Compressive Statistical Learning with Random Feature Moments

Rémi Gribonval, Gilles Blanchard, Nicolas Keriven, Yann Traonmilin

Compressive Statistical Learning with Random Feature Moments

Rémi Gribonval* remi.gribonval@inria.fr
Gilles Blanchard † gilles.blanchard@universite-paris-saclay.fr
Nicolas Keriven‡ nicolas.keriven@gipsa-lab.grenoble-inp.fr
Yann Traonmilin§ yann.traonmilin@math.u-bordeaux.fr

April 16, 2020

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 PCA, compressive clustering, and compressive Gaussian mixture Modeling with fixed known variance. The latter two are further developed in a companion paper.

Keywords: Kernel mean embedding, random features, random moments, statistical learning, dimension reduction

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, $\mathcal{R}(\pi, h) := \mathbb{E}_{X\sim\pi}\ell(X, h)$ over a parameterized family of hypotheses $\mathcal{H}$ (where π 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

*Univ Lyon, Inria, CNRS, ENS de Lyon, UCB Lyon 1, LIP UMR 5668, F-69342, Lyon, France
This work was initiated while R. Gribonval, N. Keriven and Y. Traonmilin were with Univ Rennes, Inria, CNRS, IRISA F-35000 Rennes, France
†Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, F-91405, Orsay, France.
‡CNRS, GIPSA-lab, UMR 5216, F-38400 Saint-Martin-d'Hères, France.
§CNRS, Univ. Bordeaux, Bordeaux INP, IMB, UMR 5251, F-33400 Talence, France.

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. Mairal et al., 2010], sampling techniques such as coresets [Feldman and Langberg, 2011] or Nyström's method [e.g. Rudi et al., 2015] 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 [Cormode et al., 2011] (see Section 1.3). Rooted both in the generalized method of moments [Hall, 2005] and in compressive sensing [Foucart and Rauhut, 2012], it leverages techniques from kernel methods such as kernel mean embeddings [Gretton et al., 2007, Sriperumbudur et al., 2010] and random Fourier features [Rahimi and Recht, 2007] 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. As shown in the companion paper [Gribonval et al., 2020], 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 (centered) 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 $\ell(x, h) = \|x - P_h x\|_2^2$ with P_h the orthogonal projector onto h. The matrix of second moments $\mathbf{\Sigma}_\pi := \mathbb{E}_{X \sim \pi} X X^T$ 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 [Foucart and Rauhut, 2012] allow to compress the matrix of second moments to a sketch of dimension of the order of 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 4). 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 statistical 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.

In the companion paper [Gribonval et al., 2020], we exploit this framework to establish statistical learning guarantees for compressive clustering and compressive Gaussian mixture estimation. We conclude this paper by briefly discussing 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* [Cormode et al., 2011], 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 [Cormode et al., 2011] such as the popular detection of heavy-hitters [Cormode and Muthukrishnan, 2005a,b, Cormode and Hadjieleftheriou, 2009]. Closer to our framework, sketches have been used to approximately maintain histograms [Thaper et al., 2002] or quantiles [Gilbert et al., 2002], 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 [Andrieu and Doucet, 2003, Cappé and Moulines, 2009], the k-means algorithm [Guha and al., 2000, Ailon et al., 2009], or Principal Component Analysis [Ghashami et al., 2016]. In each case, the result of the algorithm is updated as new data becomes available. 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 [Har-Peled and Mazumdar, 2004] or, more generally, subspace approximation [Feldman et al., 2010, Feldman and Langberg, 2011] and also applied to learning Gaussian Mixture Models [Feldman et al., 2011, Lucic et al., 2017]. 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 [Frahling and Sohler, 2005] already incorporates steps of Lloyd's k-means algorithm in its construction. Similar to the k-means++ algorithm [Arthur and Vassilvitskii,

2007a], many coresets have been developed as (weighted) adaptive subsampling of the data [Feldman et al., 2011, Lucic et al., 2017].

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 Har-Peled and Mazumdar, 2004, Feldman and Langberg, 2011, Lucic et al., 2017] 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 the companion paper [Gribonval et al., 2020] for k-means 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) [Landau, 1987, Hall, 2005]: 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 [Belkin and Sinha, 2010, Hsu and Kakade, 2013, Anderson et al., 2014] or when the likelihood is not available. Traditionally, a finite number of moments is considered, but modern developments give guarantees when an infinite (integral) number of generalized moments are available [Carrasco and Florens, 2000, 2014], in particular generalized moments associated to the (empirical) characteristic function [Carrasco and Florens, 2002, Feuerverger and Mureika, 1977]. Our point of view is slightly different: we consider the collection of moments as a *compressed* representation of the data and as a means 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 Models, we prove in the companion paper [Gribonval et al., 2020] 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 practitioners [see Newey and McFadden, 1994, p. 2127] by the strictly stronger Lower Restricted Isometry Property (LRIP) from the Compressive Sensing literature [Donoho, 2006, Candès et al., 2006, Baraniuk, 2007, Foucart and Rauhut, 2012]. This is achieved by considering *random* feature moments (related to random features [Rahimi and Recht, 2007, 2009, Bach, 2017] and kernel mean embeddings [Sriperumbudur et al., 2010]), 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. Sriperumbudur and Szabó, 2015, Bach, 2017], 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. An important concept is the notion of Lower Restricted Isometry Property (LRIP) using the notion of a model set (a set of "simple" probability distributions) which is futher discussed in Section 3. To illustrate the proposed framework, we detail in Section 4 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 [Arora et al., 2012, Balsubramani et al., 2013] but rather to give a first illustration. Generic techniques to establish the LRIP property for sketches of controlled size are described in Section 5. In the companion paper [Gribonval et al., 2020], we specify a sketching procedure and state the associated learning guarantees for compressive clustering and compressive Gaussian mixture estimation. We discuss in Section 6 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 Appendix.

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 (supervised or unsupervised) 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 on the measurable space $(\mathcal{Z}, \mathfrak{Z})$. In our examples, $\mathcal{Z} = \mathbb{R}^d$ endowed with the Borel sigma-algebra $\mathfrak{Z}$. 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* $\ell : (x, h) \mapsto \ell(x, h) \in \mathbb{R}$ and the *expected risk* associated to h:

$$\mathcal{R}(\pi, h) := \mathbb{E}_{X\sim\pi}\ \ell(X, h)\,,$$

where (here and in the sequel) we will always assume that we restrict our attention to probability distributions π such that $x \mapsto \ell(x, h)$ is measurable and π-integrable for all $h \in \mathcal{H}$. In the idealized learning problem, one selects a function $h_\pi^\star$ that minimizes the expected risk (we will assume existence of this minimum for a simpler presentation, although most statements to come can be transformed if needed using a sequence of approximate minimizers)

$$h_\pi^\star \in \arg\min_{h\in\mathcal{H}} \mathcal{R}(\pi, h). \tag{1}$$

We will use the shorthand $h^\star$ for $h_\pi^\star$ whenever there is no ambiguity from the context. In practice one has no access to the true risk $\mathcal{R}(\pi, h)$ since the expectation with respect to the underlying probability distribution, $\mathbb{E}_{X\sim\pi}[\cdot]$, is unavailable. Instead, methods such as *empirical risk minimization (ERM)* produce an estimated hypothesis $\hat{h}$ from the training dataset $\mathbf{X}$ by minimizing the risk $\mathcal{R}(\hat{\pi}_n, \cdot)$ (or a regularized version) associated to the *empirical probability distribution* $\hat{\pi}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ of the training samples. One expects to produce, with high probability at least $1-\zeta$ on the draw of the training set, the bound on the excess risk

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^\star) \leq \eta_n = \eta_n(\zeta), \tag{2}$$

where η_n typically decays as $1/\sqrt{n}$ or better. We will use the following running examples.

Examples:

- **PCA**: as stated in the introduction, the loss function is $\ell(x, h) = \|x - P_h x\|_2^2$ 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, $h = \{c_1, \dots, c_k\}$, and the loss is defined by the k-means cost $\ell(x, h) = \min_{1 \le l \le k} \|x - c_l\|_2^2$. The hypothesis class $\mathcal{H}$ may be further reduced by defining constraints on the considered centers (e.g., in some domain, or as detailed in the companion paper [Gribonval et al., 2020] with some separation between centers).
- **Gaussian Mixture Modeling**: each hypothesis h corresponds to the collection of weights, means and variances of a mixture of k Gaussians, whose 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 (2) 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 [Donoho, 2006, Candès et al., 2006, Foucart and Rauhut, 2012] and the generalized method of moments [Landau, 1987, Hall, 2005], but also on kernel mean embeddings [Smola et al., 2007, Sriperumbudur et al., 2010], random features [Rahimi and Recht, 2007, 2009, Bach, 2017], and streaming algorithms [Gilbert et al., 2002, Cormode and Muthukrishnan, 2005a, Cormode et al., 2011], *compressive learning* relies on the choice of a measurable (nonlinear) *feature function* $\Phi : \mathcal{Z} \to \mathbb{R}^m$ or $\mathbb{C}^m$ and has two main steps:

1. Compute generalized empirical moments using the feature function on the training collection to summarize it into a single *sketch vector*

$$\mathbf{y} := \texttt{Sketch}(\mathbf{X}) := \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \in \mathbb{R}^m \text{or } \mathbb{C}^m; \tag{3}$$

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

Overall, the goal is to design the sketching function $\Phi(\cdot)$ and the learning procedure $\texttt{Learn}(\cdot)$ given a learning task (i.e., a loss function) such that the resulting hypothesis $\hat{h}$ has controlled excess risk (2) (if Φ is drawn at random according to some specification, we want (2) to hold with high probablity also with respect to the draw of Φ). To anticipate, let us mention that learning from a sketch will take the form of a minimization problem

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} R(\mathbf{y}, h) \tag{4}$$

where in a sense $R(\mathbf{y}, \cdot)$ will play the role of a proxy for the empirical risk $\mathcal{R}(\hat{\pi}_n, \cdot)$.

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 $\mathbf{\Sigma}_\pi = \mathbb{E}_{X\sim\pi} XX^T$, 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 of the order of kd from which one can accurately estimate the best rank-k approximation to $\mathbf{\Sigma}_\pi$, 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 4 as a first illustration of our framework.

In the companion paper [Gribonval et al., 2020], 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 specific proxy (4) corresponding to a non-convex least-squares minimization) that leads to a control of the excess risk.

As described below, these results are 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 $\hat{\pi}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$:

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

where $\mathcal{A}$ is a linear operator from the set of distributions π such that Φ is integrable with respect to π, to $\mathbb{R}^m$ (or $\mathbb{C}^m$), defined by

$$\mathcal{A}(\pi) := \mathbb{E}_{X\sim\pi}\Phi(X). \tag{5}$$

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 $0 \le \theta \le 1$.

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

Conceptually, we construct the learning-from-sketch procedure $\hat{h} = \texttt{Learn}(\mathbf{y})$ in two steps:

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

$$\widetilde{\pi} = \Delta[\mathbf{y}];$$

- Find a best hypothesis from this estimate:

$$\hat{h} \in \arg\min_{h\in\mathcal{H}} \mathcal{R}(\widetilde{\pi}, h). \tag{6}$$

[1] One can extend $\mathcal{A}$ to a linear operator on the space of finite signed measures such that Φ is integrable, see Appendix A.2.

As a first coarse analysis of this scheme, notice that if the decoder step is such that a uniform approximation between the risk of $\widetilde{\pi}$ and of π holds:

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

then we will be able to control the excess risk (2) – our goal. Indeed, using (6) and the triangle inequality, it is easy to show that (7) directly implies (2). (We will see later that (7) can be too coarse and will introduce a more refined analysis based on excess risks in Section 2.5.) 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 $\mathcal{R}(\widetilde{\pi},\cdot)$ where $\widetilde{\pi}$ 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* non-parametric 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 in the standard sense of density estimation as, e.g., in [Bertin et al., 2011], but only to accurately estimate the risk $\mathcal{R}(\pi,\cdot)$. On practical examples, a natural decoder will be based on best moment matching over a parametric family of probability distributions, which will be expressed more directly as the minimization of a proxy for the risk* (4)*, cf Section 3.1.*

2.4 Statistical learning guarantees: a first control of the excess risk

In this section, for simplicity we first focus on how to establish uniform control of the risks of the form (7) using general results from linear inverse problems (we shall introduce in the next section a sharper but also slightly more notation-heavy analysis). 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 $\mathcal{G}$ of measurable functions $f : \mathcal{Z} \to \mathbb{R}$ or $\mathbb{C}$, we use the following notation throughout this work:

$$\|\pi - \pi'\|_{\mathcal{G}} := \sup_{f\in\mathcal{G}} |\mathbb{E}_{X\sim\pi} f(X) - \mathbb{E}_{X'\sim\pi'} f(X')|, \tag{8}$$

which defines a semi-norm on the space of finite signed measures (see Appendix A.2) on $(\mathcal{Z},\mathfrak{Z})$ such that all $f \in \mathcal{G}$ are integrable. In order to be explicit about the integrability assumptions in the results to come, we will call this space the set of $\mathcal{G}$-integrable finite signed measures (resp. probability distributions, when appropriate).

With this notation, we have $\sup_{h\in\mathcal{H}} |\mathcal{R}(\pi,h) - \mathcal{R}(\pi',h)| = \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})}$ where

$$\mathcal{L}(\mathcal{H}) := \{\ell(\cdot,h) : h \in \mathcal{H}\}. \tag{9}$$

We will usually abbreviate the latter notation by dropping the dependence on $\mathcal{H}$, considered fixed. The desired guarantee (7) then reads $\|\pi - \Delta[\mathbf{y}]\|_{\mathcal{L}} \leq \eta_n/2$.

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. $\mathbf{y} \approx \mathbf{A}\mathrm{vec}(\mathbf{X})$). Guarantees are achieved when $\mathbf{x}$ is close to the set of k-sparse vectors (resp. when $\mathbf{X}$ is close to the set of rank-r matrices).*

Similarly here, estimating π from $\mathbf{y} \approx \mathcal{A}(\pi)$ may require considering some model set $\mathfrak{S}$, whose choice and definition will be discussed in Section 3.

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 often obtain* distribution free *excess risk guarantees using models* derived from the loss function.

Given a model set[2] $\mathfrak{S}$ that plays the role of regularizer, and a sketching operator $\mathcal{A}$, an "ideal" decoder Δ should be robust to two different sources of error: the distribution of the data π generally does not belong to $\mathfrak{S}$ but is "close" to it, introducing some *modelling error*, and the empirical sketch is used instead of the true generalized moments, which adds some *noise*. Generalizing early formulations for sparsity regularized inverse problems, a decoder robust to both noise and modelling error is usually reffered to as *instance optimal* [Cohen et al., 2009, Bourrier et al., 2014]. Mathematically, this can be expressed as: for any distribution π, any draw of the training samples from π (embodied by the empirical distribution $\hat{\pi}_n$), with $\mathbf{y} = \mathcal{A}(\hat{\pi}_n)$ and $\widetilde{\pi} = \Delta[\mathcal{A}(\hat{\pi}_n)]$

$$\|\widetilde{\pi} - \pi\|_{\mathcal{L}} \lesssim d(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n)\|_2 \tag{10}$$

where $\lesssim$ hides multiplicative constants, and $d(\cdot, \mathfrak{S})$ is some measure of distance to the model set $\mathfrak{S}$. In the rest of the paper, we refer to this first term as "bias". A significant part of later sections will be devoted to the control of the bias and the choice of a good model set.

It turns out that general results from abstract linear inverse problems [Bourrier et al., 2014] can be adapted to already characterize the *existence* of a decoder satisfying this property. By [Bourrier et al., 2014, 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

$$\|\tau' - \tau\|_{\mathcal{L}} \leq C_{\mathcal{A}} \|\mathcal{A}(\tau') - \mathcal{A}(\tau)\|_2 \quad \forall \tau, \tau' \in \mathfrak{S}. \tag{11}$$

Conversely, the LRIP (11) implies [Bourrier et al., 2014, Theorem 7] that the following decoder (aka *ideal decoder*)

$$\Delta[\mathbf{y}] := \operatorname{argmin}_{\tau \in \mathfrak{S}} \|\mathcal{A}(\tau) - \mathbf{y}\|_2^2, \tag{12}$$

which corresponds to best moment matching, is instance optimal, i.e., (10) holds for any π and $\hat{\pi}_n$, with the particular distance

$$d^{\circ}(\pi, \mathfrak{S}) := \inf_{\tau \in \mathfrak{S}} \Big\{ \|\pi - \tau\|_{\mathcal{L}} + C_{\mathcal{A}} \|\mathcal{A}(\pi) - \mathcal{A}(\tau)\|_2 \Big\}. \tag{13}$$

As a consequence, the LRIP (11) implies a control of the excess risk achieved with the hypothesis $\hat{h}$ selected with (6), where $\widetilde{\pi} = \Delta[\mathbf{y}]$, as

$$\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^\star) \leq 4 d^{\circ}(\pi, \mathfrak{S}) + 4 C_{\mathcal{A}} \|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n)\|_2 \tag{14}$$

where we used explicit constants from [Bourrier et al., 2014, Theorem 7]. Note that in the above argument, it was never used that the data is distributed i.i.d. from π. Estimate (14) therefore holds under this form for *any* fixed data sample (in fact, for any empirical distribution $\hat{\pi}_n$, being understood that it determines $\hat{h}$) and *any* distribution π. Of course, the data distributional assumption is useful to control the second term in the bound.

2.5 Improved excess risk analysis

The analysis of the previous section has the merits of simplicity, generality, and using existing results from linear inverse problems. However it has some limitations, in particular when the bias term

[2]We will always assume that the models $\mathfrak{S}$ under consideration are such that loss and feature functions are integrable with respect to any distribution belonging to $\mathfrak{S}$, i.e. $\mathfrak{S}$ is both $\mathcal{L}$-integrable and $\{\Phi\}$-integrable, using the terminology introduced after (8).

$d^\circ(\pi, \mathfrak{S})$ is not close to zero. To emphasize this point, we consider a simple example and compare the excess risk control (for the same sketched learning procedure) obtained through the general bound (14), to a direct computation specific to this example. Consider the problem of estimating the median of a distribution on $\mathbb{R}$: we assume $\mathcal{Z} = \mathcal{H} = \mathbb{R}$, and consider the absolute value loss $\ell(x, h) = |x - h|$, whose risk minimizer under the distribution π is the median $h^\star = \mathrm{Med}(\pi)$. As a sketching operator we take simply $\Phi(x) = x$, resulting in the sketch given by the empirical mean $\mathbf{y} = \mathcal{A}(\hat{\pi}_n) = \frac{1}{n}\sum_{i=1}^n x_i$. Finally, as a model we consider the family of 1-point Dirac measures, $\mathfrak{S} = \{\delta_x, x \in \mathbb{R}\}$ (this is the model consisting of all distributions with vanishing optimal risk, see Section 3.2 for a more general discussion). Obviously, it then holds with these choices that the ideal decoder given by (12) is $\widetilde{\pi} = \Delta[\mathbf{y}] = \delta_{\mathbf{y}}$, and further $\hat{h} = \mathbf{y}$.
On the one hand, the excess risk for this sketching/decoding scheme is bounded as follows, by a simple direct calculation, putting $\mathrm{Mean}(\pi) := \mathbb{E}_{X\sim\pi}[X]$:

$$\begin{aligned}\mathcal{R}(\pi, \hat{h}) - \mathcal{R}(\pi, h^\star) &= \mathbb{E}_{X\sim\pi}[|X - \mathbf{y}| - |X - \mathrm{Med}(\pi)|] \\ &\leq \underbrace{\mathbb{E}_{X\sim\pi}[|X - \mathrm{Mean}(\pi)| - |X - \mathrm{Med}(\pi)|]}_{=:\mathcal{B}(\pi)} + |\mathbf{y} - \mathrm{Mean}(\pi)|. \end{aligned} \tag{15}$$

On the other hand, it is easy to check that the LRIP (11) holds, with equality, for 1-point Dirac measures with constant $C_{\mathcal{A}} = 1$. If we consider the general bound (14), while we recover (up to factor 4) the second term above $\|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n)\|_2 = |\mathbf{y} - \mathrm{Mean}(\pi)|$ (of order $\mathcal{O}(1/\sqrt{n})$), the first term in (14) is a *bias term* which is driven by

$$\begin{aligned} d^\circ(\pi, \mathfrak{S}) &:= \inf_{h\in\mathbb{R}} \Big\{ \sup_{h'\in\mathbb{R}} \big|\mathbb{E}_{X\sim\pi}[|X - h'|] - |h - h'|\big| + |\mathrm{Mean}(\pi) - h| \Big\} \\ &\geq \inf_{h\in\mathbb{R}} \Big\{ \mathbb{E}_{X\sim\pi}[|X - h|] + |\mathrm{Mean}(\pi) - h| \Big\} \\ &\geq \mathbb{E}_{X\sim\pi}[|X - \mathrm{Mean}(\pi)|] \\ &= \mathcal{B}(\pi) + \mathcal{R}(\pi, h^\star). \end{aligned} \tag{16}$$

The inequality in the third line above is obtained by noticing that $|x - \mathrm{Mean}(\pi)| \leq |x - h| + |\mathrm{Mean}(\pi) - h|$ for each $x, h \in \mathbb{R}$. Hence, using the general bound (14) instead of the specific direct calculation, we get an additional, unwanted term corresponding to the optimal risk $\mathcal{R}(\pi, h^\star)$ which is nonzero as soon as $\pi \notin \mathfrak{S}$, and can become arbitrary large (even if $\mathcal{B}(\pi) = 0$, e.g. if π is symmetric around its mean).

One reason for this lack of sharpness is that the analysis in the previous section concentrated first on (uniform) control of the risk difference $\|\pi - \pi'\|_{\mathcal{L}}$, to deduce only as a second step a control on the *excess risk*. It is known from the statistical learning literature that it is generally sharper to directly analyze the excess risk; and correspondingly consider the excess loss class

$$\Delta\mathcal{L}(\mathcal{H}) := \mathcal{L}(\mathcal{H}) - \mathcal{L}(\mathcal{H}) = \{g : x \mapsto g(x) = \ell(x, h) - \ell(x, h'), h, h' \in \mathcal{H}\}. \tag{17}$$

However, the risk minimizer $h^\star$ depends on the distribution π; for this reason we will consider a *family* of excess losses and risks with respect to some reference hypothesis h_0.

Definition 2.4. *The* excess risk *relative to a reference hypothesis h_0 is defined as:*

$$\Delta\mathcal{R}_{h_0}(\pi, h) := \mathcal{R}(\pi, h) - \mathcal{R}(\pi, h_0) = \mathbb{E}_{X\sim\pi}[\ell(X, h) - \ell(X, h_0)],$$

and the associated excess risk divergence *with respect to h_0 is:*

$$D_{h_0}(\pi\|\pi') := \sup_{h\in\mathcal{H}} (\Delta\mathcal{R}_{h_0}(\pi, h) - \Delta\mathcal{R}_{h_0}(\pi', h)). \tag{18}$$

Observe that $\Delta\mathcal{R}_{h_0}(\pi, h_0) = 0$ for each π, hence the latter quantity is nonnegative (although no absolute value is involved in its definition), but not symmetric in general. Yet, it satisfies an (oriented) triangle inequality: for any π, π', π''

$$D_{h_0}(\pi\|\pi') \leq D_{h_0}(\pi\|\pi'') + D_{h_0}(\pi''\|\pi').$$

It is therefore a hemimetric *(see Definition B.1 in the Appendix).*

The excess risk divergence will play a role similar to that of $\|\pi - \pi'\|_{\mathcal{L}}$, and satisfies in particular

$$\begin{aligned}\sup_{h_0\in\mathcal{H}} D_{h_0}(\pi\|\pi') &= \sup_{h_0,h\in\mathcal{H}} (\mathbb{E}_{X\sim\pi}[\ell(X,h) - \ell(X,h_0)] - \mathbb{E}_{X\sim\pi'}[\ell(X,h) - \ell(X,h_0)]) \\ &= \|\pi - \pi'\|_{\Delta\mathcal{L}} \leq 2\|\pi - \pi'\|_{\mathcal{L}}. \end{aligned} \tag{19}$$

With this setting we have the following result, which can be seen as a refinement of (14).

Theorem 2.5. *Consider a loss class $\mathcal{L}(\mathcal{H})$, a feature function Φ, and a model set $\mathfrak{S}$ such that every probability distribution $\tau \in \mathfrak{S}$ is both $\mathcal{L}$-integrable and $\{\Phi\}$-integrable. Assume that the sketching operator $\mathcal{A}$ associated to Φ satisfies the following LRIP inequality:*

$$\|\tau - \tau'\|_{\Delta\mathcal{L}} \leq C_{\mathcal{A}}\|\mathcal{A}(\tau) - \mathcal{A}(\tau')\|_2 + \eta, \qquad \forall \tau, \tau' \in \mathfrak{S}, \tag{20}$$

for some finite positive constants $C_{\mathcal{A}}$ and η.

Consider any training collection $\mathbf{X} = \{x_i\}_{i=1}^n \in \mathcal{Z}^n$, and denote $\hat{\pi}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Define

$$\mathbf{y} := \texttt{Sketch}(\mathbf{X}) = \mathcal{A}(\hat{\pi}_n), \tag{21}$$

$$\widetilde{\pi} \in \mathfrak{S} \text{ satisfying } \|\mathcal{A}(\widetilde{\pi}) - \mathbf{y}\|_2 \leq (1+\nu)\inf_{\tau\in\mathfrak{S}}\|\mathcal{A}(\tau) - \mathbf{y}\|_2 + \varepsilon, \quad \text{for some constants } \varepsilon, \nu \geq 0, \tag{22}$$

$$\hat{h} \text{ satisfying } \mathcal{R}(\widetilde{\pi}, \hat{h}) \leq \inf_{h\in\mathcal{H}} \mathcal{R}(\widetilde{\pi}, h) + \varepsilon', \quad \text{for some constant } \varepsilon' \geq 0. \tag{23}$$

Then, for any probability distribution π that is both $\mathcal{L}$-integrable and $\{\Phi\}$-integrable:

$$\forall h_0 \in \mathcal{H}: \quad \Delta\mathcal{R}_{h_0}(\pi, \hat{h}) \leq d_{h_0}(\pi, \mathfrak{S}) + (2+\nu)C_{\mathcal{A}}\|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n)\|_2 + \eta + C_{\mathcal{A}}\varepsilon + \varepsilon', \tag{24}$$

where

$$d_{h_0}(\pi, \mathfrak{S}) := \inf_{\tau\in\mathfrak{S}}(D_{h_0}(\pi\|\tau) + (2+\nu)C_{\mathcal{A}}\|\mathcal{A}(\pi) - \mathcal{A}(\tau)\|_2). \tag{25}$$

Similarly to (14), the above estimate holds regardless of any distributional assumptions on the training collection $\mathbf{X}$. Nevertheless, estimate (24) is primarily of interest when $\mathbf{X}$ is drawn i.i.d. according to π and with $h_0 = h^\star_\pi$, in which case the left-hand side is the excess risk with respect to the optimum risk, which is what one generally aims at controlling. However, in some situations it may be also helpful to consider excess risk with respect to other reference hypotheses h_0; this can include situations where $h^\star_\pi$ itself is not well-defined if the infimum of the risk is not attained.
As compared to (14), we observe that this result is more general, as it allows for a (ν, ε)-approximate decoder (22), an ε'-approximate ERM (23), an η-approximate LRIP condition (20); more importantly, the main bound (24) involves the sharper excess risk divergence rather than the loss norm $\|.\|_{\mathcal{L}}$. It may also be useful to consider $\pi = \hat{\pi}_n$, to predict the quality compared to the empirical risk minimizer.

Moreover, inequality (19) implies that the lower LRIP condition (11) considered in the previous section implies the relaxed LRIP condition (20) (with $\eta = 0$, and up to a factor 2 in the constant), so establishing (11) is sufficient in order to obtain the improved inequality (24).

The proof of Theorem 2.5 follows the structure outlined in the previous section, but requires to formally extend the result of [Bourrier et al., 2014, Section IV-A] (leading from the LRIP (11) to instance optimality (10)) to the case of a hemimetric, η-approximate LRIP and ε-approximate decoder. These technical aspects are relegated to Appendix B.

Discussion:

- Computing the sketch (21) 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 (22) may appear as a general nonparametric density estimation problem, in all the examples considered in this paper and the companion one [Gribonval et al., 2020], it is indeed a nonlinear parametric least-squares fitting problem and the existence of the minimizer follows in practice from compactness arguments.
  - For **Compressive PCA** (Section 4) 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** (cf the companion paper [Gribonval et al., 2020]), the resulting optimization problem has been empirically addressed through the CL-OMPR algorithm [Keriven et al., 2015, 2016]. Algorithmic success guarantees are an interesting challenge. This is however beyond the scope of this paper. We note that the classic (non-compressed) k-means problem by minimization of the empirical risk is known to be NP-hard [Garey et al., 1982, Aloise et al., 2009] and that guarantees for approaches such as K-means++ [Arthur and Vassilvitskii, 2007b] are only in expectation and with a logarithmic sub-optimality factor.
- In Section 3 we discuss choices of the model set $\mathfrak{S}$ that are driven by the learning task only and make the minimization problem (23) trivial to solve. With these choices the combined solution of (22)-(23) is explicitly turned into the minimization of a proxy for the risk, as in (4).
- The second term in the bound (24) of the excess risk, η_n, is the empirical estimation error $\|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n)\|_2$. It is easy to show that it decays as $1/\sqrt{n}$ when the data is drawn i.i.d. according to π, this will be done explicitly for the considered examples.

For large collection size n drawn i.i.d. according to π, the term $\|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n)\|_2$ becomes small and (24) 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 (20) holds; ideally for a "small" value of m, as we also seek to design compact sketches and, eventually, tractable algorithms to learn from them.
2. the distance $d_{h^\star_\pi}(\pi, \mathfrak{S})$ is "small"; this vague notion will be exploited in Section 3 below to guide our choice of model set $\mathfrak{S}$, and will be made more concrete on examples.

We illustrate the improvement obtained for the bias term with respect to the coarser analysis on the toy example of median estimation considered previously. In that setting we have $h^\star_\pi = \mathrm{Med}(\pi)$, and for $\tau = \delta_x \in \mathfrak{S}$:

$$\begin{aligned} D_{h^\star}(\pi\|\delta_x) &= \sup_{h\in\mathcal{H}} (\mathbb{E}_{X\sim\pi}[|X-h| - |X-\mathrm{Med}(\pi)|] - (|x-h| - |x-\mathrm{Med}(\pi)|)) \\ &= \mathbb{E}_{X\sim\pi}[|X-x|] - \mathbb{E}_{X\sim\pi}[|X-\mathrm{Med}(\pi)|] + |x-\mathrm{Med}(\pi)| \end{aligned}$$

so that

$$\begin{aligned} d_{h^\star}(\pi, \mathfrak{S}) &:= \inf_{\delta_x\in\mathfrak{S}} (D_{h^\star}(\pi\|\delta_x) + (2+\nu)C_{\mathcal{A}}\|\mathcal{A}(\pi) - \mathcal{A}(\delta_x)\|_2). \\ &= \inf_{x\in\mathbb{R}} (\mathbb{E}_{X\sim\pi}[|X-x|] - \mathbb{E}_{X\sim\pi}[|X-\mathrm{Med}(\pi)|] + |x-\mathrm{Med}(\pi)| + (2+\nu)|x-\mathrm{Mean}(\pi)|). \\ &\leq \mathcal{B}(\pi) + |\mathrm{Med}(\pi) - \mathrm{Mean}(\pi)|. \end{aligned}$$

The inequality in the third line is obtained by using $x = \mathrm{Mean}(\pi)$. Note that the presence of the last term is unavoidable (since $|x - \mathrm{Med}(\pi_0)| + |x - \mathrm{Mean}(\pi)| \geq |\mathrm{Med}(\pi_0) - \mathrm{Mean}(\pi)|$), and that it is still larger than $\mathcal{B}(\pi)$ which we recall is the only bias term appearing in the direct calculation (15). (A situation where it is much larger is the following: assume $\pi = (\frac{1}{2}+\varepsilon)\delta_0+(\frac{1}{2}-\varepsilon)\delta_1$, for $0 < \varepsilon < 1/2$. Then $\mathrm{Med}(\pi) = 0$, $\mathrm{Mean}(\pi) = \frac{1}{2}-\varepsilon$, $\mathbb{E}_{X\sim\pi}|X - \mathrm{Mean}(\pi)| = (\frac{1}{2}+\varepsilon)(\frac{1}{2}-\varepsilon)+(\frac{1}{2}-\varepsilon)(\frac{1}{2}+\varepsilon) = (1+2\varepsilon)(\frac{1}{2}-\varepsilon)$, $\mathbb{E}_{X\sim\pi}|X - \mathrm{Med}(\pi)| = \mathbb{E}_{X\sim\pi}X = \frac{1}{2}-\varepsilon$, hence $|\mathrm{Med}(\pi) - \mathrm{Mean}(\pi)| = \frac{1}{2} - \varepsilon$ while $\mathcal{B}(\pi) = 2\varepsilon(1/2-\varepsilon)$.) In this sense, even using the improved excess risk analysis, the general bound (24) can lack some tightness. It is nevertheless much sharper than the bound (16), in particular $d_{h^\star_\pi}(\pi, \mathfrak{S}) = 0$ as soon as $\mathrm{Mean}(\pi) = \mathrm{Med}(\pi)$, while $d_{h^\star_\pi}(\pi, \mathfrak{S}) > 0$ in general in this case, see (16).

3 Task-driven model sets

An important ingredient of the proposed framework is the model set $\mathfrak{S}$ and we now discuss its choice. It is of course possible to design $\mathfrak{S}$ so as to incorporate prior knowledge on the data distribution into compressive statistical learning. However, it is more common in statistical learning to seek "distribution-free" statistical guarantees, so deriving a model set from the learning task itself rather than prior knowledge may also be desirable.

For certain model sets directly derived below from the considered learning task, we further show that the abstract two-step learning mechanism (cf steps (22)-(23) in Theorem 2.5) can be written as the minimization of a more explicit proxy (4) for the empirical risk. Moreover, for certain learning tasks, the bias term in the excess risk (24) is controlled by a function of the optimal risk $\mathcal{R}(\pi, h^\star)$.

3.1 Learning from a sketch without explicit density estimation

Consider a family $\mathfrak{S}$ of $\mathcal{L}$-integrable distributions and assume for each $\pi \in \mathfrak{S}$ the risk admits a minimizer, i.e., there is $h \in \mathcal{H}$ such that $\mathcal{R}(\pi, h) = \inf_{h' \in \mathcal{H}} \mathcal{R}(\pi, h')$. When this holds the model set $\mathfrak{S}$ can be decomposed as $\mathfrak{S} = \cup_{h\in\mathcal{H}}\mathfrak{S}_h$ where for each hypothesis $h \in \mathcal{H}$ we define

$$\mathfrak{S}_h := \{\pi \in \mathfrak{S} : \mathcal{R}(\pi, h) \leq \mathcal{R}(\pi, h'), \forall h' \in \mathcal{H}\}, \tag{26}$$

and the hypothesis $\hat{h}$ selected using steps (22)-(23) in Theorem 2.5 (with $\varepsilon' = 0$) is equivalently obtained as a near-minimizer of the following proxy for the risk

$$R(\mathbf{y}, h) := \inf_{\tau\in\mathfrak{S}_h} \|\mathcal{A}(\tau) - \mathbf{y}\|_2 , \tag{27}$$

in the sense that $\inf_{\tau\in\mathfrak{S}}\|\mathcal{A}(\tau) - \mathbf{y}\|_2 = \inf_h \inf_{\tau\in\mathfrak{S}_h}\|\mathcal{A}(\tau) - \mathbf{y}\|_2$. With this expression in hand, it is possible to directly cast the estimation of $\hat{h}$ as (4). To turn this into a concrete proxy for the risk it is helpful to consider a model set $\mathfrak{S}$ such that $\mathfrak{S}_h$ has a simple characterization.

3.2 Choosing a model set: with or without prior knowledge ?

Learning tasks such as maximum likelihood estimation directly involve a natural model set for which $\mathfrak{S}_h$ as in (26) is easily characterized. Consider the loss $\ell(x, h) = -\log \pi_h(x)$ with $\{\pi_h, h \in \mathcal{H}\}$ a parameterized family of distributions with $\pi_{h'} \neq \pi_h$ for $h' \neq h$. Given $\widetilde{\pi} = \pi_h$, a minimum risk hypothesis $\hat{h}$ according to (6) minimizes the Kullback-Leibler divergence $\min_{h'} \mathrm{KL}(\pi_h||\pi_{h'})$, hence $\hat{h} = h$ [Cover and Thomas, 1991, Chapter 9]. We consider the following model set, which in this setting is nothing more than a statistical model in the usual sense:

$$\mathfrak{S}^{\mathtt{ML}}(\mathcal{H}) := \{\pi_h : h \in \mathcal{H}\}, \tag{28}$$

so we have $\mathfrak{S}_h^{\texttt{ML}} = \{\pi_h\}$ for each $h \in \mathcal{H}$, and the proxy (27) reads $R(\mathbf{y}, h) = \|\mathcal{A}(\pi_h) - \mathbf{y}\|_2$.

For many other learning tasks, the choice of the model set $\mathfrak{S}$ results from a tradeoff between several needs. On the one hand, results from compressed sensing suggest that given a model set $\mathfrak{S}$ that has proper "low-dimensional" properties, it is possible to choose a small sketch size m and design the sketching operator $\mathcal{A}$ such that the LRIP (20) holds, and the ideal decoder Δ in (12) — or its relaxed version in (22) — is guaranteed to stably recover probability distributions in $\mathfrak{S}$ 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, in the common case of compression-type tasks, as defined below, the bias term in the excess risk (24) should be small when the true optimum risk is small, and even vanish when the true optimum risk vanishes, i.e. when $\inf_{h\in\mathcal{H}} \mathcal{R}(\pi, h) = 0$.

Definition 3.1. *We call the learning task a* compression-type task *if the loss can be written as* $\ell(x, h) = d^p(x, P_h x)$*, where* d *is a metric on* $\mathcal{Z}$*,* $p > 0$*, and* $P_h : \mathcal{Z} \to \mathcal{Z}$ *is a "projection function", i.e.,*

$$P_h \circ P_h = P_h; \tag{29}$$

$$d(x, P_h x) \leq d(x, P_h x'), \quad \forall x, x' \in \mathcal{Z}. \tag{30}$$

Typical examples of compression-type tasks are PCA, k-means, and k-medians. For PCA, P_h is the orthogonal projector onto subspace h. For k-means and k-medians, P_h maps $x \in \mathcal{Z} = \mathbb{R}^d$ to the closest center c_i from $h = (c_1, \ldots, c_k)$, with ties broken arbitrarily. In other words, given an arbitrary *Voronoi partition* corresponding to k disjoint sets W_j such that $\cup_j W_j = \mathbb{R}^d$ and $d(x, c_j) = \min_l d(x, c_l)$ for each $x \in W_j$, $P_h x = c_j$ if and only if $x \in W_j$. Manifold learning tasks where P_h is a projection onto a manifold parameterized by h (with ties broken arbitrarily) would also fit under this framework.

For a compression-type task, a natural model set is the family of $\mathcal{L}$-integrable probability distributions

$$\mathfrak{S}^{\texttt{CT}}(\mathcal{H}) := \cup_{h\in\mathcal{H}} \mathfrak{S}_h^{\texttt{CT}} \qquad \text{where} \quad \mathfrak{S}_h^{\texttt{CT}} := \{\pi : \mathcal{R}(\pi, h) = 0\}. \tag{31}$$

We consider a few examples:

- **Compressive PCA**: the model set $\mathfrak{S}^{\texttt{CT}}(\mathcal{H})$ consists of all distributions which admit a matrix of second moments of rank at most k. Given any $\widetilde{\pi} \in \mathfrak{S}^{\texttt{CT}}(\mathcal{H})$, a minimum risk hypothesis according to (6) is any subspace $\hat{h}$ spanned by eigenvectors associated to the k largest eigenvalues of $\mathbf{\Sigma}_{\widetilde{\pi}}$. More details will be given shortly in Section 4.

- **Compressive k-means or k-medians**: the model set $\mathfrak{S}^{\texttt{CT}}(\mathcal{H})$ consists of mixtures of k Diracs. Given $h = \{c_1, \ldots, c_k\}$ and any $\widetilde{\pi} = \sum_{\ell=1}^k \alpha_\ell \delta_{c_\ell} \in \mathfrak{S}_h^{\texttt{CT}}$, a minimum risk hypothesis according to (6) is $\hat{h} = h$. Since $\mathcal{A}(\delta_c) = \Phi(c)$, the proxy (27) reads

$$R(\mathbf{y}, h) = \min_{\alpha\in\mathbb{S}_{k-1}} \left\| \sum_{\ell=1}^k \alpha_\ell \Phi(c_\ell) - \mathbf{y} \right\|_2 \tag{32}$$

  with $\mathbb{S}_{k-1} := \left\{\alpha \in \mathbb{R}^k : \alpha_\ell \geq 0; \sum_{\ell=1}^k \alpha_\ell = 1\right\}$ the simplex.

For compressive PCA we exhibit in Section 4 a feature function Φ so that $\mathcal{A}$ satisfies the LRIP (20) with respect to the model set $\mathfrak{S} = \mathfrak{S}^{\texttt{CT}}(\mathcal{H})$. The same is done in the companion paper [Gribonval et al., 2020] for compressive k-means, compressive k-medians and compressive Gaussian mixture modeling.

3.3 Controlling the bias term for compression-type tasks

The first term in (24) is a measure of distance to the model set $\mathfrak{S}$. For compression-type tasks, the particular model set $\mathfrak{S} = \mathfrak{S}^{\mathtt{CT}}(\mathcal{H})$ in (31) was designed so that this *bias term* –defined in (25)– vanishes when $\pi \in \mathfrak{S}$. We can further bound the bias term $d_{h^\star_\pi}(\pi, \mathfrak{S}^{\mathtt{CT}}(\mathcal{H}))$ with an increasing function of the true minimum risk, $\mathcal{R}(\pi, h^*)$. This leads to recovery guarantees providing *distribution-free* excess risk guarantees. Whether this holds for other learning tasks, or even generically, is a challenging question left to further work.

The following lemmas allow to obtain an upper bound of the bias term in function of the minimum risk in a number of relevant cases. For a probability distribution π on $\mathcal{Z}$ and P_h as in Definition 3.1, we denote $P_h\pi$ the push-forward of π through P_h, i.e., the probability distribution of a random variable $Y = P_h X$ where $X \sim \pi$. Given a loss class $\mathcal{L}(\mathcal{H})$ we recall that $h^\star_\pi = \arg\min_{h\in\mathcal{H}} \mathcal{R}(\pi, h)$.

Lemma 3.2. *Consider a compression-type task on the input space $\mathcal{Z}$. Then*

- *$\mathfrak{S}^{\mathtt{CT}}_h$ is the set of probability distributions on $X \in \mathcal{Z}$ such that $X \in \mathcal{E}_h := P_h\mathcal{Z}$ almost surely. With the model set $\mathfrak{S}^{\mathtt{CT}}(\mathcal{H})$ and the loss class $\mathcal{L}(\mathcal{H})$, the bias term (25) satisfies (for any $h_0 \in \mathcal{H}$)*

$$d_{h_0}(\pi, \mathfrak{S}^{\mathtt{CT}}(\mathcal{H})) \leq D_{h_0}(\pi\|P_{h_0}\pi) + (2+\nu)C_{\mathcal{A}}\|\mathcal{A}(\pi) - \mathcal{A}(P_{h_0}\pi)\|_2. \tag{33}$$

- *If d^p is a metric (in particular if $p \leq 1$) then $D_h(\pi\|P_h\pi) = 0$ for any $h \in \mathcal{H}$ and $\mathcal{L}$-integrable distribution π.*

Remark 3.3. *When d^p is not a metric there are $u, v, w \in \mathcal{Z}$ such that $d^p(u, v) > d^p(u, w) + d^p(w, v)$. The loss $\ell(x, h) := d^p(x, h)$, with $\mathcal{H} := \{v, w\}$, defines a compression-type task with $P_h x = h$ for all $x \in \mathcal{Z}$. Set $\pi = \delta_u$. Since $d(u, w) < d(u, v)$ we have $h^\star_\pi = w$. We also have for $h = v \in \mathcal{H}$*

$$\begin{aligned} D_{h^\star_\pi}(\pi\|P_{h^\star_\pi}\pi) \geq \Delta\mathcal{R}_w(\pi, h) - \Delta\mathcal{R}_w(P_w\pi, h) &= [d^p(u, h) - d^p(u, w)] - [d^p(w, h) - d^p(w, w)] \\ &= d^p(u, v) - d^p(u, w) - d^p(w, v) > 0. \end{aligned}$$

Hence, one cannot generically obtain $D_h(\pi\|P_h\pi) = 0$, not even with the restriction $h = h^\star_\pi$.

For $p = 2$, d the Euclidean distance on $\mathbb{R}^d$, and $h_0 = h^\star_\pi$ (which we recall is generally the primary interest case since our main bound (24) then gives a control of the excess risk with respect to the optimum), we still have $D_{h^\star_\pi}(\pi\|P_{h^\star_\pi}\pi) = 0$ for certain tasks. In light of the above remark, this is a nontrivial property which is established for PCA in Lemma E.1, and for k-means in the companion paper [Gribonval et al., 2020]. Beyond these specific situations, it is possible (under somewhat generic additional assumptions) to bound the two terms appearing in (33) by (a power of) the risk itself, as established next.

Lemma 3.4. *Consider a compression-type task where $(\mathcal{Z}, d)$ is a* separable *metric space. Then*

- *Assume that $\mathcal{Z}$ has d-diameter bounded by B. Then for any $p > 1$, $h \in \mathcal{H}$ and $\mathcal{L}$-integrable distribution π:*

$$D_h(\pi\|P_h\pi) \leq 2pB^{p-1}\mathcal{R}(\pi, h)^{\frac{1}{p}}. \tag{34}$$

- *Assume that $\Phi : (\mathcal{Z}, d) \to (\mathbb{R}^m, \|\cdot\|_2)$ (or $(\mathbb{C}^m, \|\cdot\|_2)$) is L-Lipschitz. Then, for $h \in \mathcal{H}'$ and $p \geq 1$:*

$$\|\mathcal{A}(\pi) - \mathcal{A}(P_h\pi)\|_2 \leq L \inf_{h\in\mathcal{H}} \mathcal{R}(\pi, h)^{\frac{1}{p}}. \tag{35}$$

  For $h \in \mathcal{H}$ and $p \leq 1$, if the space $\mathcal{Z}$ has d-diameter bounded by B:

$$\|\mathcal{A}(\pi) - \mathcal{A}(P_h\pi)\|_2 \leq LB^{1-p} \inf_{h\in\mathcal{H}} \mathcal{R}(\pi, h). \tag{36}$$

The proofs of the above lemmas are in Appendix D. For Lemma 3.4), optimal transport is exploited 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. The two lemmas can be combined to express an "explicit" bound on $d_{h^\star_\pi}$, this is postponed to concrete examples.

4 Illustration with 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.

Definition of the learning task. The risk associated to the PCA learning problem is defined[3] as $\mathcal{R}_{k-\mathtt{PCA}}(\pi, h) = \mathbb{E}_{X\sim\pi}\|X - P_h X\|_2^2$ with P_h the orthogonal projector onto subspace h. It is minimized by the subspace $h^\star_\pi$ associated with the k largest eigenvalues of the matrix $\mathbf{\Sigma}_\pi = \mathbb{E}_{X\sim\pi} XX^T$.

It is well established [Foucart and Rauhut, 2012] 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 a model set. The "natural" model set from (31) is $\mathfrak{S}^{\mathtt{CT}}(\mathcal{H}) = \{\pi : \mathrm{rank}(\mathbf{\Sigma}_\pi) \leq k\}$. More generally we can consider as a model set $\mathfrak{S}_r := \{\pi : \mathrm{rank}(\mathbf{\Sigma}_\pi) \leq r\}$, with $r \geq k$, so that $\mathfrak{S}_r \supset \mathfrak{S}^{\mathtt{CT}}(\mathcal{H})$.

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 RIP on low-rank matrices: for any $\mathbf{M} \in \mathbb{R}^{d\times d}$ of rank at most $2r$,

$$1 - \delta \leq \frac{\|\mathcal{M}(\mathbf{M})\|_2^2}{\|\mathbf{M}\|_F^2} \leq 1 + \delta \tag{37}$$

with $\|\cdot\|_F$ the Frobenius norm and $\delta < 1$. This is feasible with m of the order of rd, by taking the Frobenius inner product of $\mathbf{M}$ with m independent random Gaussian matrices [see e.g. Foucart and Rauhut, 2012].

Given these facts one can define the feature function as $\Phi : \mathcal{Z} = \mathbb{R}^d \to \mathbb{R}^m$ 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 (3), i.e., compute empirical estimates of random second moments of the distribution π of X.

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

$$\hat{\mathbf{\Sigma}} := \arg \min_{\mathrm{rank}(\mathbf{\Sigma})\leq r, \mathbf{\Sigma}\succcurlyeq 0} \|\mathcal{M}(\mathbf{\Sigma}) - \mathbf{y}\|_2^2. \tag{38}$$

This step estimates the rank-r matrix whose sketch best matches that 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

$$\hat{h} := \mathrm{span}(\mathbf{U}(:, 1:k)). \tag{39}$$

In Appendix E we control the excess risk of PCA by relating the excess risk divergence $D_{h_0}(\pi\|\pi')$ — with $h_0 \in \mathcal{H}$ an arbitrary hypothesis — to the Frobenius norm $\|\pi - \pi'\|_F$, and upper bounding the "bias" term (25) appearing in the generic bound of Theorem 2.5, to obtain the following result:

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

Theorem 4.1. *Consider any probability distribution π 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, for any s, $1 \leq s \leq r$:*

$$\mathcal{R}_{k-\mathtt{PCA}}(\pi, \hat{h}) - \mathcal{R}_{k-\mathtt{PCA}}(\pi, h_\pi^\star) \leq c_\delta \sqrt{\frac{k}{s}} \sum_{j \geq r-s+2} \lambda_j(\mathbf{\Sigma}_\pi) + c'_\delta \sqrt{k} \|\mathcal{M}(\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\hat{\pi}_n})\|_2, \tag{40}$$

where $\lambda_i(\mathbf{\Sigma}_\pi)$ are the eigenvalues of $\mathbf{\Sigma}_\pi$ ranked in decreasing order (with multiplicity), $c_\delta := 2\sqrt{2}\frac{\sqrt{1+\delta}}{\sqrt{1-\delta}}$, $c'_\delta := 2\sqrt{2}/\sqrt{1-\delta}$. In particular:

$$\mathcal{R}_{k-\mathtt{PCA}}(\pi, \hat{h}) - \mathcal{R}_{k-\mathtt{PCA}}(\pi, h_\pi^\star) \leq c_\delta \sqrt{\frac{k}{r-k+1}} \mathcal{R}_{k-\mathtt{PCA}}(\pi, h_\pi^\star) + c'_\delta \sqrt{k} \|\mathcal{M}(\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\hat{\pi}_n})\|_2. \tag{41}$$

Discussion:

- **Bias term.** The first term in the right hand side of (41) is a bias term that vanishes when the true risk is low. Since it is proportional to to the true risk, it leads to the (non-sharp) oracle inequality $\mathcal{R}_{k-\mathtt{PCA}}(\pi, \hat{h}) \leq C_\delta(k,r)\mathcal{R}_{k-\mathtt{PCA}}(\pi, h_\pi^\star) + c'_\delta\sqrt{k}\|\mathcal{M}(\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\hat{\pi}_n})\|_2$. We show in [Gribonval et al., 2020], (using Lemma 3.4 and (36)) that this type of property also holds for Compressive k-medians; for Compressive k-means we prove similar properties where the bias term is bounded by the square root of the true risk (using (34),(35)).
  It is notable that the bias multiplier $C_\delta(k,r)$ is of order $\sqrt{k}$ if we use the natural model set ($r = k$), but drops to a constant independent of k as soon as we choose e.g. the larger model set $\mathfrak{S}_r$ with $r = 2k$. Thus, there appears to be a clear advantage, in the sense of the obtained bound, in choosing a reconstruction model that is larger than the natural model set $\mathfrak{S}^{\mathtt{CT}}(\mathcal{H})$, while not significantly changing the magnitude of the number of required data sketches. At this point it is unclear to us if the inflation of the bias factor for the natural model is unavoidable or is just a technical artefact.

- **Sample complexity.** Regarding the second term, if we further assume that the support of π is contained in a Euclidean ball of radius R, then by the RIP (37) we have a.s. $\left\|\mathcal{M}(xx^T)\right\|_2 \leq \sqrt{1+\delta} \cdot R^2$ hence, by the vectorial Hoeffding's inequality [see e.g. Pinelis, 1992], we obtain with high probability w.r.t. data sampling that $\sqrt{k}\|\mathcal{M}(\mathbf{\Sigma}_\pi) - \mathcal{M}(\mathbf{\Sigma}_{\hat{\pi}_n})\|_2$ is of the order of $R^2\sqrt{k/n}$.

- **Root-n consistency in a high-dimensional scenario**. As noticed in the previous point, the statistical estimation error term in the sketched learning bound is of order $\sqrt{k/n}$. Consider a high-dimensional situation where d is large and growing with n, and assume a polynomial spectral decay $\lambda_j(\mathbf{\Sigma}_\pi) \leq j^{-\alpha}$ with $\alpha > 1$. Then by choosing $s = r/2$, the bias term in (40) is of order $\sqrt{kr^{-(2\alpha-1)}}$. As a consequence, it is sufficient to take r of order $\min(d, n^{\frac{1}{2\alpha-1}})$ so as to ensure that the bias term is at most of the same order as the statistical error term. This gives an advantage compared to the standard approach of storing all d^2 second order moments, as soon as $d > n^{\frac{1}{2\alpha-1}}$. For comparison, standard statistical analysis of PCA based on the uniform bound on the deviations of the empirical risk from its expectation (see e.g. Shawe-Taylor et al., 2005) leads to a control of order[4] $\sqrt{k/n}$. Hence, using the sketched approach we can reduce storage/memory imprint significantly while keeping statistical guarantees of the same order as in the standard setting.

[4]More refined techniques [Blanchard et al., 2007, Reiß and Wahl, 2016] can lead to a convergence rate of the PCA excess risk of order n^{-1} asymptotically, for k fixed, but depending on eigenvalue gaps. For the present discussion we compare ourselves to the simplest analysis available in the standard learning context.

Practical algorithms for learning and comparison to prior PCA-specific results. One can consider several relaxations of the nonconvex optimization problem (38) in order to perform compressive PCA. Beside convex relaxations using the minimization of the nuclear norm [Foucart and Rauhut, 2012, Section 4.6], Kabanava et al. [2016] showed (in a complex-valued setting) that the rank constraint in (38) 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

$$\hat{\mathbf{\Sigma}} := \arg\min_{\mathbf{\Sigma} \succcurlyeq 0} \|\mathcal{M}(\mathbf{\Sigma}) - \mathbf{y}\|_2^2, \tag{42}$$

and the corresponding hypothesis $\hat{h}$ obtained through (39). Combining [Kabanava et al., 2016, Theorem 4 with $p = 2$] with Equation (72) in Section E and Equation (63) in Section B, we have the following result: 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

$$\mathcal{R}_{k-\texttt{PCA}}(\pi, \hat{h}) - \mathcal{R}_{k-\texttt{PCA}}(\pi, h^\star) \leq \left\| \hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_\pi \right\|_F \leq D_1 \mathcal{R}_{k-\texttt{PCA}}(\pi, h^\star) + D_2 \sqrt{k} \|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n)\|_2,$$

where D_1, D_2 are positive universal constants that do not depend on k.

Hence, provided we use a model set of dimension $2r$ as discussed above, the error control (41) obtained via our general approach matches what can be obtained using directly the PCA-specific study of Kabanava et al. [2016]. Two practical advantages of the latter are (a) that (42) is a convex program, and (b) that the sketches are made using rank-one matrices, which are cheaper to store. Still, the guarantees obtained by our general approach is able to match prior results for setting-specific methods. It will further permit the study of the less trivial setting of compressive clustering and compressive Gaussian mixture estimation as shown in the companion paper [Gribonval et al., 2020].

5 Establishing the LRIP for random sketching operators

In this section, we investigate how to establish the LRIP (20) (with $\eta = 0$) when the sketching operator $\mathcal{A}$ is associated to random features. The approach uses connections with the notion of kernel mean embedding of probability distributions.

5.1 Random features and kernel mean embeddings

Definition 5.1 (Random feature map)**.** *Consider $\mathcal{F} := \{\phi_\omega\}_{\omega \in \Omega}$ a parameterized family of (real- or complex-valued) measurable functions, Λ a probability distribution Λ over the parameter set Ω (often $\Omega = \mathbb{R}^d$), and a sketch size m. A random feature map is defined by drawing m i.i.d parameters $(\omega_j)_{j=1}^m$ according to Λ and setting*

$$\Phi(x) := \frac{1}{\sqrt{m}} \big(\phi_{\omega_j}(x)\big)_{j=1,m}. \tag{43}$$

Any draw of the feature function Φ defines a positive semi-definite kernel between samples $\kappa_\Phi(x, x') := \langle \Phi(x), \Phi(x') \rangle_{\mathbb{R}^m}$ (or $\langle \Phi(x), \Phi(x') \rangle_{\mathbb{C}^m}$). Compressive learning is deeply connected to kernel mean embeddings of probability distributions, as the related sketching operator $\mathcal{A}$ defines a so-called *kernel mean embedding* between probability distributions which are $\mathcal{F}$-integrable.

Definition 5.2 (Kernel mean embedding, Mean Map Discrepancy [Gretton et al., 2007, Sriperumbudur et al., 2010])**.** *Any positive semi-definite kernel $\kappa(\cdot, \cdot)$ in the sample space is associated to a Mean Map Embedding (a kernel between distributions). By abuse of notation, we keep the notation κ for both the expression of the kernel in the sample space and of the corresponding kernel for probability distributions with appropriate integrability*

$$\kappa(\pi, \pi') := \mathbb{E}_{X \sim \pi} \mathbb{E}_{X' \sim \pi'} \kappa(X, X'). \tag{44}$$

The associated Maximum Mean Discrepancy (MMD) metric is

$$\|\pi - \pi'\|_\kappa := \sqrt{\kappa(\pi,\pi) - 2\mathrm{Re}(\kappa(\pi,\pi')) + \kappa(\pi',\pi')}. \tag{45}$$

The average kernel κ associated to $(\mathcal{F}, \Lambda)$, will play a key role in establishing the LRIP. Given $x, x' \in \mathcal{Z}$, the expectation of $\kappa_\Phi(x,x') = \frac{1}{m}\sum_{j=1}^m \phi_{\omega j}(x)\overline{\phi_{\omega j}}(x')$ over the draws of ω_j is

$$\kappa(x,x') = \mathbb{E}_{\omega\sim\Lambda}\kappa_\Phi(x,x') = \mathbb{E}_{\omega\sim\Lambda}\phi_\omega(x)\overline{\phi_\omega(x')}. \tag{46}$$

Similarly, given π, π', the squared MMD $\|\pi - \pi'\|_\kappa^2$ with this kernel is the expectation of

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

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

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

We observe that $\mathcal{A}$ satisfies the LRIP (20) (with $\eta = 0$) for a given model set $\mathfrak{S}$ if, and only if, the metric $\|\pi - \pi'\|_{\Delta\mathcal{L}}$ is dominated by $\|\pi - \pi'\|_{\kappa_\Phi}$ for $\pi, \pi' \in \mathfrak{S}$. Our overall strategy to check that a random feature function Φ defined by $\mathcal{F}$ and Λ satisfies the LRIP (20) (with $\eta = 0$) with controlled sketch dimension m will be to:

1. prove that the average kernel κ defined by (46) satisfies the *Kernel LRIP*

$$\|\tau - \tau'\|_{\Delta\mathcal{L}(\mathcal{H})} \le C_\kappa \|\tau - \tau'\|_\kappa, \quad \forall \tau, \tau' \in \mathfrak{S}; \tag{47}$$

2. 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,

$$1 - \delta \le \frac{\|\mathcal{A}(\tau) - \mathcal{A}(\tau')\|_2^2}{\|\tau - \tau'\|_\kappa^2} = \frac{\|\tau - \tau'\|_{\kappa_\Phi}^2}{\|\tau - \tau'\|_\kappa^2} \le 1 + \delta, \quad \forall \tau, \tau' \in \mathfrak{S} \tag{48}$$

so that the kernel LRIP (47) actually holds with κ_Φ instead of κ and constant $C_{\kappa_\Phi} := C_\kappa/\sqrt{1-\delta}$.

Remark 5.3. *The expression* (48) *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. Sriperumbudur and Szabó, 2015, Bach, 2017] 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.*

5.2 Ingredients to verify the Lower Restricted Isometry Property

In sight of the inequalities (47),(48) we need to prove, the analysis will focus on the so-called normalized secant set of the model $\mathfrak{S}$ with respect to the average kernel κ, defined as follows [see, e.g. Dirksen, 2016, Puy et al., 2017]:

Definition 5.4 (Normalized secant set)**.** *The* normalized secant set *of a model set* $\mathfrak{S}$ *with respect to a kernel* κ *is the following subset of the set of finite, signed measures (see Appendix A.2) with appropriate integrability*

$$\mathcal{S}_\kappa = \mathcal{S}_\kappa(\mathfrak{S}) := \left\{ \frac{\tau - \tau'}{\|\tau - \tau'\|_\kappa} : \tau, \tau' \in \mathfrak{S}, \|\tau - \tau'\|_\kappa > 0 \right\}. \tag{49}$$

Using the secant set, the LRIP (48) is equivalent to

$$\Big|\|\mathcal{A}(\mu)\|_2^2 - 1\Big| \leq \delta, \quad \forall \mu \in \mathcal{S}_\kappa \tag{50}$$

The radius of $\mathcal{S}_\kappa$ with respect to certain function norms will play an important role. Since this notion will come up repeatedly in the analysis, we introduce the following notation, which will be heavily used in the sequel. Given a norm $\|\cdot\|$ on measures, the radius of a subset $\mathcal{E}$ of finite signed measures is denoted

$$\|\mathcal{E}\| := \sup_{\mu \in \mathcal{E}} \|\mu\|, \tag{51}$$

where we recall that the metric $\|\cdot\|_{\mathcal{G}}$ is defined in (8). With these definitions we can observe that

$$\sup_{\tau,\tau' \in \mathfrak{S}, \|\tau-\tau'\|_\kappa > 0} \frac{\|\tau - \tau'\|_{\mathcal{G}}}{\|\tau - \tau'\|_\kappa} = \|\mathcal{S}_\kappa(\mathfrak{S})\|_{\mathcal{G}};$$

in particular, the constant from (47) can be equivalently rewritten as $C_\kappa := \|\mathcal{S}_\kappa\|_{\Delta\mathcal{L}}$.

Concerning 48, the strategy to establish it will rely on the following two quantities: first, a *concentration function* $c_\kappa(t)$ characterizing the pointwise (i.e. for fixed $\tau, \tau' \in \mathfrak{S}$) concentration of $\|\mathcal{A}(\tau) - \mathcal{A}(\tau')\|_2^2$ around its expectation; secondly, certain covering numbers of $\mathcal{S}_\kappa$ needed to step from pointwise to uniform concentration.

Classical arguments from compressive sensing [Baraniuk et al., 2008, Eftekhari and Wakin, 2015, Puy et al., 2017, Dirksen, 2016, Foucart and Rauhut, 2012] prove that certain random linear operators satisfy the RIP by relying on pointwise concentration inequalities. Similarly, a first step to establish that the inequalities (48) hold with high probability consists in assuming first a pointwise version of the same, i.e., for any choice of m in (43):

$$\text{for any } \mu \in \mathcal{S}_\kappa : \qquad \mathbb{P}\Big(\Big|\|\mathcal{A}(\mu)\|_2^2 - 1\Big| \geq t\Big) \leq 2\exp\Big(-\frac{m}{c_\kappa(t)}\Big), \tag{52}$$

for some *concentration function* $t \mapsto c_\kappa(t)$ that should ideally be as small as possible. The following result shows that the radius $\|\mathcal{S}_\kappa\|_{\mathcal{F}}$ can be used to control such a concentration function.

Lemma 5.5. *Consider a family of functions $\mathcal{F} := \{\phi_\omega\}_{\omega\in\Omega}$, m parameters $(\omega_j)_{j=1}^m$ drawn i.i.d. according to some distribution Λ on Ω, and $\mathcal{A}$ the (random) operator induced (see (5)) by the feature function $\Phi(x) := \frac{1}{\sqrt{m}}\big(\phi_{\omega_j}(x)\big)_{j=1}^m$. Denoting κ the associated average kernel (cf (46)) we have $\|\mathcal{S}_\kappa\|_{\mathcal{F}} \geq 1$. Moreover, if $\|\mathcal{S}_\kappa\|_{\mathcal{F}} < \infty$ then (52) holds for all $\mu \in \mathcal{S}_\kappa$, with*

$$c_\kappa(t) \leq 2t^{-2}(1 + t/3) \cdot \|\mathcal{S}_\kappa\|_{\mathcal{F}}^2, \qquad \forall t > 0. \tag{53}$$

The proof is in Appendix C. Observe that the above estimate only depends on the choice of the feature family $\mathcal{F}$, and holds for any feature sampling distribution Λ. More refined estimates for mixture models, exploiting moments of Λ rather than a uniform bound, are provided in the companion paper [Gribonval et al., 2020] and used to obtain concrete estimates for Compressive Clustering and Compressive GMM.

Finally, we will extrapolate pointwise concentration (52) to all pairs $\tau, \tau' \in \mathfrak{S}$ using covering numbers of the normalized secant set with respect to an appropriate metric.

Definition 5.6 (Covering number)**.** *The* covering number $\mathrm{N}(d(\cdot,\cdot), S, \delta)$ *of a set S with respect to a (pseudo)metric*[5] *$d(\cdot,\cdot)$ is the minimum number of closed balls of radius δ with respect to $d(\cdot,\cdot)$ with centers in S needed to cover S.*

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

As the normalized secant set is a subset of the infinite-dimensional space of finite-signed measures, it is not obvious when its covering numbers are finite. Controlling them can be nontrivial, yet this is feasible on a case by case basis as will be illustrated in the companion paper [Gribonval et al., 2020].

Covering numbers and pointwise concentration can be then combined to give rise to the following result (whose proof is in Appendix C) where the logarithm of the covering numbers somehow captures an intrinsic dimension of the considered learning task:

Theorem 5.7. *Consider $\mathcal{F} := \{\phi_\omega\}_{\omega\in\Omega}$ a family of functions, Λ a probability distribution on Ω, Φ the associated random feature function and κ the corresponding average kernel. Consider the pseudometric on $\mathcal{F}$-integrable probability distributions*[6]

$$d_{\mathcal{F}}(\pi,\pi') := \sup_{\omega\in\Omega}\Big||\mathbb{E}_{X\sim\pi}\phi_\omega(X)|^2 - |\mathbb{E}_{X'\sim\pi'}\phi_\omega(X')|^2\Big|. \tag{54}$$

Consider a model set $\mathfrak{S}$ and $\mathcal{S}_\kappa = \mathcal{S}_\kappa(\mathfrak{S})$ its normalized secant set. Assume that $\mathcal{S}_\kappa$ has finite covering numbers with respect to the pseudometric $d_{\mathcal{F}}$. For $0 < \delta, \zeta < 1$, if

$$m \geq c_\kappa(\delta/2)\cdot\log\Big(2\mathrm{N}(d_{\mathcal{F}},\mathcal{S}_\kappa,\delta/2)/\zeta\Big), \tag{55}$$

then, with probability at least $1-\zeta$ on the draw of $(\omega_j)_{j=1}^m \overset{i.i.d.}{\sim} \Lambda$, the operator $\mathcal{A}$ induced by Φ (cf (43) and (5)) satisfies

$$1-\delta \leq \frac{\|\mathcal{A}(\tau)-\mathcal{A}(\tau')\|_2^2}{\|\tau-\tau'\|_\kappa^2} \leq 1+\delta, \qquad \forall \tau,\tau' \in \mathfrak{S}. \tag{56}$$

When (56) holds, the LRIP (20) with $\eta = 0$ holds with constant $C_{\mathcal{A}} := \frac{\|\mathcal{S}_\kappa\|_{\Delta\mathcal{L}}}{\sqrt{1-\delta}}$ and $\eta = 0$.

5.3 Summary and applications

To briefly summarize the results in this section, in order to establish the LRIP property with respect to a given model $\mathfrak{S}$ in the context of a sketching operator Φ associated to a family of random features $\mathcal{F}$ and feature sampling distribution Λ we proceed as follows. After identifying the associated average kernel (44), the key quantities to estimate relative to the normalized secant $\mathcal{S}_\kappa(\mathfrak{S})$ are its radius $\|\mathcal{S}_\kappa\|_{\Delta\mathcal{L}}$ (which serves as a measure of compatibility between the kernel, the learning task, and the model set $\mathfrak{S}$), the pointwise concentration function $c_\kappa(.)$ from (52), and the covering numbers of $\mathcal{S}_\kappa$ with respect to the distance $d_{\mathcal{F}}$ from (54).

Even though the above ingredients and results may look quite abstract at this stage, we can turn them into concrete estimates on several examples. The resulting guarantees are summarized in Table 1 for the examples developed in detail in the companion paper [Gribonval et al., 2020](compressive k-Means, k-medians and GMM).

The random sketching results developed in the present section can also be used to revisit the illustrative PCA example from Section 4. Namely, while we have directly lifted from existing literature the RIP property (37) for random Gaussian sketching matrices applied to low-rank covariance matrices, the arguments used there to establish this property follow in essence the canvas of this section (pointwise concentration of the random operator to its average, then unifom concentration via appropriate covering number arguments). In this context, the squared MMD with respect to the averaged kernel is precisely the Frobenius norm between covariance matrices. The additional ingredient needed to complete the analysis is to relate the PCA excess risk to the Frobenius norm of differences of low rank matrices (see (72) in the technical Appendix E, which can be reinterpreted as a bound on $\|\mathcal{S}_\kappa\|_{\Delta\mathcal{L}}$ in the PCA setting).

[6] In fact, we consider the extension of d_Φ to finite, $\mathcal{F}$-integrable signed measures, see Appendix A.2.

Task	PCA	k-med./means ($p = 1/p = 2$)	Gaussian Mixture Model.
Hypothesis h	subspace $h \subset \mathbb{R}^d$ $\dim h = k$	k cluster centers $c_1, \ldots, c_k \in \mathbb{R}^d$	mixture π_h 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 $\ell(x, h)$	$\|x - P_h x\|_2^2$	$\min_{1 \leq l \leq k} \|x - c_l\|_2^p$	$-\log \pi_h(x)$
Feature function $\Phi(x)$	quadratic polyn. $\frac{1}{\sqrt{m}} \left(x^T \mathbf{L}_j x\right)_{j=1}^m$	weighted Fourier features $\frac{1}{\sqrt{m}} \left(\frac{e^{\jmath \omega_j^T x}}{w(\omega_j)} \right)_{j=1}^m$	Fourier features $\frac{1}{\sqrt{m}} \left(e^{\jmath \omega_j^T x} \right)_{j=1}^m$
Sampling law Λ	$\mathbb{P}(\mathbf{L}) \propto e^{-\|\mathbf{L}\|_F^2}$	$\mathbb{P}(\omega) \propto w^2(\omega) e^{-\frac{s^2 \|\omega\|^2}{2}}$	$\mathbb{P}(\omega) \propto e^{-\frac{s^2 \|\omega\|_{\mathbf{\Sigma}^{-1}}^2}{2}}$
Average kernel $\kappa(x, x')$	$\left\|xx^T - x'x'^T\right\|_F^2$	$\exp\left(-\frac{\|x - x'\|_2^2}{2s^2}\right)$	$\exp\left(-\frac{\|x - x'\|_{\mathbf{\Sigma}}^2}{2s^2}\right)$
Proxy (27) $R(h, \mathbf{y})$	Low-rank recovery	$\min_{\alpha \in \mathbb{S}_{k-1}} \|\sum_{l=1}^k \alpha_l \Phi(c_l) - \mathbf{y}\|_2$	$\|\mathcal{A}(\pi_h) - \mathbf{y}\|_2$
Restrictions on hypothesis class $\mathcal{H}$ when optim- mizing proxy	N/A	$\min_{c_l \neq c_{l'}} \|c_l - c_{l'}\|_2 \geq 2\varepsilon$ $\max_l \|c_l\|_2 \leq R$ $\varepsilon := 4s\sqrt{\log(ek)}$ $w(\omega) := 1 + \frac{s^2 \|\omega\|_2^2}{d}$	$\min_{c_l \neq c_{l'}} \|c_l - c_{l'}\|_{\mathbf{\Sigma}} \geq 2\varepsilon$ $\max_l \|c_l\|_{\mathbf{\Sigma}} \leq R$ $\varepsilon := 4\sqrt{(2 + s^2)\log(ek)}$ known covariance $\mathbf{\Sigma}_l = \mathbf{\Sigma}, \forall l$
Sketch size m	kd	$k^2 d \log(ekdR/\varepsilon) \log^2(ek)$	$k^2 d \log(ekdR) \log^2(ek)$ when $s = \sqrt{d}$, cf [Gribonval et al., 2020](Table 1) for other values of s.

Table 1: Summary of the application of the framework on our three main examples (detailed in Section 4 and the companion paper [Gribonval et al., 2020]) 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 $\mathbf{\Sigma}$. The order of the sketch size is indicated up to universal numerical multiplicative factor and logarithmic dependencies on the parameters δ and ζ from Theorem 5.7. The displayed average kernels are up to a multiplicative constant.

6 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 set up a general mathematical framework for compressive statistical learning and to demonstrate on an example (compressive PCA) that the excess risk of this procedure can be controlled, as well as the sketch size. The companion paper [Gribonval et al., 2020] completes the illustration of the framework by considering two more examples: compressive clustering and compressive Gaussian mixture estimation — with fixed known covariance.

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. A number of non-sharp oracle inequalities have been established in the course of our endeavor. A particular question is to obtain more explicit and/or tighter control of the bias term $d_{h^\star_\pi}(\pi, \mathfrak{S}^{\mathsf{CT}}(\mathcal{H}))$, and to understand whether Lemma 3.2, 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 convergence rates for the excess risk can be established for certain classical statistical learning tasks under appropriate conditions (see e.g. [Levrard, 2013] for the case of k-means), it is natural to wonder whether the same holds for compressive statistical learning.

Links with neural networks. From an algorithmic perspective, the sketching techniques we have explicitly characterized in this paper have a particular structure which is reminiscent of a one-layer (random) neural network with subsequent averaging over multiple data points. Indeed, when the sketching function Φ corresponds to random Fourier features, its computation for a given vector x involves first multiplication by the matrix $\mathbf{W} \in \mathbb{R}^{m \times d}$ whose rows are the selected frequencies $\omega_j \in \mathbb{R}^d$, then pointwise application of the $e^{\jmath \cdot}$ nonlinearity. Here we consider random Fourier *moments*, hence a subsequent *averaging* operation is performed. As we have seen, this draws a link with the MMD, as is done e.g. in the so-called MMD-GANS [Li et al., 2015, Binkowski et al., 2018], where the so-called discriminator is a neural net trained to compute MMDs over batches of samples.

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 [Shwartz-Ziv and Tishby, 2017]. 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 work on the invertibility of deep networks and pooling representations with random Gaussian weights [Estrach et al., 2014, Giryes et al., 2016, Gilbert et al., 2017]. Finally, we mention the recent popularity of networks with random weights in statistical physics [Gabrié et al., 2018] and in analyzing the initialization point of optimization algorithms with a kernel characterization [Jacot et al., 2018, Bietti and Mairal, 2019], for which information-preservation (non-degeneracy during training) is also an essential feature.

Privacy-aware learning via sketching? The reader may have noticed that, while we have defined sketching in (3) 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 $\mathcal{A}(\pi) = \mathbb{E}_{X \sim \pi} \Phi(X)$ of the underlying probability distribution. A consequence is that one can envision *other sketching mechanisms*, in particular ones more compatible with privacy-preservation constraints [Duchi et al., 2014]. 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. Preliminary results in this direction have been recently achieved Schellekens et al. [2019].

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 and its companion [Gribonval et al., 2020]. 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

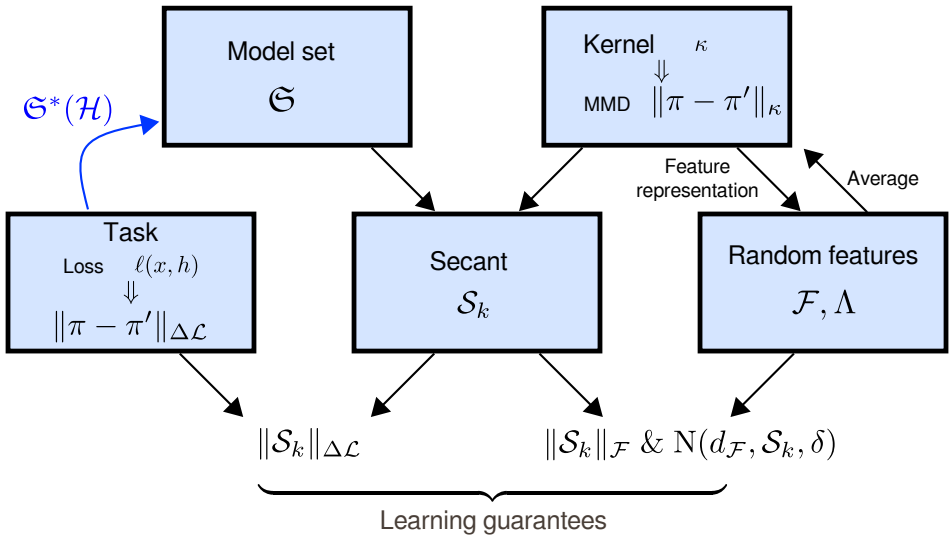

Figure 1: A representation of the links between different concepts in this paper.

and compressive regression are also natural candidate tasks in a learning setting, but seem more challenging.

Given a learning task, the main bottleneck is to find an adequate sketching function $\Phi(\cdot)$. As illustrated on Figure 1, this primarily relies on the quest for a *task-compatible kernel*, i.e., one satisfying the Kernel LRIP (47). 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 (54). On a case by case basis, one may have to identify the analog of the separation conditions apparently needed for compressive k-means, see [Gribonval et al., 2020].

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 lead to sketches linear in the product probability $\pi \otimes \pi$ instead of π. The investigation of such extended scenarios is expected to benefit from analogies with the lifting techniques used in phaseless reconstruction, see e.g. [Candès et al., 2013].

Acknowledgements

This work was supported in part by the European Research Council, PLEASE project (ERC-StG-2011-277906), and the german DFG (FOR-1735 "Structural inference in statistics", SFB-1294 "Data Assimilation"). 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, as well as Gabriel Peyré and Lorenzo Rosasco for their constructive feedback on early versions of this manuscript.

Appendix

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 **semi-norm** $\|\cdot\|$ over a vector space X satisfies the axioms of a norm except that $\|x\| = 0$ does not necessarily imply $x = 0$.*

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** $\mathrm{N}(d,Y,\delta) \in \mathbb{N} \cup \{+\infty\}$ 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 measurable sample space $(\mathcal{Z}, \mathfrak{Z})$ is a linear space that contains the set of probability distributions on $(\mathcal{Z}, \mathfrak{Z})$. By the Hahn-Jordan theorem, 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 $(\mathcal{Z}, \mathfrak{Z})$, hence $\mu_+ = \alpha\pi_+$ and $\mu_- = \beta\pi_-$ for some probability distributions π_+, π_-, and non-negative scalars $\alpha, \beta \geq 0$. A real-valued, measurable function f on $(\mathcal{Z}, \mathfrak{Z})$ is said integrable with respect to μ when it is integrable both with respect to μ_+ and μ_-. Noticing that the expectation of an integrable function f is linear in the considered probability distribution, we adopt the inner product notation for expectations:

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

being understood that we implicitly assume that f is integrable with respect to π when using this notation. This 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 $\mathcal{G}$ of measurable functions $g : \mathcal{Z} \to \mathbb{R}$ or $\mathbb{C}$ we can define

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

and check that this is a semi-norm on the linear subspace $\{\mu \in \mathfrak{M} : \forall f \in \mathcal{G}, f \text{ integrable wrt. } \mu\}$ as claimed when we introduced (8). Similarly, pseudometrics similar to (54) can be extended to finite signed measures as

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

When the functions in $\mathcal{G}$ 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 $\mathcal{B} = \{f : f \text{ is continuous and } \|f\|_\infty \leq 1\}$ [see e.g. Sriperumbudur et al., 2010] and yields a Banach structure on $\mathfrak{M}$ [see e.g. Halmos, 2013].

The mean kernel κ (cf (44)) 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

$$\kappa(\mu_1, \mu_2) := \alpha_1\alpha_2\kappa(\pi_1, \pi_2) - \alpha_1\beta_2\kappa(\pi_1, \pi_2') - \beta_1\alpha_2\kappa(\pi_1', \pi_2) + \beta_1\beta_2\kappa(\pi_1', \pi_2'), \tag{57}$$

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:

$$\|\mu\|_\kappa^2 := \iint \kappa(x, x')d\mu(x)d\mu(x') = \kappa(\mu, \mu), \tag{58}$$

that coincides with the metric of the mean kernel (45) for probability distributions.

B Proof of Theorem 2.5

In this section, we start with a suitable generalization of [Bourrier et al., 2014, Section IV-A], working with some relaxed assumptions on the considered metrics.

Definition B.1 (hemimetric)**.** *A function $d(\cdot\|\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a* hemimetric *if*

$$\begin{aligned} &d(x\|x) = 0, \quad \forall x \in \mathcal{X} \\ &d(x\|y) \geq 0, \quad \forall x, y \in \mathcal{X} \\ &d(x\|y) \leq d(x\|z) + d(z\|y), \quad \forall x, y, z \in \mathcal{X} \end{aligned}$$

A hemimetric is a pseudometric *if it is symmetric: $d(x\|y) = d(y\|x)$ for any $x, y \in \mathcal{X}$.*

Hemimetrics on basic sets such as $\mathcal{X} = \mathbb{R}^n$ will be denoted $d(x\|y)$. Hemimetrics between probability distributions will be denoted $D(\pi\|\pi')$. The notation $d(x, y)$ will preferentially be used for pseudometrics on basic sets, while $D(\pi, \pi')$ will denote pseudometrics between probability distributions.

Definition B.2 (relaxed lower restricted isometry property (rLRIP))**.** *A function $\Psi : \mathcal{X} \to \mathcal{Y}$ satisfies the lower restricted isometry property on the subset $\Sigma \subset \mathcal{X}$ with respect to the hemimetric $d_{\mathcal{X}}(\cdot\|\cdot)$ on $\mathcal{X}$ and the pseudometric $d_{\mathcal{Y}}(\cdot,\cdot)$ on $\mathcal{Y}$ with constant $\eta \geq 0$ iff*

$$d_{\mathcal{X}}(x\|x') \leq d_{\mathcal{Y}}(\Psi(x), \Psi(x')) + \eta, \quad \forall x, x' \in \Sigma. \tag{59}$$

Lemma B.3. *Assume that Ψ satisfies the rLRIP on Σ with respect to $d_{\mathcal{X}}(\cdot\|\cdot)$ and $d_{\mathcal{Y}}(\cdot,\cdot)$ with constant η. Consider $\varepsilon, \nu \geq 0$ and a decoder $\Delta : \mathcal{Y} \to \Sigma \subset \mathcal{X}$ such that*

$$d_{\mathcal{Y}}(y, \Psi(\Delta(y))) \leq (1 + \nu) \inf_{z \in \Sigma} d_{\mathcal{Y}}(y, \Psi(z)) + \varepsilon, \quad \forall y \in \mathcal{Y}. \tag{60}$$

Then Δ satisfies the instance optimality property: $\forall x^* \in \mathcal{X}, y \in \mathcal{Y}$,

$$d_{\mathcal{X}}(x^*\|\Delta(y)) \leq \mathcal{D}^*(x^*\|\Sigma) + (2 + \nu)d_{\mathcal{Y}}(y, \Psi(x^*)) + \eta + \varepsilon, \tag{61}$$

where

$$\mathcal{D}^*(x\|\Sigma) = \inf_{z \in \Sigma} \mathcal{D}^*(x\|z); \qquad \mathcal{D}^*(x\|z) := d_{\mathcal{X}}(x^*\|z) + (2 + \nu)d_{\mathcal{Y}}(\Psi(x^*), \Psi(z)) \tag{62}$$

Proof. The proof follows very closely [Bourrier et al., 2014] and is adaptated to the fact that $d_{\mathcal{X}}(\cdot\|\cdot)$ is a hemimetric. Consider $x^* \in \mathcal{X}$, $y \in \mathcal{Y}$ and $\hat{x} = \Delta(y)$. Consider any $z \in \Sigma$ and write

$$\begin{aligned} d_{\mathcal{X}}(x^*\|\hat{x}) &\leq d_{\mathcal{X}}(x^*\|z) + d_{\mathcal{X}}(z\|\hat{x}) \\ &\overset{rLRIP}{\leq} d_{\mathcal{X}}(x^*\|z) + d_{\mathcal{Y}}(\Psi(z), \Psi(\hat{x})) + \eta \\ &\leq d_{\mathcal{X}}(x^*\|z) + d_{\mathcal{Y}}(\Psi(z), y) + d_{\mathcal{Y}}(y, \Psi(\hat{x})) + \eta \\ &\overset{(60)}{\leq} d_{\mathcal{X}}(x^*\|z) + (2+\nu) d_{\mathcal{Y}}(y, \Psi(z)) + \eta + \varepsilon \\ &\leq d_{\mathcal{X}}(x^*\|z) + (2+\nu) d_{\mathcal{Y}}(y, \Psi(x^*)) + (2+\nu) d_{\mathcal{Y}}(\Psi(x^*), \Psi(z)) + \eta + \varepsilon \\ &= d_{\mathcal{X}}(x^*\|z) + (2+\nu) d_{\mathcal{Y}}(\Psi(x^*), \Psi(z)) + (2+\nu) d_{\mathcal{Y}}(y, \Psi(x^*)) + \eta + \varepsilon \end{aligned}$$

As this holds for any $z \in \Sigma$, taking the infimum yields the result. □

Remark B.4. *Conversely, when $d_{\mathcal{X}}(\cdot\|\cdot)$ is a* pseudometric*, if* some *decoder satisfies* (61) *for each $x^* \in \mathcal{X}, y \in \mathcal{Y}$ with $\mathcal{D}^*(\cdot\|\cdot)$* some *function such that $\mathcal{D}^*(x\|\Sigma) = 0$ for each $x \in \Sigma$, then the rLRIP* (59) *holds with constant $2(\eta+\varepsilon)$ with respect to $\widetilde{d_{\mathcal{Y}}}(\cdot,\cdot) = (2+\nu) d_{\mathcal{Y}}(\cdot,\cdot)$. Indeed, for $x, x' \in \Sigma$, as $\mathcal{D}^*(x\|\Sigma) = \mathcal{D}^*(x'\|\Sigma) = 0$, by* (61) *and the symmetry of $d_{\mathcal{X}}(\cdot\|\cdot)$, we have with $y := \Psi(x)$, $\hat{x} := \Delta(y)$:*

$$\begin{aligned} d_{\mathcal{X}}(x\|\hat{x}) &\leq \mathcal{D}^*(x\|\Sigma) + (2+\nu) d_{\mathcal{Y}}(y, \Psi(x)) + \eta + \varepsilon = \eta + \varepsilon \\ d_{\mathcal{X}}(\hat{x}\|x') = d_{\mathcal{X}}(x'\|\hat{x}) &\leq \mathcal{D}^*(x'\|\Sigma) + (2+\nu) d_{\mathcal{Y}}(y, \Psi(x')) + \eta + \varepsilon = \widetilde{d_{\mathcal{Y}}}(\Psi(x), \Psi(x')) + \eta + \varepsilon. \end{aligned}$$

The triangle inequality yields $d_{\mathcal{X}}(x\|x') \leq d_{\mathcal{X}}(x\|\hat{x}) + d_{\mathcal{X}}(\hat{x}\|x') \leq \widetilde{d_{\mathcal{Y}}}(\Psi(x), \Psi(x')) + 2(\gamma)\eta + \varepsilon)$.

To prove Theorem 2.5, given a fixed $h_0 \in \mathcal{H}$, we apply Lemma B.3 with the metrics $d_{\mathcal{X}}(\pi\|\pi') := D_{h_0}(\pi\|\pi')$, $\Psi(\cdot) = \mathcal{A}(\cdot)$, and $d_{\mathcal{Y}}(\cdot,\cdot) := C_{\mathcal{A}}\|\cdot - \cdot\|_2$. By (19)-(20) the rLRIP (59) is satisfied by Ψ on $\Sigma := \mathfrak{S}$. By (22) $y := \mathbf{y} = \mathcal{A}(\hat{\pi}_n)$ and $\Delta(\mathbf{y}) = \widetilde{\pi}$ satisfy (60) with ε replaced by $C_{\mathcal{A}}\varepsilon$. Since the definition (62) with yields $\mathcal{D}^*(\pi\|\Sigma) = d_{h_0}(\pi\|\mathfrak{S})$ as in (25), we get by (61) with $x^\star := \pi$:

$$D_{h_0}(\pi\|\widetilde{\pi}) \leq d_{h_0}(\pi, \mathfrak{S}) + (2+\nu) C_{\mathcal{A}} \|\mathcal{A}(\hat{\pi}_n) - \mathcal{A}(\pi)\| + \eta + C_{\mathcal{A}}\varepsilon.$$

On the other hand, using (23), we obtain

$$\Delta\mathcal{R}_{h_0}(\widetilde{\pi}, \hat{h}) = \mathcal{R}(\widetilde{\pi}, \hat{h}) - \mathcal{R}(\widetilde{\pi}, h_0) \leq \varepsilon',$$

and thus, by definition of $D_{h_0}(\pi\|\widetilde{\pi})$ (cf (18)):

$$\Delta\mathcal{R}_{h_0}(\pi, \widetilde{h}) \leq \Delta\mathcal{R}_{h_0}(\pi, \hat{h}) - \Delta\mathcal{R}_{h_0}(\widetilde{\pi}, \hat{h}) + \varepsilon' \leq D_{h_0}(\pi\|\widetilde{\pi}) + \varepsilon'. \tag{63}$$

C Proof of Lemma 5.5 and Theorem 5.7

To establish Lemma 5.5 we use Bernstein's inequality for bounded random variables, which is for example a consequence of Massart 2007, Corollary 2.10.

Lemma C.1 (Bernstein's inequality)**.** *Let $X_i \in \mathbb{R}$, $i = 1, \ldots, N$ 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*

$$P\left(\frac{1}{N}\sum_{i=1}^{N} X_i \geq t\right) \leq \exp\left(-\frac{Nt^2}{2\sigma^2 + 2Mt/3}\right). \tag{64}$$

Proof of Lemma 5.5. First, observe that for any $\mathcal{F}$-integrable probability distributions π, π'

$$\|\pi - \pi'\|_\kappa^2 = \mathbb{E}_{\omega\sim\Lambda}|\langle \pi, \phi_\omega\rangle - \langle \pi', \phi_\omega\rangle|^2 \le \sup_{\omega\sim\Lambda}|\langle \pi, \phi_\omega\rangle - \langle \pi', \phi_\omega\rangle|^2 = \|\pi - \pi'\|_{\mathcal{F}}^2$$

and that

$$\frac{\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2^2}{\|\pi - \pi'\|_\kappa^2} - 1 = \tfrac{1}{m}\sum_{j=1}^m Z(\omega_j) \qquad \text{with} \qquad Z(\omega) := \frac{|\langle \pi, \phi_\omega\rangle - \langle \pi', \phi_\omega\rangle|^2}{\|\pi - \pi'\|_\kappa^2} - 1$$

Specializing to $\tau, \tau' \in \mathfrak{S}$ we get $1 \le C := \|\tau - \tau'\|_{\mathcal{F}}/\|\tau - \tau'\|_\kappa \le \|\mathcal{S}_\kappa\|_{\mathcal{F}}$ and $-1 \le Z(\omega) \le C^2 - 1$, hence $\|\mathcal{S}_\kappa\|_{\mathcal{F}} \ge 1$ and $|Z(\omega)| \le \max\big(1, C^2 - 1\big) \le C^2 \le \|\mathcal{S}_\kappa\|_{\mathcal{F}}^2$. As $\mathbb{E}_{\omega\sim\Lambda} Z(\omega) = 0$ we obtain

$$\begin{aligned}\mathtt{Var}_{\omega\sim\Lambda}(Z(\omega)) =& \mathtt{Var}_{\omega\sim\Lambda}\left(\frac{|\langle \tau-\tau', \phi_\omega\rangle|^2}{\|\tau-\tau'\|_\kappa^2}\right) \le \frac{\mathbb{E}_{\omega\sim\Lambda}|\langle \tau - \tau', \phi_\omega\rangle|^4}{\|\tau - \tau'\|_\kappa^4} \\ \le& \frac{\mathbb{E}_{\omega\sim\Lambda}\|\tau - \tau'\|_{\mathcal{F}}{}^2 \cdot |\langle \tau - \tau', \phi_\omega\rangle|^2}{\|\tau - \tau'\|_\kappa^4} = \frac{\|\tau - \tau'\|_{\mathcal{F}}{}^2}{\|\tau - \tau'\|_\kappa^2} = C^2.\end{aligned}$$

Applying Lemma C.1 with the independent random variables $Z(\omega)$ we obtain for each $t > 0$:

$$\mathbb{P}\left(\left|\frac{\|\mathcal{A}(\tau - \tau')\|_2^2}{\|\tau - \tau'\|_\kappa^2} - 1\right| \ge t\right) \le 2\exp\left(-\frac{mt^2}{2C^2 \cdot (1 + t/3)}\right). \qquad \square$$

Lemma C.2. *Consider a family of functions $\mathcal{F} := \{\phi_\omega\}_{\omega\in\Omega}$, m parameters $(\omega_j)_{j=1}^m$ drawn i.i.d. according to some distribution Λ on Ω, and $\mathcal{A}$ the (random) operator induced (see* (5)*) by the feature function $\Phi(x) := \frac{1}{\sqrt{m}}\big(\phi_{\omega_j}(x)\big)_{j=1}^m$. and κ the associated kernel. Assume that* (52) *holds with concentration function $c_\kappa(t)$ and consider $\mathcal{S} \subset \mathcal{S}_\kappa$ and $d_{\mathcal{F}}$ the metric defined in* (54)*. For any $\delta > 0$ such that*

$$N := \mathrm{N}(d_{\mathcal{F}}, \mathcal{S}, \delta/2) < \infty, \tag{65}$$

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

$$\sup_{\mu\in\mathcal{S}}\Big|\|\mathcal{A}(\mu)\|_2^2 - 1\Big| \le \delta. \tag{66}$$

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

$$\mathbb{P}\Big(\Big|\|\mathcal{A}(\mu)\|_2^2 - 1\Big| \ge t\Big) \le 2\exp(-m/c_\kappa(t)). \tag{67}$$

This establishes a pointwise concentration result when μ is on the normalized secant set $\mathcal{S}_\kappa$. We now use a standard argument to extend this to a uniform result on $\mathcal{S}$. Let μ_i, $1 \le i \le N$ be the centers of a $\delta/2$-covering (with respect to the metric $d_{\mathcal{F}}$) of $\mathcal{S}$. Using (67) with $t = \delta/2$, the probability that there is an index i such that $\Big|\|\mathcal{A}(\mu_i)\|_2^2 - 1\Big| \ge \delta/2$ 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 $d_{\mathcal{F}}(\mu, \mu_i) \le \delta/2$:

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

Proof of Theorem 5.7. Denote $\zeta = 2N\exp(-m/c_\kappa(\delta/2))$ with $N := \mathrm{N}(d_{\mathcal{F}}, \mathcal{S}_\kappa, \delta/2)$. By Lemma C.2, the assumptions imply that with probability at least $1-\zeta$ on the draw of ω_j, $1 \le j \le m$, we have

$$\sup_{\mu \in \mathcal{S}_\kappa} \left| \|\mathcal{A}(\mu)\|_2^2 - 1 \right| \le \delta.$$

This implies (48). Since $\|\mathcal{S}_\kappa\|_{\Delta\mathcal{L}} < \infty$, the LRIP (20) holds wrt $\|\cdot\|_{\Delta\mathcal{L}}$ with $C_{\mathcal{A}} = \frac{\|\mathcal{S}_\kappa\|_{\Delta\mathcal{L}}}{\sqrt{1-\delta}}$. □

D Proof of Lemma 3.2 and Lemma 3.4

If $\pi \in \mathfrak{S}_h^{\mathtt{CT}}$ then $0 = \mathcal{R}(\pi, h) = \mathbb{E}_{X\sim\pi}\ell(X,h) = \mathbb{E}_{X\sim\pi}d^p(X, P_hX)$ hence $d(X, P_hX) = 0$ almost surely, i.e., $X = P_hX \in P_h\mathcal{Z} = \mathcal{E}_h$ almost surely. The converse is trivial. The bound (33) follows directly since for any $h \in \mathcal{H}$, $P_h\pi \in \mathfrak{S}^{\mathtt{CT}}(h) \subset \mathfrak{S}^{\mathtt{CT}}(\mathcal{H})$. This establishes the first claim of Lemma 3.2.

Let $h_0 \in \mathcal{H}$ be fixed. By (29), with $Y \sim P_{h_0}\pi$, we have $P_{h_0}Y = Y$ hence $\ell(Y, h_0) = d^p(Y, P_{h_0}Y) = 0$ and for any $h \in \mathcal{H}$:

$$\begin{aligned}\Delta\mathcal{R}_{h_0}(\pi,h) - \Delta\mathcal{R}_{h_0}(P_{h_0}\pi,h) &= \mathbb{E}_{X\sim\pi}\ell(X,h) - \mathbb{E}_{X\sim\pi}\ell(X,h_0) - \big(\underbrace{\mathbb{E}_{Y\sim P_{h_0}\pi}\ell(Y,h)}_{\mathbb{E}_{X\sim\pi}\ell(P_{h_0}X,h)} - \underbrace{\mathbb{E}_{Y\sim P_{h_0}\pi}\ell(Y,h_0)}_{0}\big)\\ &= \mathbb{E}_{X\sim\pi}\{\ell(X,h) - \ell(X,h_0) - \ell(P_{h_0}X,h)\}\\ &= \mathbb{E}_{X\sim\pi}\{d^p(X,P_hX) - d^p(X,P_{h_0}X) - d^p(P_{h_0}X, P_hP_{h_0}X)\}. \end{aligned} \tag{68}$$

For the second claim of Lemma 3.2, by (30), since d^p is a metric we have for any $x \in \mathcal{Z}$

$$d^p(x,P_hx) \le d^p(x,P_hP_{h_0}x) \le d^p(x,P_{h_0}x) + d^p(P_{h_0}x, P_hP_{h_0}x).$$

It follows using (68) that

$$\Delta\mathcal{R}_{h_0}(\pi,h) - \Delta\mathcal{R}_{h_0}(P_{h_0}\pi,h) \le 0.$$

As this holds for any h, and as equality is reached for $h = h_0$, we get $D_{h_0}(\pi\|P_{h_0}\pi) = 0$.
In particular when $p \in (0,1]$ we have $(a+b)^p \le a^p + b^p$ for any $a, b \ge 0$ hence for $u, v, w \in \mathcal{Z}$, by the triangle inequality, $d^p(u,v) \le [d(u,w) + d(w,v)]^p \le d^p(u,w) + d^p(w,v)$, showing that d^p is a metric.

For the claims of Lemma 3.4, we will exploit optimal 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 $p \ge 1$, and $\mathcal{Z}$ with d-diameter bounded by B, since $|a^p - b^p| \le \max(pa^{p-1}, pb^{p-1})|a-b|$ for any $a, b \ge 0$, we have

$$|\ell(x,h) - \ell(x',h)| \le pB^{p-1}|d(x,P_hx) - d(x',P_hx')| \le pB^{p-1}d(x,x'),$$

by the triangle inequality, hence $\mathcal{L}(\mathcal{H}) \subset \mathrm{Lip}(pB^{p-1}, d)$. Using (19) this implies that for any π, π' we have in general in the above considered setting:

$$D_{h_0}(\pi\|\pi') \le \|\pi - \pi'\|_{\Delta\mathcal{L}} \le 2\|\pi - \pi'\|_{\mathcal{L}} \le 2pB^{p-1}\|\pi - \pi'\|_{\mathrm{Lip}(1,d)}.$$

It is well-known that the 1-Wasserstein distance between two distributions can be equivalently defined in terms of optimal 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., [Dudley, 2002, Theorem 11.8.2]. By the transport characterization of the Wasserstein distance, considering the transport plan that sends x to P_hx, where $h \in \mathcal{H}'$, we conclude

$$\|\pi - P_h\pi\|_{\mathrm{Wasserstein}_1(d)} \le \mathbb{E}_{X\sim\pi}d(X, P_h(X)) \le [\mathbb{E}_{X\sim\pi}d^p(X,P_h(X))]^{\frac{1}{p}} = \mathcal{R}(\pi,h)^{\frac{1}{p}}, \tag{69}$$

by Jensen's inequality (since $p \geq 1$ here), yielding the claim (34).

For the final claim, we have

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

where $f_{\mathbf{u}}(x) := \langle \Phi(x), \mathbf{u}\rangle$. Moreover, for $\|\mathbf{u}\|_2 \leq 1$ and any x, x', since Φ is assumed L-Lipschitz:

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

i.e., $f_{\mathbf{u}}(\cdot)$ is L-Lipschitz with respect to $d(\cdot,\cdot)$. It follows that for any π, π', $\|\mathcal{A}(\pi) - \mathcal{A}(\pi')\|_2 \leq L\|\pi - \pi'\|_{\mathrm{Lip}(1,d)} = L\|\pi - \pi'\|_{\mathrm{Wasserstein}_1(d)}$.

The claim (35) when $p \geq 1$ follows by (69). When $p \leq 1$ and the space $\mathcal{Z}$ has d-diameter bounded by B, as $d(X, P_h X) = d^{1-p}(X, P_h X) d^p(X, P_h X) \leq B^{1-p} d^p(X, P_h X)$ we obtain (36) as follows

$$\|\pi - P_h\pi\|_{\mathrm{Wasserstein}_1(d)} \leq \mathbb{E}_{X\sim\pi} d(X, P_h(X)) \leq B^{1-p} \mathbb{E}_{X\sim\pi} d(X, P_h(X))^p = B^{1-p} \mathcal{R}(\pi, h).$$

E Proof of Theorem 4.1 on Compressive PCA

For Compressive PCA, we recall that k is the number of PCA components we want to estimate. The hypothesis class $\mathcal{H}$ is the set of linear subspaces of dimension k of the input space $\mathbb{R}^d$, which is in one-to-one correspondance with the space $\mathfrak{P}_k$ of orthoprojectors $\mathbf{P}$ of rank k. In the remainder of this section we therefore use directly $\mathfrak{P}_k$ as the hypothesis class, for notational convenience. We recall that for $r \geq k$, we consider the model $\mathfrak{S}_r$ consisting of probability distributions having their second moment matrix of rank at most r.

Observe that for any $\mathbf{P} \in \mathfrak{P}_k$:

$$\mathcal{R}_{k-\mathtt{PCA}}(\pi, \mathbf{P}) := \mathbb{E}_{X\sim\pi} \|X - \mathbf{P}X\|_2^2 = \langle \mathbf{\Sigma}_\pi, \mathbf{I} - \mathbf{P}\rangle_F,$$

where the inner product $\langle , \rangle_F$ is the Frobenius product, and the minimum risk is

$$\mathcal{R}_{k-\mathtt{PCA}}(\pi, \mathbf{P}_\pi^{*[k]}) = \inf_{\mathrm{rank}(\mathbf{M}) \leq k, \mathbf{M} \succcurlyeq 0} \|\mathbf{\Sigma}_\pi - \mathbf{M}\|_\star = \sum_{i>k} \lambda_i(\mathbf{\Sigma}_\pi), \tag{70}$$

where $\mathbf{P}_\pi^{*[\ell]}$ $(1 \leq \ell \leq d)$ denotes an orthoprojector onto the ℓ first eigenvectors of $\mathbf{\Sigma}_\pi$, and $\lambda_i(\mathbf{M})$ denote the eigenvalues, with multiplicity and ordered in nonincreasing sequence, of a matrix $\mathbf{M}$.

We follow the improved risk analysis of Section 2.5. In the above setting, the excess risk divergence (18) with respect to an *arbitrary* reference hypothesis $\mathbf{P}_0 \in \mathfrak{P}_k$ is given by

$$D_{\mathbf{P}_0}(\pi\|\pi') = \sup_{\mathbf{Q}\in\mathfrak{P}_k} \langle \mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\pi'}, \mathbf{P}_0 - \mathbf{Q}\rangle_F = \langle \mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\pi'}, \mathbf{P}_0\rangle_F - \inf_{\mathbf{Q}\in\mathfrak{P}_k} \langle \mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\pi'}, \mathbf{Q}\rangle_F. \tag{71}$$

Lower RIP. We start with the following bound on the excess risk divergence, holding without restriction for any distributions π, π' with existing second moments and any $\mathbf{P}_0 \in \mathfrak{P}_k$:

$$D_{\mathbf{P}_0}(\pi\|\pi') \leq \sqrt{2\min(k, d-k)} \|\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\pi'}\|_F. \tag{72}$$

Proof of (72). By the so-called Ky Fan Theorem Fan [1949], for a symmetric matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, and a positive integer $\ell \leq d$, one has

$$\sup_{\mathbf{P}\in\mathfrak{P}_\ell} \mathrm{Tr}(\mathbf{MP}) = \sum_{i=1}^{\ell} \lambda_i(\mathbf{M})\,.$$

As a result, we obtain

$$\begin{aligned}
D_{\mathbf{P}_0}(\pi\|\pi') &= \sup_{\mathbf{Q}\in\mathfrak{P}_k} \langle \mathbf{\Sigma}_{\pi'} - \mathbf{\Sigma}_\pi, \mathbf{Q}\rangle_F + \langle \mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\pi'}, \mathbf{P}_0\rangle_F \\
&\leq \sum_{i=1}^{k} \lambda_i(\mathbf{\Sigma}_{\pi'} - \mathbf{\Sigma}_\pi) + \sum_{i=1}^{k} \lambda_i(\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\pi'}) = \sum_{i=1}^{k} \lambda_i(\mathbf{\Sigma}_{\pi'} - \mathbf{\Sigma}_\pi) - \sum_{i=d-k+1}^{d} \lambda_i(\mathbf{\Sigma}_{\pi'} - \mathbf{\Sigma}_\pi) \\
&= \sum_{i=1}^{\min(k,d-k)} \lambda_i(\mathbf{\Sigma}_{\pi'} - \mathbf{\Sigma}_\pi) - \sum_{i=d-\min(k,d-k)+1}^{d} \lambda_i(\mathbf{\Sigma}_{\pi'} - \mathbf{\Sigma}_\pi) \\
&\leq \sqrt{2\min(k,d-k)}\sqrt{\sum_{i=1}^{d} \lambda_i^2(\mathbf{\Sigma}_{\pi'} - \mathbf{\Sigma}_\pi)} = \sqrt{2\min(k,d-k)}\|\mathbf{\Sigma}_{\pi'} - \mathbf{\Sigma}_\pi\|_F.
\end{aligned}$$

□

Since the right-hand side of (72) does not depend of $\mathbf{P}_0$, we get from (19) in particular that for any π, π' with finite second moments

$$\|\pi - \pi'\|_{\Delta\mathcal{L}} \leq \sqrt{2k}\|\mathbf{\Sigma}_{\pi'} - \mathbf{\Sigma}_\pi\|_F.$$

Hence, since $\mathcal{M}$ is a linear operator having a RIP (37) on matrices of rank lower than $2r$, $\mathcal{M}$ induces (in the way described in Section 4) a sketching operator $\mathcal{A} : \pi \mapsto \mathcal{A}(\pi) := \mathcal{M}(\mathbf{\Sigma}_\pi)$ that has the lower RIP described in (20) with constants $C_{\mathcal{A}} = \frac{\sqrt{2k}}{\sqrt{1-\delta}}, \eta = 0$ on model $\mathfrak{S}_r$.

Ideal decoder and generic excess risk control. The ideal decoder (12) writes

$$\Delta[\mathbf{y}] := \mathrm{argmin}_{\tau\in\mathfrak{S}_r}\|\mathcal{A}(\tau) - \mathbf{y}\|_2^2,$$

and is equivalent to (38), i.e.

$$\hat{\mathbf{\Sigma}} := \mathrm{argmin}_{\mathbf{\Sigma}:\mathrm{rank}(\mathbf{\Sigma})\leq r;\mathbf{\Sigma}\succcurlyeq 0}\|\mathcal{M}(\mathbf{\Sigma}) - \mathbf{y}\|_2^2.$$

Formally, $\Delta[\mathbf{y}]$ can then be taken as any distribution having second moment matrix $\hat{\mathbf{\Sigma}}$. This last step can naturally be shunted since whatever the choice of such a representative distribution, the associated estimated hypothesis $\hat{h}$ is directly given by (39).

By Theorem 2.5, this decoder is instance optimal yielding (24) with $\nu = \eta = \varepsilon = \varepsilon' = 0$, i.e., the excess risk of $\hat{h}$ of the procedure of Section 4 when the true data distribution is π is controlled by

$$d_{h_\pi^\star}(\pi, \mathfrak{S}_r) + 2C_{\mathcal{A}}\|\mathcal{A}(\pi) - \mathcal{A}(\hat{\pi}_n)\|_2 = d_{h_\pi^\star}(\pi, \mathfrak{S}_r) + 2C_{\mathcal{A}}\|\mathcal{M}(\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_{\hat{\pi}_n})\|_2, \tag{73}$$

with the bias term defined in (25).

Control of the bias term. The next lemma improves over Lemma 3.2 in the special case of PCA.

Lemma E.1. *Consider a probability distribution π with finite second moments, $\mathbf{P}^* := \mathbf{P}_\pi^{*[k]}$, and $\tau_r \in \mathfrak{S}_r$ (any) probability distribution with covariance $\mathbf{\Sigma}_\pi^{[r]} := \mathbf{P}^{*[r]}\mathbf{\Sigma}_\pi\mathbf{P}^{*[r]}$. We have $D_{\mathbf{P}^*}(\pi\|\tau_r) = 0$.*

Proof. Since $\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_\pi^{[r]}$ is a nonnegative matrix and $r \geq k$, (71) yields

$$D_{\mathbf{P}^*}(\pi\|\tau_r) = \underbrace{\left\langle \mathbf{\Sigma}_\pi - \mathbf{\Sigma}_\pi^{[r]}, \mathbf{P}^* \right\rangle_F}_{=0} - \inf_{\mathrm{Q}\in\mathfrak{P}_k} \underbrace{\left\langle \mathbf{\Sigma}_\pi - \mathbf{\Sigma}_\pi^{[r]}, \mathbf{Q} \right\rangle_F}_{\geq 0} = 0. \qquad \square$$

As a result the "bias" term in the bound (73) is upper bounded as

$$d_{h_\pi^\star}(\pi, \mathfrak{S}_r) = \inf_{\tau\in\mathfrak{S}_r} (D_{\mathbf{P}^*}(\pi\|\tau) + 2C_{\mathcal{A}}\|\mathcal{A}(\pi) - \mathcal{A}(\tau)\|_2) \leq 2C_{\mathcal{A}} \left\| \mathcal{M}(\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_\pi^{[r]}) \right\|_2. \tag{74}$$

From now on we assume $d > k$ (otherwise the bias is trivially 0). We now use the following lemma to bound $\left\| \mathcal{M}(\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_\pi^{[r]}) \right\|_2$.

Lemma E.2. *Let $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ be symmetric p.s.d. and $\mathcal{M} : \mathbb{R}^{d\times d} \to \mathbb{R}^m$ satisfy the upper RIP in (37), i.e. $\|\mathcal{M}(\mathbf{M})\|_2^2/\|\mathbf{M}\|_F^2 \leq 1+\delta$ for all matrices of rank at most $2r$. For $1 \leq \ell \leq d$, consider $\mathbf{\Sigma}^{[\ell]}$ a best rank ℓ approximation to $\mathbf{\Sigma}$. Then for any $1 \leq s \leq \min(\ell, 2r)$ we have*

$$\left\| \mathcal{M}(\mathbf{\Sigma} - \mathbf{\Sigma}^{[\ell]}) \right\|_2 \leq \sqrt{1+\delta}\frac{\sigma_{\ell-s+1}}{\sqrt{s}},$$

where $\sigma_j := \sigma_j(\mathbf{\Sigma}) := \sum_{i=j+1}^d \lambda_i(\mathbf{\Sigma})$.

Proof. This proof follows mainly the ideas of Candès [2008]. By definition of $\mathbf{\Sigma}^{[\ell]}$ there is an eigendecomposition $\mathbf{\Sigma} = U\Lambda U^T$, where Λ is a diagonal matrix containing the eigenvalues of $\mathbf{\Sigma}$ with multiplicity in decreasing order, such that $\mathbf{\Sigma}^{[\ell]} = U\Lambda^{[\ell]}U^T$ where $\Lambda^{[\ell]}$ contains the first ℓ eigenvalues. Decompose Λ into blocks, $\Lambda = \sum_{j\geq 0}\Lambda_j$, where Λ_0 contains the first $\ell - s$ eigenvalues, and Λ_j, $j \geq 1$ are the next blocks of s eigenvalues in decreasing order (the last block is of size $\leq s$); that is, for $j \geq 1$ the block Λ_j contains eigenvalues of indices m such that $\ell + (j-2)s < m \leq \ell + (j-1)s$. Let us also denote Λ_j^+, for $j \geq 1$, the blocks of eigenvalues of size s starting one index later, that is, Λ_j contains eigenvalues of indices m such that $\ell + (j-2)s + 1 < m \leq \ell + (j-1)s + 1$. Let $S_j = U\Lambda_j U^T$, so that $\mathbf{\Sigma} - \mathbf{\Sigma}^{[\ell]} = \sum_{j\geq 2} S_j$. As $s \leq 2r$, $\mathrm{rank}(S_j) \leq 2r$, $j \geq 2$ hence by the upper RIP property,

$$\begin{aligned}
\tfrac{1}{\sqrt{1+\delta}}\left\| \mathcal{M}(\mathbf{\Sigma} - \mathbf{\Sigma}^{[\ell]}) \right\|_2 &= \tfrac{1}{\sqrt{1+\delta}}\left\| \sum_{j\geq 2} \mathcal{M}(S_j) \right\|_2 \leq \tfrac{1}{\sqrt{1+\delta}} \sum_{j\geq 2} \|\mathcal{M}(S_j)\|_2 \leq \sum_{j\geq 2} \|S_j\|_F \\
&= \sum_{j\geq 2} \|\Lambda_j\|_2 \leq \sum_{j\geq 2} \sqrt{s}\|\Lambda_j\|_\infty \leq \sum_{j\geq 2} \sqrt{s}\frac{\|\Lambda_{j-1}^+\|_1}{s} = \sum_{j\geq 1} \frac{\|\Lambda_j^+\|_1}{\sqrt{s}} = \frac{\sigma_{\ell-s+1}}{\sqrt{s}}.
\end{aligned}$$

$\square$

The above lemma (with $\ell := r$) allows us to control $\|\mathcal{A}(\pi) - \mathcal{A}(\pi_r)\| = \left\| \mathcal{M}(\mathbf{\Sigma}_\pi - \mathbf{\Sigma}_\pi^{[r]}) \right\|_2$ using σ_{r-s+1} for $1 \leq s \leq r$. This, in combination with (74) and (73), establishes (40). Moreover, choosing $s := r - k + 1$, we get $\sigma_{r-s+1} = \sigma_k = \mathcal{R}_{k-\mathtt{PCA}}(\pi, \mathbf{P}_\pi^{*[k]})$, leading to (41).This proves Theorem 4.1.

Table of notations

$x \in \mathcal{Z}$	sample and sample space
$\mathbf{y}$	sketch vector (3)
Φ	sketching function (3), (43),
$\mathcal{A}$	sketching operator (5)
π, τ	probabilities on sample space
μ, ν	measures on sample space
$\langle \pi, f \rangle$	$\mathbb{E}_{X\sim\pi} f(X)$
$\langle \mu, f \rangle$	$\int f(x) d\mu(x)$ (App. A.2)
h	hypothesis
$\mathcal{H}$	class of hypotheses
$\ell(\cdot, h)$	loss function
$\mathcal{R}, \Delta\mathcal{R}_h$	risk (1), excess risk (Def. 2.4)
$h^\star = h^\star_\pi$	best hypothesis (1)
R	proxy for the risk (4), (27)
$\hat{h}$	learned hypothesis (4)
P_h	projection function for comp.-type task (Def. 3.1)
$\mathcal{L} = \mathcal{L}(\mathcal{H})$	class of loss functions (9)
$\Delta\mathcal{L} = \Delta\mathcal{L}(\mathcal{H})$	class of loss differences (17)
$\mathcal{F} = \{\phi_\omega\}_{\omega\in\Omega}$	class of features (Def. 5.1)
Λ	probability distribution of feature parameters ω (Def. 5.1)
$\kappa(x, x')$	psd kernel (Def. 5.2)
$\kappa(\pi, \pi')$	kernel mean embedding (44)
$C_\mathcal{A}$,C_κ, $C_{\kappa\Phi}$	kernel constants (11); (47); (48)
$\|\mu\|_\mathcal{G}$	$\sup_{f\in\mathcal{G}} \lvert\langle \mu, f\rangle\rvert$ ((8), App. A.2)
$\|\mu\|_\kappa$	MMD norm (45), (58)
$\|\cdot\|_\mathcal{L}$, $\|\cdot\|_{\Delta\mathcal{L}}$	task-driven norms (9); (19)
$D_h(\pi\|\pi')$	excess-risk divergence (18)
$d_h(\pi, \mathfrak{S})$,	bias term wrt. model (25)
$d_\mathcal{F}(\pi, \pi')$	feature-based metric (54)
$\mathfrak{S}$	model set (of probabilities)
$\mathfrak{S}_h$	probabilities s.t. h optimal (26)
$\mathfrak{S}_h^{\mathtt{CT}}$, $\mathfrak{S}^{\mathtt{CT}}(\mathcal{H})$	compression-type model set (31)
$\mathfrak{S}_h^{\mathtt{ML}}$, $\mathfrak{S}^{\mathtt{ML}}(\mathcal{H})$	max. likelihood model set (28)
$\mathcal{S} = \mathcal{S}_\kappa(\mathfrak{S})$	normalized secant set (49)
$\|\mathcal{E}\|$	radius of a set of measures (51)
$c_\kappa(t)$	concentration function (52)
$\mathrm{N}(\|\cdot\|, A, \varepsilon)$	covering numbers (Def. 5.6)
$\mathcal{B}$	Ball (Def. A.2)

References

N. Ailon, R. Jaiswal, and C. Monteleoni. Streaming k -means approximation. *Advances in Neural Information Processing Systems (NIPS)*, pages 10–18, 2009.

D. Aloise, A. Deshpande, P. Hansen, and P. Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009. ISSN 08856125. doi: 10.1007/s10994-009-5103-0.

J. Anderson, M. Belkin, N. Goyal, L. Rademacher, and J. Voss. The more, the merrier: the blessing of dimensionality for learning large gaussian mixtures. In *Conference on Learning Theory*, pages 1135–1164, 2014.

C. Andrieu and A. Doucet. Online expectation-maximization type algorithms for parameter estimation in general state space models. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 6, pages VI–69, 2003.

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. doi: 10.1109/Allerton.2012.6483308.

D. Arthur and S. Vassilvitskii. k-means++: The Advantages of Careful Seeding. In *ACM-SIAM symposium on Discrete algorithms*, pages 1027—-1035, 2007a.

D. Arthur and S. Vassilvitskii. k-means++ - the advantages of careful seeding. *SODA*, 2007b.

F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017. URL `http://jmlr.org/papers/v18/15-178.html`.

A. Balsubramani, S. Dasgupta, and Y. 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. URL `http://papers.nips.cc/paper/5132-the-fast-convergence-of-incremental-pca.pdf`.

R. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4):118–121, 2007.

R. Baraniuk, M. Davenport, R. A. DeVore, and M. B. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28(3):253–263, 2008.

M. Belkin and K. Sinha. Polynomial learning of distribution families. In *IEEE 51st Annual Symposium on Foundations of Computer Science*. Ieee, 2010. ISBN 978-1-4244-8525-3. doi: 10.1109/FOCS.2010.16.

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, Feb. 2011.

A. Bietti and J. Mairal. On the Inductive Bias of Neural Tangent Kernels. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1–23, 2019. URL `http://arxiv.org/abs/1905.12173`.

M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. pages 1–30, 2018.

G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2-3):259–294, 2007.

A. Bourrier, M. 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. ISSN 0018-9448.

E. Candès, T. Strohmer, and V. Voroninski. PhaseLift: Exact and Stable Signal Recovery from Magnitude Measurements via Convex Programming. *Comm. Pure Appl. Math*, 66(8):1241–1274, 2013.

E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9-10):589–592, 2008.

E. J. Candès, J. Romberg, and T. Tao. Stable Signal Recovery from Incomplete and Inaccurate Measurements. *Comm. Pure Appl. Math*, 59:1207–1223, 2006.

O. Cappé and E. Moulines. Online EM Algorithm for Latent Data Models. *Journal of the Royal Statistical Society*, 71(3):593–613, 2009. ISSN 13697412. doi: 10.1111/j.1467-9868.2009.00698.x.

M. Carrasco and J.-P. Florens. Generalization of GMM to a continuum of moment conditions. *Econometric Theory*, 2000.

M. Carrasco and J.-P. Florens. Efficient GMM estimation using the empirical characteristic function. *IDEI Working Paper*, 140, 2002.

M. Carrasco and J.-P. Florens. On The Asymptotic Efficiency Of GMM. *Econometric Theory*, 30(02): 372–406, 2014. ISSN 0266-4666. doi: 10.1017/S0266466613000340.

A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc*, 22(1), 2009.

G. Cormode and M. Hadjieleftheriou. Methods for finding frequent items in data streams. *The VLDB Journal*, 19(1):3–20, 2009. ISSN 1066-8888. doi: 10.1007/s00778-009-0172-z. URL `http://link.springer.com/10.1007/s00778-009-0172-z`.

G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *J. Algorithms*, 55(1):58–75, 2005a.

G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005b. ISSN 01966774. doi: 10.1016/j.jalgor.2003.12.001. URL `http://linkinghub.elsevier.com/retrieve/pii/S0196677403001913`.

G. Cormode, M. Garofalakis, P. J. Haas, and C. Jermaine. Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches. *Foundations and Trends in Databases*, 4(xx):1–294, 2011. ISSN 1931-7883. doi: 10.1561/1900000004.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. Wiley-Interscience, 1991.

S. Dirksen. Dimensionality reduction with subgaussian matrices: a unified theory. *Foundations of Computational Mathematics*, 16(5):1367–1396, 2016.

D. L. Donoho. Compressed sensing. *IEEE Trans. Information Theory*, 52(4):1289–1306, 2006.

J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Privacy Aware Learning. *Journal of the ACM*, 61 (6), Nov. 2014.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

A. Eftekhari and M. B. Wakin. New analysis of manifold embeddings and signal recovery from compressive measurements. *Applied and Computational Harmonic Analysis*, 39(1):67–109, 2015.

J. B. Estrach, A. Szlam, and Y. LeCun. Signal recovery from Pooling Representations. *ICML*, 2014.

K. Fan. On a Theorem of Weyl Concerning Eigenvalues of Linear Transformations I. *Proc. Nat. Aca. Sci.*, 35(11):652–655, Nov. 1949.

D. Feldman and M. 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. ISSN 07378017. doi: 10.1145/1993636.1993712.

D. Feldman, M. Monemizadeh, C. Sohler, and D. P. Woodruff. Coresets and Sketches for High Dimensional Subspace Approximation Problems. 1:630–649, 2010.

D. Feldman, M. Faulkner, and A. Krause. Scalable Training of Mixture Models via Coresets. *Proceedings of Neural Information Processing Systems*, pages 1–9, 2011.

A. Feuerverger and R. A. Mureika. The Empirical Characteristic Function and Its Applications. *Annals of Statistics*, 5(1):88–97, Jan. 1977.

S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, May 2012.

G. Frahling and C. Sohler. A fast k -means implementation using coresets. *Proceedings of the twenty-second annual symposium on Computational geometry (SoCG)*, 18(6):605–625, 2005. ISSN 0218-1959. doi: 10.1142/S0218195908002787.

M. Gabrié, A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, and L. Zdeborová. Entropy and mutual information in models of deep neural networks. In *Advances in Neural Information and Processing Systems (NIPS)*, 2018. doi: 10.1016/0960-1686(92)90180-S. URL `http://arxiv.org/abs/1805.09785`.

M. R. Garey, D. S. Johnson, and H. S. Witsenhausen. The complexity of the generalized Lloyd - Max problem. *IEEE Trans. Inf. Theory*, 28(2):255–256, 1982.

M. Ghashami, D. Perry, and J. M. Phillips. Streaming Kernel Principal Component Analysis. *International Conference on Artificial Intelligence and Statistics*, 41:1–16, 2016. URL `http://arxiv.org/abs/1512.05059`.

A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. 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.

A. C. Gilbert, Y. Zhang, K. Lee, Y. Zhang, and H. Lee. Towards understanding the invertibility of convolutional neural networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI17, page 17031710. AAAI Press, 2017. ISBN 9780999241103.

R. Giryes, G. Sapiro, and A. M. Bronstein. Deep Neural Networks with Random Gaussian Weights - A Universal Classification Strategy? *IEEE Trans. Signal Processing*, 2016.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A Kernel Method for the Two-Sample Problem. In *Advances in Neural Information Processing Systems (NIPS)*, pages 513–520, 2007. ISBN 0-262-19568-2.

R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin. Statistical Learning Guarantees for Compressive Clustering and Compressive Mixture Modeling. 2020. URL https://hal.inria.fr/hal-02536818.

S. Guha and E. al. Clustering Data Streams. 2000.

A. R. Hall. *Generalized method of moments*. 2005. ISBN 0198775210.

P. R. Halmos. *Measure theory*, volume 18. Springer, 2013.

S. Har-Peled and S. 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. ISBN 1581138520. doi: 10.1145/1007352.1007400. URL http://ukpmc.ac.uk/abstract/CIT/664454.

D. Hsu and S. M. Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Conference on Innovations in Theoretical Computer Science*, 2013. URL http://dl.acm.org/citation.cfm?id=2422439.

A. Jacot, F. Gabriel, and C. Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. doi: arXiv:1806.07572v2. URL http://arxiv.org/abs/1806.07572.

M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege. Stable low-rank matrix recovery via null space properties. *Information and Inference*, 5(4):405–441, 2016.

N. Keriven, A. Bourrier, R. Gribonval, and P. Pérèz. Sketching for Large-Scale Learning of Mixture Models. In *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2015. ISBN 2011277906.

N. Keriven, A. Bourrier, R. Gribonval, and P. Pérèz. Sketching for Large-Scale Learning of Mixture Models. *arXiv preprint arXiv:1606.02838*, pages 1–50, 2016. ISSN 15206149. doi: 10.1109/ICASSP.2016.7472867.

H. J. Landau. *Moments in mathematics*. 1987. URL http://books.google.com/books?hl=en{&}lr={&}id=IEo2Iu{_}uogUC{&}oi=fnd{&}pg=PA1{&}dq=Moments+in+Mathematics{&}ots=GoVRyR9tmp{&}sig=5rD8iF93S5{_}rgXR7JiNCjdI{_}5-E.

C. Levrard. Fast rates for empirical vector quantization. *Electronic Journal of Statistics*, 7(0):1716–1746, 2013.

Y. Li, K. Swersky, and R. Zemel. Generative Moment Matching Networks. *Proceedings of The 32nd International Conference on Machine Learning*, 37:1718–1727, 2015.

M. Lucic, M. Faulkner, A. Krause, and D. Feldman. Training Mixture Models at Scale via Coresets. 2017. URL https://arxiv.org/pdf/1703.08110.pdf.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online Learning for Matrix Factorization and Sparse Coding. *Journal of Machine Learning Research*, 11(1):19–60, Jan. 2010.

P. Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, 2007.

W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. 1994. ISBN 9780444887665. doi: 10.1016/S1573-4412(05)80005-4.

I. Pinelis. An approach to inequalities for the distributions of infinite-dimensional martingales. In R. Dudley, M. Hahn, and J. Kuelbs, editors, *Probability in Banach Spaces, 8, Proceedings of the 8th International Conference*, volume 30, pages 128–134. Birkäuser, 1992.

G. Puy, M. E. Davies, and R. Gribonval. Recipes for Stable Linear Embeddings From Hilbert Spaces to $\mathbb{R}^m$. *IEEE Trans. Information Theory*, 63(4):2171–2187, 2017.

A. Rahimi and B. Recht. Random Features for Large Scale Kernel Machines. *Advances in Neural Information Processing Systems (NIPS)*, (1):1–8, 2007.

A. Rahimi and B. 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. URL `http://papers.nips.cc/paper/3495-weighted-sums-of-random-kitchen-sinks-replacing-minimization-with-randomization-in-learning`.

M. Reiß and M. Wahl. Non-asymptotic upper bounds for the reconstruction error of pca. *arXiv preprint arXiv:1609.03779*, 2016.

A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.

V. Schellekens, A. Chatalic, F. Houssiau, Y.-A. De Montjoye, L. Jacques, and R. Gribonval. Differentially Private Compressive k-Means. In *ICASSP 2019 - 44th International Conference on Acoustics, Speech, and Signal Processing*, pages 7933–7937, Brighton, United Kingdom, May 2019. IEEE. doi: 10.1109/ICASSP.2019.8682829. URL `https://hal.inria.fr/hal-02060208`.

J. Shawe-Taylor, C. K. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the gram matrix and the generalization error of kernel-pca. *IEEE Transactions on Information Theory*, 51 (7):2510–2522, 2005.

R. Shwartz-Ziv and N. Tishby. Opening the Black Box of Deep Neural Networks via Information. (Dl), 2017. URL `https://arxiv.org/pdf/1703.00810.pdf`.

A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert Space Embedding for Distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31, 2007. ISBN 978-3-540-75224-0. doi: 10.1007/978-3-540-75225-7_5. URL `http://eprints.pascal-network.org/archive/00003987/`.

B. K. Sriperumbudur and Z. Szabó. Optimal Rates for Random Fourier Features. *NIPS*, 2015.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11: 1517–1561, 2010. URL `http://dl.acm.org/citation.cfm?id=1859901`.

N. Thaper, S. Guha, P. Indyk, and N. 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.