Compressive Statistical Learning with Random Feature Moments

Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, Yann Traonmilin

Compressive Statistical Learning with Random Feature Moments

R. Gribonval, G. Blanchard, N. Keriven, Y. Traonmilin

June 21, 2017

Abstract

We describe a general framework –*compressive statistical learning*– for resource-efficient large-scale learning: the training collection is compressed in one pass into a low-dimensional *sketch* (a vector of random empirical generalized moments) that captures the information relevant to the considered learning task. A near-minimizer of the risk is computed from the sketch through the solution of a nonlinear least squares problem. We investigate sufficient sketch sizes to control the generalization error of this procedure. The framework is illustrated on compressive clustering, compressive Gaussian mixture Modeling with fixed known variance, and compressive PCA.

1 Introduction

Large-scale machine learning faces a number of fundamental computational challenges, triggered both by the high dimensionality of modern data and the increasing availability of very large training collections. Besides the need to cope with high-dimensional features extracted from images, volumetric data, etc., a key challenge is to develop techniques able to fully leverage the information content and learning opportunities opened by large training collections of millions to billions or more items, with controlled computational resources.

Such training volumes can severely challenge traditional statistical learning paradigms based on batch empirical risk minimization. Statistical learning offers a standardized setting where learning problems are expressed as the optimization of an expected loss, or risk, $\mathcal{R}(\pi_0, h) := \mathbb{E}_{X\sim\pi_0}\ell(X, h)$ over a parameterized family of hypotheses $\mathcal{H}$ (where π_0 is the probability distribution of the training collection). This risk is empirically estimated on a training collection, and parameters that empirically minimize it are seeked, possibly with some regularization. Empirical minimization typically requires access to the whole training collection, either in batch mode or iteratively with one or more passes of stochastic gradient. This can become prohibitively costly when the collection is large and each iteration has non-negligible cost. An alternative is to sub-sample the collection, but this may come at the price of neglecting some important items from the collection. Besides online learning (e.g., [70]), sampling techniques such as coresets [47] or Nyström's method (e.g., [75]) have emerged to circumvent computational bottlenecks and preserve the ability to exploit latent information from large collections.

Can we design an alternate learning framework, with the ability to compress the training collection before even starting to learn? We advocate a possible route, *compressive statistical learning*, which is inspired by the notion of *sketching* and is endowed with favorable computational features especially in the context of the streaming and distributed data model [27] (see Section 1.3). Rooted both in the generalized method of moments [57] and in compressive sensing [50], it leverages recent techniques from kernel methods such as kernel mean embeddings [79] and random Fourier features [73] to obtain innovative statistical guarantees.

As a trivial example, assume x, h belong to $\mathbb{R}^d$, and consider the squared loss $\ell(x,h) = \|x-h\|^2$, whose risk minimizer is $\mathbb{E}[X]$. In this specific example, keeping only the d empirical averages of the coordinates of X is obviously sufficient. The vision developed in this paper is that, for certain learning problems, all the necessary information can be captured in a *sketch*: a vector of empirical (generalized) moments of the collection that captures the information relevant to the considered learning task. Computing the sketch is then feasible in one pass, and a near-minimizer of the risk can be computed from the sketch with controlled generalization error.

This paper is dedicated to show how this phenomenon can be generalized: roughly speaking, can the sketch size be taken to be proportional to the number of "intrinsic parameters" of the learning task? Another fundamental requirement for the sketching operation is to be online. When recording the training collection, it should be possible to update the sketch at almost no additional cost. The original training collection can then be discarded and learning can be performed from the sketch only, potentially leading to privacy-preservation. We will see that a sketching procedure based on random generalized moments meets these requirement for clustering and Gaussian mixture estimation.

1.1 Inspiration from compressive sensing

Another classical example of learning task is Principal Component Analysis (PCA). In this setting, $x \in \mathbb{R}^d$, h is an arbitrary linear subspace of dimension k, and the loss is $\ell(x,h) = \|x - \mathrm{P}_h x\|_2^2$ with P_h the orthogonal projector onto h. The matrix of second moments $\mathbf{\Sigma}_\pi := \mathbb{E}_{X\sim\pi} XX^T$ is known to summarize all the information needed to select the best subspace for a training collection. It thus constitutes a natural sketch (of finite dimension d^2) of the training set.

A much smaller sketch can in fact be computed. Results from compressive sensing and low-rank matrix completion [50] allow to compress the matrix of second moments to a sketch of dimension $\mathcal{O}(kd)$ (much smaller that d^2 when $k \ll d$) from which the best rank-k approximation to $\mathbf{\Sigma}_\pi$ can be accurately estimated (this rank-k approximation allows to calculate the PCA with appropriate learning guarantees, as we will see Section 3). This compression operation is made using random linear projections on $\mathbf{\Sigma}_\pi$, which can be seen as random second order moments of the training collection.

We propose to generalize such a sketching procedure to arbitrary random generalized moments. Given a learning task and training collection, we study the following questions:

- How can we perform learning from a sketch of the training collection?
- What learning guarantees can we obtain with such a procedure?

1.2 Contributions

In this paper, we present a general compressive learning framework.

- We describe a generic **sketching mechanism** with random generalized moments and provide a theoretical learning procedure from the sketched data.
- We derive general **learning guarantees** for sketching with random generalized moments.
- We apply this framework to **compressive clustering**, demonstrating that a sketch of dimension $\mathcal{O}\left(k^2d^2 \cdot (1 + \log(kd) + \log(R/\varepsilon))\right)$, with k the prescribed number of clusters, R a bound on the norm of the centroids, and ε the separation between them, is sufficient to obtain statistical guarantees. To the best of our knowledge this is the first time that such guarantees are given.
- We apply this framework to **compressive Gaussian mixture estimation** with known covariance. In this case, we identify a finite sketch size sufficient to obtain statistical guarantees under a separation assumption between means expressed in the Mahalanobis norm associated to the

known covariance matrix. A parameter embodies the tradeoff between sketch size and separation. At one end of the spectrum the sketch size is polynomial in k and exponential in d and guarantees are given for means that can be separated in $\mathcal{O}\left(\sqrt{\log k}\right)$, which compares favorably to existing literature [1, 83] (recent works make use of more complex conditions that theoretically permits arbitrary separation [13], however all these approaches use the full data while we consider a compressive approach that uses only a sketch of the data), while at the other end the sketch size is polynomial in k and d but the required separation is in $\mathcal{O}\left(\sqrt{d \log k}\right)$.

We finally briefly discuss the potential impact of the proposed framework and its extensions in terms of privacy-aware learning and of the insight it may bring on the information-theoretic properties of certain convolutional neural networks.

1.3 Related work

Sketching and streaming methods. *Sketches* are closely linked with the development of *streaming methods* [27], in which data items are seen once by the user then discarded. A sketch is a small summary of the data seen at a given time, that can be queried for a particular piece of information about the data. As required by the streaming context, when the database is modified, e.g. by inserting or deleting an element, the subsequent update of the sketch must be very fast. In practice, sketches are often applied in context where the data are stored in multiple places. In this heavily distributed framework, a popular class of sketch is that of *linear* sketches, i.e. structures such that the sketch of the union of two databases is the sum of their sketches – then the sketch of a database distributed over several parts is simply the sum of all their sketches. The sketch presented in this work is indeed a linear sketch (when considered without the normalization constant $1/n$) and as such, updates operations are excessively simple and fast. Sketches have been used for a large variety of operations [27] such as the popular detection of heavy-hitters [30, 31, 29]. Closer to our framework, sketches have been used to approximately maintain histograms [80] or quantiles [53], however these methods are subject to the well-known curse of dimensionality and are unfeasible even in moderate dimension.

Learning in a streaming context. Various learning algorithms have also been directly adapted to a streaming context. Examples include the Expectation-Maximization algorithm [5, 21], the k-means algorithm [56, 3], or Principal Component Analysis [52]. In each case, the result of the algorithm is updated as data arrive. However these algorithms do not fully benefit from the many advantages of sketches. Sketches are simpler to merge in a distributed context, update operations are more immediate, and the learning step can be delocalized and performed on a dedicated machine.

Coresets. Another popular class of structures that summarize a database for learning is called *coresets*. Coresets were initially developed for k-means [59] or, more generally, subspace approximation [48, 47] and also applied to learning Gaussian Mixture Models [46, 69]. In a sense, the philosophy behind coresets is situated halfway between sketches and streaming learning algorithms. Like the sketching approaches, coresets methods construct a compressed representation of the database (or "coreset"), but are somehow closer to already approximately performing the learning task. For instance, the coreset described in [51] already incorporates steps of Lloyd's k-means algorithm in its construction. Similar to the k-means++ algorithm [7], many coresets have been developed as (weighted) adaptive subsampling of the data [46, 69].

Linear sketches *vs* Coresets. It is in general difficult to compare sketching and coresets methods (including the sketching method presented in this paper) in terms of pure performance or theoretical guarantees, since they are very different approaches that can be more or less adapted to certain contexts. We can however outline some differences. Unlike sketches, coresets are not specifically build

for the streaming context, and they may require several passes over the data. Nevertheless they can still be adapted to streams of data, as described e.g. in [59, 47, 69], by using a merge-and-reduce hierarchical strategy: for each batch of data that arrives sequentially, the user builds a coreset, then groups these coresets and build a coreset of coresets, and so on. This update method is clearly less direct than updating a linear sketch, and more importantly the user must balance between keeping many coresets and letting the size of the overall summary grow with the number of points in the database, or keeping only highest-level coresets at the cost of losing precision in the theoretical guarantees each time the height of the hierarchical structure increases. As a comparison, the sketch presented in this work for k-means (Section 4) does not have these limitations: like any linear sketch, updates are totally independent of previous events, and for a fixed sketch size the ability to perform the learning task strictly increases with the number of points.

Generalized Method of Moments and Compressive Sensing. The methodology that we employ to develop the proposed sketching framework is similar to a Generalized Method of Moments (GeMM) [65, 57]: the parameters θ of a model are learned by matching a collection of theoretical generalized moments from the distribution π_θ with empirical ones from the data. GeMM is often seen as an alternative to Maximum Likelihood estimation, to obtain different identifiability guarantees [12, 60, 4] or when the likelihood is not available. Traditionally, a finite number of moments is considered, but recent developments give guarantees when an infinite (integral) number of generalized moments are available [22, 24], in particular generalized moments associated to the (empirical) characteristic function [23, 49]. Our point of view is slightly different: we consider the collection of moments as a *compressed* representation of the data and as a mean to achieve a learning task.

Compared to the guarantees usually obtained in GeMM such as consistency and efficiency of the estimator $\hat{\theta}$, the results that we obtain are more akin to Compressive Sensing and Statistical Learning. For instance, when learning Gaussian Mixture Model (Section 5), we prove that learning is robust to modeling error (the true distribution of the data is not exactly a GMM but close to one), which is generally overlooked is GeMM. In the proof technique, this is done by replacing the so-called "global identifiability condition", (i.e. injectivity of the moment operator, which is a classical condition in GeMM but is already difficult to prove and sometimes simply assumed by practitioner, see [71, p. 2127]) by the strictly stronger Lower Restricted Isometry Property (LRIP) from the Compressive Sensing literature [37, 20, 10, 50]. This is achieved by considering *random* feature moments (related to random features [73, 74, 8] and kernel mean embeddings [79]), so in a sense the resulting Compressive Statistical Learning framework could be considered as a *Method of Random Feature Moments*. While the LRIP is reminiscent of certain kernel approximation guarantees with random features (see e.g. [78, 8], it is in fact of a different nature, and none seems to be a direct consequence of the other.

1.4 Outline

Section 2 describes our general framework for compressive statistical learning. We define here statistical learning guarantees, introduce the required notions and state our general Theorem for statistical learning guarantees for compressive learning. To familiarize the reader with the proposed framework, we detail in Section 3 a procedure for Compressive PCA, where we do not intend to match the latest developments in the domain of PCA such as stochastic and incremental PCA [6, 9] but rather to give a first illustration. Section 4 (respectively Section 5) specifies a sketching procedure and states the associated learning guarantees for compressive clustering (respectively for compressive Gaussian mixture estimation). Section 6 describes precisely how learning guarantees can be obtained when estimation of mixtures of elementary distributions are involved as in the two examples of compressive clustering and compressive Gaussian mixture estimation. We discuss in Section 7 possible extensions of the proposed framework as well as the insight it may bring on the information flow across one layer of a convolutive neural network with average pooling. Finally, all proofs are stated in the Annex.

2 A general compression framework for statistical learning

This section is dedicated to the introduction of our compressive learning framework.

2.1 Statistical learning

Statistical learning offers a standardized setting where many learning problems can be expressed as the optimization of an expected risk over a parameterized family of functions. Formally, we consider a training collection $\mathbf{X} = \{x_i\}_{i=1}^n \in \mathcal{Z}^n$ drawn i.i.d. from a probability distribution π_0 on the set $\mathcal{Z}$. In our examples, $\mathcal{Z} = \mathbb{R}^d$. One wishes to select a hypothesis h from a hypothesis class $\mathcal{H}$ to perform the task at hand. How well the task can be accomplished with the hypothesis h is typically measured through a *loss function* $\ell : (x, h) \mapsto \ell(x, h)$ and the *expected risk* associated to h:

$$\mathcal{R}(\pi_0, h) := \mathbb{E}_{X\sim\pi_0}\, \ell(X, h)\,. \tag{1}$$

In the idealized learning problem, one selects the function $h^\star$ that minimizes the expected risk

$$h^\star \in \arg\min_{h\in\mathcal{H}} \mathcal{R}(\pi_0, h). \tag{2}$$

In practice one has no access to the true risk $\mathcal{R}(\pi_0, h)$ since the expectation with respect to the underlying probability distribution, $\mathbb{E}_{X\sim\pi_0}[\cdot]$, is unavailable. Instead, methods such as *empirical risk minimization (ERM)* produce an estimated hypothesis $\hat{h}$ from the training dataset $\mathbf{X}$. One expects to produce, with high probability at least $1 - \zeta$ on the draw of the training set, the bound on the excess risk

$$\mathcal{R}(\pi_0, \hat{h}) - \mathcal{R}(\pi_0, h^\star) \leq \eta_n = \eta_n(\zeta), \tag{3}$$

where η_n is a control that has good behavior with respect to n. We will use three running examples.

Examples:

- **PCA**: as stated in the introduction, the loss function is $\ell(x, h) = \|x - \mathrm{P}_h x\|_2^2$ where P_h is the orthogonal projection onto the subspace hypothesis h of prescribed dimension k.
- **k-means clustering**: each hypothesis corresponds to a set of k candidate cluster centers, $h = \{c_1, \ldots, c_k\}$, and the loss is defined by the k-means cost $\ell(x, h) = \min_{1\leq l\leq k} \|x - c_l\|_2^2$. The hypothesis class $\mathcal{H}$ may be further reduced by defining constraints on the considered centers (e.g., in some domain, or as we will see with some separation between centers).
- **Gaussian Mixture Modeling**: each hypothesis h corresponds to the collection of weights, means and variances of mixture of k Gaussians, which probability density function is denoted $\pi_h(x)$. The loss function is based on the maximum likelihood $\ell(x, h) = -\log \pi_h(x)$.

2.2 Compressive learning

Our aim, and one of the major achievements of this paper, is to control the excess risk (3) using an estimate $\hat{h}$ *obtained from the sole knowledge of a sketch of the training collection*. As we will see, the resulting philosophy for large scale learning is, instead of addressing an ERM optimization problem of size proportional to the number of training samples, to first compute a sketch vector *of size driven by the complexity of the task*, then to address a nonlinear least-squares optimization problem associated to the *Generalized Method of Moments (GeMM)* on this sketch.

Taking its roots in compressive sensing [37, 20, 50] and the generalized method of moments [65, 57], but also on kernel mean embeddings [76, 79], random features [73, 74, 8], and streaming algorithms [53, 30, 28], *compressive learning* has three main steps:

1. Choose (at random) a (nonlinear) *feature function* $\Phi : \mathcal{Z} \mapsto \mathbb{R}^m$ or $\mathbb{C}^m$.

2. Compute (random) generalized moments using the feature function of the training collection to summarize it into a single *sketch vector*

$$\mathbf{y} := \text{Sketch}(\mathbf{X}) := \frac{1}{n}\sum_{i=1}^{n} \Phi(x_i) \in \mathbb{R}^m \text{or } \mathbb{C}^m; \tag{4}$$

3. Produce an hypothesis from the sketch using an appropriate learning procedure: $\hat{h} = \text{Learn}(\mathbf{y})$.

Overall, the goal is to design the sketching function $\Phi(\cdot)$ and the learning procedure $\text{Learn}(\cdot)$ given a learning task (i.e., a loss function) such that the resulting hypothesis $\hat{h}$ has controlled excess risk (3).

Trivial examples.

- Estimation of the mean: Assume x, h belong to $\mathbb{R}^d$, and consider the squared loss $\ell(x, h) = \|x - h\|^2$, whose risk minimizer is $\mathbb{E}[X]$. In this specific example, it is obviously sufficient to keep only the d empirical averages of the coordinates of X, i.e., to use $\Phi(x) := x$.
- PCA: As the principal components are calculated from the eigenvalue decomposition of the matrix of second moments of the samples, we can simply use $\Phi(x) := xx^T$.

A less trivial example is *Compressive PCA*. Instead of estimating the full matrix $\mathbf{\Sigma}_{\pi_0}$, of size $d \times d$, it is known that computing m random gaussian linear measurements of this matrix makes it possible to manipulate a vector $\mathbf{y}$ of dimension $m = \mathcal{O}(kd)$ from which one can accurately estimate the best rank-k approximation to $\mathbf{\Sigma}_{\pi_0}$, that gives the k first principal components. Nuclear norm minimization is typically used to produce this low rank approximation given the vector $\mathbf{y}$. We will describe this procedure in details in Section 3 as a first illustration of our framework.

In Sections 4 and 5, for the more challenging examples of *Compressive k-means* and *Compressive Gaussian Mixture Modeling*, we provide a feature function Φ and a method "Learn" (based on a non-convex least-squares minimization) that leads to a control of the excess risk. This is achieved by establishing links with the formalism of linear inverse problems and low complexity recovery (i.e., sparse/structured vector recovery, low-rank matrix recovery) and extending theoretical tools to the setting of compressive statistical learning.

2.3 Compressive learning as a linear inverse problem

The most immediate link with linear inverse problems is the following. The sketch vector $\mathbf{y}$ can be seen as a *linear* function of the *empirical probability distribution* $\hat{\pi}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ of the training samples:

$$\mathbf{y} = \frac{1}{n}\sum_{i=1}^{n} \Phi(x_i) = \mathcal{A}(\hat{\pi}_n), \tag{5}$$

where $\mathcal{A}$ is a linear operator from the space of distributions to $\mathbb{R}^m$ (or $\mathbb{C}^m$) defined by

$$\mathcal{A}(\pi) := \mathbb{E}_{X\sim\pi}\Phi(X). \tag{6}$$

This is linear in the sense that[1] $\mathcal{A}(\theta\pi + (1-\theta)\pi') = \theta\mathcal{A}(\pi) + (1-\theta)\mathcal{A}(\pi')$ for any π, π' and $0 \leq \theta \leq 1$.

Since for large n we should have $\mathcal{A}(\hat{\pi}_n) \approx \mathcal{A}(\pi_0)$, the sketch $\mathbf{y}$ can be viewed as a noisy linear observation of the underlying probability distribution π_0. This viewpoint allows to formally leverage the general methodology of linear inverse problems to produce an hypothesis from the sketch $\mathbf{y}$.

Conceptually, we will be able to control the excess risk (3) –our goal– if we can:

[1] One can indeed extend $\mathcal{A}$ to a linear operator on the space of finite signed measures, see Annex A.2.

- Define a so-called *decoder* Δ that finds a probability distribution $\hat{\pi}$ given $\mathbf{y}$:

$$\hat{\pi} = \Delta[\mathbf{y}]$$

such that the risk with $\hat{\pi}$ uniformly approximates the risk with π_0:

$$\sup_{h \in \mathcal{H}} |\mathcal{R}(\pi_0, h) - \mathcal{R}(\hat{\pi}, h)| \leq \tfrac{1}{2}\eta_n \tag{7}$$

- Deduce the best hypothesis from this estimate:

$$\hat{h} = \Psi(\hat{\pi}) \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\hat{\pi}, h) \tag{8}$$

Indeed, using (8) and the triangle inequality, it is easy to show that (7) directly implies (3). In a way, this is very similar to ERM except that instead of using the empirical risk $\mathcal{R}(\hat{\pi}_n, \cdot)$, we use an estimate of the risk $\mathcal{R}(\hat{\pi}, \cdot)$ where $\hat{\pi}$ is deduced directly from the sketch $\mathbf{y}$.

Remark 2.1. *At first sight, the above conceptual view may wrongly suggest that compressive learning replaces statistical learning with the much more difficult problem of* density estimation. *Fortunately, as we will see, this is not the case, thanks to the fact that our objective is never to accurately estimate π_0 in the standard sense of density estimation [14], but only to accurately estimate the risk $\mathcal{R}(\pi_0, \cdot)$.*

2.4 Statistical learning guarantees: control of the excess risk

To leverage the links between compressive learning and general inverse problems, we further notice that $\sup_{h \in \mathcal{H}} |\mathcal{R}(\pi, h) - \mathcal{R}(\pi', h)|$ can be viewed as a metric on probability distributions. Given a class $\mathcal{F}$ of measurable functions $f : \mathcal{Z} \to \mathbb{R}$ or $\mathbb{C}$, one can indeed define

$$\|\pi - \pi'\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{E}_{X \sim \pi} f(X) - \mathbb{E}_{X' \sim \pi'} . f(X')| \tag{9}$$

which defines a semi-norm on the space of finite signed measures (see Annex A.2) on $\mathcal{Z}$. With this definition

$$\sup_{h \in \mathcal{H}} |\mathcal{R}(\pi, h) - \mathcal{R}(\pi', h)| = \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \tag{10}$$

where $\mathcal{L}(\mathcal{H}) := \{\ell(\cdot, h) : h \in \mathcal{H}\}$. The desired guarantee (7) then reads $\|\pi_0 - \Delta[\mathbf{y}]\|_{\mathcal{L}(\mathcal{H})} \leq \eta_n/2$.

In the usual context of linear inverse problems, producing an accurate estimate from noisy underdetermined linear observations requires some "regularity" assumption Such an assumption often takes the form of a "low-dimensional" model set that the quantity to estimate is close to.

Example 2.2. *In the case of sparse vector recovery (respectively low-rank matrix recovery), one wishes to estimate $\mathbf{x} \in \mathbb{R}^n$ (resp. $\mathbf{X} \in \mathbb{R}^{n \times n}$) from $\mathbf{y} \approx \mathbf{A}\mathbf{x}$ (resp. $\mathbf{y} \approx \mathbf{A}\mathrm{vec}(\mathbf{X})$, and obtains guarantees provided that $\mathbf{x}$ is close to the set of k-sparse vectors (resp. that $\mathbf{X}$ is close to the set of rank-r matrices).*

Similarly here, estimating $\hat{\pi}$ from $\mathbf{y} \approx \mathcal{A}(\pi_0)$ may require considering some model set $\mathfrak{S}$, which choice and definition will be discussed in Section 2.5.

Remark 2.3. *While in classical compressive sensing the model set plays the role of prior knowledge on the data distribution that completes the observations, in the examples considered here we will obtain* distribution free *excess risk guarantees using models* derived from the loss function.

Given a model set $\mathfrak{S}$ and a sketching operator $\mathcal{A}$, an ideal decoder Δ should satisfy recovery guarantees that can be expressed as: for any distribution π_0, any draw of the training samples from π_0 (embodied by the empirical distribution $\hat{\pi}_n$), with $\mathbf{y} = \mathcal{A}(\hat{\pi}_n)$ and $\hat{\pi} = \Delta[\mathcal{A}(\hat{\pi}_n)]$

$$\|\hat{\pi} - \pi_0\|_{\mathcal{L}(\mathcal{H})} \lesssim d(\pi_0, \mathfrak{S}) + \|\mathcal{A}(\pi_0 - \hat{\pi}_n)\|_2 \tag{11}$$

where $\lesssim$ hides multiplicative constants, and $d(\cdot, \mathfrak{S})$ is some measure of distance to the model set $\mathfrak{S}$.

It turns out that general results from abstract linear inverse problems [16] can be adapted to characterize the *existence* of a decoder satisfying this property, which is a form of *instance optimality* generalizing early formulations for sparsity regularized inverse problems [26]. By [16, Section IV-A], if a decoder with the above property exists then a so-called *lower Restricted Isometry Property (LRIP)* must hold: there is a finite constant $C_{\mathcal{A}} < \infty$ such that

$$\|\pi' - \pi\|_{\mathcal{L}(\mathcal{H})} \leq C_{\mathcal{A}} \|\mathcal{A}(\pi' - \pi)\|_2 \quad \forall \pi, \pi' \in \mathfrak{S}. \tag{12}$$

Conversely, the LRIP (12) implies [16, Theorem 7] that the following decoder (aka *ideal decoder*)

$$\Delta[\mathbf{y}] := \operatorname{argmin}_{\pi \in \mathfrak{S}} \|\mathcal{A}(\pi) - \mathbf{y}\|_2^2 . \tag{13}$$

is instance optimal, i.e., (11) holds for any π_0 and $\hat{\pi}_n$, with the particular distance

$$D(\pi_0, \mathfrak{S}) := \inf_{\sigma \in \mathfrak{S}} \Big\{ \|\pi_0 - \sigma\|_{\mathcal{L}(\mathcal{H})} + 2C_{\mathcal{A}} \|\mathcal{A}(\pi_0 - \sigma)\|_2 \Big\}. \tag{14}$$

As a consequence, the LRIP (12) implies a control of the excess risk achieved with the hypothesis $\hat{h}$ selected with (8), where $\hat{\pi} = \Delta[\mathbf{y}]$, as

$$\mathcal{R}(\pi_0, \hat{h}) - \mathcal{R}(\pi_0, h^\star) \leq 2D(\pi_0, \mathfrak{S}) + 4C_{\mathcal{A}} \|\mathcal{A}(\pi_0 - \hat{\pi}_n)\|_2 \tag{15}$$

where we used explicit constants from [16, Theorem 7].

For large collection size n, the term $\|\mathcal{A}(\pi_0 - \hat{\pi}_n)\|_2$ becomes small and (15) shows that compressive learning will benefit from accurate excess risk guarantees provided the model $\mathfrak{S}$ and the feature function Φ (or equivalently the sketching operator $\mathcal{A}$) are chosen so that:

1. the LRIP (12) holds;
2. the distance $D(\pi_0, \mathfrak{S})$ is "small"; this vague notion will be exploited in Section 2.5 below to guide our choice of $\mathfrak{S}$, and will be made more concrete on examples in Sections 4 and 5;
3. this holds for a "small" value of m, as we also seek to design compact sketches and, eventually, tractable algorithms to learn from them.

All the above considerations will guide our choice of model set $\mathfrak{S}$ and feature function Φ.

2.5 Choice of a model set

Learning tasks such as maximum likelihood estimation directly involve a natural model set: if the loss is $\ell(x, h) = -\log \pi_h(x)$ for $h \in \mathcal{H}$, then a natural model set is $\mathfrak{S}_{\mathcal{H}} := \{\pi_h : h \in \mathcal{H}\}$.

For many other learning tasks, the choice of the model set $\mathfrak{S}$ results from a tradeoff between several needs, and can primarily be guided by the loss function itself.

On the one hand, results from compressed sensing tell us that given a model set $\mathfrak{S}$ that has proper "low-dimensional" properties, it is possible to choose a small m and design $\mathcal{A}$ such that the LRIP holds, and the ideal decoder Δ has stable recovery guarantees of elements of $\mathfrak{S}$ from their compressed version obtained with $\mathcal{A}$. This calls for the choice of a "small" model set.

On the other hand, and perhaps more importantly, the model set should not be "too small" in order to ensure that the obtained control of the excess risk is nontrivial. Ideally, when the loss is non-negative, the bias term $D(\pi_0, \mathfrak{S})$ in the excess risk (15) should be small when the true optimum risk is small, and even vanish when the true optimum risk vanishes, i.e. when $\mathcal{R}(\pi_0, h^\star) = \inf_{h\in\mathcal{H}} \mathcal{R}(\pi_0, h) = 0$. The "smallest" model with this property is the collection

$$\mathfrak{S}_{\mathcal{H}} := \{\pi : \exists h \in \mathcal{H}, \mathcal{R}(\pi, h) = 0\}, \tag{16}$$

and any model such that $\mathfrak{S} \supset \mathfrak{S}_{\mathcal{H}}$ also has this property.

Finally, given an estimate $\hat{\pi} \in \mathfrak{S}$, obtained either with the ideal decoder (13) or more realistically with a heuristic procedure, we need to select the minimum risk hypothesis according to (8), i.e. to find a minimizer

$$\arg\min_{h\in\mathcal{H}} \mathcal{R}(\hat{\pi}, h). \tag{17}$$

In our examples this procedure is trivial when $\mathfrak{S} = \mathfrak{S}_{\mathcal{H}}$ as h is simply the parameterization of $\hat{\pi} \in \mathfrak{S}_{\mathcal{H}}$.

Examples: The resulting model sets are the following

- **Compressive PCA**: the model set $\mathfrak{S}_{\mathcal{H}}$ consists of all distributions which admit a matrix of second moments of rank at most k. Given any $\hat{\pi} \in \mathfrak{S}_{\mathcal{H}}$, a minimum risk hypothesis is any subspace $\hat{h}$ spanned by eigenvectors associated to the k largest eigenvalues of $\mathbf{\Sigma}_{\hat{\pi}}$.
- **Compressive k-means**: the model set $\mathfrak{S}_{\mathcal{H}}$ consists of all mixtures of k Diracs, possibly with constraints on the Dirac locations. Given any $\hat{\pi} = \sum_{l=1}^{k} \alpha_l \delta_{c_l} \in \mathfrak{S}_{\mathcal{H}}$, a minimum risk hypothesis is $\hat{h} = \{c_1, \dots, c_k\}$.
- **Compressive Gaussian Mixture Modeling** : the model set $\mathfrak{S}_{\mathcal{H}}$ consists of all mixtures π_h of k Gaussians, where the mixture parameters h may further be constrained. Given any $\hat{\pi} = \pi_h \in \mathfrak{S}_{\mathcal{H}}$, a minimum risk hypothesis indeed minimizes the Kullback-Leibler divergence[2] $\min_{h'} \mathrm{KL}(\pi_{h'} || \pi_h)$, hence $\hat{h} = h$. This also holds for more general density models.

For these examples and certain choices of hypothesis class $\mathcal{H}$ we will exhibit in Section 2.5 below and in Sections 3, 4 and 5 a feature function Φ so that $\mathcal{A}$ satisfies the LRIP (12) with $\mathfrak{S} = \mathfrak{S}_{\mathcal{H}}$.

2.6 Choice of a feature function

Compressive learning is deeply connected to kernel mean embeddings of probability distributions, as any feature function Φ (and the related sketching operator $\mathcal{A}$) defines a kernel (i.e. an inner product) between probability distributions:

$$\kappa_{\mathcal{A}}(\pi, \pi') := \langle \mathcal{A}(\pi), \mathcal{A}(\pi') \rangle = \mathbb{E}_{X\sim\pi} \mathbb{E}_{X'\sim\pi'} \kappa_\Phi(X, X')$$

where the explicit kernel between samples is $\kappa_\Phi(x, x') := \langle \Phi(x), \Phi(x') \rangle_{\mathbb{R}^m}$ (or $\langle \Phi(x), \Phi(x') \rangle_{\mathbb{C}^m}$). In fact, any kernel $\kappa(\cdot, \cdot)$ in the sample space is associated to a Mean Map Embedding (a kernel between distributions) [79]. By abuse of notation, we keep the notation κ for both the expression of the kernel in the sample space and of the kernel for probability distributions,

$$\kappa(\pi, \pi') := \mathbb{E}_{X\sim\pi} \mathbb{E}_{X'\sim\pi'} \kappa(X, X'). \tag{18}$$

The associated Maximum Mean Discrepancy (MMD) metric is

$$\|\pi - \pi'\|_\kappa := \sqrt{\kappa(\pi, \pi) - 2\kappa(\pi, \pi') + \kappa(\pi', \pi')}. \tag{19}$$

[2] see Section 5 for reminders on the Kullback-Leibler divergence.

Designing Φ (resp. $\mathcal{A}$) that satisfies the LRIP (12) for a given model set $\mathfrak{S}$ thus amounts to designing a kernel $\kappa(x, x')$ so that the metric $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})}$ is dominated by the metric $\|\pi - \pi'\|_\kappa$ for $\pi, \pi' \in \mathfrak{S}$.

In practice, choosing $\mathcal{A}$ will often amount to choosing a set of (real- or complex-valued) functions $\Phi = \{\phi_\omega\}_{\omega\in\Omega}$ and a probability distribution Λ over a set Ω (often $\Omega = \mathbb{R}^d$) and drawing m independent functions $(\phi_{\omega_j})_{j=1,m}$ from Λ to calculate the feature function[3]:

$$\Phi(x) := \tfrac{1}{\sqrt{m}} \left(\phi_{\omega_j}(x)\right)_{j=1,m} \tag{20}$$

The couple (Φ, Λ) is an *integral representation* of a kernel κ through the relation

$$\kappa(x, x') = \mathbb{E}_{\omega\sim\Lambda} \phi_\omega(x)\overline{\phi_\omega(x')}. \tag{21}$$

Examples:

- **Compressive PCA**: $\phi_{\omega_j}(x) := \langle \mathbf{L}_j, xx^T\rangle_F = x^T\mathbf{L}_j x$, where $\mathbf{L}_j$ is a random matrix in $\mathbb{R}^{d\times d}$ and $\langle \mathbf{A}, \mathbf{B}\rangle_F := \mathrm{Tr}(\mathbf{A}^T\mathbf{B})$ is the Frobenius inner product between matrices. The kernel mean embedding $\kappa(\pi, \pi')$, which is implicitly determined by the distribution of the random matrix $\mathbf{L}$, is a weighted inner product between the matrices $\mathbf{\Sigma}_\pi$ and $\mathbf{\Sigma}_{\pi'}$.

- **Compressive k-means**: $\phi_{\omega_j}(x) := Ce^{\jmath\langle\omega_j, x\rangle}/w(\omega_j)$ is a weighted random Fourier feature, with $\omega_j \in \mathbb{R}^d$, $\jmath$ the imaginary unit and $w(\omega)$ a weighting function. The implicit kernel $\kappa(x, x')$ is determined by the distribution of the random frequency vector ω. It is shift-invariant hence with a standard abuse of notation it can be written as $\kappa(x - x')$. The MMD takes the form $\|\pi - \pi'\|_\kappa = \|\kappa \star \pi - \kappa \star \pi'\|_{L^2(\mathbb{R}^d)}$.

- **Compressive Gaussian Mixture Modeling**: $\phi_{\omega_j}(x) := e^{\jmath\langle\omega_j, x\rangle}$ is a plain random Fourier feature. The implicit kernel κ is again determined by the distribution of the random frequency vector ω, and shift-invariant.

A characterization of the MMD that we will leverage throughout this paper is that for any π, π',

$$\|\pi - \pi'\|_\kappa^2 = \mathbb{E}_{\omega\sim\Lambda} \left|\mathbb{E}_{X\sim\pi}\phi_\omega(X) - \mathbb{E}_{X'\sim\pi'}\phi_\omega(X')\right|^2. \tag{22}$$

Hence $\|\pi - \pi'\|_\kappa^2$ is the expectation with respect to the distribution of the random ω of the quantity

$$\|\mathcal{A}(\pi - \pi')\|_2^2 = \frac{1}{m}\sum_{j=1}^m \left|\mathbb{E}_{X\sim\pi}\phi_{\omega_j}(X) - \mathbb{E}_{X'\sim\pi'}\phi_{\omega_j}(X')\right|^2,$$

and our overall strategy to design a feature function Φ satisfying the LRIP (12) with controlled sketch dimension m will be to:

1. identify an (implicit) kernel κ that satisfies the *Kernel LRIP*

$$\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \le C_\kappa \|\pi - \pi'\|_\kappa, \quad \forall \pi, \pi' \in \mathfrak{S}; \tag{23}$$

2. in the spirit of *random features*, exploit an integral representation of κ to design a random finite-dimensional Φ associated with an explicit kernel κ_Φ approximating κ;

[3]Note the distinct fonts denoting the feature function Φ and the family of functions Φ from which it is built.

3. in the spirit of compressive sensing theory, use concentration of measure and covering arguments to show that for any $0 < \delta < 1$, for large enough m, with high probability on the draw of ω_j,

$$1 - \delta \leq \frac{\|\mathcal{A}(\pi - \pi')\|_2^2}{\|\pi - \pi'\|_\kappa^2} = \frac{\|\pi - \pi'\|_{\kappa_\mathcal{A}}^2}{\|\pi - \pi'\|_\kappa^2} \leq 1 + \delta, \quad \forall \pi, \pi' \in \mathfrak{S} \tag{24}$$

so that the kernel LRIP (23) actually holds with $\kappa_\mathcal{A}$ instead of κ and an adapted constant.

Remark 2.4. *The LRIP* (24) *expresses the control of the* relative error *of approximation* of the MMD*,* restricted *to certain distributions. This contrasts with state of the art results on random features (see e.g. [78, 8] that control uniformly the error $|\kappa_\Phi(\cdot,\cdot) - \kappa(\cdot,\cdot)|$. These two types of controls are indeed of a different nature, and none seems to be a direct consequence of the other.*

2.7 Verifying the Lower Restricted Isometry Property

This strategy will be achieved through the estimation of three quantities: first, a constant C_κ characterizing the compatibility between a kernel, a task, and a model set; second, a constant W_κ characterizing the concentration of $\|\mathcal{A}(\pi - \pi')\|_2^2$ around its expectation; and finally certain covering numbers.

2.7.1 Compatibility between a kernel, a learning task, and a model set

Definition 2.5 (Compatibility constant)**.** *Consider a kernel κ and a learning task defined by the loss functions $\ell(\cdot, h),\ h \in \mathcal{H}$. The* compatibility constant *between this kernel, this task, and the model set $\mathfrak{S}$ is*

$$C_\kappa(\mathcal{L}(\mathcal{H}), \mathfrak{S}) := \sup_{\pi,\pi' \in \mathfrak{S}, \|\pi - \pi'\|_\kappa > 0} \frac{\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})}}{\|\pi - \pi'\|_\kappa}. \tag{25}$$

where we recall that the metric $\|\cdot\|_{\mathcal{L}(\mathcal{H})}$ is defined in (10)*. The compatibility constant measures the suitability of the kernel κ for performing the considered learning task in terms of a Kernel LRIP* (23)*. The case of $\mathfrak{S} = \mathfrak{S}_\mathcal{H}$ is of particular interest, giving the compatibility constant between the kernel and the task, that we will denote for short $C_\kappa = C_\kappa(\mathcal{L}(\mathcal{H}), \mathfrak{S}_\mathcal{H})$.*

Examples:

- **Compressive PCA**: we will show in Section 3 that for the considered kernel, both $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})}$ and $\|\pi - \pi'\|_\kappa$ are indeed norms on the matrix $\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\pi'} \in \mathbb{R}^{d \times d}$. The existence of a finite compatibility constant C_κ will simply follow from the equivalence of all norms in finite dimension, and more explicit bounds will be provided.
- **Compressive k-means** (resp. **Compressive Gaussian Mixture Modeling**): for certain shift-invariant kernels (such as the Gaussian kernel $\kappa(x, x') := \exp(-\|x - x'\|_2^2 / 2\sigma^2)$), we provide finite bounds on the compatibility constant in Section 4 for k-mixtures of Diracs (resp. in Section 5 for k-mixtures of Gaussians) with ε-*separation* assumptions between centroids (resp. between Gaussian means). These assumptions ensure the boundedness of

$$\sup_{\pi,\pi' \in \mathfrak{S}, \|\pi - \pi'\|_\kappa > 0} \frac{\|\pi - \pi'\|_{\mathcal{BL}}}{\|\pi - \pi'\|_\kappa} < \infty \tag{26}$$

where $\mathcal{BL}$ is a class of regular functions (see Definition 6.2 in Section 6 for a precise definition of $\mathcal{BL}$). Up to some fixed rescaling the considered loss functions belong to $\mathcal{BL}$, hence $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \leq C \|\pi - \pi'\|_{\mathcal{BL}}$ which yields the desired bound.

2.7.2 Concentration of the empirical kernel to its expectation on the model

Classical arguments from compressive sensing [11, 42, 72, 36, 50] prove that certain random linear operators satisfy the RIP by relying on pointwise concentration inequalities. Similarly, a first step to establish that the inequalities (24) hold with high probability consists in assuming that for any $\pi, \pi' \in \mathfrak{S}$, $t > 0$, and $m \geq 1$

$$\mathbb{P}\left(\left|\frac{\|\mathcal{A}(\pi - \pi')\|_2^2}{\|\pi - \pi'\|_\kappa^2} - 1\right| \geq t\right) \leq 2\exp\left(-\frac{m}{c_\kappa(t)}\right) \tag{27}$$

for some *concentration function* $t \mapsto c_\kappa(t)$ that should ideally be as small as possible. The concrete estimates we will get exploit a Lemma based on Bernstein's inequality that we prove in Annex C.

Lemma 2.6. *Let κ be a kernel with integral representation (Φ, Λ) where we recall that Λ is a distribution over the random variable $\omega \in \Omega$. Consider m parameters $(\omega_j)_{j=1}^m$ drawn i.i.d. according to Λ and the feature function*

$$\Phi(x) := \tfrac{1}{\sqrt{m}}\left[\phi_{\omega_j}(x)\right]_{j=1}^m. \tag{28}$$

Consider π, π' such that $\|\pi - \pi'\|_\kappa > 0$ and $W = W(\pi - \pi') := \frac{\|\pi - \pi'\|_\Phi}{\|\pi - \pi'\|_\kappa} < \infty$. For any $t > 0$ we have

$$\mathbb{P}\left(\left|\frac{\|\mathcal{A}(\pi - \pi')\|_2^2}{\|\pi - \pi'\|_\kappa^2} - 1\right| \geq t\right) \leq 2\exp\left(-\frac{mt^2}{2W^2 \cdot (1 + t/3)}\right). \tag{29}$$

This suggests the following definition.

Definition 2.7 (Concentration constant)**.** *The* concentration constant W_κ *of the integral representation (Φ, Λ) of the kernel κ with respect to the model $\mathfrak{S}$ is defined by*

$$W_\kappa := \sup_{\pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\|_\kappa > 0} \frac{\|\pi - \pi'\|_\Phi}{\|\pi - \pi'\|_\kappa}. \tag{30}$$

By Lemma 2.6 this constant gives a bound on the best possible concentration function*: for any $t > 0$*

$$c_\kappa(t) \leq W_\kappa^2 \cdot \frac{2(1 + t/3)}{t^2}. \tag{31}$$

The reader may have noticed the similarity between the definition of the concentration constant W_κ in (30) and of the compatibility constant C_κ in (25). To bound the concentration constant W_κ for **Compressive k-means** and **Compressive Gaussian Mixture Modeling**, we will indeed reuse the bound (26) with a well-chosen integral representation (Φ, Λ) of the kernel such that $\Phi \subset \mathcal{BL}$ (see Definition 6.2 in Section 6 for the definition of $\mathcal{BL}$).

2.7.3 Measuring the model "size" through coverings of its normalized secant set

Finally, one can extrapolate pointwise concentration (27) to all pairs $\pi, \pi' \in \mathfrak{S}$ using covering numbers of the so-called *normalized secant set* of the model $\mathfrak{S}$ with an appropriate metric (see, e.g., [36, 72]).

Definition 2.8 (Normalized secant set; covering number; yet another metric)**.**

- *The* normalized secant set *of the model set $\mathfrak{S}$ with respect to a seminorm $\|\cdot\|$ is the following subset of the set of finite signed measures (see Annex A.2)*

$$\mathcal{S}_{\|\cdot\|} := \left\{\frac{\pi - \pi'}{\|\pi - \pi'\|} : \pi, \pi' \in \mathfrak{S}, \|\pi - \pi'\| > 0\right\}. \tag{32}$$

- *The* covering number $\mathcal{N}\left(d(\cdot,\cdot), S, \delta\right)$ *of a set* S *with respect to to a (pseudo)metric*[4] $d(\cdot,\cdot)$ *is the minimum number of closed balls of radius* δ *with respect to* $d(\cdot,\cdot)$ *with centers in* S *needed to cover* S*. The set* S *has finite upper box-counting dimension smaller than* s *if*

$$\liminf_{\delta \to 0} \frac{\log \mathcal{N}\left(d(\cdot,\cdot), S, \delta\right)}{\log 1/\delta} \leq s.$$

  This holds as soon as the covering numbers are bounded by $(C/\delta)^s$ *for small enough* δ*.*

- *We focus on covering numbers of the normalized secant set* $\mathcal{S}_{\|\cdot\|_\kappa}$ *with respect to the following pseudometric*[5]

$$d_\Phi(\pi, \pi') := \sup_{\omega \in \Omega} \left| |\mathbb{E}_{X \sim \pi} \phi_\omega(X)|^2 - |\mathbb{E}_{X' \sim \pi'} \phi_\omega(X')|^2 \right|. \tag{33}$$

  where (Φ, Λ) *is an integral representation of the kernel* κ*.*

Examples: In our examples, we obtain the following covering results with the relevant metrics:

- **Compressive PCA**: The normalized secant set associated with the set of matrices with rank lower than k has upper-box counting dimension $s = \mathcal{O}(kd)$.
- **Compressive k-means**: The normalized secant set associated with the set of mixtures of k ε-separated Diracs in a bounded domain has upper-box counting dimension $s = \mathcal{O}(kd)$.
- **Compressive Gaussian Mixture Modeling**: The normalized secant set associated with the set of mixtures of k Gaussians with the same covariance and 2ε-separated means in a bounded domain has upper-box counting dimension $s = \mathcal{O}(kd)$.

2.8 Compressive statistical learning guarantees

Even though the above ingredients may look quite abstract at this stage, we will turn them into concrete estimates on several examples. Let us first see how they can be combined to yield *compressive statistical learning guarantees*. The proof of the following theorem is in Annex C.

Theorem 2.9. *Consider a kernel* $\kappa(x, x')$ *with integral representation* (Φ, Λ) *and finite compatibility constant with a model set* $\mathfrak{S}$*,* $C_\kappa < \infty$*, and* $0 < \delta < 1$ *such that:*

- *the concentration function is finite,* $c_\kappa(\delta/2) < \infty$*;*
- *the normalized secant of the model set* $\mathfrak{S}$ *has finite covering number* $\mathcal{N}\left(d_\Phi, \mathcal{S}_{\|\cdot\|_\kappa}, \delta/2\right) < \infty$*.*

Fix any probability level $0 < \zeta < 1$ *and a sketch size such that*

$$m \geq c_\kappa(\delta/2) \cdot \log\left(2\mathcal{N}\left(d_\Phi, \mathcal{S}_{\|\cdot\|_\kappa}, \delta/2\right)/\zeta\right) \tag{34}$$

Draw ω_j*,* $1 \leq j \leq m$*, i.i.d. from the distribution* Λ *and define the feature function*

$$\Phi(x) := \tfrac{1}{\sqrt{m}} \left(\phi_{\omega_j}(x)\right)_{j=1,m} \tag{35}$$

Then, with probability at least $1 - \zeta$ *on the draw of* $(\omega_j)_{j=1}^m$*, the induced sketching operator* $\mathcal{A}$ *satisfies the LRIP* (12) *with constant* $C_\mathcal{A} = \frac{C_\kappa}{\sqrt{1-\delta}}$*.*

[4] Further reminders on metrics, pseudometrics, and covering numbers are given in Annex A.

[5] Or rather with respect to the extension of d_Φ to finite signed measures, see Annex A.2.

In turn, assume that $\mathcal{A}$ satisfies the LRIP (12) *with constant $C_{\mathcal{A}} = \frac{C_\kappa}{\sqrt{1-\delta}}$. Consider* any *probability distribution π_0 and* any training collection $\mathbf{X} = \{x_i\}_{i=1}^n \in \mathcal{Z}^n$ *(possibly drawn i.i.d. from π_0 but not necessarily), and denote $\hat{\pi}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Consider*

$$\mathbf{y} := \text{Sketch}(\mathbf{X}) = \mathcal{A}(\hat{\pi}_n) \tag{36}$$

$$\hat{\pi} \in \arg\min_{\pi \in \mathfrak{S}} \|\mathcal{A}(\pi) - \mathbf{y}\|_2^2 \tag{37}$$

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \mathcal{R}(\hat{\pi}, h). \tag{38}$$

We have the guarantee

$$\mathcal{R}(\pi_0, \hat{h}) - \mathcal{R}(\pi_0, h^\star) \leq \eta_n := 2D(\pi_0, \mathfrak{S}) + 4C_{\mathcal{A}} \cdot \|\mathcal{A}(\pi_0 - \hat{\pi}_n)\|_2 \tag{39}$$

where $D(\pi_0, \mathfrak{S}) := \inf_{\sigma \in \mathfrak{S}} \left\{ \|\pi_0 - \sigma\|_{\mathcal{L}(\mathcal{H})} + 2C_{\mathcal{A}} \|\mathcal{A}(\pi_0 - \sigma)\|_2 \right\}$.

Discussion :

- As, we will see in our examples, a concentration function with good behaviour is primarily obtained when the concentration constant W_κ is small enough.
- Computing the sketch (36) is highly parallelizable and distributable. Multiple sketches can be easily aggregated and updated as new data become available.
- As discussed in Remark 2.1, while (37) may appear as a general nonparametric density estimation problem, in all our examples it is indeed a nonlinear parametric least-squares fitting problem when the model set is $\mathfrak{S} = \mathfrak{S}_{\mathcal{H}}$, and the existence of the minimizer follows in practice from compactness arguments.
  - For **Compressive PCA** it is a low-rank matrix reconstruction problem. Provably good algorithms to estimate its solution have been widely studied.
  - For **Compressive k-means** and **Compressive Gaussian Mixture Modeling**, the problem has been empirically addressed with success through the CL-OMPR algorithm [62, 63]. Algorithmic success guarantees are an interesting challenge. This is however beyond the scope of this paper.
- In the Examples of Section 2.5, solving the minimization problem (38) is trivial when $\mathfrak{S} = \mathfrak{S}_{\mathcal{H}}$.
- The first term in the bound (39) of the excess risk, η_n, is the empirical estimation error $\|\mathcal{A}(\pi_0 - \hat{\pi}_n)\|_2$. It is easy to bound it as $\mathcal{O}(1/\sqrt{n})$, this will be done explicitly for the considered examples.

2.9 Controlling the bias term $D(\pi_0, \mathfrak{S}_{\mathcal{H}})$ for certain learning tasks

The second term in (39) is the distance to the model set $\mathfrak{S}$, $D(\pi_0, \mathfrak{S})$. The particular model set $\mathfrak{S} = \mathfrak{S}_{\mathcal{H}}$ was designed so that this *bias term* vanishes when $\pi_0 \in \mathfrak{S}$. For certain learning tasks such as Compressive Clustering, we can further bound the bias term $D(\pi_0, \mathfrak{S}_{\mathcal{H}})$ defined in (14) with an increasing function of the true minimum risk, $\mathcal{R}(\pi_0, h^*)$. These recovery guarantees provide *distribution free* excess risk guarantees. Whether this holds for other learning tasks, or even generically, is a challenging question left to further work.

Lemma 2.10. *Assume that $(\mathcal{Z}, d)$ is a separable metric space and consider a loss that can be written as $\ell(x, h) = d^p(x, \mathcal{E}_h)$ where $1 \leq p < \infty$ and $\mathcal{E}_h \subset \mathcal{Z}$ for each $h \in \mathcal{H}$. Assume further that $\Phi : (\mathcal{Z}, d) \to (\mathbb{R}^m, \|\cdot\|_2)$ (or $(\mathbb{C}^m, \|\cdot\|_2)$) is L-Lipschitz. Then for any value of $C_{\mathcal{A}}$ the distance defined by* (14) *satisfies*

- *case $p = 1$: for any π_0*
$$D(\pi_0, \mathfrak{S}_{\mathcal{H}}) \leq (1 + 2L \cdot C_{\mathcal{A}}) \cdot \inf_{h \in \mathcal{H}} \mathcal{R}(\pi_0, h) \tag{40}$$

- *case $p > 1$: denoting $B := \sup_{x \in \mathcal{Z}, h \in \mathcal{H}} d(x, V_h)$ we have for any π_0*
$$D(\pi_0, \mathfrak{S}_{\mathcal{H}}) \leq (p \cdot B^{p-1} + 2L \cdot C_{\mathcal{A}}) \cdot \inf_{h \in \mathcal{H}} \mathcal{R}^{1/p}(\pi_0, h) \tag{41}$$

The proof (in Annex D) exploits optimal transport through connections between the considered norms and the norm $\|\pi - \pi'\|_{\mathrm{Lip}(L,d)} = L \cdot \|\pi - \pi'\|_{\mathrm{Lip}(1,d)}$, where $\mathrm{Lip}(L, d)$ denotes the class of functions $f : (\mathcal{Z}, d) \to \mathbb{R}$ that are L-Lipschitz.

2.10 Summary

Given a learning task, embodied by a family of loss functions $\ell(x, h)$, $h \in \mathcal{H}$, and a (random) feature function Φ used to define a sketching procedure, establishing compressive statistical learning guarantees involves several steps. Overall, to determine whether the task and the sketching function are compatible one needs

1. to determine a model set $\mathfrak{S}_{\mathcal{H}}$ associated to the learning task, and to identify the metric $\|\cdot\|_{\mathcal{L}(\mathcal{H})}$;
2. to identify the kernel $\kappa(x, x')$ associated to the (random) feature function, and the MMD $\|\cdot\|_{\kappa}$;
3. to check whether the Kernel LRIP (23) holds, through the characterization of a *compatibility constant* between the kernel, the learning task, and the model set $\mathfrak{S}_{\mathcal{H}}$.
4. to characterize the concentration (27) of the empirical MMD toward the MMD $\|\cdot\|_{\kappa}$ for distributions in the model set $\mathfrak{S}_{\mathcal{H}}$;
5. to control a certain covering dimension of the normalized secant set $\mathcal{S}$ of $\mathfrak{S}_{\mathcal{H}}$ with respect to the MMD.

Gathering all these steps, one can prove that the sketching operator $\mathcal{A}$ associated to Φ satisfies a LRIP (12) with high probability for a sketch of controlled finite size, and that the solution of a certain nonlinear least squares problem (37) yields an hypothesis (38) with controlled excess risk. These steps, and the resulting guarantees are summarized in Table 1 for the three examples developed in the next sections.

Conversely, to construct a suitable random feature function given *only* a learning task, it will be relevant to first find an appropriate kernel and to approximate it by random feature sampling, see Section 7 for a discussion of perspectives in this direction.

3 A first illustration: Compressive PCA

As a first simple illustration, this general compressive statistical framework can be applied to the example of PCA, where most of the tools already exist. Our aim is essentially illustrative, and focuses on controlling the excess risk, rather than to compare the results with state-of-the art PCA techniques.

Task	PCA	k-med. ($p=1$) / k-means ($p=2$)	Gaussian Mixture Model.
Hypothesis h	k-dim. subspace $h \subset \mathbb{R}^d$, $\dim h = k$	k cluster centers $\{c_1, \dots, c_k\} \subset \mathbb{R}^d$	param. of k Gaussians means $c_l \in \mathbb{R}^d$ covar. $\mathbf{\Sigma}_l \in \mathbb{R}^{d\times d}$ mixture parameters α_l
Loss $\ell(x,h)$	$\|x - \mathrm{P}_h x\|_2^2$	$\min_{1\le l\le k} \|x - c_l\|_2^p$	$-\log \pi_h(x)$
Model set $\mathfrak{S}_\mathcal{H}$	$\{\pi : \mathrm{rank}(\mathbf{\Sigma}_\pi) \le k\}$	$\{\pi : \text{mixt. of } k \text{ Diracs}\}$	$\{\pi : \text{mixt. of } k \text{ Gaussians}\}$
Feature function $\Phi(x)$	quadratic polyn. $\left(x^T \mathbf{L}_j x\right)_{j=1}^m$	weighted Fourier features $\left(e^{\mathrm{J}\omega_j^T x}/w(\omega_j)\right)_{j=1}^m$	Fourier features $\left(e^{\mathrm{J}\omega_j^T x}\right)_{j=1}^m$
Sampling law Λ	$\mathbb{P}(\mathbf{L}) \propto e^{-\|\mathbf{L}\|_F^2}$	$\mathbb{P}(\omega) \propto w^2(\omega) e^{-\frac{\|\omega\|^2}{2\lambda^2}}$	$\mathbb{P}(\omega) \propto e^{-\frac{\|\omega\|_{\mathbf{\Sigma}}^2}{2\lambda^2}}$
Kernel $\kappa(x,x')$	$\left\|xx^T - x'x'^T\right\|_F^2$	$\exp(-\lambda^2 \|x-x'\|_2^2/2)$	$\exp\left(-\lambda^2 \|x-x'\|_{\mathbf{\Sigma}}^2/2\right)$
Learning step $\hat{h} = \mathrm{Learn}(\mathbf{y})$	Low-rank recovery	$\arg\min_{h\in\mathcal{H}} \min_{\alpha\in\mathbb{S}_{k-1}} \|\sum_{l=1}^k \alpha_l \Phi(c_l) - \mathbf{y}\|_2$	$\arg\min_{h\in\mathcal{H}} \|\mathcal{A}(\pi_h) - \mathbf{y}\|_2$
Assumptions	N/A	$\min_{l\ne l'} \|c_l - c_{l'}\|_2 \ge 2\varepsilon$ $\max_l \|c_l\|_2 \le R$ $1/\varepsilon = \mathcal{O}\left(\lambda/\sqrt{\log k}\right)$	$\min_{l\ne l'} \|c_l - c_{l'}\|_{\mathbf{\Sigma}} \ge 2\varepsilon_\lambda$ $\max_l \|c_l\|_{\mathbf{\Sigma}} \le R$ known covariance $\mathbf{\Sigma}_l = \mathbf{\Sigma}, \forall l$ [see Table 5.2 for expr. of ε_λ]
Compat. C_κ Concent. $c_\kappa(\cdot)$ Covering dim. s	$\mathcal{O}\left(\sqrt{k}\right)$ $\mathcal{O}(1)$ $\mathcal{O}(kd)$	$\mathcal{O}\left(\sqrt{k}R^p\right)$ $\mathcal{O}(kd\log k)$ $\mathcal{O}(kd)$	$\mathcal{O}\left(\sqrt{k}R^2\right)$ $\mathcal{O}(kM_\lambda)$ $\mathcal{O}(kd)$
Sketch size $m \gtrsim c_\kappa \cdot s$	$\mathcal{O}(kd)$	$\mathcal{O}\left(k^2 d^2 \log k \cdot \log(kdR/\varepsilon)\right)$	$\mathcal{O}\left(k^2 dM_\lambda \cdot \log(kM_\lambda R/\varepsilon_\lambda)\right)$ [see Table 5.2 for expr. of M_λ]
Bias term $D(\pi_0, \mathfrak{S}_\mathcal{H})$	$\mathcal{O}\left(\sqrt{k}\right) \cdot \mathcal{R}_{\mathtt{PCA}}(\pi_0, h^*)$	$\mathcal{O}\left(\sqrt{k\log k}R^p/\varepsilon\right) \cdot \mathcal{R}_{\mathtt{clust.}}^{1/p}(\pi_0, h^\star)$ [$p=2$: assuming $\|X\|_2 \le R$ a.s.]	N/A

Table 1: Summary of the application of the framework on our three main examples (detailed in Sections 3, 4 and 5) in $\mathcal{Z} = \mathbb{R}^d$. $\mathbb{S}_{k-1}$ denotes the $(k-1)$-dimensional simplex (i.e. the sphere with respect to the ℓ^1-norm in the non-negative orthant of $\mathbb{R}^k$), and $\|x\|_{\mathbf{\Sigma}} = x^T \mathbf{\Sigma}^{-1} x$ the Mahalanobis norm associated to the positive definite covariance matrix $\mathbf{\Sigma}$. Stricly speaking we should write $\log(ek) = 1 + \log k$ instead of $\log k$ to cover the case $k = 1$. For PCA, improved bounds can be obtained with specialized arguments, see Section 3. This suggests that improved constants may be achievable also for the other considered tasks.

Definition of the learning task. The risk associated to the PCA learning problem is defined[6] as $\mathcal{R}_{\mathtt{PCA}}(\pi, h) = \mathbb{E}_{X\sim\pi} \|X - \mathrm{P}_h X\|_2^2$. It is minimized by the subspace associated with the k largest eigenvalues of the matrix $\mathbf{\Sigma}_\pi = \mathbb{E}_{X\sim\pi} XX^T$.

[6]for simplicity we assume centered distributions $\mathbb{E}_{X\sim\pi} X = 0$ and don't empirically recenter the data.

It is well established [50] that matrices that are approximately low rank can be estimated from partial linear observations under a certain Restricted Isometry Property (RIP). This leads to the following natural way to perform Compressive PCA.

Choice of feature function. Choose (at random) a linear operator $\mathcal{M} : \mathbb{R}^{d\times d} \to \mathbb{R}^m$ satisfying (with high probability) the following RIP on low-rank matrices: for any $\mathbf{M} \in \mathbb{R}^{d\times d}$ of rank at most $2k$,

$$1 - \delta \leq \frac{\|\mathcal{M}(\mathbf{M})\|_2^2}{\|\mathbf{M}\|_F^2} \leq 1 + \delta \tag{42}$$

with $\|\cdot\|_F$ the Frobenius norm and $\delta < 1$. This is feasible with $m = \mathcal{O}(kd)$, see e.g. [50]. Define the feature function $\Phi : \mathcal{Z} = \mathbb{R}^d \to \mathbb{R}^m$ by $\Phi(x) := \mathcal{M}(xx^T)$.

Sketch computation. Given sample points $x_1, \ldots, x_n$ in $\mathbb{R}^d$, compute the sketch $\mathbf{y}$ as in (4), i.e., compute empirical estimates of random second moments of the distribution π_0 of X. These random moments are well-defined provided that π_0 admits second moments, i.e., that it is not too heavy-tailed.

Learning from a sketch. Given a sketch vector $\mathbf{y}$, estimate a solution of the optimization problem over semi-definite positive symmetric matrices ($\mathbf{\Sigma} \succcurlyeq 0$)

$$\hat{\mathbf{\Sigma}} := \arg \min_{\mathrm{rank}(\mathbf{\Sigma})\leq k, \mathbf{\Sigma}\succcurlyeq 0} \|\mathcal{M}(\mathbf{\Sigma}) - \mathbf{y}\|_2^2 . \tag{43}$$

This step estimates the rank-k matrix whose sketch best matches the sketch of the empirical matrix of second moments, in the least squares sense. Compute the eigen-decomposition $\hat{\mathbf{\Sigma}} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ and output

$$\hat{h} := \mathrm{span}(\mathbf{U}(:, 1:k)). \tag{44}$$

In Annex H we control the excess risk of PCA through the characterization of $\|\pi' - \pi\|_{\mathcal{L}(\mathcal{H})}$ and the proof (cf Eq. (199)) that $\|\pi' - \pi\|_{\mathcal{L}(\mathcal{H})} \leq \|\mathbf{\Sigma}_{\pi'} - \mathbf{\Sigma}_\pi\|_\star$ with $\|\cdot\|_\star$ the nuclear norm.

Theorem 3.1. *Consider any probability distribution π_0 with finite second moments and any draw of x_i, $1 \leq i \leq n$ (represented by the empirical distribution $\hat{\pi}_n$). Applying the above approach yields*

$$\mathcal{R}_{PCA}(\pi_0, \hat{h}) - \mathcal{R}_{PCA}(\pi_0, h^\star) \leq \eta_n := C_1 \mathcal{R}_{PCA}(\pi_0, h^\star) + C_2 \|\mathcal{A}(\pi_0 - \hat{\pi}_n)\|_2 . \tag{45}$$

where $C_1 = 2 + \frac{4\sqrt{2k}\sqrt{1+\delta}}{\sqrt{1-\delta}}$ and $C_2 = \frac{4\sqrt{2k}}{\sqrt{1-\delta}}$.

Discussion :

- **Bias term.** The first term in the right hand side of (45) is a bias term that vanishes when the true risk is low. Remarkably, it is also proportional to the true risk, hence leading to a (non-sharp) oracle inequality $\mathcal{R}_{\mathtt{PCA}}(\pi_0, \hat{h}) \leq (1 + C_1)\mathcal{R}_{\mathtt{PCA}}(\pi_0, h^\star) + C_2 \|\mathcal{A}(\pi_0 - \hat{\pi}_n)\|_2$. We will show that this remarkable property also holds for Compressive k-medians (a variant of k-means). For Compressive k-means we will prove similar properties where the bias term is essentially the square root of the true risk.

- **Sample complexity.** Regarding the second term, if we further assume that the support of π_0 is contained in a Euclidean ball of radius R, then by the RIP (42) we have a.s. $\|\Phi(x)\|_2 \leq \sqrt{1+\delta}\cdot R^2$ hence, by the vectorial Hoeffding's inequality [74], we obtain with high probability w.r.t. data sampling that $C_2 \|\mathcal{A}(\pi_0 - \hat{\pi}_n)\|_2 \lesssim R^2\sqrt{k(1+\delta)/(1-\delta)}/\sqrt{n}$.

Improved guarantees and practical algorithms for learning. One can consider several relaxations of the nonconvex optimization problem (43) in order to perform compressive PCA. Beside convex relaxations using the minimization of the nuclear norm [50, Section 4.6], Kabanava et al. [61] showed (in a complex setting) that the rank constraint in (43) can be relaxed when $\mathcal{M}$ is made of random rank one projections, i.e. when $\Phi(x) = \frac{1}{\sqrt{m}}(|\langle a_j, x\rangle|^2)_{j=1,m}$ where $a_j \in \mathbb{C}^d$ are independent standard complex Gaussian vectors. In this setting, let

$$\hat{\mathbf{\Sigma}} := \arg\min_{\mathbf{\Sigma}\succcurlyeq 0} \|\mathcal{M}(\mathbf{\Sigma}) - \mathbf{y}\|_2^2, \tag{46}$$

and the corresponding hypothesis $\hat{h}$ obtained through (44). We have the following result ([61, Theorem 4 with $p = 1$] combined with Equation (199) in Section H): if $m \geq Ckd$ where C is a universal constant, then with high probability on the draw of the a_j, for any $x_1, \ldots, x_n$, we have the control

$$\eta_n := D_1 \mathcal{R}_{\texttt{PCA}}(\pi, h^\star) + D_2 \|\mathcal{A}(\pi - \hat{\pi}_n)\|_2$$

where D_1, D_2 are positive universal constants that no longer depend on k.

Hence, the constant C_1 from Theorem 3.1 seems pessimistic as it grows with $\sqrt{k}$. This may be due to the generality of our approach, where we lose a factor (by using a RIP in the Frobenius norm) compared to the more precise study of [61]. However, the general approach developed here permits the study of the less trivial setting of compressive clustering and compressive Gaussian mixture estimation as shown in the next sections.

4 Compressive clustering

We consider here two losses that measure clustering performance: the k-**means** and k-**medians** losses.

4.1 Application of the theoretical framework

Definition of the learning task. For k-means (resp. k-medians), we consider as sample space the Euclidean space $\mathcal{Z} = \mathbb{R}^d$, and hypotheses are sets of size lower than k : $h = \{c_1, \ldots, c_{k_1}\}$ where $c_l \in \mathbb{R}^d$ are the so-called centers of clusters and $k_1 \leq k$. The loss function for the clustering task is

$$\ell(x, h) := \min_{1\leq l\leq k} \|x - c_l\|_2^p \tag{47}$$

with $p = 2$ for k-means (resp. $p = 1$ for k-medians) and $\mathcal{R}_{\texttt{clust.}}(\pi, h) = \mathbb{E}_{X\sim\pi} \min_{1\leq l\leq k} \|X - c_l\|_2^p$.

Model set $\mathfrak{S}_\mathcal{H}$ and best hypothesis for $\pi \in \mathfrak{S}_\mathcal{H}$. For compressive clustering with k-means or k-medians with a hypothesis class $\mathcal{H} \subset (\mathbb{R}^d)^k$, distributions such that $\mathcal{R}_{\texttt{clust.}}(\pi, h^*) = 0$ are precisely mixtures of k Diracs,

$$\mathfrak{S}_\mathcal{H} = \left\{ \sum_{l=1}^k \alpha_l \delta_{c_l} : \ \{c_l\}_{l=1}^k \in \mathcal{H}, \alpha \in \mathbb{S}_{k-1} \right\} \tag{48}$$

where $\mathbb{S}_{k-1} := \left\{\alpha \in \mathbb{R}^k : \ \alpha_l \geq 0, \sum_{l=1}^k \alpha_l = 1\right\}$ denotes the $(k-1)$-dimensional simplex. Moreover, for any distribution in this model set, $\hat{\pi} = \sum_{l=1}^k \alpha_l \delta_{c_l} \in \mathfrak{S}_\mathcal{H}$, the optimum of minimization (17) is $\hat{h} = \{c_1, \ldots, c_k\}$ (hypothesis resulting from the probability density $\hat{\pi}$) .

Separation assumption. Because one can show (see Section F.8) it is a necessary condition for the derivation of theoretical risk control with smooth shift invariance kernels, we impose a minimum separation between Diracs, as well as a bounded domain, hence we consider a particular set of k-tuples

$$\mathcal{H}_{k,\varepsilon,R} := \left\{ \{c_l\}_{l=1}^{k_1} \subset \mathbb{R}^d : \quad k_1 \leq k, \quad \min_{l \neq l'} \|c_l - c_{l'}\|_2 \geq \varepsilon, \quad \max_l \|c_l\|_2 \leq R \right\}. \tag{49}$$

The parameters ε and R represent the resolution at which we cluster the data.

Choice of feature function: weighted random Fourier features. Given that the model set $\mathfrak{S}_{\mathcal{H}}$ consists of mixture of Diracs, and by analogy with compressive sensing where random Fourier sensing yields RIP guarantees, compressive clustering with random Fourier moments has been recently proposed [64]. To establish our theoretical guarantees we rely on a reweighted version: sample m random frequencies $\omega_1, \ldots, \omega_m$ in $\mathbb{R}^d$ independently from the distribution with density

$$\Lambda(\omega) = \Lambda_{w,\lambda}(\omega) \propto w^2(\omega) e^{-\frac{\|\omega\|^2}{2\lambda^2}}, \tag{50}$$

with weights $w(\omega)$ and scale parameter λ, and define the associated feature function $\Phi : \mathbb{R}^d \to \mathbb{C}^m$:

$$\Phi(x) := \frac{C_\Lambda}{\sqrt{m}} \left[\frac{e^{\jmath \omega_j^T x}}{w(\omega_j)} \right]_{j=1,\ldots,m}. \tag{51}$$

with

$$w(\omega) := 1 + \frac{\|\omega\|_2^2}{\lambda^2 d} \tag{52}$$

$$C_\Lambda := \left(4 + 2/d^2\right)^{1/2} \tag{53}$$

This sketching operator is based on a reweighting of Random Fourier Features $x \mapsto e^{\jmath \omega^T x}$ [73]. These weights $w(\omega)$ are mainly required for technical reasons (see general proof strategy in Section 6 and proofs in Annex F) but may be an artefact of our proof technique. The constant scaling in front of Φ is of course irrelevant for the algorithm itself which is invariant by rescaling of Φ, but is included for coherence with the theory and in particular the bounds below which involve Φ.

Sketch computation. Given sample points $x_1, \ldots, x_n$ in $\mathbb{R}^d$, compute the sketch $\mathbf{y}$ as in (4), i.e., compute a weighted version of samples of the conjugate of the empirical characteristic function [49] of the distribution π_0 of X. In contrast to the case of Compressive PCA, the characteristic function and its empirical counterpart are always well-defined, even if π_0 is very heavy-tailed.

Learning from a sketch. Given a sketch vector $\mathbf{y}$ and a class of hypotheses $\mathcal{H}$, estimate a solution of the following nonlinear least-squares problem

$$\hat{h} = \{\hat{c}_1, \ldots, \hat{c}_{k_1}\} := \arg \min_{\{c_1,\ldots,c_{k'}\} \in \mathcal{H}} \min_{\alpha \in \mathbb{S}_{k-1}} \left\| \sum_{l=1}^{k'} \alpha_l \Phi(c_l) - \mathbf{y} \right\|_2^2, \tag{54}$$

This learn-from-sketch part finds the k-mixture of Diracs (under R-boundedness and centroid 2ε-separation constraints) whose sketch best matches the empirical sketch (in the least squares sense): it corresponds exactly to the calculation of the minimizers of (37) and (38) in Theorem 2.9. We have the following guarantees.

Theorem 4.1. *Build the sketching function as in* (51) *where the* ω_j *are drawn according to* (50). *Define* $\varepsilon := 1/(\lambda\sigma_k)$ *where for* $k \geq 1$, $\sigma_k := (2.4(\log(2k-1)+10))^{-1/2}$, *and consider* $R > \varepsilon$ *large enough so that* $\mathcal{H}_{k,2\varepsilon,R}$ *is non-empty.*

There is a universal constant $C > 0$ *such that, for any* $\zeta, \delta \in (0,1)$, *when the sketch size* m *satisfies*

$$m \geq C\delta^{-2}kd(1+\log k)\cdot\left[kd\cdot\left(1+\log kd+\log\tfrac{R}{\varepsilon}+\log\tfrac{1}{\delta}\right)+\log\tfrac{1}{\zeta}\right], \tag{55}$$

we have with probability at least $1-\zeta$ *on the draw of the random Fourier frequencies* $(\omega_j)_{j=1}^m$: *for any* $\mathcal{H} \subset \mathcal{H}_{k,2\varepsilon,R}$, *any source distribution* π_0 *on* $\mathcal{Z} = \mathbb{R}^d$, *any samples* $x_i \in \mathbb{R}^d$, $1 \leq i \leq n$ *(represented by the empirical distribution* $\hat{\pi}_n$*), denoting* $\hat{h}$ *a* k*-vector of centroids obtained by minimization* (54) *and* $h^\star \in \arg\min_{h\in\mathcal{H}} \mathcal{R}_{\texttt{clust.}}(\pi_0, h)$:

$$\mathcal{R}_{\texttt{clust.}}(\pi_0, \hat{h}) - \mathcal{R}_{\texttt{clust.}}(\pi_0, h^\star) \leq \eta_n \tag{56}$$

$$\eta_n \leq 2D_{\texttt{clust.}}(\pi_0, \mathfrak{S}_{\mathcal{H}}) + \tfrac{C_\kappa}{\sqrt{1-\delta}}\left\|\mathcal{A}(\pi_0 - \hat{\pi}_n)\right\|_2 \tag{57}$$

where $C_\kappa \leq 8\sqrt{6}\sqrt{k}R^p$, *with* $p = 2$ *for* k*-means, resp.* $p = 1$ *for* k*-medians, and* $D_{\texttt{clust.}}$ *is the instantiation of* (14) *for the norm* (10) *associated to the considered clustering task.*

Remark 4.2. *Note that this holds with the sample space* $\mathcal{Z} = \mathbb{R}^d$, *i.e., we only restrict the centroids,* not the data *to the Euclidean ball of radius* R, $\mathcal{B}_{\mathbb{R}^d,\|\cdot\|_2}(0,R)$.

Given λ and k, Theorem 4.1 sets a separation condition $\varepsilon = 1/(\lambda\sigma_k)$ sufficient to ensure compressive statistical learning guarantees with the proposed sketching procedure. Vice-versa, to target a given separation ε with a given number k of components, choosing $\lambda = 1/(\varepsilon\sigma_k) = \mathcal{O}\left(\sqrt{1+\log k}/\varepsilon\right)$ is sufficient.

Sketch of the proof of Theorem 4.1. We apply Theorem 2.9. The kernel resulting from the choice of random features is

$$\kappa(x,x') = \exp\left(-\tfrac{\lambda^2\|x-x'\|_2^2}{2}\right).$$

The control of the compatibility constant C_κ, the concentration constant W_κ, and the covering numbers $\mathcal{N}(d_\Phi, \mathcal{S}, \delta)$ for the model $\mathfrak{S}_0 := \mathfrak{S}_{\mathcal{H}_{k,2\varepsilon,R}}$ and its secant $\mathcal{S}$ is obtained with a general strategy for the more general case of mixture models with ε-separation that is described in Section 6 and will be reused for Gaussian mixtures. The specifics of this strategy for compressive clustering are proved in Annex F, and yield:

- **Compatibility and concentration.** The compatibility constant is bounded as

$$C_\kappa = \mathcal{O}\left(\sqrt{k}R^p\right), \tag{58}$$

  with $p = 2$ for k-means (resp. $p = 1$ for k-medians). We also control the concentration constant

$$W_\kappa = \mathcal{O}\left(\sqrt{kd\log(ek)}\right), \tag{59}$$

  yielding

$$c_\kappa(\delta/2) = \mathcal{O}\left(\delta^{-2}kd\log(ek)\right). \tag{60}$$

- **Covering numbers.** We establish that the covering numbers satisfy

$$\log\mathcal{N}(d_\Phi, \mathcal{S}, \delta) = \mathcal{O}\left(kd\cdot\log\left(ekd\cdot\tfrac{1}{\delta}\cdot\tfrac{R}{\varepsilon}\right)\right). \tag{61}$$

Given these results, we can invoke Theorem 2.9 to obtain that the sketching operator satisfies with high probability the LRIP with the claimed constants for the model $\mathfrak{S}_0$. Whenever $\mathcal{H} \subset \mathcal{H}_{k,2\varepsilon,R}$, we have $\mathfrak{S}_{\mathcal{H}} \subset \mathfrak{S}_0$ hence the LRIP also holds for all models $\mathfrak{S}_{\mathcal{H}}$ such that $\mathcal{H} \subset \mathcal{H}_{k,2\varepsilon,R}$ and the conclusion follows as in Theorem 2.9. □

4.2 Sample complexity and distribution free result.

The second term in the bound (57) measures the estimation error and can be easily controlled since $\|\Phi(x)\|_2 \leq C_\Lambda \leq \sqrt{6}$ by construction (51). By the vectorial Hoeffding's inequality [74], with high probability w.r.t. data sampling it holds that $\|\mathcal{A}(\pi_0 - \hat{\pi}_n)\|_2 \lesssim C_\Lambda/\sqrt{n} = \mathcal{O}(1/\sqrt{n})$.

The first term in the bound (57) is a bias term $D_{\texttt{clust.}}(\pi_0, \mathfrak{S}_\mathcal{H})$ which means that the quality of the bound depends on whether π_0 is well modeled by $\mathfrak{S}_\mathcal{H}$ or not. However, in the case of Compressive PCA we have shown that this bias term is in fact controlled by the risk of the optimal hypothesis, $\mathcal{R}(\pi_0, h^*)$, yielding a (non-sharp) oracle inequality and a *distribution free* guarantee. Does this also hold for Compressive Clustering ?

Since the loss is a power of a metric $d(x, x') = \|x - x'\|_2$, we can leverage Lemma 2.10 to bound the bias term $D_{\texttt{clust.}}(\pi_0, \mathfrak{S}_\mathcal{H})$ provided $\Phi : (\mathcal{Z}, \|\cdot\|_2) \to (\mathbb{C}^m, \|\cdot\|_2)$ is Lipschitz. This is indeed the case thanks to the following lemma (proof in Annex F.6).

Lemma 4.3. *Build the sketching function as in* (51) *where the* ω_j *are drawn according to* (50)*. For any* $0 < \zeta < 1$*,* $t \geq 0$ *when the sketch size satisfies*

$$m \geq C_\Lambda^2 t^{-2} \left(2d + 4 + \tfrac{d}{6} \cdot t\right) \cdot \log\left(\frac{d}{\zeta}\right) \tag{62}$$

we have: with probability at least $1 - \zeta$ *on the draw of the random Fourier frequencies* $(\omega_j)_{j=1}^m$*, the function* $\Phi : (\mathcal{Z}, \|\cdot\|_2) \to (\mathbb{C}^m, \|\cdot\|_2)$ *is L-Lipschitz with* $L = \lambda\sqrt{1+t}$*.*

Combining with Lemma 2.10 and recalling that $C_\mathcal{A} = \mathcal{O}\left(\sqrt{k}R^p\right)$, we obtain for k-medians ($p = 1$) and any π_0

$$D_{k-\texttt{medians}}(\pi_0, \mathfrak{S}_\mathcal{H}) \leq C_1 \cdot \mathcal{R}_{k-\texttt{medians}}(h)$$

with $C_1 = \mathcal{O}\left(1 + \sqrt{k}\lambda R\right) = \mathcal{O}\left(\sqrt{k \log k} R/\varepsilon\right)$, while for k-means ($p = 2$), under the additional assumption that when $X \sim \pi_0$ we have $\|X\|_2 \leq R$ a.s., we get

$$D_{k-\texttt{means}}(\pi_0, \mathfrak{S}_\mathcal{H}) \leq C_2 \cdot \sqrt{\mathcal{R}_{k-\texttt{means}}(h)}.$$

with $C_2 = \mathcal{O}(RC_1) = \mathcal{O}\left(\sqrt{k \log k} R^2/\varepsilon\right)$.

Remark 4.4. *Given the assumptions of Lemma 2.10 we seem to require a bound B on the samples to control the bias term for k-means (not k-medians).*

Combined with Theorem 4.1, this yields with probability $1 - 2\zeta$ a *uniform, distribution free* guarantee valid for any π_0 (resp. for any π_0 such that X is a.s. bounded by B) provided m is large enough. In particular for k-medians we obtain a (non-sharp) oracle inequality

$$\mathcal{R}(\pi_0, \hat{h}) \lesssim \mathcal{R}(\pi_0, h^\star) + 1/\sqrt{n}.$$

We leave possible tightening of the hidden constants (which may be large) to future work.

4.3 Learning algorithm ?

For compressive clustering, learning in the sketched domain means solving problem (54), which is analogous to the classical finite-dimensional least squares problem under a sparsity constraint. The latter is NP-hard, yet, under RIP conditions, provably good and computationally efficient algorithms (either greedy or based on convex relaxations) have been derived [50].

It was shown practically in [64] (with some reproduced results in Section 4.3) that a heuristic based on orthogonal matching pursuit (which neglects the separation and boundedness constraint associated

to the class $\mathcal{H}_{k,2\varepsilon,R}$) is able to recover sums of Diracs from sketches of the appropriate size. It must be noted that recovering sums of Diracs from Fourier observations has been studied in the case of regular low frequency measurements. In this problem, called super-resolution, it was shown that a convex proxy (convexity in the space of distributions using total variation regularization) for the non-convex optimization (54) is able to recover sufficiently separated Diracs [19, 35, 41]. However, these methods rely on semi-definite relaxation of dual optimization followed by root finding in dimension d and their extension to weighted random Fourier measurements in not straigthforward. Consequently, we leave this direction for future research.

4.4 Discussion

Improved sketch size guarantees? Although Theorem 4.1 only provides guarantees when $m \geq \mathcal{O}\left(k^2 d^2\right)$ (up to logarithmic factors), the observed empirical phase transition pattern [64] strongly hints that $m \geq \mathcal{O}\left(kd\right)$ is in fact sufficient. This is intuitively what one would expect the "dimensionality" of the problem to be, since this is the number of parameters of the model $\mathfrak{S}_{\mathcal{H}}$.

In fact, as the parameters live in the cartesian product of k balls of radius R in $\mathbb{R}^d$ and the "resolution" associated to the separation assumption is ε, a naive approach to address the problem would consist in discretizing the parameter space into $N = \mathcal{O}\left((R/\varepsilon)^d\right)$ bins. Standard intuition from compressive sensing would suggest a sufficient number of measures $m \geq \mathcal{O}\left(k \log N\right) = \mathcal{O}\left(kd \log \frac{R}{\varepsilon}\right)$.

Given that the covering dimension captured by our results seems of the right order, we expect that improved compressed statistical learning guarantees for compressive k-means should primarily result from a more subtle control of the concentration function $c_\kappa(t)$, to ideally obtain $c_\kappa(\delta/2) = \mathcal{O}\left(\delta^{-2}\right)$ instead of $\mathcal{O}\left(\delta^{-2} kd \log k\right)$. We leave a possible refinement of our analysis, trying to capture the empirically observed phase transition, for future work. This may also lead to improved estimates of the compatibility constant removing certain logarithmic dependencies.

Role of separation Although the separation ε is important in the definition of the sketch and in the derivation of learning guarantees, its role is less stringent than it may appear at first sight. In Theorem 4.1, both the estimated and optimal hypotheses $\hat{h}$ and $h^\star$ are defined under a separation (and boundedness) constraint $\mathcal{H}_{k,2\varepsilon,R}$. In fact, if the optimal hypothesis *without separation constraint* (denote it $h_0^\star$) happens to be indeed 2ε-separated, then $h^\star = h_0^\star$ and Theorem 4.1 does provide guarantees with respect to the corresponding risk $\mathcal{R}_{\texttt{clust.}}(\pi_0, h_0^\star)$. Besides, we can show that the separation hypothesis causes a maximum deviation 2ε in the risk estimation. We have the following Lemma:

Lemma 4.5. *Let* $h_0^\star \in \arg\min_{h \in \mathcal{H}_{k,0,R}} \mathcal{R}_{\texttt{clust.}}(\pi_0, h)$ *be an optimal hypothesis* without *separation constraint, and* $h^\star \in \arg\min_{h \in \mathcal{H}_{k,2\varepsilon,R}} \mathcal{R}_{\texttt{clust.}}(\pi_0, h)$ *an optimal hypothesis* with *the separation constraint.*

- *For k-medians* ($p = 1$) *we have:*

$$\mathcal{R}_{k-\texttt{medians}}(\pi_0, h^\star) \leq \mathcal{R}_{k-\texttt{medians}}(\pi, h_0^\star) + 2\varepsilon. \tag{63}$$

- *For k-means* ($p = 2$) *we have:*

$$\sqrt{\mathcal{R}_{k-\texttt{means}}(\pi_0, h^\star)} \leq \sqrt{\mathcal{R}_{k-\texttt{means}}(\pi_0, h_0^\star)} + 2\varepsilon. \tag{64}$$

This Lemma, which is proved in a slightly more general version (Lemma F.5) in Section F.7, allows to compare the risk of the separation-constrained estimation $\hat{h}$ (with method (54)) and the optimal risk $h_0^\star$ without separation, e.g. in the k-medians case,

$$\mathcal{R}_{\texttt{clust.}}(\pi_0, \hat{h}) - \mathcal{R}_{\texttt{clust.}}(\pi_0, h_0^\star) \leq \eta_n + 2\varepsilon. \tag{65}$$

Whether one can similarly relate the solutions $\hat{h}$ and $\hat{h}_0$ of (54) with and without separation constraint is an interesting question left to future work.

Separation *vs* sketch size *vs* range of frequencies Assume we seek a given number of clusters k in a ball of given radius R in dimension d. Practically, given a reasonable minimium separation between cluster centers ε one has to chose a sketch size $m \geq \mathcal{O}(\log(1/\varepsilon))$. Conversely, if the maximum sketch size m is fixed, learning guarantees are available for separations $\varepsilon \geq e^{-\mathcal{O}(m)}$. A linear increase of the sketch size can thus be very valuable since it decreases exponentially one of the terms appearing the excess risk, see (65). Since the distribution of frequencies ω_j to be considered is parameterized by $\lambda = \mathcal{O}(1/\varepsilon)$, decreasing ε also means exploring higher frequencies.

5 Compressive Gaussian Mixture Modeling

We consider here Gaussian Mixture Modeling with known covariance.

5.1 Application of the framework

Definition of the learning task. We consider Gaussian Mixture Modeling on the sample space $\mathcal{Z} = \mathbb{R}^d$, with k Gaussian components with *fixed, known invertible covariance* matrix $\mathbf{\Sigma} \in \mathbb{R}^d$. We denote $\pi_c = \mathcal{N}(c, \mathbf{\Sigma})$.

An hypothesis $h = (c_1, ..., c_k, \alpha_1, ..., \alpha_k)$ contains the means and weights of the components of a GMM denoted $\pi_h = \sum_{l=1}^{k} \alpha_l \pi_{c_l}$, with $c_l \in \mathbb{R}^d$ and $\alpha \in \mathbb{S}_{k-1}$. The loss function for a density fitting problem is the negative log-likelihood:

$$\ell(x, h) = -\log \pi_h(x) \tag{66}$$

and correspondingly $\mathcal{R}_{\texttt{GMM}}(\pi, h) = \mathbb{E}_{X\sim\pi}(-\log \pi_h(X))$. As recalled in Annex A.3, the risk can also be written $\mathcal{R}_{\texttt{GMM}}(\pi, h) = \mathrm{KL}(\pi||\pi_h) + \mathrm{H}(\pi)$ with $\mathrm{KL}(\cdot||\cdot)$ the Kullback-Leibler divergence and $\mathrm{H}(\cdot)$ the differential entropy.

Model set $\mathfrak{S}_\mathcal{H}$ and best hypothesis for $\pi \in \mathfrak{S}_\mathcal{H}$. A natural model set for density fitting maximum log likelihood is precisely the model of all parametric densities:

$$\mathfrak{S}_\mathcal{H} = \{\pi_h : h \in \mathcal{H}\} \tag{67}$$

Separation assumption. Similar to the compressive clustering framework case of Section 4, we enforce a minimum separation between the means of the components of a GMM. We denote

$$\mathcal{H}_{k,\varepsilon,R} = \left\{(c_1, ..., c_k, \alpha_1, ..., \alpha_k) : c_l \in \mathbb{R}^d,\ \|c_l\|_{\mathbf{\Sigma}} \leq R,\ \|c_l - c_p\|_{\mathbf{\Sigma}} \geq \varepsilon,\ \alpha \in \mathbb{S}_{k-1}\right\} \tag{68}$$

where

$$\|c\|_{\mathbf{\Sigma}} := \sqrt{c^T \mathbf{\Sigma}^{-1} c} \tag{69}$$

is the Mahalanobis norm associated to the known covariance $\mathbf{\Sigma}$. Moreover, for any distribution $\hat{\pi} = \pi_{h_0}$ in the model set $\mathfrak{S}_\mathcal{H}$, the optimum of minimization (17) is $\hat{h} = h_0$ as it corresponds up to an offset to minimizing $\mathrm{KL}(\hat{\pi}||\pi_h)$.

Choice of feature function: random Fourier features. Compressive learning of GMMs with random Fourier features has been recently studied [17, 62]. Unlike compressive clustering we do not need to define a reweighted version of the Fourier features, and we directly sample m frequencies $\omega_1, \dots, \omega_m$ in $\mathbb{R}^d$ i.i.d from the distribution with density

$$\Lambda = \Lambda_\lambda = \mathcal{N}(0, \lambda^2 \mathbf{\Sigma}^{-1}) \tag{70}$$

with scale parameter λ. Define the associate feature function $\Phi : \mathbb{R}^d \to \mathbb{C}^m$:

$$\Phi(x) := \frac{C_\lambda}{\sqrt{m}} \left[e^{\mathrm{J} \omega_j^T x} \right]_{j=1,\dots,m}. \tag{71}$$

with

$$C_\lambda := (1 + 2\lambda^2)^{d/4} \leq e^{d\lambda^2/2}. \tag{72}$$

The constant C_λ multiplying Fourier features has no incidence on the recovery procedure but is included for coherence with the theory. As we will see in Section 6, with our approach both compressive clustering and compressive GMM are dealt with using the same general mathematical tools.

Computing and learning from a sketch. Given sample points $x_1, \dots, x_n$ in $\mathbb{R}^d$, compute the sketch $\mathbf{y}$ as in (4), i.e. a sampling of the conjugate of the empirical characteristic function [49] of the distribution π_0 of X. Then, given a vector $\mathbf{y}$ and a hypothesis class $\mathcal{H}$, estimate a solution of the following nonlinear least-squares problem:

$$\hat{h} := \arg\min_{h \in \mathcal{H}} \|\mathcal{A}(\pi_h) - \mathbf{y}\|_2^2 \tag{73}$$

where $\mathcal{A}$ is defined as (6). In particular, up to a scaling factor $C_\lambda / \sqrt{m}$, $\mathcal{A}(\pi_h)$ is a sampling of the conjugate characteristic function of the mixture of Gaussians π_h, which has here a closed form expression

$$\mathcal{A}(\pi_h) = \left[\frac{C_\lambda}{\sqrt{m}} \left(\sum_{l=1}^k \alpha_l e^{\mathrm{J} \omega_j^T c_l} \right) e^{-\frac{1}{2} \omega_j^T \mathbf{\Sigma} \omega_j} \right]_{j=1,\dots,m}.$$

We have the following guarantees.

Theorem 5.1. *Build the sketching function as in* (71) *where the* ω_j *are drawn according to* (70)*. Define the separation*

$$\varepsilon_\lambda := \sqrt{\frac{2 + 1/\lambda^2}{\sigma_k^2}} = \mathcal{O}\left(\sqrt{(1 + 1/\lambda^2) \log(ek)} \right) \tag{74}$$

and consider $R > \varepsilon_\lambda$ *large enough so that* $\mathcal{H}_{k, 2\varepsilon_\lambda, R}$ *is non-empty. Denote*

$$M_\lambda := (\varepsilon_\lambda C_\lambda)^2 = \frac{(1 + 2\lambda^2)^{\frac{d}{2}+1}}{\lambda^2 \sigma_k^2} \leq \frac{e^{d\lambda^2}(1 + 2\lambda^2)}{\lambda^2 \sigma_k^2}. \tag{75}$$

There is a universal constant $C > 0$ *such that, for any* $\zeta, \delta \in (0, 1)$*, when the sketch size* m *satisfies*

$$m \geq \delta^{-2} k^2 d M_\lambda \cdot \left[\log(ekM_\lambda) + \log \tfrac{R}{\varepsilon_\lambda} + \log \tfrac{1}{\delta} \right] + \delta^{-2} k M_\lambda \cdot \log \tfrac{1}{\zeta} \tag{76}$$

we have with probability at least $1 - \zeta$ *on the draw of the random Fourier frequencies* $(\omega_j)_{j=1}^m$*: for any* $\mathcal{H} \subset \mathcal{H}_{k, 2\varepsilon, R}$*, any source distribution* π_0 *on* $\mathcal{Z} = \mathbb{R}^d$*, any samples* $x_i \in \mathbb{R}^d$*,* $1 \leq i \leq n$ *(represented by*

the empirical distribution $\hat{\pi}_n$), denoting $\hat{h}$ the parameters of a GMM obtained by minimization (73) *and $h^\star \in \arg\min_{h\in\mathcal{H}} \mathcal{R}_{\texttt{GMM}}(\pi_0, h)$:*

$$\mathrm{KL}(\pi_0||\pi_{\hat{h}}) - \mathrm{KL}(\pi_0||\pi_{h^\star}) = \mathcal{R}_{\texttt{GMM}}(\pi_0, \hat{h}) - \mathcal{R}_{\texttt{GMM}}(\pi_0, h^\star) \leq \eta_n$$

where

$$\eta_n \leq 2D_{\texttt{GMM}}(\pi_0, \mathfrak{S}_{\mathcal{H}}) + \tfrac{C_\kappa}{\sqrt{1-\delta}} \left\|\mathcal{A}(\pi_0 - \hat{\pi}_n)\right\|_2 \tag{77}$$

where $C_\kappa = 4\sqrt{3}\sqrt{k}R^2$ and $D_{\texttt{GMM}}$ is the instantiation of (14) *for the norm* (10) *associated to the considered learning task.*

Remark 5.2. *Note that this holds with the sample space $\mathcal{Z} = \mathbb{R}^d$, i.e., we only restrict the means of the GMM,* not the data*, to the ball of radius R, $\mathcal{B}_{\mathbb{R}^d, \|\cdot\|_{\mathbf{\Sigma}}}(0, R)$.*

The proof follows the same steps as for Compressive Clustering, exploiting the generic results for mixtures that we detail in Section 6. The details are provided in Annex G. The major elements are to control the compatibility constant as $C_\kappa \leq 4\sqrt{3}\sqrt{k}R^2$, the concentration constant as $W_\kappa \leq 4\sqrt{k}\varepsilon_\lambda C_\lambda$, the concentration function as $c_\kappa(\delta/2) = \mathcal{O}\left(\delta^{-2}W_\kappa^2\right) = \mathcal{O}\left(\delta^{-2}kM_\lambda\right)$, and the covering numbers as $\log \mathcal{N}\left(d_\Phi, \mathcal{S}, \delta\right) = \mathcal{O}\left(kd \cdot \log\left(ekM_\lambda \cdot \frac{R}{\varepsilon} \cdot \frac{1}{\delta}\right)\right)$.

5.2 Discussion

Estimation error. The second term in the bound (77) is easy to control: since from (71) it holds that $\|\Phi(x)\|_2 = C_\lambda$, by applying the vectorial Hoeffding's inequality [74], with high probability w.r.t. data sampling it holds that $C_\kappa \left\|\mathcal{A}(\pi_0 - \hat{\pi}_n)\right\|_2 = \mathcal{O}\left(C_\lambda\sqrt{k}R^2/\sqrt{n}\right)$. To reach a given precision we thus need $n \gtrsim C_\lambda^2 kR^4$. Notice that when $\lambda = \mathcal{O}(1/d)$ we have $C_\lambda = \mathcal{O}(1)$. However C_λ can grow exponentially with d when λ is of order one or more, potentially requiring n to grow exponentially with d to have a small second term in (77).

Separation assumption. Given λ and k, Theorem 5.1 sets a separation condition ε_λ sufficient to ensure compressive statistical learning guarantees with the proposed sketching procedure, as well as a sketch size driven by M_λ. Contrary to the case of Compressive Clustering, one cannot target an arbitrary small separation as we have $\varepsilon_\lambda \geq \sqrt{2}/\sigma_k$ which is of the order of $\mathcal{O}\left(\sqrt{\log(ek)}\right)$. Reaching guarantees for this level of separation requires choosing λ of the order of one ($1/\lambda = \mathcal{O}(1)$). As we have just seen, this may require exponentially many training samples to reach a small estimation error, which is not necessarily surprising as such a level of separation is smaller that one can generally be found in the litterature, see e.g. [1, 34, 83]. For smaller values of λ the separation required for our results to hold is larger.

Sketch size. Contrary to the case of Compressive Clustering, the choice of λ also impacts the sketch size required for the guarantees of Theorem 5.1 to hold, through the value of M_λ. An easy function study shows that M_λ is minimum when $\lambda^2 = \frac{1}{d}$, leading to $M_\lambda \leq (d+2)e/\sigma_k^2 = \mathcal{O}(d\log(ek))$, and $C_\lambda = \mathcal{O}(1)$, however at the price of a larger separation condition $\varepsilon_\lambda = \mathcal{O}\left(\sqrt{d\log(ek)}\right)$. For larger values of λ, the required separation is smaller, but the estimation error captured by C_λ^2 increases as well as the sketch size under which we have guarantees. Choosing $\lambda^2 \ll 1/d$ does not seem to pay off.

Tradeoffs. Overall we observe a trade-off between the required sketch size, the required separation of the means in the considered class of GMMs, and the sample complexity. When λ increases, more high frequencies are sampled (or, equivalently, the analysis kernel is more precise), and the required separation of means decreases. As a price, more frequencies are required, and the sketch size increases as well as the estimation error factor C_λ^2. Overall, parameterizing $\lambda^2 = \frac{\gamma}{d}$ we get $\varepsilon_\lambda = \mathcal{O}\left(\sqrt{(\frac{d}{\gamma}+1)\log(ek)}\right)$, $C_\lambda^2 = \mathcal{O}\left(e^{\gamma/2}\right)$, $M_\lambda = \mathcal{O}\left((\frac{d}{\gamma}+1)\log(ek)e^{\gamma/2}\right)$, and $m \gtrsim \mathcal{O}\left(k^2 d(\frac{d}{\gamma}+1)e^{\gamma/2}\mathrm{polylog}(k,d,\gamma,d/\gamma)\right)$. We

Variance λ^2	Separation ε_λ	Estimation error factor C_λ^2	$M_\lambda = (\varepsilon_\lambda C_\lambda)^2$	Sketch size $k^2 d M_\lambda \log(ekM_\lambda)$
$\frac{1}{d}$	$\mathcal{O}\left(\sqrt{d\log(ek)}\right)$	$\mathcal{O}(1)$	$\mathcal{O}(d\log(ek))$	$\mathcal{O}\left(k^2d^2\mathrm{polylog}(k,d)\right)$
$\frac{\log(ek)}{d}$	$\mathcal{O}\left(\sqrt{d+\log(ek)}\right)$	$\mathcal{O}(k)$	$\mathcal{O}(k(d+\log(ek)))$	$\mathcal{O}\left(k^3d^2\mathrm{polylog}(k,d)\right)$
$\frac{1}{2}$	$\mathcal{O}\left(\sqrt{\log(ek)}\right)$	$2^{d/2}$	$\mathcal{O}\left(2^{d/2}\log(ek)\right)$	$\mathcal{O}\left(k^2d^2\ 2^{d/2}\mathrm{polylog}(k,d)\right)$

Table 2: Some tradeoffs between separation assumption, estimation error factor, and sketch size guarantees obtained using Theorem 5.1 for various values of the variance λ^2 of the frequency distribution (70).

give some particular values for λ in Table 5.2. The regime $\lambda = 1/2$ may be useful to resolve close Gaussians in moderate dimensions (typically $d \leq 10$) where the factor $2^{d/2}$ in sample complexity and sketch size remains tractable.

Learning algorithm and improved sketch size guarantees ? Again, although Theorem 5.1 only provides guarantees when $m \geq \mathcal{O}\left(k^2d^2\right)$ (up to logarithmic factors), the observed empirical phase transition pattern [63] (using an algorithm to adress (73) with a greedy heuristic) suggests that $m \geq \mathcal{O}(kd)$, of the order of the covering dimension, is in fact sufficient. We expect that this can be achieved theoretically with improved estimates of the concentration function $c_\kappa(t)$, which is likely to be $\mathcal{O}(1)$ rather than $\mathcal{O}(kd)$. Also, while Theorem 5.1 only handles mixtures of Gaussians with fixed known covariance matrix, the same algorithm has been observed to behave well for mixtures of Gaussians with unknown diagonal covariance.

Controlling the bias term ? Controlling the bias term

$$D_{\texttt{GMM}}(\pi_0, \mathfrak{S}_\mathcal{H}) = \inf_{h\in\mathfrak{S}_\mathcal{H}} \left\{ \|\pi_0 - \pi_h\|_{\mathcal{L}(\mathcal{H})} + C_\mathcal{A} \|\mathcal{A}(\pi_0 - \pi_h)\|_2 \right\} \tag{78}$$

for Gaussian Mixture Modeling seems more delicate than for clustering, as we can no longer rely on our Lemma 2.10. As the sketching functions are uniformly bounded, using Pinsker's inequality (see Annex A.3) we have $\|\mathcal{A}(\pi_0 - \pi_h)\|_2 \lesssim \|\pi_0 - \pi_h\|_{\mathrm{TV}} \lesssim \sqrt{2\mathrm{KL}(\pi_0||\pi_h)}$. Its infimum over h is thus bounded by a constant times $\sqrt{\mathrm{KL}(\pi_0||\pi_{h^\star})}$. As $\mathcal{R}_{\texttt{GMM}}(\pi_0, h)$ is, up to an additive offset, equal to $\mathrm{KL}(\pi_0||\pi_h)$, this is reminiscent of the type of distribution free control obtained for k-means. Establishing such a result would however require controlling the term $\|\pi_0 - \pi_h\|_{\mathcal{L}(\mathcal{H})}$, which is left to future work.

6 Recovery guarantees for general mixture models

Given the hypotheses of our main Theorem 2.9, for a loss class $\mathcal{L}(\mathcal{H})$ (see Section 2.4) and a model $\mathfrak{S}$ (see Section 2.5), our goal is to find a kernel κ along with its integral representation (Φ, Λ) such that the compatibility constant C_κ (see (25)) is finite; the concentration constant W_κ (see (30)) is finite;

the normalized secant set $\mathcal{S}_{\|\cdot\|_\kappa}$ (see (32)) has finite covering numbers. As the distance $\|\pi - \pi'\|_\kappa$ is the denominator of all these expressions, most difficulties arise when $\|\pi - \pi'\|_\kappa$ is small ($\pi, \pi' \in \mathfrak{S}$ get "close" to each other) and we primarily have to control the ratio $\|\pi - \pi'\| / \|\pi - \pi'\|_\kappa$ for various norms when $\|\pi - \pi'\|_\kappa \to 0$.

In this section, we develop a framework to control these quantities when the model $\mathfrak{S}$ is a mixture model, which covers mixtures of Diracs (see Section 4) and mixtures of Gaussians (see Section 5).

Definition 6.1 (Mixture model)**.** *Given $\mathcal{T} = \{\pi_\theta : \theta \in \mathbf{\Theta}\}$ a family of probability distributions (e.g. $\mathcal{T}$ may be a family of Diracs or of Gaussians), and an integer $k > 0$, we define the* mixture model

$$\mathfrak{S}_k(\mathcal{T}) := \left\{ \sum_{l=1}^k \alpha_l \pi_l : \alpha_l \geq 0,\ \sum_{l=1}^k \alpha_l = 1,\ \pi_l \in \mathcal{T} \right\}.$$

The existence of a finite compatibility constant means that for any $\pi, \pi' \in \mathfrak{S}_k(\mathcal{T})$ we must have $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \leq C_\kappa \|\pi - \pi'\|_\kappa$, so in particular for any $\theta, \theta' \in \mathbf{\Theta}$ and any loss function $f(\cdot) := \ell(\cdot, h)$, $h \in \mathcal{H}$, we must have $\left|\mathbb{E}_{X\sim\pi_\theta} f(X) - \mathbb{E}_{X\sim\pi_{\theta'}} f(X)\right| \leq \|\pi_\theta - \pi_{\theta'}\|_{\mathcal{L}(\mathcal{H})} \leq C_\kappa \|\pi_\theta - \pi_{\theta'}\|_\kappa$. For the particular case of mixtures of Diracs this reads $|f(\theta) - f(\theta')| \leq C_\kappa \varrho_\kappa(\theta, \theta')$ where $\varrho_\kappa(\theta, \theta') := \|\delta_\theta - \delta_{\theta'}\|_\kappa$, i.e., f is Lipschitz with respect to a certain metric $\varrho_\kappa(\cdot, \cdot)$ between parameters, which is Hilbertian after embedding the parameters in an appropriate Hilbert space.

6.1 Bounded and Lipschitz functions "in expectation"

Vice-versa, for certain types of kernels (see Section 6.3), we will control the ratio $\|\pi - \pi'\| / \|\pi - \pi'\|_\kappa$ for various norms by controlling $\|\pi - \pi'\|_{\mathcal{BL}} / \|\pi - \pi'\|_\kappa$ where $\mathcal{BL}$ is the following class of functions.

Definition 6.2 ("Bounded and Lipschitz in expectation" functions)**.** *A function $f : \mathcal{Z} \to \mathbb{C}$ is "bounded and Lipschitz (with respect to a metric $\varrho(\cdot, \cdot)$ on the parameter space $\mathbf{\Theta}$) in expectation" on the basic set $\mathcal{T} = \{\pi_\theta : \theta \in \mathbf{\Theta}\}$ if there exists $D, L < \infty$ such that for all $\theta, \theta' \in \mathbf{\Theta}$,*

$$\left|\mathbb{E}_{X\sim\pi_\theta} f(X)\right| \leq D, \tag{79}$$

$$\left|\mathbb{E}_{X\sim\pi_\theta} f(X) - \mathbb{E}_{X'\sim\pi_{\theta'}} f(X')\right| \leq L \cdot \varrho(\theta, \theta'). \tag{80}$$

We denote $\mathcal{BL}(D, L, \mathcal{T}, \varrho)$ (or in short $\mathcal{BL}(D, L)$) the set of all functions satisfying (79)-(80)*.*

6.2 Separated mixtures models and dipoles

We will be interested in a restricted mixture model taking into account a notion of separation between components. Given ϱ be a metric on the parameter space $\mathbf{\Theta}$ of the form

$$\varrho(\theta, \theta') = \|\psi(\theta) - \psi(\theta')\|_2 \tag{81}$$

with $\psi(\cdot)$ a mapping from $\mathbf{\Theta}$ to some Euclidean space, we define the mixture set with ε-separation

$$\mathfrak{S}_{k,\varepsilon,\varrho}(\mathcal{T}) := \left\{ \sum_{l=1}^k \alpha_l \pi_{\theta_l} : \alpha_l \geq 0,\ \sum_{l=1}^k \alpha_l = 1,\ \pi_{\theta_l} \in \mathcal{T},\ \varrho(\theta_l, \theta_{l'}) \geq \varepsilon\ \forall l \neq l' \right\} \tag{82}$$

or simply $\mathfrak{S}_{k,\varepsilon}(\mathcal{T})$ for short when there is no ambiguity on which metric ϱ is used.

Remark 6.3. *In the following, for lighter notations we will incorporate ε into the metric and work with a constant 2-separation, i.e., with $\mathfrak{S}_{k,2,\varrho}(\mathcal{T})$. For example, in the case of Diracs, choosing $\varrho(c, c') = \|c - c'\|_2 / \varepsilon$ yields $\mathfrak{S}_{k,2,\varrho}(\mathcal{T}) = \mathfrak{S}_{k,2\varepsilon,\|\cdot\|_2}(\mathcal{T})$ with the desired 2ε-separation in Euclidean norm. In the case of Gaussians, the same holds with $\varrho(c, c') = \|c - c'\|_{\mathbf{\Sigma}} / \varepsilon$.*

The notion of *dipoles* will turn out to be particularly useful in our analysis.

Definition 6.4 (Dipoles, separation of dipoles)**.** *A finite signed measure*[7] ν *is called a* **dipole** *with respect to the metric* ϱ *and the set* $\mathcal{T}$ *if it admits a decomposition as* $\nu = \alpha_1 \pi_{\theta_1} - \alpha_2 \pi_{\theta_2}$ *where* $\pi_{\theta_1}, \pi_{\theta_2} \in \mathcal{T}$*, such that*

$$\varrho(\theta_1, \theta_2) \leq 1$$

and $0 \leq \alpha_i \leq 1$ *for* $i = 1, 2$*. Note that the coefficients* α_i*'s are not normalized to* 1*, and that any of them can be put to* 0 *to yield a* monopole *as a special case of dipole.*

Two dipoles ν, ν' *are called* **1-separated** *if they admit a decomposition* $\nu = \alpha_1 \pi_{\theta_1} - \alpha_2 \pi_{\theta_2}$*,* $\nu' = \alpha'_1 \pi_{\theta'_1} - \alpha'_2 \pi_{\theta'_2}$ *such that* $\varrho(\theta_i, \theta'_j) \geq 1$ *for all* $i, j \in \{1, 2\}$*.*

With these definitions, elements of the secant set of the model $\mathfrak{S}_{k,2}(\mathcal{T})$ are sums of $2k$ pairwise 1-separated dipoles.

Lemma 6.5. *Let* $\pi, \pi' \in \mathfrak{S}_{k,2}(\mathcal{T})$*. It holds that*

$$\pi - \pi' = \sum_{l=1}^{2k} \nu_l$$

where the measures ν_l *are dipoles that are pairwise* 1*-separated.*

Proof. Using the 2-separation in π and π' and the triangle inequality, for the metric ϱ each parameter θ_i in π is 1-close to *at most* one parameter θ'_j in π', and 1-separated from all other components in both π and π'. Hence $\pi - \pi'$ can be decomposed into a sum of (at most) $2k$ dipoles (which may also be monopoles). Adding zeros if needed, we obtain exactly $2k$ dipoles. □

6.3 RBF-like Mean Map Embeddings

We focus on kernels $\kappa(\cdot, \cdot)$ such that the Mean Map Embedding (18) can be expressed as:

$$\kappa(\pi_\theta, \pi_{\theta'}) = K(\varrho(\theta, \theta')), \quad \forall \theta, \theta' \in \boldsymbol{\Theta} \tag{83}$$

where $K : \mathbb{R}_+ \to \mathbb{R}_+$. For the particular case where ϱ is the Euclidean distance, this corresponds to assuming that $\kappa(\pi_\theta, \pi_{\theta'}) = K(\|\theta - \theta'\|_2)$, i.e., the Mean Map Embedding is a radial basis function (RBF) of the parameters. Hence, (83) characterizes a family of *RBF-like Mean Map Embeddings*. Such embeddings can "distinguish" separated dipoles, under some assumptions on the function $K(\cdot)$.

Definition 6.6. *The class* $\mathcal{E}(A, B, C, c)$ *consists of all functions* $K : \mathbb{R}_+ \to \mathbb{R}_+$ *that satisfy*

i) over the interval $[0, 1]$*:*

- $K(0) = 1$ *;*
- $K(u) \leq 1 - \frac{cu^2}{2}$ *for all* $u \leq 1$ *;*

ii) over the interval $[1, \infty)$*:*

- K *is bounded:* $0 \leq K(u) \leq A$*, for all* $u \geq 1$ *;*
- K *is differentiable with bounded derivative:* $|K'(u)| \leq B$*, for all* $u \geq 1$ *;*
- K' *is* C*-Lipschitz:* $|K'(u) - K'(v)| \leq C\,|u - v|$*, for all* $u, v \geq 1$ *.*

[7] See Annex A.2

Definition 6.7. *The class $\mathcal{E}_k(c)$ consists of all functions $K : \mathbb{R}_+ \to \mathbb{R}_+$ such that $K(\cdot) \in \mathcal{E}(A, B, C, c)$ for some A, B, C such that*

$$1 \leq k \leq \tfrac{1}{2} + \frac{3\min(c,1)}{64\max(A, 2(B+C))}. \tag{84}$$

For any integer k, one can indeed design a function $K(\cdot) \in \mathcal{E}_k(1)$ using the Gaussian kernel. The proof is in Annex E.1.

Lemma 6.8. *Define for any $\sigma > 0$ and any $k \geq 1$,*

$$K_\sigma(u) := e^{-\frac{u^2}{2\sigma^2}}, \quad u \geq 0 \tag{85}$$

$$\sigma_k^2 := \frac{1}{2.4(\log(2k-1)+10)}. \tag{86}$$

We have $0 < \sigma_k^2 \leq 1/24 < 1$, and for any $0 < \sigma \leq \sigma_k$, $K_\sigma(\cdot) \in \mathcal{E}_k(1)$.

6.4 Mutual coherence between separated dipoles

The motivation for choosing RBF-like kernels in the class $\mathcal{E}_k(c)$ is our first main technical lemma below which bounds what can be considered as the *mutual coherence* (with respect to the inner product defined by the RBF-like Mean Map Embedding) between any pair of dipoles that are 1-separated and uses Gersgorin's disc theorem to handle sums of mutually 1-separated dipoles. The proof is in Annex E.2.

Lemma 6.9. *Consider a function $K(\cdot) \in \mathcal{E}(A, B, C, c)$. For any kernel $\kappa(\cdot, \cdot)$ and any set $\mathcal{T} = \{\pi_\theta : \theta \in \mathbf{\Theta}\}$ such that the Mean Map Embedding* (18) *satisfies* (83) *with some metric ϱ of the form* (81)*, the following holds:*

- *for any two dipoles (with respect to $\mathcal{T}$ and ϱ) that are 1-separated from each other, ν, ν', we have*[8]

$$\frac{|\kappa(\nu, \nu')|}{\|\nu\|_\kappa \|\nu'\|_\kappa} \leq M = M(A, B, C, c) := \tfrac{8\max(A, 2(B+C))}{\min(c,1)}; \tag{87}$$

- *for any ℓ dipoles ν_l that are pairwise 1-separated, we have*

$$1 - M \cdot (\ell - 1) \leq \frac{\left\|\sum_{l=1}^\ell \nu_l\right\|_\kappa^2}{\sum_{l=1}^\ell \|\nu_l\|_\kappa^2} \leq 1 + M \cdot (\ell - 1). \tag{88}$$

Specializing to $\ell = 2k$, condition (84) *reads $k \leq \frac{1}{2}(1 + \frac{3}{4M})$ which is equivalent to $1 - M \cdot (2k-1) \geq 1/4$.*

Remark 6.10. *The reader familiar with classic results on sparse recovery will find this lemma highly reminiscent of the classical link between the coherence of a dictionary and its restricted isometry property, see e.g. [50, Theorem 5.13]. To handle incoherence in a continuous "off the grid" setting (such as mixtures of separated Diracs in section F, which also appear in super-resolution imaging scenarios [19, 35, 41]), the apparently new trick is to consider incoherence* between dipoles *rather than between monopoles.*

[8] We properly define in Annex A.2 the extension of the Mean Map Embedding to finite signed measures, to make sense of the notation $\kappa(\nu, \nu')$.

6.5 Compatibility constant, concentration constant

Our main tool to control the compatibility constant C_κ and concentration constant W_κ is a bound on $\|\pi - \pi'\|_{\mathcal{BL}(D,L)} / \|\pi - \pi'\|_\kappa$ when π, π' are separated mixtures. Then, on a case-by-case basis, obtaining the desired results will amount to proving that:

a) the family of loss functions $\mathcal{L}(\mathcal{H})$ characterizing the learning task is a subset of $\mathcal{BL}(D_\mathcal{L}, L_\mathcal{L})$ for some constants $D_\mathcal{L}, L_\mathcal{L}$;

b) the family of functions $\Phi = \{\phi_\omega\}$ used to define the random feature function Φ in (20) is also a subset of $\mathcal{BL}(D_\Phi, L_\Phi)$ for some other constants D_Φ, L_Φ.

This will yield explicit bounds on the compatibility constant C_κ and the concentration constant W_κ.

We first bound $\|\cdot\|_{\mathcal{BL}(D,L)} / \|\cdot\|_\kappa$ for dipoles.

Lemma 6.11. *Consider a function $K(\cdot) \in \mathcal{E}(A, B, C, c)$, a set $\mathcal{T} = \{\pi_\theta : \theta \in \mathbf{\Theta}\}$, and a metric ϱ on $\mathbf{\Theta}$ of the form* (81)*. For any kernel $\kappa(\cdot, \cdot)$ such that the Mean Map Embedding* (18) *satisfies* (83)*, the following holds:*

- *for any dipole μ and any $D, L > 0$ we have*

$$\frac{\|\mu\|_{\mathcal{BL}(D,L,\mathcal{T},\varrho)}}{\|\mu\|_\kappa} \leq \left(L^2/c + 2D^2\right)^{\frac{1}{2}}. \tag{89}$$

NB: the constants A, B, C from the definition of $\mathcal{E}(A, B, C, c)$ do not play a role in this lemma.

The proof is is Annex E.3. Using the dipole decomposition of Lemma 6.5, and the bounded mutual coherence between separated dipoles, we can obtain the desired bound.

Theorem 6.12. *Consider $K(\cdot) \in \mathcal{E}_k(c)$ with k the number of mixture components, a set $\mathcal{T} = \{\pi_\theta : \theta \in \mathbf{\Theta}\}$, and a metric ϱ on $\mathbf{\Theta}$ of the form* (81)*. For any kernel $\kappa(\cdot, \cdot)$ such that the Mean Map Embedding* (18) *satisfies* (83)*, the following holds:*

- *for all 2-separated mixtures $\pi,\ \pi' \in \mathfrak{S}_{k,2,\varrho}(\mathcal{T})$ and any $D, L > 0$, we have*

$$\|\pi - \pi'\|_{\mathcal{BL}(D,L,\mathcal{T},\varrho)} \leq 2\left(L^2/c + 2D^2\right)^{\frac{1}{2}} \sqrt{2k}\, \|\pi - \pi'\|_\kappa \tag{90}$$

Proof. Let $\pi,\ \pi' \in \mathfrak{S}_{k,2,\varrho}(\mathcal{T})$. Using Lemma 6.5 we obtain a decomposition $\pi - \pi' = \sum_{i=1}^{2k} \nu_i$ where the ν_i's are dipoles that are pairwise 1-separated. By the triangle inequality and Lemma 6.11 we have

$$\|\pi - \pi'\|_{\mathcal{BL}(D,L)} \leq \sum_{i=1}^{2k} \|\nu_i\|_{\mathcal{BL}(D,L)} \leq \left(L^2/c + 2D^2\right)^{\frac{1}{2}} \sum_{i=1}^{2k} \|\nu_i\|_\kappa \leq \left(L^2/c + 2D^2\right)^{\frac{1}{2}} \sqrt{2k} \left(\sum_{i=1}^{2k} \|\nu_i\|_\kappa^2\right)^{\frac{1}{2}}$$

Since $K \in \mathcal{E}_k(c)$, (84) holds which is equivalent to $1 - M(2k-1) \geq 1/4$. By Lemma 6.9 we have

$$\|\pi - \pi'\|_{\mathcal{BL}(D,L)} \leq \frac{\left(L^2/c + 2D^2\right)^{\frac{1}{2}}}{\sqrt{1 - M(k-1)}} \sqrt{2k} \left\|\sum_{i=1}^{2k} \nu_i\right\|_\kappa \leq 2\left(L^2/c + 2D^2\right)^{\frac{1}{2}} \sqrt{2k}\, \|\pi - \pi'\|_\kappa .$$

□

In Annex F (resp. Annex G) we characterize $L_\mathcal{H}$, $D_\mathcal{H}$, L_Φ, D_Φ, for mixtures of Diracs (resp. of Gaussians), leading to bounds on the compatibility constant C_κ and the concentration constant W_κ.

6.6 Covering numbers of the secant set

Thanks again to the decomposition of $\pi - \pi'$ into separated dipoles, it is sufficient to control the covering numbers (with respect to the metric $\|\cdot\|_\Phi$ instead of d_Φ) of the set of *normalized dipoles*.

$$\mathcal{D} := \left\{ \frac{\nu}{\|\nu\|_\kappa} : \nu \text{ is a dipole}, \ \|\nu\|_\kappa > 0 \right\} \tag{91}$$

Theorem 6.13. *Consider $\kappa(\cdot,\cdot)$, $\mathcal{T} = \{\pi_\theta : \theta \in \mathbf{\Theta}\}$, and a metric ϱ of the form* (81) *such that*

1. *the Mean Map Embedding* (18) *satisfies* (83) *with some $K(\cdot) \in \mathcal{E}_k(c)$;*
2. *the kernel $\kappa(x,x')$ admits an integral representation* (21) *(Φ,Λ) with $\Phi \subset \mathcal{BL}(D,L,\mathcal{T},\varrho)$.*

With $W := \left(L^2/c + 2D^2\right)^{\frac{1}{2}}$, we have for any $\delta > 0$

$$\mathcal{N}\left(d_\Phi, \mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S}_{k,2}(\mathcal{T})), \delta\right) \leq \left[\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}, \tfrac{\delta}{C_0}\right) \cdot \max\left(1, \tfrac{C_1}{\delta}\right)\right]^{2k} \tag{92}$$

where $C_0 := 64kW$, $C_1 := 256kW^2$.

The proof is in Annex E.4.

Controlling the covering numbers of the set $\mathcal{D}$ of normalized dipoles can be done in part for dipoles associated to π and π' "far away" from one another. This leads to a control in terms of the covering numbers of $\mathbf{\Theta}$, which are often relatively easy to characterize. Getting a complete control requires a finer study of the kernel metric $\|\pi_\theta - \pi_{\theta'}\|_\kappa$ when θ and θ' are close to each other with respect to the metric $\varrho(\theta,\theta')$. A relevant notion is that of a *tangent approximation*.

Definition 6.14 (Tangent approximation)**.** *Consider a kernel $\kappa(\cdot,\cdot)$ and a norm $\|\cdot\|$ on finite signed measures. The set $\mathcal{V}$ made of tempered distributions is said to be a* tangent approximation *to the set of normalized dipoles $\mathcal{D}$ with respect to the metrics $\|\cdot\|_\kappa$, $\|\cdot\|$, with constants $t, T > 0$ if for any $\theta, \theta' \in \mathbf{\Theta}$ such that $\|\pi_\theta - \pi_{\theta'}\|_\kappa \leq t$, there is $\nu \in \mathcal{V}$ such that*

$$\left\| \frac{\pi_\theta - \pi_{\theta'}}{\|\pi_\theta - \pi_{\theta'}\|_\kappa} - \nu \right\| \leq T \|\pi_\theta - \pi_{\theta'}\|_\kappa . \tag{93}$$

We have the following Theorem.

Theorem 6.15. *Under the assumptions of Theorem 6.13, assume that $\mathcal{V}$ is a tangent approximation to the set of normalized dipoles $\mathcal{D}$ with constants $t, T > 0$ with respect to $\|\cdot\|_\Phi$. Denote $W := \left(L^2/c + 2D^2\right)^{\frac{1}{2}}$ and*

$$\mathcal{V}' := \{\alpha\nu + \beta\pi_\theta : \ \nu \in \mathcal{V}, \theta \in \mathbf{\Theta}, \ 0 \leq \alpha \leq 2, \ 0 \leq \beta \leq 1\} . \tag{94}$$

For any $\delta \leq 16T\min(3/4, t/2)$, we have

$$\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}, \delta\right) \ \leq \ \mathcal{N}\left(\varrho, \mathbf{\Theta}, \tfrac{\delta^2}{C_2}\right)^2 \cdot \max\left(1, \left(\tfrac{C_3}{\delta}\right)^2\right) + \mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{V}', \tfrac{\delta}{C_4}\right) \tag{95}$$

with $C_2 := 256WLT$, $C_3 := 16\sqrt{WDT}$, $C_4 := 4$.

The proof is in Annex E.5. In Sections F and G we prove on a case-by-case basis for Diracs and Gaussians the existence of such a tangent approximation, with $\|\cdot\| = \|\cdot\|_\Phi$ the metric associated to a representation (Φ,Λ) of the kernel. We further control the needed covering numbers.

6.7 Summary

Given a separated mixture model set $\mathfrak{S}_{k,2,\varrho}(\mathcal{T})$ with basic distributions $\mathcal{T} = \{\pi_\theta : \theta \in \mathbf{\Theta}\}$, bounding the compability and concentration constants and covering numbers (i.e. obtaining learning guarantees) amounts to finding a kernel $\kappa(\cdot,\cdot)$ with an integral representation $(\Phi = \{\phi_\omega : \omega \in \Omega\}, \Lambda)$ and a metric ϱ of the form (81) such that

1. the Mean Map embedding satisfies $\kappa(\pi_\theta, \pi_{\theta'}) = K(\varrho(\theta, \theta'))$ where $K \in \mathcal{E}_k(c)$;
2. the random features ϕ_ω are Bounded and Lipschitz in expectation: $\Phi \subset \mathcal{BL}(D_\Phi, L_\Phi, \mathcal{T}, \varrho)$;
3. the loss functions $\ell(\cdot, h)$ are also Bounded and Lipschitz in expectation: $\mathcal{L}(\mathcal{H}) \subset \mathcal{BL}(D_\mathcal{L}, L_\mathcal{L}, \mathcal{T}, \varrho)$;
4. the parameter set $\mathbf{\Theta}$ has controlled covering numbers $\mathcal{N}(\varrho, \mathbf{\Theta}, \delta)$;
5. the dipole set $\mathcal{D}$ has a tangent approximation $\mathcal{V}$ with controlled covering numbers $\mathcal{N}(\|\cdot\|_\Phi, \mathcal{V}, \delta)$.

As a consequence we get:

- applying Theorem 6.12 on the class Φ of random features, the concentration constant is finite
$$W_\kappa = 2\sqrt{2k}(2D_\Phi^2 + L_\Phi^2/c)^{\frac{1}{2}}; \tag{96}$$
- applying Theorem 6.12 on the loss class $\mathcal{L}(\mathcal{H})$, the compatibility constant is finite
$$C_\kappa = 2\sqrt{2k}(2D_\mathcal{L}^2 + L_\mathcal{L}^2/c)^{\frac{1}{2}}; \tag{97}$$
- applying Theorems 6.13-6.15, the secant set $\mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S}_{k,2,\varrho}(\mathcal{T}))$ has controlled covering numbers with respect to d_Φ.

This allows us to leverage Theorem 2.9. This is precisely the strategy we follow in Annexes F-G to establish concrete results for compressive clustering and compressive Gaussian Mixture Modeling.

7 Conclusion and perspectives

The principle of compressive statistical learning is to learn from large-scale collections by first summarizing the collection into a sketch vector made of empirical (random) moments, before solving a nonlinear least squares problem. The main contribution of this paper is to setup a general mathematical framework for compressive statistical learning and to demonstrate on three examples (compressive PCA, compressive clustering and compressive Gaussian mixture estimation –with fixed known covariance) that the excess risk of this procedure can be controlled, as well as the sketch size. In a sense, the random feature moments that constitute the sketch vector play the role of sufficient statistics for the considered learning task.

Sharpened estimates ? Our demonstration of the validity of the compressive statistical learning framework for certain tasks is, in a sense, qualitative, and we expect that many bounds and constants are sub-optimal. This is the case for example of the estimated sketch sizes for which statistical learning guarantees have been established, and an immediate theoretical challenge is to sharpen these guarantees to match the empirical phase transitions observed empirically for compressive clustering and compressive GMM [63, 64]. As our control of the sketch size combines an estimate of the covering dimension of the secant set –which seems to be of the right order of magnitude– with a concentration constant (27), such a tightening will likely follow from better concentration estimates for mixture models.

A number of non-sharp oracle inequalities have been established in the course of our endeavor. For mixture models, as our proof technique involves Geshgorin's disc theorem, it is natural to wonder to what extent the involved constants can be tightened to get closer to sharp oracle inequalities, possibly at the price of larger sketch sizes. A related problem is to obtain more explicit and/or tighter control of the bias term $D(\pi_0, \mathfrak{S}_{\mathcal{H}})$, in particular for k-means and GMM, and to understand whether Lemma 2.10, which relates this bias term to the optimal risk, can be tightened and/or extended to other loss functions.

In the same vein, as fast rates (see e.g. [67] for the case of k-means) on the estimation error can be established for certain classical statistical learning task (under appropriate margin conditions), it is natural to wonder whether the same holds for compressive statistical learning.

Overall, an important question to benchmark the quality of the established bounds (on achievable sketch sizes, on the separation assumptions used for k-mixtures, etc.) is of course to investigate corresponding lower-bounds.

Provably-good algorithms of bounded complexity? As the control of the excess risk relies on the minimizer of a nonlinear least-squares problem (37), the results in this paper are essentially information-theoretic. Can we go beyond the heuristic optimization algorithms derived for compressive K-means and compressive GMM [63, 64] and characterize provably good, computationally efficient algorithms to obtain this minimizer ?

Promising directions revolve around recent advances in super-resolution imaging and low-rank matrix recovery. For compressive clustering (resp. compressive GMM), the similarity between problem (54) (resp. (73)) and super-resolution imaging suggests to explore TV-norm minimization –a *convex* problem– techniques [19, 35, 41] and to seek generalized RIP guarantees [81]. Further, to circumvent the difficulties of optimization (convex or not) in the space of finite signed measures, it may also be possible to adapt the recent guarantees obtained for certain nonconvex problems that directly leverage a convex "lifted" problem [68] without incurring the cost of actually computing in the lifted domain.

Finally, the computational cost of sketching itself could be further controlled by replacing random Gaussian weights where possible with fast approximations [66, 25, 15]. This is likely to also result in accelerations of the learning stage wherever matrix multiplications are exploited. To conduct the theoretical analysis of the resulting sketching procedure, one will need to analyze the kernels associated to these fast approximations.

Links with convolutional neural networks. From an algorithmic perspective, the sketching techniques we have explicitly characterized in this paper have a particular structure which is reminiscent of one layer of a (random) convolutive neural network with average pooling. Indeed, when the sketching function Φ corresponds to (weighted or not) random Fourier features, its computation for a given vector x involves first multiplication by the matrix $\mathbf{W} \in \mathbb{R}^{m \times d}$ which rows are the selected frequencies $\omega_j \in \mathbb{R}^d$, then pointwise application of the $e^{\mathrm{j}\cdot}$ nonlinearity.

Here we consider random Fourier *moments*, hence a subsequent *average pooling* operation computing $\frac{1}{n}\sum_{i=1}^n \Phi(x_i)$. Up to the choice of nonlinearity and (random) weights, this is exactly the computational structure of a convolutional neural network where the x_i would be the collection of all patches from an image or frames from a time series.

This suggests that our analysis could help analyze the tradeoffs between reduction of the information flow (dimension reduction) across multiple layers of such networks and the preservation of statistical information. For example, this could explain why the pooled output of a layer is rich enough to cluster the input patches. Given the focus on drastic dimension reduction, this seems very complementary to the recent work on the invertibility of deep networks and pooling representations with random Gaussian weights [43, 55, 54].

Privacy-aware learning via sketching ? The reader may have noticed that, while we have defined sketching in (4) as the empirical average of (random) features $\Phi(x_i)$ over the training collection (or in fact the training *stream*), the essential feature of the sketching procedure is to provide a good empirical estimator of the sketch vector $\mathcal{A}(\pi_0) = \mathbb{E}_{X\sim\pi_0}\Phi(X)$ of the underlying probability distribution. A consequence is that one can envision *other sketching mechanisms*, in particular ones more compatible with privacy-preservation constraints [39]. For example, one could average $\Phi(x_i + \xi_i)$, or $\Phi(x_i) + \xi_i$, or $\mathbf{D}_i\Phi(x_i)$, etc., where ξ_i is a heavy-tailed random vector drawn independently from x_i, and $\mathbf{D}_i$ is a diagonal "masking" matrix with random Bernoulli $\{0, 1\}$ entries. An interesting perspective is to characterize such schemes in terms of tradeoffs between differential privacy and ability to learn from the resulting sketch.

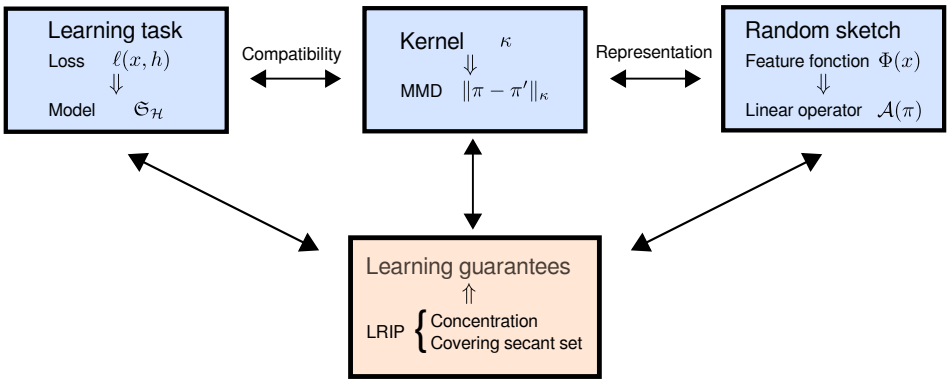

Figure 1: A representation of the links between different concepts in this paper.

Recipes to design sketches for other learning tasks through kernel design? Given the apparent genericity of the proposed compressive statistical learning framework, a particular challenge is to extend it beyond the learning tasks considered in this paper. Kernel versions of these tasks (*kernel* PCA, *kernel* k-means, or *spectral* clustering) appear as the most likely immediate extensions. They are expected to lead to sketching architectures reminiscent of *two-layer* convolutional neural networks with additive pooling.

Compressive supervised classification and compressive regression seem more challenging. Given a learning task, the main challenge is to find an adequate sketching function $\Phi(\cdot)$.

As illustrated on Figure 1, this primarily relies on the quest for a *compatible kernel*, i.e., one satisfying the Kernel LRIP (23). Subsequent technical steps would rely on the identification of an integral representation of this kernel using random features with the right concentration properties, and establishing that the associated secant set has finite covering dimension with respect to the feature-based metric (33). On a case by case basis, one may have to identify the analog of the separation conditions apparently need for compressive k-means.

Vice-versa, one could wonder which family of learning tasks is compatible with a given kernel. In other words, how "universal" is a kernel, and how much can be learned from a single sketched representation of a database ? We expect that tasks such as compressive ranking, which involve pairs, triples, etc. of training samples, may require further extensions of the compressive statistical learning framework, to design sketches based on U-statistics rather than plain moments. These would be lead to sketches linear in the product probability $\pi_0 \otimes \pi_0$ instead of π_0. The investigation of such extended scenarios is expected to benefit from analogies with the lifting techniques used in phaseless reconstruction, see e.g. [18].

Acknowledgements

This work was supported in part by the European Research Council, PLEASE project (ERC-StG-2011-277906). Rémi Gribonval is very grateful to Michael E. Davies for many enlightening discussions around the idea of compressive statistical learning since this project started several years ago. The authors also wish to warmly thank Bernard Delyon and Adrien Saumard for their constructive feedback on early versions of this manuscript.

Annex

We begin by introducing notations and useful results. We then provide general properties on covering numbers, followed by properties that are shared by any model of mixtures of distributions that are sufficiently separated $\mathfrak{S} = \mathfrak{S}_{k,\varepsilon}$. We then apply these results to mixtures of Diracs and both k-medians and k-means risks, and to Gaussian Mixture Models with fixed known covariance for maximum likelihood estimation. We conclude with the case of Compressive PCA.

A Notations, definitions

In this section we group all notations and some useful classical results.

A.1 Metrics and covering numbers

Definition A.1. *A **pseudometric** d over a set X satisfies all the axioms of a metric, except that $d(x, y) = 0$ does not necessarily imply $x = y$. Similarly, a **seminorm** $\|\cdot\|$ over a vector space X satisfies the axioms of a norm except that $\|x\| = 0$ does not necessarily imply $x = 0$.*

The **radius** of a subset Y of a seminormed vector space $(X, \|\cdot\|)$ is denoted $\mathrm{rad}_{\|\cdot\|}(Y) := \sup_{x\in Y} \|x\|$. The **diameter** of a pseudometric set (X, d) is denoted $\mathrm{diam}_d(X) := \sup_{x,x'\in X} d(x, x')$.

Definition A.2 (Ball, δ-covering, Covering number)**.** *Let (X, d) be a pseudometric space. For any $\delta > 0$ and $x \in X$, we denote $\mathcal{B}_{X,d}(x, \delta)$ the **ball** of radius δ centered at the point x:*

$$\mathcal{B}_{X,d}(x, \delta) = \{y \in X,\ d(x, y) \leq \delta\}.$$

*Let $Y \subseteq X$ be a subset of X. A subset $Z \subseteq Y$ is a **δ-covering** of Y if $Y \subseteq \bigcup_{z\in Z} \mathcal{B}_{X,d}(z, \delta)$. The **covering number** $\mathcal{N}(d, Y, \delta) \in \mathbb{N} \cup \{+\infty\}$ is the smallest k such that there exists an δ-covering of Y made of k elements $z_i \in Y$.*

A.2 Finite signed measures

The space $\mathfrak{M}$ of finite signed measures on the sample space $\mathcal{Z}$ is a linear space that contains the set of probability distributions on $\mathcal{Z}$. Any finite signed measure $\mu \in \mathfrak{M}$ can be decomposed into a positive and a negative part, $\mu = \mu_+ - \mu_-$, where both μ_+ and μ_- are non-negative finite measures on $\mathcal{Z}$, hence $\mu_+ = \alpha\pi_+$ and $\mu_- = \beta\pi_-$ for some probability distributions π_+, π_-, and non-negative scalars $\alpha, \beta \geq 0$. Noticing that the expectation of a bounded function f is linear in the considered probability distribution, we adopt the inner product notation for expectations:

$$\langle \pi, f \rangle := \mathbb{E}_{X\sim\pi} f(X).$$

This notation extends to finite signed-measures: given a decomposition of $\mu \in \mathfrak{M}$ as $\mu = \alpha\pi - \beta\pi'$ with π, π' two probability distributions and $\alpha, \beta \geq 0$, we denote

$$\langle \mu, f \rangle := \alpha \langle \pi, f \rangle - \beta \langle \pi', f \rangle$$

which can be checked to be independent of the particular choice of decomposition of μ. With these notations, given a class $\mathcal{F}$ of measurable functions $f : \mathcal{Z} \to \mathbb{R}$ or $\mathbb{C}$ we can define

$$\|\mu\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\langle \mu, f \rangle|$$

and check that this is a semi-norm as claimed when we introduced (9). Similarly the metric (33) can be extended to finite signed measures as

$$d_{\mathcal{F}}(\mu, \mu') := \sup_{f \in \mathcal{F}} \left| |\langle \mu, f \rangle|^2 - |\langle \mu', f \rangle|^2 \right|.$$

When the functions in $\mathcal{F}$ are smooth these quantities can be extended to tempered distributions.

The total variation norm is defined on $\mathfrak{M}$ as $\|\cdot\|_{\mathrm{TV}} = \|\cdot\|_{\mathcal{B}}$ with $\mathcal{B} = \{f : \|f\|_\infty \leq 1\}$ (see e.g. [79]) and yields a Banach structure on $\mathfrak{M}$ (see e.g. [58]).

The mean kernel κ (cf (18)) can naturally be extended from probability distributions to finite signed measures. Let $\mu_1, \mu_2 \in \mathfrak{M}$ and $\pi_1, \pi_1', \pi_2, \pi_2', \alpha_1, \alpha_2, \beta_1, \beta_2$ such that $\mu_1 = \alpha_1\pi_1 - \beta_1\pi_1'$ and $\mu_2 = \alpha_2\pi_2 - \beta_2\pi_2'$ (decompositions as differences of probability measures). Provided that $\kappa(\cdot, \cdot)$ is well defined on the corresponding probability distributions, we can define

$$\kappa(\mu_1, \mu_2) := \alpha_1\alpha_2\kappa(\pi_1, \pi_2) - \alpha_1\beta_2\kappa(\pi_1, \pi_2') - \beta_1\alpha_2\kappa(\pi_1', \pi_2) + \beta_1\beta_2\kappa(\pi_1', \pi_2') \tag{98}$$

which can be checked to be independent of the particular choices of decomposition.

By linearity of the integral and the definition of the kernel for probability distributions, we obtain a *pseudonorm* $\|\cdot\|_\kappa$ associated to the mean kernel:

$$\|\mu\|_\kappa^2 := \int\int \kappa(x, x') d\mu(x) d\mu(x') = \kappa(\mu, \mu) \tag{99}$$

that coincides with the metric of the mean kernel (19) for probability distributions.

A.3 Kullback-Leibler divergence and differential entropy

For two probability distributions π, π' that admit probability densities with respect to the Lebesgue measure, the Kullback-Leibler divergence is defined as

$$\mathrm{KL}(\pi||\pi') := \mathbb{E}_{X \sim \pi} \log \frac{\pi(X)}{\pi'(X)} \tag{100}$$

and the differential entropy is

$$\mathrm{H}(\pi) := \mathbb{E}_{X \sim \pi} - \log \pi(X). \tag{101}$$

A fundamental property of the Kullback-Leibler divergence is that $\mathrm{KL}(\pi||\pi') \geq 0$ with equality if, and only if $\pi = \pi'$. Details on these notions can be found, e.g., in [32, Chapter 9]. Pinsker's inequality [45] relates the Kullback-Leibler divergence to the total variation norm

$$\|\pi - \pi'\|_{\mathrm{TV}} \leq \sqrt{2\mathrm{KL}(\pi||\pi')}. \tag{102}$$

B Generalities on covering numbers

In this section we formulate generic results on covering numbers.

B.1 Basic properties

The definition used in this paper is that of *internal* covering numbers, meaning that the centers z_i of the covering balls are required to be included in the set Y being covered. Somehow counter-intuitively these covering numbers (for a fixed radius δ) are not necessarily increasing with the inclusion of sets: for instance, consider a set A formed by two points, included in set B which is a ball of radius δ. Suppose those two points diametrically opposed in B. We have $A \subset B$, but two balls of radius δ are required to cover A (since their centers have to be in A), while only one such ball is sufficient to cover B. Yet as shown by the following lemma the covering numbers of include sets still behave in a controlled manner.

Lemma B.1. *Let $A \subseteq B \subseteq X$ be subsets of a pseudometric set (X, d), and $\delta > 0$. Then,*

$$\mathcal{N}(d, A, \delta) \leq \mathcal{N}(d, B, \delta/2) \tag{103}$$

Proof. Let $b_1, ..., b_N$ be a δ-covering of B. We construct a δ-covering a_i of A in the following way. Each b_i is either: a) in the set A, in which case we take $a_i = b_i$; b) at distance less than $\delta/2$ of a point $a \in A$, in which case we take $a_i = a$ and note that the ball of radius δ centered on a_i covers at least as much as the ball of radius $\delta/2$ centered in b_i, i.e. $\mathcal{B}_{X,d}(b_i, \delta/2) \subset \mathcal{B}_{X,d}(a_i, \delta)$; c) in none of these cases and we discard it. There are less a_i's than b_i's, and the union of balls of radius δ with centers a_i covers A (and in fact even B): for any $a \in B$, there is an index i such that $a \in \mathcal{B}_{X,d}(b_i, \delta/2)$; by construction the corresponding ball $\mathcal{B}_{X,d}(a_i, \delta)$ also contains a. Therefore the set of a_i's is a δ-covering of B, and of A. □

Lemma B.2. *Let (X, d) and (X', d') be two pseudometric sets, and $Y \subseteq X$, $Y' \subseteq X'$. If there exists a surjective function $f : Y \to Y'$ which is L-Lipschitz with $L > 0$, i.e. such that*

$$\forall x, y \in Y,\ d'(f(x), f(y)) \leq L d(x, y),$$

then for all $\delta > 0$ we have

$$\mathcal{N}(d', Y', \delta) \leq \mathcal{N}(d, Y, \delta/L). \tag{104}$$

Proof. Define $\delta_2 = \delta/L$, denote $N = \mathcal{N}(d, Y, \delta_2)$, and let $y_i \in Y$, $i = 1, ..., N$ be a δ_2-covering of Y. Consider $y' \in Y'$. There exists $y \in Y$ such that $f(y) = y'$ since f is surjective. For some $1 \leq i \leq N$ we have $d(y, y_i) \leq \delta_2$, hence we have

$$d'(y', f(y_i)) = d'(f(y), f(y_i)) \leq L d(y, y_i) \leq L\delta_2 = \delta.$$

Thus $\{f(y_i)\}_{i=1,...,N}$ is a δ-covering of Y', and we have $\mathcal{N}(d', Y', \delta) \leq N$. □

Lemma B.3. *Let Y, Z be two subsets of a pseudometric set (X, d) such that the following holds:*

$$\forall z \in Z,\ \exists y \in Y,\ d(z, y) \leq \epsilon \tag{105}$$

where $\epsilon \geq 0$. Then for all $\delta > 0$

$$\mathcal{N}(d, Z, 2(\delta + \epsilon)) \leq \mathcal{N}(d, Y, \delta). \tag{106}$$

Proof. Denote $N = \mathcal{N}(d, Y, \delta)$ and let $y_1, ..., y_N \in Y$ be a δ-covering of Y. For all $z \in Z$, by the assumption (105) there is $y \in Y$ such that $d(z, y) \leq \epsilon$, and subsequently there is an index i such that $d(z, y_i) \leq d(z, y) + d(y, y_i) \leq \delta + \epsilon$. This implies $Z \subset \bigcup_{i=1}^{N} \mathcal{B}_{X,d}(y_i, \delta + \epsilon)$, hence by Lemma B.1

$$\mathcal{N}(d, Z, 2(\delta + \epsilon)) \leq \mathcal{N}\left(d, \bigcup_{i=1}^{N} \mathcal{B}_{X,d}(y_i, \delta + \epsilon), \delta + \epsilon\right) \leq N.$$

□

Lemma B.4 ([33], Prop. 5)**.** *Let $(X, \|\cdot\|)$ be a Banach space of finite dimension d. Then for any $x \in X$ and $R > 0$ we have for any $\delta > 0$*

$$\mathcal{N}\left(\|\cdot\|, \mathcal{B}_{X,\|\cdot\|}(x, R), \delta\right) \leq \max\left(1, \left(\frac{4R}{\delta}\right)^d\right) \tag{107}$$

NB: The result in [33, Prop. 5] does not include $\max(1, \cdot)$. This obviously cannot hold for $\delta > 4R$ since the left hand side is at least one. The proof of [33, Prop. 5] yields the result stated here.

B.2 "Extruded" Secant set

To control the covering numbers of the normalized secant set (32) of certains model sets, it will be convenient to control those of a subset which we propose to call the **extruded** normalized secant set. Considering a subset Y of a seminormed vector space $(X, \|\cdot\|)$, its extruded normalized secant set is

$$\mathcal{S}^{\eta}_{\|\cdot\|} := \left\{ \frac{y - y'}{\|y - y'\|} \;\middle|\; y, y' \in Y, \;\; \|y - y'\| > \eta \right\},$$

The covering numbers of $\mathcal{S}^{\eta}_{\|\cdot\|}$ can be controlled by those of Y itself when $\eta > 0$. In the following Lemma, we deliberately control the covering numbers of a *subset* $\mathcal{S}$ of the extruded normalized secant set $\mathcal{S}^{\eta}_{\|\cdot\|_b}(Y)$ instead of $\mathcal{S}^{\eta}_{\|\cdot\|_b}(Y)$ itself, to avoid the possible subsequent use of Lemma B.1.

Lemma B.5. *Let X be a vector space and consider a subset $Y \subset X$ and two seminorms $\|\cdot\|_a, \|\cdot\|_b$ (possibly only defined on subspaces $X_a, X_b \subset X$) such that, for some constants $0 < A \leq B < \infty$,*

$$\forall y, y' \in Y, \;\; A\|y - y'\|_b \leq \|y - y'\|_a \leq B\|y - y'\|_b < \infty. \tag{108}$$

Let $\eta > 0$, and $\mathcal{S} \subset \mathcal{S}^{\eta}_{\|\cdot\|_b}$ be a subset of the extruded normalized secant set of Y. For any $\delta > 0$ we have

$$\mathcal{N}(\|\cdot\|_a, \mathcal{S}, \delta) \leq \mathcal{N}^2\left(\|\cdot\|_a, Y, \frac{\delta\eta}{4(1 + B/A)}\right). \tag{109}$$

Proof. Define the (semi)norm on Y^2:

$$\|(y_1, y_2) - (y_1', y_2')\|_a = \|y_1 - y_1'\|_a + \|y_2 - y_2'\|_a$$

and note that we have trivially $\mathcal{N}\left(\|\cdot\|_a, Y^2, \delta\right) \leq \mathcal{N}^2\left(\|\cdot\|_a, Y, \delta/2\right)$. Consider the set:

$$\mathcal{Q} = \left\{ (y_1, y_2) \in Y^2 : \|y_1 - y_2\|_b > \eta, \frac{y_1 - y_2}{\|y_1 - y_2\|_b} \in \mathcal{S} \right\} \subset Y^2. \tag{110}$$

By definition the function $f : (\mathcal{Q}, \|\cdot\|_a) \to (\mathcal{S}, \|\cdot\|_b)$ such that $f(y_1, y_2) = \frac{y_1 - y_2}{\|y_1 - y_2\|_b}$ is surjective. Let us show that f is Lipschitz continuous, and conclude with Lemma B.2. For $(y_1, y_2), (y_1', y_2') \in \mathcal{Q}$, we have

$$\begin{aligned}\|f(y_1, y_2) - f(y_1', y_2')\|_a &= \left\| \frac{y_1 - y_2}{\|y_1 - y_2\|_b} - \frac{y_1' - y_2'}{\|y_1' - y_2'\|_b} \right\|_a, \\ &\leq \left\| \frac{y_1 - y_2}{\|y_1 - y_2\|_b} - \frac{y_1' - y_2'}{\|y_1 - y_2\|_b} \right\|_a + \left\| \frac{y_1' - y_2'}{\|y_1 - y_2\|_b} - \frac{y_1' - y_2'}{\|y_1' - y_2'\|_b} \right\|_a.\end{aligned}$$

Since $\|y_1 - y_2\|_b > \eta$, the first term is bounded by

$$\frac{1}{\eta}\Big(\|y_1 - y_1'\|_a + \|y_2 - y_2'\|_a \Big) = \frac{1}{\eta} \|(y_1, y_2) - (y_1', y_2')\|_a,$$

while the second term is bounded by

$$\begin{aligned}\|y_1' - y_2'\|_a \left| \frac{1}{\|y_1 - y_2\|_b} - \frac{1}{\|y_1' - y_2'\|_b} \right| &\overset{(108)}{\leq} B \left| \frac{\|y_1' - y_2'\|_b}{\|y_1 - y_2\|_b} - 1 \right| \leq \frac{B}{\eta} \left| \|y_1' - y_2'\|_b - \|y_1 - y_2\|_b \right| \\ &\leq \frac{B}{\eta}\Big(\|y_1 - y_1'\|_b + \|y_2 - y_2'\|_b \Big), \\ &\overset{(108)}{\leq} \frac{B}{A\eta} \left(\|y_1 - y_1'\|_a + \|y_2 - y_2'\|_a \right), \\ &= \frac{B}{A\eta} \|(y_1, y_2) - (y_1', y_2')\|_a.\end{aligned}$$

Hence we have

$$\|f(y_1, y_2) - f(y_1', y_2')\|_a \leq \frac{1 + B/A}{\eta} \|(y_1, y_2) - (y_1', y_2')\|_a.$$

The function f is Lipshitz continuous with constant $L = (1 + B/A)/\eta$, and therefore for all $\delta > 0$:

$$\mathcal{N}\left(\|\cdot\|_a, \mathcal{S}, \delta\right) \overset{\text{Lemma B.2}}{\leq} \mathcal{N}\left(\|\cdot\|_a, \mathcal{Q}, \delta/L\right) \overset{\text{Lemma B.1}}{\leq} \mathcal{N}\left(\|\cdot\|_a, Y^2, \frac{\delta}{2L}\right) \leq \mathcal{N}^2\left(\|\cdot\|_a, Y, \frac{\delta}{4L}\right).$$

□

B.3 Mixture set

Let $(X, \|\cdot\|)$ be a vector space over $\mathbb{R}$ and $Y \subset X,\ Y \neq \emptyset$. Let $k > 0$ and $\mathcal{W} \subset \mathbb{R}^k$. For $k > 0$ and a bounded set $\mathcal{W} \subset \mathbb{R}^k,\ \mathcal{W} \neq \emptyset$, denote

$$Y_{k,\mathcal{W}} = \left\{ \sum_{i=1}^{k} \alpha_i y_i : \boldsymbol{\alpha} \in \mathcal{W},\ y_i \in Y \right\} \tag{111}$$

Lemma B.6. *For all $\delta > 0$ the set $Y_{k,\mathcal{W}}$ satisfies*

$$\mathcal{N}\left(\|\cdot\|, Y_{k,\mathcal{W}}, \delta\right) \leq \min_{\tau \in]0,\ 1[} \mathcal{N}\left(\|\cdot\|_1, \mathcal{W}, \frac{(1-\tau)\delta}{\mathrm{rad}_{\|\cdot\|}(Y)}\right) \cdot \mathcal{N}^k\left(\|\cdot\|, Y, \frac{\tau\delta}{\mathrm{rad}_{\|\cdot\|_1}(\mathcal{W})}\right). \tag{112}$$

If the seminorm $\|\cdot\|$ is indeed a norm and Y and $\mathcal{W}$ are compact, then $Y_{k,\mathcal{W}}$ is also compact.

Proof. Let $\delta > 0$ and $\tau \in]0;1[$. Denote $\delta_1 = \tau\delta/\operatorname{rad}_{\|\cdot\|_1}(\mathcal{W})$ and $\delta_2 = (1-\tau)\delta/\operatorname{rad}_{\|\cdot\|}(Y)$. Also denote $N_1 = \mathcal{N}(\|\cdot\|, Y, \delta_1)$ and let $\mathcal{C}_1 = \{x_1, ..., x_{N_1}\}$ be a δ_1-covering of Y. Similarly, denote $N_2 = \mathcal{N}(\|\cdot\|_1, \mathcal{W}, \delta_2)$, let $\mathcal{C}_2 = \{\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_{N_2}\}$ be a δ_2-covering of $\mathcal{W}$. The cardinality of the set

$$Z = \left\{ \sum_{j=1}^{k} \alpha_j x_j : x_j \in \mathcal{C}_1,\ \boldsymbol{\alpha} \in \mathcal{C}_2 \right\} \tag{113}$$

is $|Z| \leq N_1^k N_2$. We will show that Z is a δ-covering of $Y_{k,\mathcal{W}}$.

Consider $y = \sum_{j=1}^{k} \alpha_j y_j \in Y_{k,\mathcal{W}}$. By definition, there is $\bar{\boldsymbol{\alpha}} \in \mathcal{C}_2$ so that $\|\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}\|_1 \leq \delta_2$, and for all $j = 1...k$, there is $\bar{y}_j \in \mathcal{C}_1$ so that $\|y_j - \bar{y}_j\| \leq \delta_1$. Denote $\bar{y} = \sum_{j=1}^{k} \bar{\alpha}_j \bar{y}_j \in Z$. We have

$$\begin{aligned}
\|y - \bar{y}\| &= \left\| \sum_{j=1}^{k} \alpha_j y_j - \sum_{j=1}^{k} \bar{\alpha}_j \bar{y}_j \right\| \leq \left\| \sum_{j=1}^{k} \alpha_j y_j - \sum_{j=1}^{k} \alpha_j \bar{y}_j \right\| + \left\| \sum_{j=1}^{k} \alpha_j \bar{y}_j - \sum_{j=1}^{k} \bar{\alpha}_j \bar{y}_j \right\| \\
&\leq \sum_{j=1}^{k} |\alpha_j| \, \|y_j - \bar{y}_j\| + \sum_{j=1}^{k} |\alpha_j - \bar{\alpha}_j| \, \|\bar{y}_j\| \\
&\leq \|\boldsymbol{\alpha}\|_1 \delta_1 + \|\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}\|_1 \operatorname{rad}_{\|\cdot\|}(Y) \leq \operatorname{rad}_{\|\cdot\|_1}(\mathcal{W}) \delta_1 + \delta_2 \operatorname{rad}_{\|\cdot\|}(Y) = \delta,
\end{aligned} \tag{114}$$

and Z is indeed a δ-covering of $Y_{k,\mathcal{W}}$. Therefore, we have the bound (for all τ)

$$\mathcal{N}(\|\cdot\|, Y_{k,\mathcal{W}}, \delta) \leq |Z| \leq N_1^k N_2.$$

Furthermore, in equation (114), we have shown in particular that the embedding $(y_1, ..., y_k, \boldsymbol{\alpha}) \to \sum_{j=1}^{k} \alpha_j y_j$ from $Y^k \times \mathcal{W}$ to $Y_{k,\mathcal{W}}$ is continuous. Hence if Y and $\mathcal{W}$ are compact $Y_{k,\mathcal{W}}$ is the continuous image of a compact set and is compact. □

C Proof of Theorem 2.9

We first prove Lemma 2.6 using Bernstein's inequality in the following simple version[77]:

Lemma C.1 (Bernstein's inequality ([77], Thm. 6))**.** *Let $X_i \in \mathbb{R}$, $i = 1, ..., N$ be i.i.d. bounded random variables such that $\mathbb{E}X_i = 0$, $|X_i| \leq M$ and $Var(X_i) \leq \sigma^2$ for all i's. Then for all $t > 0$ we have*

$$P\left(\frac{1}{N} \sum_{i=1}^{N} X_i \geq t \right) \leq \exp\left(-\frac{Nt^2}{2\sigma^2 + 2Mt/3} \right). \tag{115}$$

Proof of Lemma 2.6. Observe that

$$\frac{\|\mathcal{A}(\pi - \pi')\|_2^2}{\|\pi - \pi'\|_\kappa^2} - 1 = \tfrac{1}{m} \sum_{j=1}^{m} Y(\omega_j)$$

with

$$Y(\omega) := \frac{|\langle \pi, \phi_\omega \rangle - \langle \pi', \phi_\omega \rangle|^2}{\|\pi - \pi'\|_\kappa^2} - 1$$

Observe that $\mathbb{E}_{\omega \sim \Lambda} Y(\omega) = 0$ and that $-1 \leq Y(\omega) \leq \frac{\|\pi - \pi'\|_\Phi^2}{\|\pi - \pi'\|_\kappa^2} - 1$ which implies

$$|Y(\omega)| \leq \max\left(1, \tfrac{\|\pi - \pi'\|_\Phi^2}{\|\pi - \pi'\|_\kappa^2} - 1\right) \leq \tfrac{\|\pi - \pi'\|_\Phi^2}{\|\pi - \pi'\|_\kappa^2} = W^2. \tag{116}$$

Moreover, we have

$$\begin{aligned}\mathtt{Var}_{\omega\sim\Lambda}(Y(\omega)) =&\mathtt{Var}_{\omega\sim\Lambda}\left(\frac{\left|\langle \pi-\pi',\phi_\omega\rangle\right|^2}{\|\pi-\pi'\|_\kappa^2}\right) \leq \frac{\mathbb{E}_{\omega\sim\Lambda}\left|\langle \pi-\pi',\phi_\omega\rangle\right|^4}{\|\pi-\pi'\|_\kappa^4}\\ \leq&\frac{\mathbb{E}_{\omega\sim\Lambda}\|\pi-\pi'\|_\Phi^2\cdot\left|\langle \pi-\pi',\phi_\omega\rangle\right|^2}{\|\pi-\pi'\|_\kappa^4} = \frac{\|\pi-\pi'\|_\Phi^2}{\|\pi-\pi'\|_\kappa^2} = W^2 \end{aligned} \tag{117}$$

Applying Bernstein's inequality with the independant random variables $Y(\omega)$ (Lemma C.1) we obtain for any $t>0$

$$\mathbb{P}\left(\left|\frac{\|\mathcal{A}(\pi-\pi')\|_2^2}{\|\pi-\pi'\|_\kappa^2}-1\right|\geq t\right)\leq 2\exp\left(-\frac{mt^2}{2W^2\cdot(1+t/3)}\right). \tag{118}$$

□

Lemma C.2. *Let κ and (Λ,Φ) with concentration function $c_\kappa(t)$. Consider m parameters $(\omega_j)_{j=1}^m$ drawn i.i.d. according to Λ and the sketching operator*

$$\mathcal{A}(\pi) := \tfrac{1}{\sqrt{m}}\left[\langle \pi,\phi_{\omega_j}\rangle\right]_{j=1}^m. \tag{119}$$

Let $\mathcal{S}$ be a subset of the normalized secant set $\mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S})$. For any $\delta>0$ such that

$$N := \mathcal{N}\left(d_\Phi,\mathcal{S},\delta/2\right)<\infty, \tag{120}$$

we have, with probability at least $1-2N\exp(-m/c_\kappa(\delta/2))$:

$$\sup_{\mu\in\mathcal{S}}\left|\|\mathcal{A}(\mu)\|_2^2-1\right|\leq\delta. \tag{121}$$

Proof of Lemma C.2. Consider $\mu=(\pi-\pi')/\|\pi-\pi'\|_\kappa$ with $\pi,\pi'\in\mathfrak{S}$. By definition of the concentration function, for any $t>0$ and $m\geq 1$

$$\mathbb{P}\left(\left|\|\mathcal{A}(\mu)\|_2^2-1\right|\geq t\right)\leq 2\exp\left(-m/c_\kappa(t)\right). \tag{122}$$

This establishes a pointwise concentration result when μ is on the normalized secant set $\mathcal{S}_{\|\cdot\|_\kappa}$. We now use a standard argument to extend this to a uniform result on $\mathcal{S}$. Let μ_i, $1\leq i\leq N$ be the centers of a $\delta/2$-covering (with respect to the metric d_Φ) of $\mathcal{S}$. Using (122) with $t=\delta/2$, the probability that there is an index i such that $\left|\|\mathcal{A}(\mu_i)\|_2^2-1\right|\geq\delta/2$ is at most $\zeta=2N\exp(-m/c_\kappa(\delta/2))$. Hence, with probability at least $1-\zeta$, we have: for any $\mu\in\mathcal{S}$, with i an index chosen so that $d_\Phi(\mu,\mu_i)\leq\delta/2$:

$$\begin{aligned}\left|\|\mathcal{A}(\mu)\|_2^2-1\right| &\leq \left|\|\mathcal{A}(\mu)\|_2^2-\|\mathcal{A}(\mu_i)\|_2^2\right|+\left|\|\mathcal{A}(\mu_i)\|_2^2-1\right|\\ &\leq \frac{1}{m}\left|\sum_{j=1}^m\left(\left|\langle\mu,\phi_{\omega_j}\rangle\right|^2-\left|\langle\mu_i,\phi_{\omega_j}\rangle\right|^2\right)\right|+\delta/2\\ &\leq d_\Phi(\mu,\mu_i)+\delta/2\leq\delta.\end{aligned}$$

□

Proof Theorem 2.9. Denote $\zeta=2N\exp(-m/c_\kappa(\delta/2))$. By Lemma C.2, the assumptions imply that with probability at least $1-\zeta$ on the draw of ω_j, $1\leq j\leq m$, we have

$$\inf_{\mu\in\mathcal{S}_{\|\cdot\|_\kappa}}\|\mathcal{A}(\mu)\|_2^2\geq 1-\delta.$$

This implies (24) and since $C_\kappa<\infty$, the LRIP (12) holds with respect to $\|\cdot\|_{\mathcal{L}(\mathcal{H})}$ and constant $C_\mathcal{A}=\frac{C_\kappa}{\sqrt{1-\delta}}$. As a result the ideal decoder (13) satisfies the instance optimality property (11) with the distance (14), yielding (15). Expliciting and specializing to π_0 and $\hat{\pi}_n$ yields the result. □

D Proof of Lemma 2.10

The proof exploits transport through connections between the considered norms and the norm $\|\cdot\|_{\mathrm{Lip}(L,d)} = L \cdot \|\cdot\|_{\mathrm{Lip}(1,d)}$, where $\mathrm{Lip}(L,d)$ denotes the class of functions $f : (\mathcal{Z}, d) \to \mathbb{R}$ that are L-Lipschitz.

For $p = 1$ we have $\mathcal{L}(\mathcal{H}) \subset \mathrm{Lip}(1,d)$ since by the triangle inequality $|\ell(x,\mathcal{E}_h) - \ell(x',\mathcal{E}_h)| \le d(x,x')$ for any x, x'. For $p \ge 1$, since $|a^p - b^p| \le \max(pa^{p-1}, pb^{p-1})\,|a-b|$ we have $|\ell(x,\mathcal{E}_h) - \ell(x',\mathcal{E}_h)| \le pB^{p-1}d(x,x')$ hence $\mathcal{L}(\mathcal{H}) \subset \mathrm{Lip}(pB^{p-1}, d)$. This implies that for any π, π' we have

$$\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \le pB^{p-1} \|\pi - \pi'\|_{\mathrm{Lip}(1,d)} .$$

Next, we have

$$\|\mathcal{A}(\pi - \pi')\|_2 = \sup_{\|\mathbf{u}\|_2 \le 1} |\langle \mathcal{A}(\pi - \pi'), \mathbf{u}\rangle| = \sup_{\|\mathbf{u}\|_2 \le 1} |E_{X\sim\pi} f_{\mathbf{u}}(X) - E_{X\sim\pi'} f_{\mathbf{u}}(X)|$$

where $f_{\mathbf{u}}(x) := \langle \Phi(x), \mathbf{u}\rangle$. On the other hand, for $\|\mathbf{u}\|_2 \le 1$ and any x, x':

$$|f_{\mathbf{u}}(x) - f_{\mathbf{u}}(x')|^2 = \langle \Phi(x) - \Phi(x'), \mathbf{u}\rangle^2 \le \|\Phi(x) - \Phi(x')\|_2^2 \le L^2 d^2(x,x')$$

i.e., $f_{\mathbf{u}}(\cdot)$ is L-Lipschitz with respect to $d(\cdot,\cdot)$. It follows that for any π, π', $\|\mathcal{A}(\pi - \pi')\|_2 \le L \|\pi - \pi'\|_{\mathrm{Lip}(1,d)}$.

Gathering the above bounds we have for any π_0 and $\sigma \in \mathfrak{S}_{\mathcal{H}}$,

$$\|\pi_0 - \sigma\|_{\mathcal{L}(\mathcal{H})} + 2C_{\mathcal{A}} \|\mathcal{A}(\pi_0 - \sigma)\|_2 \le (pB^{p-1} + 2C_{\mathcal{A}} L) \cdot \|\pi_0 - \sigma\|_{\mathrm{Lip}(1,d)} .$$

It is well-known that the Wasserstein distance between two distributions can be equivalently defined in terms of transport (so-called "earth mover's distance") but also as

$$\|\pi - \pi'\|_{\mathrm{Wasserstein}_1(d)} = \|\pi - \pi'\|_{\mathrm{Lip}(1,d)}$$

as soon as $(\mathcal{Z}, d)$ is a separable metric space, see, e.g., [40, Theorem 11.8.2]. For a given π_0, let $\sigma \in \mathfrak{S}_{\mathcal{H}}$ be the distribution of $P_{\mathcal{E}_h} X$, where $X \sim \pi_0$ and $P_V x \in \arg\min_{y \in V} d(x,y)$ is the projection[9] onto $\mathcal{E}_h$. By the transport characterization of the Wasserstein distance, considering the transport plan that sends x to $P_{\mathcal{E}_h} x$, we conclude

$$\|\pi_0 - \sigma\|_{\mathrm{Wasserstein}_1(d)} \le \mathbb{E}_{X\sim\pi} d(X, P_{\mathcal{E}_h}(X)) = \mathbb{E}_{X\sim\pi} d(X, \mathcal{E}_h).$$

This yields the result for $p = 1$. The result for $1 \le p < \infty$ is a consequence of Jensen's inequality

$$\mathbb{E}_{X\sim\pi} d(X, \mathcal{E}_h) \le [\mathbb{E}_{X\sim\pi} d(X, \mathcal{E}_h)^p]^{\frac{1}{p}} .$$

E Proofs on mixtures of distributions

We gather here all proofs related to results stated in Section 6.

E.1 Proof of Lemma 6.8

Consider $\sigma^2 \le 1/2$. An easy function study of $h(t) := (1 - t/2)\exp(\frac{t}{2\sigma^2})$ shows that h is non-decreasing on $[0,1]$ with $h(0) = 1$, implying that $1 - u^2/2 \ge K(u)$ for $0 \le u \le 1$. This verifies (i) in the definition of $\mathcal{E}(A,B,C,c)$ (Def. 6.6), with $c = 1$.

[9] Ties are breaked arbitrarily in the argmin, and if needed the proof can we adapted with $P_V x$ a $(1+\epsilon)$-minimizer.

By an easy study of K'', K' is negative and increasing for $u^2 \geq \sigma^2$. Thus, considering $\sigma^2 \leq 1$, $|K'(u)|$ is decreasing for $u \geq 1$ and we can set $B = |K'(1)| = \exp(-\frac{1}{2\sigma^2})/\sigma^2$. Since $B > A = K(1)$ we have $\max(A, 2(C+B)) = 2(C+B)$ and the condition (84) reads $2(B+C) \leq \frac{3}{64(2k-1)}$.

Similarly, an easy study of $K^{(3)}$ shows that K'' is positive and decreasing for $u^2 \geq 3\sigma^2$. Considering $\sigma^2 \leq 1/3$, K'' is positive decreasing for $u \geq 1$ and we can set $C = K''(1) = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2} - 1\right)\exp(-\frac{1}{2\sigma^2})$. As a result $(B+C) = \exp(-1/2\sigma^2)/\sigma^4$ and for $\sigma^2 \leq 1/3$ the condition $2(B+C) \leq 3/(64(2k-1))$ reads as:

$$\exp(-\tfrac{1}{2\sigma^2})/\sigma^4 \leq 3/(128(2k-1)). \tag{123}$$

As the left hand side is a increasing function of σ when $\sigma^2 \leq 1/2$, and as σ_k defined by (86) satisfies $\sigma_k^2 \leq 1/3$, the result will be established provided we show that this σ_k satisfies (123) or equivalently that $\sigma_k^4 e^{1/2\sigma_k^2} \geq 128(2k-1)/3$.

To show this, we write $\sigma_k^2 = \frac{1}{2}(a\ln(2k-1)+b)^{-1}$ and rewrite the desired property as

$$\frac{(2k-1)^a e^b}{4(a\ln(2k-1)+b)^2} \geq 128(2k-1)/3$$
$$\frac{(2k-1)^{a-1}}{(a\ln(2k-1)+b)^2} \geq 512e^{-b}/3$$

Consider $f(t) := \ln\left(t^{a-1}/(a\ln t + b)^2\right) = (a-1)\ln t - 2\ln(a\ln t + b)$. A quick function study shows that its derivative is positive if $\ln t \geq 2/(a-1) - b/a$. As soon as $2/(a-1) - b/a \leq 0$, i.e.,

$$a \geq \frac{b}{b-2}, \tag{124}$$

the function f is therefore increasing for $t \geq 1$, with a minimum at $t = 1$, $f(1) = 1/b^2$, and the desired property holds if and only if $1/b^2 \geq 512e^{-b}/3$, i.e.,

$$b - 2\ln b - \ln\frac{512}{3} \geq 0.$$

The latter holds true for $b = 12$, and (124) holds for $a \geq b/(b-2) = 1.2$, which proves the result.

E.2 Proof of Lemma 6.9

To prove, Lemma 6.9, we will need the following intermediary results.

Lemma E.1. *Assume $h : \mathbb{R}_+ \to \mathbb{R}$ is differentiable and that $h'(t)$ is C-Lipschitz. Then for any $x, y \geq 0$:*

$$|h(0) - h(x) - h(y) + h(x+y)| \leq 2xyC\,.$$

Proof. Assume without loss of generality that $x = \min(x, y)$. Write $h(x) - h(0) = h'(c_1)x$ for some $c_1 \in [0, x]$ and $h(x+y) - h(y) = h'(c_2)x$ for some $c_2 \in [y, x+y]$, thus

$$|h(0) - h(x) - h(y) + h(x+y)| = |x(h'(c_2) - h'(c_1))| \leq Cx\,|c_2 - c_1|\,,$$

bounded in absolute value by $2xyC$, since $|c_1 - c_2| \leq x + y \leq 2y$. □

Lemma E.2. *Let $\nu = \pi_{\theta_1} - \pi_{\theta_2}$ and $\nu' = \pi_{\theta_3} - \pi_{\theta_4}$ be two dipoles (with unit coefficients) that are 1-separated, denote $d_{ij} = \varrho(\theta_i, \theta_j)$. Let $K \in \mathcal{E}(A, B, C, c)$. Then we have:*

$$K(d_{13}) - K(d_{23}) - K(d_{14}) + K(d_{24}) \leq 2(B+C)d_{12}d_{34} \tag{125}$$

Proof. Assume without loss of generality that $d_{13} = \min(d_{13}, d_{23}, d_{14}, d_{24})$ and write

$$\begin{aligned}|K(d_{13}) - K(d_{23}) - K(d_{14}) + K(d_{24})| \leq\ & |K(d_{13}) - K(d_{23}) - K(d_{14}) + K(d_{23} + d_{14} - d_{13})| \\ & + |K(d_{24}) - K(d_{23} + d_{14} - d_{13})|\,. \end{aligned} \tag{126}$$

To bound the first term of the right hand side of (126), since we assumed without loss of generality that $d_{13} = \min(d_{13}, d_{23}, d_{14}, d_{24})$, we can apply Lemma E.1 with $h(t) := K(d_{13}+t)$, $x := d_{23} - d_{13} \geq 0$, $y := d_{14} - d_{13} \geq 0$, leading to

$$|K(d_{13}) - K(d_{23}) - K(d_{14}) + K(d_{23} + d_{14} - d_{13})| \leq 2C\,|(d_{23} - d_{13})(d_{14} - d_{13})| \leq 2Cd_{12}d_{34}\,.$$

To bound the second term in (126), let $g(u) := K(\sqrt{u})$ and note that $g'(u) = K'(\sqrt{u})/2\sqrt{u}$. Since $K \in \mathcal{E}(A, B, C, c)$, we have $g'(u^2) \leq B/2$ for $u \geq 1$. By the separation assumption we have $1 \leq d_{23} \leq d_{23} + d_{14} - d_{13}$ and $1 \leq d_{24}$. We write

$$K(d_{24}) - K(d_{23} + d_{14} - d_{13}) = g(d_{24}^2) - g((d_{23} + d_{14} - d_{13})^2) \leq \tfrac{B}{2}\left|d_{24}^2 - (d_{23} + d_{14} - d_{13})^2\right|$$

where the last inequality follows from Rolle's theorem Now, it holds

$$d_{24}^2 - (d_{23} + d_{14} - d_{13})^2 = d_{24}^2 - d_{23}^2 - d_{14}^2 + d_{13}^2 - 2(d_{13} - d_{23})(d_{13} - d_{14})\,,$$

and by the reversed triangle inequality $|d_{ij} - d_{il}| \leq d_{jl}$ for any i, j, l so that the last product is bounded in absolute value by $2d_{12}d_{34}$. It is also easy to check by expanding the squared norms $d_{ij}^2 = \|\psi(\theta_i) - \psi(\theta_j)\|_2^2$ that

$$\left|d_{24}^2 - d_{23}^2 - d_{14}^2 + d_{13}^2\right| = 2\,|\langle \psi(\theta_1) - \psi(\theta_2), \psi(\theta_3) - \psi(\theta_4)\rangle| \leq 2d_{12}d_{34}\,.$$

Gathering everything we get the desired result. □

We can now prove Lemma 6.9.

Proof of Lemma 6.9. Denote $\nu = \alpha_1\pi_{\theta_1} - \alpha_2\pi_{\theta_2}$ and $\nu' = \alpha_3\pi_{\theta_3} - \alpha_4\pi_{\theta_4}$ two dipoles that are 1-separated, and without lost of generality suppose that $\alpha_1 = \alpha_3 = 1$, $\alpha_2 = a \leq 1$, $\alpha_4 = b \leq 1$. Our goal is to prove that $\frac{|\kappa(\nu,\nu')|}{\|\nu\|_\kappa \|\nu'\|_\kappa}$ is bounded. Denote $d_{ij} = \varrho(\theta_i, \theta_j)$ and $K_{ij} = K(d_{ij}) = \kappa(\pi_{\theta_i}, \pi_{\theta_j})$. We have

$$\begin{aligned}\frac{|\kappa(\nu,\nu')|}{\|\nu\|_\kappa \|\nu'\|_\kappa} =& \frac{|K_{13} - aK_{23} - bK_{14} + abK_{24}|}{\sqrt{1 - 2aK_{12} + a^2}\sqrt{1 - 2bK_{34} + b^2}} \\ \leq& \frac{|K_{13} - K_{23} - K_{14} + K_{24}| + |(1-a)(K_{23} - K_{24})| + |(1-b)(K_{14} - K_{24})| + |(a-1)(b-1)K_{24}|}{\sqrt{(1-a)^2 + 2a(1-K_{12})}\sqrt{(1-b)^2 + 2b(1-K_{34})}}\end{aligned}$$

Applying Lemma E.2 we get:

$$|K_{13} - K_{23} - K_{14} + K_{24}| \leq 2(B + C)d_{12}d_{34}$$

and by Rolle's theorem and the assumption on K we have as well as

$$\begin{aligned} |K_{23} - K_{24}| &\leq Bd_{34} \quad (\text{since } d_{23} \geq 1 \text{ and } d_{24} \geq 1) \\ |K_{14} - K_{24}| &\leq Bd_{12} \quad (\text{since } d_{14} \geq 1 \text{ and } d_{24} \geq 1) \\ |K_{24}| &\leq A \quad (\text{since } d_{24} \geq 1) \\ 2(1 - K_{12}) &\geq cd_{12}^2 \quad (\text{since } d_{12} \leq 1) \\ 2(1 - K_{34}) &\geq cd_{34}^2 \quad (\text{since } d_{34} \leq 1) \end{aligned}$$

Therefore, denoting $D := \max(2(B+C), A)$ and $g(x,y) := \frac{x+y}{\sqrt{x^2+(1-x)y^2}}$ for $0 \leq x, y \leq 1$, we have

$$\begin{aligned}
\frac{|\kappa(\nu,\nu')|}{\|\nu\|_\kappa \|\nu'\|_\kappa} &\leq D \cdot \frac{d_{12}d_{34} + (1-a)d_{34} + (1-b)d_{12} + (1-a)(1-b)}{\sqrt{(1-a)^2 + acd_{12}^2}\sqrt{(1-b)^2 + bcd_{34}^2}} \\
&= D \cdot \frac{d_{12}+1-a}{\sqrt{(1-a)^2+acd_{12}^2}} \cdot \frac{d_{34}+1-b}{\sqrt{(1-b)^2+bcd_{34}^2}} \\
&\leq D \cdot \frac{d_{12}+1-a}{\sqrt{\min(c,1)}\sqrt{(1-a)^2+ad_{12}^2}} \cdot \frac{d_{34}+1-b}{\sqrt{\min(c,1)}\sqrt{(1-b)^2+bd_{34}^2}} \\
&= \frac{D}{\min(c,1)} \cdot g(1-a, d_{12}) g(1-b, d_{34}).
\end{aligned}$$

As we have for any $0 \leq x, y \leq 1$

$$\begin{aligned}
g(x,y) = \frac{x+y}{\sqrt{x^2+(1-x)y^2}} &\leq \sqrt{2} \cdot \frac{x+y}{x+\sqrt{(1-x)y}} \leq \sqrt{2} \cdot \frac{x+y}{x+(1-x)y} \\
&= \sqrt{2}\left(1 + \frac{xy}{x+y-xy}\right) = \sqrt{2}\left(1 + \frac{1}{1/y+1/x-1}\right) \leq 2\sqrt{2},
\end{aligned}$$

gathering everything, we obtain

$$\frac{|\kappa(\nu,\nu')|}{\|\nu\|_\kappa \|\nu'\|_\kappa} \leq \frac{8D}{\min(c,1)}.$$

The result for mixtures of k mutually separated dipoles is a simple application of Gersgorin's disc lemma, see e.g. [50, Theorem 5.3]. □

E.3 Proof of Lemma 6.11

Let $\nu = \alpha_1 \pi_{\theta_1} - \alpha_2 \pi_{\theta_2}$ be a 1-dipole, with $\varrho(\theta_1, \theta_2) \leq 1$, and denote $\boldsymbol{\alpha} := [\alpha_1,\ \alpha_2]^T$ and $\mathbf{K}$ the 2×2 matrix with entries $K_{ij} := \kappa(\pi_{\theta_i}, \pi_{\theta_j}) = K(d_{ij})$ where $d_{ij} = \varrho(\theta_i, \theta_j)$.

Consider any $f \in \mathcal{BL}(D, L)$ and denote $\mathbf{F}$ the 2×2 matrix with entries $F_{ij} := f_i \overline{f_j}$ where $f_i := \mathbb{E}_{X \sim \pi_{\theta_i}} f(X)$. By the assumptions on f we have $|f_i| \leq D$ and $|f_1 - f_2| \leq L d_{12}$. For any $W \geq 0$ we have

$$W^2 \|\nu\|_\kappa^2 - |\langle \nu, f \rangle|^2 = \boldsymbol{\alpha}^T (W^2 \mathbf{K} - \mathbf{F}) \boldsymbol{\alpha}.$$

Therefore it is sufficient to prove that there is some real value W, that does not depend on the choice of function $f \in \mathcal{BL}(D, L)$, such that the matrix $\mathbf{Q} := W^2 \mathbf{K} - \mathbf{F}$ is positive semi-definite. It is the case if its trace and determinant are non-negative. We have $\operatorname{tr}(\mathbf{Q}) = 2W^2 - |f_1|^2 - |f_2|^2 \geq 2(W^2 - D^2)$. A sufficient condition for $\operatorname{tr}(\mathbf{Q}) \geq 0$ is therefore

$$W \geq D \tag{127}$$

We further have:

$$\begin{aligned}
\det(\mathbf{Q}) &= \left(W^2 - |f_1|^2\right)\left(W^2 - |f_2|^2\right) - \left|W^2 K(d_{12}) - f_1 \overline{f_2}\right|^2 \\
&= W^4 - W^2\left(|f_1|^2 + |f_2|^2\right) + |f_1|^2 |f_2|^2 - \left(W^2 K(d_{12}) - \operatorname{Re}\left(f_1 \overline{f_2}\right)\right)^2 - \left(\operatorname{Im}\left(f_1 \overline{f_2}\right)\right)^2.
\end{aligned}$$

Using $|f_1|^2 |f_2|^2 = \frac{1}{4}\left(\left(|f_1|^2 + |f_2|^2\right)^2 - \left(|f_1|^2 - |f_2|^2\right)^2\right)$, we get

$$\begin{aligned}
\det(\mathbf{Q}) = &\left(W^2 - \tfrac{1}{2}\left(|f_1|^2 + |f_2|^2\right)\right)^2 - \left(W^2 K(d_{12}) - \operatorname{Re}\left(f_1 \overline{f_2}\right)\right)^2 \\
&- \tfrac{1}{4}\left[\left(|f_1|^2 - |f_2|^2\right)^2 + 4\left(\operatorname{Im}\left(f_1 \overline{f_2}\right)\right)^2\right].
\end{aligned}$$

On the one hand, we have

$$\begin{aligned}\left(W^2-\tfrac{1}{2}\left(|f_1|^2+|f_2|^2\right)\right)^2-\left(W^2K(d_{12})-\mathrm{Re}\left(f_1\overline{f_2}\right)\right)^2 &= \left(W^2-\tfrac{1}{2}\left(|f_1|^2+|f_2|^2\right)-W^2K(d_{12})+\mathrm{Re}\left(f_1\overline{f_2}\right)\right)\\ &\quad\times\left(W^2-\tfrac{1}{2}\left(|f_1|^2+|f_2|^2\right)+W^2K(d_{12})-\mathrm{Re}\left(f_1\overline{f_2}\right)\right)\\ &\overset{(a)}{\geq}\left(W^2\left(1-K(d_{12})\right)-\tfrac{1}{2}|f_1-f_2|^2\right)\left(W^2-2D^2\right)\\ &\overset{(b)}{\geq} d_{12}^2\left(W^2c-L^2\right)\left(\tfrac{W^2}{2}-D^2\right)\end{aligned}$$

where in (a) we used that $K(\cdot)\geq 0$ and $|f_i|\leq D$, and in (b) that $K(d_{12})\leq 1-c\frac{d_{12}^2}{2}$ and $d_{12}\leq 1$, and $|f_1-f_2|\leq Ld_{12}$. On the other hand,

$$\begin{aligned}\left(|f_1|^2-|f_2|^2\right)^2+4\left(\mathrm{Im}\left(f_1\overline{f_2}\right)\right)^2 &= \left(|f_1|^2-|f_2|^2\right)^2+4|f_1|^2|f_2|^2-4\left(\mathrm{Re}\left(f_1\overline{f_2}\right)\right)^2\\ &= \left(|f_1|^2+|f_2|^2\right)^2-4\left(\mathrm{Re}\left(f_1\overline{f_2}\right)\right)^2\\ &= \left(|f_1|^2+|f_2|^2-2\mathrm{Re}\left(f_1\overline{f_2}\right)\right)\left(|f_1|^2+|f_2|^2+2\mathrm{Re}\left(f_1\overline{f_2}\right)\right)\\ &= |f_1-f_2|^2|f_1+f_2|^2\leq 4D^2L^2d_{12}^2\end{aligned}$$

since $|f_1+f_2|\leq 2D$ and $|f_1-f_2|\leq Ld_{12}$. Gathering everything, we have

$$\det(\mathbf{Q})\geq d_{12}^2\left(\left(W^2c-L^2\right)\left(\tfrac{W^2}{2}-D^2\right)-D^2L^2\right)=d_{12}^2\tfrac{W^2c}{2}\left(W^2-L^2/c-2D^2\right)$$

and therefore it is sufficient that

$$W\geq\left(L^2/c+2D^2\right)^{\frac{1}{2}} \tag{128}$$

We conclude by observing that (128) implies (127).

E.4 Proof of Theorem 6.13

First, we rely on the characterization of the normalized secant set as a mixture of dipoles.

Lemma E.3. *Consider a function $K(\cdot)\in\mathcal{E}(A,B,C,c)$. For any kernel $\kappa(x,x')$ and any set $\mathcal{T}=\{\pi_\theta:\theta\in\mathbf{\Theta}\}$ such that the Mean Map Embedding* (18) *satisfies* (83) *with some metric ϱ of the form* (81)*, with the mutual coherence M defined in* (87) *we have: if $M(2k-1)<1$, then the normalized secant of the set $\mathfrak{S}_{k,2}(\mathcal{T})$ of 2-separated mixtures is made of mixtures of $2k$ normalized dipoles*

$$\mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S}_{k,2}(\mathcal{T}))\subset\left\{\sum_{l=1}^{2k}\alpha_l\mu_l:\|\boldsymbol{\alpha}\|_2\leq(1-M(2k-1))^{-1/2},\ \mu_l\in\mathcal{D}\right\} \tag{129}$$

Proof. By definition any $\mu\in\mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S}_{k,2}(\mathcal{T}))$ can be written as $\mu=(\pi-\pi')/\|\pi-\pi'\|_\kappa$ with $\pi,\pi'\in\mathfrak{S}_{k,2}(\mathcal{T})$. By Lemma 6.5 we have $\pi-\pi'=\sum_{l=1}^{\ell}\nu_l$ where the ν_l are non-zero dipoles that are 1-separated from one another and $\ell\leq 2k$. With $\alpha_l:=\frac{\|\nu_l\|_\kappa}{\left\|\sum_{i=1}^{l}\nu_i\right\|_\kappa}$, $\mu_l:=\nu_l/\|\nu_l\|_\kappa$ we can write

$$\mu=\frac{\sum_{l=1}^{\ell}\nu_l}{\left\|\sum_{l=1}^{\ell}\nu_l\right\|_\kappa}=\sum_{l=1}^{\ell}\frac{\|\nu_l\|_\kappa}{\left\|\sum_{l=1}^{\ell}\nu_l\right\|_\kappa}\cdot\frac{\nu_l}{\|\nu_l\|_\kappa}=\sum_{l=1}^{\ell}\alpha_l\cdot\mu_l$$

By construction $\mu_l \in \mathcal{D}$, and by Lemma 6.9 we have $|\kappa(\mu_l, \mu_{l'})| \leq M$ for $l \neq l'$ and

$$\|\boldsymbol{\alpha}\|_2^2 = \sum_{i=1}^{l} \alpha_i^2 = \frac{\sum_{i=1}^{l} \|\nu_i\|_\kappa^2}{\left\|\sum_{i=1}^{l} \nu_i\right\|_\kappa^2} \leq \frac{1}{1 - M \cdot (2k-1)}.$$

If needed, we add to μ arbitrary normalized dipoles μ_l with $\alpha_l = 0$ for $l = \ell + 1 \dots 2k$. □

We now begin the proof of Theorem 6.13 by establishing a few properties. By the assumption $K(\cdot) \in \mathcal{E}_k(c)$, we have (84) which reads $M(2k-1) < 3/4$, hence any $\boldsymbol{\alpha} \in \mathbb{R}^{2k}$ such that $\|\boldsymbol{\alpha}\|_2 \leq (1 - M(2k-1))^{-1/2}$ satisfies $\|\boldsymbol{\alpha}\|_1 \leq \sqrt{2k}\,\|\boldsymbol{\alpha}\|_2 \leq 2\sqrt{2k}$. By Lemma E.3 we thus have the inclusion

$$\mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S}_{k,2}(\mathcal{T})) \subset [\mathcal{D}]_{2k,\mathcal{B}}$$

where we recall that $[Y]_{k,\mathcal{W}}$ is the set of k-mixtures of elements in Y with weights in $\mathcal{W}$ (see (111)), and here $\mathcal{B} := \mathcal{B}_{\mathbb{R}^{2k},\|\cdot\|_1}(0, 2\sqrt{2k})$ is the closed ball of center 0 and radius $2\sqrt{2k}$ with respect to $\|\cdot\|_1$ in $\mathbb{R}^{2k}$. Furthermore, by Lemma 6.11 and the assumption that $\Phi \subset \mathcal{BL}(D, L, \boldsymbol{\Theta}, \varrho)$ we have

$$R = \mathrm{rad}_{\|\cdot\|_1}(\mathcal{B}) \quad = \quad 2\sqrt{2k} \tag{130}$$

$$\mathrm{rad}_{\|\cdot\|_\Phi}(\mathcal{D}) \quad := \quad \sup_{\mu \in \mathcal{D}_0} \|\mu\|_\Phi = \sup_{\nu \text{ dipole}} \frac{\|\nu\|_\Phi}{\|\nu\|_\kappa} \leq W = \left(L^2/c + 2D^2\right)^{1/2} \tag{131}$$

and by Theorem 6.12 the representation (Φ, Λ) has concentration constant $W_\kappa \leq 2\sqrt{2k}W$. Then, we show that we can replace the metric $d_\Phi(\cdot,\cdot)$ by the metric $\|\cdot\|_\Phi$ using the following Lemma, which holds beyond the case of mixture models.

Lemma E.4. *Assume that the integral representation (Φ, Λ) of the kernel $\kappa(x, x')$ has finite concentration constant W_κ with respect to the model $\mathfrak{S}$. Then for any $\delta > 0$,*

$$\mathcal{N}\left(d_\Phi, \mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S}), \delta\right) \leq \mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S}), \delta/2W_\kappa\right). \tag{132}$$

Proof. By definition of W_κ, for all $\mu = (\pi - \pi')/\|\pi - \pi'\|_\kappa \in \mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S})$ and all $\phi_\omega \in \Phi$ we have

$$|\langle \mu, \phi_\omega \rangle| \leq \|\pi - \pi'\|_\Phi / \|\pi - \pi'\|_\kappa \leq W_\kappa.$$

For $\mu_i = (\pi_i - \pi_i')/\|\pi_i - \pi_i'\|_\kappa$, $i = 1, 2$ in $\mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S})$ we have

$$\begin{aligned} d_\Phi(\mu_1, \mu_2) = \sup_\omega \left| |\langle \mu_1, \phi_\omega \rangle|^2 - |\langle \mu_2, \phi_\omega \rangle|^2 \right| &= \sup_\omega \Big| |\langle \mu_1, \phi_\omega \rangle| + |\langle \mu_2, \phi_\omega \rangle| \Big| \cdot \Big| |\langle \mu_1, \phi_\omega \rangle| - |\langle \mu_2, \phi_\omega \rangle| \Big| \\ &\leq \sup_\omega 2W_\kappa \cdot |\langle \mu_1 - \mu_2, \phi_\omega \rangle| = 2W_\kappa \|\mu_1 - \mu_2\|_\Phi. \end{aligned}$$

We conclude using Lemma B.2. □

We are now ready to exploit our generic lemmas on covering numbers. For any $\delta > 0$,

$$\mathcal{N}\left(d_\Phi, \mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S}_{k,2}(\mathcal{T})), \delta\right) \overset{\text{Lemma E.4}}{\leq} \mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S}_{k,2}(\mathcal{T})), \tfrac{\delta}{2W_\kappa}\right) \tag{133}$$

$$\overset{\text{Lemma B.1}}{\leq} \mathcal{N}\left(\|\cdot\|_\Phi, [\mathcal{D}]_{2k,\mathcal{B}}, \tfrac{\delta}{4W_\kappa}\right)$$

$$\overset{\text{Lemma B.6 with } \tau=\frac{1}{2}\&(130)\&(131)}{\leq} \mathcal{N}\left(\|\cdot\|_1, \mathcal{B}, \tfrac{\delta}{8WW_\kappa}\right) \cdot \mathcal{N}^{2k}\left(\|\cdot\|_\Phi, \mathcal{D}, \tfrac{\delta}{16\sqrt{2k}W_\kappa}\right)$$

$$\overset{\text{Lemma B.4}}{\leq} \left[\max\left(1, \tfrac{64WW_\kappa\sqrt{2k}}{\delta}\right) \cdot \mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}, \tfrac{\delta}{16\sqrt{2k}W_\kappa}\right)\right]^{2k}. \tag{134}$$

We conclude by using that $W_\kappa = 2\sqrt{2k}W$.

E.5 Proof of Theorem 6.15

For a constant $\eta \geq 0$, consider the following subset of normalized dipoles

$$\mathcal{D}_\eta = \left\{ \frac{\nu}{\|\nu\|_\kappa} : \nu \text{ is a dipole, } \|\nu\|_\kappa > \eta \right\} \tag{135}$$

so that $\mathcal{D} = \mathcal{D}_0$. We bound the covering numbers of $\mathcal{D}$ by splitting it into two parts, $\mathcal{D} = \mathcal{D}_\eta \cup \mathcal{D}_\eta^c$ where $\mathcal{D}_\eta^c$ is the complementary of $\mathcal{D}_\eta^c$ in $\mathcal{D}$ for some $\eta > 0$ that we shall precise later. This yields for any $\delta > 0$

$$\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}, \delta\right) \leq \mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}_\eta, \delta\right) + \mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}_\eta^c, \delta\right) \tag{136}$$

First we establish some useful properties. Since $\Phi \subset \mathcal{BL}(D, L, \mathbf{\Theta}, \varrho)$ we have

$$\mathrm{rad}_{\|\cdot\|_\Phi}(\mathcal{T}) := \sup_{\theta \in \mathbf{\Theta}} \|\pi_\theta\|_\Phi \leq D. \tag{137}$$

and the embedding $\varphi : \mathbf{\Theta} \to \mathcal{T}$ defined as $\varphi(\theta) = \pi_\theta$ is surjective and L-Lipschitz: for any $\theta, \theta' \in \mathbf{\Theta}$

$$\|\pi_\theta - \pi_{\theta'}\|_\Phi \leq L\varrho(\theta, \theta'). \tag{138}$$

Moreover, for any $y, y' \in \mathcal{T}' := \{\alpha\pi_\theta : 0 \leq \alpha \leq 1,\ \pi_\theta \in \mathcal{T}\}$, $y - y'$ is a dipole hence by the fact that (Φ, Λ) is a representation of the kernel, and by Lemma 6.11, we have

$$\|y - y'\|_\kappa \quad \leq \quad \|y - y'\|_\Phi \leq W \|y - y'\|_\kappa \tag{139}$$

with $W := \left(L^2/c + 2D^2\right)^{1/2}$. Note that this implies $W \geq 1$.

The first term in (136) is bounded as

$$\begin{aligned}
\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}_\eta, \delta\right) &\overset{(139)\&\text{Lemma B.5}}{\leq} \mathcal{N}^2\left(\|\cdot\|_\Phi, \mathcal{T}', \tfrac{\delta\eta}{4(1+W)}\right) \overset{W\geq 1}{\leq} \mathcal{N}^2\left(\|\cdot\|_\Phi, \mathcal{T}', \tfrac{\delta\eta}{8W}\right) \\
&\overset{\text{Lemma B.6 with } Y=\mathcal{T}, \mathcal{W}=[0\ 1], k=1,\ \tau=\frac{1}{2}\ \&(137)}{\leq} \left[\mathcal{N}\left(\|\cdot\|_1, [0,\ 1], \tfrac{\delta\eta}{8WD}\right) \cdot \mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{T}, \tfrac{\delta\eta}{16W}\right)\right]^2 \\
&\overset{\text{Lemma B.4 with } \mathcal{B}_{\mathbb{R}^1, \|\cdot\|_1}(1/2,1/2)=[0,1]}{\leq} \left[\max\left(1, \tfrac{16WD}{\delta\eta}\right) \cdot \mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{T}, \tfrac{\delta\eta}{16W}\right)\right]^2 \\
&\overset{(138)\&\text{LemmaB.2}}{\leq} \left[\max\left(1, \tfrac{16WD}{\delta\eta}\right) \cdot \mathcal{N}\left(\varrho, \mathbf{\Theta}, \tfrac{\delta\eta}{16WL}\right)\right]^2.
\end{aligned} \tag{140}$$

where we used that $\mathrm{rad}_{\|\cdot\|_1}(\mathcal{W}) = 1$ for $\mathcal{W} = [0,\ 1]$.

To control the second term in (136) we use the following representation.

Lemma E.5. *Assume $\mathcal{V}$ is a tangent approximation to $\mathcal{D}$ with constants $t, T > 0$ with respect to $\|\cdot\|_\kappa$ and $\|\cdot\|$ where the kernel κ is such that the Mean Map Embedding* (18) *satisfies* (83) *with some $K(\cdot) \in \mathcal{E}_k(c)$. Consider $\eta \leq \min(3/4, t/2)$. For any nonzero element $\mu \in \mathcal{D}_\eta^c$ there exists $\nu \in \mathcal{V}$, $\theta' \in \mathbf{\Theta}$ and $0 \leq \alpha \leq 2$, $0 \leq \beta \leq 1$ such that*

$$\|\mu - \alpha\nu - \beta\pi_{\theta'}\| \leq 4T\eta. \tag{141}$$

Proof. By definition, there exists $\alpha, \alpha' \geq 0$ and $\theta, \theta' \in \mathbf{\Theta}$ with $\varrho(\theta, \theta') \leq 1$ such that $\mu = \frac{\alpha\pi_\theta - \alpha'\pi_{\theta'}}{\|\alpha\pi_\theta - \alpha'\pi_{\theta'}\|_\kappa}$. Without loss of generality, $\alpha \geq \alpha'$ and $\mu = \frac{\pi_\theta - (1-\epsilon)\pi_{\theta'}}{\|\pi_\theta - (1-\epsilon)\pi_{\theta'}\|_\kappa}$ where $0 \leq 1 - \epsilon := \alpha'/\alpha \leq 1$. Since by definition $\mu \notin \mathcal{D}_\eta$, we have $\eta_0 := \|\pi_\theta - (1-\epsilon)\pi_{\theta'}\|_\kappa \leq \eta$.

As $K_{12} := K(\varrho(\theta,\theta')) \leq 1$ we have

$$\eta_0^2 = 1 - 2(1-\epsilon)K_{12} + 1 + \epsilon^2 - 2\epsilon = 2(1-\epsilon)(1-K_{12}) + \epsilon^2$$

hence $\epsilon \leq \eta_0 \leq \eta$. Since $\eta \leq 3/4$, we have $1/(1-\eta) \leq 4$ and we further have

$$\|\pi_\theta - \pi_{\theta'}\|_\kappa^2 = 2(1-K_{12}) = \eta_0^2 \frac{2(1-K_{12})}{\eta_0^2} \leq \eta^2 \frac{2(1-K_{12})}{2(1-K_{12})(1-\epsilon)} = \frac{\eta^2}{1-\epsilon} \leq \frac{\eta^2}{1-\eta} \leq 4\eta^2.$$

Since $\eta \leq t/2$, this yields $\|\pi_\theta - \pi_{\theta'}\|_\kappa \leq 2\eta \leq t$ and by definition of a tangent approximation there is $\nu \in \mathcal{V}$ such that

$$\left\| \frac{\pi_\theta - \pi_{\theta'}}{\|\pi_\theta - \pi_{\theta'}\|_\kappa} - \nu \right\| \leq T \|\pi_\theta - \pi_{\theta'}\|_\kappa$$

Observe that $\alpha := \frac{\|\pi_\theta - \pi_{\theta'}\|_\kappa}{\|\pi_\theta - (1-\epsilon)\pi_{\theta'}\|_\kappa}$ and $\beta := \frac{\epsilon}{\|\pi_\theta - (1-\epsilon)\pi_{\theta'}\|_\kappa} = \frac{\epsilon}{\eta_0}$ satisfy $\alpha \geq 0$, $0 \leq \beta \leq 1$. Moreover we also have

$$\alpha^2 = \frac{2(1-K_{12})}{2(1-K_{12})(1-\epsilon)+\epsilon^2} \leq \frac{1}{1-\epsilon} \leq \frac{1}{1-\eta} \leq 4$$

hence $\alpha \leq 2$, and since $\beta/\alpha = \epsilon / \|\pi_\theta - \pi_{\theta'}\|_\kappa$

$$\begin{aligned} \|\mu - \alpha\nu - \beta\pi_{\theta'}\| &= \alpha \cdot \left\| \frac{\pi_\theta - (1-\epsilon)\pi_{\theta'}}{\|\pi_\theta - \pi_{\theta'}\|_\kappa} - \nu - \frac{\beta}{\alpha}\pi_{\theta'} \right\| = \alpha \cdot \left\| \frac{\pi_\theta - \pi_{\theta'}}{\|\pi_\theta - \pi_{\theta'}\|_\kappa} - \nu \right\| \\ &\leq 2T \|\pi_\theta - \pi_{\theta'}\|_\kappa \leq 4T\eta. \end{aligned}$$

□

Denote

$$\mathcal{V}' := \{\alpha\nu + \beta\pi_{\theta'} : \ \nu \in \mathcal{V},\ \theta' \in \mathbf{\Theta},\ 0 \leq \alpha \leq 2,\ 0 \leq \beta \leq 1\}.$$

Applying the above with $\|\cdot\| = \|\cdot\|_\Phi$, by Lemma B.3 with $\epsilon = 4T\eta$, $Z = \mathcal{D}$, $Y = \mathcal{V}'$ (and X a space of tempered distribution containing both Y and Z where $\|\cdot\|_\Phi$ is well defined) it follows that for any $\delta' > 0$ and $\eta \leq \min(3/4, t/2)$, we obtain

$$\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}_\eta^c, 2(\delta' + 4T\eta)\right) \leq \mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{V}', \delta'\right)$$

As $\delta \leq 16T\min(3/4, t/2)$ we have $\eta := \frac{\delta}{16T} \leq \min(3/4, t/2)$, and we get with $\delta' := 4T\eta = \delta/4$

$$\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}_{\delta/(16T)}^c, \delta\right) \leq \mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{V}', \delta/4\right) \tag{142}$$

Combining (136), (140) and (142) with $\eta := \delta/16T$ yields

$$\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}, \delta\right) \leq \mathcal{N}\left(\varrho, \mathbf{\Theta}, \tfrac{\delta^2}{C_2}\right)^2 \cdot \max\left(1, \left(\tfrac{C_3}{\delta}\right)^2\right) + \mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{V}', \tfrac{\delta}{C_4}\right)$$

with $C_2 := 256WLT$, $C_3 := 16\sqrt{WDT}$, $C_4 := 4$.

F Proof of Theorem 4.1 on Compressive Clustering

The proof of Theorem 4.1 combines Theorem 2.9 with the generic strategy of Section 6 applied the specific case of separated mixtures of Diracs. For the reader's convenience we recall that for a hypothesis $h = \{c_1, ..., c_{k_1}\} \in \mathcal{H} \subset \mathbb{R}^d$ with $k_1 \leq k$, the loss function reads $\ell(x, h) = \min_{c \in h} \|x - c\|_2^p$, where $p = 2$ for k-means and $p = 1$ for k-medians. The model set $\mathfrak{S}_{\mathcal{H}}$ is precisely the set of mixtures of Diracs with location parameters in $\mathcal{H}$.

Denote $\mathbf{\Theta} = \mathcal{B}_{\mathbb{R}^d, \|\cdot\|_2}(0, R)$ the Euclidean ball of radius R and $\mathcal{T} = \{\pi_\theta = \delta_\theta;\ \theta \in \mathbf{\Theta}\}$. With the separation assumption $\mathcal{H} \subset \mathcal{H}_{k,2\varepsilon,R}$, the model $\mathfrak{S}_{\mathcal{H}} \subset \mathfrak{S}_{k,2\varepsilon,\|\cdot\|_2}(\mathcal{T}) = \mathfrak{S}_{k,2,\varrho_\varepsilon}(\mathcal{T})$ consists of mixtures of Diracs that are 2ε-separated for the Euclidean norm, or 2-separated with the metric ϱ_ε, where

$$\varrho_\beta(\theta, \theta') := \|\theta - \theta'\|_2 / \beta. \tag{143}$$

F.1 Kernel and features

We recall that the sketching function is built with weighted random Fourier features where frequency vectors ω_j are drawn according to the distribution with probability density function given by (50), $\Lambda(\omega) = \Lambda_{w,\lambda}(\omega) = \frac{w^2(\omega)}{C_\Lambda^2} \cdot p_{\mathcal{N}(0,\lambda^2 \mathbf{I}_d)}(\omega)$ where $p_{\mathcal{N}(0,\lambda^2 \mathbf{I}_d)}(\omega)$ denotes the Gaussian pdf and

$$C_\Lambda := \sqrt{\mathbb{E}_{\omega \sim \mathcal{N}(0,\lambda^2 \mathbf{I}_d)} w^2(\omega)} \tag{144}$$

to ensure Λ is a proper probability density function.

Denoting K_σ the Gaussian kernel (85) and $\phi_\omega(x) := \frac{C_\Lambda}{w(\omega)} \cdot e^{\mathrm{J}\omega^T x}$, we have for any $\theta, \theta' \in \mathbb{R}^d$

$$\begin{aligned}
\kappa(\pi_\theta, \pi_{\theta'}) = \kappa(\theta, \theta') &= \mathbb{E}_{\omega \sim \Lambda} \phi_\omega(\theta) \overline{\phi_\omega(\theta')} = \int_{\omega \in \mathbb{R}^d} \frac{C_\Lambda^2}{w^2(\omega)} \cdot e^{\mathrm{J}\omega^T(\theta - \theta')} \cdot \frac{w^2(\omega)}{C_\Lambda^2} \cdot p_{\mathcal{N}(0,\lambda^2 \mathbf{I}_d)}(\omega) d\omega \\
&= \mathbb{E}_{\omega \sim \mathcal{N}(0,\lambda^2 \mathbf{I}_d)} e^{\mathrm{J}\omega^T(\theta - \theta')} \overset{(*)}{=} \exp\left(-\frac{\lambda^2 \|\theta - \theta'\|_2^2}{2}\right) \\
&= \exp\left(-\frac{\|(\theta - \theta')/\varepsilon\|_2^2}{2(1/\lambda\varepsilon)^2}\right) = K_{\frac{1}{\lambda\varepsilon}}(\varrho_\varepsilon(\theta, \theta'))
\end{aligned}$$

where (*) follows from the expression of the characteristic function of the Gaussian. The assumption $\varepsilon = 1/(\lambda\sigma_k)$ implies $\sigma^2 = (1/(\lambda\varepsilon))^2 = \sigma_k^2$ hence by Lemma 6.8 we have $K_{\frac{1}{\lambda\varepsilon}} \in \mathcal{E}_k(1)$. This holds for any $\lambda, w(\cdot)$ such that C_Λ is finite.

Below we control the compatibility constant, concentration constant, and covering numbers for any weight $w(\cdot)$ that furthermore satisfies

$$A_w := \sup_\omega \frac{1}{w(\omega)} < \infty; \quad B_w := \sup_\omega \frac{\|\omega\|_2}{w(\omega)} < \infty; \quad C_w := \sup_\omega \frac{\|\omega\|_2^2}{w(\omega)} < \infty. \tag{145}$$

At the end of the section we consider the specific choice of $w(\cdot)$ expressed in the Theorem.

F.2 Compatibility constant

Since the basic set is made of Diracs we have $\mathbb{E}_{x \sim \delta_\theta} f(x) = f(\theta)$ hence the notion of "bounded and Lipschitz property in expectation" boils down to standard boundedness and Lipschitz property.

Lemma F.1. *Denote $\mathcal{B} := \mathcal{B}_{\mathbb{R}^d,\|\cdot\|_2}(0,R)$. For any $\mathcal{H} \subset \mathcal{B}^k$ and $\mathbf{\Theta} \subset \mathcal{B}$, the loss class $\mathcal{L}(\mathcal{H})$ associated to k-means (resp. k-medians) satisfies for any $\beta > 0$: $\mathcal{L}(\mathcal{H}) \subset \mathcal{BL}(D_\mathcal{L}, L_\mathcal{L}, \mathbf{\Theta}, \varrho_\beta)$ with*

$$D_\mathcal{L} := (2R)^p \tag{146}$$
$$L_\mathcal{L} := \beta(4R)^{p-1} \tag{147}$$

where $p = 2$ for k-means, $p = 1$ for k-medians.

Proof. For any $\theta \in \mathbf{\Theta}$ and all $h = \{c_1, \dots, c_{k_1}\} \in \mathcal{H}$, $k_1 \leq k$, by the triangle inequality $\|\theta - c_l\|_2 \leq \|\theta\|_2 + \|c_l\|_2 \leq 2R$, we have $\ell(\theta, h) \leq (2R)^p$ where we recall that $p = 2$ for k-means and $p = 1$ for k-medians.

Consider now $h = \{c_1, \dots, c_{k_1}\} \in \mathcal{H}$, $k_1 \leq k$. Given $\theta_1, \theta_2 \in \mathbf{\Theta}$, let $l^\star$ be an index such that $\ell(\theta_2, h) = \min_{1\leq l\leq k_1} \|\theta_2 - c_l\|_2^p = \|\theta_2 - c_{l^\star}\|_2^p = \ell(\theta_2, \{c_{l^\star}\})$. By definition, $\ell(\theta_1, h) = \min_{1\leq l\leq k_1} \|\theta_1 - c_l\|_2^p \leq \|\theta_1 - c_{l^\star}\|_2^p = \ell(\theta_1, \{c_{l^\star}\})$ hence

$$\begin{aligned}\ell(\theta_1, h) - \ell(\theta_2, h) &\leq \ell(\theta_1, \{c_{l^\star}\}) - \ell(\theta_2, h) = \ell(\theta_1, \{c_{l^\star}\}) - \ell(\theta_2, \{c_{l^\star}\}) \\ &= \|\theta_1 - c_{l^\star}\|_2^p - \|\theta_2 - c_{l^\star}\|_2^p .\end{aligned}$$

For k-medians, $p = 1$ and the reversed triangle inequality further yields

$$\|\theta_1 - c_{l^\star}\|_2 - \|\theta_2 - c_{l^\star}\|_2 \leq \|\theta_1 - \theta_2\|_2 .$$

In the case of k-means, $p = 2$ and we use

$$\|\theta_1 - c_{l^\star}\|_2^2 - \|\theta_2 - c_{l^\star}\|_2^2 = (\|\theta_1 - c_{l^\star}\|_2 + \|\theta_2 - c_{l^\star}\|_2)(\|\theta_1 - c_{l^\star}\|_2 - \|\theta_2 - c_{l^\star}\|_2) \leq 4R \|\theta_1 - \theta_2\|_2 .$$

By symmetry we obtain $|\ell(\theta_1, h) - \ell(\theta_2, h)| \leq (4R)^{p-1} \|\theta_1 - \theta_2\|_2 = (4R)^{p-1} \cdot \beta \cdot \varrho_\beta(\theta_1, \theta_2)$. □

Since we consider $\mathcal{H} \subset \mathcal{H}_{k,2\varepsilon,R} \subset \mathcal{B}_{\mathbb{R}^d,\|\cdot\|_2}(0,R)$ with $0 < \varepsilon < R$, we get with $\beta = \varepsilon$: $D_\mathcal{L} = 2^p R^p$, $L_\mathcal{L} \leq 4^{p-1} R^p$. Moreover the model set $\mathfrak{S}_\mathcal{H}$ consists of 2ε-separated mixtures with respect to the Euclidean metric and therefore of 2-separated mixtures with respect to ϱ_ε. As $K_{1/\lambda\varepsilon} \in \mathcal{E}_k(1)$, we can apply Theorem 6.12 to bound the compatibility constant as

$$C_\kappa \leq 2\sqrt{2k}\sqrt{L_\mathcal{L}^2 + 2D_\mathcal{L}^2} = 8\sqrt{6k}R^p. \tag{148}$$

F.3 Concentration constant

The considered integral representation of the kernel $\kappa(x, x')$ involves the class

$$\Phi := \left\{\phi_\omega(x) := \tfrac{C_\Lambda}{w(\omega)} \cdot e^{\mathrm{J}\omega^T x}; \omega \in \mathbb{R}^d\right\}.$$

We have the following result.

Lemma F.2. *Assume $A_w < \infty$ and $B_w < \infty$. Then for any $\mathbf{\Theta} \subset \mathbb{R}^d$ we have for any $\beta > 0$: $\Phi \subset \mathcal{BL}(D_\Phi, L_\Phi, \mathbf{\Theta}, \varrho_\beta)$ with*

$$D_\Phi := A_w C_\Lambda \tag{149}$$
$$L_\Phi := \beta B_w C_\Lambda. \tag{150}$$

Note that multiplying the weights $w(\omega)$ by a constant factor rescales C_Λ accordingly, leading to unchanged features and unchanged $A_w C_\Lambda$, $B_w C_\Lambda$.

Proof. It is immediate that $\sup_\omega \sup_\theta |\phi_\omega(\theta)| = A_w C_\Lambda$ and since for any $a \le b$, $\left|e^{\jmath a} - e^{\jmath b}\right| = \left|\int_a^b \jmath e^{\jmath u} du\right| \le \int_a^b |\jmath e^{\jmath u}|\, du = b - a$, we have

$$\begin{aligned}|\phi_\omega(\theta) - \phi_\omega(\theta')| &= \frac{C_\Lambda}{w(\omega)} \cdot \left|e^{\jmath \omega^T \theta} - e^{\jmath \omega^T \theta'}\right| \le \frac{C_\Lambda}{w(\omega)} \cdot \|\omega\|_2 \|\theta - \theta'\|_2 \\ &\le B_w C_\Lambda \|\theta - \theta'\|_2 = \beta B_w C_\Lambda \varrho_\beta(\theta, \theta').\end{aligned}$$

□

With $\beta = \varepsilon$ we get the shorthand $W_\Phi := \sqrt{L_\Phi^2 + 2D_\Phi^2} = \sqrt{A_w^2 + 2\varepsilon^2 B_w^2}\, C_\Lambda$ that will be reused in several places, and we can again combine with Theorem 6.12 to bound the concentration constant as

$$W_\kappa \le 2\sqrt{2k} W_\Phi = 2\sqrt{2(A_w^2 + 2\varepsilon^2 B_w^2)k} \cdot C_\Lambda. \tag{151}$$

F.4 Covering numbers

To control the covering numbers we need first to control those of the parameter set. Since $\mathbf{\Theta} = \mathcal{B}_{\mathbb{R}^d, \|\cdot\|_2}(0, R)$ we have for all $\delta > 0$

$$\mathcal{N}(\varrho_\varepsilon, \mathbf{\Theta}, \delta) = \mathcal{N}(\|\cdot\|_2/\varepsilon, \mathbf{\Theta}, \delta) = \mathcal{N}(\|\cdot\|_2, \mathbf{\Theta}, \delta\varepsilon) \overset{\text{Lemma B.4}}{\le} \max\left(1, \left(\tfrac{4R}{\delta\varepsilon}\right)^d\right) = \max\left(1, \left(\tfrac{4R/\varepsilon}{\delta}\right)^d\right). \tag{152}$$

We now establish the existence of a tangent approximation $\mathcal{V}$ to the set of dipoles.

Consider a dipole $\mu = \pi_\theta - \pi_{\theta'}$ where by definition we have $\varrho_\varepsilon(\theta, \theta') \le 1$. Given any tempered distribution ν we have, with $f(\omega) := C_\Lambda / w(\omega)$:

$$\left\|\frac{\mu}{\|\mu\|_\kappa} - \nu\right\|_\Phi = \sup_\omega f(\omega) \cdot \left|\frac{e^{\jmath\omega^T\theta} - e^{\jmath\omega^T\theta'}}{\|\pi_\theta - \pi_{\theta'}\|_\kappa} - \psi_\nu(\omega)\right| = \sup_\omega f(\omega) \cdot \left|\frac{e^{\jmath\omega^T(\theta-\theta')} - 1}{\|\pi_\theta - \pi_{\theta'}\|_\kappa} - e^{-\jmath\omega^T\theta'}\psi_\nu(\omega)\right| \tag{153}$$

where $\psi_\nu(\omega) = \int e^{\jmath\omega^T x} d\nu(x)$ is the characteristic function of ν. We will exploit the fact that, by a Taylor expansion, for any $y \in \mathbb{R}$, $t > 0$

$$\left|\frac{e^{\jmath y t} - 1}{t} - \jmath y\right| \le \sup_{0 \le \tau \le t} \left|\frac{d^2}{d\tau^2} e^{\jmath y \tau}\right| \frac{t}{2} = y^2 \frac{t}{2}. \tag{154}$$

Setting $t := \|\pi_\theta - \pi_{\theta'}\|_\kappa$, $\Delta := (\theta - \theta')/t$, and $y := \omega^T \Delta$, we have $\omega^T(\theta - \theta') = yt$. We seek ν such that $\psi_\nu(\omega) = e^{\jmath\omega^T\theta'} \jmath \omega^T \Delta$. Denote $\delta'_{a,u}$ the derivative of the Dirac function at position a along direction u, which is defined by its action on test functions $g \mapsto \frac{d}{dt}_{|t=0} g(a + tu)$. Specializing to $g(x) := e^{\jmath\omega^T x}$ we get $\psi_{\delta_{a,u}}(\omega) = j\omega^T u \cdot e^{\jmath\omega^T a}$. Considering $\nu := \delta'_{\theta', \Delta}$, we get $\psi_\nu(\omega) = e^{\jmath\omega^T\theta'} \jmath \omega^T \Delta$.

Since $K_{\frac{1}{\lambda\varepsilon}}(\cdot) \in \mathcal{E}_k(1)$ and $\varrho_\varepsilon(\theta, \theta') \le 1$ we have $\|\pi_\theta - \pi_{\theta'}\|_\kappa^2 = 2(1 - K_{\frac{1}{\lambda\varepsilon}}(\varrho_\varepsilon(\theta, \theta'))) \ge \varrho_\varepsilon^2(\theta, \theta')$ i.e.

$$\varrho_\varepsilon(\theta, \theta') \le \|\pi_\theta - \pi_{\theta'}\|_\kappa. \tag{155}$$

By Cauchy-Schwarz, $y^2 \le \|\omega\|_2^2 \|\Delta\|_2^2$, hence

$$\begin{aligned}\sup_\omega f(\omega)\cdot\left|\frac{e^{\jmath\omega^T(\theta-\theta')}-1}{\|\pi_\theta-\pi_{\theta'}\|_\kappa}-e^{-\jmath\omega^T\theta'}\psi_\nu(\omega)\right| &= \sup_\omega f(\omega)\cdot\left|\frac{e^{\jmath\omega^T\Delta t}-1}{t}-\jmath\omega^T\Delta\right| \\ &\le \sup_\omega f(\omega)\cdot\|\omega\|_2^2\|\Delta\|_2^2\cdot\frac{t}{2} \\ &\overset{(*)}{=} C_wC_\Lambda\frac{\|\theta-\theta'\|_2^2}{2t} \\ &= C_wC_\Lambda\frac{\varepsilon^2\varrho_\varepsilon^2(\theta,\theta')}{2\|\pi_\theta-\pi_{\theta'}\|_\kappa} \overset{(155)}{\le} \tfrac{C_wC_\Lambda\varepsilon^2}{2}\|\pi_\theta-\pi_{\theta'}\|_\kappa\end{aligned}$$

where in (*) we used that $\sup_\omega f(\omega) = C_wC_\Lambda$ (recall the definition (145) of $C_w := \sup_\omega \|\omega\|_2^2/w(\omega)$).

As this holds with no constraint on $\|\pi_\theta-\pi_{\theta'}\|_\kappa$ (other than the trivial bound $\|\pi_\theta-\pi_{\theta'}\|_\kappa \le 2$), and as $\|\Delta\|_2 = \varepsilon\varrho_\varepsilon(\theta,\theta') \le \varepsilon$ we have just proved that the set of tempered distributions

$$\mathcal{V} := \left\{\nu = \delta'_{\theta',-\Delta} :\ \theta'\in\mathbf{\Theta},\ \Delta\in\mathbb{R}^d, \|\Delta\|_2\le\varepsilon\right\}$$

provides a tangent approximation to the set of dipoles, with constants $t=2$,

$$T = \varepsilon^2\sup_\omega f(\omega)/2 = \varepsilon^2 C_wC_\Lambda/2. \tag{156}$$

In order to use Theorem 6.15, we proceed to control the covering numbers (with respect to $\|\cdot\|_\Phi$) of

$$\mathcal{V}' := \left\{\alpha\delta_{\theta',-\Delta}+\beta\pi_\theta |\ \theta,\theta'\in\mathbf{\Theta},\ \Delta\in\mathbb{R}^d, \|\Delta\|_2\le\varepsilon,\ 0\le\alpha\le 2,\ 0\le\beta\le 1\right\}.$$

Since $\alpha\delta_{\theta',-\Delta} = \delta_{\theta',-\alpha\Delta}$ we can write

$$\mathcal{V}' = \left\{\delta_{\theta',\Delta}+\beta\pi_\theta |\ \theta,\theta'\in\mathbf{\Theta},\ \Delta\in\mathbb{R}^d, \|\Delta\|_2\le 2\varepsilon,\ 0\le\beta\le 1\right\}$$

Consider the product space $X := \mathbf{\Theta}^2\times\mathcal{B}_{\mathbb{R}^d,\|\cdot\|_2}(0,2\varepsilon)\times[0,\ 1]$. Given $x=(\theta_1,\theta_2,\Delta,\beta)\in X$, define the function $\varphi: X\mapsto\mathcal{V}'$ by $\varphi(x) = \delta'_{\theta_1,\Delta}+\beta\pi_{\theta_2}$. For $x=(\theta_1,\theta_2,\Delta,\beta)$ and $x'=(\theta'_1,\theta'_2,\Delta',\beta')$ in X, we have

$$\begin{aligned}\|\varphi(x)-\varphi(x')\|_\Phi &= \left\|\delta'_{\theta_1,\Delta}+\beta\pi_{\theta_2}-\delta'_{\theta'_1,\Delta'}-\beta'\pi_{\theta'_2}\right\|_\Phi \\ &\le \left\|\delta'_{\theta_1,\Delta}-\delta'_{\theta'_1,\Delta}\right\|_\Phi + \left\|\delta'_{\theta'_1,\Delta}-\delta'_{\theta'_1,\Delta'}\right\|_\Phi + \|\beta\pi_{\theta_2}-\beta'\pi_{\theta_2}\|_\Phi + \left\|\beta'\pi_{\theta_2}-\beta'\pi_{\theta'_2}\right\|_\Phi\end{aligned}$$

We bound each of these terms, with $f(\omega) := C_\Lambda/w(\omega)$:

i) We have

$$\begin{aligned}\left\|\delta'_{\theta_1,\Delta}-\delta'_{\theta'_1,\Delta}\right\|_\Phi &= \sup_\omega f(\omega)\left|\omega^T\Delta\right|\left|e^{\jmath\omega^T\theta_1}-e^{\jmath\omega^T\theta'_1}\right| \le \sup_\omega f(\omega)\left|\omega^T\Delta\right|\left|\omega^T(\theta_1-\theta'_1)\right| \\ &\le \sup_\omega f(\omega)\|\omega\|_2^2\cdot 2\varepsilon\|\theta_1-\theta'_1\|_2 = L_1\|\theta_1-\theta'_1\|_2\end{aligned}$$

where $L_1 = 2\varepsilon\sup_\omega f(\omega)\|\omega\|_2^2 = 2\varepsilon C_wC_\Lambda$.

ii) We have

$$\left\|\delta'_{\theta'_1,\Delta}-\delta'_{\theta'_1,\Delta'}\right\|_\Phi = \sup_\omega f(\omega)\left|\omega^T(\Delta-\Delta')\right| \le \sup_\omega f(\omega)\|\omega\|_2\|\Delta-\Delta'\|_2 = L_2\|\Delta-\Delta'\|_2$$

where $L_2 = \sup_\omega f(\omega)\|\omega\|_2 = B_wC_\Lambda$.

iii) We have

$$\|\beta\pi_{\theta_2} - \beta'\pi_{\theta_2}\|_\Phi \leq \|\pi_{\theta_2}\|_\Phi |\beta - \beta'| \leq L_3 |\beta - \beta'|$$

with $L_3 = D_\Phi = A_w C_\Lambda$.

iv) Finally, we have

$$\begin{aligned}\left\|\beta'\pi_{\theta_2} - \beta'\pi_{\theta_2'}\right\|_\Phi \leq \left\|\pi_{\theta_2} - \pi_{\theta_2'}\right\|_\Phi &= \sup_\omega f(\omega) \left|e^{\jmath\omega^T\theta_2} - e^{\jmath\omega^T\theta_2'}\right| \leq \sup_\omega f(\omega) \left|\omega^T(\theta_2 - \theta_2')\right| \\ &\leq \sup_\omega f(\omega) \|\omega\|_2 \|\theta_2 - \theta_2'\|_2 = L_4 \|\theta_2 - \theta_2'\|\end{aligned}$$

where $L_4 = L_2$.

Denote $\mathcal{C}_1$ a $\frac{\delta}{4L_1}$-covering of $\boldsymbol{\Theta}$, $\mathcal{C}_2$ a $\frac{\delta}{4L_2}$-covering of $\mathcal{B}_{\mathbb{R}^d,\|\cdot\|_2}(0, 2\varepsilon)$, $\mathcal{C}_3$ a $\frac{\delta}{4L_3}$-covering of $[0,\ 1]$ and $\mathcal{C}_4$ a $\frac{\delta}{4L_2}$-covering of $\boldsymbol{\Theta}$. For any $x \in X$ there exists an element $x' \in \mathcal{C}_1 \times \mathcal{C}_2 \times \mathcal{C}_3 \times \mathcal{C}_4$ such that $\|\varphi(x) - \varphi(x')\|_\Phi \leq \delta$. Thus, for any $\delta > 0$,

$$\begin{aligned}\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{V}', \delta\right) &\leq |\mathcal{C}_1| \cdot |\mathcal{C}_2| \cdot |\mathcal{C}_3| \cdot |\mathcal{C}_4| \\ &\leq \mathcal{N}\left(\|\cdot\|_2, \boldsymbol{\Theta}, \frac{\delta}{4L_1}\right) \cdot \mathcal{N}\left(\|\cdot\|_2, \boldsymbol{\Theta}, \frac{\delta}{4L_2}\right) \\ &\quad \cdot \mathcal{N}\left(\|\cdot\|_2, \mathcal{B}_{\mathbb{R}^d,\|\cdot\|_2}(0, 2\varepsilon), \frac{\delta}{4L_2}\right) \cdot \mathcal{N}\left(|\cdot|, [0,\ 1], \frac{\delta}{4L_3}\right) \\ &\overset{\text{Lemma B.4}}{\leq} \max\left(1, \left(\tfrac{16L_1R}{\delta}\right)^d\right) \cdot \max\left(1, \left(\tfrac{16L_2R}{\delta}\right)^d\right) \cdot \max\left(1, \left(\tfrac{32L_2\varepsilon}{\delta}\right)^d\right) \cdot \max\left(1, \tfrac{8L_3}{\delta}\right). \quad (157)\end{aligned}$$

F.5 Summary: establishing the LRIP

We can now establish that with high probability on the draw of frequencies, the sketching operator $\mathcal{A}$ satisfies the LRIP simultaneously for all models $\mathcal{H} \subset \mathcal{H}_{k,2\varepsilon,R}$.

Consider first $\mathcal{H} = \mathcal{H}_{k,2\varepsilon,R}$. The compatibility constant (148) with respect to the model $\mathfrak{S}_0 := \mathfrak{S}_{\mathcal{H}_{k,2\varepsilon,R}}$ is

$$C_\kappa \leq 8\sqrt{6}\sqrt{k}R^p.$$

The above results hold for any weights satisfying $A_w, B_w, C_w < \infty$ (cf the definition (145) of these constants) and $C_\Lambda < \infty$. Without loss of generality we can normalize the weights so that $A_w = 1$ (the constant C_Λ will be rescaled accordingly, and the features will not change; the products $A_w C_\Lambda$, $B_w C_\Lambda$ and $C_w C_\Lambda$ will also be unchanged). Choosing

$$w(\omega) \quad := \quad 1 + \tfrac{\|\omega\|_2^2}{\lambda^2 d} \quad (158)$$

yields $A_w = 1$, $B_w = \sqrt{\lambda^2 d}/2$, and $C_w = \lambda^2 d$. By (144), as $\mathbb{E}_{\omega\sim\mathcal{N}(0,\lambda^2\mathbf{I}_d)} \|\omega\|_2^2 = \lambda^2 d$ and $\mathbb{E}_{\omega\sim\mathcal{N}(0,\lambda^2\mathbf{I}_d)} \|\omega\|_2^4 = \lambda^4 d(d+2)$, we obtain

$$C_\Lambda = \sqrt{1 + 2\tfrac{\mathbb{E}_\omega\|\omega\|_2^2}{\lambda^2 d} + \tfrac{\mathbb{E}_\omega\|\omega\|_2^4}{\lambda^4 d^2}} = \sqrt{4 + 2/d^2} \leq \sqrt{6} = \mathcal{O}(1) \quad (159)$$

hence $A_w C_\Lambda = \sqrt{4 + 2/d^2} \leq \sqrt{6} = \mathcal{O}(1)$, $B_w C_\Lambda = \lambda\sqrt{d}\sqrt{1 + \frac{1}{2d^2}} = \mathcal{O}\left(\lambda\sqrt{d}\right)$ and $C_w C_\Lambda = \mathcal{O}\left(\lambda^2 d\right)$.

We recall that[10] $D_\Phi := A_w C_\Lambda = \mathcal{O}(1)$, $L_\Phi := \varepsilon B_w C_\Lambda = \mathcal{O}\left(\lambda\varepsilon\sqrt{d}\right) = \mathcal{O}\left(\sqrt{d\log(ek)}\right)$, $1 \leq W_\Phi := \sqrt{L_\Phi^2 + 2D_\Phi^2} = \mathcal{O}\left(\sqrt{d\log(ek)}\right)$.

The concentration constant (151) with respect to $\mathfrak{S}_0$ is thus

$$W_\kappa = \mathcal{O}\left(\sqrt{k}W\right) = \mathcal{O}\left(\sqrt{dk\log(ek)}\right).$$

Combined with (156), the assumption $\varepsilon = 1/\lambda\sigma_k$ and the fact that $1/\sigma_k^2 = \mathcal{O}(\log(ek))$ yields $T := \varepsilon^2 C_w C_\Lambda/2 = \lambda^2\varepsilon^2 d C_\Lambda/2 = \frac{dC_\Lambda}{2\sigma_k^2} = \mathcal{O}(d\log(ek))$. As $\sigma_k < 1$ and $C_\Lambda \geq 2$ we have $T \geq 1$, and as $t = 2$ we obtain $16T\min(3/4, t/2) = 12T = 6\frac{dC_\Lambda}{\sigma_k^2} \geq 1$. By Theorem 6.15 we have: for any $0 < \delta \leq 1$

$$\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}, \delta\right) \leq \mathcal{N}\left(\varrho_\varepsilon, \mathbf{\Theta}, \tfrac{\delta^2}{C_2}\right)^2 \cdot \max\left(1, \left(\tfrac{C_3}{\delta}\right)^2\right) + \mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{V}', \tfrac{\delta}{C_4}\right) \tag{160}$$

with $C_2 := 256 W_\Phi L_\Phi T = \mathcal{O}\left((d\log(ek))^2\right)$, $C_3 := 16\sqrt{W_\Phi D_\Phi T} = \mathcal{O}\left((d\log(ek))^{3/2}\right)$, $C_4 := 4$.

By (152) we have for any $0 < \delta \leq 1$, with $C \geq 1$ denoting some universal constant that may change from equation to equation,

$$\mathcal{N}\left(\varrho_\varepsilon, \mathbf{\Theta}, \delta^2/C_2\right) \leq \max\left(1, (4C_2)^d (R/\varepsilon)^d (1/\delta^2)^d\right) \leq (d\log(ek))^{2d} \left(\tfrac{R}{\varepsilon}\right)^d \left(\tfrac{C}{\delta}\right)^{2d}.$$

hence for any $0 < \delta \leq 1$

$$\left(\mathcal{N}\left(\varrho_\varepsilon, \mathbf{\Theta}, \delta^2/C_2\right)\right)^2 \cdot \max\left(1, (C_3/\delta)^2\right) \leq (d\log(ek))^{4d+3} \left(\tfrac{R}{\varepsilon}\right)^{2d} \left(\tfrac{C}{\delta}\right)^{4d+2} \tag{161}$$

Further, as

$$\begin{aligned}
L_1 R &= 2\varepsilon C_w C_\Lambda R = 2\varepsilon^2 C_w C_\Lambda \tfrac{R}{\varepsilon} = \mathcal{O}\left(\varepsilon^2\lambda^2 d\tfrac{R}{\varepsilon}\right) = \mathcal{O}\left(d\log(ek)\cdot\tfrac{R}{\varepsilon}\right); \\
L_2 R &= B_w C_\Lambda R = \varepsilon B_w C_\Lambda \tfrac{R}{\varepsilon} = \mathcal{O}\left(\varepsilon\lambda\sqrt{d}\tfrac{R}{\varepsilon}\right) = \mathcal{O}\left(\sqrt{d\log(ek)}\cdot\tfrac{R}{\varepsilon}\right) \\
L_2\varepsilon &= B_w C_\Lambda \varepsilon = \mathcal{O}\left(\sqrt{d\log(ek)}\right) \\
L_3 &= A_w C_\Lambda = \mathcal{O}(1) \qquad (\text{and } L_3 \geq 1)
\end{aligned}$$

we have $(16L_1R)(16L_2R)(32L_2\varepsilon) = \mathcal{O}\left((d\log(ek))^2\left(\tfrac{R}{\varepsilon}\right)^2\right)$. By (157) we have for any $0 < \delta \leq 1$

$$\begin{aligned}
\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{V}', \delta/C_4\right) &\leq (d\log(ek))^{2d}\left(\tfrac{R}{\varepsilon}\right)^{2d}\left(\tfrac{C}{\delta}\right)^{3d+1} \\
&= \underbrace{(d\log(ek))^{-2d-3}(\delta/C)^{d+1}}_{\leq 1 \text{ when } \delta\leq 1}(d\log(ek))^{4d+3}\left(\tfrac{R}{\varepsilon}\right)^{2d}\left(\tfrac{C}{\delta}\right)^{4d+2}
\end{aligned} \tag{162}$$

Combining (160) with (161)-(162) we get for any $0 < \delta \leq 1$

$$\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}, \delta\right) \leq (d\log(ek))^{4d+3}\left(\tfrac{R}{\varepsilon}\right)^{2d}\left(\tfrac{C}{\delta}\right)^{4d+2} \tag{163}$$

By Theorem 6.13 we have for any $\delta > 0$

$$\mathcal{N}\left(d_\Phi, \mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S}_{k,2}(\mathcal{T})), \delta\right) \leq \left[\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}, \tfrac{\delta}{C_0}\right)\cdot\max\left(1, \tfrac{C_1}{\delta}\right)\right]^{2k} \tag{164}$$

[10] NB: All logarithms are expressed in base e. We write $\log ek$ instead of $\log k$ in the $\mathcal{O}(\cdot)$ notations to cover the case $k = 1$ where $\log k = 0$ while $\log ek = 1$.

where $C_0 := 64kW_\Phi \geq 1$, $C_1 := 256kW_\Phi^2 \geq 1$.

As $C_0 = \mathcal{O}\left(k\sqrt{d\log(ek)}\right)$ and $C_1 = \mathcal{O}\left(kd\log(ek)\right)$ we have $C_0^{4d+2}\cdot C_1 = \mathcal{O}\left(k^{4d+3}(d\log(ek))^{2d+2}\right)$. Moreover, for $0<\delta\leq 1$, we have $\delta/C_1 \leq 1$ hence we can apply (163) to obtain

$$\begin{aligned}\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}, \tfrac{\delta}{C_0}\right)\cdot\max\left(1,\tfrac{C_1}{\delta}\right) &\leq (d\log(ek))^{4d+3}\left(\tfrac{R}{\varepsilon}\right)^{2d}\left(\tfrac{C}{\delta}\right)^{4d+2}\cdot C_0^{4d+2}\cdot C_1/\delta \\ &\leq k^{4d+3}(d\log(ek))^{6d+5}\left(\tfrac{R}{\varepsilon}\right)^{2d}\left(\tfrac{C}{\delta}\right)^{4d+3} \qquad (165)\end{aligned}$$

Overall this shows that the covering dimension of the normalized secant set if $s=\mathcal{O}(kd)$, and we get for any $0<\zeta<1$

$$\log 2\mathcal{N}\left(d_\Phi, \mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S}_{k,2}(\mathcal{T})),\delta\right)/\zeta = \mathcal{O}\left(kd\cdot\left[1+\log k+\log d+\log\log(ek)+\log\tfrac{R}{\varepsilon}+\log\tfrac{1}{\delta}\right]+\log\tfrac{1}{\zeta}\right)$$

Using (31) with $0<\delta\leq 1$ we get $c_\kappa(\delta/2) = \mathcal{O}\left(\delta^{-2}W_\kappa^2\right) = \mathcal{O}\left(\delta^{-2}kd\log(ek)\right)$. As $\log\log(ek) = \log(1+\log k) = \mathcal{O}(\log k)$, by Theorem 2.9 for $\mathfrak{S}=\mathfrak{S}_0$ we obtain for $\delta<1$: if $m\geq m_0$, where

$$m_0 = \mathcal{O}\left(\delta^{-2}kd\log(ek)\cdot\left[kd\cdot\left(1+\log kd+\log\tfrac{R}{\varepsilon}+\log\tfrac{1}{\delta}\right)+\log\tfrac{1}{\zeta}\right]\right)$$

then with probability at least $1-\zeta$ on the draw of frequencies $(\omega_j)_{j=1}^m$, the sketching operator satisfies the LRIP (12) on $\mathfrak{S}_0$ with constant $C_\mathcal{A} = \frac{C_\kappa}{\sqrt{1-\delta}} = \frac{8\sqrt{6}}{\sqrt{1-\delta}}\sqrt{k}R^p$.

Extension to $\mathcal{H}\subset\mathcal{H}_{k,2\varepsilon,R}$. When $\mathcal{H}\subset\mathcal{H}_{k,2\varepsilon,R}$ we have $\mathfrak{S}_\mathcal{H}\subset\mathfrak{S}_0$ and the LRIP for $\mathfrak{S}_0$ implies the LRIP for $\mathfrak{S}_\mathcal{H}$. Hence, with probability at least $1-\zeta$ the sketching operator satisfies the LRIP simultaneously for all models associated to $\mathcal{H}\subset\mathcal{H}_{k,2\varepsilon,R}$.

F.6 Control of the bias term: Proof of Lemma 4.3

To control the Lipschitz constant L of Φ we observe that

$$\begin{aligned}\|\Phi(x)-\Phi(x')\|_2^2 &= \tfrac{1}{m}\sum_{j=1}^m \tfrac{C_\Lambda^2}{w^2(\omega_j)}\left|e^{\mathrm{J}\omega_j^T x}-e^{\mathrm{J}\omega_j^T x'}\right|^2 = \tfrac{1}{m}\sum_{j=1}^m \tfrac{C_\Lambda^2}{w^2(\omega_j)}\left|e^{\mathrm{J}\omega_j^T(x-x')}-1\right|^2 \\ &\leq \tfrac{1}{m}\sum_{j=1}^m \tfrac{C_\Lambda^2}{w^2(\omega_j)}\left|\omega_j^T(x'-x)\right|^2 = (x'-x)^T\left[\tfrac{1}{m}\sum_{j=1}^m \tfrac{C_\Lambda^2}{w^2(\omega_j)}\omega_j\omega_j^T\right](x'-x) \\ &\leq \left\|\tfrac{1}{m}\sum_{j=1}^m \tfrac{C_\Lambda^2}{w^2(\omega_j)}\omega_j\omega_j^T\right\|_{2\to 2}\cdot\|x'-x\|_2^2 .\end{aligned}$$

Since

$$\mathbb{E}_{\omega\sim\Lambda}\tfrac{C_\Lambda^2}{w^2(\omega)}\omega\omega^T = \int_{\omega\in\mathbb{R}^d}\tfrac{C_\Lambda^2}{w^2(\omega)}\omega\omega^T\tfrac{w^2(\omega)}{C_\Lambda^2}p_{\mathcal{N}(0,\lambda^2\mathbf{I}_d)}(\omega)d\omega = \mathbb{E}_{\omega\sim\mathcal{N}(0,\lambda^2\mathbf{I}_d)}\omega\omega^T = \lambda^2\mathbf{I}_d$$

we will show that, for large enough m, we have $L^2 := \left\|\frac{1}{m}\sum_{j=1}^m \frac{C_\Lambda^2}{w^2(\omega_j)}\omega_j\omega_j^T\right\|_{2\to 2} \lesssim \lambda^2$ with high probability. This will follow from a Bernstein-type matrix concentration inequality.

Theorem F.3 (Theorem 1.4,[82]). *Consider a finite sequence $\{\mathbf{X}_j\}_{j=1}^N$ of independent, random, self-adjoint matrices with dimension d. Assume that each random matrix satisfies $\mathbb{E}\mathbf{X}_j = \mathbf{0}$ and $\|\mathbf{X}_j\|_{2\to 2}\leq M$ almost surely. Then, for all $t\geq 0$,*

$$\mathbb{P}\left\{\left\|\sum_{j=1}^N \mathbf{X}_j\right\|_{2\to 2}\geq t\right\} \leq d\cdot e^{-\frac{t^2/2}{\sigma^2+Mt/3}} \quad \text{where } \sigma^2 := \left\|\sum_{j=1}^N \mathbb{E}\mathbf{X}_j^2\right\|_{2\to 2}. \qquad (166)$$

To exploit this concentration inequality we let $\mathbf{Y}_j := \frac{C_\Lambda^2}{w^2(\omega_j)}\omega_j\omega_j^T$ and $\mathbf{X}_j := \mathbf{Y}_j - \lambda^2\mathbf{I}_d$ so that $\mathbb{E}\mathbf{X}_j = \mathbf{0}$. As $0 \preccurlyeq \mathbf{Y}_j \preccurlyeq C_\Lambda^2 \frac{\|\omega_j\|_2^2}{w^2(\omega_j)} \cdot \mathbf{I}_d$ and

$$\sup_{\omega\in\mathbb{R}^d} \frac{\|\omega\|_2^2}{w^2(\omega)} = \sup_{r\geq 0} \frac{r^2}{(1+\frac{r^2}{\lambda^2 d})^2} = \frac{1}{\inf_{r>0}(\frac{1}{r}+\frac{r}{\lambda^2 d})^2} = \frac{1}{\left(\frac{2}{\lambda\sqrt{d}}\right)^2} = \frac{\lambda^2 d}{4},$$

we have $-\lambda^2\mathbf{I}_d \preccurlyeq \mathbf{X}_j \preccurlyeq \lambda^2\left(\frac{C_\Lambda^2 d}{4}-1\right)\mathbf{I}_d$, hence $\|\mathbf{X}_j\|_{2\to 2} \leq \lambda^2 \cdot \max\left(1, \frac{C_\Lambda^2 d}{4}-1\right) \leq \lambda^2\frac{C_\Lambda^2 d}{4} =: M$. Moreover $\sigma^2 := \left\|\sum_{j=1}^m \mathbb{E}\mathbf{X}_j^2\right\|_{2\to 2} = m\left\|\mathbb{E}\mathbf{X}^2\right\|_{2\to 2}$. It follows that for any $t \geq 0$

$$\mathbb{P}\left\{\left\|\frac{1}{m}\sum_{j=1}^m \frac{C_\Lambda^2}{w^2(\omega_j)}\omega_j\omega_j^T - \lambda^2\mathbf{I}_d\right\|_{2\to 2} \geq t\lambda^2\right\} = \mathbb{P}\left\{\left\|\sum_{j=1}^m \mathbf{X}_j\right\|_{2\to 2} \geq mt\lambda^2\right\} \leq d\cdot e^{-\frac{m^2t^2\lambda^4/2}{\sigma^2+Mmt\lambda^2/3}}. \quad (167)$$

Noticing that

$$\frac{m^2t^2\lambda^4/2}{\sigma^2+Mmt\lambda^2/3} = m\cdot\frac{t^2}{2\left(\|\mathbb{E}\mathbf{X}^2/\lambda^4\|_{2\to 2}+\frac{C_\Lambda^2 d}{12}t\right)}$$

We now bound $\left\|\mathbb{E}\mathbf{X}^2/\lambda^4\right\|_{2\to 2}$. We have $\mathbb{E}\mathbf{X}^2 = \mathbb{E}\mathbf{Y}^2 - (\mathbb{E}\mathbf{Y})^2 = \mathbb{E}\mathbf{Y}^2 - \lambda^4\mathbf{I}_d$ and

$$\begin{aligned}
\mathbb{E}\mathbf{Y}^2 &= \int_{\omega\in\mathbb{R}^d}\left(\frac{C_\Lambda^2}{w^2(\omega)}\right)^2\|\omega\|_2^2\,\omega\omega^T\frac{w^2(\omega)}{C_\Lambda^2}p_{\mathcal{N}(0,\lambda^2\mathbf{I}_d)}(\omega)d\omega = \mathbb{E}_{\omega\sim\mathcal{N}(0,\lambda^2\mathbf{I}_d)}\frac{C_\Lambda^2}{w^2(\omega)}\|\omega\|_2^2\,\omega\omega^T \\
&= \lambda^4C_\Lambda^2\cdot\mathbb{E}_{\omega'\sim\mathcal{N}(0,\mathbf{I}_d)}\frac{\|\omega'\|_2^2\omega'\omega'^T}{(1+\|\omega'\|_2^2/d)^2} \overset{(*)}{=} \lambda^4C_\Lambda^2\cdot\mathbb{E}_{r\sim\chi_d}\frac{r^4}{(1+r^2/d)^2}\cdot\mathbb{E}_{\mathbf{u}\sim\mathcal{U}_d}\mathbf{u}\mathbf{u}^T \\
&= \lambda^4C_\Lambda^2\cdot\mathbb{E}_{r\sim\chi_d}\frac{r^4}{(1+r^2/d)^2}\cdot\frac{1}{d}\cdot\mathbf{I}_d
\end{aligned}$$

where in (*) we used $\mathcal{U}_d$ the uniform distribution on the unit sphere in $\mathbb{R}^d$. Furthermore

$$\mathbb{E}_{r\sim\chi_d}\frac{r^4}{(1+r^2/d)^2} \leq \mathbb{E}_{r\sim\chi_d}r^4 = d(d+2).$$

As a result $\left\|\mathbb{E}\mathbf{X}^2\lambda^{-4}\right\|_{2\to 2} \leq C_\Lambda^2(d+2)$ and $2\left(\left\|\mathbb{E}\mathbf{X}^2\lambda^{-4}\right\|_{2\to 2}+\frac{C_\Lambda^2 d}{12}t\right) \leq C_\Lambda^2\left(2d+4+\frac{d}{6}\cdot t\right)$. Combining the above it follows that $L^2 \leq \lambda^2(1+t)$ with probability at least $1-\zeta$ provided that

$$m \geq C_\Lambda^2 t^{-2}\left(2d+4+\tfrac{d}{6}\cdot t\right)\cdot\log\left(\frac{d}{\zeta}\right).$$

F.7 Link between risks with and without ε-separation

We define a "distance" (not symmetric) between hypotheses $h = \{c_1,..,c_k\}$ and $h' = \{c_1',..,c_l'\}$ two hypotheses.

$$d(h,h') := \max_{c_i\in h} d(c_i,h') = \max_{c_i\in h}\min_{c_j'\in h'}\left\|c_i - c_j'\right\|_2 \quad (168)$$

For any hypothesis class $\mathcal{H}$, we further define

$$d(h,\mathcal{H}) := \inf_{h'\in\mathcal{H}} d(h,h'). \quad (169)$$

We have the following Lemma :

Lemma F.4. *Let $R > 0$. For any $h \in \mathcal{H}_{k,0,R}$ an hypothesis, there is $h' \in \mathcal{H}_{k,\varepsilon,R}$ such that $d(h, h') \leq \epsilon$.*

Proof. We prove this statement by induction on k.

For $k = 1$, consider $h = \{c_1\}$. Taking $h' := \{c_1\} \in \mathcal{H}_{k,\varepsilon,R}$ we have $d(h, h') = 0 \leq \epsilon$.

Consider now $k \geq 2$ and suppose the statement true for $k - 1$.

Let $h = \{c_1, .., c_{k_1}\} \in \mathcal{H}_{k,0,R}$, with $k_1 \leq k$. If $k_1 \leq k - 1$, then indeed $h \in \mathcal{H}_{k-1,0,R}$ hence by the induction hypothesis there is $h' \in \mathcal{H}_{k-1,\varepsilon,R} \subset \mathcal{H}_{k,\varepsilon,R}$ such that $d(h, h') \leq \varepsilon$. Otherwise (that is to say if $k_1 = k \geq 2$, and $c_i \neq c_j$ for $i \neq j$), since $\{c_1, \ldots, c_{k_1-1}\} \in \mathcal{H}_{k-1,0,R}$, by the induction hypothesis there is $\left\{c'_1, .., c'_{k'_2}\right\} \in \mathcal{H}_{k-1,\varepsilon,R}$ ($k'_2 \leq k - 1$) such that $d(\{c_1, .., c_{k_1-1}\}, \left\{c'_1, .., c'_{k'_2}\right\}) \leq \varepsilon$. We now distinguish two cases:

- If $\min_{1 \leq j \leq k_2} \left\|c_{k_1} - c'_j\right\|_2 > \varepsilon$, set $c'_{k_2+1} := c_{k_1}$ and $h' := \left\{c'_1, .., c'_{k_2+1}\right\}$. We have $h' \in \mathcal{H}_{k,\varepsilon,R}$ and $d(h, h') \leq \varepsilon$.
- If $\min_{1 \leq j \leq k_2} \left\|c_{k_1} - c'_j\right\|_2 \leq \varepsilon$. Take $h' := \left\{c'_1, .., c'_{k_2}\right\}$. We have $h' \in \mathcal{H}_{k-1,\varepsilon,R} \subset \mathcal{H}_{k,\varepsilon,R}$ and $d(h, h') \leq \varepsilon$.

□

Lemma F.5. *Let $h_0^\star \in \arg\min_{h \in \mathcal{H}_{k,0,R}} \mathcal{R}_{\texttt{clust.}}(\pi_0, h)$ be an optimal hypothesis* without *separation constraint, and $h^\star \in \arg\min_{h \in \mathcal{H}_{k,2\varepsilon,R}} \mathcal{R}_{\texttt{clust.}}(\pi_0, h)$ an optimal hypothesis* with *the separation constraint.*

- *For k-medians ($p = 1$) we have:*

$$\mathcal{R}_{\texttt{clust.}}(\pi_0, h^\star) \leq \mathcal{R}_{\texttt{clust.}}(\pi, h_0^\star) + d(h_0^\star, \mathcal{H}_{k,2\varepsilon,R}) \leq \mathcal{R}_{\texttt{clust.}}(\pi, h_0^\star) + 2\varepsilon. \tag{170}$$

- *For k-means ($p = 2$) we have:*

$$\sqrt{\mathcal{R}_{\texttt{clust.}}(\pi_0, h^\star)} \leq \sqrt{\mathcal{R}_{\texttt{clust.}}(\pi_0, h_0^\star)} + d(h_0^\star, \mathcal{H}_{k,2\varepsilon,R}) \leq \sqrt{\mathcal{R}_{\texttt{clust.}}(\pi_0, h_0^\star)} + 2\varepsilon. \tag{171}$$

Proof. For any $\alpha > 0$ there is $h' \in \mathcal{H}_{k,2\varepsilon,R}$ such that $d(h_0^\star, h') \leq d(h_0^\star, \mathcal{H}_{k,2\varepsilon,R}) + \alpha$. Considering any $x \in \mathcal{Z}$, $c_{i_0} := \arg\min_{c_i \in h_0^\star} \|x - c_i\|_2$ and $c'_{j_0} \in h'$ such that $\left\|c'_{j_0} - c_{i_0}\right\|_2 \leq d(h_0^\star, h')$, with the reverse triangle inequality we have

$$\begin{aligned} \|x - c_{i_0}\|_2 = \left\|x - c'_{j_0} + c'_{j_0} - c_{i_0}\right\|_2 &\geq \left\|x - c'_{j_0}\right\|_2 - \left\|c'_{j_0} - c_{i_0}\right\|_2 \geq \left\|x - c'_{j_0}\right\|_2 - d(h_0^\star, h') \\ &\geq \min_{c'_j \in h'} \left\|x - c'_j\right\|_2 - d(h_0^\star, h'). \end{aligned}$$

- for $p = 1$ it follows that

$$\begin{aligned} \mathcal{R}_{\texttt{clust.}}(\pi_0, h') = \mathbb{E}_{X \sim \pi_0} \min_{c'_j \in h'} \left\|X - c'_j\right\|_2 &\leq \mathbb{E}_{X \sim \pi_0} \min_{c_i \in h_0^\star} \|X - c_i\|_2 + d(h_0^\star, h') \\ &\leq \mathcal{R}_{\texttt{clust.}}(\pi_0, h_0^\star) + d(h_0^\star, \mathcal{H}_{k,2\varepsilon,R}) + \alpha. \end{aligned}$$

  Now by definition of $h^\star$,

$$\mathcal{R}_{\texttt{clust.}}(\pi_0, h^\star) \leq \mathcal{R}_{\texttt{clust.}}(\pi_0, h') \leq \mathcal{R}_{\texttt{clust.}}(\pi, h_0^\star) + d(h_0^\star, \mathcal{H}_{k,2\varepsilon,R}) + \alpha.$$

  As this holds for any $\alpha > 0$, this yields (170).

- for $p = 2$, we have instead

$$\mathcal{R}_{\texttt{clust.}}(\pi_0, h') \leq \mathbb{E}_{X\sim\pi_0}\left(\min_{c_i\in h_0^\star} \|X - c_i\|_2 + d(h_0^\star, h')\right)^2$$
$$= \mathcal{R}_{\texttt{clust.}}(\pi_0, h_0^\star) + 2\left(\mathbb{E}_{X\sim\pi_0} \min_{c_i\in h_0^\star} \|X - c_i\|_2\right) d(h_0^\star, h') + d(h_0^\star, h')^2$$

With Jensen's inequality we have $\mathbb{E}_{X\sim\pi_0} \min_{c_i\in h_0^\star} \|X - c_i\|_2 \leq \sqrt{\mathcal{R}_{\texttt{clust.}}(\pi_0, h_0^\star)}$, yielding

$$\sqrt{\mathcal{R}_{\texttt{clust.}}(\pi_0, h')} \leq \sqrt{\mathcal{R}_{\texttt{clust.}}(\pi_0, h_0^\star)} + d(h_0^\star, h').$$

Similarly as the $p = 1$ case, this gives eventually (171).

Finally, with Lemma F.4, we have $d(h_0^\star, \mathcal{H}_{k,2\varepsilon,R}) \leq 2\epsilon$. □

F.8 Necessity of the separation assumption for certain kernels

As we show now, the separation assumption for the Compressive Clustering method is in fact necessary under mild smoothness assumption on the shift invariant kernel $\kappa(.)$ (including the Gaussian kernel used in Section 4).

Lemma F.6. *Consider a loss $\ell(x, \{c_l\}) = \min_l \|x - c_l\|_2^p$ with $0 < p < \infty$ associated to a clustering task (this includes k-means, $p = 2$, or k-medians, $p = 1$), and a shift invariant kernel $\kappa(x, x') = \kappa(x - x')$. Assume that there is at least one direction $\theta_0 \in \mathbb{R}^d$ such that $f : t \mapsto \kappa(t\theta_0)$ is twice differentiable at zero. Then there is no finite constant $C < \infty$ such that for all $\pi, \pi' \in \mathfrak{S}_{\mathcal{H}_{k,0,R}}$ it holds that*

$$\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}_{k,0,R})} \leq C \|\pi - \pi'\|_\kappa$$

Proof. Consider $\theta_+ := \frac{\varepsilon}{2}\theta_0, \theta_- = -\theta_+$ (hence $\|\theta_+\| \leq R/2$, $\|\theta_-\| \leq R/2$ for small enough ε). Observe that $\varepsilon = \|\theta_+ - \theta_-\|_2$. Define two mixtures $\pi = \frac{1}{2}(\delta_{\theta_+} + \delta_{\theta_-})$, $\pi' = \delta_{(\theta_+ + \theta_-)/2} = \delta_0 \in \mathfrak{S}_{\mathcal{H}_{k,0,R}}$. Setting $\alpha := \frac{R}{2\varepsilon}$, define the set of centroids $h = \{c_+, c_-\}$ as

$$c_+ = \theta_+ + \alpha(\theta_+ - \theta_-), \quad c_- = \theta_- + \alpha(\theta_- - \theta_+)$$

As $\ell(\theta_+, h) = \ell(\theta_-, h) = (\alpha\varepsilon)^p = (R/2)^p$ and $\ell\left(\frac{\theta_+ + \theta_-}{2}, h\right) = (1/2 + \alpha)^p \varepsilon^p = (R/2)^p (1 + \varepsilon/R)^p$, we have

$$|\langle \pi - \pi', \ell(\cdot, h)\rangle| = (R/2)^p |(1 + \varepsilon/R)^p - 1| = (R/2)^p \left(\tfrac{p}{R}\varepsilon + o(\varepsilon)\right) \tag{172}$$

We also have $\|c_+\| \leq R$, $\|c_-\| \leq R$ hence $h \in \mathcal{H}_{k,0,R}$, and $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}_{k,0,R})} \geq |\langle \pi - \pi', \ell(\cdot, h)\rangle|$.

Now, since κ is a kernel we have $\kappa(x, x') = \kappa(x', x)$ for any x, x', hence f is an even function ($f(t) = f(-t)$). Since f is differentiable, this implies $f'(0) = 0$. We have $\kappa(\pi, \pi) = \frac{1}{4}[2\kappa(0) + 2\kappa(\theta_+ - \theta_-)] = \frac{1}{2}[\kappa(0) + \kappa(\varepsilon\theta_0)] = \frac{1}{2}[f(0) + f(\varepsilon)]$, $\kappa(\pi', \pi') = \kappa(0) = f(0)$, and $\kappa(\pi, \pi') = \frac{1}{2}[\kappa((\theta_+ - \theta_-)/2) + \kappa((\theta_- - \theta_+)/2)] = \kappa((\theta_+ - \theta_-)/2) = \kappa(\varepsilon\theta_0/2) = f(\varepsilon/2)$. Hence

$$\|\pi - \pi'\|_\kappa^2 = \kappa(\pi, \pi) + \kappa(\pi', \pi') - 2\kappa(\pi, \pi') = \tfrac{1}{2}[f(0) + f(\varepsilon)] + f(0) - 2f(\varepsilon/2)$$
$$= \tfrac{1}{2}[f(\varepsilon) - f(0)] - 2[f(\varepsilon/2) - f(0)] = \tfrac{1}{2}\left[\tfrac{f''(0)}{2}\varepsilon^2 + o(\varepsilon^2)\right] - 2\left[\tfrac{f''(0)}{2}\tfrac{\varepsilon^2}{4} + o(\varepsilon^2)\right] = o(\varepsilon^2).$$

As a result $\|\pi - \pi'\|_\kappa = o(\varepsilon)$. Given (172) we obtain $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H}_{k,0,R})} / \|\pi - \pi'\|_\kappa \underset{\varepsilon\to 0}{\to} \infty$, which contradicts the existence of a constant C_κ such that compatibility holds. □

G Proof of Theorem 5.1 on Compressive GMM

The proof of Theorem 5.1 combines Theorem 2.9 with the generic strategy of Section 6 applied the specific case of separated mixtures of Gaussians. Consider $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ a fixed known full-rank covariance, $\|\cdot\|_{\mathbf{\Sigma}}$ the associated Mahalanobis norm (69), and a set of Gaussian distributions $\mathcal{T} = \{\pi_\theta = \mathcal{N}(\theta, \mathbf{\Sigma});\ \theta \in \mathbf{\Theta}\}$ where $\mathbf{\Theta} = \mathcal{B}_{\mathbb{R}^d,\|\cdot\|_{\mathbf{\Sigma}}}(0, R)$ is the ball of radius R for the Mahalanobis norm. The sample space is $\mathcal{Z} = \mathbb{R}^d$.

For a hypothesis $h = (\theta_1, ..., \theta_k, \alpha) \in \mathcal{H} \subset \mathbf{\Theta}^k \times \mathbb{S}_{k-1}$, the loss function for density fitting is the negative log-likelihood $\ell(x, h) = -\log \sum_{l=1}^k \alpha_l \pi_{\theta_l}(x)$. The model set $\mathfrak{S}_{\mathcal{H}}$ is precisely the set of mixtures of Gaussians with parameters in $\mathcal{H}$.

With the separation assumption, the hypothesis class is $\mathcal{H} \subset \mathcal{H}_{k,2\varepsilon,R}$ where for GMM we consider $\mathcal{H}_{k,2\varepsilon,R}$ defined in (68). The model set is $\mathfrak{S}_{\mathcal{H}} = \mathfrak{S}_{k,2\varepsilon,\|\cdot\|_{\mathbf{\Sigma}}}(\mathcal{T}) = \mathfrak{S}_{k,2,\varrho_\varepsilon}(\mathcal{T})$, consisting of mixtures of Gaussians that are 2ε-separated for the Mahalanobis norm, or 2-separated with the metric ϱ_ε, where

$$\varrho_\beta(\theta, \theta') := \|\theta - \theta'\|_{\mathbf{\Sigma}} / \beta. \tag{173}$$

G.1 Kernel and features

We recall that the sketching function is built with scaled random Fourier features where frequency vectors ω_j are drawn according to (70). With $\phi_\omega(x) := C_\lambda e^{\mathrm{J}\omega^T x}$, we have for any $x, x' \in \mathbb{R}^d$

$$\kappa(x, x') \quad = \quad \mathbb{E}_{\omega\sim\Lambda}\phi_\omega(x)\overline{\phi_\omega(x')} = C_\lambda^2 \cdot \mathbb{E}_{\omega\sim\mathcal{N}(0,\lambda^2\mathbf{\Sigma}^{-1})} e^{\mathrm{J}\omega^T(x-x')} \overset{(*)}{=} C_\lambda^2 \cdot \exp\left(-\frac{\|x - x'\|^2_{\lambda^{-2}\mathbf{\Sigma}}}{2}\right)$$

where (*) follows from the expression of the characteristic function of the Gaussian. With the following Lemma we can characterize the associated Mean Map kernel for Gaussian distributions that do not necessarily have a fixed known covariance.

Lemma G.1. *Consider a Gaussian kernel $\kappa(x, x') := \exp\left(-\frac{1}{2}\|x - x'\|^2_{\mathbf{\Sigma}_\kappa}\right)$ where $\mathbf{\Sigma}_\kappa$ is an arbitrary invertible covariance matrix. For any two Gaussians $\pi_1 = \mathcal{N}(\theta_1, \mathbf{\Sigma}_1)$, $\pi_2 = \mathcal{N}(\theta_2, \mathbf{\Sigma}_2)$, the mean kernel* (18) *is*

$$\kappa(\pi_1, \pi_2) = \frac{\sqrt{\det(\mathbf{\Sigma}_\kappa)}}{\sqrt{\det(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 + \mathbf{\Sigma}_\kappa)}} \exp\left(-\tfrac{1}{2}\|\theta_1 - \theta_2\|^2_{\mathbf{\Sigma}_1+\mathbf{\Sigma}_2+\mathbf{\Sigma}_\kappa}\right) \tag{174}$$

Proof. We use a property from [2] on product of Gaussians:

$$\int \pi_1(x)\pi_2(x)dx = \frac{1}{\sqrt{\det(2\pi(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2))}} \exp\left(-\tfrac{1}{2}\|\theta_1 - \theta_2\|^2_{\mathbf{\Sigma}_1+\mathbf{\Sigma}_2}\right) \tag{175}$$

We can write the kernel

$$\kappa(x, x') = \exp\left(-\tfrac{1}{2}\|x - x'\|^2_{\mathbf{\Sigma}_\kappa}\right) = \sqrt{\det(2\pi\mathbf{\Sigma}_\kappa)} \cdot \pi_\kappa(x - x')$$

where $\pi_\kappa = \mathcal{N}(0, \mathbf{\Sigma}_\kappa)$. Hence we have

$$\begin{aligned}\kappa(\pi_1, \pi_2) =& \sqrt{\det(2\pi\mathbf{\Sigma}_\kappa)} \int_x \pi_1(x) \left(\int_{x'} \pi_2(x')\pi_\kappa(x - x')dx'\right) dx \\ =& \sqrt{\det(2\pi\mathbf{\Sigma}_\kappa)} \int_x \pi_1(x)\pi_{2,\kappa}(x)dx,\end{aligned}$$

by convolution, where $\pi_{2,\kappa} = \mathcal{N}(\theta_2, \mathbf{\Sigma}_2 + \mathbf{\Sigma}_\kappa)$. Using (175) we get the desired result. ∎

According to Lemma G.1, denoting $K_\sigma(\cdot)$ the Gaussian kernel (85), we have

$$\begin{aligned}\kappa(\pi_\theta, \pi_{\theta'}) =& C_\lambda^2 \frac{\sqrt{\det\left(\lambda^{-2}\boldsymbol{\Sigma}\right)}}{\sqrt{\det\left((2+\lambda^{-2})\boldsymbol{\Sigma}\right)}} \exp\left(-\tfrac{1}{2}\cdot\|\theta-\theta'\|^2_{(2+\lambda^{-2})\boldsymbol{\Sigma}}\right) \\ =& C_\lambda^2 \left(\frac{\lambda^{-2}}{2+\lambda^{-2}}\right)^{d/2} \exp\left(-\frac{1}{2\sigma_k^2}\cdot\frac{\sigma_k^2}{2+\lambda^{-2}}\|\theta-\theta'\|^2_{\boldsymbol{\Sigma}}\right) = K_{\sigma_k}(\varrho_\varepsilon(\theta,\theta'))\end{aligned}$$

with $\varepsilon = \varepsilon_\lambda := \sqrt{2+\lambda^{-2}}/\sigma_k$ and $C_\lambda := (2\lambda^2+1)^{d/4}$. Notice that as $\sigma_k \leq 1$ we have $\varepsilon \geq \sqrt{2}$, and that $C_\lambda \geq 1$. By Lemma 6.8 we have $K_{\sigma_k} \in \mathcal{E}_k(1)$ and we conclude that the Mean Map kernel has the desired form $\kappa(\pi_\theta, \pi_{\theta'}) = K(\varrho_\varepsilon(\theta,\theta'))$ with $K(\cdot) \in \mathcal{E}_k(1)$.

G.2 Concentration constant

The considered integral representation of the kernel $\kappa(x,x')$ involves the class

$$\Phi := \left\{\phi_\omega(x) := C_\lambda \cdot e^{\jmath\omega^T x}; \omega \in \mathbb{R}^d\right\}.$$

Unlike for compressive clustering, no weights are needed on these features to establish that they are bounded and Lipschitz in expectation.

Lemma G.2. *For any $\beta > 0$ we have $\Phi \subset \mathcal{BL}(D_\Phi, L_\Phi, \mathcal{T}, \varrho_\beta)$ with $D_\Phi = C_\lambda$ and $L_\Phi = C_\lambda\beta$.*

Proof. Using the expression of the characteristic function of a Gaussian, $\sup_\omega \sup_\theta |\mathbb{E}_{x\sim\pi_\theta}\phi_\omega(x)| = C_\lambda$, and by Lemma G.3 and the definition of $\|\cdot\|_{\mathrm{TV}} = \|\cdot\|_{\mathcal{B}}$ with $\mathcal{B} = \{f : \|f\|_\infty \leq 1\}$, we have for any θ, θ'

$$|\langle \pi_\theta - \pi_{\theta'}, \phi_\omega\rangle| \leq C_\lambda \|\pi_\theta - \pi_{\theta'}\|_{\mathrm{TV}} \leq C_\lambda \|\theta-\theta'\|_{\boldsymbol{\Sigma}} = C_\lambda\beta\varrho_\beta(\theta,\theta').$$

□

Lemma G.3. *Consider two Gaussians with the same covariance $\pi_\theta = \mathcal{N}(\theta, \boldsymbol{\Sigma})$, $\pi_{\theta'} = \mathcal{N}(\theta', \boldsymbol{\Sigma})$. We have*

$$\|\pi_\theta - \pi_{\theta'}\|_{\mathrm{TV}} \leq \|\theta-\theta'\|_{\boldsymbol{\Sigma}} \tag{176}$$

Proof. Using Pinsker's inequality (102) we have $\|\pi_\theta - \pi_{\theta'}\|_{\mathrm{TV}} \leq \sqrt{2\mathrm{KL}(\pi_\theta||\pi_{\theta'})}$. The Kullback-Leibler divergence has a closed form expression in the case of multivariate Gaussians [38]:

$$\mathrm{KL}(\mathcal{N}(\theta_1,\boldsymbol{\Sigma}_1)||\mathcal{N}(\theta_2,\boldsymbol{\Sigma}_2)) = \tfrac{1}{2}\left[\log\frac{\det(\boldsymbol{\Sigma}_2)}{\det(\boldsymbol{\Sigma}_1)} + \mathrm{tr}\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right) - d + (\theta_2-\theta_1)^T\boldsymbol{\Sigma}_2^{-1}(\theta_2-\theta_1)\right]. \tag{177}$$

In our case, with fixed covariance, it yields $\mathrm{KL}(\pi_\theta||\pi_{\theta'}) = \frac{1}{2}\|\theta-\theta'\|^2_{\boldsymbol{\Sigma}}$. □

We get with $\beta = \varepsilon$: $D_\Phi = C_\lambda$, $L_\Phi = \varepsilon C_\lambda$. As we will use this constant in several places, we introduce the shorthand $W_\Phi := \sqrt{L_\Phi^2 + 2D_\Phi^2} = C_\lambda\sqrt{2+\varepsilon^2} \leq \sqrt{2}\varepsilon C_\lambda$ (since $\varepsilon \geq \sqrt{2}$). Moreover the model set $\mathfrak{S}_{\mathcal{H}}$ consists of 2ε-separated mixtures with respect to the Mahalanobis metric and therefore of 2-separated mixtures with respect to ϱ_ε. As $K_{\sigma_k} \in \mathcal{E}_k(1)$, we can apply Theorem 6.12 to bound the concentration constant as

$$W_\kappa \leq 2\sqrt{2k}W_\Phi = 2\sqrt{4+2\varepsilon^2}\sqrt{k}C_\lambda \leq 4\sqrt{k}\varepsilon C_\lambda. \tag{178}$$

G.3 Compatibility constant

Controlling the compatibility constant by showing that the loss function is bounded and Lipschitz in expectation is a bit more delicate. In particular, in order to minimize the bounding constant $D_{\mathcal{L}}$, we will use the fact that the learning task (minimizing the risk) and the associated norm $\|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})}$ are unchanged when an offset is added to the loss $\ell(\cdot, h) = -\log \pi_h(\cdot)$. The following lemma, which applies to any family of probability distributions $\mathcal{T} = \{\pi_\theta : \theta \in \mathbf{\Theta}\}$, will be soon specialized to Gaussians with fixed known covariance.

Lemma G.4. *Consider a family of probability distributions $\mathcal{T} = \{\pi_\theta : \theta \in \mathbf{\Theta}\}$. Assume that*

$$\mathrm{H}_{\min} := \inf_{\theta \in \mathbf{\Theta}} \mathrm{H}(\pi_\theta) > -\infty \tag{179}$$

$$\mathrm{H}_{\max} := \sup_{\theta, \theta' \in \mathbf{\Theta}} \mathrm{H}(\pi_\theta) + \mathrm{KL}(\pi_\theta || \pi_{\theta'}) < \infty. \tag{180}$$

For any $\pi_h := \sum_{l=1}^k \alpha_l \pi_{\theta_l}$, where $\boldsymbol{\alpha} \in \mathbb{S}_{k-1}$, $\theta_l \in \mathbf{\Theta}$ we have for any $\theta \in \mathbf{\Theta}$:

$$\mathrm{H}_{\min} \leq \mathbb{E}_{X \sim \pi_\theta} \left[-\log \pi_h(X) \right] \leq \mathrm{H}_{\max}.$$

The lower and the upper bounds are both tight.

Proof. By the properties of the Kullback-Leibler divergence and the convexity of $-\log(\cdot)$ we have

$$\begin{aligned}
\mathrm{H}(\pi_\theta) \leq \mathrm{H}(\pi_\theta) + \mathrm{KL}(\pi_\theta || \pi_h) &= \mathbb{E}_{X \sim \pi_\theta} \left[-\log \pi_h(X) \right] \\
&= \mathbb{E}_{X \sim \pi_\theta} \left[-\log \left(\sum_{l=1}^k \alpha_l \pi_{\theta_l}(X) \right) \right] \leq \sum_{l=1}^k \alpha_l \mathbb{E}_{X \sim \pi_\theta} \left[-\log \pi_{\theta_l}(X) \right] \\
&\leq \sum_{l=1}^k \alpha_l \left[\mathrm{H}(\pi_\theta) + \mathrm{KL}(\pi_\theta || \pi_{\theta_l}) \right] \leq \mathrm{H}(\pi_\theta) + \sup_{\theta' \in \mathbf{\Theta}} \mathrm{KL}(\pi_\theta || \pi_{\theta'})
\end{aligned}$$

For a given θ, both the lower and the upper bound are tight. The conclusion immediately follows. □

This translates into a concrete result for Gaussian mixtures with fixed known covariance.

Lemma G.5. *Consider $\mathcal{T} = \left\{ \pi_\theta : \theta \in \mathcal{B}_{\mathbb{R}^d, \|\cdot\|_{\mathbf{\Sigma}}}(0, R) \right\}$, where $\pi_\theta := \mathcal{N}(\theta, \mathbf{\Sigma})$. Consider loss functions with an offset*

$$\bar{\ell}(x, h) := -\log \pi_h(x) - \tfrac{1}{2} \log \det (2\pi e \mathbf{\Sigma}) - R^2$$

The loss class $\mathcal{L}(\mathcal{H}) = \left\{ \bar{\ell}(\cdot, h) : h \in \mathcal{H} \right\}$ satisfies

$$\mathcal{L}(\mathcal{H}) \subset \mathcal{BL}(D_{\mathcal{L}}, L_{\mathcal{L}}, \mathcal{T}, \varrho_\beta) \tag{181}$$

with

$$D_{\mathcal{L}} := R^2 \tag{182}$$

$$L_{\mathcal{L}} := 2R\beta. \tag{183}$$

Proof. To control $D_{\mathcal{L}}$ we exploit Lemma G.4. The entropy of a Gaussian is $\mathrm{H}(\pi_\theta) = \frac{1}{2} \log \det (2\pi e \mathbf{\Sigma})$ which is independent of θ, hence $\mathrm{H}_{\min} = \frac{1}{2} \log \det (2\pi e \mathbf{\Sigma})$. The Kullback-Leibler divergence (177) is $\mathrm{KL}(\pi_\theta || \pi_{\theta'}) = \frac{1}{2} \|\theta - \theta'\|_{\mathbf{\Sigma}}^2$. As a result $\mathrm{H}_{\max} = \mathrm{H}_{\min} + \sup_{\theta, \theta' \in \mathcal{B}(0, R, \|\cdot\|_{\mathbf{\Sigma}})} \frac{1}{2} \|\theta - \theta'\|_{\mathbf{\Sigma}}^2 = \mathrm{H}_{\min} + 2R^2$. Hence

$$\bar{\ell}(x, h) = -\log \pi_h(x) - \tfrac{\mathrm{H}_{\max} + \mathrm{H}_{\min}}{2}$$

and by Lemma G.4

$$D_{\mathcal{L}} := \sup_{\theta\in\mathbf{\Theta}} \sup_{h\in\mathcal{H}} \left|\mathbb{E}_{X\sim\pi_\theta}\bar{\ell}(X,h)\right| = \tfrac{\mathrm{H}_{\max}-\mathrm{H}_{\min}}{2} = R^2.$$

For the Lipschitz part, since $\pi_\theta(x) = \pi_0(x-\theta) = \pi_0(\theta - x)$, denoting $\ell_h(x) := -\log \pi_h(x)$ we have

$$\begin{aligned}
f(\theta) &:= \mathbb{E}_{X\sim\pi_\theta}\ell_h(X) = \int \pi_\theta(x)\ell_h(x)dx = \int \pi_0(\theta-x)\ell_h(x)dx = \int \pi_0(x)\ell_h(\theta-x)dx \\
\nabla f(\theta) &= \int \pi_0(x)\nabla\ell_h(\theta-x)dx = \int \pi_0(\theta-x)\nabla\ell_h(x)dx = \mathbb{E}_{X\sim\pi_\theta}\nabla\ell_h(X) \\
\nabla\ell_h(x) &= -\frac{\sum_{l=1}^k \alpha_l \nabla\pi_{\theta_l}(x)}{\sum_{l=1}^k \alpha_l\pi_{\theta_l}(x)} = -\frac{\sum_{l=1}^k \alpha_l\pi_{\theta_l}(x)\cdot\frac{\nabla\pi_{\theta_l}(x)}{\pi_{\theta_l}(x)}}{\sum_{l=1}^k \alpha_l\pi_{\theta_l}(x)} = -\sum_{l=1}^k \frac{\alpha_l\pi_{\theta_l}(x)}{\sum_{l=1}^k \alpha_l\pi_{\theta_l}(x)}\cdot\frac{\nabla\pi_{\theta_l}(x)}{\pi_{\theta_l}(x)} \\
&= -\sum_{l=1}^k \beta_l(x)\cdot\nabla\log\pi_{\theta_l}(x)
\end{aligned}$$

where $\beta_l(x) := \frac{\alpha_l\pi_{\theta_l}(x)}{\sum_{l=1}^k \alpha_l\pi_{\theta_l}(x)} \geq 0$ satisfies $\sum_l \beta_l(x) = 1$. Since $\nabla\log\pi_{\theta_l}(x) = -\mathbf{\Sigma}^{-1}(x-\theta_l)$, we have

$$\begin{aligned}
\mathbb{E}_{X\sim\pi_\theta}\nabla\ell_h(X) &= \mathbb{E}_{X\sim\pi_\theta}\sum_{l=1}^k \beta_l(X)\cdot\mathbf{\Sigma}^{-1}(X-\theta_l) = \mathbf{\Sigma}^{-1}\mathbb{E}_{X\sim\pi_\theta}\left(X - \sum_{l=1}^k \beta_l(X)\theta_l\right) \\
&= \mathbf{\Sigma}^{-1}\left(\theta - \sum_{l=1}^k \gamma_l\cdot\theta_l\right)
\end{aligned}$$

with $\gamma_l := \mathbb{E}_{X\sim\pi_\theta}\beta_l(X) \geq 0$, $\sum_{l=1}^k \gamma_l = 1$. It follows that

$$\|\nabla f(\theta)\|_{\mathbf{\Sigma}^{-1}} = \left\|\mathbf{\Sigma}^{1/2}\nabla f(\theta)\right\|_2 = \left\|\mathbf{\Sigma}^{-1/2}(\theta - \sum_{l=1}^k \gamma_l\cdot\theta_l)\right\|_2 = \left\|\theta - \sum_{l=1}^k \gamma_l\cdot\theta_l\right\|_{\mathbf{\Sigma}} \leq 2R$$

were we used that θ, θ_l are in the ball of radius R with respect to $\|\cdot\|_{\mathbf{\Sigma}}$. To conclude, given θ, θ', defining $\theta(t) := \theta + t(\theta'-\theta)$ we have

$$\begin{aligned}
|f(\theta') - f(\theta)| &= \left|\int_0^1 \frac{d}{dt}f(\theta(t))dt\right| = \left|\int_0^1 \langle\nabla f(\theta(t)), \theta'-\theta\rangle\, dt\right| \leq \int_0^1 \|\nabla f(\theta(t))\|_{\mathbf{\Sigma}^{-1}}\|\theta'-\theta\|_{\mathbf{\Sigma}}\, dt \\
&\leq 2R\|\theta'-\theta\|_{\mathbf{\Sigma}} = 2R\beta\varrho_\beta(\theta,\theta').
\end{aligned}$$

□

As above we get with $\beta = \varepsilon$: $D_{\mathcal{L}} = R^2$, $L_{\mathcal{L}} = 2R\varepsilon \leq 2R^2$ since $R \geq \varepsilon$. Moreover the model set $\mathfrak{S}_{\mathcal{H}}$ consists of 2ε-separated mixtures with respect to the Mahalanobis metric and therefore of 2-separated mixtures with respect to ϱ_ε. As $K_{\sigma_k} \in \mathcal{E}_k(1)$, we can apply Theorem 6.12 to bound the compatibility constant as

$$C_\kappa \leq 2\sqrt{2k}\sqrt{L_{\mathcal{L}}^2 + 2D_{\mathcal{L}}^2} \leq 4\sqrt{3}\sqrt{k}R^2. \tag{184}$$

G.4 Covering numbers

To control the covering numbers we need first to control those of the parameter set. Since $\mathbf{\Theta} = \mathcal{B}_{\mathbb{R}^d,\|\cdot\|_{\mathbf{\Sigma}}}(0,R)$ we have for all $\delta > 0$, exactly as for the case of Diracs,

$$\mathcal{N}(\varrho_\varepsilon, \mathbf{\Theta}, \delta) = \mathcal{N}(\|\cdot\|_{\mathbf{\Sigma}}/\varepsilon, \mathbf{\Theta}, \delta) = \mathcal{N}(\|\cdot\|_{\mathbf{\Sigma}}, \mathbf{\Theta}, \delta\varepsilon) \overset{\text{Lemma B.4}}{\leq} \max\left(1, \left(\tfrac{4R}{\delta\varepsilon}\right)^d\right) = \max\left(1, \left(\tfrac{4R/\varepsilon}{\delta}\right)^d\right). \tag{185}$$

We now establish the existence of a tangent approximation $\mathcal{V}$ to the set of dipoles.

Consider a dipole $\mu = \pi_\theta - \pi_{\theta'}$ where by definition we have $\varrho_\varepsilon(\theta, \theta') \leq 1$. The proof is a minor variant of the technique used for Diracs: we primarily observe that given any tempered distribution ν we have

$$\left\| \frac{\mu}{\|\mu\|_\kappa} - \nu \right\|_\Phi = \sup_\omega C_\lambda \cdot \left| \frac{e^{\jmath\omega^T\theta} - e^{\jmath\omega^T\theta'}}{\|\pi_\theta - \pi_{\theta'}\|_\kappa} e^{-\frac{\|\omega\|^2_{\mathbf{\Sigma}^{-1}}}{2}} - \psi_\nu(\omega) \right| = \sup_\omega f(\omega) \cdot \left| \frac{e^{\jmath\omega^T(\theta-\theta')} - 1}{\|\pi_\theta - \pi_{\theta'}\|_\kappa} - e^{-\jmath\omega^T\theta'} e^{\frac{\|\omega\|^2_{\mathbf{\Sigma}^{-1}}}{2}} \psi_\nu(\omega) \right| \tag{186}$$

with $f(\omega) := C_\lambda e^{-\frac{\|\omega\|^2_{\mathbf{\Sigma}^{-1}}}{2}}$ and ψ_ν the characteristic function of ν (up to a normalization):

Considering $\nu := \delta'_{\theta',-\Delta} \star \pi_0$ the derivative of the Gaussian distribution with mean θ' and covariance $\mathbf{\Sigma}$ along direction $-\Delta$, we get $\psi_\nu(\omega) = e^{\jmath\omega^T\theta'} e^{-\frac{\|\omega\|^2_{\mathbf{\Sigma}^{-1}}}{2}} \jmath\omega^T\Delta$. All the reasoning done with Diracs can be adapted to show that the set of smooth functions

$$\mathcal{V} \quad := \quad \left\{ \nu = \delta'_{\theta',-\Delta} * \pi_0 : \ \theta' \in \mathbf{\Theta}, \ \Delta \in \mathbb{R}^d, \|\Delta\|_2 \leq \varepsilon \right\}$$

is a tangent approximation to the set of dipoles, with constants $t = 2$ and

$$T := \varepsilon^2 \sup_\omega f(\omega)/2 = \varepsilon^2 C_\lambda/2. \tag{187}$$

We remarked in Section G.1 that $\varepsilon \geq \sqrt{2}$ and $C_\lambda \geq 1$ hence $T \geq 1$. Similarly, one can control the covering numbers (with respect to $\|\cdot\|_\Phi$) of

$$\mathcal{V}' \quad := \quad \left\{ \alpha\delta_{\theta',-\Delta} \star \pi_0 + \beta\pi_\theta | \ \theta, \theta' \in \mathbf{\Theta}, \ \Delta \in \mathbb{R}^d, \|\Delta\|_2 \leq \varepsilon, \ 0 \leq \alpha \leq 2, \ 0 \leq \beta \leq 1 \right\}.$$

as in the case of Diracs, simply by replacing the bound $\left|\omega^T\mathbf{v}\right| \leq \|\omega\|_2 \|\mathbf{v}\|_2$ with $\left|\omega^T\mathbf{v}\right| \leq \|\omega\|_{\mathbf{\Sigma}^{-1}} \|\mathbf{v}\|_{\mathbf{\Sigma}}$ and using

$$L_1 \quad = \quad 2\varepsilon \sup_\omega f(\omega) \|\omega\|^2_{\mathbf{\Sigma}^{-1}} = 2\varepsilon C_\lambda \sup_{u>0} u e^{-u/2} = 4\varepsilon C_\lambda/e, \tag{188}$$

$$L_2 \quad = \quad \sup_\omega f(\omega) \|\omega\|_{\mathbf{\Sigma}^{-1}} = C_\lambda \sup_{u>0} u e^{-u^2/2} = C_\lambda/\sqrt{e}, \tag{189}$$

$$L_3 \quad = \quad D_\Phi = C_\lambda \tag{190}$$

$$L_4 \quad = \quad L_2. \tag{191}$$

This yields a control of the covering number of $\mathcal{V}'$ as in (157) with these values of L_i.

G.5 Summary: establishing the LRIP

With the above we can establish that with high probability on the draw of frequencies, the sketching operator $\mathcal{A}$ satisfies the LRIP simultaneously for all models $\mathcal{H} \subset \mathcal{H}_{k,2\varepsilon,R}$.

Consider first $\mathcal{H} = \mathcal{H}_{k,2\varepsilon,R}$. The compatibility constant (184) with respect to $\mathfrak{S}_0 := \mathfrak{S}_{\mathcal{H}_{k,2\varepsilon,R}}$ is

$$C_\kappa \leq 4\sqrt{3}\sqrt{k}R^2.$$

The concentration constant (178) with respect to $\mathfrak{S}_0$ is

$$W_\kappa \leq 4\sqrt{k}\varepsilon C_\lambda.$$

and $D_\Phi := C_\lambda$, $L_\Phi := \varepsilon C_\lambda$, $W_\Phi \leq \sqrt{2}\varepsilon C_\lambda$, and

From (187) we have $T := \varepsilon^2 C_\lambda/2 \geq 1$, hence $16T\min(3/4, t/2) = 12T \geq 1$, and by Theorem 6.15 we have: for any $0 < \delta \leq 1$

$$\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}, \delta\right) \leq \mathcal{N}\left(\varrho_\varepsilon, \mathbf{\Theta}, \tfrac{\delta^2}{C_2}\right)^2 \cdot \max\left(1, \left(\tfrac{C_3}{\delta}\right)^2\right) + \mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{V}', \tfrac{\delta}{C_4}\right) \tag{192}$$

with $C_2 := 256 W_\Phi L_\Phi T$, $C_3 := 16\sqrt{W_\Phi D_\Phi T}$, $C_4 := 4$.

As $C_2 = \mathcal{O}\left(\varepsilon^4 C_\lambda^3\right)$ and $\varepsilon \geq 1$, $C_\lambda \geq 1$, by (185) we have for any $0 < \delta \leq 1$, with $C \geq 1$ denoting some universal constant that may change from equation to equation,

$$\mathcal{N}\left(\varrho_\varepsilon, \mathbf{\Theta}, \delta^2/C_2\right) \leq \max\left(1, (4C_2)^d (R/\varepsilon)^d (1/\delta)^{2d}\right) \leq \varepsilon^{4d} C_\lambda^{3d} (R/\varepsilon)^d (C/\delta)^{2d}.$$

As $C_3 \leq 16\sqrt{\varepsilon^3 C_\lambda^3/\sqrt{2}} = \mathcal{O}\left((\varepsilon C_\lambda)^{3/2}\right)$, we get for any $0 < \delta \leq 1$

$$\left(\mathcal{N}\left(\varrho_\varepsilon, \mathbf{\Theta}, \delta^2/C_2\right)\right)^2 \cdot \max\left(1, (C_3/\delta)^2\right) \leq \varepsilon^{8d+3} C_\lambda^{6d+3} (R/\varepsilon)^{2d} (C/\delta)^{4d+2} \tag{193}$$

Further, as

$$\begin{aligned} 16L_1 R &= \tfrac{2^6}{e} \cdot C_\lambda \cdot \varepsilon^2 \cdot \tfrac{R}{\varepsilon} \\ 16L_2 R &= \tfrac{2^4}{\sqrt{e}} \cdot C_\lambda \cdot \varepsilon \cdot \tfrac{R}{\varepsilon} \\ 32L_2\varepsilon &= \tfrac{2^5}{\sqrt{e}} \cdot C_\lambda \cdot \varepsilon \\ 8L_3 &= 2^3 C_\lambda. \end{aligned}$$

we have $(16L_1R)(16L_2R)(32L_2\varepsilon) = \mathcal{O}\left(\varepsilon^4 C_\lambda^3 (R/\varepsilon)^2\right)$. By (157) we obtain for any $0 < \delta \leq 1$

$$\begin{aligned} \mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{V}', \delta/C_4\right) &\leq \varepsilon^{4d} C_\lambda^{3d} (R/\varepsilon)^{2d} \cdot C_\lambda \cdot (C/\delta)^{3d+1} \\ &= \underbrace{\left(\varepsilon^{-4d-3} C_\lambda^{-3d-2} (\delta/C)^{d+1}\right)}_{\leq 1} \cdot \varepsilon^{8d+3} C_\lambda^{6d+3} (R/\varepsilon)^{2d} (C/\delta)^{4d+2} \end{aligned} \tag{194}$$

Combining (193)-(194) with (192), we obtain that for any $0 < \delta \leq 1$

$$\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}, \delta\right) \leq \varepsilon^{8d+3} C_\lambda^{6d+3} (R/\varepsilon)^{2d} (C/\delta)^{4d+2} \tag{195}$$

By Theorem 6.13, we have for any $\delta > 0$

$$\mathcal{N}\left(d_\Phi, \mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S}_{k,2}(\mathcal{T})), \delta\right) \leq \left[\mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}, \tfrac{\delta}{C_0}\right) \cdot \max\left(1, \tfrac{C_1}{\delta}\right)\right]^{2k} \tag{196}$$

where $C_0 := 64 k W_\Phi \geq 1$, $C_1 := 256 k W_\Phi^2 \geq 1$.

As $C_0 = \mathcal{O}\left(k\varepsilon C_\lambda\right)$ and $C_1 = \mathcal{O}\left(k\varepsilon^2 C_\lambda^2\right)$, we have $C_0^{4d+2} C_1 = \mathcal{O}\left(C^{4d+3} k^{4d+3} \varepsilon^{4d+4} C_\lambda^{4d+4}\right)$. Moreover, for $0 < \delta \leq 1$ we have $\delta/C_1 \leq 1$ hence we can apply (195) to obtain

$$\begin{aligned} \mathcal{N}\left(\|\cdot\|_\Phi, \mathcal{D}, \tfrac{\delta}{C_0}\right) \cdot \max\left(1, \tfrac{C_1}{\delta}\right) &\leq \varepsilon^{8d+3} C_\lambda^{6d+3} (R/\varepsilon)^{2d} (C/\delta)^{4d+2} \cdot C_0^{4d+2} \cdot C_1/\delta \\ &\leq k^{4d+3} \varepsilon^{12d+7} C_\lambda^{10d+7} (R/\varepsilon)^{2d} (C/\delta)^{4d+3} \end{aligned} \tag{197}$$

Overall this shows that the covering dimension of the normalized secant set if $s = \mathcal{O}(kd)$, and we get for any $0 < \zeta < 1$

$$\log 2\mathcal{N}\left(d_\Phi, \mathcal{S}_{\|\cdot\|_\kappa}(\mathfrak{S}_{k,2}(\mathcal{T})), \delta\right)/\zeta = \mathcal{O}\left(kd \cdot \left[\log(ek) + \log\tfrac{R}{\varepsilon} + \log(\varepsilon C_\lambda)^2 + \log\tfrac{1}{\delta}\right] + \log\tfrac{1}{\zeta}\right)$$

Using (31) with $0 < \delta \leq 1$ we get $c_\kappa(\delta/2) = \mathcal{O}\left(\delta^{-2} W_\kappa^2\right) = \mathcal{O}\left(\delta^{-2} k (\varepsilon C_\lambda)^2\right)$. By Theorem 2.9 for $\mathfrak{S} = \mathfrak{S}_0$ we obtain for $\delta < 1$: if $m \geq m_0$, where

$$m_0 = \mathcal{O}\left(\delta^{-2} k^2 d (\varepsilon C_\lambda)^2 \cdot \left[\log(ek) + \log \tfrac{R}{\varepsilon} + \log(\varepsilon C_\lambda)^2 + \log \tfrac{1}{\delta}\right] + \delta^{-2} k (\varepsilon C_\lambda)^2 \cdot \log \tfrac{1}{\zeta}\right)$$

then with probability at least $1 - \zeta$ on the draw of frequencies $(\omega_j)_{j=1}^m$, the sketching operator satisfies the LRIP (12) on $\mathfrak{S}_0$ with constant $C_{\mathcal{A}} = \frac{C_\kappa}{\sqrt{1-\delta}} = \frac{4\sqrt{3}}{\sqrt{1-\delta}} \cdot \sqrt{k} R^2$.

Extension to $\mathcal{H} \subset \mathcal{H}_{k,2\varepsilon,R}$. When $\mathcal{H} \subset \mathcal{H}_{k,2\varepsilon,R}$ we have $\mathfrak{S}_\mathcal{H} \subset \mathfrak{S}_0$ and the LRIP for $\mathfrak{S}_0$ implies the LRIP for $\mathfrak{S}_\mathcal{H}$. Hence, with probability at least $1 - \zeta$ the sketching operator satisfies the LRIP simultaneously for all models associated to $\mathcal{H} \subset \mathcal{H}_{k,2\varepsilon,R}$.

H Proof of Theorem 3.1 on Compressive PCA

For Compressive PCA, observe that $\mathcal{R}_{\texttt{PCA}}(\pi, h) := \mathbb{E}_{X \sim \pi} \left\| X - \mathrm{P}_h X \right\|_2^2 = \mathrm{Tr}\left(\mathbf{\Sigma}_\pi \mathrm{P}_{h^\perp}\right)$ with $h^\perp$ the orthogonal complement of the subspace h, and the minimum risk is

$$\mathcal{R}_{\texttt{PCA}}(\pi, h^\star) = \inf_{\mathrm{rank}(\mathbf{M}) \leq k, \mathbf{M} \succcurlyeq 0} \left\| \mathbf{\Sigma}_\pi - \mathbf{M} \right\|_\star . \tag{198}$$

Model set $\mathfrak{S}_\mathcal{H}$ and best hypothesis for $\pi \in \mathfrak{S}_\mathcal{H}$. As $\mathcal{R}_{\texttt{PCA}}(\pi, h^\star) = 0$ if, and only if, the covariance matrix $\mathbf{\Sigma}_\pi := \mathbb{E}_{X \sim \pi} X X^T$ has rank at most k, the model set is $\mathfrak{S}_\mathcal{H} = \{\pi : \quad \mathrm{rank}(\mathbf{\Sigma}_\pi) \leq k\}$. For $\pi \in \mathfrak{S}_\mathcal{H}$, the eigen-value decomposition $\mathbf{\Sigma}_\pi = \mathbf{U}\mathbf{D}\mathbf{U}^T$ yields an optimum of (17), $\hat{h} = \mathrm{span}(\mathbf{U}(:, 1:k))$.

Metric associated to the learning task. To design sketching operators satisfying the LRIP (12) on $\mathfrak{S}_\mathcal{H}$, we leverage the well-established body of work on the recovery of low-rank matrices from random projections and establish the apparently new inequality

$$\left\| \pi' - \pi \right\|_{\mathcal{L}(\mathcal{H})} = \max\left(\sum_{i=1}^{d-k} \lambda_i(\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\pi'}), \sum_{i=1}^{d-k} \lambda_i(\mathbf{\Sigma}_{\pi'} - \mathbf{\Sigma}_\pi) \right) \leq \left\| \mathbf{\Sigma}_{\pi'} - \mathbf{\Sigma}_\pi \right\|_\star \tag{199}$$

with $\left\| \cdot \right\|_\star$ the nuclear norm.

Proof of (199). By the definition (10), for any probability distributions π, π' with finite second moments

$$\left\| \pi' - \pi \right\|_{\mathcal{L}(\mathcal{H})} = \sup_{h \in \mathcal{H}} \left| \mathrm{Tr}\left(\left(\mathbf{\Sigma}_{\pi'} - \mathbf{\Sigma}_\pi \right) \mathrm{P}_{h^\perp} \right) \right| . \tag{200}$$

By the so-called Ky Fan Theorem [44], for a symmetric matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$, and a positive integer $\ell \leq d$, one has

$$\sup_{\dim(V) \leq \ell} \mathrm{Tr}(\mathbf{M}\, \mathrm{P}_V) = \sum_{i=1}^{\ell} \lambda_i(\mathbf{M}) ,$$

where $\lambda_i(\mathbf{M})$ denote the eigenvalues, with multiplicity and ordered in nonincreasing sequence, of $\mathbf{M}$. As a result the seminorm defined in (200) is

$$\max\left(\sum_{i=1}^{d-k} \lambda_i(\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\pi'}), \sum_{i=1}^{d-k} \lambda_i(\mathbf{\Sigma}_{\pi'} - \mathbf{\Sigma}_\pi) \right) .$$

Denoting $\sigma_i(\mathbf{M})$ the singular values, ordered in decreasing sequence, of a matrix $\mathbf{M}$, we further have

$$\sum_{i=1}^{d-k} \lambda_i(\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\pi'}) \leq \sum_{i=1}^{d} \sigma_i(\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\pi'}) = \|\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\pi'}\|_\star .$$

This establishes (199). □

Compatibility constant. For any rank-r matrix $\mathbf{M}$, we have $\|\mathbf{M}\|_\star \leq \sqrt{r}\,\|\mathbf{M}\|_F$ where $\|\cdot\|_F$ denotes the Frobenius norm. This implies that if $\pi, \pi' \in \mathfrak{S}_\mathcal{H}$ then

$$\|\pi' - \pi\|_{\mathcal{L}(\mathcal{H})} \leq \sqrt{2k}\,\|\mathbf{\Sigma}_{\pi'} - \mathbf{\Sigma}_\pi\|_F \tag{201}$$

Hence any linear operator $\mathcal{M}$ on matrices having a lower RIP (42) on matrices of rank lower than $2k$ induces (in the way described in Section 3) a sketching operator $\mathcal{A} : \pi \mapsto \mathcal{A}(\pi) := \mathcal{M}(\mathbf{\Sigma}_\pi)$ that has the lower RIP described in (12) with constant $C_\mathcal{A} = \frac{\sqrt{2k}}{\sqrt{1-\delta}}$.

Concentration constant and covering dimension. It is well known that $\mathcal{M}(\mathbf{M}) := (\langle \mathbf{L}_i, \mathbf{M}\rangle_F)_{i=1}^m$ where $\mathbf{L}_i$ has properly standardized i.i.d. (sub)Gaussian entries satisfy the required RIP with constant δ ($\mathcal{M}$ satisfies inequations (42)) with high probability provided the sketch size is of the order $m \gtrsim \mathcal{O}(kd)$. Technically this is proved by establishing that for a given δ, $c_\kappa(\delta/2) = \mathcal{O}(1)$ and that the covering dimension of the normalized secant set is $s = \mathcal{O}(kd)$.

Ideal decoder and generic excess risk control. The ideal decoder (13) writes

$$\Delta[\mathbf{y}] := \mathrm{argmin}_{\pi \in \mathfrak{S}_\mathcal{H}} \|\mathcal{A}(\pi) - \mathbf{y}\|_2^2 := \mathrm{argmin}_{\mathbf{\Sigma}:\mathrm{rank}(\mathbf{\Sigma})\leq k; \mathbf{\Sigma}\succcurlyeq 0} \|\mathcal{M}(\mathbf{\Sigma}) - \mathbf{y}\|_2^2$$

which matches (43). By the lower RIP on $\mathcal{A}$, this decoder is instance optimal yielding (15) with the bias term defined in (14), i.e., the excess risk of $\hat{h}$ from the procedure of Section 3 is controlled with

$$\eta_n \leq 2D(\pi_0, \mathfrak{S}_\mathcal{H}) + 4C_\mathcal{A}\,\|\mathcal{A}(\pi - \hat{\pi}_n)\|_2 .$$

Control of the bias term. Using (199) yields

$$D(\pi_0, \mathfrak{S}_\mathcal{H}) \leq \inf_{\mathrm{rank}(\mathbf{\Sigma})\leq k, \mathbf{\Sigma}\succcurlyeq 0} \{\|\mathbf{\Sigma}_\pi - \mathbf{\Sigma}\|_\star + 2C_\mathcal{A}\,\|\mathcal{M}(\mathbf{\Sigma}_\pi - \mathbf{\Sigma})\|_2\}$$

By the upper RIP (the rhs inequality in (42)), we have for any $\mathbf{M}$:

$$\|\mathcal{M}(\mathbf{\Sigma}_\pi - \mathbf{M})\|_2 \leq \sqrt{1+\delta}\,\|\mathbf{\Sigma}_\pi - \mathbf{M}\|_\star \tag{202}$$

Proof of (202). Decompose $\mathbf{\Sigma}_\pi - \mathbf{M}$ as a sum of orthogonal rank-1 matrices $\mathbf{\Sigma}_\pi - \mathbf{M} = \sum_i \mathbf{M}_i$ (the SVD of $\mathbf{\Sigma}_\pi - \mathbf{M}$). With the triangle inequality and the upper RIP in (42), we have $\|\mathcal{M}(\mathbf{\Sigma}_\pi - \mathbf{M})\|_2 \leq \sqrt{1+\delta}\sum_i \|\mathbf{M}_i\|_F = \sqrt{1+\delta}\,\|\mathbf{\Sigma}_\pi - \mathbf{M}\|_\star$. □

As a result, the bias term is bounded by $(1 + 2C_\mathcal{A}\sqrt{1+\delta})\inf_{\mathrm{rank}(\mathbf{\Sigma})\leq k, \mathbf{\Sigma}\succcurlyeq 0} \|\mathbf{\Sigma}_\pi - \mathbf{\Sigma}\|_\star$. Given the expression of the minimum risk (198) and the fact that $C_\mathcal{A} = \frac{\sqrt{2k}}{\sqrt{1-\delta}}$, we obtain

$$\eta_n \leq 2(1 + \tfrac{2\sqrt{2k}}{\sqrt{1-\delta}}\sqrt{1+\delta})\mathcal{R}_{\texttt{PCA}}(\pi, h^\star) + 4\tfrac{\sqrt{2k}}{\sqrt{1-\delta}}\,\|\mathcal{A}(\pi - \hat{\pi}_n)\|_2 .$$

This proves Theorem 3.1.

References

[1] Dimitris Achlioptas and Frank Mcsherry. On spectral learning of mixtures of distributions. *Learning Theory*, pages 458–469, 2005.

[2] Peter Ahrendt. The Multivariate Gaussian Probability Distribution. Technical report, IMM, Technical University of Denmark, 2005.

[3] Nir Ailon, Ragesh Jaiswal, and Claire Monteleoni. Streaming k -means approximation. *Advances in Neural Information Processing Systems (NIPS)*, pages 10–18, 2009.

[4] Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James Voss. The more, the merrier: the blessing of dimensionality for learning large gaussian mixtures. *arXiv:1311.2891*, 35:1–30, 2013.

[5] Christophe Andrieu and Arnaud Doucet. Online expectation-maximization type algorithms for parameter estimation in general state space models. *ICASSP (6)*, (4), 2003.

[6] R. Arora, A. Cotter, K. Livescu, and N. Srebro. Stochastic optimization for pca and pls. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 861–868, Oct 2012.

[7] David Arthur and Sergei Vassilvitskii. k-means++: The Advantages of Careful Seeding. In *ACM-SIAM symposium on Discrete algorithms*, pages 1027—-1035, 2007.

[8] Francis Bach. On the Equivalence between Kernel Quadrature Rules and Random Feature Expansions. pages 1–38, February 2015.

[9] Akshay Balsubramani, Sanjoy Dasgupta, and Yoav Freund. The fast convergence of incremental pca. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3174–3182. Curran Associates, Inc., 2013.

[10] Richard Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4):118–121, 2007.

[11] Richard Baraniuk, Mark Davenport, Ronald A DeVore, and Michael B Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28(3):253–263, 2008.

[12] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *IEEE 51st Annual Symposium on Foundations of Computer Science*. Ieee, 2010.

[13] Mikhail Belkin and Kaushik Sinha. Toward Learning Gaussian Mixtures with Arbitrary Separation. In *Conference On Learning Theory (COLT)*, 2010.

[14] K Bertin, E Le Pennec, and V Rivoirard. Adaptive Dantzig density estimation. *Annales De L Institut Henri Poincare-Probabilites Et Statistiques*, 47(1):43–74, February 2011.

[15] M Bojarski, A Choromanska, and K Choromanski. Structured adaptive and random spinners for fast machine learning computations. *arXiv*, 2016.

[16] A. Bourrier, M.E. Davies, T. Peleg, P. Perez, and R. Gribonval. Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems. *Information Theory, IEEE Transactions on*, 60(12):7928–7946, Dec 2014.

[17] Anthony Bourrier, Remi Gribonval, and Patrick Perez. Compressive Gaussian Mixture Estimation. In *ICASSP*, pages 6024–6028, Vancouver, Canada, March 2013.

[18] Emmanuel Candès, Thomas Strohmer, and Vladislav Voroninski. PhaseLift: Exact and Stable Signal Recovery from Magnitude Measurements via Convex Programming. *Comm. Pure Appl. Math*, 66(8):1241–1274, 2013.

[19] Emmanuel J Candès and Carlos Fernandez-Granda. Super-resolution from noisy data. *Journal of Fourier Analysis and Applications*, 19(6):1229–1254, 2013.

[20] Emmanuel J Candès, Justin Romberg, and Terence Tao. Stable Signal Recovery from Incomplete and Inaccurate Measurements. *Comm. Pure Appl. Math*, 59:1207–1223, 2006.

[21] Olivier Cappé and Eric Moulines. Online EM Algorithm for Latent Data Models. *Journal of the Royal Statistical Society*, 71(3):593–613, 2009.

[22] Marine Carrasco and Jean-Pierre Florens. Generalization of GMM to a continuum of moment conditions. *Econometric Theory*, 2000.

[23] Marine Carrasco and Jean-Pierre Florens. Efficient GMM estimation using the empirical characteristic function. *IDEI Working Paper*, 140, 2002.

[24] Marine Carrasco and Jean-Pierre Florens. On The Asymptotic Efficiency Of Gmm. *Econometric Theory*, 30(02):372–406, 2014.

[25] Choromanski, Krzysztof and Sindhwani, Vikas. Recycling Randomness with Structure for Sublinear time Kernel Expansions. *ICML*, May 2016.

[26] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc*, 22(1), 2009.

[27] Graham Cormode, Minos Garofalakis, Peter J. Haas, and Chris Jermaine. Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches. *Foundations and Trends in Databases*, 4(xx):1–294, 2011.

[28] Graham Cormode, Minos N Garofalakis, Peter J Haas, and Chris Jermaine. Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches. *FTDB) 4(1-3*, 4(1-3):1–294, 2012.

[29] Graham Cormode and Marios Hadjieleftheriou. Methods for finding frequent items in data streams. *The VLDB Journal*, 19(1):3–20, 2009.

[30] Graham Cormode and S Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *J. Algorithms*, 55(1):58–75, 2005.

[31] Graham Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.

[32] T M Cover and J A Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. Wiley-Interscience, 1991.

[33] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

[34] Sanjoy Dasgupta and Leonard J. Schulman. A Two-Round Variant of EM for Gaussian Mixtures. *Uncertainty in Artificial Intelligence*, pages 152–159, 2000.

[35] Y De Castro, F Gamboa, Didier Henrion, and J-B Lasserre. Exact solutions to super resolution on semi-algebraic domains in higher dimensions. *arXiv preprint arXiv:1502.02436*, 2015.

[36] Sjoerd Dirksen. Dimensionality reduction with subgaussian matrices: a unified theory. *arXiv*, February 2014.

[37] David L Donoho. Compressed sensing. *IEEE Trans. Information Theory*, 52(4):1289–1306, 2006.

[38] John Duchi. Derivations for Linear Algebra and Optimization. Technical report, 2007.

[39] John C Duchi, Michael I Jordan, and Martin J Wainwright. Privacy Aware Learning. October 2012.

[40] R M Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

[41] Vincent Duval and Gabriel Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.

[42] Armin Eftekhari and Michael B Wakin. New Analysis of Manifold Embeddings and Signal Recovery from Compressive Measurements. *arXiv*, June 2013.

[43] Joan Bruna Estrach, Arthur Szlam, and Yann LeCun. Signal recovery from Pooling Representations. *ICML*, 2014.

[44] K Fan. On a Theorem of Weyl Concerning Eigenvalues of Linear Transformations I. *Proc. Nat. Aca. Sci.*, 35(11):652–655, November 1949.

[45] Alexei A Fedotov, Peter Harremoës, and Flemming Topsøe. Refinements of Pinsker's Inequality. *IEEE Trans. Inf. Theor.*, 49(6):1491–1498, 2003.

[46] Dan Feldman, Matthew Faulkner, and Andreas Krause. Scalable Training of Mixture Models via Coresets. *Proceedings of Neural Information Processing Systems*, pages 1–9, 2011.

[47] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. *Proceedings of the forty-third annual ACM symposium on Theory of computing*, (46109):569–578, 2011.

[48] Dan Feldman, Morteza Monemizadeh, Christian Sohler, and David P Woodruff. Coresets and Sketches for High Dimensional Subspace Approximation Problems. 1:630–649, 2010.

[49] Andrey Feuerverger and Roman A Mureika. The Empirical Characteristic Function and Its Applications. *Annals of Statistics*, 5(1):88–97, January 1977.

[50] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, May 2012.

[51] Gereon Frahling and Christian Sohler. A fast k -means implementation using coresets. *Proceedings of the twenty-second annual symposium on Computational geometry (SoCG)*, 18(6):605–625, 2005.

[52] Mina Ghashami, Daniel Perry, and Jeff M. Phillips. Streaming Kernel Principal Component Analysis. *International Conference on Artificial Intelligence and Statistics*, 41:1–16, 2016.

[53] Anna C Gilbert, Yannis Kotidis, S Muthukrishnan, and Martin J Strauss. How to summarize the universe: dynamic maintenance of quantiles. In *VLDB '02: Proceedings of the 28th international conference on Very Large Data Bases*, pages 454–465. VLDB Endowment, 2002.

[54] Anna C Gilbert, Yi Zhang, Kibok Lee, Yuting Zhang, and Honglak Lee. Towards Understanding the Invertibility of Convolutional Neural Networks. May 2017.

[55] Raja Giryes, Guillermo Sapiro, and Alexander M Bronstein. Deep Neural Networks with Random Gaussian Weights - A Universal Classification Strategy? *IEEE Trans. Signal Processing*, 2016.

[56] Sudipto Guha and Et al. Clustering Data Streams. 2000.

[57] Alastair R. Hall. *Generalized method of moments*. 2005.

[58] Paul R Halmos. *Measure theory*, volume 18. Springer, 2013.

[59] Sariel Har-Peled and Soham Mazumdar. Coresets for k-Means and k-Median Clustering and their Applications. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 291—-300, 2004.

[60] Daniel Hsu and Sham M. Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Conference on Innovations in Theoretical Computer Science*, 2013.

[61] Maryia Kabanava, Richard Kueng, Holger Rauhut, and Ulrich Terstiege. Stable low-rank matrix recovery via null space properties. *Information and Inference*, 5(4):405–441, 2016.

[62] Nicolas Keriven, Anthony Bourrier, Rémi Gribonval, and Patrick Pérèz. Sketching for Large-Scale Learning of Mixture Models. In *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2015.

[63] Nicolas Keriven, Anthony Bourrier, Rémi Gribonval, and Patrick Pérèz. Sketching for Large-Scale Learning of Mixture Models. *arXiv preprint arXiv:1606.02838*, pages 1–50, 2016.

[64] Nicolas Keriven, Nicolas Tremblay, Yann Traonmilin, and Rémi Gribonval. Compressive k-means. *ICASSP 2017*, 2017.

[65] Henry J. Landau. *Moments in mathematics*. 1987.

[66] Le, Quoc, Sarlós, Tamás, and Smola, Alex J. Fastfood — Approximating Kernel Expansions in Loglinear Time. *arXiv*, 28(1):1–29, 2013.

[67] Clément Levrard. Fast rates for empirical vector quantization. *Electronic Journal of Statistics*, 7(0):1716–1746, 2013.

[68] Qiuwei Li and Gongguo Tang. The Nonconvex Geometry of Low-Rank Matrix Optimizations with General Objective Functions. November 2016.

[69] Mario Lucic, Matthew Faulkner, Andreas Krause, and Dan Feldman. Training Mixture Models at Scale via Coresets. 2017.

[70] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online Learning for Matrix Factorization and Sparse Coding. *Journal of Machine Learning Research*, 11(1):19–60, January 2010.

[71] Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. 1994.

[72] Gilles Puy, Michael E Davies, and Remi Gribonval. Recipes for Stable Linear Embeddings From Hilbert Spaces to $\mathbb{R}^m$. *IEEE Trans. Information Theory*, 63(4):2171–2187, 2017.

[73] Ali Rahimi and Benjamin Recht. Random Features for Large Scale Kernel Machines. *Advances in Neural Information Processing Systems (NIPS)*, (1):1–8, 2007.

[74] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in Neural Information Processing Systems (NIPS)*, 1(1):1–8, 2009.

[75] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is More: Nyström Computational Regularization. *arXiv*, July 2015.

[76] Alexander J. Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert Space Embedding for Distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31, 2007.

[77] K Sridharan. A gentle introduction to concentration inequalities. *Dept Comput Sci*, 2002.

[78] Bharath K Sriperumbudur and Zoltán Szabó 0001. Optimal Rates for Random Fourier Features. *NIPS*, 2015.

[79] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

[80] Nitin Thaper, Sudipto Guha, Piotr Indyk, and Nick Koudas. Dynamic multidimensional histograms. In *SIGMOD '02: Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, pages 428–439, New York, NY, USA, 2002. ACM.

[81] Yann Traonmilin and Rémi Gribonval. Stable recovery of low-dimensional cones in hilbert spaces: One RIP to rule them all. *Applied and Computational Harmonic Analysis*, in press, 2016.

[82] Joel A Tropp. User-friendly tail bounds for Sums of Random Matrices. *Foundations of Computational Mathematics*, 2011.

[83] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.