \sqrt{\mathbb{E}[(Z')^2]} - \sqrt{2} C(\varepsilon) \frac{\|\delta\| M}{\sqrt{N-1}}$. Putting together the two bounds, we get a first lower bound for Z': for $u \ge 1$ and probability greater than $1 - e^{-u}$:

$$Z' \ge \sqrt{\delta^T \Sigma_1 \delta} \frac{1-\varepsilon}{1+\varepsilon} - C(\varepsilon) \left(\sqrt{2} \frac{1-\varepsilon}{1+\varepsilon} + 1 \right) \frac{\|\delta\| M}{\sqrt{N-1}} u. \tag{V.92}$$

Let us choose $\varepsilon = 1/3$ to conclude. □

Classical concentration inequalities for Gaussian random variables.

Lemma V.39. *Let $X \sim \mathcal{N}(m, \sigma^2)$, then for all $x \ge 0$:*

$$\mathbb{P}\Big[|X - m| \ge \sqrt{2\sigma^2 x} \Big] \le 2e^{-x}$$

Proof. It is a direct consequence of the Chernoff bound (Chernoff, 1952). □

Lemma V.40. *[Concentration of Gaussian vectors] Let $X \sim \mathcal{N}(\mu, \Sigma)$, then for all $x \geq 0$:*

$$\mathbb{P}\Big[\|X\|^2 \geq \Big(\|\mu\|^2 + \operatorname{Tr}\Sigma\Big) + 2\sqrt{(\operatorname{Tr}\Sigma^2 + 2\mu^T\Sigma\mu)x} + 2\|\Sigma\|_\infty x\Big] \leq e^{-x}\,,$$

$$\mathbb{P}\Big[\|X\|^2 \leq \Big(\|\mu\|^2 + \operatorname{Tr}\Sigma\Big) - 2\sqrt{(\operatorname{Tr}\Sigma^2 + 2\mu^T\Sigma\mu)x}\Big] \leq e^{-x}\,,$$

The above is a reformulation of Lemma 2 in Laurent et al. (2012) and can be seen as a consequence of combining the arguments of Lemma 1 of Laurent and Massart (2000) and Lemma 8.1 of Birgé (2001).

Lemma V.41. *[Lower bound for χ^2] Let $Z \sim \chi^2(n)$, then for all $x \geq 0$:*

$$\mathbb{P}\Big[Z \leq ne^{-(1+2x/n)}\Big] \leq e^{-x}\,.$$

Moreover, for all $x \geq 0$:

$$\mathbb{P}\Big[Z \leq ne^{-2(\sqrt{x/n}+x/n)}\Big] \leq e^{-x}\,.$$

Proof. Let $\delta \in (0,1)$, $\lambda \in \mathbb{R}_+$:

$$\mathbb{P}[Z \leq n\delta] = \mathbb{P}\big[e^{-\lambda Z} \geq e^{-n\lambda\delta}\big] \leq \mathbb{E}\big[e^{-\lambda Z}\big]e^{n\lambda\delta} = \exp\Big(-\frac{n}{2}(\log(1+2\lambda) - 2\lambda\delta)\Big)$$

where the inequality is due to Markov. Fix $\lambda = (-1+\delta^{-1})/2 > 0$, then:

$$\mathbb{P}[Z \leq n\delta] \leq \exp\Big(-\frac{n}{2}(-\log(\delta) + \delta - 1)\Big) \leq \exp\Big(-\frac{n}{2}(-\log(\delta) - 1)\Big)$$

Let us choose $\delta = \exp(-1-2x/n)$ to obtain the first concentration bound. For the second one, we can choose $\delta = \exp\Big(-2\sqrt{x/n} - 2x/n\Big)$ and then:

$$-\log(\delta) + \delta - 1 = 2\sqrt{x/n} + 2x/n + \exp\Big(-2\sqrt{x/n} - 2x/n\Big) - 1 \geq 2x/n\,.$$

The inequality is trivially verified for $2\sqrt{x/n} \geq 1$ and otherwise we can use that $-u - u^2/2 \geq \log(1-u)$ with $u = 2\sqrt{x/n}$. □

Classical concentration inequalities for bounded random variables.

Lemma V.42. *[Bernstein's concentration inequality] Let $X_1, \ldots, X_N$ i.i.d. real centred random variables bounded by M such that $\mathbb{E}\big[X_1^2\big] \leq \sigma^2$, then for all $x \geq 0$:*

$$\mathbb{P}\left[\sum_{i=1}^N X_i \geq \sqrt{2N\sigma^2 x} + \frac{2Mx}{3}\right] \leq e^{-x}$$

Proof. See for instance Vershynin (2018), Exercise 2.8.5. □

Theorem V.43. *[Talagrand's inequality] Let $X_1^t, ..., X_n^t$ independant random variables indexed by $t \in T$ (T countable) in $\mathbb{R}$ and $L > 0$ such that for all $t \in T$, $i \leq n$,*

$$\mathbb{E}\big[X_i^t\big] = 0\,, \quad |X_i^t| \leq L \tag{V.93}$$

Let

$$Z := \sup_{t\in T}\sum_{i=1}^n X_i^t\,, \quad \sigma^2 = \sup_{t\in T}\sum_{i=1}^n \mathbb{E}\big[(X_i^t)^2\big]$$

then for all $x \geq 0$ and $\varepsilon \in (0,1)$:

$$\mathbb{P}\Big[Z \geq \mathbb{E}[Z](1+\varepsilon) + 2\sqrt{2\sigma^2 x} + 2Lx(1+8\varepsilon^{-1})\Big] \leq e^{-x} \tag{V.94}$$

$$\mathbb{P}\Big[Z \leq \mathbb{E}[Z](1-\varepsilon) - 2\sqrt{4\sigma^2 x} - 4Lx(1+8\varepsilon^{-1})\Big] \leq e^{-x} \tag{V.95}$$

Proof. See for instance Massart (2000). □

V.9.7 Proofs for Section V.5

Proof of Proposition V.18 This proof follows the same scheme as the Pinsker's bound (Pinsker, 1980 or see Tsybakov, 2008 for a recent version).

The proof is provided for $V = B$ but can be directly adapted for $V < B$ by assuming μ_k independent of μ_1 for $k > V$ when constructing the distribution $\mathbb{Q}$ (V.97).

Let us first restrict ourselves to the case where μ_1 is in a ball around 0:

$$\inf_{\widehat{\mu}_1} \sup_{\mu_i \in B(\mu_1, \sqrt{\tau} s_1)} R_1(\widehat{\mu}_1) \geq \inf_{\widehat{\mu}_1} \sup_{\substack{\mu_1 \in B(0, \sqrt{\beta} s_1) \\ \mu_i \in B(\mu_1, \sqrt{\tau} s_1)}} R_1(\widehat{\mu}_1).$$

Then the infimum over the estimators is now attained for an estimator $\widehat{\mu}_1$ bounded by $2\sqrt{\beta} s_1$. Indeed, any estimator $\widehat{\mu}$ further perform less well than the deterministic estimator $\widehat{\mu} = 0$. If $\|\widehat{\mu}\| > 2\sqrt{\beta} s_1$:

$$\|\widehat{\mu} - \mu_1\| \geq \|\widehat{\mu}\| - \|\mu_1\| > \sqrt{\beta} s_1 > \|0 - \mu_1\|. \tag{V.96}$$

We introduce now the probability measure $\mathbb{Q}$:

$$\mu_1 \overset{\mathbb{Q}}{\sim} \mathcal{N}(0, \alpha\beta s_1^2 \Sigma), \quad \mu_2 = \ldots = \mu_B = \mu_\circ \overset{\mathbb{Q}}{\sim} \mathcal{N}(\mu_1, \alpha\tau s_1^2 \Sigma), \tag{V.97}$$

where $\beta > 0$ and $\alpha \in (0, 1)$. Let A be the event $\{\|\mu_1\|^2 \leq \beta s_1^2, \|\mu_\circ - \mu_1\|^2 \leq \tau s_1^2\}$ and $\mathbb{E}_\mathbb{Q}$ denote the expectation over the distribution $\mathbb{Q}$, then:

$$\begin{aligned}
\inf_{\widehat{\mu}_1} \sup_{\mu_i \in B(\mu_1, \tau s_1)} R_1(\widehat{\mu}_1) &\geq \inf_{\widehat{\mu}_1 : \|\widehat{\mu}_1\| \leq 2\sqrt{\beta} s_1} \sup_{\substack{\mu_1 \in B(0, \sqrt{\beta} s_1) \\ \mu_i \in B(\mu_1, \sqrt{\tau} s_1)}} R_1(\widehat{\mu}_1) \\
&\geq \inf_{\widehat{\mu}_1 : \|\widehat{\mu}_1\| \leq 2\sqrt{\beta} s_1} \frac{1}{\mathbb{Q}(A)} \int_A R_1(\widehat{\mu}_1) d\mathbb{Q}(\nu, \mu_1, \ldots, \mu_B) \\
&\geq \inf_{\widehat{\mu}_1} \mathbb{E}_\mathbb{Q}[R_1(\widehat{\mu}_1)] - \sup_{\widehat{\mu}_1 : \|\widehat{\mu}_1\| \leq 2\sqrt{\beta} s_1} \mathbb{E}_\mathbb{Q}[R_1(\widehat{\mu}_1) 1_{A^c}] \\
&=: I - r,
\end{aligned} \tag{V.98}$$

Let us now bound I and r.

Lower bound for I : The first infimum (term I) is attained for $\widehat{\mu}_1 = \mathbb{E}\Big[\mu_1 | X_\bullet^{(1)}, \ldots, X_\bullet^{(B)}\Big]$. Let us calculate $\widehat{\mu}_1$.

$$\mathbb{E}\Big[\mu_1 | \mu_\circ, X_\bullet^{(1)}, \ldots, X_\bullet^{(B)}\Big] = \mathbb{E}\Big[\mu_1 | \mu_\circ, X_\bullet^{(1)}\Big] = \big((\alpha\beta)^{-1} + 1 + (\alpha\tau)^{-1}\big)^{-1} \left(\widehat{\mu}_1^{\texttt{NE}} + \frac{1}{\alpha\tau}\mu_\circ\right),$$

$$\mathbb{E}\Big[\mu_\circ | \mu_1, X_\bullet^{(1)}, \ldots, X_\bullet^{(B)}\Big] = \big((\alpha\tau)^{-1} + \|\rho\|^2\big)^{-1} \left(\frac{1}{\alpha\tau}\mu_1 + \sum_{k=2}^B \rho_k^2 \widehat{\mu}_k^{\texttt{NE}}\right)$$

where $\rho = (s_1/s_k)_{k \neq 1}$ and $\|\rho\|^2 = \sum_{k=2}^B \rho_k^2$. Combining these two expressions we get:

$$\begin{aligned}
\mathbb{E}\Big[\mu_1 | X_\bullet^{(1)}, \ldots, X_\bullet^{(B)}\Big] &= \big((\alpha\beta)^{-1} + 1 + (\alpha\tau)^{-1}\big)^{-1} \\
&\quad \times \left(\widehat{\mu}_1^{\texttt{NE}} + \frac{1}{\alpha\tau}\big((\alpha\tau)^{-1} + \|\rho\|^2\big)^{-1} \left(\frac{1}{\alpha\tau}\mathbb{E}\Big[\mu_1 | X_\bullet^{(1)}, \ldots, X_\bullet^{(B)}\Big] + \sum_{k=2}^B \rho_k^2 \widehat{\mu}_k^{\texttt{NE}}\right)\right),
\end{aligned}$$

and then:

$$\mathbb{E}\Big[\mu_1 | X_\bullet^{(1)}, \ldots, X_\bullet^{(B)}\Big] = \left((\alpha\beta)^{-1} + 1 + \frac{\|\rho\|^2}{1 + \alpha\tau\|\rho\|^2}\right)^{-1} \left(\widehat{\mu}_1^{\texttt{NE}} + \frac{1}{1 + \alpha\tau\|\rho\|^2} \sum_{k=2}^B \rho_k^2 \widehat{\mu}_k^{\texttt{NE}}\right),$$

Let us first notice that:

$$\mathbb{E}\Big[\mu_1|X_\bullet^{(\cdot)}\Big]-\mu_1=\left((\alpha\beta)^{-1}+1+\frac{\|\rho\|^2}{1+\alpha\tau\|\rho\|^2}\right)^{-1}\\ \times\left[(\widehat{\mu}_1^{\texttt{NE}}-\mu_1)+\frac{1}{1+\alpha\tau\|\rho\|^2}\sum_{k=2}^{B}\rho_k^2(\widehat{\mu}_k^{\texttt{NE}}-\mu_\circ)+\frac{\|\rho\|^2}{1+\alpha\tau\|\rho\|^2}(\mu_\circ-\mu_1)-\frac{1}{\alpha\beta}\mu_1\right]$$

Using that $\widehat{\mu}_1^{\texttt{NE}}-\mu_1$, $\widehat{\mu}_k^{\texttt{NE}}-\mu_\circ$ (for $k\neq 1$), $\mu_\circ-\mu_1$ and μ_1 are pairwise independent we get that:

$$\frac{\mathbb{E}\big[\|\widehat{\mu}_1-\mu_1\|^2\big]}{s_1^2}=\left((\alpha\beta)^{-1}+1+\frac{\|\rho\|^2}{1+\alpha\tau\|\rho\|^2}\right)^{-2}\\ \times\left[1+\frac{1}{(1+\alpha\tau\|\rho\|^2)^2}\sum_{k=2}^{B}\rho_k^4\rho_k^{-2}+\frac{\alpha\tau\|\rho\|^4}{(1+\alpha\tau\|\rho\|^2)^2}+\frac{1}{\alpha\beta}\right]$$

After simplification:

$$I=s_1^2\left((\alpha\beta)^{-1}+1+\frac{\|\rho\|^2}{1+\alpha\tau\|\rho\|^2}\right)^{-1} \tag{V.99}$$

Upper bound for r: Using the triangle and Cauchy-Schwartz inequalities we have:

$$\begin{aligned} r&=\sup_{\widehat{\mu}_1:\|\widehat{\mu}_1\|\leq 2\sqrt{\beta}s_1}\mathbb{E}\Big[\|\widehat{\mu}_1(X_\bullet^{(k)},k\in[\![B]\!])-\mu_1\|^2 1_{A^c}\Big] \qquad \text{(V.100)}\\ &\leq \mathbb{E}\big[2\big(4\beta s_1^2+\|\mu_1\|^2\big)1_{A^c}\big]\\ &\leq 8\beta s_1^2\mathbb{P}[A^c]+2\sqrt{\mathbb{E}[\|\mu_1\|^4]\mathbb{P}[A^c]}\\ &\leq 2s_1^2\Big(4\beta+\sqrt{3}\alpha\beta\Big)\sqrt{\mathbb{P}[A^c]}\leq 20\beta s_1^2\sqrt{\mathbb{P}[A^c]}\end{aligned}$$

It stays to show the exponential decrease of $\mathbb{P}[A^c]$. Let $\xi\sim\mathcal{N}(0,\Sigma)$:

$$\mathbb{P}\big[\|\mu_1\|^2\geq\beta s_1^2\big]=\mathbb{P}\big[\|\mu_\circ-\mu_1\|^2\geq\tau s_1^2\big]=\mathbb{P}\big[\|\xi\|^2\geq\alpha^{-1}\big]\\ \leq\exp\left(-\frac{d_1^{\mathrm{e}}}{2}\left(\sqrt{\frac{2}{\alpha}-1}-1\right)\right).$$

This follows from the concentration of the norm of Gaussian vectors (Lemma V.40). By union bound we get that:

$$r\leq 30s_1^2\beta\exp\left(-\frac{d_1^{\mathrm{e}}}{4}\left(\sqrt{\frac{2}{\alpha}-1}-1\right)\right).$$

Conclusion :

The lower bound finally obtained is :

$$\inf_{\widehat{\mu}_1}\sup_{\mu_i\in B(\mu_1,\tau s_1)}\frac{R_1(\widehat{\mu}_1)}{s_1^2}\geq\left((\alpha\beta)^{-1}+1+\frac{\|\rho\|^2}{1+\alpha\tau\|\rho\|^2}\right)^{-1}\\ -30\beta\exp\left(-\frac{d_1^{\mathrm{e}}}{4}\left(\sqrt{\frac{2}{\alpha}-1}-1\right)\right),$$

where $\alpha\in(0,1)$ and $\beta\in\mathbb{R}_+$ are two free parameters. We can choose $\beta=d_1^{\mathrm{e}}/\log d_1^{\mathrm{e}}$ and $\alpha=\frac{2}{1+(1+8\beta^{-1})^2}$, then:

$$\beta\exp\left(-\frac{d_1^{\mathrm{e}}}{4}\left(\sqrt{\frac{2}{\alpha}-1}-1\right)\right)=\beta\exp\left(-\frac{2d_1^{\mathrm{e}}}{\beta}\right)=\frac{1}{d_1^{\mathrm{e}}\log d_1^{\mathrm{e}}}$$

$$\left((\alpha\beta)^{-1}+1+\frac{\|\rho\|^2}{1+\alpha\tau\|\rho\|^2}\right)^{-1}-\left(1+\frac{\|\rho\|^2}{1+\alpha\tau\|\rho\|^2}\right)^{-1}\geq-(\alpha\beta)^{-1}\geq-41\frac{\log d_1^{\mathrm{e}}}{d_1^{\mathrm{e}}}.$$

and

$$\left(1+\frac{\|\rho\|^2}{1+\alpha\tau\|\rho\|^2}\right)^{-1}-\left(1+\frac{\|\rho\|^2}{1+\tau\|\rho\|^2}\right)^{-1}=-(1-\alpha)\frac{\tau\|\rho\|^2}{1+\tau\|\rho\|^2}\frac{\frac{\|\rho\|^2}{1+\alpha\tau\|\rho\|^2}}{1+\frac{\|\rho\|^2}{1+\alpha\tau\|\rho\|^2}}\frac{1}{1+\frac{\|\rho\|^2}{1+\tau\|\rho\|^2}}$$
$$\geq -(1-\alpha)\geq -40\frac{\log d_1^{\mathrm{e}}}{d_1^{\mathrm{e}}}$$

where we recall $\|\rho\|^2=\sum_{i=1}^B \frac{s_1^2}{s_i^2}-1=(\nu(V_\tau))^{-1}-1$. Hence:

$$\left(1+\frac{\|\rho\|^2}{1+\tau\|\rho\|^2}\right)^{-1}=\mathcal{B}(\tau,\nu(V_\tau))$$

By combining these three inequalities, we get that:

$$\inf_{\widehat{\mu}_1}\sup_{\mu_i\in B(\mu_1,\tau s_1)}\frac{R_1(\widehat{\mu}_1)}{s_1^2}\geq \mathcal{B}(\tau,\nu(V_\tau))-111\frac{\log d_1^{\mathrm{e}}}{d_1^{\mathrm{e}}}$$

Proof of Proposition V.20 Let $\mathcal{C}$ be a fixed J-partition of the means $(\mu_k)_{k\in[\![B]\!]}$ and denote $\zeta=\mathtt{diam}(\mathcal{C})$. Let us focus first on a specific group $j\in[\![J]\!]$ and task $k\in\mathcal{C}_j$. Denote $\tau_{j,k}=\zeta_j^2/s_k^2$ and $\nu_{j,k}=s^2(\mathcal{C}_j)/s_k^2$. Consider the vector of oracle weights $\boldsymbol{\omega}_k^*$ given by (V.13), wherein the target task 1 is replaced by k everywhere, and the subset of neighbouring tasks is taken as $\mathcal{C}_j\subseteq V_{\tau_{j,k}}$. Lemma V.6 then states $R_k(\boldsymbol{\omega}_k^*)/s_k^2\leq\mathcal{B}(\tau_{j,k},\nu_{j,k})$. As a consequence, according to Proposition V.12, it holds

$$\frac{R_k(\widehat{\boldsymbol{\omega}}_k)}{s_k^2}\leq(1+CBe^{-u_0})\left(\mathcal{B}(\tau_{j,k},\nu_{j,k})+C\sqrt{u_0}\frac{Q_k(\boldsymbol{\omega}_k^*)}{s_k^2}\right)+C\frac{u_0}{\sqrt{d_1^\bullet}}.$$

The rest of the proof is dedicated to bounding the terms $Q_k(\boldsymbol{\omega}_k^*)s_k^{-1}$ (and their sum over $k\in\mathcal{C}_j$). Denote $\omega_{k,\ell}^*$ the ℓ-th component of $\boldsymbol{\omega}_k^*$. It holds

$$\begin{aligned}\frac{Q_k(\boldsymbol{\omega}_k^*)}{s_k^2}&=s_k^{-2}\sum_{\ell\in\mathcal{C}_j\setminus\{k\}}\omega_{k,\ell}^*\sqrt{\frac{(\mu_\ell-\mu_k)^T\Sigma_k(\mu_\ell-\mu_k)}{N_k}+\frac{\operatorname{Tr}\Sigma_\ell\Sigma_k}{N_\ell N_k}}\\
&\leq s_k^{-2}\sum_{\ell\in\mathcal{C}_j\setminus\{k\}}\omega_{k,\ell}^*\frac{\|\Sigma_k\|_\infty^{1/2}}{\sqrt{N_k}}\sqrt{\zeta_j^2+s_\ell^2}\\
&\leq\frac{1}{\sqrt{d_k^{\mathrm{e}}}}\left((1-\omega_{k,k}^*)\sqrt{\tau_{j,k}}+\frac{\nu_{j,k}s_k}{1+\tau_{j,k}(1-\nu_{j,k})}\sum_{\ell\in\mathcal{C}_j\setminus\{k\}}s_\ell^{-1}\right)\\
&\leq\frac{1}{\sqrt{d_k^{\mathrm{e}}}}\left((1-\omega_{k,k}^*)\sqrt{\tau_{j,k}}+\nu_{j,k}s_k\sum_{\ell\in\mathcal{C}_j}s_\ell^{-1}\right),\end{aligned}\tag{V.101}$$

where we have used: $\|\mu_\ell-\mu_k\|\leq\zeta_j$ as tasks k and ℓ are in the group $\mathcal{C}_j$; $(\|\Sigma_k\|_\infty/N_k)^{1/2}=s_k/\sqrt{d_k^{\mathrm{e}}}$; and the explicit expression (V.13) for the oracle weights $\omega_{k,\ell}^*$ for group $\mathcal{C}_j$. For the first term of (V.101), for all $k\in\mathcal{C}_j$ we have:

$$(1-\omega_{k,k}^*)\sqrt{\tau_{j,k}}=\frac{1-\nu_{j,k}}{1+\tau_{j,k}(1-\nu_{j,k})}\sqrt{\tau_{j,k}}\leq\frac{\sqrt{\tau_{j,k}}}{1+\tau_{j,k}}\leq 1\,.$$

For the second term of (V.101), introduce the vector $\boldsymbol{\rho}:=(s_\ell^{-1})_{\ell\in\mathcal{C}_j}$ and observe that $\nu_{j,k}=\rho_k^2/\|\boldsymbol{\rho}\|_2^2$, thus, when summing over $k\in\mathcal{C}_j$:

$$\sum_{k\in\mathcal{C}_j}\left(\nu_{j,k}s_k\sum_{\ell\in\mathcal{C}_j}s_\ell^{-1}\right)=\sum_{k\in\mathcal{C}_j}\rho_k\frac{\|\boldsymbol{\rho}\|_1}{\|\boldsymbol{\rho}\|_2^2}=\frac{\|\boldsymbol{\rho}\|_1^2}{\|\boldsymbol{\rho}\|_2^2}\leq|\mathcal{C}_j|.$$

We deduce from the above estimates:

$$\sum_{k\in\mathcal{C}_j}\frac{Q_k(\boldsymbol{\omega}_k^*)}{s_k^2}\le\frac{2|\mathcal{C}|_j}{\min_k(d_k^{\mathrm{e}})^{1/2}},$$

implying

$$\frac{1}{B}\sum_{k=1}^{B}\frac{Q(\boldsymbol{\omega}_k^*)}{s_k^2}\le\frac{2}{\min_k(d_k^{\mathrm{e}})^{1/2}}.$$

Therefore for any J-partition $\boldsymbol{\mathcal{C}}$, since $d_k^\bullet\ge d_k^{\mathrm{e}}$:

$$\frac{1}{B}\sum_{k=1}^{B}\frac{R_k(\widehat{\boldsymbol{\omega}}_k)}{s_k^2}\le\left(1+CBe^{-u_0}\right)\left(\frac{1}{B}\sum_{j=1}^{J}\sum_{j\in\mathcal{C}_j}\mathcal{B}(\tau_{j,k},\nu_{j,k})+C'\frac{u_0}{\min_{k\in[\![B]\!]}(d_k^{\mathrm{e}})^{1/2}}\right).$$

□

Proof of Proposition V.22 The proof follows the same steps as the proof of Proposition V.18. Let $\mathcal{C}$ a J-partition of $[\![B]\!]$, $\boldsymbol{\zeta}\in\mathbb{R}_+^J$ and Σ a definite positive matrix in $\mathbb{R}^{d\times d}$. W.l.g. we can assume that $\operatorname{Tr}\Sigma=1$. In a first time, we are going to lower bound the minimax risk for the estimation of μ_1 that we can assume to be in the cluster 1 ($1\in\mathcal{C}_1$).

If for $j\in[\![J]\!]$ the means of $\mathcal{C}_j$ are in a ball of radius $\zeta_j/2$, then two means are at a distance at most ζ_j:

$$\inf_{\widehat{\mu}_1}\sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{mult}}(\mathcal{C},\boldsymbol{\zeta},\Sigma,\boldsymbol{s^2})}R_1(\widehat{\mu}_1)\ge\inf_{\widehat{\mu}_1}\sup_{\substack{\exists\nu_1,\dots,\nu_J\in\mathbb{R}^d\\ \mu_k\in B(\nu_j,\zeta_j/2),\forall k\in\mathcal{C}_j}}R_1(\widehat{\mu}_1).$$

For simplicity, the supremum over the vectors means μ_k is used to denote the supremum over the Gaussian distributions $\mathbb{P}_k=\mathcal{N}(\mu_k,s_k^2\Sigma)$.
We can restrict ourself in the case where the centres ν_j are in a ball around 0 of radius $\sqrt{\beta}$:

$$\inf_{\widehat{\mu}_1}\sup_{\substack{\exists\nu_1,\dots\nu_J\in\mathbb{R}^d\\ \mu_k\in B(\nu_j,\zeta_j/2),\forall k\in\mathcal{C}_j}}R_1(\widehat{\mu}_1)\ge\inf_{\widehat{\mu}_1}\sup_{\substack{\exists\nu_1,\dots\nu_J\in B(0,\sqrt{\beta})\\ \mu_k\in B(\nu_j,\zeta_j/2),\forall k\in\mathcal{C}_j}}R_1(\widehat{\mu}_1)$$

Let $\alpha\in(0,1),\beta>0$, we introduce now the probability measure $\mathbb{Q}=\mathbb{Q}(\alpha,\beta)$ on $(\mathbb{R}^d)^{B+J}$ such that a random vector $(\nu_1,\dots,\nu_J,\mu_1,\dots,\mu_B)\in(\mathbb{R}^d)^{B+J}$ follows the distribution $\mathbb{Q}$ if:

$$\nu_j\overset{\mathbb{Q}}{\sim}\mathcal{N}(0,\alpha\beta\Sigma)\text{ for }k\in[\![\mathcal{N}]\!],\qquad\mu_k\overset{\mathbb{Q}}{\sim}\mathcal{N}(\nu_j,\alpha\frac{\zeta_j^2}{4}\Sigma)\text{ for }k\in\mathcal{C}_j.$$

Hence, considering the events $H_j:=\{\|\nu_j\|^2\le\beta,\|\mu_k-\nu_j\|^2\le\zeta_j^2/4,k\in\mathcal{C}_j\}$, $H:=\cap_{j=1}^J H_j$, as in the equations (V.98):

$$\inf_{\widehat{\mu}_1}\sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{mult}}(\mathcal{C},\boldsymbol{\zeta},\Sigma,\boldsymbol{s^2})}R_1(\widehat{\mu}_1)\ge\inf_{\widehat{\mu}_1}\mathbb{E}_{\mathbb{Q}}[R_1(\widehat{\mu}_1)|H].$$

The distribution $\mathbb{Q}$ can be decomposed into a product of J probability measure: $\mathbb{Q}=\bigotimes_{j=1}^J\mathbb{Q}_j$ where $\mathbb{Q}_j$ is the distribution of $(\nu_j,(\mu_k)_{k\in\mathcal{C}_j})$. By independence, the Bayes estimator of μ_1 only consider the means of $\mathcal{C}_1$ and following equations (V.98) we get:

$$\inf_{\widehat{\mu}_1}\sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{mult}}(\mathcal{C},\boldsymbol{\zeta},\Sigma,\boldsymbol{s^2})}R_1(\widehat{\mu}_1)\ge\inf_{\widehat{\mu}_1}\mathbb{E}_{\mathbb{Q}_1}[R_1(\widehat{\mu}_1)|H_1]\ge\frac{1}{\mathbb{Q}(H_1)}(I_1-r_1),$$

where

$$I_1:=\inf_{\widehat{\mu}_1}\mathbb{E}_{\mathbb{Q}_1}[R_1(\widehat{\mu}_1)],\qquad r_1:=\sup_{\widehat{\mu}_1:\|\widehat{\mu}_1\|\le2\sqrt{\beta}+\zeta_1}\mathbb{E}_{\mathbb{Q}}\left[R_1(\widehat{\mu}_1)1_{H_1^c}\right]\tag{V.102}$$

We have used that the infimum is attained for an estimator $\widehat{\mu}_1$ bounded by $2\sqrt{\beta}+\zeta_1$, because the estimator $\widehat{\mu}=0$ beats the estimators outside that ball (as in (V.96)).

Lower bound for I_1 : The infimum is attained for $\widehat{\mu}_1 = \mathbb{E}\Big[\mu_1 | X_\bullet^{(k)}\ k\in\mathcal{C}_1\Big]$. Let us calculate $\widehat{\mu}_1$. We will denote in the rest of the proof $\tilde{\zeta}_j := \zeta_j/2$:

$$\mathbb{E}\Big[\mu_1|\nu_1, X_\bullet^{(k)}\ k\in\mathcal{C}_1\Big] = \mathbb{E}\Big[\mu_1|\nu_1, X_\bullet^{(1)}\Big] = \frac{\alpha\widetilde{\zeta}_1^2}{s_1^2+\alpha\widetilde{\zeta}_1^2}\widehat{\mu}_1^{\mathtt{NE}} + \frac{s_1^2}{s_1^2+\alpha\widetilde{\zeta}_1^2}\nu_1\,,$$

$$\mathbb{E}\Big[\nu_1|X_\bullet^{(k)}\ k\in\mathcal{C}_1\Big] = \left((\alpha\beta)^{-1}+\sum_{i\in\mathcal{C}_1}\left(\alpha\widetilde{\zeta}_1^2+s_k^2\right)^{-1}\right)^{-1}\sum_{k\in\mathcal{C}_1}\frac{1}{\alpha\widetilde{\zeta}_1^2+s_k^2}\widehat{\mu}_k^{\mathtt{NE}}$$

Combining these two expressions:

$$\mathbb{E}\Big[\mu_1|X_\bullet^{(k)}\ k\in\mathcal{C}_1\Big] = \frac{\alpha\widetilde{\zeta}_1^2}{s_1^2+\alpha\widetilde{\zeta}_1^2}\widehat{\mu}_1^{\mathtt{NE}} + \frac{s_1^2}{s_1^2+\alpha\widetilde{\zeta}_1^2}\left((\alpha\beta)^{-1}+\sum_{k\in\mathcal{C}_1}\left(\alpha\widetilde{\zeta}_1^2+s_k^2\right)^{-1}\right)^{-1}\sum_{k\in\mathcal{C}_1}\frac{1}{\alpha\widetilde{\zeta}_1^2+s_k^2}\widehat{\mu}_k^{\mathtt{NE}}$$

Let $\kappa_1 := \left((\alpha\beta)^{-1}+\sum_{k\in\mathcal{C}_1}\left(\alpha\widetilde{\zeta}_1^2+s_k^2\right)^{-1}\right)^{-1}$, we can first notice that:

$$\begin{aligned}\mathbb{E}\Big[\mu_1|X_\bullet^{(\cdot)}\Big]-\mu_1 =& \left[\frac{\alpha\widetilde{\zeta}_1^2}{s_1^2+\alpha\widetilde{\zeta}_1^2}+\frac{\kappa_1 s_1^2}{(s_1^2+\alpha\widetilde{\zeta}_1)^2}\right](\widehat{\mu}_1^{\mathtt{NE}}-\mu_1)\\ &+\frac{\kappa_1 s_1^2}{s_1^2+\alpha\widetilde{\zeta}_1^2}\sum_{k\in\mathcal{C}_1\setminus\{1\}}\frac{1}{\alpha\widetilde{\zeta}_1^2+s_k^2}(\widehat{\mu}_k^{\mathtt{NE}}-\nu_1)\\ &-\frac{s_1^2}{s_1^2+\alpha\widetilde{\zeta}_1^2}\left(1-\frac{\kappa_1 s_1^2}{s_1^2+\alpha\widetilde{\zeta}_1^2}\right)(\mu_1-\nu_1)-\frac{\kappa_1 s_1^2}{s_1^2+\alpha\widetilde{\zeta}_1^2}\frac{1}{\alpha\beta}\nu_1\,.\end{aligned}$$

Using that $\widehat{\mu}_k^{\mathtt{NE}}-\nu_1$ for $k\in\mathcal{C}_1\setminus\{1\}$, $\widehat{\mu}_1^{\mathtt{NE}}-\mu_1$, $\mu_1-\nu_1$ and ν_1 are pairwise independent we get that:

$$\begin{aligned}\mathbb{E}\big[\|\widehat{\mu}_1-\mu_1\|^2\big] =& \left[\frac{\alpha\widetilde{\zeta}_1^2}{s_1^2+\alpha\widetilde{\zeta}_1^2}+\frac{\kappa_1 s_1^2}{(s_1^2+\alpha\widetilde{\zeta}_1)^2}\right]^2 s_1^2+\frac{\kappa_1^2 s_1^4}{(s_1^2+\alpha\widetilde{\zeta}_1^2)^2}\sum_{k\in\mathcal{C}_1\setminus\{1\}}\frac{1}{\alpha\widetilde{\zeta}_1^2+s_k^2}\\ &+\frac{s_1^4}{(s_1^2+\alpha\widetilde{\zeta}_1^2)^2}\left(1-\frac{\kappa_1 s_1^2}{s_1^2+\alpha\widetilde{\zeta}_1^2}\right)^2\alpha\widetilde{\zeta}_1^2+\frac{\kappa_1^2 s_1^4}{(s_1^2+\alpha\widetilde{\zeta}_1^2)^2}\frac{1}{\alpha\beta}\,.\end{aligned}$$

After simplification:

$$\frac{I_1}{s_1^2} = \frac{\alpha\widetilde{\zeta}_1^2}{s_1^2+\alpha\widetilde{\zeta}_1^2}+\frac{\kappa_1 s_1^2}{(s_1^2+\alpha\widetilde{\zeta}_1^2)^2} \tag{V.103}$$

Upper bound for r_1 : By the same arguments of equations (V.100):

$$\sup_{\widehat{\mu}_1:\|\widehat{\mu}_1\|\le 2\sqrt{\beta}+\zeta_1}\mathbb{E}\Big[\|\widehat{\mu}_1(X_\bullet^{(k)},k\in\mathcal{C}_1)-\mu_1\|^2 1_{H_1^c}\Big] \le 20(\beta+\zeta_1^2)\sqrt{\mathbb{P}[H_1^c]}$$

From Lemma V.40, for all $k\in\mathcal{C}_1$:

$$\mathbb{P}\big[\|\nu_1\|^2\ge\beta\big] = \mathbb{P}\big[\|\mu_k-\nu_1\|^2\ge\zeta_1^2/2\big] \le \exp\left(-\frac{d^{\mathrm{e}}}{2}\left(\sqrt{\frac{2}{\alpha}-1}-1\right)\right),$$

and by union bound we get that :

$$r_1 \le 20(\beta+\zeta_1^2)\sqrt{|\mathcal{C}_1|+1}\exp\left(-\frac{d^{\mathrm{e}}}{4}\left(\sqrt{\frac{2}{\alpha}-1}-1\right)\right).$$

where $d^{\mathrm{e}} = \operatorname{Tr}\Sigma/\|\Sigma\|_\infty$.

Compound bound We recall that $\mathbb{Q} = \bigotimes_{j=1}^J \mathbb{Q}_j$ where $\mathbb{Q}_j$ is the distribution of $(\nu_j, \mu_k \text{ for } k \in \mathcal{C}_j)$. Then let $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_k)_{k\in[\![B]\!]} \in (\mathbb{R}^d)^B$ be an estimator of the vectors $(\boldsymbol{\mu}_k)_{k\in[\![B]\!]}$:

$$\begin{aligned}\inf_{\widehat{\boldsymbol{\mu}}} \sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{mult}}(\mathcal{C},\boldsymbol{\zeta},\Sigma,\boldsymbol{s^2})} \frac{1}{B}\sum_{k=1}^B \frac{R_k(\widehat{\mu}_k)}{s_k^2} &\geq \inf_{\widehat{\boldsymbol{\mu}}} \frac{1}{\mathbb{Q}(H)}\int_H \frac{1}{B}\sum_{k=1}^B \frac{R_k(\widehat{\mu}_k)}{s_k^2} d\mathbb{Q}(\nu_1,\dots,\nu_{\mathcal{N}},\mu_1,\dots,\mu_B)\\ &= \inf_{\widehat{\boldsymbol{\mu}}} \frac{1}{B}\sum_{j=1}^J \sum_{k\in\mathcal{C}_j} \frac{\mathbb{Q}(H_{-j})}{\mathbb{Q}(H)}\int_{H_j} \frac{R_k(\widehat{\mu}_k)}{s_k^2} d\mathbb{Q}_j(\nu_j,(\mu_\ell)_{\ell\in\mathcal{C}_j})\end{aligned}$$

where we recall $H_j = \{\|\nu_j\|^2 \leq \beta, \|\mu_k - \nu_j\|^2 \leq \widetilde{\zeta}_j^2, \forall k \in \mathcal{C}_j\}$, $H = \bigcap_{j=1}^J H_j$ and $H_{-j} = \bigcap_{\ell\neq j} H_\ell$. Using that $\mathbb{Q}(H_{-j})/\mathbb{Q}(H) = \mathbb{Q}_j(H_j)^{-1} \geq 1$ and that the infimum over estimators $\widehat{\boldsymbol{\mu}}$ of the sum is the sum of the infimum over estimators $\widehat{\mu}_k$, we get that:

$$\begin{aligned}&\inf_{\widehat{\boldsymbol{\mu}}} \sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{mult}}(\mathcal{C},\boldsymbol{\zeta},\Sigma,\boldsymbol{s^2})} \frac{1}{B}\sum_{k=1}^B \frac{R_k(\widehat{\mu}_k)}{s_k^2} \geq \frac{1}{B}\sum_{j=1}^J\sum_{k\in\mathcal{C}_j}(I_k - r_k)\\ &\geq \frac{1}{B}\sum_{j=1}^J\sum_{k\in\mathcal{C}_j}\frac{\alpha\widetilde{\zeta}_j^2}{s_k^2+\alpha\widetilde{\zeta}_j^2} + \frac{\kappa_j s_k^2}{(s_k^2+\alpha\widetilde{\zeta}_j^2)^2} - \frac{20}{B}\left(\sum_{j=1}^J |\mathcal{C}_j|^{3/2}\frac{\beta+\widetilde{\zeta}_j^2}{s^2(\mathcal{C}_j)}\right)\exp(-d^{\mathrm{e}}c(\alpha)) \qquad \text{(V.104)}\end{aligned}$$

where $\kappa_j = \left((\alpha\beta)^{-1} + \sum_{k\in\mathcal{C}_j}\left(\alpha\widetilde{\zeta}_j^2 + s_k^2\right)^{-1}\right)^{-1}$ and $c(\alpha) = \left(\sqrt{\frac{2}{\alpha}-1}-1\right)/4$.

Conclusion :

Let $d^{\mathrm{e}} \to \infty$ in (V.104), then:

$$\lim_{d^{\mathrm{e}}\to\infty}\inf_{\widehat{\boldsymbol{\mu}}}\sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{mult}}(\mathcal{C},\boldsymbol{\zeta},\Sigma,\boldsymbol{s^2})}\frac{1}{B}\sum_{k=1}^B\frac{R_k(\widehat{\mu}_k)}{s_k^2} \geq \frac{1}{B}\sum_{j=1}^J\sum_{k\in\mathcal{C}_j}\frac{\alpha\widetilde{\zeta}_j^2}{s_k^2+\alpha\widetilde{\zeta}_j^2} + \frac{\kappa_j s_k^2}{(s_k^2+\alpha\widetilde{\zeta}_j^2)^2}$$

Let $\alpha \to 1$ and $\beta \to \infty$, then:

$$\lim_{d^{\mathrm{e}}\to\infty}\inf_{\widehat{\boldsymbol{\mu}}}\sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{mult}}(\mathcal{C},\boldsymbol{\zeta},\Sigma,\boldsymbol{s^2})}\frac{1}{B}\sum_{k=1}^B\frac{R_k(\widehat{\mu}_k)}{s_k^2} \geq \frac{1}{B}\sum_{j=1}^J\sum_{k\in\mathcal{C}_j}\frac{\widetilde{\zeta}_j^2}{s_k^2+\widetilde{\zeta}_j^2} + \frac{s_k^2}{s_k^2+\widetilde{\zeta}_j^2}\frac{1}{\sum_{\ell\in C_j}\frac{s_k^2+\widetilde{\zeta}_j^2}{s_\ell^2+\widetilde{\zeta}_j^2}} \qquad \text{(V.105)}$$

We conclude by remarking that for all $j \in [\![J]\!]$:

$$\sum_{\ell\in\mathcal{C}_j}\frac{s_k^2+\widetilde{\zeta}_j^2}{s_\ell^2+\widetilde{\zeta}_j^2} = 1 + \sum_{\ell\in\mathcal{C}_j\setminus\{k\}}\frac{s_k^2+\widetilde{\zeta}_j^2}{s_\ell^2+\widetilde{\zeta}_j^2} \leq 1 + \sum_{\ell\in\mathcal{C}_j\setminus\{k\}}\frac{s_k^2+\widetilde{\zeta}_j^2}{s_\ell^2}$$

Then:

$$\begin{aligned}&\frac{1}{B}\sum_{j=1}^J\sum_{k\in\mathcal{C}_j}\frac{\widetilde{\zeta}_j^2}{s_k^2+\widetilde{\zeta}_j^2} + \frac{s_k^2}{s_k^2+\widetilde{\zeta}_j^2}\frac{1}{\sum_{\ell\in C_j}\frac{s_k^2+\widetilde{\zeta}_j^2}{s_\ell^2+\widetilde{\zeta}_j^2}}\\ &\qquad\geq \frac{1}{B}\sum_{j=1}^J\sum_{k\in\mathcal{C}_j}\frac{\widetilde{\zeta}_j^2}{s_k^2+\widetilde{\zeta}_j^2} + \frac{s_k^2}{s_k^2+\widetilde{\zeta}_j^2}\frac{1}{1+\sum_{\ell\in\mathcal{C}_j\setminus\{k\}}\frac{s_k^2+\widetilde{\zeta}_j^2}{s_\ell^2}} = \mathcal{L}^*(\boldsymbol{s},\boldsymbol{\mathcal{C}},\boldsymbol{\zeta}/2).\end{aligned}$$

□

Proof of Proposition V.23 We start with the following elementary bounds on the function $\mathcal{B}$ (for $\tau \geq 0, \nu \in [0,1]$)

$$\mathcal{B}(\tau,\nu) \leq \frac{\tau+\nu}{1+\tau} \leq \max\left(1, \frac{\tau}{1+\tau} + \nu\right). \tag{V.106}$$

Now consider the quantity $A_j := |\mathcal{C}_j|^{-1} \sum_{k\in\mathcal{C}_j} \mathcal{B}(\tau_{j,k}, \nu_{j,k})$. Observe that $\sum_{k\in\mathcal{C}_j} \nu_{j,k} = 1$ and $\tau_{j,k} = \nu_{j,k} B_j$, where $B_j := \zeta_j^2 / s^2(\mathcal{C}_j)$. Thus

$$A_j := |\mathcal{C}_j|^{-1} \sum_{k\in\mathcal{C}_j} \mathcal{B}(B_j\nu_{j,k}, \nu_{j,k}) \leq (B_j+1)|\mathcal{C}_j|^{-1} \sum_{k\in\mathcal{C}_j} \frac{\nu_{j,k}}{1+B_j\nu_{j,k}}.$$

where we have used the first inequality in (V.106). By concavity of $t \mapsto t/(1+t)$ we conclude to

$$A_j \leq \frac{B_j|\mathcal{C}_j|^{-1} + |\mathcal{C}_j|^{-1}}{1+B_j|\mathcal{C}_j|^{-1}} = \frac{\bar{\tau}_j + |\mathcal{C}_j|^{-1}}{1+\bar{\tau}_j},$$

and thus to (V.45) by summation over $j \in [\![J]\!]$. Now using the second inequality in (V.106), we obtain

$$\sum_{j\in[\![J]\!]} \frac{|\mathcal{C}_j|}{B} \frac{\bar{\tau}_j + |\mathcal{C}_j|^{-1}}{1+\bar{\tau}_j} \leq \sum_{j\in[\![J]\!]} \frac{|\mathcal{C}_j|}{B} \min\left(1, \frac{\bar{\tau}_j}{1+\bar{\tau}_j} + |\mathcal{C}_j|^{-1}\right) \leq \min\left(1, \frac{\bar{\tau}_*}{1+\bar{\tau}_*} + \frac{J}{B}\right),$$

where we have used the second inequality in (V.106) and the biconcave character of the function $(x,y) \mapsto \min(1, y + x/(1+x))$; thus establishing (V.46). Assume now that all risks and the diameters are equal, i.e. $s_k^2 = s^2$ and $\zeta_j = \zeta$ for $k \in [\![B]\!]$ and $j \in [\![J]\!]$. Then for all $j \in [\![J]\!]$ and $k \in [\![B]\!]$, $\bar{s}^2(\mathcal{C}_j) = s^2$, $\bar{\tau}_{j,k} = \zeta^2/s^2 = \bar{\tau}$ and $\nu_{j,k} = |\mathcal{C}_j|^{-1}$. Using the elementary bound

$$\mathcal{B}(\tau,\nu) \geq \frac{\tau}{1+\tau} + \frac{\nu}{(1+\tau)^2},$$

we thus have in this case

$$\begin{aligned}
\mathcal{L}^*(\boldsymbol{s}, \boldsymbol{\mathcal{C}}, \boldsymbol{\zeta}) &= \frac{1}{B}\sum_{j=1}^{J}\sum_{k\in C_j} \mathcal{B}(\tau_{j,k}, \nu_{j,k}) = \frac{1}{B}\sum_{j=1}^{J}\sum_{k\in C_j} \mathcal{B}(\bar{\tau}, |\mathcal{C}_j|^{-1}) \\
&\geq \frac{1}{B}\sum_{j=1}^{J}\sum_{k\in C_j} \left(\frac{\bar{\tau}}{1+\bar{\tau}} + \frac{|\mathcal{C}_j|^{-1}}{(1+\bar{\tau})^2}\right) \\
&= \frac{\bar{\tau}}{1+\bar{\tau}} + \frac{J}{B}\frac{1}{(1+\bar{\tau})^2},
\end{aligned} \tag{V.107}$$

Finally, since for $\tau \geq 0, \nu \in [0,1]$:

$$\begin{aligned}
\frac{\tau}{1+\tau} + \frac{\nu}{(1+\tau)^2} &\geq \max\left(\frac{\tau}{1+\tau}, \frac{1}{(1+\tau)^2}\left(\frac{\tau}{1+\tau} + \nu\right)\right) \\
&\geq \max\left(\frac{\tau}{1+\tau}, \frac{1}{(1+\tau)^2}\right) \min\left(1, \frac{\tau}{1+\tau} + \nu\right) \\
&\geq 0.38 \min\left(1, \frac{\tau}{1+\tau} + \nu\right),
\end{aligned}$$

we conclude that in the case of equal risks and diameters the upper bound (V.46) and the lower bound (V.107) differ by a factor at most $1/0.38 \leq 2.7$.

□

V.10 About the constant in the translation-invariant kernel setting

In this section, we investigate the distribution-dependent constant $\phi = M^2/(\operatorname{Tr}\Sigma)$ in the **(BS)** setting (i.e., for data bounded in norm by the constant M). This constant comes into play in the risk bounds for our methods, in relation to sufficient sample sizes, see e.g. Props. V.31, V.16. Rewriting $\operatorname{Tr}\Sigma = \mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right]$ yields a direct interpretation of ϕ, namely it is the ratio between the known bound on $\|X\|$ and the "variance" of X; in other words, ϕ is all the bigger as the variable X is more concentrated in relation to the size of its support.

We are interested in an understanding more detailed than this simple observation in the situation of kernel mean embedding (KME), which was our primary motivation for investigating the **(BS)** setting. Namely, in that situation the user might choose between different kernels and their associated Hilbert space mappings, in particular choosing or tuning the "kernel bandwidth". Even if kernels under consideration are all bounded by the same constant, different kernels may give rise to different constants ϕ for the same underlying data distribution.

We look into this issue under the following general conditions:

(K1) the original data takes values in $\mathcal{Z} = \mathbb{R}^\ell$, and the data whose means we wish to estimate have been obtained via a Hilbert space mapping $X = \Phi_\kappa(Z)$, $\Phi_\kappa : \mathbb{R}^\ell \to \mathcal{H}$, associated to the kernel $\kappa(z, z') = \langle \Phi_\kappa(z), \Phi_\kappa(z') \rangle$.

(K2) κ is a translation-invariant kernel on $\mathbb{R}^\ell$, of the form $\kappa(z, z') = F(z - z')$, where $F : \mathbb{R}^\ell \to \mathbb{R}$, with $M^2 := F(0)$.

(K3) For any $u \in \mathbb{R}^\ell$, the function $\lambda \mapsto F(\lambda u)$ is nonincreasing on $\mathbb{R}_+$. Furthermore, there exist constants $h > 0, c \leq 1$ such that

$$F(u) \leq M^2 \left(1 - c\frac{\|u\|^2}{h^2}\right), \text{ for all } u \in \mathbb{R}^\ell \text{ s.t. } 0 \leq \|u\| \leq h. \tag{V.108}$$

Observe that (K1)-(K2) imply that the mapped data X satisfies **(BS)**; as for (K3), it means that the kernel is locally upper bounded by a strongly concave function in a neighbourhood of 0 of size h. The latter quantity can therefore interpreted as a proxy bandwidth for the kernel; and if F_1 satisfies (V.108) for $h = 1$ then the rescaled kernel function $F_h(u) := F_1(u/h)$ satisfies (V.108) for the bandwidth parameter $h > 0$. The classical Gaussian, exponential, and Matérn kernels, for example, satisfy such conditions.

Proposition V.44. *Assume (K1)-(K2)-(K3) hold, and that the distribution P of the original data Z in $\mathbb{R}^\ell$ satisfies the following norm moment condition for some $p \geq 1, C > 0$:*

$$\frac{\mathbb{E}\left[\xi^{2p}\right]}{\mathbb{E}[\xi^p]^2} \leq C, \qquad \textit{where } \xi := \|Z - \mathbb{E}[Z]\|. \tag{V.109}$$

Then it holds

$$\phi = \frac{M^2}{\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right]} \leq \frac{4.2^{\frac{2}{p}+2p}C}{c} \max\left(1, \frac{h}{2\mathbb{E}[\|Z - \mathbb{E}[Z]\|^p]^{\frac{1}{p}}}\right)^2.$$

Assume $p = 2$ to simplify (we allowed for other values of p in the moment condition (V.109) mainly with the possible value $p = 1$ in mind, which makes the condition weaker; the discussion below can be readily adapted to other values of p). This result shows that, provided the bandwidth parameter h is chosen of the order of $\sigma_Z := \mathbb{E}\left[\|Z - \mathbb{E}[Z]\|^2\right]^{\frac{1}{2}}$ or smaller, the constant ϕ for the mapped data is bounded independently of h. The bound depends on (1) the strong concavity parameter c of the upper bound on the (unit scaled) kernel function in a neighbourhood of the origin, and (2) the norm moment ratio (V.109) of the original

data distribution. Since $\mathbb{E}\big[\xi^4\big] \leq \mathbb{E}\big[\xi^2\big]\|\xi\|_{L^\infty}^2$, in the case where the original data is itself bounded in norm by a constant R, (V.109) holds with $C = (R/\sigma_Z)^2$. Thus, if the original X data is bounded, the distribution of the mapped data Z under the above conditions "inherits" the constant ϕ from that of the original data, up to factors. However, the norm moment condition is much milder than a boundedness condition and can also accommodate unbounded distributions with heavy tails of the original data.

Proof of Proposition V.44.

For $Z, Z' \sim \mathbb{P}$ independent, denote $D := \|Z - Z'\|$, $\theta := \min\Big(\frac{h^p}{\mathbb{E}[D^p]}, \frac{1}{2}\Big)$, and $t^p := \theta\mathbb{E}[D^p] = \min\Big(h^p, \frac{\mathbb{E}[D^p]}{2}\Big)$, it holds

$$\begin{aligned}
\|\mathbb{E}[\Phi_\kappa(Z)]\|^2/M^2 &= M^{-2}\mathbb{E}[\langle\Phi_\kappa(Z), \Phi_\kappa(Z')\rangle] \\
&= M^{-2}\mathbb{E}[F(Z - Z')] \\
&\leq 1 - c\frac{t^2}{h^2}\mathbb{P}[D^p > t^p] \\
&\leq 1 - c\frac{\mathbb{E}[D^p]^{\frac{2}{p}}}{h^2}\theta^{\frac{2}{p}}(1-\theta)^2\frac{\mathbb{E}[D^p]^2}{\mathbb{E}[D^{2p}]} \\
&\leq \frac{\mathbb{E}\Big[\|\Phi_\kappa(Z)\|^2\Big]}{M^2} - \frac{c}{4}\min\left(1, \frac{\mathbb{E}[D^p]^{\frac{2}{p}}}{2^{\frac{2}{p}}h^2}\right)\frac{\mathbb{E}[D^p]^2}{\mathbb{E}[D^{2p}]},
\end{aligned}$$

where the first inequality stems from (K3); the second comes from the Paley-Zygmund inequality; and we used $\theta \leq \frac{1}{2}$ for the third. Since $\mathbb{E}\Big[\|\Phi_\kappa(Z)\|^2\Big] - \|\mathbb{E}[\Phi_\kappa(Z)]\|^2 = \mathbb{E}\Big[\|X\|^2\Big] - \|\mathbb{E}[X]\|^2 = \mathbb{E}\Big[\|X - \mathbb{E}[X]\|^2\Big]$, we deduce

$$\frac{M^2}{\mathbb{E}\Big[\|X - \mathbb{E}[X]\|^2\Big]} \leq \frac{4.2^{\frac{2}{p}}}{c}\max\left(1, \frac{h}{\|D\|_{L^p(P)}}\right)^2\left(\frac{\mathbb{E}\big[D^{2p}\big]}{\mathbb{E}[D^p]^2}\right).$$

Finally, note that

$$\mathbb{E}\big[D^{2p}\big] = \mathbb{E}\Big[\|Z - Z'\|^{2p}\Big] \leq \mathbb{E}\big[(\|Z - \mathbb{E}[Z]\| + \|Z' - \mathbb{E}[Z']\|)^{2p}\big] \leq 2^{2p}\mathbb{E}\Big[\|Z - \mathbb{E}[Z]\|^{2p}\Big],$$

and by Jensen's inequality

$$\mathbb{E}[\|Z - \mathbb{E}[Z]\|^p] = \mathbb{E}\big[\|Z - \mathbb{E}[Z']\|^p\big] \leq \mathbb{E}\big[\|Z - Z'\|^p\big] = \mathbb{E}[D^p].$$

(Observe that the equality $\mathbb{E}\Big[\|Z - Z'\|^2\Big] = 2\mathbb{E}\Big[\|Z - \mathbb{E}[Z]\|^2\Big]$ holds, so the constants in the first, resp. second inequality above can be improved for the special cases $p = 1$, resp. $p = 2$.) □

V.11 Description of the tested methods

The tested methods propose KME estimations of the form

$$\widehat{\mu}_i^{\mathtt{m}} := \sum_{j \in [\![B]\!]} \omega_{ij}^{\mathtt{m}} \cdot \widehat{\mu}_j^{\mathtt{NE}},$$

where the definition of the weighting $\omega_{ij}^{\mathtt{m}}$ depends on the applied method $\mathtt{m}$.

V.11.1 State-of-the-Art Approaches

(i) `NE` considers each bag individually.

$$\omega_{ij}^{\mathtt{NE}} = \begin{cases} 1, & \text{for } i = j \\ 0, & \text{otherwise.} \end{cases}$$

(ii) `R-KMSE` (Muandet et al., 2016) estimates each KME individually but shrinks it towards $\mathbf{0}$. The amount of shrinkage is data dependent

$$\omega_{ij}^{\texttt{R-KMSE}} = \begin{cases} 1 - \frac{\lambda_i}{1+\lambda_i}, & \text{for } i = j \\ 0, & \text{otherwise} \end{cases}$$

where

$$\lambda_i = \frac{\varrho_i - \rho_i}{(1/N_i - 1)\varrho_i + (N_i - 1)\rho_i}$$

with $\varrho_i = 1/N_i \sum_{n=1}^{N_i} \kappa(Z_n^{(i)}, Z_n^{(i)})$ and $\rho_i = 1/N_i^2 \sum_{n,n'=1}^{N_i} \kappa(Z_n^{(i)}, Z_{n'}^{(i)})$.

(iii) `MTA const` (Feldman et al., 2014) was initially proposed for the estimation for multiple real means. We adapted the approach such that it can be applied to the estimation of multiple kernel means

$$\omega_{ij}^{\texttt{MTA const}} = \left(\left(I + \frac{\gamma}{B} \hat{S} \cdot L(A) \right)^{-1} \right)_{ij}. \tag{V.110}$$

Here, $\hat{S} = \operatorname{diag}\big((\hat{s}_i^2)_{i \in [\![B]\!]}\big)$, as defined in (V.3), can be estimated as

$$\hat{s}_i^2 = \frac{1}{2N_i^2(N_i - 1)} \sum_{n \neq n'}^{N_i} \kappa\Big(Z_n^{(i)}, Z_n^{(i)}\Big) - 2\kappa\Big(Z_n^{(i)}, Z_{n'}^{(i)}\Big) + \kappa\Big(Z_{n'}^{(i)}, Z_{n'}^{(i)}\Big), \tag{V.111}$$

which corresponds to (V.30), and $L(A)$ denotes the graph Laplacian of task-similarity matrix A. For `MTA const` the similarity is assumed to be constant, i.e., $A = a \cdot (\mathbf{1}\mathbf{1}^T)$ with

$$a = \frac{1}{B(B-1)} \sum_{i,j \in [\![B]\!]} \big\| \widehat{\mu}_i^{\texttt{NE}} - \widehat{\mu}_j^{\texttt{NE}} \big\|_{\mathcal{H}}^2.$$

The optimal value for model parameter γ may be found using model optimization. As default value we propose $\gamma = 1$.

V.11.2 `AGG` Approaches

The aggregation approaches form a convex combination of possibly all bags whose weights are found directly by minimization of quantities related to the squared risk.

(iv) `AGG orth` is based on the constraint optimization problem

$$\omega_{i\cdot} = \underset{w_{i\cdot}}{\operatorname{argmin}} \left\{ \mathbb{E} \left\| \sum_{j \in [\![B]\!]} w_{ij} \widehat{\mu}_j^{\texttt{NE}} - \mu_i \right\|_{\mathcal{H}}^2 \right\} \text{ s.t. } \sum_{j \in [\![B]\!]} \omega_{ij} = 1\,, \ \forall i, j \in [\![B]\!] : \omega_{ij} \geq 0.$$

Using Lagrangian multipliers the optimal solution can be derived as

$$\omega_{i\cdot} \simeq \left(S + \Lambda^{(i)} \right)^{(-1)} \mathbf{1} \tag{V.112}$$

where $S = \operatorname{diag}\big((s_i^2)_{i \in [\![B]\!]}\big)$ and $\Lambda^{(i)} \in \mathbb{R}^{B \times B}$ with $\Lambda_{j,j'}^{(i)} = \langle \mu_j - \mu_i, \mu_{j'} - \mu_i \rangle_{\mathcal{H}}$. Central assumption of `AGG orth` is $\Lambda_{j,j'}^{(i)} = \langle \mu_j - \mu_i, \mu_{j'} - \mu_i \rangle_{\mathcal{H}} = 0$ for all $j \neq j'$ such that $\Lambda^{(i)}$ becomes a diagonal matrix with $\Lambda^{(i)} = \operatorname{diag}\left(\left(\|\mu_j - \mu_i\|_{\mathcal{H}}^2 \right)_{j \in [\![B]\!]} \right)$. An unbiased estimation of $\|\mu_j - \mu_i\|_{\mathcal{H}}^2$ is given by (V.16) which in the kernel setting translates to

$$\begin{aligned} \widehat{\mathrm{MMD}}^2(\mu_i, \mu_j) = &\sum_{n \neq n'}^{N_i} \frac{\kappa(Z_n^{(i)}, Z_{n'}^{(i)})}{N_i(N_i - 1)} + \sum_{m \neq m'}^{N_j} \frac{\kappa(Z_m^{(j)}, Z_{m'}^{(j)})}{N_j(N_j - 1)} \\ &- 2 \sum_{n=1}^{N_i} \sum_{m=1}^{N_j} \frac{\kappa(Z_n^{(i)}, Z_m^{(j)})}{N_i N_j}. \end{aligned} \tag{V.113}$$

Eq. (V.112) reduces to

$$\omega_{ij}^{\texttt{AGG orth}} = \frac{1}{\hat{s}_j^2 + \gamma \cdot \widehat{\mathsf{MMD}}^2(\mu_i, \mu_j)}.$$

We add a multiplicative constant γ for more flexibility, whose value is either found by model optimization or $\gamma = 13$ taken as default. If the distances between bags is inhomogeneous, e.g., the data set contains close but also far distant unrelated bags, higher values of γ might be advisable. Finally the weights are normalised such that they sum to one.

(v) `AGG egd` is based on Q-Aggregation and resembles (V.37)

$$\omega_{i\cdot}^{\texttt{AGG egd}} = \underset{w_{i\cdot}}{\operatorname{argmin}} \left\{ \widehat{L}_i + c_q \widehat{Q}_i + c_1 \sum_{j \in [\![B]\!]} w_{ij} \frac{\sqrt{\mathsf{Tr}\Sigma_j^2}}{N_j} + c_2 \sum_{j \in [\![B]\!]} w_{ij}^2 \frac{\sqrt{\mathsf{Tr}\Sigma_j^2}}{N_j} \right\},$$

$$\widehat{L}_i = \left\| \sum_{j \in [\![B]\!]} w_{ij} \left(\widehat{\mu}_j^{\texttt{NE}} - \widehat{\mu}_i^{\texttt{NE}} \right) \right\|_{\mathcal{H}}^2 + s_i^2 (2w_{ii} - 1), \; \widehat{Q}_i = \sum_{j \in [\![B]\!]} w_{ij} \sqrt{\frac{\hat{\Delta}_j^T \Sigma_i \hat{\Delta}_j}{N_i}},$$

such that $\sum_{j \in [\![B]\!]} \omega_{ij} = 1$ and $\forall i, j \in [\![B]\!] : \omega_{ij} \geq 0$. There is no instantiation of $\widehat{Q}^{\mathbf{BS}}$. It is required for the theoretical results to hold on bounded data which is less regularised than Gaussian data. In practice, we add two regularization terms instead. The c_1 term favours sparse results whereas the c_2 regularization leads to diffuse, small weights. Their effect can be compared to that of ℓ_1- and ℓ_2-regularization respectively. Distant means are penalised by the c_q term.

The optimization over the probability simplex is done by exponentiated gradient descent (Kivinen and Warmuth, 1997) with gradient

$$\nabla \omega_{i\cdot}^{\texttt{AGG egd}} = 2\Big(\Lambda^{(i)} + c_2 \operatorname{diag}(\vartheta)\Big)\omega_{i\cdot} + 2S_{i\cdot} + c_q \varrho^{(i)} + c_1 \vartheta,$$

where $S_{i\cdot}$ denotes the i-th column of matrix $S = \operatorname{diag}((s_i^2)_{i \in [\![B]\!]})$, ϑ and $\varrho^{(i)}$ are B-dimensional vectors and defined as $\vartheta_j = \sqrt{\mathsf{Tr}\Sigma_j^2}/N_j$ and $\varrho_j^{(i)} = \sqrt{\hat{\Delta}_j^T \Sigma_i \hat{\Delta}_j / N_i}$. We propose the following estimators for these terms: $\hat{s}_i^2$ is estimated as shown in (V.111). Matrix $\check{\Lambda}^{(i)}$ is a biased estimator of $\Lambda^{(i)}$ with $\check{\Lambda}_{j,j'}^{(i)} = \langle \widehat{\mu}_j^{\texttt{NE}} - \widehat{\mu}_i^{\texttt{NE}}, \widehat{\mu}_{j'}^{\texttt{NE}} - \widehat{\mu}_i^{\texttt{NE}} \rangle$ that can be computed as

$$\check{\Lambda}_{j,j'}^{(i)} = \begin{cases} 0, & \text{for } i = j, \text{ or } i = j', \text{ or } i = j = j' \\ \frac{1}{N_j N_{j'}} \sum_m^{N_j} \sum_{m'}^{N_{j'}} \kappa(Z_m^{(j)}, Z_{m'}^{(j')}) \\ \quad - \frac{1}{N_j N_i} \sum_m^{N_j} \sum_n^{N_i} \kappa(Z_m^{(j)}, Z_n^{(i)}) \\ \quad - \frac{1}{N_i N_{j'}} \sum_n^{N_i} \sum_{m'}^{N_{j'}} \kappa(Z_n^{(i)}, Z_{m'}^{(j')}) \\ \quad + \frac{1}{N_i N_i} \sum_n^{N_i} \sum_{n'}^{N_i} \kappa(Z_n^{(i)}, Z_{n'}^{(i)}) \end{cases}, \quad \text{otherwise.} \tag{V.114}$$

Vector ϑ is based on $\mathsf{Tr}\Sigma_j^2$. Let X_1, X_2, X_3, X_4 denote independent copies, then

$$\Sigma = \mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^T] = \frac{1}{2}\mathbb{E}[(X_1 - X_2)(X_1 - X_2)^T]$$

such that

$$\Sigma^2 = \frac{1}{4}\mathbb{E}[(X_1 - X_2)(X_1 - X_2)^T (X_3 - X_4)(X_3 - X_4)^T].$$

By linearity of the trace, we then have

$$\begin{aligned} \mathsf{Tr}(\Sigma^2) &= \frac{1}{4}\mathbb{E}\Big[\mathsf{Tr}\Big((X_1 - X_2)(X_1 - X_2)^T (X_3 - X_4)(X_3 - X_4)^T\Big)\Big] \\ &= \frac{1}{4}\mathbb{E}[(X_1 - X_2)^T (X_3 - X_4) \cdot (X_3 - X_4)^T (X_1 - X_2)] \\ &= \mathbb{E}[\langle X_1, X_3 \rangle^2 - \langle X_1, X_3 \rangle \langle X_2, X_3 \rangle - \langle X_1, X_3 \rangle \langle X_1, X_4 \rangle + \langle X_1, X_3 \rangle \langle X_2, X_4 \rangle]. \end{aligned}$$

For $N_i \geq 4$ an unbiased estimation for $\mathrm{Tr}(\Sigma_i^2)$ is then given by

$$\begin{aligned}
&\frac{1}{N_i(N_i-1)} \sum_{n_1 \neq n_2}^{N_i} \kappa(Z_{n_1}^{(i)}, Z_{n_2}^{(i)})^2 \\
&- \frac{2}{N_i(N_i-1)(N_i-2)} \sum_{n_1 \neq n_2 \neq n_3}^{N_i} \kappa(Z_{n_1}^{(i)}, Z_{n_2}^{(i)}) \kappa(Z_{n_1}^{(i)}, Z_{n_3}^{(i)}) \\
&+ \frac{1}{N_i(N_i-1)(N_i-2)(N_i-3)} \sum_{n_1 \neq n_2 \neq n_3 \neq n_4}^{N_i} \kappa(Z_{n_1}^{(i)}, Z_{n_2}^{(i)}) \kappa(Z_{n_3}^{(i)}, Z_{n_4}^{(i)}),
\end{aligned}$$

and we recover (V.56). However this estimator has computational complexity $\mathcal{O}(N_i^4)$ and is infeasible in practice. Instead, we propose in Algorithm 1 a subsampling strategy that gives an approximation which operates in $\mathcal{O}(N_i)$.

Algorithm 1 Approximation of estimation of $\mathrm{Tr}(\Sigma_i^2)$

Require: data $Z_\bullet^{(i)}$, bag size N_i, number of repetitions r

```
#  initialise
t_1 ← 0
t_2 ← 0
t_3 ← 0
#  first term can be calculated directly in linear time
t_1 ← Σ_{n,n'}^{N_i} k(Z_n^(i), Z_n'^(i))^2 − Σ_n^{N_i} k(Z_n^(i), Z_n^(i))^2
#  other terms are approximated in r iterations
for 1 to r do
    #  select four distinct samples
    n_1, n_2, n_3, n_4 ← randint(1, N_i, 4)
    #  approximate second and third term
    t_2 ← t_2 + κ(Z_{n_1}^(i), Z_{n_2}^(i)) · κ(Z_{n_1}^(i), Z_{n_3}^(i))
    t_3 ← t_3 + κ(Z_{n_1}^(i), Z_{n_2}^(i)) · κ(Z_{n_3}^(i), Z_{n_4}^(i))
end for
#  normalise and add
trS_i ← t_1/(N_i(N_i − 1)) − 2t_2/r + t_3/r
return trS_i
```

For the vector $\varrho^{(i)}$ we need an estimation of $\hat{\Delta}_j^T \Sigma_i \hat{\Delta}_j$ for which we propose a biased estimate

$$\begin{aligned}
\mathsf{dSd}_j^{(i)} =& \frac{1}{N_i - 1} \sum_{n=1}^{N_i} \left(\frac{1}{N_j} \sum_{m=1}^{N_j} \kappa(Z_n^{(i)}, Z_m^{(j)}) - \frac{1}{N_i} \sum_{n\prime=1}^{N_i} \kappa(Z_n^{(i)}, Z_{n\prime}^{(i)}) \right)^2 \\
&- \frac{N_i}{N_i - 1} \left(\frac{1}{N_i N_j} \sum_{n=1}^{N_i} \sum_{m=1}^{N_j} \kappa(Z_n^{(i)}, Z_m^{(j)}) - \frac{1}{N_i N_i} \sum_{n=1}^{N_i} \sum_{n\prime=1}^{N_i} \kappa(Z_n^{(i)}, Z_{n\prime}^{(i)}) \right)^2 \qquad \text{(V.115)}
\end{aligned}$$

Note that estimator $\mathsf{dSd}_j^{(1)}$ is a rewriting of $\widehat{q}_j$ (V.31) in the kernel setting. For translation invariant

kernels we obtain a less biased estimate

$$\begin{aligned}\mathrm{dSd}_j^{(i)} =& \frac{1}{N_i - 1} \sum_{n=1}^{N_i} \left(\frac{1}{N_j} \sum_{m=1}^{N_j} \kappa(Z_n^{(i)}, Z_m^{(j)}) - \frac{1}{N_i - 2} \sum_{n\prime=1}^{N_i} \kappa(Z_n^{(i)}, Z_{n\prime}^{(i)}) \right)^2 \\ & - \frac{N_i}{N_i - 1} \left(\frac{1}{N_i N_j} \sum_{n=1}^{N_i} \sum_{m=1}^{N_j} \kappa(Z_n^{(i)}, Z_m^{(j)}) - \frac{1}{N_i(N_i - 2)} \sum_{n=1}^{N_i} \sum_{n\prime=1}^{N_i} \kappa(Z_n^{(i)}, Z_{n\prime}^{(i)}) \right)^2 .\end{aligned}$$

Its computational complexity is in $\mathcal{O}({N_i}^2)$.

The final procedure of `AGG egd` is shown in Algorithm 2. We suggest $c_q = 1.4, c_1 = 1, c_2 = 4$ and

Algorithm 2 `AGG egd`

Require: matrix $\check{\Lambda}^{(i)}$ (Eq. (V.114)), vectors trS (Alg. 1), $\mathrm{dSd}^{(i)}$ (Eq. (V.115)), $(\hat{s}_j)_{j \in [\![B]\!]}$ (Eq. (V.111)), model parameters c_q, c_1, c_2, learning rate η, maximum nr. of iterations $t_{\max}$

```
#  initialise
```
$\tilde{\vartheta}_j \leftarrow (\mathrm{trS}_j)^{1/2}/N_j \ , \ \forall j \in [\![B]\!]$

$\check{\varrho}_j^{(i)} \leftarrow (\mathrm{dSd}_j^{(i)}/N_i)^{1/2} \ , \ \forall j \in [\![B]\!]$

$\omega_{i\cdot}^{(0)} \leftarrow \mathbf{1}$

```
#  until maximum nr. of iterations or convergence
```
while $t \leq t_{\max}$ and $\left(\omega_i^{(t-1)} - \omega_i^{(t)}\right)^2 > 10^{-8}$ **do**

```
    #  compute gradient
```
$\nabla\omega_{i\cdot}^{(t-1)} \leftarrow 2\left(\check{\Lambda}^{(i)} + c_2 \operatorname{diag}(\tilde{\vartheta})\right)\omega_{i\cdot}^{(t-1)} + 2\hat{S}_{i\cdot} + c_q\check{\varrho}^{(i)} + c_1\tilde{\vartheta}$

```
    #  perform exponentiated gradient descent
```
$\omega_{i\cdot}^{(t)} \leftarrow \omega_{i\cdot}^{(t-1)} \cdot \exp\{-\eta^{(t)} \cdot \nabla\omega_{i\cdot}^{(t-1)}\}$

```
    #  normalise
```
$\omega_{i\cdot}^{(t)} \leftarrow \frac{\omega_{i\cdot}^{(t)}}{\mathbf{1}^T\omega_{i\cdot}^{(t)}}$

end while

```
#  estimated optimal weight vector for bag i
```
return $\omega_{i\cdot}^{(t)}$

$r = 100, t_{\max} = 500, \eta^{(t)} = 50/(1 + (t/B))$ as default parameter values.

V.11.3 `STB` Approaches

The similarity test based approaches shrink the estimation only towards neighbouring means. Neighbors are found as described in Cor. V.9,

$$\begin{aligned} W_i &= \left\{ j \in [\![B]\!] : \sqrt{\mathrm{Tr}(\Sigma_j^2)}/N_j \leq 5 \cdot \sqrt{\mathrm{Tr}(\Sigma_i^2)}/N_i \right\} \\ V_i &= \left\{ j \in W_i : \|\mu_i - \mu_j\|_{\mathcal{H}}^2 \leq \tau \cdot s_i^2 \right\}. \end{aligned} \tag{V.116}$$

In practice the quantities are estimated. Alg. 1 provides an approximation of $\mathrm{Tr}(\Sigma_i^2)$. Eq. (V.111) shows an unbiased estimate for s_i^2 and (V.113) for $\|\mu_i - \mu_j\|_{\mathcal{H}}^2$.

(vi) STB weight (Section III) assigns a uniform weight to all neighbours except for ω_{ii} which is higher

$$\omega_{ij}^{\texttt{STB weight}} = \begin{cases} \gamma + \frac{1-\gamma}{|V_i|}, & \text{for } i = j \\ \frac{1-\gamma}{|V_i|}, & \text{for } i \neq j, j \in V_i \\ 0, & \text{otherwise.} \end{cases}$$

We recall that STB weight was proposed for balanced bags and under independence of test and data. The optimal values of τ and γ are found by model optimization or $\tau = 2.2, \gamma = 0.2$ taken as default. Larger values of τ allow higher distances between μ_i and its neighbours, thus, potentially increase the number of neighbours and the bias of the estimation. Higher γ values put emphasis on μ_i, i.e., $\omega_{ii} > \omega_{ij}$ for $i \neq j$, and the solution reduces to NE for $\gamma = 1$.

(vii) STB opt corresponds to Lemma V.6 and minimizes an upper bound on the risk

$$\omega_{i\cdot}^{\texttt{STB opt}} = \underset{w_{i\cdot}}{\text{argmin}} \left\{ \tau s_i^2 (1 - w_{ii})^2 + \sum_{j \in V_i} w_{ij}^2 s_j^2 \right\},$$

such that $\sum_{j \in [\![B]\!]} \omega_{ij} = 1$ and $\forall i, j \in [\![B]\!]. \omega_{ij} \geq 0$. Using Lagrangian multipliers the optimal solution is (cf. (V.13))

$$\omega_{ij}^{\texttt{STB opt}} = \begin{cases} \lambda_i \nu_i + (1 - \lambda_i), & \text{for } i = j \\ \lambda_i \nu_j, & \text{for } i \neq j, j \in V_i \\ 0, & \text{otherwise.} \end{cases}$$

where $\nu_j := {s_j^{-2}}/{\sum_{j' \in V_i} s_{j'}^{-2}}$ and $\lambda_i := (1 + \gamma\tau(1 - \nu_i))^{(-1)}$. An unbiased estimator for s_i^2 is given in (V.111). The additional multiplicative constant γ allows for more flexibility and tends to put emphasis on ω_{ii}. Model optimization can be used to find suitable values for τ and γ. Otherwise, we recommend $\tau = 2.2, \gamma = 0.2$ as default values.

(viii) STB orth performs the similarity test and applies AGG orth on neighbouring means

$$\omega_{ij}^{\texttt{STB orth}} = \begin{cases} \omega_{ij}^{\texttt{AGG orth}}, & \text{for } j \in V_i \\ 0, & \text{otherwise.} \end{cases}$$

The similarity test merely functions as a safeguard here and excludes high distant neighbours and does not play such a central role as for the other STB methods. Therefore, τ can be fixed to a large value, e.g., $\tau := 5$. Even though $\omega_{ij}^{\texttt{AGG orth}}$ is reduced when $\|\mu_i - \mu_j\|_{\mathcal{H}}^2$ is high, AGG orth does not perform well when there are many high distant neighbours. Their weights accumulate and reduce the weights of important bags because of the normalization step. The similarity test alleviates this problem.

Either model optimization can be used to find suitable values for τ and γ, or their default values $\tau := 5, \gamma = 3$ can be chosen. Note that, compared to STB weight and STB opt, τ is larger which highlights the fact that here the similarity test only excludes distant bags. Because of this safeguard, γ, which penalises large distances, can be reduced ($\gamma = 2.2$ vs $\gamma = 13$ for AGG orth).

(ix) STB egd performs the similarity test and applies AGG egd on neighbouring means

$$\omega_{ij}^{\texttt{STB egd}} = \begin{cases} \omega_{ij}^{\texttt{AGG egd}}, & \text{for } j \in V_i \\ 0, & \text{otherwise.} \end{cases}$$

Analogous to the discussion of STB orth the similarity test functions as a safeguard to exclude high distant neighbours. It can also be seen as another instrument to replace $\widehat{Q}^{\mathbf{BS}}$ (see also discussion

of `AGG egd`). `STB egd` relies on several model parameters. We recommend to set $r = 100, t_{\max} = 500, \eta^{(t)} = 50/(1 + (t/B))$ and $\tau := 5, c_q = 1, c_1 = 1, c_2 = 5$ as default. Compared to the default values of `AGG egd`, diffuse weights should be favoured whereas regularization based on the distances (c_q) or sparse weights (c_1) become less important because of the preselection of neighbouring means.

VI A high dimensional analysis of attention

We present in this section some first results on the self-attention mechanism used recently in the Transformers neural networks. We propose to interpret the self-attention mechanism as a procedure of noise reduction of the data embeddings and link it to the previous considerations about multi-task averaging and multiple means estimation. For this purpose, we assume that the inputs of the attention mechanism are noisy observations of some underlying true vectors, belonging to a sphere and having a lower dimensional structure (small covering or belonging to a smaller dimensional subspace). Then we show that, in high dimension, for simplified forms of attention, the points built by the attention are better estimators of these vectors than the original points. Thanks to these considerations, we propose a modification of the attention which are more robust, more flexible and similar performance on toy data.

Contents

VI.1 Introduction

The attention mechanism (Bahdanau et al., 2014), popularized by the Transformers architecture (Vaswani et al., 2017), has drawn considerable attention thanks to its high performance in numerous domain such as natural language processing (Brown et al., 2020; Devlin et al., 2019) or computer vision (Dosovitskiy et al., 2020). Among its applications, it is impossible not to mention the chatbot ChatGPT (OpenAI et al., 2024), which seems poised to have a significant impact on our societies.

A self attention block takes as input a set of N points $X_1, \dots, X_N$ in $\mathbb{R}^d$, for instance each point can be the embedding of a token like a word of a sentence or a sub-zone of an image. For each of them, a new point $a_s(X_i)$ is constructed as a convex combinaison of all the points:

$$a_s(X_i) = \sum_{j=1}^{N} \omega_{ij} X_j \,, \quad \text{where} \quad \omega_{i\cdot} = \texttt{Softmax}(s(X_i, X_\cdot)) \in \mathcal{S}_N \,, \tag{VI.1}$$

where the function $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ will be called a *similarity* function and has been learned during the training. For $v \in \mathbb{R}^N$, the softmax operation is defined as $\texttt{Softmax}(v)_i = e^{v_i} / \left(\sum_{j=1}^N e^{v_j} \right)$ and the resulting vector belongs to the simplex $\mathcal{S}_N = \left\{ \omega \in [0,1]^N : \sum_{j=1}^N \omega_j = 1 \right\}$. The similarity function s is interpreted as a notion of distance between the points learned by the neural network, which permits to mix up each point with similar points. In natural language processing, this distance is easy to interpret as it has been observed that the learned similarity functions associates words of the sentence with similar meanings (Vaswani et al., 2017). The commonly chosen similarity function s_o is a scalar product between a *key* and *query* value $s_o(x, y) := \langle Qx, Ky \rangle$ for some matrices $Q, K \in \mathbb{R}^{d \times d}$ learned during the training phase.

Another one that we will also consider is the quadratic distance between the query and key values known as L2-attention (Kim et al., 2021). This similarity function is defined by $s_{L^2}(x,y) := -\|Qx - Ky\|_2^2/2$.

These new points $a_s(X_i)$ are then combined with the original ones X_i by a linear combination and a normalization, the resulting points are then:

$$A_{s,V}(X_i) := \frac{X_i + Va_s(X_i)}{\|X_i + Va_s(X_i)\|}, \quad \text{for} \quad 1 \leq i \leq N, \tag{VI.2}$$

for some learned matrix $V \in \mathbb{R}^{d\times d}$. One interpretation is that this operation combines the own information of the point with some context information derived from the others. Similarly as for the ResNets which have been linked to ordinary differential equations (Chen et al., 2018), a flow map can be derived from (VI.2) by omitting the normalization and permit to analyze dynamically the attention (Geshkovski et al., 2024; Sander et al., 2022). However this normalization is essential in practice to avoid a divergence of the points and begins to be also considered theoretically (see Geshkovski et al., 2023 for a dynamical point of view).

Thus, we propose in this section to assume that the points X_i are noisy observations of some true embedding μ_i belonging to a sphere. Then the attention operators a_s or $A_{s,V}$ can be seen as multiple vector estimators of these true embeddings μ_i. By analogy with the previous model (III.1), we will sometimes call the vectors μ_i the "means" of the vectors X_i. We show that in high dimension, if the set of these means have some underlying structure, the self attention operator has a denoising effect on the points X_i individually and in average. We consider simplified versions of attention where the matrices Q and K are equal and proportional to identity ($Q = K = I_d/\sqrt{h}$ for some scale $h > 0$). We then study the dependence in the dimension d of the scale h to avoid degenerate case and to get an improvement of the estimation of the true embeddings or means μ_i-s.

Remark VI.1. We point out that in practice the attention operations are performed separately on subgroups of coordinates with different learned similarity functions. Each subgroup is called a head and the whole is referred as multi-head attention. We will however focus here on the single head case.

VI.1.1 Discussion of the model

Before considering the cases of Transformers, it is important to notice that the main task of a neural network is to learn a good representation of the data. A multilayer perceptron in its simpler form (L layer of same width p) can be summarised as a function f_θ parameterized by $\theta = (W^L, \dots, W^1) \in \mathbb{R}^{d_{out}\times p} \times (\mathbb{R}^{p\times p})^{L-2} \times \mathbb{R}^{p\times d_{in}}$ and defined for an input $x \in \mathbb{R}^{d_{in}}$ by:

$$f_\theta(x) = W^L r(W^{L-1} r(\dots r(W^1 x))) \in \mathbb{R}^{d_{out}},$$

where r is some non-linear function applied coordinates by coordinates (e.g. the ReLU function defined by $r(x) = \max(x, 0)$ for $x \in \mathbb{R}$). We can rewrite the function f_θ as $f_\theta(x) = W^L \Phi(x)$ where Φ is the output of the first $L-1$ layers. With this notation, the neural network can be condensed in two part, first it learns a representation Φ of the data x and then realizes a linear regression in this feature space. Thus, we recover the idea of kernel methods but with a learned kernel. Equivalence has besides been proven between them in an infinite width regime and for a specific initialisation (NTK regime see Jacot et al., 2018). We can also note that the transfer learning methods, which use the first layers of previous neural network for a new task, reuse in fact the representation learned by the first network.

In a Transformer, the input is a set of vectors $X_1, \dots, X_N$ which have dependencies (words of a sentence, parts of an image,...). The block of attention permits to the neural network to learn a representation of each points depending of the others by combining their representations. To model this, we suppose that the high dimensional points X_i are noisy observations of some true representations μ_i but with independent

noise corruption. The points X_i belong to a high dimensional space but we will suppose the vectors μ_i to have a lower dimensional structure. The self attention operation permits to simplify the representation of the data by recovering this underlying structure.

A first link with vectors estimation appears by supposing the noise to be Gaussian (i.e. $X_i \sim \mathcal{N}(\mu_i, \Sigma)$) and that the vectors μ_i are drawn from a common unknown distribution. Then the Bayes estimator of μ_i is:

$$\mathbb{E}[\mu_i | X_i] = X_i + \Sigma \frac{\nabla g(X_i)}{g(X_i)}\,, \tag{VI.3}$$

where g is the marginal density of the distribution of X_1 (Brown, 1971). From this identity, Brown and Greenshtein (2009) propose to estimate μ_i by replacing g by its kernel density estimator $\widehat{g}_h$ from the sample $X_1, \ldots, X_N$. Using a Gaussian kernel, then $\widehat{g}_h(x) = C_{N,h} \sum_{j=1}^{N} e^{-\|x - X_j\|^2/(2h)}$ and after injecting it into (VI.3), we get an estimator $\widehat{\mu}_i$ close to the attention vector with the L^2 similarity function (for $Q = K = I_d/\sqrt{h}$):

$$\widehat{\mu}_i = X_i + \Sigma \frac{\nabla \widehat{g}_h(X_i)}{\widehat{g}_h(X_i)} = \left(I_d - \frac{\Sigma}{h}\right) X_i + \frac{\Sigma}{h} a_{s_{L^2}}(X_i)\,. \tag{VI.4}$$

This estimator is a shrinkage estimator (with a matricial factor) of the vector X_i to the reference point $a_{s_{L^2}}(X_i)$ built from the other observations. Up to a normalization and a matrix multiplication, we recover the formula of the normalized step of attention $A_{s_{L^2},V}$ for some specific matrix V:

$$A_{s_{L^2},V}(X_i) = \frac{(I_d - \Sigma/h)^{-1}\widehat{\mu}_i}{\|(I_d - \Sigma/h)^{-1}\widehat{\mu}_i\|} \overset{\Sigma=\sigma^2 I_d}{=} \frac{\widehat{\mu}_i}{\|\widehat{\mu}_i\|}\,, \tag{VI.5}$$

where the first equality is satisfied for $V = (h\Sigma^{-1} - I_d)^{-1}$ and the second for $\Sigma = \sigma^2 I_d$. This last equation justifies one main assumption of this work: as the outputs of the self attention are normalized in a Transformer, we consider that only the directions of the vectors matter in the learned representation and then suppose that the true representations μ_i-s are of same norm. Under these conditions we can wonder if the attention operator a_s improves the estimation of these vectors for each point, i.e.

$$\mathbb{E}\Big[\|a_s(X_1) - \mu_1\|^2\Big] < \mathbb{E}\Big[\|X_1 - \mu_1\|^2\Big] = d\sigma^2\,,$$

or in average:

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\Big[\|a_s(X_i) - \mu_i\|^2\Big] < d\sigma^2\,.$$

We choose to focus on a_s and not $A_{s,V}$ to avoid to make additional assumptions on V. In fact, considering that the neural network only learns in this phase a representation of the data to apply a linear regression, two representations are equivalent up to a linear transform. Then, in view of (VI.5), for some matrix V depending on the covariance Σ, $A_{s,V}$ will then be an estimator of μ_i or of a dilatation of this vector with error smaller than an estimator using only X_i. Such a matrix could have been learned in the training phase. Although these questions are fundamental to a deeper analysis, we elude them for the moment by concentrating on $a_s(X_i)$ and showing that it is a better estimator of μ_i than X_i.

Notation: For some quantities a_d and b_d depending of the dimension, $a_d = O(b_d)$ and $a_d = o(b_d)$ denotes respectively that the ratio a_d/b_d is upper bounded or goes to 0 as the dimension increases; $a_d = \Omega(b_d)$ and $a_d = \omega(b_d)$ are respectively equivalent to $b_d = O(a_d)$ and $b_d = o(a_d)$; $a_d = \Theta(b_d)$ means that the ratio a_d/b_d is lower and upper bounded. The notation C denotes constants independent of d and N whose values can differ between equations.

VI.2 Theoretical results

In Section VI.2.1, we study the effect of high dimension denoising of the self attention mechanism for different classical similarity functions. In Section VI.2.2 we propose two slightly modified versions of attention which are more flexible and robust and achieve a denoising effect similar to the originals.

VI.2.1 Denoising by attention

As announced, we suppose that each point X_i is a noisy observation of an underlying mean $\mu_i \in \mathbb{S}_{d-1}(R)$ where $\mathbb{S}_{d-1}(R)$ is the d-dimensional sphere of radius R of $\mathbb{R}^d$. Formally, we suppose that

$$X_i \sim \mathcal{N}(\mu_i, \sigma^2 I_d), \quad \|\mu_i\|_2 = R \quad 1 \leq i \leq N. \tag{VI.6}$$

We point out that we assume the true means of same norm but not on the unit sphere. Indeed as we are interested in the high dimensional behavior of the self attention ($d \to \infty$), we need to suppose the squared radius R^2 and the quadratic risk $\sigma^2 d$ of same order to avoid trivial cases. Indeed if $R^2 \ll d\sigma^2$ the signal is indistinguishable from the noise or conversely if $R^2 \gg d\sigma^2$, there is no noise to reduce. That leads us to the following assumption.

Assumption VI.2. *$R^2 = \Theta(\sigma^2 d)$ i.e. there exists a constant $C_R > 0$ such that $C_R^{-1} \leq R^2/d\sigma^2 \leq C_R$.*

As we only want that the ratio $R^2/\sigma^2 d$ is lower and upper bounded as the dimension increases, we could have also assumed that R^2 is fixed and that the noise σ^2 goes to 0 when the dimension increases. We find however more natural to consider the noise σ^2 fixed.

Since the means μ_i are on the sphere, we can remark that the scalar products between them behave as the quadratic distance, indeed $\langle \mu_i, \mu_j \rangle = R^2 - \|\mu_i - \mu_j\|^2/2$. Then for any $h > 0$, the softmax weights of the scalar product and of the quadratic distance are the same:

$$\texttt{Softmax}\left[\left(\frac{\langle \mu_1, \mu_i \rangle}{h}\right)_i\right] = \texttt{Softmax}\left[\left(-\frac{\|\mu_i - \mu_1\|^2}{2h}\right)_i\right].$$

However this equality is not anymore verified if the vectors μ_i are replaced by their noisy observations X_i. This consideration leads us to consider as similarity functions the scalar product and the squared norm. We will also consider the *projected* similarity function s_p which is the scalar product of the vectors projected on the sphere and which gives the same weights after the softmax as the quadratic distance between them. We will observe similar behavior for these three functions. So, in this section, we consider the following similarity functions.

Definition VI.3. For all $x, y \in \mathbb{R}^d$, we consider the similarity functions:

$$\begin{aligned} s_o(x,y) &= \langle x, y \rangle, && \textit{(original)}, \\ s_{L^2}(x,y) &= -\frac{\|x-y\|^2}{2}, && \textit{(L2)}, \\ s_p(x,y) &= \left\langle \frac{Rx}{\|x\|}, \frac{Ry}{\|y\|} \right\rangle, && \textit{(projected)}. \end{aligned}$$

For s a similarity function and a scale $h > 0$, we will denote in the rest of this section $a_s(X_i)$ the attention points (VI.1) associated to the similarity function s/h which are then defined for $1 \leq i \leq N$ by:

$$a_s(X_i) := a_{s,h}(X_i) = \sum_{j=1}^{N} \omega_{ij}^{s,h} X_j \quad \text{where} \quad \left(\omega_{ij}^{s,h}\right)_j = \texttt{Softmax}\left(\left(\frac{s(X_i, X_j)}{h}\right)_j\right) \in \mathcal{S}_N\,. \tag{VI.7}$$

We will sometimes omit the dependence in h or s for the attention $a_{s,h}$ or the weights $\omega^{s,h}$. For these similarity functions, Lemma VI.4 gives the right rate for the scale h in the dimension such that the weights of the self attention are not trivial.

Lemma VI.4. *Let s be either s_o, s_{L^2} or s_p and suppose Assumption VI.2 satisfied. Then, if $h = o(d)$, for $1 \leq i \leq N$:*

$$\omega_{ii}^{s,h} \xrightarrow[d\to\infty]{a.s.} 1\,,$$

where ω is defined in (VI.1). Conversely, if $d = o(h)$, then for $1 \leq i, j \leq N$:

$$\left|\omega_{ij}^{s,h} - \frac{1}{N}\right| \xrightarrow[d\to\infty]{a.s.} 0\,.$$

In the first regime ($h = o(d)$) all the weights of the self attention tend to be given to the point itself. Then the transformation converges to the identity and does not use any similarity between the points. Inversely, for a large scaling h, a same weight is given to all the points but nothing is learned either. Only an empirical mean is done without trying to exclude different points in the convex combination. Hence, for these similarity functions, a right regime to consider is $h = \Theta(d)$. For this scaling, we observe that if the means μ_i belong to a lower dimensional space, the resulting points are brought closer to this subspace.

Proposition VI.5. *Let s be either s_o or s_{L^2}, there exists an absolute constant $C_1 > 0$ such that for $d \geq \log N$, if R satisfies Assumption VI.2 with constant C_1 and $h \geq C_1 d\sigma^2$, then for $1 \leq i \leq N$:*

$$\frac{\mathbb{E}\left[\delta^2(a_s(X_i), \mathcal{M})\right]}{\mathbb{E}[\delta^2(X_i, \mathcal{M})]} \leq C \max\left(\frac{1}{N}, \frac{\log N}{\sqrt{d-m}}\right), \tag{VI.8}$$

where $\delta(x, \mathcal{M}) := \inf_{y\in\mathcal{M}} \|x - y\|$ denotes the Euclidean distance of a point x to the subspace $\mathcal{M}$ and C is an absolute constant.

This result is verified for $h = \Omega(d)$, but is only relevant for $h = \Theta(d)$. Indeed, if $h = \omega(d)$, as presented in Lemma VI.4, the attention points $a(X_i)$ tend to be the empirical mean and then directly the noise would be reduced by a factor N^{-1}. However the empirical mean can be far from the means μ_i as soon as the vectors μ_i are different. Effectively, Proposition VI.5 only considers the noise in the orthogonal of the set of means. So an important remaining question is to know if the resulting points $a_s(X_i)$ get really closer to their means μ_i. In fact, for the similarity functions of Definition VI.3 we did not manage to exhibit a theoretical bound for the improvement in high dimension in all generality. However, in Proposition VI.6, we present that in some cases, the attention cannot lead to an improvement. This situation happens for example when all the means are distributed among orthogonal directions such as the canonical basis.

Proposition VI.6. *Let s be either s_o or s_{L^2}, there exists some means $\mu_1, \ldots, \mu_N$ on the sphere $R\mathbb{S}_d$ and an absolute constant $C_1 > 0$ such that for $d \geq \log N$, if R satisfies Assumption VI.2 with constant C_1, then for all $h \geq C_1 d\sigma^2$ and $1 \leq i \leq N$:*

$$\frac{\mathbb{E}\left[\|a_s(X_i) - \mu_i\|^2\right]}{\mathbb{E}\left[\|X_i - \mu_i\|^2\right]} \geq c\left(1 - C\max\left(\frac{1}{N}, \frac{1}{d}, \frac{\log N}{\sqrt{d}}\right)\right) \tag{VI.9}$$

for some absolute constants $C, c > 0$.

To interpret this result correctly, it is important to keep in mind that we are analyzing the error under a dimensional asymptotic point of view. Although the constant c may be less than 1 in the bound (VI.9), it does not depend on either d or N. Thus, even for a number of data N tending to infinity with the dimension (under the weak condition $\log N \leq d$), we have exhibited a situation where the improvement remains lower bounded by a constant. In practice this constant c seems to be higher than 1 as no improvement is observed in our experiments (see Figure 6).

This impossibility is due to the limitation on the scale h. As $h = \Theta(d) = \Theta(R^2)$, the attention cannot succeed in excluding the points whose means are far from the target one. The condition $h = \Theta(d)$, needed to avoid trivial cases (Lemma VI.4), is in fact due to a form of bias in the weights. To resolve this problem, we propose in next section two slightly modified similarity functions for which a wider field of scale h is possible while avoiding these trivial cases.

VI.2.2 Debiased attention

We have observed in Lemma VI.4 that the scale h needs to be proportional to the space dimension for the similarity functions s_o, s_{L^2} and s_p. This phenomenon is caused by a form of "bias" of the diagonal weights. Indeed, the similarity functions considered can be seen as estimators of the distance between the true means μ_i, for $j \neq i$:

$$\mathbb{E}[s_o(X_i, X_j) - s_o(X_i, X_i)] = \mathbb{E}[s_{L^2}(X_i, X_j) - s_{L^2}(X_i, X_j)] = -\|\mu_i - \mu_j\|^2/2 - d\sigma^2 \,.$$

It is indeed completely equivalent to consider the similarity function $\widetilde{s}(x, y) := s(x, y) - s(x, x)$ for some similarity function s as the weights given by each of them are the same ($\texttt{Softmax}[(s(X_i, X_j)/h)_j] = \texttt{Softmax}[(\widetilde{s}(X_i, X_j)/h)_j]$). As the non-diagonal weights will be proportional to $e^{-d\sigma^2/h}$, choosing $h = \Theta(d)$ is mandatory to avoid the trivial cases (we recover Lemma VI.4). To avoid this bias, we propose two slightly modified similarity functions such that $s(X_i, X_j) - s(X_i, X_i)$ is an unbiased estimator of the distance between μ_i and μ_j.

Definition VI.7. The functions $s_m, s_{db} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ are respectively the *modified* and *debiased* similarity functions and are defined by:

$$s_m(x, y) = s_{db}(x, y) = \langle x, y \rangle \quad \text{for} \quad x \neq y \in \mathbb{R}^d \,,$$
$$s_m(x, x) = R^2 \,, \quad s_{db}(x, x) = \|x\|^2 - d\sigma^2 \quad \text{for} \quad x \in \mathbb{R}^d \,.$$

The similarity functions s_m and s_{db} only differ from s_o by their value on the diagonal. Our intuition is as $s_o(X_i, X_i)$ approximates $\|\mu_i\|^2 = R^2$ with a bias of $d\sigma^2$, we can either substract this bias (like in s_{db}) or just replace this value by its known true value R^2 (like in s_m). Hence, for all i, j:

$$\mathbb{E}[s_m(X_i, X_j) - s_m(X_i, X_i)] = \mathbb{E}[s_{db}(X_i, X_j) - s_{db}(X_i, X_j)] = -\|\mu_i - \mu_j\|^2/2 \,.$$

For these similarity functions, the self attention improves the estimation for a wider range of h. Effectively, we observe a noise reduction of the means μ_i if these are covered by a small number of balls. We show in this case that for each point and in average over all the points, the points $a_s(X_i)$ are closer to the mean μ_i than the initial point X_i.

Proposition VI.8. *Let s be either s_m or s_{db} and $h = C_h \sigma^2 d^\beta$ for $\beta \geq 1/2$, there exists an absolute constant $C_1 > 0$ such that for $d \geq \log N$, if R satisfies Assumption VI.2 with constant C_1 and if $C_h > C_1$, then for $1 \leq i \leq N$:*

$$\frac{\mathbb{E}\left[\|a(X_i) - \mu_i\|^2\right]}{\mathbb{E}\left[\|X_i - \mu_i\|^2\right]} \leq C \max\left(\frac{1}{V_i}, \frac{\log N}{d^{1-\beta}}\right), \tag{VI.10}$$

where $V_i = \sum_{j=1}^N e^{-\|\Delta_{ij}\|^2/(2h)} \geq 1$ and $\Delta_{ij} = \mu_i - \mu_j$. Under the same conditions, if $\mathcal{N}_h$ is the covering number of the set of the means $\{\mu_i\}_{i \in [\![N]\!]}$ by balls of radius $\sqrt{h}$, then:

$$\frac{1}{N}\sum_{i=1}^N \frac{\mathbb{E}\left[\|a(X_i) - \mu_i\|^2\right]}{\mathbb{E}\left[\|X_i - \mu_i\|^2\right]} \leq C \max\left(\frac{\mathcal{N}_h}{N}, \frac{1 + \log \mathcal{N}_h}{d^{1-\beta}}, \sqrt{\frac{\log N}{d}}\right). \tag{VI.11}$$

The quantity C denotes an absolute constant possibly different between lines.

These bounds illustrate that the proposed modified similarity functions are more flexible as h can be taken down to $\Theta(\sqrt{d})$ and still induce a potential noise reduction. As the bound decreases only if $h = o(d)$, we get that h can take a value between $\Theta(\sqrt{d})$ and $\Theta(d)$.The improvement depends on a trade-off between

the bias induced by a large scale h and the variance reduction induced by a high number of close means (V_i high or $\mathcal{N}_h$ small). If all the vectors μ_i are equal, we get an improvement up to $\frac{1}{N}$ (for d large enough) as an empirical mean. Inversely, in a worst case where all the means are isolated for instance at a distance $R \simeq \sqrt{d}$, then the scale h needs to be of order $\Theta(d)$ for having a covering number smaller than N. Then the second term of the maximum will be greater than 1 and the bound does not predict any improvement. This worst case is similar as the one considered in Proposition VI.6. This higher flexibility is observed in the following experiments.

VI.3 Experiments

In this section we validate the observations made in the theoretical part on synthetic data. The points considered are Gaussian random vectors with means on the sphere $R\mathbb{S}_d$:

$$X_i \sim \mathcal{N}(\mu_i, \sigma^2 I_d)\,, \quad \|\mu_i\| = R\,, \quad 1 \leq i \leq N.$$

For all the experiments, the radius will be fixed to $R = \sqrt{d}$ and the noise will be equal to $\sigma^2 = 0.5$ such that an improvement is possible as the noise is smaller than the radius $R > \sigma\sqrt{d}$. We consider three different settings for the distribution of the means. In a first one, the means are clustered into $\mathcal{N}_{\text{mean}}$ groups, in a second one, the means are still clustered but belong to a m dimensional subspace and the last one is the negative setting of Proposition VI.6 which defeats the first methods. In this last setting the means are equally distributed between each directions. Let P_d be the projection on the sphere $\mathbb{S}_{d-1}$, the settings considered for the means are the following :

1. **Clustered setting:** The means are clustered around $\mathcal{N}_{\text{mean}}$ points δ_j. The repartition between the clusters is uniform. Formally: $\mu_i = R\,P_d\Big(\delta_{J_i} + \varepsilon_i\Big)$ where $J_i \sim \mathcal{U}(\llbracket \mathcal{N}_{\text{mean}} \rrbracket)$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2_{\text{mean}} I_d)$ with $\sigma^2_{\text{mean}} > 0$ and $\delta_j \sim \mathcal{N}(0, I_d)$ for $i \in \llbracket N \rrbracket$ and $j \in \llbracket \mathcal{N}_{\text{mean}} \rrbracket$.

2. **Low dimensional setting:** Same setting as the clustered setting but the means belong to a m-dimensional subspace $\mathcal{M}$.

3. **Worst setting:** The means are separated into d clusters of size $\lfloor N/d \rfloor$ or $\lceil N/d \rceil$ and for a cluster k, each mean of the cluster is equal to Re_k where $(e_k)_{1 \leq \leq d}$ is the canonical basis of $\mathbb{R}^d$.

The parameters $\mathcal{N}_{\text{mean}}$, σ^2_{mean} for the clustered setting will be fixed and independent of the ambient dimension d. We will call *original*, *L2*, *modified* and *debiased* the respective self attention methods induced by the similarity function s_o, s_{L^2}, s_m and s_{db}, i.e. for s one of these functions and for $1 \leq i \leq N$:

$$a_s(X_i) = \sum_{j=1}^{N} \omega_{ij} X_j \quad \text{where} \quad \omega_{ij} = \frac{e^{s(X_i, X_j)/h}}{\sum_{\ell=1}^{N} e^{s(X_i, X_\ell)/h}}\,, \tag{VI.12}$$

for some scale h. The last method called *projected* method, is the original method applied to the points projected on the sphere, i.e. for $1 \leq i \leq N$:

$$a_{s_p}(X_i) = \sum_{j=1}^{N} \omega_{ij} \frac{R\,X_j}{\|X_j\|} \quad \text{where} \quad \omega_{ij} = \frac{e^{s_p(X_i, X_j)/h}}{\sum_{\ell=1}^{N} e^{s_p(X_i, X_\ell)/h}}\,. \tag{VI.13}$$

We evaluate the efficiency of these five methods by computing the ratios of the quadratic distance to the true means of the projected points after attention and the projected points before. Knowing that the means are on the sphere, we consider that the reference estimator of a vector μ_i with one observation is the

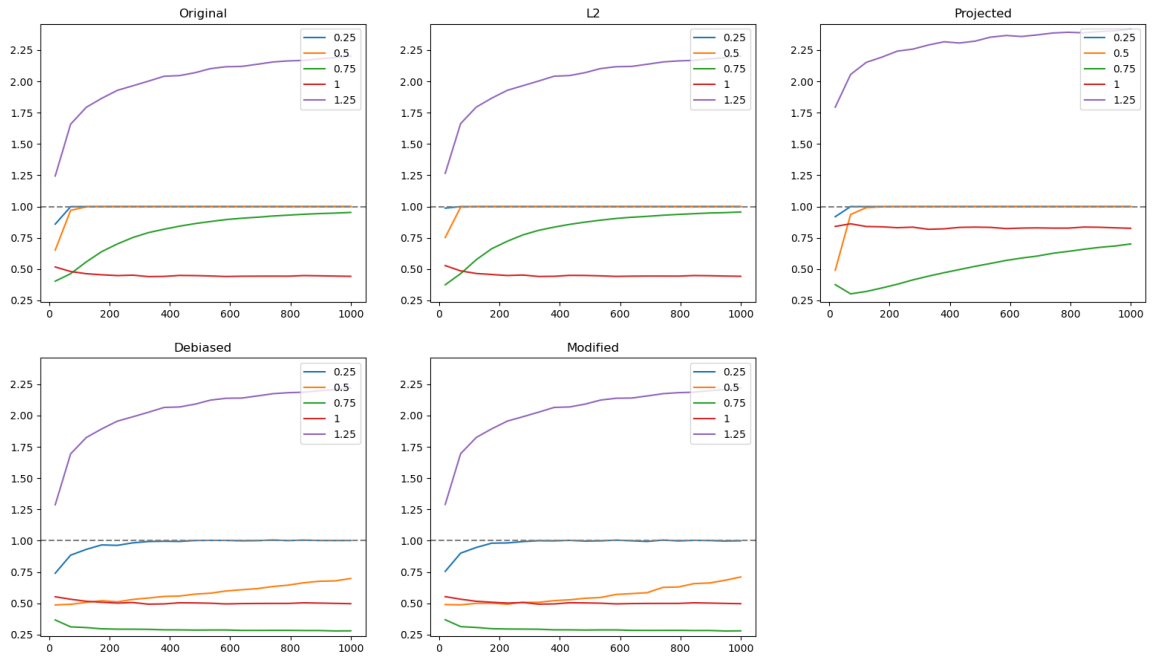

Figure 4: Relative risk $\bar{\mathcal{R}}$ in the **clustered setting** in function of the dimension for the different methods and for different scales $h = \sigma^2 d^\beta$ for $\beta \in \{0.25, 0.5, 0.75, 1, 1.25\}$ in function of the space dimension with $N = 100$, $\sigma^2_{\text{mean}} = 0.1$, $\sigma^2 = 0.5$ and $\mathcal{N}_{\text{mean}} = 4$. Lower is better.

projection on the sphere of this observation. For each similarity functions s and a N sample of observations and means $(X_\bullet, \mu_\bullet)$, we compute:

$$\widehat{\mathcal{R}}_{s,d}(X_\bullet, \mu_\bullet) := \frac{1}{N}\sum_{i=1}^{N} \frac{\|RP_d(a_s(X_i)) - \mu_i\|^2}{\|RP_d(X_i) - \mu_i\|^2}\,.$$

As $\mathbb{E}\left[\|RP_d(X_i) - \mu_i\|^2\right] \simeq \mathbb{E}\left[\|X_i - \mu_i\|^2\right] \simeq d\sigma^2$ (see Lemma VI.12), this measure of performance is similar to the ones considered in the theoretical results. We repeat this operation M times for each dimensions d and average the improvements $\bar{\mathcal{R}}_{s,d} = \frac{1}{M}\sum_{\ell=1}^{M} \widehat{\mathcal{R}}_s(X^\ell_\bullet, \mu^\ell_\bullet)$ and call this quantity the relative risk of a similarity function s for a dimension d.

Figure 4 presents the average risk $\bar{\mathcal{R}}$ for the different methods in the clustered setting. As expected by Lemma VI.4, for $h = o(d)$ the relative risk tends to 1 for the original, L2 and projected similarity functions as the attention operator converge to the identity. An improvement is only obtained for a scale $h = \Theta(d)$. For higher scale, the attention is worse than the simple individual point projection as it tends to an empirical mean of all points. For similarity functions s_{db} and s_m, we effectively get a stable improvement with the dimension for a scale h between $\sqrt{d}$ and d. For a smaller scale, no more improvement is obtained and for higher, the estimation is worsen than simple projection as was the case for the original s_o and s_{L^2} attention.

For the low dimensional setting, we group the improvements given by the original and L2 on one side and the debiased and modified on the other as they give similar results. In addition to the relative risk $\bar{\mathcal{R}}_s$, we evaluate in this setting the relative distance to the subspace computed as the ratio of the distance to the subspace $\mathcal{M}$ before and after attention:

$$\widehat{\delta}^2_s(X_\bullet, \mu_\bullet) := \frac{1}{N}\sum_{i=1}^{N} \frac{\delta^2(RP_d(a_s(X_i)), \mathcal{M})}{\delta^2(RP_d(X_i), \mathcal{M})}\,,$$

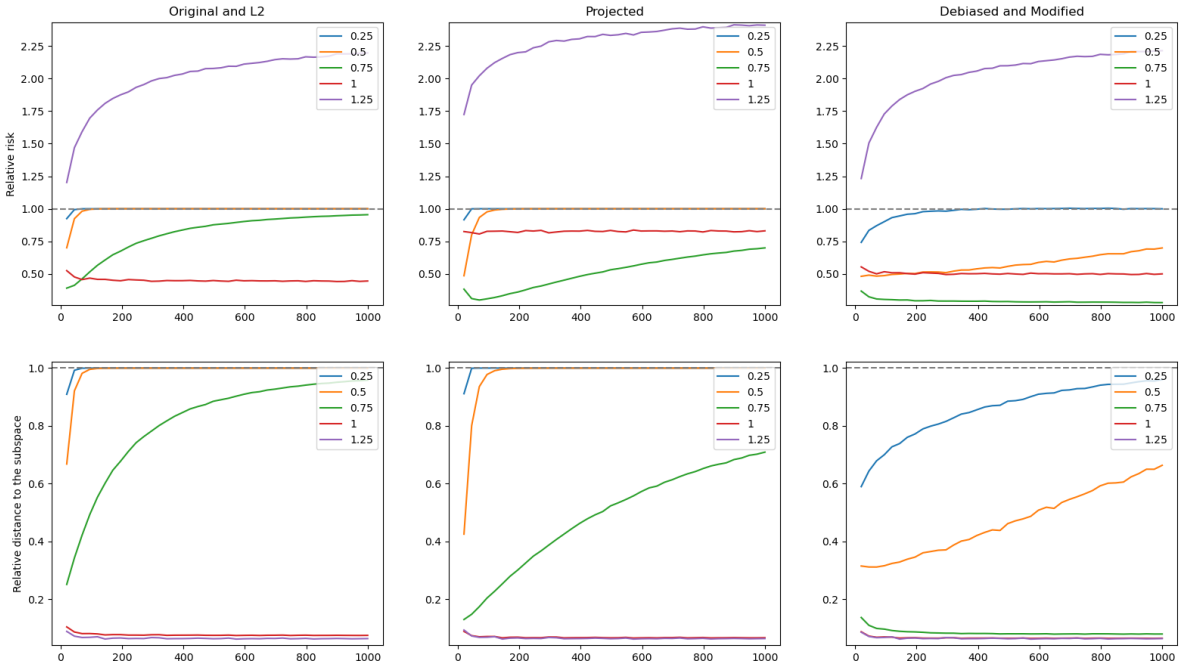

Figure 5: Relative risk $\bar{\mathcal{R}}$ (upper row) and relative distance to the subspace (lower row) in the **low dimensional setting** in function of the dimension for the different methods and for different scales $h = \sigma^2 d^\beta$ for $\beta \in \{0.25, 0.5, 0.75, 1, 1.25\}$ in function of the space dimension with $m = d/2$, $N = 100$, $\sigma^2_{\text{mean}} = 0.1$, $\sigma^2 = 0.5$ and $\mathcal{N}_{\text{mean}} = 4$. Lower is better.

where $\delta(x) = \min_{y \in \mathcal{M}} \|x - y\|$ as defined in Proposition VI.5. We can observe in Figure 5 that for the original, L2 and projected similarity functions, we recover the three regimes $h = o(d), h = \Theta(d)$ and $h = \omega(d)$ where we get respectively no improvement, a sensible improvement both in the distance to the means and to the subspace and only an improvement in distance of the subspace $\mathcal{M}$. For the debiased and modified attention, for any scale h, the attention brings the points towards the lower dimensional subspace. The estimations of the means are improved again for $\sqrt{d} \leq h \leq d$. We can remark that the scale $h = \Theta(\sqrt{d})$ is a bit smaller, and the relative risk get close to 1 with the dimension. We point out that in this setting the dimension of the subspace $\mathcal{M}$ grows with the ambient dimension ($\dim \mathcal{M} = d/2$).

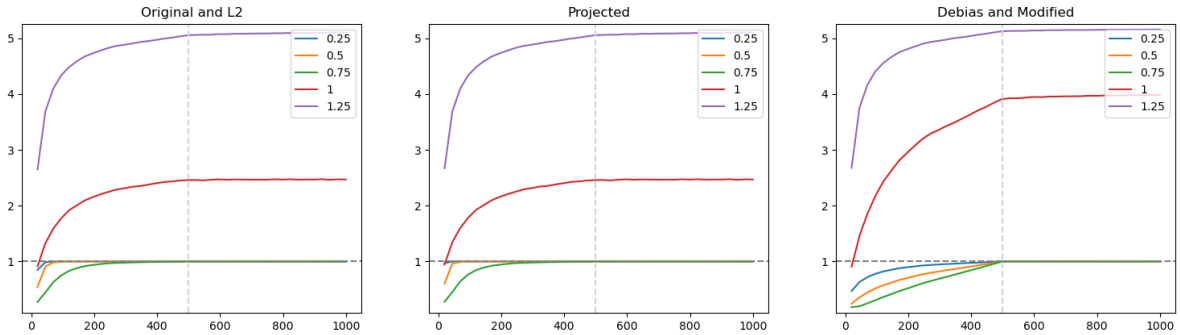

Figure 6: Relative risk $\bar{\mathcal{R}}$ in the **worst setting** in function of the dimension for the different methods and for different scales $h = \sigma^2 d^\beta$ for $\beta \in \{0.25, 0.5, 0.75, 1, 1.25\}$ in function of the space dimension with $N = 500$ and $\sigma^2 = 0.5$. The vertical line indicates $N = d$. Lower is better.

The so-called worst setting (Figure 6) makes all methods fail for $h = \Omega(d)$. This is a problem for

the original, L2 and projected methods as an improvement is only possible for $h = \Theta(d)$ as witnessed in previous experiments (Figure 4 and Figure 5). However for the debiased and modified methods, for a scale h of smaller order, improvements are possible and in these cases the estimation is not worsened and is even still improved for lower dimension ($d \leq N$). These results illustrate the greater flexibility of these slightly modified similarity functions.

To sum up, we have evaluated the denoising effect of the attention by comparing the quadratic distances of the projected points to their means before and after the attention operation.

- For the original, L2 and projected similarity functions, we have observed a lack of robustness of the methods. The scale h needs to be of order $\Theta(d)$ to obtain a denoising effect in some cases, but in others, with this scale, the observations are deteriorated.
- For the modified and debiased similarity functions, a whole range of scales h provides sensible improvements and, in bad cases where the means are far from each others, the observations are preserved contrary to the other similarity functions.

VI.4 Conclusion

We have analyzed the attention mechanism as a multiple estimation of vectors and have shown theoretically and experimentally that simplified versions of the self attention mechanism has a denoising effect on these vectors. This effect is noticeable when the true means have a structure and we have considered the case where they belong to a low dimensional subspace or can be covered by a small number of balls. A rough interpretation of this effect is that the attention mechanism extract from the set of points some underlying true information which can be parameterized by less than Nd parameters. Our simplified attention functions are only dependent of one real parameter h and we exhibit different behavior of the attention in function of this parameter. Thanks to this analysis, we proposed two natural modified similarity functions more flexible and with better performance on synthetic data.

These first results need to be further explored in more general cases. Indeed we have only considered some simplified form of attention, neglected the learning phase and modeled the embedding by independent Gaussian vectors which is debatable. To take them into account, we could adopt the in-context modelization as some recent works have done to study attention (Garg et al., 2022) and hence, suppose that the means are themselves drawn from an underlying distribution. This would fit with the empirical Bayes interpretation (Brown and Greenshtein, 2009, see Section VI.1.1). In this model, we expect that the attention would learn a similarity function adapted to the distribution.

On the experimental part, the effect of our modified versions of attention on Transformer should be investigated. Since a larger range of values is possible for the scale, we expect a faster learning phase. Our versions can also be adapted to general query and key matrices and in that case, the attention would not used the Euclidean distance but a distance defined by these matrices. With this different geometry, or with a non isotropic noise, we can expect the appearance of a notion of effective dimension.

VI.5 Proofs for Section VI

VI.5.1 Proof of Lemma VI.4

The proof is made for each of the different similarity functions. We begin y the result for small scale h and continue by the second part of the result. In each cases the weights are controlled by concentration inequalities. In this proof, $\varepsilon_i := X_i - \mu_i$ for all $1 \leq i \leq N$.

Case $h = o(d)$: To show the convergence of the weights, we will show in each case that the difference $(s(X_i, X_i) - s(X_i, X_j))/h$ decreases to minus infinity for all i, j. Hence the diagonal weights ω_{ii} converge

to 1.

Proof for $s = s_o$: According to Lemma 8.1 of Birgé (2001), using an union bound, for all $t \geq 0$, with probability at least $1 - Ne^{-t}$:

$$\|X_i\|^2 \geq d\sigma^2 + R^2 - 2\sigma\sqrt{(2R^2 + \sigma^2)t}\,, \quad \text{for} \quad 1 \leq i \leq N. \tag{VI.14}$$

Moreover, for all $i \neq j$:

$$\langle X_i, X_j \rangle = \langle \mu_i, \mu_j \rangle + \langle \varepsilon_i, \mu_j \rangle + \langle \varepsilon_j, \mu_i \rangle + \langle \varepsilon_i, \varepsilon_j \rangle\,.$$

Then according to Proposition III.7 and the concentration of a normal random variable, for all $t \geq 0$, with probability at least $1 - 3N(N-1)e^{-t}$, for all $1 \leq i, j \leq N$:

$$\langle X_i, X_j \rangle \leq \langle \mu_i, \mu_j \rangle + 2R\sigma\sqrt{2t} + \sigma^2\sqrt{2dt} + \sigma^2 t \leq R^2 + C\sigma\sqrt{(R^2 + \sigma^2 d)t} + \sigma^2 t\,, \tag{VI.15}$$

for some absolute constant $C > 0$. Then after combining these two bounds, we get a lower bound on ω_{ii} for all $1 \leq i \leq N$. With probability at least $1 - 3N^2e^{-t}$:

$$\begin{aligned}\omega_{ii} = \frac{e^{\|X_i\|^2/h}}{e^{\|X_i\|^2/h} + \sum_{j\neq i} e^{\langle X_i, X_j\rangle/h}} &\geq \left(1 + (N-1)\exp\left[\left(-d\sigma^2 + C\sigma\sqrt{(R^2+\sigma^2 d)t} + \sigma^2 t\right)/h\right]\right)^{-1} \\ &\geq 1 - (N-1)\exp\left[\left(-d\sigma^2 + C\sigma\sqrt{(R^2+\sigma^2 d)t} + \sigma^2 t\right)/h\right].\end{aligned}$$

The second term goes to 0 as $d \to \infty$. Then for all $\varepsilon \in (0,1)$,

$$\mathbb{P}\big[\max_{1\leq i\leq N} |\omega_{ii} - 1| \geq \varepsilon\big] \leq 3N^2 e^{-t_\varepsilon}$$

where $t_\varepsilon > 0$ is solution of

$$(N-1)\exp\left[\left(-d\sigma^2 + C\sigma\sqrt{(R^2+\sigma^2 d)t_\varepsilon} + \sigma^2 t_\varepsilon\right)/h\right] = \varepsilon\,.$$

Quantity t_ε is well defined for d large enough as by assumption $R^2 = \Theta(d)$ and $h = o(d)$. In particular $t_\varepsilon = \Omega(d)$ and we can conclude using Borel-Cantelli Lemma.

Proof for $s = s_{L^2}$: According to Lemma 8.1 of Birgé (2001), using an union bound, for all $t \geq 0$, with probability at least $1 - N^2e^{-t}$:

$$\|X_i - X_j\|^2 \geq 2d\sigma^2 - C\sigma\sqrt{(R^2 + \sigma^2 d)t}\,, \quad \text{for} \quad 1 \leq i \neq j \leq N, \tag{VI.16}$$

for some absolute constant $C > 0$. We have lower bounded $\|\mu_i - \mu_j\|^2$ by 0 and upper bounded by CR^2. Then with probability at least $1 - N^2e^{-t}$, for all $1 \leq i \leq N$.:

$$\begin{aligned}\omega_{ii} = \left(1 + \sum_{j\neq i} e^{-\|X_i - X_j\|^2/2h}\right)^{-1} &\geq \left(1 + (N-1)\exp\left[\left(-d\sigma^2 + C\sigma\sqrt{(R^2+\sigma^2 d)t}\right)/h\right]\right)^{-1} \\ &\geq 1 - (N-1)\exp\left[\left(-d\sigma^2 + C\sigma\sqrt{(R^2+\sigma^2 d)t}\right)/h\right].\end{aligned}$$

Similarly as the previous case we get the almost sure convergence using by instance Borel-Cantelli Lemma.

Proof for $s = s_p$: Let us denote $\delta(t) := \sigma\sqrt{(R^2 + \sigma^2 d)t}$ for all $t \geq 0$. Then combining concentration bounds (VI.14) and (VI.15), with probability greater than $1 - 3N^2e^{-t}$, for all $1 \leq i, j \leq N$:

$$\frac{\langle X_i, X_j\rangle}{\|X_i\|\|X_j\|} \leq \frac{R^2 + \delta(t) + \sigma^2 t}{\|X_i\|\|X_j\|} \leq \frac{R^2 + C\delta(t) + \sigma^2 t}{d\sigma^2 + R^2 - C\delta(t)}$$

for some absolute constant $C > 0$ and t smaller than d (otherwise the bound vanishes). Then we can lower bound the diagonal weights ω_{ii}, with probability greater than $1 - 3N^2e^{-t}$, for all $1 \leq i \leq N$:

$$\omega_{ii} = \frac{e^{R^2/h}}{e^{R^2/h} + \sum_{j\neq i} e^{\frac{R^2}{h}\frac{\langle X_i, X_j\rangle}{\|X_i\|\|X_j\|}}} \geq \left(1 + (N-1)\exp\left(-\frac{R^2}{h}\frac{d\sigma^2 - \delta(t) - \sigma^2 t}{d\sigma^2 + R^2 - \delta(t)}\right)\right)^{-1}.$$

For t/d small enough, the right term goes to 1 as the dimension increases. Indeed for $\lambda > 0$ small enough, $\delta(\lambda d) < d\sigma^2$ and by assumption R^2/h goes to infinity. We conclude similarly as the previous cases.

Case $d = o(h)$ **:** To prove the convergence of the weights to $1/N$, for all similarity functions, we will show that for all pairs of points (X_i, X_j), the scaled similarity between them $s(X_i, X_j)/h$ goes to 0.

Proof for $s = s_o$ *:* According to (VI.14), with probability greater than $1 - Ne^{-t}$, for all $1 \leq i, j \leq N$:

$$\left|\frac{\langle X_i, X_j\rangle}{h}\right| \leq \frac{\|X_i\|\|X_j\|}{h} \leq \frac{d\sigma^2 + R^2 + \delta(t) + C\sigma^2 t}{h} \underset{d\to\infty}{\longrightarrow} 0\,,$$

where $\delta(t) := \sigma\sqrt{(R^2 + \sigma^2 d)t}$. Then, for all $1 \leq i, j \leq N$:

$$\omega_{ij} \leq \frac{1}{N}\exp\left(2\frac{d\sigma^2 + R^2 + \delta(t) + C\sigma^2 t}{h}\right),$$

and then:

$$\omega_{ij} \leq \frac{1}{N} + \frac{1}{N}\left(\exp\left(2\frac{d\sigma^2 + R^2 + \delta(t) + C\sigma^2 t}{h}\right) - 1\right).$$

As the vector $(\omega_{ij})_j$ belongs to the simplex $\mathcal{S}_N$, we can deduce from the upper bound above on the weights the following lower bound. With probability greater than $1 - Ne^{-t}$, for all $1 \leq i, j \leq N$:

$$\begin{aligned}\omega_{ij} = 1 - \sum_{\ell\neq j}\omega_{i\ell} &\geq \left(1 - \frac{N-1}{N}\right) - \frac{N-1}{N}\left(\exp\left(2\frac{d\sigma^2 + R^2 + \delta(t) + C\sigma^2 t}{h}\right) - 1\right)\\ &\geq \frac{1}{N} - \left(\exp\left(2\frac{d\sigma^2 + R^2 + \delta(t) + C\sigma^2 t}{h}\right) - 1\right).\end{aligned}$$

By inverting the bound, the get that

$$\mathbb{P}\left[\max_{ij}\left|\omega_{ij} - \frac{1}{N}\right| > \varepsilon\right] \leq Ne^{-t_\varepsilon}$$

where we can show that $t_\varepsilon = \Omega(h)$. We conclude again by Borel-Cantelli lemma.

Proof for $s = s_{L^2}$ *:* using again (VI.16), with probability at least $1 - N^2e^{-t}$:

$$\|X_i - X_j\|^2 \geq 2d\sigma^2 - C\sigma\sqrt{(R^2 + \sigma^2 d)t}\,, \quad \text{for} \quad 1 \leq i \neq j \leq N\,.$$

We get, with probability at least $1 - N^2e^{-t}$, that for all $1 \leq i, j \leq N$:

$$\omega_{ij} \leq \frac{1}{N}\exp\left(\frac{2d\sigma^2 - C\sigma\sqrt{(R^2 + \sigma^2 d)t}}{h}\right).$$

Similarly as for the similarity function s_o, using that the weights are in the simplex, we get a high probability lower bound on them. With probability at least $1 - N^2e^{-t}$, that for all $1 \leq i, j \leq N$:

$$\omega_{ij} = 1 - \sum_{\ell\neq j}\omega_{i\ell} \geq \frac{1}{N} - \left(\exp\left(\frac{2d\sigma^2 - C\sigma\sqrt{(R^2 + \sigma^2 d)t}}{h}\right) - 1\right).$$

We deduce from this last concentration inequality that almost sure convergence of the weights to $1/N$.

Proof for $s = s_p$: This case is the simplest as the weights can be bounded without concentration inequality. Indeed, for all $1 \leq i, j \leq N$:

$$\left| \frac{R^2}{h} s_p(X_i, X_j) \right| \leq \frac{R^2}{h} \underset{d \to \infty}{\longrightarrow} 0 \,,$$

as $R^2 = \Theta(d)$ and $d = o(h)$. Then for all $1 \leq i, j \leq N$, $\omega_{ij} \underset{d \to \infty}{\longrightarrow} N^{-1}$. □

VI.5.2 Proofs of Proposition VI.5, Proposition VI.6 and Proposition VI.8

Let us first make a general assumption on the similarity functions.

Assumption VI.9. *Let $X \sim \mathcal{N}(\mu, \sigma^2 I_d)$ and $Y \sim \mathcal{N}(\nu, \sigma^2 I_d)$ be two independent Gaussian vectors, then for all $t \geq 0$:*

$$\mathbb{P}\Big[\Big| s(X, Y) - s(X, X) + \|\mu - \nu\|^2/2 \Big| \geq \gamma(t) \Big] \leq e^{-t} \tag{VI.17}$$

for some non-negative function $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$.

In particular, we will show in Lemma VI.11 that s_o, s_{L^2} and s_p satisfy this assumption. This assumption allows us to state a general result.

Lemma VI.10. *Let s be a similarity function satisfying Assumption VI.9. Assume Assumption VI.2 satisfied for a constant $C_1 > 0$, then for all $t \geq 0$, with probability greater than $1 - e^{-t}$:*

$$\|a(X_1) - \mu_1\|^2 \leq e^{4\gamma(t_N)/h} \left(\frac{d\sigma^2}{1 + V_1} + 2h \log\left(\frac{N}{1 + V_1} \right) \right) + C\sigma^2 \left(\sqrt{dt_N} + t_N \right) \tag{VI.18}$$

here $V_1 = \sum_{i=2}^N e^{-\|\Delta_i\|^2/(2h)}$, $t_N = t + 2(1 + \log N)$ and $C > 0$ an absolute constant.

Proof. Using Lemma 8.1 of Birgé (2001), Proposition III.7 and Assumption VI.9, after an union bound, with probability greater than $1 - 3N^2 e^{-t}$:

$$\forall i, \quad \|X_i - \mu_1\|^2 \leq d\sigma^2 + \|\Delta_i\|^2 + \delta(t) \,, \tag{VI.19}$$

$$\forall i, j, \quad \langle X_i - \mu_1, X_j - \mu_1 \rangle \leq \langle \Delta_i, \Delta_j \rangle + \delta(t) \,, \tag{VI.20}$$

$$\forall i, \quad \left| s(X_i, X_1) - s(X_1, X_1) + \frac{\|\Delta_i\|^2}{2} \right| \leq \gamma(t) \,. \tag{VI.21}$$

where $\Delta_i := \mu_1 - \mu_i$ and $\delta(t) = C\sigma\sqrt{(R^2 + \sigma^2 d)t} + \sigma^2 t \leq C\sigma^2 \left(\sqrt{dt} + t \right)$ for some absolute constant C (independent of d and N). Assuming these events are satisfied, we can now bound the quadratic distance between $a(X_1)$ and μ_1:

$$\begin{aligned}
\|a(X_1) - \mu_1\|^2 &= \sum_{i=1}^N \omega_i^2 \|X_i - \mu_1\|^2 + \sum_{i \neq j} \omega_i \omega_j \langle X_i - \mu_1, X_j - \mu_1 \rangle \\
&\leq \sum_{i=1}^N \omega_i^2 (d\sigma^2 + \|\Delta_i\|^2) + \sum_{i \neq j} \omega_i \omega_j \langle \Delta_i, \Delta_j \rangle + \delta(t) \\
&= d\sigma^2 \sum_{i=1}^N \omega_i^2 + \left\| \sum_{i=1}^N \omega_i \Delta_i \right\|^2 + \delta(t) \leq d\sigma^2 \sum_{i=1}^N \omega_i^2 + \sum_{i=1}^N \omega_i \|\Delta_i\|^2 + \delta(t) \,,
\end{aligned} \tag{VI.22}$$

where $\omega_i = \frac{\exp(s(X_i,X_1)/h)}{\sum_{j=1}^N \exp(s(X_j,X_1)/h)}$. We have used the bounds (VI.19) and (VI.20) in the first inequality and the convexity of the squared norm for the last. Using (VI.21), we can bound the weights. For $i \neq 1$:

$$\begin{aligned}
\omega_i &= \exp\Big((s(X_i,X_1) - s(X_1,X_1))/h\Big)\Big(1 + \sum_{j=2}^N \exp\Big((s(X_i,X_1) - s(X_1,X_1)/h)\Big)\Big)^{-1} \\
&\leq \exp\Big(- \|\Delta_i\|^2/(2h) + \gamma(t)/h\Big)\Big(1 + \sum_{j=2}^N \exp\Big(- \|\Delta_j\|^2/(2h) - \gamma(t)/h\Big)\Big)^{-1} \\
&\leq e^{2\gamma(t)/h} \frac{e^{-\|\Delta_i\|^2/(2h)}}{\sum_{j=1}^N e^{-\|\Delta_j\|^2/(2h)}} \,.
\end{aligned}$$

We have used for the last inequality that $1 \geq e^{-\|\Delta_1\|^2/(2h)} e^{-\gamma(t)/h}$. Similarly for the weight ω_1 :

$$\begin{aligned}
\omega_1 &= \left(1 + \sum_{j=2}^N \exp((s(X_i,X_1) - s(X_1,X_1)/h))\right)^{-1} \\
&\leq \left(1 + \sum_{j=2}^N \exp\Big(-\|\Delta_j\|^2/(2h) - \gamma(t)/h\Big)\right)^{-1} \leq e^{\gamma(t)/h} \frac{e^{-\|\Delta_1\|^2/(2h)}}{\sum_{j=1}^N e^{-\|\Delta_j\|^2/(2h)}} \,.
\end{aligned}$$

We recall that $V_i := \sum_{j\neq i} e^{-\|\Delta_j\|^2/(2h)}$. Injecting the bounds of ω into (VI.22), we get:

$$\begin{aligned}
\|a(X_1) - \mu_1\|^2 &\leq d\sigma^2 e^{4\gamma(t)/h} \sum_{i=1}^N \frac{e^{-\|\Delta_i\|^2/h}}{(1+V_1)^2} + e^{2\gamma(t)/h} \frac{\sum_{i=2}^N e^{-\|\Delta_i\|^2/(2h)} \|\Delta_i\|^2}{1+V_1} + \delta(t) \\
&\leq e^{4\gamma(t)/h} \frac{d\sigma^2}{1+V_1} + e^{2\gamma(t)/h} \frac{N}{1+V_1} \frac{2h}{N} \sum_{i=2}^N f(u_i) + \delta(t) \,, \qquad \text{(VI.23)}
\end{aligned}$$

where $f : x \mapsto x\log(x^{-1})$ and $u_i := e^{-\|\Delta_i\|^2/(2h)}$ for the second term. For the first term, we have bounded $e^{-\|\Delta_i\|^2/h}$ by $e^{-\|\Delta_i\|^2/(2h)}$. Using the concavity of f and that $u_1 = 1$ and $f(u_1) = 0$, we can bound the second term of (VI.23):

$$\frac{1}{N}\sum_{i=2}^N f(u_i) = \frac{1}{N}\sum_{i=1}^N f(u_i) \leq f\left(\frac{1}{N}\sum_{i=1}^N u_i\right) = f\left(\frac{1+V_1}{N}\right) = \frac{1+V_1}{N}\log\left(\frac{N}{1+V_1}\right).$$

Combining this bound with (VI.23), we get that with probability at least $1 - e^{-t}$:

$$\|a(X_1) - \mu_1\|^2 \leq e^{4\gamma(\widetilde{t}_N)/h}\left(\frac{d\sigma^2}{1+V_1} + 2h\log\left(\frac{N}{1+V_1}\right)\right) + C\sigma^2\left(\sqrt{d\widetilde{t}_N} + \widetilde{t}_N\right),$$

where $\widetilde{t}_N = t + \log(3N^2) \leq t + 2(1+\log N) = t_N$. □

Proof of Proposition VI.5 Let us first compute $\mathbb{E}\big[\delta^2(X_1,\mathcal{M})\big]$. By invariance by rotation of a Gaussian noise, we can assume that for all $1 \leq i \leq N$, the points X_i can be rewritten as:

$$X_i = X_i^{\mathcal{M}} + \xi_i \,,$$

where $X_i^{\mathcal{M}}$ is Gaussian vector in $\mathcal{M}$ and ξ_i is a Gaussian vector in the orthogonal of $\mathcal{M}$ of distribution $\mathcal{N}(0,\sigma^2 I_{d-m})$ (when restriced to the orthogonal of $\mathcal{M}$). Then for $1 \leq i \leq N$:

$$\mathbb{E}\big[\delta^2(X_i,\mathcal{M})\big] = \mathbb{E}\Big[\|\xi_i\|^2\Big] = \sigma^2(d-m) \,.$$

Let us now consider $\mathbb{E}\big[\delta^2(a_s(X_1),\mathcal{M})\big]$. With the same notation, the distance of $a_s(X_1)$ to the subspace $\mathcal{M}$ is the norm of its orthogonal part, which is:

$$\delta^2(a_s(X_1),\mathcal{M}) = \left\|\sum_{j=1}^N \omega_j \xi_j\right\|^2 ,$$

where $\omega_j := \omega_{1j}^s$. Similarly as in the proof of Lemma VI.10, from Lemma 8.1 of Birgé (2001), Proposition III.7 and Lemma VI.11, after an union bound, with probability greater than $1-3N^2e^{-t}$:

$$\forall i, \quad \|\xi_i\|^2 \le (d-m)\sigma^2 + \delta(t)\,, \tag{VI.24}$$

$$\forall i,j, \quad \langle \xi_i, \xi_j\rangle \le \delta(t)\,, \tag{VI.25}$$

$$\forall i, \quad \left| s(X_i,X_1) - s(X_1,X_1) + \frac{\|\Delta_i\|^2}{2}\right| \le \gamma(t)\,. \tag{VI.26}$$

where $\Delta_i := \mu_1 - \mu_i$ and $\delta(t) = C\sigma^2\sqrt{(d-m)t} + \sigma^2 t$ for some absolute constant $C>0$ and for $\gamma(t)$ defined in Lemma VI.11. Then, assuming these events satisfied, we get:

$$\left\|\sum_{j=1}^N \omega_j\xi_j\right\|^2 = \sum_{i=1}^N \omega_i^2\|\xi_i\|^2 + \sum_{i\neq j}\omega_i\omega_j\langle\xi_i,\xi_j\rangle \le (d-m)\sigma^2\sum_{i=1}^N\omega_i^2 + \delta(t) \tag{VI.27}$$

Using the inequality of proof of Lemma VI.10 combining with $\|\Delta_i\|^2 \le 4R^2 \le 4C_1 d\sigma^2 \le 4h$, we get that for $i\neq 1$:

$$\omega_i \le e^{2\gamma(t)/h}\frac{e^{-\|\Delta_i\|^2/(2h)}}{\sum_{j=1}^N e^{-\|\Delta_j\|^2/(2h)}} \le e^{2\gamma(t)/h}\frac{e^2}{N}\,.$$

Similarly for the weight ω_1 :

$$\omega_1 \le e^{\gamma(t)/h}\frac{e^{-\|\Delta_1\|^2/(2h)}}{\sum_{j=1}^N e^{-\|\Delta_j\|^2/(2h)}} \le e^{2\gamma(t)/h}\frac{e^2}{N}\,.$$

After injecting into (VI.27), we obtain that there exists an exponential random variable $\xi\sim\mathcal{E}(1)$ such that almost surely:

$$\delta^2(a_s(X_1),\mathcal{M}) \le C\frac{(d-m)\sigma^2}{N}e^{4\gamma(\xi+\log 3N^2)/h} + C\sigma^2\sqrt{(d-m)}(\xi+1+\log N)\,. \tag{VI.28}$$

According to Lemma VI.11 for s be either s_o or s_{L^2}, as $R^2\le C_1 d\sigma^2$ we have:

$$\begin{aligned}\frac{\gamma(\xi+\log 3N^2)}{h} &\le \frac{d\sigma^2}{h} + \frac{C\max(1,\sqrt{C_1})\sigma^2\sqrt{d(\xi+1+\log N)} + C\sigma^2(\xi+1+\log N)}{h}\\ &\le CC_1^{-1}\left(1+\frac{\max(1,\sqrt{C_1})}{\sqrt{d}}\sqrt{\xi}+\frac{\xi}{d}+\sqrt{\frac{1+\log N}{d}}+\frac{1+\log N}{d}\right) \qquad \text{(VI.29)}\\ &\le CC_1^{-1}\left(1+\frac{\max(1,\sqrt{C_1})}{\sqrt{d}}\sqrt{\xi}+\frac{\xi}{d}\right) \qquad \text{(VI.30)}\end{aligned}$$

using that $h\ge C_1 d\sigma^2$ for (VI.29) and that $d\ge \log N$ for (VI.30). Then for C_1 large enough ($C_1\ge C$), the expectation $\mathbb{E}\left[e^{4\gamma(\xi+\log 3N^2)/h}\right]$ is finite and upper bounded for all d. Then taking the expectation of (VI.28) we conclude the proof. □

Proof of Proposition VI.6 In this proof the constant C can differ between equations and can depend of C_1. Let us partition the set $[\![N]\!]$ into d subsets $\mathcal{C}_1, \dots, \mathcal{C}_d$ of size $\lceil N/d \rceil$ or $\lfloor N/d \rfloor$. For $i \in \mathcal{C}_k$, let $\mu_i := Re_k$ where $(e_1, \dots, e_d)$ is an orthonormal basis of $\mathbb{R}^d$. Then for $\Delta_i := \mu_i - \mu_1$, we have:

$$\langle \Delta_i, \Delta_j \rangle = \begin{cases} 0 & \text{if } i \in \mathcal{C}_1 \text{ or } j \in \mathcal{C}_1, \\ 2R^2 & \text{if } i,j \in \mathcal{C}_k \text{ for } k \neq 1, \\ R^2 & \text{otherwise.} \end{cases}$$

Then, similarly as the previous proofs, with probability greater than $1 - 3N^2 e^{-t}$:

$$\forall i, \quad \|X_i - \mu_1\|^2 \geq d\sigma^2 + \|\Delta_i\|^2 - \delta(t)\,, \tag{VI.31}$$

$$\forall i,j, \quad \langle X_i - \mu_1, X_j - \mu_1 \rangle \geq \langle \Delta_i, \Delta_j \rangle - \delta(t)\,, \tag{VI.32}$$

$$\forall i, \quad \left| s(X_i, X_1) - s(X_1, X_1) + \frac{\|\Delta_i\|^2}{2} \right| \leq \gamma(t)\,. \tag{VI.33}$$

where $\delta(t) = C\sigma\sqrt{(R^2 + \sigma^2 d)t} + \sigma^2 t \leq C\sigma^2\Big(\sqrt{dt} + t\Big)$ for some constant C depending of C_1 and for $\gamma(t)$ defined in Lemma VI.11. Assuming these events are satisfied and that $1 \in \mathcal{C}_1$, we have

$$\begin{aligned} \|a_s(X_1) - \mu_1\|^2 &= \sum_{i=1}^N \omega_i^2 \|X_i - \mu_1\|^2 + \sum_{i \neq j} \omega_i \omega_j \langle X_i - \mu_1, X_j - \mu_1 \rangle \\ &\geq \sum_{i=1}^N \omega_i^2 (d\sigma^2 + \|\Delta_i\|^2) + \sum_{i \neq j} \omega_i \omega_j \langle \Delta_i, \Delta_j \rangle - \delta(t) \\ &\geq \sum_{i \notin \mathcal{C}_1} \omega_i^2 (0 + 2R^2) + \sum_{i \neq j \notin \mathcal{C}_1} \omega_i \omega_j R^2 - \delta(t) = R^2 \left(\sum_{i \notin \mathcal{C}_1} \omega_i \right)^2 - \delta(t)\,, \end{aligned} \tag{VI.34}$$

where $\omega_i = \omega_{i1}^s$. The first inequality is obtained using bounds (VI.31) and (VI.32) and the second by replacing $\langle \Delta_i, \Delta_j \rangle$ by its value. Using now (VI.33), we get for $i \notin \mathcal{C}_1$:

$$\omega_i \geq e^{-2\gamma(t)/h} \frac{e^{-\|\Delta_i\|^2/(2h)}}{\sum_{j=1}^N e^{-\|\Delta_j\|^2/(2h)}} \geq e^{-2\gamma(t)/h} \frac{e^{-2}}{N}\,.$$

We have used that $\|\Delta_i\|^2 \leq 4R^2 \leq 4h$ by assumption. By combining the two previous bounds and replacing t by $t_N = t + \log(3N^2)$ into (VI.34), with probability at least $1 - e^{-t}$ we get:

$$\begin{aligned} \|a_s(X_1) - \mu_1\|^2 &\geq CR^2 e^{-4\gamma(t_N)/h} \left((N - |\mathcal{C}_1|) \frac{C}{N} \right)^2 - \delta(t_N) \\ &\geq CR^2 e^{-4\gamma(t_N)/h} \left(1 - \frac{1}{N} - \frac{1}{d} \right)_+^2 - \delta(t_N)\,, \end{aligned}$$

where we have used that $|\mathcal{C}_1| \leq \frac{N}{d} + 1$ and where $(\cdot)_+$ denotes the positive part. Using Eq.(VI.30), we get that for C_1 large enough:

$$\frac{\gamma(t + \log 3N^2)}{h} \leq C\Big(1 + \sqrt{t} + t\Big)\,.$$

Then, for $\xi \sim \mathcal{E}(1)$ an exponential random variable, we have:

$$\begin{aligned} \mathbb{E}\Big[\|a_s(X_1) - \mu_1\|^2\Big] &\geq CR^2 \mathbb{E}\Big[e^{-C(1+\sqrt{\xi}+\xi)}\Big] \left(1 - \frac{1}{N} - \frac{1}{d} \right)_+^2 - C\sigma^2 \mathbb{E}\Big[\sqrt{d(\xi + \log 3N^2)} + \xi + \log 3N^2\Big] \\ &\geq CC_1^{-1} d\sigma^2 \left(1 - \frac{1}{N} - \frac{1}{d} \right)_+^2 - C\sigma^2 \sqrt{d(1 + \log N)}\,. \end{aligned}$$

We have used that $\log N \leq \sqrt{d \log N}$.

□

Proof of Proposition VI.8 According to Lemma VI.11, the similarity functions s_m and s_{db} verify Assumption VI.9 for γ satisfying for $t \geq 0$:

$$\begin{aligned}\gamma(t + C(1+\log N)) &= C\max(1,\sqrt{C_1})\sigma^2\sqrt{d(t+C(1+\log N))} + C\sigma^2(t+C(1+\log N)) \\ &\leq \gamma(t) + C\max(1,\sqrt{C_1})\sqrt{d(1+\log N)}\,.\end{aligned}$$

as $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $d \geq \log N$. Moreover, using that $h \geq C_1\sigma^2\sqrt{d}$:

$$\frac{\gamma(t)}{h} \leq CC_1^{-1}\left(\max(1,\sqrt{C_1})\sqrt{t} + \frac{t}{\sqrt{d}}\right),$$

and then, for C_1 large enough ($C_1 \geq C/\sqrt{d}$), $\mathbb{E}\left[e^{4\gamma(\xi)/h}\right]$ is upper bounded for $\xi \sim \mathcal{E}(1)$ an exponential random variable. From here, in this proof, the notation C denotes a constant depending potentially of C_1. By combining these bounds with Lemma VI.10, we get that for $1 \leq i \leq N$:

$$\begin{aligned}\mathbb{E}\left[\|a(X_i) - \mu_i\|^2\right] &\leq \mathbb{E}\left[e^{4\gamma(\xi)/h}\right] e^{C\sigma^2\sqrt{d\log N}/h}\left(\frac{d\sigma^2}{1+V_i} + 2h\log\left(\frac{N}{1+V_i}\right)\right) + C\sigma^2\sqrt{d\log N} \\ &\leq C\left(\frac{d\sigma^2}{1+V_i} + 2h\log\left(\frac{N}{1+V_i}\right)\right) + C\sigma^2\sqrt{d\log N}\,.\end{aligned}$$

for $V_i = \sum_{\substack{j=1\\ j\neq i}}^{N} e^{-\|\Delta_{ij}\|^2/(2h)}$. We have bounded again $\log N$ by $\sqrt{d\log N}$. As $V_i \geq 0$ and $h > \sigma^2\sqrt{d}\log N$, we get the individual bound:

$$\mathbb{E}\left[\|a(X_i) - \mu_i\|^2\right] \leq C\left(\frac{d\sigma^2}{1+V_i} + h\log N\right).$$

The $\mathcal{N}_h$ balls induce a partition $\mathcal{C}_1 \sqcup \ldots \sqcup \mathcal{C}_{\mathcal{N}_h} = [\![N]\!]$ of the means such that for $i,j \in \mathcal{C}_k$, $\|\Delta_{ij}\| \leq 2\sqrt{h}$. Then for $i \in \mathcal{C}_k$, $V_i \geq (|\mathcal{C}_k| - 1)e^{-2}$. Hence

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\left(\frac{d\sigma^2}{1+V_i} + 2h\log\left(\frac{N}{1+V_i}\right)\right) &\leq \frac{1}{N}\sum_{k=1}^{\mathcal{N}_h}|\mathcal{C}_k|\left[\frac{d\sigma^2}{|\mathcal{C}_k|}e^2 + 2h\log\left(\frac{N}{1+|\mathcal{C}_k|e^{-2}}\right)\right] \\ &\leq C\left[d\sigma^2\frac{\mathcal{N}_h}{N} + h + h\sum_{k=1}^{\mathcal{N}_h}\frac{|\mathcal{C}_k|}{N}\log\left(\frac{N}{|\mathcal{C}_k|}\right)\right] \\ &\leq C\left[d\sigma^2\frac{\mathcal{N}_h}{N} + h + h\log\mathcal{N}_h\right], \qquad \text{(VI.35)}\end{aligned}$$

using the concavity of the function $x \mapsto x\log x^{-1}$ for the last inequality. □

Lemma VI.11. *Assume $R^2/d\sigma^2 \leq C_R$, then the similarity functions s_o, s_{L^2} satisfy Assumption VI.9 for $\gamma(t) = d\sigma^2 + \delta(t)$ and the similarity functions s_m and s_{db} satisfy it for $\gamma(t) = \delta(t)$ where for $t \geq 0$:*

$$\delta(t) = C\max(1, C_R)\sigma^2\sqrt{d(t+1)} + C\sigma^2(t+1)$$

where C is an absolute constant.

Proof. Let us recall that for all $x, y \in \mathbb{R}^d$:

$$\begin{gathered}s_o(x,y) = \langle x, y\rangle\,, \quad s_{L^2}(x,y) = -\frac{\|x-y\|^2}{2}\,; \\ s_m(x,y) = \langle x, y\rangle \quad \text{for } x \neq y, \text{ and} \quad s_m(x,x) = R^2\ : \\ s_{db}(x,y) = \langle x, y\rangle \quad \text{for } x \neq y, \text{ and} \quad s_{db}(x,x) = \|x\|^2 - d\sigma^2\,.\end{gathered}$$

Using the concentration bounds on squared norm and scalar product of Gaussian random vectors (Birgé, 2001 and Proposition III.7) we get the following deviations. These concentration bonds are obtained by direct applications of these results or by combining them using union bounds. Let $t \geq 0$, each inequality is true with probability at least $1 - e^{-t}$:

$$\left| s_o(X,Y) - s_o(X,X) + \frac{\|\mu-\nu\|^2}{2} \right| = \left| \langle X, Y\rangle - \|X\|^2 + R^2 - \langle \mu, \nu\rangle \right| \leq d\sigma^2 + \delta(t)\,;$$

$$\left| s_{L^2}(X,Y) - s_{L^2}(X,X) + \frac{\|\mu-\nu\|^2}{2} \right| = \left| -\frac{\|X-Y\|^2}{2} + \frac{\|\mu-\nu\|^2}{2} \right| \leq d\sigma^2 + \delta(t)\,;$$

$$\left| s_m(X,Y) - s_m(X,X) + \frac{\|\mu-\nu\|^2}{2} \right| = \left| \langle X, Y\rangle - \langle \mu, \nu\rangle \right| \leq \delta(t)\,;$$

$$\left| s_{db}(X,Y) - s_{db}(X,X) + \frac{\|\mu-\nu\|^2}{2} \right| = \left| \langle X, Y\rangle - \|X\|^2 + d\sigma^2 + R^2 - \langle \mu, \nu\rangle \right| \leq \delta(t)\,;$$

where $\delta(t) = C\sigma\Big(\sqrt{(R^2 + d\sigma^2)(t+1)} + \sigma(t+1)\Big)$ for some absolute constant C. The deviation δ is a polynomial of $t+1$ and not of t due to the use of union bounds. We conclude by upper bounding R^2 by $C_R d\sigma^2$.

□

VI.5.3 Estimation on the sphere

The following lemma lower bounds the distance of a projected point to its mean. We observe that the risk of this estimator is of same order of the distance without projection.

Lemma VI.12. *Let $d \geq 2$, $\mu \in \mathbb{R}^d$, such that $\|\mu\| = R$ and $X \sim \mathcal{N}(\mu, \sigma^2 I_d)$. Assume that $R^2 \geq \sigma^2 d$, then:*

$$\mathbb{E}\left[\left\| \frac{RX}{\|X\|} - \mu \right\|^2 \right] \geq C(d-1)\sigma^2, \tag{VI.36}$$

for some constant $C \geq 0.1$.

Proof. Let us first remark that:

$$\left\| \frac{RX}{\|X\|} - \mu \right\|^2 = 2R^2 - 2R\left\langle \frac{X}{\|X\|}, \mu \right\rangle.$$

Let us denote $\varepsilon := X - \mu \sim \mathcal{N}(0, \sigma^2 I_d)$. It exists a random variable Z^2 such that $Z^2/\sigma^2 \sim \chi^2(d-1)$, is independent of $\langle \varepsilon, \mu\rangle$ and $\|\varepsilon\|^2 = Z^2 + \langle \varepsilon, \mu\rangle^2 / R^2$. Then:

$$\left\langle \frac{X}{\|X\|}, \mu \right\rangle = \frac{R^2 + \langle \varepsilon, \mu\rangle}{\sqrt{Z^2 + \frac{\langle \varepsilon, \mu\rangle^2}{R^2} + 2\langle \varepsilon, \mu\rangle + R^2}}\,.$$

As the function $f : x \mapsto \frac{R^2 + x}{\sqrt{Z^2 + x^2/R^2 + 2x + R^2}}$ is concave, then by Jensen's inequality conditionally to Z^2:

$$\mathbb{E}\left[\left\langle \frac{X}{\|X\|}, \mu \right\rangle \right] = \mathbb{E}\big[\mathbb{E}\big[f(\langle \varepsilon, \mu\rangle) | Z^2\big]\big] \leq \mathbb{E}\big[f(\mathbb{E}\big[\langle \varepsilon, \mu\rangle | Z^2\big])\big] = \mathbb{E}[f(\mathbb{E}[\langle \varepsilon, \mu\rangle])] = \mathbb{E}[f(0)] = \mathbb{E}\left[\frac{R^2}{\sqrt{R^2 + Z^2}} \right].$$

Using that for all $z > 0$:

$$1 - \frac{1}{\sqrt{1+z}} \geq \frac{z}{2(1+z)},$$

and that it exists a random variable $\xi \sim \mathcal{E}(1)$ such that almost surely $Z^2 \geq (d-1)\sigma^2 e^{-(1+\xi/(d-1))}$ (Lemma V.41), we get:

$$\begin{aligned}
\mathbb{E}\left[2R^2\left(1 - \frac{1}{\sqrt{1+Z^2/R^2}}\right)\right] \geq \mathbb{E}\left[\frac{Z^2}{1+Z^2/R^2}\right] &\geq \sigma^2(d-1)\mathbb{E}\left[\frac{e^{-(1+\xi/(d-1))}}{1+(d-1)\sigma^2/R^2 e^{-(1+\xi/(d-1))}}\right] \\
&\geq \sigma^2(d-1)\mathbb{E}\left[\frac{e^{-(1+\xi/(d-1))}}{1+e^{-(1+\xi/(d-1))}}\right] \\
&\geq \sigma^2(d-1)\mathbb{E}\left[\frac{e^{-(1+\xi/(d-1))}}{1+e^{-1}}\right]
\end{aligned}$$

We have used that by assumption $R^2 \geq \sigma^2 d \geq \sigma^2(d-1)$. We conclude by integrating.

$$\mathbb{E}\left[\frac{e^{-(1+\xi/(d-1))}}{1+e^{-1}}\right] = \frac{1}{1+(d-1)^{-1}}\frac{1}{e+1} \geq \frac{1}{2}\frac{1}{e+1} \geq 0.1$$

□

VII Conclusion and future directions

Taking as starting point the classical analysis of the influence of dimension for Gaussian isotropic data, we have studied closeness testing, multiple means estimation and the attention mechanism on a high dimensional framework and proposed new methods adapted to this setting.

Closeness testing: In Section IV, we recovered the phase transition of the separation rate of the test already known for isotropic distributions. This separation rate evolves with the targeted proximity: for a small distance, the test is harder than for a large one with an error ranging from $\Theta(\sqrt{d^*})$ to $\Theta(1)$. The upper bound on the separation rate is obtained by building tests based on estimators of covariance moments for which we provide non-asymptotic bounds. Our test are build for Gaussian or bounded data which permits to evaluate the separation rate of a two sample test in term of MMD distance. These mathematical results were also a first step to be able to consider the multi task averaging problem for heterogeneous distributions.

Multiple means estimation: The non-asymptotic quantification of the (effective) dimensional dependencies of test errors have enabled us to build better estimators than empirical means for multiple means estimation. We adapted our first method (Section III), fit to homogeneous distributions, to highly heterogeneous distributions with guarantees outside the framework of finite-dimensional isotropic Gaussian data. The average improvement depends of the unknown structure of the set of means, e.g. the possible improvement is high if the set is covered by a small number of balls. This estimation improvement is obtained by aggregating the empirical mean of each sample with the empirical means of samples with closed mean and relatively smaller test error. This last constraint is unnecessary for homogeneous distributions but is crucial to consider heterogeneous distributions as in Section V. In this setting, we had to take into account both that the coordinates are no longer independent and that the different distributions can have different covariance structure and then different effective dimensions. These methods select these samples explicitly by testing or implicitly by minimizing a penalized empirical risk. The simplicity of the multitask averaging problem relatively to a general multitask problem allowed us to obtain a sharp control of the improvement. This noise reduction is limited at $(d^\bullet)^{-1/2} = \sqrt{d^*}/d^{\mathrm{e}}$ which is roughly the ratio of the estimation error and the test error. Moreover, we show that our Q-aggregation method does not deteriorate the estimation even in a worst case where the means are far from each others as our upper bounds on the improvement is upper bounded by 1 when the dimension goes to infinity ($\mathcal{B}(\tau, \nu) \leq 1$). We adapted all our methods to bounded distributions to be able to apply them to the estimation of multiple kernel mean embeddings.

Self-attention: These considerations about denoising empirical means have led us to study the self attention mechanism as a problem of multiple mean estimation. With this new point of view, we have exhibited a noise reduction effect of this operation in high dimension and have proposed some modifications which make it more robust in our simplified setting of Gaussian distributions. This analysis has been made by reducing the attention matrices to an unique real parameter h for which we study the dependence in the dimension. For classical attention mechanisms the scale h needs to be of order $\Theta(d)$, but with this scale in some worst cases, the attention can deteriorate the observations. Our modifications of the attention are more robust as they accept a wider range of scale h while supporting these worst cases.

These different works leave us with many unanswered questions and potential future developments.

- **Improvement of the minimax bound.** For the separation rate of the closeness testing and for the minimax risk of multiple means estimation, a gap remains between upper and lower bounds for bounded data. Indeed, for now, our upper bounds in the bounded setting are compared with Gaussian lower bounds and only match with some assumption on the maximum norm of the data or on the sample size (ϕ needs to be controlled for instance in Proposition V.16). We think that

these conditions are not caused by our methods but are intrinsic of the problem and that a specific minimax analysis of the bounded setting could justify them. The gap between the lower and upper bounds is also present in Gaussian setting for the problem of multiple means estimation. This gap is located in two quantities, the additive term $\log B/\sqrt{d^\bullet}$ and the set of neighbors $W_{(\varsigma)}$ appearing in the upper bound and not in the lower bounds (comparing Corollary V.13 and Proposition V.18). We interpret $\log B/\sqrt{d^\bullet}$ as the error of the detection of the closed means and $W_{(\varsigma)}$ as the error induced by the covariance heterogeneity of the distributions. The term $\log B/\sqrt{d^\bullet}$ appears indeed in the homogeneous setting (see Proposition III.3) contrary to $W_{(\varsigma)}$ only present in the methods of Section V. In fact, in our actual lower bounds, the closed means are assumed known and the distributions are supposed to share a common covariance structure (see Definition V.17). We think that taking them into account in the minimax analysis could fill this gap.

- **Generalization of the multiple means estimation.** One other open question is the extension of our approach to problems more general than mean vectors estimation. In particular, we are interested to consider problems of regression and understand how to adapt our approaches such that they could apply for such problems with high dimensional data. One possibility would be to adapt our methods in order to apply them to the embeddings of the data and hope to denoise them before the regression. In case of a distribution regression problem, the embeddings could be the kernel mean embedding of the distributions which fits with our bounded setting. In this case and more generally when the outputs are high dimensional vectors, an other possibility could be to combine directly the regressors of each bags, which comes close to an expert aggregation problem.

- **Extension of the denoising analysis of the self attention:** Our new interpretation of this central mechanism of the Transformers have been justified in this thesis in an artificial Gaussian model with simplified forms of attention, but leaves us with many unanswered questions. It would be interesting to understand whether this phenomenon extends to a general form of attention, if in this case an other data structure is captured by the attention mechanism, and whether a notion of effective dimension is relevant when the noise is not isotropic. For such a non-isotropic noise, we expect that the attention matrices, used in the general form of attention, will learn its covariance and use it to denoise the sample. Some recent results in others simplified form of attention have already point to this direction (Zhang et al., 2024). The linked question is to know if our attention modifications can be exported to this more general setting. For a general couple of matrices (Q, K), the debiased similarity function can be adapted by substracting the trace of the matrix but an adaptation is less evident for the "modified" similarity function. Another question is to include in our analysis the multi-head attention used in practice and which apply the attention mechanism to subgroups of the vectors coordinates. One possibility is that this mechanism may capture the block structure of the means or of the covariance. Furthermore, the modifications of attention that have been proposed require to be tested in a neural network and with real data to observe if the wider range of scale can really improve the network on performance or maybe on training time.

References

Adan, A., Alizada, G., Kiraz, Y., Baran, Y., and Nalbant, A. (2017). Flow cytometry: basic principles and applications. *Critical reviews in biotechnology* 37 (2), 163–176.

Aizerman, A. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control* 25, 821–837.

Anderson, T. (2003). *An introduction to multivariate statistical analysis*. Third edition. Wiley series in probability and mathematical statistics. Wiley.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society* 68 (3), 337–404.

Assouad, P. (1979). Etude d'une dimension metrique liee a la possibilite de plongements dans R^ n. *CR Acad. Sci. Paris Sér. AB* 288, A731–A734.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473.

Baker, C. R. (1973). Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society* 186, 273–289.

Balasubramanian, K., Li, T., and Yuan, M. (2021). On the optimality of kernel-embedding based goodness-of-fit tests. *Journal of Machine Learning Research* 22 (1), 1–45.

Baranchik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *The Annals of Mathematical Statistics* 41 (2), 642–645.

Baranchik, A. J. (1964). *Multiple regression and estimation of the mean of a multivariate normal distribution*. 51. Department of Statistics, Stanford University Stanford, CA.

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* 8 (5), 577–606.

Baxter, J. (1997). A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning* 28, 7–39.

Bellman, R. (1966). Dynamic programming. *Science* 153 (3731), 34–37.

Beran, R. (1996). Madan Puri Festschrift. In: ed. by E. Brunner and M. Denker. VSP Editors, Zeist. Chap. Stein estimation in high dimensions: a retrospective.

Berger, J. O. (1976). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *The Annals of Statistics*, 223–226.

Berger, J. O. and Delampady, M. (1987). Testing Precise Hypotheses. *Statistical Science* 2 (3), 317–335.

Bickel, P. J. (1969). A distribution free version of the Smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics* 40 (1), 1–23.

Birgé, L. (2001). An alternative point of view on Lepski's method. In: *State of the art in probability and statistics (Leiden, 1999)*. Vol. 36. IMS Lecture Notes Monogr. Ser. Inst. Math. Statist., 113–133.

Blanchard, G., Carpentier, A., and Gutzeit, M. (2018). Minimax Euclidean separation rates for testing convex hypotheses in $\mathbb{R}^d$. *Electronic Journal of Statistics* 12 (2), 3713–3735.

Blanchard, G. and Fermanian, J.-B. (2023). Nonasymptotic One- and Two-Sample Tests in High Dimension with Unknown Covariance Structure. *Foundations of Modern Statistics (Festschrift in Honor of Vladimir Spokoiny)*. Ed. by D. Belomestny, C. Butucea, E. Mammen, E. Moulines, M. Reiß, and V. V. Ulyanov. Springer, 121–162.

Blanchard, G., Fermanian, J.-B., and Marienwald, H. (2024). Estimation of multiple mean vectors in high dimension. *arXiv preprint arXiv:2403.15038*.

Bock, M. E. (1975). Minimax estimators of the mean of a multivariate normal distribution. *The Annals of Statistics*, 209–218.

Bonilla, E. V., Chai, K., and Williams, C. (2007). Multi-task Gaussian process prediction. *Advances in neural information processing systems* 20.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22 (14), e49–e57.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, 144–152.

Boucheron, S., Lugosi, G., and Bousquet, O. (2004). Concentration Inequalities. In: *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*. Ed. by O. Bousquet, U. von Luxburg, and G. Rätsch. Berlin, Heidelberg: Springer Berlin Heidelberg, 208–240.

Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematiques de l'Académie des Sciences* 334 (6), 495–500.

Brehmer, J. and Cranmer, K. (2020). Flows for simultaneous manifold learning and density estimation. *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 442–453.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *The Annals of Mathematical Statistics* 42 (3), 855–903.

Brown, L. D. and Greenshtein, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, 1685–1704.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems* 33, 1877–1901.

Carriere, M., Cuturi, M., and Oudot, S. (2017). Sliced Wasserstein kernel for persistence diagrams. *International conference on machine learning*. PMLR, 664–673.

Caruana, R. (1997). Multitask learning. *Machine learning* 28 (1), 41–75.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician* 39 (2), 83–87.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. *Advances in neural information processing systems* 31.

Cheng, Y., Diakonikolas, I., and Ge, R. (2019). High-dimensional robust mean estimation in nearly-linear time. *Proceedings of the thirtieth annual ACM-SIAM symposium on discrete algorithms*. SIAM, 2755–2771.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 493–507.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. *Proc. of the 33rd International Conference on Machine Learning (ICML 2016)*. Vol. 48, 2606–2615.

Cohn, D. L. (1980). *Measure theory*. English. Birkhauser Boston.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning* 20, 273–297.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Dasgupta, S. and Freund, Y. (2008). Random projection trees and low dimensional manifolds. *Proceedings of the fortieth annual ACM symposium on Theory of computing*, 537–546.

Depersin, J. and Lecué, G. (2022). Robust sub-Gaussian estimation of a mean vector in nearly linear time. *The Annals of Statistics* 50 (1), 511–536.

Dette, H., Kokot, K., and Aue, A. (2020a). Functional data analysis in the Banach space of continuous functions. *The Annals of Statistics* 48 (2), 1168–1192.

Dette, H., Kokot, K., and Volgushev, S. (2020b). Testing relevant hypotheses in functional time series via self-normalization. *Journal of the Royal Statistical Society: Series B* 82 (3), 629–660.

Dette, H. and Munk, A. (1998). Nonparametric comparison of several regression functions: exact and asymptotic theory. *The Annals of Statistics* 26 (6), 2339–2368.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics*.

Devroye, L., Lerasle, M., Lugosi, G., and Oliveira, R. I. (2016). Sub-Gaussian mean estimators. *The Annals of Statistics* 44 (6), 2695–2725.

Doran, G., Muandet, K., Zhang, K., and Schölkopf, B. (2014). A Permutation-Based Kernel Conditional Independence Test. *UAI*, 132–141.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Duan, Y. and Wang, K. (2023). Adaptive and robust multi-task learning. *The Annals of Statistics* 51 (5), 2015–2039.

Dussap, B., Blanchard, G., and Chérief-Abdellatif, B.-E. (2023). Label shift quantification with robustness guarantees via distribution feature matching. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 69–85.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via Maximum Mean Discrepancy optimization. *Conference on Uncertainty in Artificial Intelligence*.

Efron, B. and Morris, C. (1972). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika* 59 (2), 335–347.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association* 68 (341), 117–130.

Efron, B. and Morris, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *The Annals of Statistics* 4 (1), 22–32.

Ermakov, M. S. (1991). Minimax detection of a signal in a Gaussian white noise. *Theory of Probability & Its Applications* 35 (4), 667–679.

Evgeniou, T., Micchelli, C. A., and Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research* 6 (21), 615–637.

Fakoor, R., Chaudhari, P., Mueller, J., and Smola, A. J. (2020). Trade: Transformers for density estimation. *arXiv preprint arXiv:2004.02441*.

Fathi, M., Goldstein, L., Reinert, G., and Saumard, A. (2020). Relaxing the Gaussian assumption in shrinkage and SURE in high dimension. *The Annals of Statistics*.

Feldman, S., Gupta, M. R., and Frigyik, B. A. (2014). Revisiting Stein's paradox: multi-task averaging. *Journal of Machine Learning Research* 15 (106), 3621–3662.

Filippi, S., Flaxman, S., Sejdinovic, D., and Cunningham, J. (2016). Bayesian learning of kernel embeddings. *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*. Association for Computing Machinery.

Finak, G., Langweiler, M., Jaimes, M., Malek, M., Taghiyar, J., Korin, Y., Raddassi, K., Devine, L., Obermoser, G., Pekalski, M. L., et al. (2016). Standardizing flow cytometry

immunophenotyping analysis from the Human Immunophenotyping Consortium. *Scientific reports* 6, 1–11.

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 697–717.

Fromont, M., Laurent, B., Lerasle, M., and Reynaud-Bouret, P. (2012). Kernels based tests with non-asymptotic bootstrap approaches for two-sample problems. *Proceedings of the 25th Annual Conference on Learning Theory*. Ed. by S. Mannor, N. Srebro, and R. C. Williamson. Vol. 23. Proceedings of Machine Learning Research, 1–23.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007). Kernel measures of conditional dependence. *Advances in neural information processing systems* 20.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. (2022). What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems* 35, 30583–30598.

Gärtner, T. (2003). A survey of kernels for structured data. *ACM SIGKDD explorations newsletter* 5 (1), 49–58.

Genevay, A., Peyré, G., and Cuturi, M. (2018). Learning generative models with sinkhorn divergences. *International Conference on Artificial Intelligence and Statistics*. PMLR, 1608–1617.

George, E. I. (1986). Minimax multiple shrinkage estimation. *The Annals of Statistics* 14 (1), 188–205.

Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. (2023). A mathematical perspective on Transformers. *arXiv preprint arXiv:2312.10794*.

Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. (2024). The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems* 36.

Giraud, C. (2021). *Introduction to high-dimensional statistics*. Chapman and Hall/CRC.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research* 13 (25), 723–773.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007). A kernel statistical test of independence. *Advances in Neural Information Processing Systems (NeurIPS 2007)*.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., Schölkopf, B., et al. (2005). Kernel methods for measuring independence.

Gruber, M. (1998). *Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators*. Vol. 156. CRC Press.

Gusfield, D. (1997). Algorithms on stings, trees, and sequences: Computer science and computational biology. *Acm Sigact News* 28 (4), 41–60.

Hagrass, O., Sriperumbudur, B. K., and Li, B. (2022). Spectral regularized kernel two-sample tests. *arXiv preprint arXiv:2212.09201*.

Hausdorff, F. (1918). Dimension und äußeres Maß. *Mathematische Annalen* 79 (1), 157–179.

Heinonen, J. (2001). *Lectures on analysis on metric spaces*. Springer Science & Business Media.

Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* 58 (301), 13–30.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12 (1), 55–67.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* 24 (6), 417.

Houdré, C. and Reynaud-Bouret, P. (2003). Exponential inequalities, with constants, for U-statistics of order two. In: *Stochastic Inequalities and Applications*. Progress in Probability 56, 55–69.

Hsu, D., Kakade, S., and Zhang, T. (2012). Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electronic Communications in Probability* 17 (none), 1–13.

Ingster, Y. I. (1982). Minimax nonparametric detection of signals in white Gaussian noise. *Problems of Information Transmission* 18 (2), 130–140.

Ingster, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives I-II-III. *Mathematical Methods of Statistics* 2 (2–4), 85–114, 171–189, 249–268.

Ingster, Y. and Suslina, I. A. (2012). *Nonparametric goodness-of-fit testing under Gaussian models*. Lecture Notes in Statistics 169. Springer.

Ingster, Y. I. and Suslina, I. A. (1998). Minimax detection of a signal for Besov bodies and balls. *Problems of Information Transmission* 34 (1), 48–59.

Isham, V. (1993). Statistical aspects of chaos: a review. *Networks and Chaos-Statistical and Probabilistic Aspects*, 124–200.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems* 31.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. of the 4th Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1, 361–379.

Jegelka, S., Gretton, A., Schölkopf, B., Sriperumbudur, B. K., and Von Luxburg, U. (2009). Generalized clustering via kernel embeddings. *Annual Conference on Artificial Intelligence (KI 2009)*. Springer, 144–152.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics* 37 (4), 1647–1684.

Jirak, M. and Wahl, M. (2018). Perturbation bounds for eigenspaces under a relative gap condition. *Proceedings of the American Mathematical Society*.

Joulin, A., Grave, É., Bojanowski, P., and Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431.

Juditsky, A. and Nemirovski, A. (2002). On nonparametric tests of positivity / monotonicity / convexity. *The Annals of Statistics* 30 (2), 498–527.

Kim, H., Papamakarios, G., and Mnih, A. (2021). The lipschitz constant of self-attention. *International Conference on Machine Learning*. PMLR, 5562–5571.

Kim, I., Balakrishnan, S., and Wasserman, L. A. (2020). Minimax optimality of permutation tests. *The Annals of Statistics*.

Kivinen, J. and Warmuth, M. K. (1997). Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation* 132 (1), 1–63.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge didistribuzione. *Giorn Dell'inst Ital Degli Att* 4, 89–91.

Koltchinskii, V. and Lounici, K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 52 (4), 1976–2013.

Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* 23 (1), 110–133.

Kpotufe, S. (2011). k-NN regression adapts to local intrinsic dimension. *Advances in neural information processing systems* 24.

Lam-Weil, J., Carpentier, A., and Sriperumbudur, B. K. (2022). Local minimax rates for closeness testing of discrete distributions. *Bernoulli* 28 (2), 1179–1197.

Laurent, B. and Massart, P. (2000). Adaptive Estimation of a Quadratic Functional by Model Selection. *The Annals of Statistics* 28 (5), 1302–1338.

Laurent, B., Loubes, J.-M., and Marteau, C. (2012). Non asymptotic minimax rates of testing in signal detection with heterogeneous variances. *Electronic Journal of Statistics* 6 (none), 91–122.

Lecué, G. and Rigollet, P. (2014). Optimal learning with Q-aggregation. *The Annals of Statistics* 42 (1).

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Vol. 23. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Lepski, O. V. and Spokoiny, V. G. (1999). Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli* 5 (2), 333–358.

Lerasle, M., Szabo, Z., Mathieu, T., and Lecue, G. (2019). MONK Outlier-Robust Mean Embedding Estimation by Median-of-Means. *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 3782–3793.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. (2017). Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems* 30.

Li, T. and Yuan, M. (2019). On the optimality of Gaussian kernel based nonparametric tests against smooth alternatives. *arXiv preprint arXiv:1909.03302*.

Li, Y., Swersky, K., and Zemel, R. (2015). Generative moment matching networks. *International conference on machine learning*. PMLR, 1718–1727.

Lugosi, G. and Mendelson, S. (2019a). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics* 19 (5), 1145–1190.

Lugosi, G. and Mendelson, S. (2019b). Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics* 47 (2), 783–794.

Lugosi, G. and Mendelson, S. (2020). Multivariate mean estimation with direction-dependent accuracy. *Journal of the European Mathematical Society*.

Marienwald, H., Fermanian, J.-B., and Blanchard, G. (2021). High-Dimensional Multi-Task Averaging and Application to Kernel Mean Embedding. *International Conference on Artificial Intelligence and Statistics AISTATS*. PMLR, 1963–1971.

Martínez-Rego, D. and Pontil, M. (2013). Multi-task averaging via task clustering. *Similarity-Based Pattern Recognition: Second International Workshop, SIMBAD 2013, York, UK, July 3-5, 2013. Proceedings 2*. Springer, 148–159.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability* 28 (2), 863–884.

Massart, P. (2003). *Concentration Inequalities and Model Selection*. Springer.

Massart, P. (2007). *Concentration inequalities and model selection*. Vol. 1896. Lecture notes in mathematics. Springer.

Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M., Gretton, A., and Muandet, K. (2021). Proximal causal learning with kernels: Two-stage estimation and moment restriction. *International conference on machine learning*. PMLR, 7512–7523.

McAuley, J. (2022). *Personalized machine learning*. Cambridge University Press.

McDiarmid, C. et al. (1989). On the method of bounded differences. *Surveys in combinatorics* 141 (1), 148–188.

McDiarmid, C. (1998). Concentration. In: *Probabilistic Methods for Algorithmic Discrete Mathematics*. Ed. by M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed. Berlin, Heidelberg: Springer Berlin Heidelberg, 195–248.

McKinnon, K. M. (2018). Flow cytometry: an overview. *Current protocols in immunology* 120 (5), 1–11.

Micchelli, C. and Pontil, M. (2004). Kernels for Multi–task Learning. *Advances in neural information processing systems* 17.

Minsker, S. (2019). Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics* 13 (2), 5213–5252.

Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. (2012). Learning from distributions via support measure machines. *Advances in Neural Information Processing Systems 25 (NeurIPS 2012)*, 1–9.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Gretton, A., and Schölkopf, B. (2014). Kernel mean estimation and Stein effect. *International Conference on Machine Learning (ICML 2014)*. PMLR, 10–18.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel mean embedding of distributions: a review and beyond. *Foundations and Trends in Machine Learning* 10 (1-2), 1–141.

Muandet, K., Sriperumbudur, B., Fukumizu, K., Gretton, A., and Schölkopf, B. (2016). Kernel mean shrinkage estimators. *Journal of Machine Learning Research* 17 (48), 1–41.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in applied probability* 29 (2), 429–443.

Munk, A. and Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 60 (1), 223–241.

Naumov, A., Spokoiny, V. G., and Ulyanov, V. (2019). Bootstrap confidence sets for spectral projectors of sample covariance. *Probability Theory and Related Fields* 174 (3), 1091–1132.

Nussbaum, M. (1996). Encyclopedia of Statistical Sciences. In: ed. by S. Kotz. Wiley. Chap. The Pinsker bound: A review.

Oliva, J., Póczos, B., and Schneider, J. (2013). Distribution to distribution regression. *International Conference on Machine Learning*. PMLR, 1049–1057.

OpenAI et al. (2024). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs.CL].

Ostrovskii, D. M., Ndaoud, M., Javanmard, A., and Razaviyayn, M. (2020). *Near-Optimal Model Discrimination with Non-Disclosure*. arXiv: 2012.02901 [math.ST].

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50 (302), 157–175.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2 (11), 559–572.

Pinelis, I. and Sakhanenko, A. I. (1986). Remarks on Inequalities for Large Deviation Probabilities. *Theory of Probability & Its Applications* 30 (1), 143–148.

Pinsker, M. S. (1980). Optimal filtering of square-integrable signals in Gaussian noise. *Problemy Peredachi Informatsii* 16 (2), 52–68.

Rio, E. (2002). Une inégalité de Bennett pour les maxima de processus empiriques. *Annales de l'IHP Probabilités et statistiques* 38 (6), 1053–1057.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. *Proceedings of the second Berkeley symposium on mathematical statistics and probability*. Vol. 2. University of California Press, 131–149.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics* 35 (1), 1–20.

Rudelson, M. and Vershynin, R. (2007). Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)* 54 (4), 21–es.

Sander, M. E., Ablin, P., Blondel, M., and Peyré, G. (2022). Sinkformers: Transformers with doubly stochastic attention. *International Conference on Artificial Intelligence and Statistics*. PMLR, 3515–3530.

Saunders, C., Gammerman, A., and Vovk, V. (1998). Ridge Regression Learning Algorithm in Dual Variables. *International Conference on Machine Learning*.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation* 10 (5), 1299–1319.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. (2023). MMD Aggregated Two-Sample Test. *Journal of Machine Learning Research* 24 (194), 1–81.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The annals of statistics*, 2263–2291.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Simpson, E. H. (1949). Measurement of diversity. *Nature* 163 (4148), 688–688.

Singh, R., Sahani, M., and Gretton, A. (2019). Kernel instrumental variable regression. *Advances in Neural Information Processing Systems* 32.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A Hilbert space embedding for distributions. *Proc. International Conference on Algorithmic Learning Theory (ALT 2007)*, 13–31.

Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets. *Annals of Statistics* 24 (6), 2477–2498.

Spokoiny, V. G. (2012). Parametric estimation. Finite sample theory. *The Annals of Statistics* 40 (6), 2877–2909.

Spokoiny, V. G. and Dickhaus, T. (2015). *Basics of modern mathematical statistics*. Springer Texts in Statistics. Springer.

Spokoiny, V. G. and Zhilova, M. (2013). Sharp deviation bounds for quadratic forms. *Mathematical Methods of Statistics* 22 (2), 100–113.

Spokoiny, V. G. and Zhilova, M. (2015). Bootstrap confidence sets under model misspecification. *Annals of Statistics* 43 (6), 2653–2675.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. (2011). Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research* 12 (7).

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Lanckriet, G., and Schölkopf, B. (2008). Injective Hilbert space embeddings of probability measures. *21st annual conference on learning theory (COLT 2008)*. Omnipress, 111–122.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research* 11, 1517–1561.

Sriperumbudur, B. K. and Sterge, N. (2022). Approximate kernel pca: Computational versus statistical trade-off. *The Annals of Statistics* 50 (5), 2713–2736.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955*. Vol. 1, 197–206.

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research* 2 (Nov), 67–93.

Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. (2016). Learning theory for distribution regression. *Journal of Machine Learning Research* 17 (152), 1–40.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones mathematicae* 126, 505–563.

Tan, A. Z., Yu, H., Cui, L., and Yang, Q. (2022). Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Tolstikhin, I., Sriperumbudur, B. K., Mu, K., et al. (2017). Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research* 18 (86), 1–47.

Tolstikhin, I. O., Sriperumbudur, B. K., and Schölkopf, B. (2016). Minimax estimation of maximum mean discrepancy with radial kernels. *Advances in Neural Information Processing Systems* 29.

Tropp, J. A. et al. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning* 8 (1-2), 1–230.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. 1st. Springer Publishing Company, Incorporated.

van Handel, R. (2017). Structured random matrices. In: *Convexity and concentration*. Springer, 107–156.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems* 30.

Verma, N., Kpotufe, S. K., and Dasgupta, S. (2009). Which spatial partition trees are adaptive to intrinsic dimension? *25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*.

Vershynin, R. (2018). *High-Dimensional Probability: an introduction with applications to data science*. Cambridge series in statistical and probabilistic mathematics 47. Cambridge University Press.

Wang, Z., Lan, L., and Vucetic, S. (2011). Mixture model for multiple instance regression and applications in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing* 50 (6), 2226–2237.

Wellek, S. (2002). *Testing statistical hypotheses of equivalence*. Chapman and Hall/CRC.

Xia, X., Shan, S., Gong, M., Wang, N., Gao, F., Wei, H., and Liu, T. (2022). Sample-Efficient Kernel Mean Estimator with Marginalized Corrupted Data. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2110–2119.

Young, L.-S. (1982). Dimension, entropy and Lyapunov exponents. *Ergodic theory and dynamical systems* 2 (1), 109–124.

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods. *Annals of Statistics*, 379–390.

Zhang, J., Marszałek, M., Lazebnik, S., and Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision* 73, 213–238.

Zhang, R., Frei, S., and Bartlett, P. L. (2024). Trained transformers learn linear models in-context. *Journal of Machine Learning Research* 25 (49), 1–55.

Zhang, Y. and Yang, Q. (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering* 34 (12), 5586–5609.