TOWARD A UNIFIED THEORY OF SPARSE DIMENSIONALITY REDUCTION IN EUCLIDEAN SPACE

JEAN BOURGAIN, SJOERD DIRKSEN, AND JELANI NELSON

Abstract. Let $\Phi \in \mathbb{R}^{m\times n}$ be a sparse Johnson-Lindenstrauss transform [KN14] with s non-zeroes per column. For a subset T of the unit sphere, $\varepsilon \in (0, 1/2)$ given, we study settings for m, s required to ensure

$$\mathop{\mathbb{E}}_{\Phi} \sup_{x\in T} \left| \|\Phi x\|_2^2 - 1 \right| < \varepsilon,$$

i.e. so that Φ preserves the norm of every $x \in T$ simultaneously and multiplicatively up to $1 + \varepsilon$. We introduce a new complexity parameter, which depends on the geometry of T, and show that it suffices to choose s and m such that this parameter is small. Our result is a sparse analog of Gordon's theorem, which was concerned with a dense Φ having i.i.d. Gaussian entries. We qualitatively unify several results related to the Johnson-Lindenstrauss lemma, subspace embeddings, and Fourier-based restricted isometries. Our work also implies new results in using the sparse Johnson-Lindenstrauss transform in numerical linear algebra, classical and model-based compressed sensing, manifold learning, and constrained least squares problems such as the Lasso.

Contents

J.B. partially supported by NSF grant DMS-1301619.

S.D. partially supported by SFB grant 1060 of the Deutsche Forschungsgemeinschaft (DFG).

J.N. supported by NSF grant IIS-1447471 and CAREER award CCF-1350670, ONR grant N00014-14-1-0632, and a Google Faculty Research Award. Part of this work done while supported by NSF grants CCF-0832797 and DMS-1128155.

1. Introduction

Dimensionality reduction is a ubiquitous tool across a wide array of disciplines: machine learning [WDL$^+$09], high-dimensional computational geometry [Ind01], privacy [BBDS12], compressed sensing [CT05], spectral graph theory [SS11], interior point methods for linear programming [LS13], numerical linear algebra [Sar06], computational learning theory [BB05, BBV06], manifold learning [HWB07, Cla08], motif-finding in computational biology [BT02], astronomy [CM12], and several others. Across all these disciplines one is typically faced with data that is not only massive, but each data item itself is represented as a very high-dimensional vector. For example, when learning spam classifiers a data point is an email, and it is represented as a high-dimensional vector indexed by dictionary words [WDL$^+$09]. In astronomy a data point could be a star, represented as a vector of light intensities measured over various points sampled in time [KZM02, VJ11]. Dimensionality reduction techniques in such applications provide the following benefits:

- Smaller storage consumption.
- Speedup during data analysis.
- Cheaper signal acquisition.
- Cheaper transmission of data across computing clusters.

The technical guarantees required from a dimensionality reduction routine are application-specific, but typically such methods must reduce dimension while still preserving point geometry, e.g. inter-point distances and angles. That is, one has some point set $X \subset \mathbb{R}^n$ with n very large, and we would like a dimensionality-reducing map $f : X \to \mathbb{R}^m$, $m \ll n$, such that

$$\forall x, y \in X,\ (1-\varepsilon)\|x-y\| \le \|f(x)-f(y)\| \le (1+\varepsilon)\|x-y\| \tag{1.1}$$

for some norm $\|\cdot\|$. Note also that for unit vectors x, y, $\cos(\angle(x,y)) = (1/2)(\|x\|_2^2 + \|y\|_2^2 - \|x-y\|_2^2)$, and thus f also preserves angles with additive error if it preserves Euclidean norms of points in $X \cup (X - X)$.

A powerful tool for achieving Eq. (1.1), used in nearly all the applications cited above, is the *Johnson-Lindenstrauss (JL) lemma* [JL84].

Theorem 1 (JL lemma)**.** *For any subset X of Euclidean space and $0 < \varepsilon < 1/2$, there exists $f : X \to \ell_2^m$ with $m = O(\varepsilon^{-2}\log|X|)$ providing Eq. (1.1) for $\|\cdot\| = \|\cdot\|_2$.*

This bound on m is nearly tight: for any $n \ge 1$ Alon exhibited a point set $X \subset \ell_2^n$, $|X| = n+1$, such that any such JL map f must have $m = \Omega(\varepsilon^{-2}(\log n)/\log(1/\varepsilon))$ [Alo03]. In fact, all known proofs of the JL lemma provide linear f, and the JL lemma is tight up to a constant factor in m when f must be linear [LN14]. Unfortunately, for actual applications such *worst-case* understanding is unsatisfying. Rather we could ask: if given a distortion parameter ε and point set X as input (or a succinct description of it if X is large or even infinite, as in some applications), what is the best target dimension $m = m(X, \varepsilon)$ such that a JL map exists for X with this particular ε? That is, in practice we are more interested in moving beyond worst case analysis and being as efficient as possible *for our particular data X*.

Unfortunately the previous question seems fairly difficult. For the related question of computing the optimal distortion for embedding X into a line (i.e. $m = 1$), it is computationally hard to approximate the optimal distortion even up to a multiplicative factor polynomial in $|X|$ [BCIS05]. In practice, however, typically f cannot be chosen arbitrarily as a function of X anyway. For example, when employing certain learning algorithms such as stochastic gradient descent on dimensionality-reduced data, it is at least required that f is differentiable on $\mathbb{R}^n$ (where $X \subset \mathbb{R}^n$) [WDL$^+$09]. For several applications it is also crucial that f is linear, e.g. in numerical linear algebra [Sar06] and compressed sensing [CT05, Don06]. In one-pass streaming applications [CW09] and data structural problems such as nearest neighbor search [HIM12], it is further required that f is chosen randomly without knowing X. For any particular X, a random f drawn from some distribution must satisfy the JL guarantee with good probability. In streaming applications this is because X is not fully known up front, but is gradually observed in a stream. In data structure applications this is because f must preserve distances to some future query points, which are not known at the time the data structure is constructed.

Due to the considerations discussed, in practice typically f is chosen as a random linear map drawn from some distribution with a small number of parameters (in some cases simply the parameter m). For example, popular choices of f include a random matrix with independent Gaussian [HIM12] or Rademacher [Ach03] entries. While worst case bounds inform us how to set parameters to obtain the JL guarantee for worst case X, we typically can obtain better parameters by exploiting prior knowledge about X. Henceforth we only discuss linear f, so we write $f(x) = \Phi x$ for $\Phi \in \mathbb{R}^{m\times n}$. Furthermore by linearity, rather than preserving Euclidean distances in X it is equivalent to discuss preserving norms of all vectors in $T = \{(x-y)/\|x-y\|_2 : x, y \in X\} \subset S^{n-1}$, the set of all normalized difference vectors amongst points in X. Thus up to changing ε by roughly a factor of 2, Eq. (1.1) is equivalent to

$$\sup_{x\in T} \left| \|\Phi x\|^2 - 1 \right| < \varepsilon. \tag{1.2}$$

Furthermore, since we consider Φ chosen at random, we more specifically want

$$\mathop{\mathbb{E}}_{\Phi} \sup_{x\in T} \left| \|\Phi x\|^2 - 1 \right| < \varepsilon. \tag{1.3}$$

Instance-wise understanding for achieving Eq. (1.3) was first provided by Gordon [Gor88], who proved that a random Φ with i.i.d. Gaussian entries satisfies Eq. (1.3) as long as $m \gtrsim (g^2(T) + 1)/\varepsilon^2$, where we write $A \gtrsim B$ if $A \geq CB$ for a universal constant $C > 0$. Denoting by g a standard n-dimensional Gaussian vector, the parameter $g(T)$ is defined as the *Gaussian mean width*

$$g(T) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{g} \sup_{x\in T} \langle g, x\rangle.$$

One can think of $g(T)$ as describing the ℓ_2-geometric complexity of T. It is always true that $g^2(T) \lesssim \log|T|$, and thus Gordon's theorem implies the JL lemma. In fact for all T we know from applications, such as for the restricted isometry property from compressed sensing [CT05] or subspace embeddings from numerical linear algebra [Sar06], the best bound on m is a corollary of Gordon's theorem. Later works extended Gordon's theorem to other distributions for Φ, such as Φ having independent subgaussian entries (e.g. Rademachers) [KM05, MPTJ07, Dir14].

Although Gordon's theorem gives a good understanding for m in most scenarios, it suffers from the fact that it analyzes a *dense* random Φ, which means that performing the dimensionality reduction $x \mapsto \Phi x$ is dense matrix-vector multiplication, and is thus slow. For example, in some numerical linear algebra applications (such as least squares regression [Sar06]), multiplying a dense unstructured Φ with the input turns out to be slower than obtaining the solution of the original, high-dimensional problem! In compressed sensing, certain iterative recovery algorithms such as CoSamp [NT09] and Iterative Hard Thresholding [BD08] involve repeated multiplications by Φ and Φ^*, the conjugate transpose of Φ, and thus Φ supporting fast matrix-vector multiply are desirable in such applications as well.

The first work to provide Φ with small m supporting faster multiplication is the Fast Johnson-Lindenstrauss Transform (FJLT) of [AC09] for finite T. The value of m was still $O(\varepsilon^{-2}\log|T|)$, with the time to multiply Φx being $O(n\log n + m^3)$. [AL09] later gave an improved construction with time $O(n\log n + m^{2+\gamma})$ for any small constant $\gamma > 0$. Most recently several works gave nearly linear embedding time in n, independent of m, at the expense of increasing m by a $(\log n)^c$ factor [AL13, KW11, NPW14]. In all these works Φ is the product of some number of very sparse matrices and Fourier matrices, with the speed coming from the Fast Fourier Transform (FFT) [CT65]. It is also known that this FFT-based approach can be used to obtain fast RIP matrices for compressed sensing [CT06, RV08, CGV13] and fast oblivious subspace embeddings for numerical linear algebra applications [Sar06] (see also [Tro11, LDFU13] for refined analyses in the latter case).

Another line of work, initiated in [Ach03] and greatly advanced in [DKS10], sought fast embedding time by making Φ sparse. If Φ is drawn from a distribution over matrices having at most s non-zeroes per column, then Φx can be computed in time $s \cdot \|x\|_0$. After some initial improvements [KN10, BOR10], the best known achievable value of s to date for the JL lemma while still maintaining $m \lesssim \varepsilon^{-2} \log |T|$ is the sparse Johnson-Lindenstrauss Transform (SJLT) of [KN14], achieving $s \lesssim \varepsilon^{-1} \log |T| \lesssim \varepsilon m$. Furthermore, an example of a set T exists which requires this bound on s up to $O(\log(1/\varepsilon))$ for any linear JL map [NN13b]. Note however that, again, this is an understanding of the *worst-case* parameter settings over all T.

In summary, while Gordon's theorem gives us a good understanding of instance-wise bounds on T for achieving good dimensionality reduction, it only does so for dense, slow Φ. Meanwhile, our understanding for efficient Φ, such as the SJLT with small s, has not moved beyond the worst case. In some very specific examples of T we do have good bounds for settings of s, m that suffice, such as T the unit norm vectors in a d-dimensional subspace [CW13, MM13, NN13a], or all elements of T having small ℓ_∞ norm [Mat08, DKS10, KN10, BOR10]. However, our understanding for general T is non-existent. This brings us to the main question addressed in this work, where S^{n-1} denotes the ℓ_2-unit sphere $\{x \in \mathbb{R}^n : \|x\|_2 = 1\}$.

Question 2. Let $T \subseteq S^{n-1}$ and Φ be the SJLT. What relationship must s, m satisfy, in terms of the geometry of T, to ensure (1.3)?

We also note that while the FFT-based and sparse Φ approaches seem orthogonal at first glance, the two are actually connected, as pointed out before [AC09, Mat08, NPW14]. The FJLT sets $\Phi = SP$ where P is some random preconditioning matrix that makes T "nice" with high probability, and S is a random sparse matrix. We point out that although the SJLT is not the same as the matrix S typically used in the FFT-based literature, one could replace S with the SJLT and hope for similar algorithmic outcome if s is small: nearly linear embedding time.

The answer to the analog of Question 2 for a standard Gaussian matrix depends only on the ℓ_2-metric structure of T. Indeed, since both ℓ_2-distances and Gaussian matrices are invariant under orthogonal transformations, so is (1.3) in this case. This is reflected in Gordon's theorem, where the embedding dimension m is governed by the Gaussian width, which is invariant under orthogonal transformations. We stress that in sharp contrast, a resolution of Question 2 cannot solely depend on the ℓ_2-metric structure of T. Indeed, we require that Φ be sparse *in a particular basis* and is therefore not invariant under orthogonal transformations. Thus an answer to Question 2 must be more nuanced (see our main theorem, Theorem 3).

Our Main Contribution: We provide a general theorem which answers Question 2. Specifically, for every $T \subseteq S^{n-1}$ analyzed in previous work that we apply our general theorem to here, we either (1) qualitatively recover or improve the previous best known result, or (2) prove the first non-trivial result for dimensionality reduction with sparse Φ. We say "qualitatively" since applying our general theorem to these applications loses a factor of $\log^c(n/\varepsilon)$ in m and $\log^c(n/\varepsilon)/\varepsilon$ in s.

In particular for (2), our work is the first to imply that non-trivially sparse dimensionality reducing linear maps can be applied for gain in model-based compressed sensing [BCDH10], manifold learning [TdSL00, DG13], and constrained least squares problems such as the popular Lasso [Tib96].

Specifically, we prove the following theorem.

Theorem 3 (Main Theorem). *Let $T \subset S^{n-1}$ and Φ be an SJLT with column sparsity s. Define the complexity parameter*

$$\kappa(T) \stackrel{\text{def}}{=} \kappa_{s,m}(T) = \max_{q \le \frac{m}{s}\log s} \Big\{ \frac{1}{\sqrt{qs}} \Big(\mathop{\mathbb{E}}_{\eta} \Big(\mathop{\mathbb{E}}_{g} \sup_{x \in T} |\sum_{j=1}^{n} \eta_j g_j x_j| \Big)^q \Big)^{1/q} \Big\},$$

where (g_j) are i.i.d. standard Gaussian and (η_j) i.i.d. Bernoulli with mean $qs/(m \log s)$. If

$$m \gtrsim (\log m)^3 (\log n)^5 \cdot \frac{(g^2(T)+1)}{\varepsilon^2}$$
$$s \gtrsim (\log m)^6 (\log n)^4 \cdot \frac{1}{\varepsilon^2}.$$

Then (1.3) *holds as long as s, m furthermore satisfy the condition*

$$(\log m)^2 (\log n)^{5/2} \kappa(T) < \varepsilon. \tag{1.4}$$

The complexity parameter $\kappa(T)$ may seem daunting at first, but Section 8 shows it can be controlled quite easily for all the T we have come across in applications.

1.1. **Applications.** Here we describe various T and their importance in certain applications. Then we state the consequences that arise from our theorem. In order to highlight the qualitative understanding arising from our work, we introduce the notation $A \lesssim_* B$ if $A \le B \cdot \varepsilon^{-2} (\log n)^c$. A summary of our bounds is in Figure 1.

Finite $|T|$: This is the setting $|T| < \infty$, for which the SJLT satisfies Eq. (1.3) with $s \lesssim \varepsilon^{-1} \log |T|$, $m \lesssim \varepsilon^{-2} \log |T|$ [KN14]. If also $T \subset B_{\ell_\infty^n}(\alpha)$, i.e. $\|x\|_\infty \le \alpha$ for all $x \in T$, [Mat08] showed it is possible to achieve $m \lesssim \varepsilon^{-2} \log |T|$ with a Φ that has an *expected* $O(\varepsilon^{-2}(\alpha \log |T|)^2)$ non-zeroes per column.

Our theorem implies $s, m \lesssim_* \log |T|$ suffices in general, and $s \lesssim_* 1 + (\alpha \log |T|)^2, m \lesssim_* \log |T|$ in the latter case, qualitatively matching the above.

Linear subspace: Here $T = \{x \in E : \|x\|_2 = 1\}$ for some d-dimensional linear subspace $E \subset \mathbb{R}^n$, for which achieving Eq. (1.3) with $m \lesssim d/\varepsilon^2$ is possible [AHK06, CW13]. A distribution satisfying Eq. (1.3) for any d-dimensional subspace E is known as an *oblivious subspace embedding (OSE)*. The paper [Sar06] pioneered the use of OSE's for speeding up approximate algorithms for numerical linear algebra problems such as low-rank approximation and least-squares regression. More applications have since been found to approximating leverage scores [DMIMW12], k-means clustering [BZMD11, CEM$^+$14], canonical correlation analysis [ABTZ13], support vector machines [PBMID13], ℓ_p-regression [CDMI$^+$13, WZ13], ridge regression [LDFU13], streaming approximation of eigenvalues [AN13], and speeding up interior point methods for linear programming [LS13]. In many of these applications there is some input $A \in \mathbb{R}^{n \times d}$, $n \gg d$, and the subspace E is for example the column space of A. Often an exact solution requires computing the singular value decomposition (SVD) of A, but using OSE's the running time is reduced to that for computing ΦA, plus computing the SVD of the smaller matrix ΦA. The work [CW13] showed $s = 1$ with small m is sufficient, yielding algorithms for least squares regression and low-rank approximation whose running times are linear in the number of non-zero entries in A for sufficiently lopsided rectangular matrices.

Our theorem implies that $s \lesssim_* 1$ and $m \lesssim_* d$ suffices to satisfy Eq. (1.3), which is qualitatively correct. Furthermore, a subset of our techniques reveals that if the maximum incoherence $\max_{1\le i\le n} \|P_E e_i\|_2$ is at most $poly(\varepsilon/\log n)$, then $s = 1$ suffices (Theorem 5). This was not known in previous work. A random d-dimensional subspace has incoherence $\sqrt{d/n}$ w.h.p. for $d \gtrsim \log n$ by the JL lemma, and thus is very incoherent if $n \gg d$.

Closed convex cones: For $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, and $\mathcal{C} \subseteq \mathbb{R}^d$ a closed convex set, consider the constrained least squares problem of minimizing $\|Ax - b\|_2^2$ subject to $x \in \mathcal{C}$. A popular choice is the Lasso [Tib96], in which the constraint set $\mathcal{C} = \{x \in \mathbb{R}^d : \|x\|_1 \le R\}$ encourages sparsity of x. Let x_* be an optimal solution, and let $T_{\mathcal{C}}(x)$ be the tangent cone of $\mathcal{C}$ at some point $x \in \mathbb{R}^d$ (see Appendix B for a definition). Suppose we wish to accelerate approximately solving the constrained least squares problem by instead computing a minimizer $\hat{x}$ of $\|\Phi Ax - \Phi b\|_2^2$ subject to $x \in \mathcal{C}$. The work [PW14] showed that to guarantee $\|A\hat{x} - b\|_2^2 \le (1+\varepsilon)\|Ax_* - b\|_2^2$, it suffices that Φ satisfy two conditions, one of which is Eq. (1.2) for $T = AT_{\mathcal{C}}(x_*) \cap S^{n-1}$. The paper [PW14] then analyzed dense random matrices for sketching constrained least squares problems. For example, for the Lasso if we are promised that the optimal solution x_* is k-sparse, [PW14] shows that it suffices to set

$$m \gtrsim \frac{1}{\varepsilon^2} \max_{j=1,\ldots,d} \frac{\|A_j\|_2^2}{\sigma_{\min,k}^2} k \log d,$$

where A_j is the jth column of A and $\sigma_{\min,k}$ the smallest ℓ_1-restricted eigenvalue of A:

$$\sigma_{\min,k} = \inf_{\|y\|_2=1,\ \|y\|_1\le 2\sqrt{k}} \|Ay\|_2.$$

Our work also applies to such T (and we further show the SJLT with small s, m satisfies the second condition required for approximate constrained least squares; see Theorem 16). For example for the Lasso, we show that again it suffices that

$$m \gtrsim_* \max_{j=1,\ldots,d} \frac{\|A_j\|_2^2}{\sigma_{\min,k}^2} k \log d$$

but where we only require from s that

$$s \gtrsim_* \max_{\substack{1\le i\le n\\ 1\le j\le d}} \frac{A_{ij}^2}{\sigma_{\min,k}^2} k$$

That is, the sparsity of Φ need only depend on the largest entry in A as opposed to the largest column norm in A. The former can be much smaller.

Unions of subspaces: Define $T = \cup_{\theta\in\Theta} E_\theta \cap S^{n-1}$, where Θ is some index set and each $E_\theta \subset \mathbb{R}^n$ is a d-dimensional linear subspace. A case of particular interest is when $\theta \in \Theta$ ranges over all k-subsets of $\{1, \ldots, n\}$, and E_θ is the subspace spanned by $\{e_j\}_{j\in\theta}$ (so $d = k$). Then T is simply the set of all k-sparse unit vectors of unit Euclidean norm: $S_{n,k} \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : \|x\|_2 = 1, \|x\|_0 \le k\}$ for $\|\cdot\|_0$ denoting support size. Φ satisfying (1.2) is then referred to as having the *restricted isometry property (RIP) of order* k with restricted isometry constant $\varepsilon_k = \varepsilon$ [CT05]. Such Φ are known to exist with $m \lesssim \varepsilon_k^{-2} k \log(n/k)$, and furthermore it is known that $\varepsilon_{2k} < \sqrt{2} - 1$ implies that any (approximately) k-sparse $x \in \mathbb{R}^n$ can be (approximately) recovered from Φx in polynomial time by solving a certain linear program [CT05, Can08]. Unfortunately it is known for $\varepsilon = \Theta(1)$ that *any* RIP matrix Φ with such small m must have $s \gtrsim m$ [NN13b]. Related is the case of vectors sparse in some other basis, i.e. $T = \{Dx \in \mathbb{R}^n : \|Dx\|_2 = 1, \|x\|_0 \le k\}$ for some so-called "dictionary" D (i.e. the subspaces are acted on by D), or when T only allows for some subset of all $\binom{n}{k}$ sparsity patterns in *model-based* compressed sensing [BCDH10] (so that $|\Theta| < \binom{n}{k}$).

Our main theorem also implies RIP matrices with $s, m \lesssim_* k\log(n/k)$. More generally when a dictionary D is involved such that the subspaces $\mathrm{span}(\{De_j\}_{j\in\theta})$ are all α-incoherent (as defined above), the sparsity can be further improved to $s \lesssim_* 1 + (\alpha k\log(n/k))^2$. That is, to satisfy the RIP with dictionaries yielding incoherent subspaces, we can keep m qualitatively the same while making s much smaller. For the general problem of unions of d-dimensional subspaces, our theorem implies one can either set $m \lesssim_* d + \log|\Theta|, s \lesssim_* \log|\Theta|$ or $m \lesssim_* d + \log|\Theta|, s \lesssim_* 1 + (\alpha\log|\Theta|)^2$. Previous work required m to depend on the *product* of d and $(\log|\Theta|)^c$ instead of the *sum* [NN13a], including a nice recent improvement by Cohen [Coh14], and is thus unsuitable for the application to the set of k-sparse vectors (RIP matrices with $\lesssim_* k^2$ rows are already attainable via simpler methods using incoherence; e.g. see [BDF+11, Proposition 1]). Iterative recovery algorithms such as CoSamp can also be used in model-based sparse recovery [BCDH10], which again involves multiplications by Φ, Φ^*, and thus sparse Φ is relevant for faster recovery. Our theorem thus shows, for the first time, that the benefit of model-based sparse recovery is not just a smaller m, but rather that the measurement matrix Φ can be made much sparser if the model is simple (i.e. $|\Theta|$ is small). For example, in the *block-sparse model* one wants to (approximately) recover a signal $x \in \mathbb{R}^n$ based on m linear measurements, where x is (approximately) k-block-sparse. That is, the n coordinates are partitioned into n/b blocks of size b each, and each block is either "on" (at least one coordinate in that block non-zero), or "off" (all non-zero). A k-block-sparse signal has at most k/b blocks on (thus $\|x\|_0 \le k$). Thus $s \lesssim_* \log|\Theta| = \log(\binom{n/b}{k/b}) \lesssim (k/b)\log(n/k)$. Then as long as $b = \omega(\log(n/k))$, our results imply non-trivial column-sparsity $s \ll m$. Ours is the first result yielding non-trivial sparsity in a model-RIP Φ for any model with a number of measurements qualitatively matching the optimal bound (which is on the order of $m \lesssim k + (k/b)\log(n/k)$ [ADR14]). We remark that for model-based RIP_1, where one wants to approximately preserve ℓ_1-norms of k-block-sparse vectors, which is useful for ℓ_1/ℓ_1 recovery, [IR13] have shown a much better sparsity bound of $O(\lceil\log_b(n/k)\rceil)$ non-zeroes per column in their measurement matrix. However, they have also shown that any model-based RIP_1 matrix for block-sparse signals must satisfy the higher lower bound of $m \gtrsim k + (k/\log b)\log(n/k)$ (which is tight).

Previous work also considered $T = HS_{n,k}$, where H is any bounded orthonormal system, i.e. $H \in \mathbb{R}^{n\times n}$ is orthogonal and $\max_{i,j}|H_{ij}| = O(1/\sqrt{n})$ (e.g. the Fourier matrix). Work of [CT06, RV08, CGV13] shows Φ can then be a sampling matrix (one non-zero per row) with $m \lesssim \varepsilon^{-2}k\log(n)(\log k)^3$. Since randomly flipping the signs of every column in an RIP matrix yields JL [KW11], this also gives a good implementation of an FJLT. Our theorem recovers a similar statement, but using the SJLT instead of a sampling matrix, where we show $m \lesssim_* k$ and $s \lesssim_* 1$ suffice for orthogonal H satisfying the weaker requirement $\max_{i,j}|H_{ij}| = O(1/\sqrt{k})$.

Smooth manifolds: Suppose we are given several images of a human face, but with varying lighting and angle of rotation. Or perhaps the input is many sample handwritten images of letters. In these examples, although input data is high-dimensional (n is the number of pixels), we imagine all possible inputs come from a set of low intrinsic dimension. That is, they lie on a d-dimensional manifold $\mathcal{M} \subset \mathbb{R}^n$ where $d \ll n$. The goal is then, given a large number of manifold examples, to learn the parameters of $\mathcal{M}$ to allow for nonlinear dimensionality reduction (reducing just to the few parameters of interest). This idea, and the first successful algorithm (ISOMAP) to learn a manifold from sampled points is due to [TdSL00]. Going back to the example of human faces, [TdSL00] shows that different images of a human face can be well-represented by a 3-dimensional manifold, where the parameters are brightness, left-right angle of rotation, and up-down angle of rotation. Since [TdSL00], several more algorithms have been developed to handle more general classes of manifolds than ISOMAP [RS00, DG13].

Baraniuk and Wakin [BW09] proposed using dimensionality-reducing maps to first map $\mathcal{M}$ to $\Phi\mathcal{M}$ and then learn the parameters of interest in the reduced space (for improved speed). Sharper analyses were later given in [Cla08, EW13, Dir14]. Of interest are both that (1) any C^1 curve in $\mathcal{M}$ should have length approximately preserved in $\Phi\mathcal{M}$, and (2) Φ should be a *manifold embedding*, in the sense that all C^1 curves $\gamma' \in \Phi\mathcal{M}$ should satisfy that the preimage of γ' (in $\mathcal{M}$) is a C^1 curve in $\mathcal{M}$. Then by (1) and (2), *geodesic* distances are preserved by Φ.

To be concrete, let $\mathcal{M} \subset \mathbb{R}^n$ be a d-dimensional manifold obtained as the image $\mathcal{M} = F(B_{\ell_2^d})$, for smooth $F : B_{\ell_2^d} \to \mathbb{R}^n$ (B_X is the unit ball of X). We assume $\|F(x)-F(y)\|_2 \simeq \|x-y\|_2$ (where $A \simeq B$ denotes that both $A \lesssim B$ and $A \gtrsim B$), and that the map sending $x \in \mathcal{M}$ to the tangent plane at x, E_x, is Lipschitz from $d_{\mathcal{M}}$ to ρ_{Fin}. Here $\rho_{\mathcal{M}}$ is geodesic distance on $\mathcal{M}$, and $\rho_{\mathrm{Fin}}(E_x, E_y) = \|P_{E_x} - P_{E_y}\|_{\ell_2^n \to \ell_2^n}$ is the *Finsler distance*, where P_E is the orthogonal projection onto E.

We want Φ satisfying $(1-\varepsilon)|\gamma| \le |\Phi(\gamma)| \le (1+\varepsilon)|\gamma|$ for all C^1 curves $\gamma \subset \mathcal{M}$. Here $|\cdot|$ is curve length. To obtain this, it suffices that Φ satisfy Eq. (1.2) for $T = \bigcup_{x\in\mathcal{M}} E_x \cap S^{n-1}$ [Dir14], an infinite union of subspaces. Using this observation, [Dir14] showed this property is satisfied for $s = m \lesssim d/\varepsilon^2$ with a *dense* matrix of subgaussian entries. For F as given above, preservation of geodesic distances is also satisfied for this m.

Our main theorem implies that to preserve curve lengths one can set $m \lesssim_* d$ for $s \lesssim_* 1 + (\alpha d)^2$, where α is the largest incoherence of any tangent space E_x for $x \in \mathcal{M}$. That is, non-trivial sparsity with $m \lesssim_* d$ is possible for any $\alpha \ll 1/\sqrt{d}$. Furthermore, we show that this is *optimal* by constructing a manifold with maximum incoherence of a tangent space approximately $1/\sqrt{d}$ such that *any* linear map Φ preserving curve lengths with $m \gtrsim_* d$ must have $s \gtrsim_* d$ (see Remark 20). We also show that Φ is a manifold embedding with large probability if the weaker condition $m \gtrsim_* d, s \gtrsim_* 1$ is satisfied. Combining these observations, we see that for the SJLT to preserve geodesic distances it suffices to set $m \lesssim_* d$ and $s \lesssim_* 1 + (\alpha d)^2$.

As seen above, not only does our answer to Question 2 qualitatively explain all known results, but it gives new results not known before with implications in numerical linear algebra, compressed sensing (model-based, and with incoherent dictionaries), constrained least squares, and manifold learning. We also believe it is possible for future work to sharpen our analyses to give asymptotically correct parameters for all the applications; see the discussion in Section 9.

We now end the introduction with an outline for the remainder. Section 2 defines the notation for the rest of the paper. Section 3 provides an overview of the proof of our main theorem, Theorem 3. Section 4 is a warmup that applies a subset of the proof ideas for Theorem 3 to the special case where $T = E \cap S^{n-1}$, E a linear subspace of dimension d. In fact the proof of our main theorem reduces the case of general T to several linear subspaces of varying dimensions, and thus this case serves as a useful warmup. Section 5 applies a different subset of our proof ideas to the special case where the norm $|||\cdot|||_T$ defined by $|||y|||_T = \sup_{x\in T} |\langle x, y\rangle|$ has a small type-2 constant, which is relevant for analyzing the FFT-based approaches to RIP matrices. Section 6 shows how similar ideas can be applied to constrained least squares problems, such as Lasso. Section 7 states and proves our most general theorem for arbitrary T. Section 8 shows how to apply Theorem 3 to obtain good

set T to preserve	**our m**	**our s**	**previous m**	**previous s**	**ref**
$\lvert T\rvert < \infty$	$\log \lvert T\rvert$	$\log \lvert T\rvert$	$\log \lvert T\rvert$	$\log \lvert T\rvert$	[JL84]
$\lvert T\rvert < \infty, \forall x \in T \lVert x\rVert_\infty \le \alpha$	$\log \lvert T\rvert$	$\lceil \alpha \log \lvert T\rvert \rceil^2$	$\log \lvert T\rvert$	$\lceil \alpha \log \lvert T\rvert \rceil^2$	[Mat08]
$E, \dim(E) \le d$	d	1	d	1	[NN13a]
$S_{n,k}$	$k\log(n/k)$	$k\log(n/k)$	$k\log(n/k)$	$k\log(n/k)$	[CT05]
$HS_{n,k}$	$k\log(n/k)$	1	$k\log(n/k)$	1	[RV08]
tangent cone for Lasso	$\max_j \frac{\lVert A_j\rVert_2^2}{\sigma_{min,k}^2} k$	$\max_{i,j} \frac{A_{ij}^2}{\sigma_{min,k}^2} k$	same as here	$s = m$	[PW14]
$\lvert\Theta\rvert < \infty$ $\forall E \in \Theta, \dim(E) \le d$	$d + \log\lvert\Theta\rvert$	$\log\lvert\Theta\rvert$	$d \cdot (\log\lvert\Theta\rvert)^6$	$(\log\lvert\Theta\rvert)^3$	[NN13a]
$\lvert\Theta\rvert < \infty$ $\forall E \in \Theta, \dim(E) \le d$ $\max_{\substack{1\le j\le n\\ E\in\Theta}} \lVert P_E e_j\rVert_2 \le \alpha$	$d + \log\lvert\Theta\rvert$	$\lceil \alpha \log\lvert\Theta\rvert \rceil^2$	—	—	—
$\lvert\Theta\rvert$ infinite $\forall E \in \Theta, \dim(E) \le d$	see appendix	see appendix (non-trivial)	similar to this work	m	[Dir14]
$\mathcal{M}$ a smooth manifold	d	$1 + (\alpha d)^2$	d	d	[Dir14]

FIGURE 1. The m, s that suffice when using the SJLT with various T as a consequence of our main theorem, compared with the best known bounds from previous work. All bounds shown hide $\mathrm{poly}(\varepsilon^{-1}\log n)$ factors. One row is blank in previous work due to no non-trivial results being previously known. For the case of the Lasso, we assume k is the sparsity of the true optimal solution.

bounds for various T, albeit losing a $\log^c(n/\varepsilon)$ factor in m and a $\log^c(n/\varepsilon)/\varepsilon$ factor in s as mentioned above. Finally, Section 9 discusses avenues for future work.

In the appendix, in Appendix A for the benefit of the reader we review many probabilistic tools that are used throughout this work . Appendix B reviews some introductory material related to convex analysis, which is helpful for understanding Section 6 on constrained least squares problems. In Appendix C we also give a direct analysis for using the FJLT for sketching constrained least squares programs, providing quantitative benefits over some analyses in [PW14].

2. Preliminaries

We fix some notation that we will be used throughout this paper. For a positive integer t, we set $[t] = \{1, \ldots, t\}$. For $a, b \in \mathbb{R}$, $a \lesssim b$ denotes $a \le Cb$ for some universal constant $C > 0$, and $a \simeq b$ signifies that both $a \lesssim b$ and $b \lesssim a$ hold. For any $x \in \mathbb{R}^n$ and $1 \le p \le \infty$, we let $\|x\|_p$ denote its ℓ_p-norm. To any set $S \subset \mathbb{R}^n$ we associate a semi-norm $|||\cdot|||_S$ defined by $|||z|||_S = \sup_{x\in S} |\langle z, x\rangle|$. Note that $|||z|||_S = |||z|||_{\mathrm{conv}(S)}$, where $\mathrm{conv}(S)$ is the closed convex hull of S, i.e., the closure of the set of all convex combinations of elements in S. We use $(e_i)_{1\le i\le n}$ and $(e_{ij})_{1\le i\le m, 1\le j\le n}$ to denote the standard basis in $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$, respectively.

If $\eta = (\eta_i)_{i\ge 1}$ is a sequence of random variables, we let $(\Omega_\eta, \mathbb{P}_\eta)$ denote the probability space on which it is defined. We use $\mathbb{E}_\eta$ and L^p_η to denote the associated expected value and L^p-space, respectively. If ζ is another sequence of random variables, then $\|\cdot\|_{L^p_\eta, L^q_\zeta}$ means that we first take the L^p_η-norm and afterwards the L^q_ζ-norm. We reserve the symbol g to denote a sequence $g = (g_i)_{i\ge 1}$ of i.i.d. standard Gaussian random variables; unless stated otherwise, the covariance matrix is the identity.

If A is an $m \times n$ matrix, then we use $\|A\|$ or $\|A\|_{\ell_2^n \to \ell_2^m}$ to denote its operator norm. Moreover we let Tr be the trace operator and use $\|A\|_F = (\mathrm{Tr}(A^*A))^{1/2}$ to denote the Frobenius norm.

In the remainder, we reserve the letter ρ to denote (semi-)metrics. If ρ corresponds to a (semi-)norm $\|\cdot\|_X$, then we let $\rho_X(x, y) = \|x - y\|_X$ denote the associated (semi-)metric. Also, we use $d_\rho(S) = \sup_{x,y\in S} \rho(x,y)$ to denote the diameter of a set S with respect to ρ and write d_X instead of d_{ρ_X} for brevity. So, for example, $\rho_{\ell_2^n}$ is the Euclidean metric and $d_{\ell_2^n}(S)$ the ℓ_2-diameter of S. From here on, T is always a fixed subset of $S^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$, and $\varepsilon \in (0, 1/2)$ the parameter appearing in (1.3).

We make use of chaining results in the remainder, so we define some relevant quantities. For a bounded set $S \subset \mathbb{R}^n$, $g(S) = \mathbb{E}_g \sup_{x\in S} \langle g, x\rangle$ is the *Gaussian mean width* of S, where $g \in \mathbb{R}^n$ is a Gaussian vector with identity covariance matrix. For a (semi-)metric ρ on $\mathbb{R}^n$, Talagrand's γ_2-functional is defined by

$$\gamma_2(S,\rho) = \inf_{\{S_r\}_{r=0}^\infty} \sup_{x\in S} \sum_{r=0}^\infty 2^{r/2}\cdot \rho(x, S_r) \tag{2.1}$$

where $\rho(x, S_r)$ is the distance from x to S_r, and the infimum is taken over all collections $\{S_r\}_{r=0}^\infty$, $S_0 \subset S_1 \subset \ldots \subseteq S$, with $|S_0| = 1, |S_r| \le 2^{2^r}$. If ρ corresponds to a (semi-)norm $\|\cdot\|_X$, then we usually write $\gamma_2(S, \|\cdot\|_X)$ instead of $\gamma_2(S, \rho_X)$. It is known that for any bounded $S \subset \mathbb{R}^n$, $g(S)$ and $\gamma_2(S, \|\cdot\|_2)$ differ multiplicatively by at most a universal constant [Fer75, Tal05]. Whenever γ_2 appears without a specified norm, we imply use of ℓ_2 or $\ell_2 \to \ell_2$ operator norm. We frequently use the entropy integral estimate (see [Tal05])

$$\gamma_2(S,\rho) \lesssim \int_0^\infty \log^{1/2} \mathcal{N}(S,\rho,u)du. \tag{2.2}$$

Here $\mathcal{N}(S, \rho, u)$ denotes the minimum number of ρ-balls of radius u centered at points in S required to cover S. If ρ corresponds to a (semi-)norm $\|\cdot\|_X$, then we write $\mathcal{N}(S, \|\cdot\|_X, u)$ instead of $\mathcal{N}(S, \rho_X, u)$.

Let us now introduce the sparse Johnson-Lindenstrauss transform in detail. Let $\sigma_{ij} : \Omega_\sigma \to \{-1, 1\}$ be independent Rademacher random variables, i.e., $\mathbb{P}(\sigma_{ij} = 1) = \mathbb{P}(\sigma_{ij} = -1) = 1/2$. We consider random variables $\delta_{ij} : \Omega_\delta \to \{0, 1\}$ with the following properties:

- For fixed j the δ_{ij} are negatively correlated, i.e.

$$\forall 1 \le i_1 < i_2 < \ldots < i_k \le m,\ \mathbb{E}\Big(\prod_{t=1}^k \delta_{i_t,j}\Big) \le \prod_{t=1}^k \mathbb{E}\,\delta_{i_t,j} = \Big(\frac{s}{m}\Big)^k; \tag{2.3}$$

- For any fixed j there are exactly s nonzero δ_{ij}, i.e., $\sum_{i=1}^m \delta_{ij} = s$;
- The vectors $(\delta_{ij})_{i=1}^m$ are independent across different $1 \le j \le n$.

We emphasize that the σ_{ij} and δ_{ij} are independent, as they are defined on different probability spaces. The *sparse Johnson-Lindenstrauss transform* Φ *with column sparsity* s, or *SJLT* for short, is defined by

$$\Phi_{ij} = \frac{1}{\sqrt{s}}\sigma_{ij}\delta_{ij}. \tag{2.4}$$

The work [KN14] gives two implementations of such a Φ satisfying the above conditions. In one example, the columns are independent, and in each column we choose exactly s locations uniformly at random, without replacement, to specify the δ_{ij}. The other example is essentially the CountSketch of [CCFC04]. In this implementation, the rows of Φ are partitioned arbitrarily into s groups of size m/s each. Then each column of Φ is chosen independently, where in each column and for each block we pick a random row in that block to be non-zero for that column (the rest are zero).

We say that Φ is an ε*-restricted isometry* on T if Eq. (1.2) holds. We define

$$\varepsilon_{\Phi,T} = \sup_{x\in T} |\|\Phi x\|_2^2 - 1|$$

as the *restricted isometry constant* of Φ on T. In the following we will be interested in estimating $\varepsilon_{\Phi,T}$. For this purpose, we use the following L^p-bound for the supremum of a second order Rademacher chaos from [KMR14] (see [Dir15, Theorem 6.5] for the refinement stated here).

Theorem 4. *Let $\mathcal{A} \subset \mathbb{R}^{m\times n}$ and let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher random variables. For any $1 \le p < \infty$*

$$\Big(\mathop{\mathbb{E}}_{\sigma} \sup_{A\in\mathcal{A}} \Big|\|A\sigma\|_2^2 - \mathbb{E}\,\|A\sigma\|_2^2\Big|^p\Big)^{1/p} \lesssim \gamma_2^2(\mathcal{A}, \|\cdot\|_{\ell_2\to\ell_2}) + d_F(\mathcal{A})\gamma_2(\mathcal{A}, \|\cdot\|_{\ell_2\to\ell_2}) + \sqrt{p}d_F(\mathcal{A})d_{\ell_2\to\ell_2}(\mathcal{A}) + pd^2_{\ell_2\to\ell_2}(\mathcal{A}). \quad (2.5)$$

For $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$, let $u\otimes v : \mathbb{R}^m \to \mathbb{R}^n$ be the operator $(u\otimes v)w = \langle v, w\rangle u$. Then, for any $x \in \mathbb{R}^n$ we can write $\Phi x = A_{\delta,x}\sigma$, where

$$A_{\delta,x} := \frac{1}{\sqrt{s}}\sum_{i=1}^m\sum_{j=1}^n \delta_{ij}x_j e_i\otimes e_{ij} = \frac{1}{\sqrt{s}}\begin{bmatrix} -x^{(\delta_{1,\cdot})}- & 0 & \cdots & 0 \\ 0 & -x^{(\delta_{2,\cdot})}- & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & -x^{(\delta_{m,\cdot})}- \end{bmatrix}. \quad (2.6)$$

for $x_j^{(\delta_{i,\cdot})} = \delta_{ij}x_j$. Note that $\mathbb{E}\,\|\Phi x\|_2^2 = \|x\|_2^2$ for all $x \in \mathbb{R}^n$ and therefore

$$\sup_{x\in T} |\|\Phi x\|_2^2 - \|x\|_2^2| = \sup_{x\in T} |\|A_{\delta,x}\sigma\|_2^2 - \mathbb{E}\,\|A_{\delta,x}\sigma\|_2^2|.$$

Associated with $\delta = (\delta_{ij})$ we define a random norm on $\mathbb{R}^n$ by

$$\|x\|_\delta = \frac{1}{\sqrt{s}}\max_{1\le i\le m}\Big(\sum_{j=1}^n \delta_{ij}x_j^2\Big)^{1/2}. \quad (2.7)$$

With this definition, we have for any $x, y \in T$,

$$\|A_{\delta,x} - A_{\delta,y}\| = \|x-y\|_\delta, \qquad \|A_{\delta,x} - A_{\delta,y}\|_F = \|x-y\|_2. \quad (2.8)$$

Therefore, (2.5) implies that

$$\Big(\mathop{\mathbb{E}}_{\sigma}\sup_{x\in T}\Big|\|\Phi x\|_2^2 - \|x\|_2^2\Big|^p\Big)^{1/p} \lesssim \gamma_2^2(T, \|\cdot\|_\delta) + \gamma_2(T, \|\cdot\|_\delta) + \sqrt{p}d_\delta(T) + pd_\delta^2(T). \quad (2.9)$$

Taking $L_p(\Omega_\delta)$-norms on both sides yields

$$\Big(\mathop{\mathbb{E}}_{\delta,\sigma}\sup_{x\in T}\Big|\|\Phi x\|_2^2 - \|x\|_2^2\Big|^p\Big)^{1/p} \lesssim (\mathop{\mathbb{E}}_{\delta}\gamma_2^{2p}(T, \|\cdot\|_\delta))^{1/p} + (\mathop{\mathbb{E}}_{\delta}\gamma_2^p(T, \|\cdot\|_\delta))^{1/p} + \sqrt{p}(\mathbb{E}\,d_\delta^p(T))^{1/p} + p(\mathbb{E}\,d_\delta^{2p}(T))^{1/p}. \quad (2.10)$$

Thus, to find a bound for the expected value of the restricted isometry constant of Φ on T, it suffices to estimate $\mathbb{E}_\delta\,\gamma_2^2(T, \|\cdot\|_\delta)$ and $\mathbb{E}_\delta\,d_\delta^2(T)$. If we can find good bounds on $(\mathbb{E}_\delta\,\gamma_2^p(T, \|\cdot\|_\delta))^{1/p}$ and $(\mathbb{E}\,d_\delta^p(T))^{1/p}$ for all $p \ge 1$, then we in addition obtain a high probability bound for the restricted isometry constant.

Unless explicitly stated otherwise, from here on Φ always denotes the SJLT with s non-zeroes per column.

3. Overview of proof of main theorem

Here we give an overview of the proof of Theorem 3. To illustrate the ideas going into our proof, it is simplest to first consider the case where T is the set of all unit vectors in a d-dimensional linear subspace $E \subset \mathbb{R}^n$. By Eq. (2.10) for $p = 1$ we have to bound for example $\mathbb{E}_\delta \gamma_2(T, \|\cdot\|_\delta)$. Standard estimates give, up to $\log d$,

$$\gamma_2(T, \|\cdot\|_\delta) \le \gamma_2(B_E, \|\cdot\|_\delta) \ll \sup_{t>0} t[\log \mathcal{N}(B_E, \|\cdot\|_\delta, t)]^{1/2}. \tag{3.1}$$

for B_E the unit ball of E. Let $U \in \mathbb{R}^{n\times d}$ have columns forming an orthonormal basis for E. Then dual Sudakov minoration [BLM89, Proposition 4.2], [PTJ86] states that

$$\sup_{t>0} t[\log \mathcal{N}(B_E, \|\cdot\|_\delta, t)]^{1/2} \le \mathop{\mathbb{E}}_g \|Ug\|_\delta \tag{3.2}$$

for a Gaussian vector g. From this point onward one can arrive at a result using the non-commutative Khintchine inequality [LP86, LPP91] and other standard Gaussian concentration arguments (see appendix).

Unfortunately, Eq. (3.2) is very specific to unit balls of linear subspaces and has no analog for a general set T. At this point we use a statement about the duality of entropy numbers [BPSTJ89, Proposition 4]. This is the principle that for two symmetric convex bodies K and D, $\mathcal{N}(K, D)$ and $\mathcal{N}(D^\circ, aK^\circ)$ are roughly comparable for some constant a ($\mathcal{N}(K, D)$ is the number of translates of D needed to cover K). Although it has been an open conjecture for over 40 years as to whether this holds in the general case [Pie72, p. 38], the work [BPSTJ89, Proposition 4] shows that these quantities are comparable up to logarithmic factors as well as a factor depending on the type-2 constant of the norm defined by D (i.e. the norm whose unit vectors are those on the boundary of D). In our case, this lets us relate $\log \mathcal{N}(\tilde{T}, \|\cdot\|_\delta, t)$ with $\log \mathcal{N}(\mathrm{conv}(\cup_{i=1}^m B_{J_i}), \|\cdot\|_T, \sqrt{s}t/8)$, losing small factors. Here $\tilde{T}$ is the convex hull of $T \cup -T$, and B_{J_i} is the unit ball of $\mathrm{span}\{e_j : \delta_{ij} = 1\}$. We next use Maurey's lemma, which is a tool for bounding covering numbers of the set of convex combinations of vectors in various spaces. This lets us relate $\log \mathcal{N}(\mathrm{conv}(\cup_{i=1}^m B_{J_i}), \|\cdot\|_T, \epsilon)$ to quantities of the form $\log \mathcal{N}(\frac{1}{k}\sum_{i\in A} B_{J_i}, \|\cdot\|_T, \epsilon)$, where $A \subset [m]$ has size $k \lesssim 1/\epsilon^2$ (see Lemma 18). For a fixed A, we bucket $j \in [n]$ according to $\sum_{i\in A} \delta_{ij}$ and define $U_\alpha = \{j \in [n] : \sum_{i\in A} \delta_{ij} \simeq 2^\alpha\}$. Abusing notation, we also let U_α denote the coordinate subspace spanned by $j \in U_\alpha$. This leads to (see Eq. (7.11))

$$\log \mathcal{N}\Big(\frac{1}{k}\sum_{i\in A} B_{J_i}, \|\cdot\|_T, \epsilon\Big) \lesssim \sum_\alpha \log \mathcal{N}\Big(B_{U_\alpha}, \|\cdot\|_T, \sqrt{\frac{k}{2^\alpha}}\frac{\epsilon}{\log m}\Big) \tag{3.3}$$

Finally we are in a position to apply dual Sudakov minoration to the right hand side of Eq. (3.3), after which point we apply various concentration arguments to yield our main theorem.

4. The case of a linear subspace

Let $E \subset \mathbb{R}^n$ be a d-dimensional linear subspace, $T = E \cap S^{n-1}$, B_E the unit ℓ_2-ball of E. We use P_E to denote the orthogonal projection onto E. The values $\|P_E e_j\|_2$, $j = 1, \ldots, n$, are typically referred to as the *leverage scores* of E in the numerical linear algebra literature. We denote the maximum leverage score by

$$\mu(E) = \max_{1\le j\le n} \|P_E e_j\|_2,$$

which is sometimes referred to as the *incoherence* $\mu(E)$ of E.

In our proof we make use of the well-known Gaussian concentration of Lipschitz functions [Pis86, Corollary 2.3]: if g is an n-dimensional standard Gaussian vector and $\phi : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz from ℓ_2^n to $\mathbb{R}$, then for all $p \ge 1$

$$\|\phi(g) - \mathbb{E}\,\phi(g)\|_{L_g^p} \lesssim L\sqrt{p}. \tag{4.1}$$

Theorem 5. *For any $p \ge \log m$ and any $0 < \epsilon < 1$,*

$$(\mathbb{E}\,\gamma_2^{2p}(T, \|\cdot\|_\delta))^{1/p} \lesssim \epsilon^2 + \frac{(d+\log m)\log^2(d/\epsilon)}{m} + \frac{p\log^2(d/\epsilon)\log m}{s}\mu(E)^2 \tag{4.2}$$

and

$$(\mathbb{E}\, d_\delta^{2p}(T))^{1/p} \lesssim \frac{d}{m} + \frac{p}{s}\mu(E)^2. \tag{4.3}$$

As a consequence, if $\eta \le 1/m$ and

$$\begin{aligned} m &\gtrsim ((d+\log m)\min\{\log^2(d/\varepsilon), \log^2(m)\} + d\log(\eta^{-1}))/\varepsilon^2\\ s &\gtrsim (\log(m)\log(\eta^{-1})\min\{\log^2(d/\varepsilon),\log^2(m)\} + \log^2(\eta^{-1}))\mu(E)^2/\varepsilon^2 \end{aligned} \tag{4.4}$$

then Eq. (1.2) *holds with probability at least $1-\eta$.*

Proof. By dual Sudakov minoration (Lemma 29 in Appendix A)

$$\log \mathcal{N}(B_E, \|\cdot\|_\delta, t) \lesssim \frac{\mathbb{E}_g \|Ug\|_\delta}{t}, \qquad \text{for all } t > 0,$$

with $U \in \mathbb{R}^{n\times d}$ having columns $\{f_k\}_{k=1}^d$ forming an orthonormal basis for E and g a random Gaussian vector. Let $U^{(i)}$ be U but where each row j is multiplied by δ_{ij}. Then by Eq. (2.7) and taking $\ell = \log m$,

$$\begin{aligned} \mathop{\mathbb{E}}_g \|Ug\|_\delta &= \frac{1}{\sqrt{s}} \mathop{\mathbb{E}}_g \max_{1\le i\le m} \Big[\sum_{j=1}^n \delta_{ij}\Big|\sum_{k=1}^d g_k\,\langle f_k, e_j\rangle\Big|^2\Big]^{1/2}\\ &= \frac{1}{\sqrt{s}} \mathop{\mathbb{E}}_g \max_{1\le i\le m} \|U^{(i)}g\|_2\\ &\le \frac{1}{\sqrt{s}}\Big(\max_{1\le i\le m}\mathop{\mathbb{E}}_g \|U^{(i)}g\|_2 + \mathop{\mathbb{E}}_g \max_{1\le i\le m}\Big|\|U^{(i)}g\|_2 - \mathop{\mathbb{E}}_g\|U^{(i)}g\|_2\Big|\Big)\\ &\le \frac{1}{\sqrt{s}}\Big(\max_{1\le i\le m}\|U^{(i)}\|_F + \Big(\sum_{i=1}^m \mathop{\mathbb{E}}_g\Big|\|U^{(i)}g\|_2 - \mathop{\mathbb{E}}_g\|U^{(i)}g\|_2\Big|^\ell\Big)^{1/\ell}\Big)\\ &\lesssim \frac{1}{\sqrt{s}}\Big(\max_{1\le i\le m}\|U^{(i)}\|_F + \sqrt{\ell}\cdot\max_{1\le i\le m}\|U^{(i)}\|_{\ell_2^d\to\ell_2^n}\Big) \end{aligned} \tag{4.5}$$

where Eq. (4.5) used Gaussian concentration of Lipschitz functions (cf. Eq. (4.1)), noting that $g \mapsto \|U^{(i)}g\|_2$ is $\|U^{(i)}\|$-Lipschitz.

Since $\|\cdot\|_\delta \le (1/\sqrt{s})\|\cdot\|_2$, we can use for small t the bound (cf. Eq. (A.6))

$$\begin{aligned} \log\mathcal{N}(T, \|\cdot\|_\delta, t) &\le \log \mathcal{N}(B_E, \|\cdot\|_2, t\sqrt{s})\\ &< d\cdot\log\Big(1 + \frac{2}{t\sqrt{s}}\Big) \end{aligned} \tag{4.6}$$

Using Eq. (4.6) for small t and noting $\|U^{(i)}\|_F = (\sum_{j=1}^n \delta_{ij}\|P_E e_j\|_2^2)^{1/2}$ for P_E the orthogonal projection onto E, Eq. (2.2) yields for $t^* = (\epsilon/d)/\log(d/\epsilon)$

$$\begin{aligned} &\gamma_2(T, \|\cdot\|_\delta)\\ &\quad\lesssim \int_0^{t^*} \sqrt{d}\cdot\Big[\log\Big(2+\frac{1}{t\sqrt{s}}\Big)\Big]^{1/2}dt + \int_{t^*}^{1/\sqrt{s}} \frac{\mathbb{E}_g\|Ug\|_\delta}{t}dt\\ &\quad\lesssim \sqrt{d}t_*\Big[\log\Big(\frac{1}{t_*\sqrt{s}}\Big)\Big]^{1/2} + \mathop{\mathbb{E}}_g\|Ug\|_\delta \log\Big(\frac{1}{t_*\sqrt{s}}\Big)\\ &\quad\lesssim \epsilon + \frac{\log(d/\epsilon)}{\sqrt{s}}\cdot\Big[\max_{1\le i\le m}\Big[\sum_{j=1}^n \delta_{ij}\|P_E e_j\|_2^2\Big]^{1/2} + \sqrt{\log m}\max_{1\le i\le m}\|U^{(i)}\|_{\ell_2^d\to\ell_2^n}\Big] \end{aligned} \tag{4.7}$$

As a consequence,

$$(\mathop{\mathbb{E}}_{\delta}\gamma_2^{2p}(T,\|\cdot\|_\delta))^{1/p} \lesssim \epsilon^2 + \frac{\log^2(d/\epsilon)}{s}\Big[\Big(\mathop{\mathbb{E}}_{\delta}\max_{1\le i\le m}\Big[\sum_{j=1}^n \delta_{ij}\|P_E e_j\|_2^2\Big]^p\Big)^{1/p} + \log(m)\Big(\mathop{\mathbb{E}}_{\delta}\max_{1\le i\le m}\|U^{(i)}\|_{\ell_2^d\to\ell_2^n}^{2p}\Big)^{1/p}\Big] \quad (4.8)$$

We estimate the first non-trivial term on the right hand side. Since $p \ge \log m$,

$$\Big(\mathop{\mathbb{E}}_{\delta}\max_{1\le i\le m}\Big[\sum_{j=1}^n \delta_{ij}\|P_E e_j\|_2^2\Big]^p\Big)^{1/p} \le \Big(\sum_{i=1}^m \mathop{\mathbb{E}}_{\delta}\Big|\sum_{j=1}^n \delta_{ij}\|P_E e_j\|_2^2\Big|^p\Big)^{1/p} \le e\cdot\max_{1\le i\le m}\Big(\mathop{\mathbb{E}}_{\delta}\Big[\sum_{j=1}^n \delta_{ij}\|P_E e_j\|_2^2\Big]^p\Big)^{1/p}$$

Note that for fixed i, the $(\delta_{ij})_{1\le j\le n}$ are i.i.d. variables of mean s/m and therefore

$$\mathbb{E}\sum_{j=1}^n \delta_{ij}\|P_E e_j\|_2^2 = \frac{s}{m}\sum_{j=1}^n\langle P_E e_j, e_j\rangle = \frac{s}{m}\mathrm{Tr}(P_E) = \frac{sd}{m}.$$

Let $(r_j)_{1\le j\le n}$ be a vector of independent Rademachers. By symmetrization (Eq. (A.4)) and Khintchine's inequality (Eq. (A.2)),

$$\begin{aligned}\Big\|\sum_{j=1}^n \delta_{ij}\|P_E e_j\|_2^2\Big\|_{L^p_\delta} &\le \frac{ds}{m} + \Big\|\sum_{j=1}^n(\delta_{ij}-\mathbb{E}\,\delta_{ij})\|P_E e_j\|_2^2\Big\|_{L^p_\delta}\\ &\le \frac{ds}{m} + 2\Big\|\sum_{j=1}^n r_j\delta_{ij}\|P_E e_j\|_2^2\Big\|_{L^p_{\delta,r}}\\ &\lesssim \frac{ds}{m} + \sqrt{p}\Big\|\Big(\sum_{j=1}^n \delta_{ij}\|P_E e_j\|_2^4\Big)^{1/2}\Big\|_{L^p_\delta}\\ &\le \frac{ds}{m} + \sqrt{p}\cdot\max_{1\le j\le n}\|P_E e_j\|_2\cdot\Big\|\sum_{j=1}^n \delta_{ij}\|P_E e_j\|_2^2\Big\|_{L^p_\delta}^{1/2}\end{aligned}$$

Solving this quadratic inequality yields

$$\Big\|\sum_{j=1}^n \delta_{ij}\|P_E e_j\|_2^2\Big\|_{L^p_\delta} \lesssim \frac{ds}{m} + p\cdot\max_{1\le j\le n}\|P_E e_j\|_2^2 \quad (4.9)$$

We conclude that

$$\Big(\mathop{\mathbb{E}}_{\delta}\max_{1\le i\le m}\Big[\sum_{j=1}^n \delta_{ij}\|P_E e_j\|_2^2\Big]^p\Big)^{1/p} \lesssim \frac{ds}{m} + p\max_j\|P_E e_j\|_2^2 \quad (4.10)$$

We now estimate the last term on the right hand side of Eq. (4.8). As $p \ge \log m$,

$$\Big(\mathop{\mathbb{E}}_{\delta}\max_{1\le i\le m}\|U^{(i)}\|_{\ell_2^d\to\ell_2^n}^{2p}\Big)^{1/p} \lesssim \max_{1\le i\le m}\Big(\mathbb{E}\,\|U^{(i)}\|_{\ell_2^d\to\ell_2^n}^{2p}\Big)^{1/p} = \max_{1\le i\le m}\Big(\mathbb{E}\,\|(U^{(i)})^*U^{(i)}\|_{\ell_2^d\to\ell_2^d}^{p}\Big)^{1/p}.$$

If u_j denotes the jth row of U, then

$$(U^{(i)})^*U^{(i)} = \sum_{j=1}^n \delta_{ij}u_j u_j^*.$$

By symmetrization and the non-commutative Khintchine inequality (cf. Eq. (A.3)),

$$\begin{aligned}\Big\|\sum_{j=1}^n \delta_{ij}u_ju_j^*\Big\|_{L_\delta^p} &\le \frac{s}{m} + \Big\|\sum_{j=1}^n \delta_{ij}u_ju_j^* - \frac{s}{m}I\Big\|_{L_\delta^p}\\ &\le \frac{s}{m} + 2\Big\|\sum_{j=1}^n r_j\delta_{ij}u_ju_j^*\Big\|_{L_{\delta,r}^p}\\ &\lesssim \frac{s}{m} + \sqrt{p}\Big\|\sum_{j=1}^n \delta_{ij}\|u_j\|_2^2 u_ju_j^*\Big\|_{L_\delta^{p/2}}^{1/2}\\ &\le \frac{s}{m} + \sqrt{p}\max_{1\le j\le n}\|P_Ee_j\|_2\cdot\Big\|\sum_{j=1}^n \delta_{ij}u_ju_j^*\Big\|_{L_\delta^p}^{1/2}\end{aligned}$$

Solving this quadratic inequality, we find

$$\Big\|\sum_{j=1}^n \delta_{ij}u_ju_j^*\Big\|_{L_\delta^p} \lesssim \frac{s}{m} + p\cdot\max_{1\le j\le n}\|P_Ee_j\|_2^2$$

and as a consequence,

$$\Big(\mathop{\mathbb{E}}_\delta \max_{1\le i\le m}\|U^{(i)}\|_{\ell_2^d\to\ell_2^n}^{2p}\Big)^{1/p} \lesssim \frac{s}{m} + p\cdot\max_{1\le j\le n}\|P_Ee_j\|_2^2$$

Applying this estimate together with Eq. (4.10) in Eq. (4.8) yields the asserted estimate in Eq. (4.2). For the second assertion, note by Cauchy-Schwarz that

$$\begin{aligned}d_\delta(T) &= \frac{1}{\sqrt{s}}\max_{1\le i\le m}\sup_{x\in T}\Big[\sum_{j=1}^n \delta_{ij}x_j^2\Big]^{1/2}\\ &= \frac{1}{\sqrt{s}}\max_{1\le i\le m}\sup_{x\in T}\Big[\sum_{j=1}^n \delta_{ij}\Big(\sum_{k=1}^d \langle x, f_k\rangle\langle f_k, e_j\rangle\Big)^2\Big]^{1/2}\\ &\le \frac{1}{\sqrt{s}}\max_{1\le i\le m}\Big[\sum_{j=1}^n \delta_{ij}\|P_Ee_j\|_2^2\Big]^{1/2}.\end{aligned}$$

Therefore,

$$(\mathop{\mathbb{E}}_\delta d_\delta^{2p}(T))^{1/p} \le \frac{1}{s}\Big(\mathop{\mathbb{E}}_\delta \max_{1\le i\le m}\Big[\sum_{j=1}^n \delta_{ij}\|P_Ee_j\|_2^2\Big]^p\Big)^{1/p}$$

and Eq. (4.3) immediately follows from Eq. (4.10).

To prove the final assertion, we combine (2.10), Eq. (4.2) and Eq. (4.3) to obtain

$$\begin{aligned}&\Big(\mathop{\mathbb{E}}_{\delta,\sigma}\sup_{x\in T}\Big|\|\Phi x\|^2 - \|x\|^2\Big|^p\Big)^{1/p}\\ &\quad\lesssim \epsilon^2 + \frac{(d+\log m)\log^2(d/\epsilon)}{m} + \frac{p\log^2(d/\epsilon)\log m}{s}\max_{1\le j\le n}\|P_Ee_j\|_2^2\\ &\qquad + \Big[\epsilon^2 + \frac{(d+\log m)\log^2(d/\epsilon)}{m} + \frac{p\log^2(d/\epsilon)\log m}{s}\max_{1\le j\le n}\|P_Ee_j\|_2^2\Big]^{1/2}\\ &\qquad + \frac{pd}{m} + \frac{p^2}{s}\max_{1\le j\le n}\|P_Ee_j\|_2^2 + \Big[\frac{pd}{m} + \frac{p^2}{s}\max_{1\le j\le n}\|P_Ee_j\|_2^2\Big]^{1/2}.\end{aligned}$$

Now apply Lemma 27 of Appendix A to obtain, for any $w \ge \log m$,

$$\mathbb{P}_{\delta,\sigma}\Big(\varepsilon_{\Phi,T} \gtrsim \epsilon^2 + \frac{(d+\log m)\log^2(d/\epsilon)}{m} + \frac{w\log^2(d/\epsilon)\log m}{s}\max_{1\le j\le n}\|P_Ee_j\|_2^2$$

$$+\Big[\epsilon^2 + \frac{(d+\log m)\log^2(d/\epsilon)}{m} + \frac{w\log^2(d/\epsilon)\log m}{s}\max_{1\le j\le n}\|P_E e_j\|_2^2\Big]^{1/2}$$
$$+\frac{wd}{m} + \frac{w^2}{s}\max_{1\le j\le n}\|P_E e_j\|_2^2 + \Big[\frac{wd}{m} + \frac{w^2}{s}\max_{1\le j\le n}\|P_E e_j\|_2^2\Big]^{1/2}\Big) \le e^{-w}.$$

Now set $w = \log(\eta^{-1})$, choose $\epsilon = \varepsilon/C$ (with C a large enough absolute constant) and $\epsilon = d/m$, and use the assumptions on m and s to conclude $\mathbb{P}(\varepsilon_{\Phi,T} \ge \varepsilon) \le \eta$. □

Theorem 5 recovers a similar result in [NN13a] but via a different method, less logarithmic factors in the setting of m, and the revelation that s can be taken smaller if all the leverage scores $\|P_E e_j\|_2$ are small (note if $\|P_E e_j\|_2 \ll (\log d \cdot \log m)^{-1}$ for all j, we may take $s = 1$, though this is not true in general). Our dependence on $1/\varepsilon$ in s is quadratic instead of the linear dependence in [NN13a], but as stated earlier, in most applications of OSE's ε is taken a constant.

Remark 6. As stated, one consequence of the above is we can set $s = 1$, and m to nearly linear in d as long as μ is at most inverse polylogarithmic in d. It is worth noting that if $d \gtrsim \log n$, then a random d-dimensional subspace E has $\mu(E) \lesssim \sqrt{d/n}$ by the JL lemma, and thus most subspaces do have even much lower coherence. Note that the latter bound is sharp. Indeed, let E be any d-dimensional subspace and let $U \in \mathbb{R}^{n\times d}$ have orthonormal columns such that E is the column space of U. Then

$$\max_{v\in E, \|v\|_2=1}\max_{1\le j\le n} |\langle v, e_j\rangle|$$

is the maximum ℓ_2 norm of a row u_j of U. Since U has n rows and $\|U\|_F^2 = d$, there exists a j such that $\|u_j\|_2 \ge \sqrt{d/n}$ and in particular $\mu(E) \ge \sqrt{d/n}$.

It is also worth noting that [AMT10, Theorem 3.3] observed that if $H \in \mathbb{R}^{n\times n}$ is a bounded orthonormal system (i.e. H is orthogonal and $\max_{i,j}|H_{ij}| = O(1/\sqrt{n})$), and D is a diagonal matrix with independent Rademacher entries, then HDE has incoherence $\mu \lesssim \sqrt{d(\log n)/n}$ with probability $1 - 1/\mathrm{poly}(n)$. They then showed that incoherence μ implies that a sampling matrix $S \in \mathbb{R}^{m\times n}$ suffices to achieve Eq. (1.3) with $m \lesssim n\mu^2\log m/\varepsilon^2$. Putting these two facts together implies SHD gives Eq. (1.3) with $m \lesssim d(\log n)(\log m)/\varepsilon^2$. Eq. (4.4) combined with [AMT10, Theorem 3.3] implies a statement of a similar flavor but with an arguably stronger implication for applications: ΦHD with $s = 1$ satisfies Eq. (1.3) for $T = E \cap S^{n-1}$ without increasing m as long as $n \gtrsim d(\log(d/\varepsilon))^c/\varepsilon^2$.

5. The type-2 case

Recall that the *type-2 constant* $T_2(\|\cdot\|)$ of a semi-norm $\|\cdot\|$ is defined as the best constant (if it exists) such that the inequality

$$\mathbb{E}_g\Big\|\sum_\alpha g_\alpha x_\alpha\Big\| \le T_2\Big(\sum_\alpha \|x_\alpha\|^2\Big)^{1/2} \tag{5.1}$$

holds, for all finite systems of vectors $\{x_\alpha\}$ and i.i.d. standard Gaussian (g_α).

Given $T \subset S^{n-1}$, define the semi-norm

$$|||x|||_T = \sup_{y\in T}|\langle x, y\rangle|$$

It will be convenient to use the following notation. For $i = 1, \ldots, m$ we let J_i be the set of 'active' indices, i.e., $J_i = \{j = 1, \ldots, n : \delta_{ij} = 1\}$. We can then write

$$\|x\|_\delta = \frac{1}{\sqrt{s}}\max_{1\le i\le m}\|P_{J_i}x\|_2, \tag{5.2}$$

where we abuse notation and let P_{J_i} denote orthogonal projection onto the coordinate subspace specified by J_i.

Lemma 7. *Let $\|\cdot\|$ be any semi-norm satisfying $|||x|||_T \le \|x\|$ for all $x \in \mathbb{R}^n$. Then,*

$$\gamma_2(T, \|\cdot\|_\delta) \lesssim \frac{1}{\sqrt{s}} T_2(\|\cdot\|)(\log n)^3 (\log m)^{1/2} \times \Big(\max_{1\le i\le m} \mathop{\mathbb{E}}_g \Big\| \sum_{j=1}^n g_j \delta_{ij} e_j \Big\| + (\log m)^{1/2} d_{\|\cdot\|}\Big(\bigcup_{i=1}^m B_{J_i}\Big)\Big),$$

where $T_2(\|\cdot\|)$ is the type-2 constant and for $1 \le i \le m$,

$$B_{J_i} = \Big\{ \sum_{j\in J_i} x_j e_j : \sum_{j \in J_i} x_j^2 \le 1 \Big\}.$$

Note also that $d_\delta(T)$ can be bounded as

$$\begin{aligned} d_\delta(T) &= \frac{1}{\sqrt{s}} \max_{1\le i\le m} \sup_{x\in T} \Big(\sum_{j=1}^n \delta_{ij} \langle x, e_j\rangle^2 \Big)^{1/2} \\ &= \frac{1}{\sqrt{s}} \max_{1\le i\le m} \sup_{x\in T} \Big(\mathop{\mathbb{E}}_g \Big| \sum_{j=1}^n \delta_{ij} g_j \langle x, e_j\rangle \Big|^2 \Big)^{1/2} \\ &\lesssim \frac{1}{\sqrt{s}} \max_{1\le i\le m} \sup_{x\in T} \mathop{\mathbb{E}}_g \Big| \Big\langle x, \sum_{j=1}^n \delta_{ij} g_j e_j \Big\rangle \Big| \le \frac{1}{\sqrt{s}} \max_{1\le i\le m} \mathop{\mathbb{E}}_g \Big\| \sum_{j=1}^n \delta_{ij} g_j e_j \Big\|. \end{aligned} \tag{5.3}$$

Remark 8. Replacing $|||\cdot|||_T$ by a stronger norm $\|\cdot\| \ge |||\cdot|||_T$ in this lemma may reduce the type-2 constant. The application to k-sparse vectors will illustrate this.

Proof of Lemma 7. By Eq. (2.2), to bound $\gamma_2(T, \|\cdot\|_\delta)$ it suffices to evaluate the covering numbers $\mathcal{N}(T, \|\cdot\|_\delta, t)$ for $t < d_\delta^*(T)/\sqrt{s}$, where we set

$$d_\delta^*(T) = \sqrt{s} d_\delta(T) = \sup_{x\in T} \max_{1\le i\le m} \|P_{J_i} x\|_2.$$

For large values of t we use duality for covering numbers. It will be convenient to work with $\tilde{T} = \mathrm{conv}(T \cup (-T))$, which is closed, bounded, convex and symmetric. Clearly,

$$\mathcal{N}(T, \|\cdot\|_\delta, t) \le \mathcal{N}(\tilde{T}, \|\cdot\|_\delta, t)$$

and in the notation of Appendix A,

$$|||x|||_T = \sup_{y \in T\cup\{-T\}} \langle x, y\rangle = \sup_{y\in\tilde{T}} \langle x, y \rangle = \|x\|_{\tilde{T}^\circ}.$$

Here $\tilde{T}^\circ$ denotes the polar (see Appendix B for a definition). Lemma 30, applied with $U = \tilde{T}$, $V = \cap_{i=1}^m \{x \in \mathbb{R}^n \ : \ s^{-1/2}\|P_{J_i} x\|_2 \le 1\}$, $\varepsilon = t$ and $\theta = d_\delta^*(T)/\sqrt{s}$, implies

$$\log \mathcal{N}(\tilde{T}, \|\cdot\|_\delta, t) \le \Big(\log \frac{d_\delta^*(T)}{t\sqrt{s}} \Big) A \log \mathcal{N}\Big(\frac{1}{\sqrt{s}} \mathrm{conv}\Big(\bigcup_{i=1}^m B_{J_i} \Big), |||\cdot|||_T, \frac{t}{8} \Big)$$

where $A \lesssim T_2(\|\cdot\|_\delta)^2 \lesssim \log m$ (by (5.2)) and we used that

$$B_{J_i} = \Big\{ \sum_{j\in J_i} x_j e_j : \sum_{j\in J_i} x_j^2 \le 1 \Big\} = \{ x \in \mathbb{R}^n \ : \ \|P_{J_i} x\|_2 \le 1\}^\circ.$$

Hence,

$$\log \mathcal{N}(\tilde{T}, \|\cdot\|_\delta, t) \lesssim \Big(\log \frac{d_\delta^*(T)}{t\sqrt{s}} \Big) (\log m) \log \mathcal{N}\Big(\mathrm{conv}\Big(\bigcup_{i=1}^m B_{J_i} \Big), \|\cdot\|, \frac{1}{8}\sqrt{s} t \Big) \tag{5.4}$$

To estimate the right hand side we use the following consequence of Maurey's lemma (Lemma 31 of Appendix A).

Claim 9. *Fix $0 < R < \infty$. Let $\|\cdot\|$ be a semi-norm on $\mathbb{R}^n$ and $\Omega \subset \{x \in \mathbb{R}^n : \|x\| \le R\}$. Then for $0 < \epsilon < R$,*

$$\epsilon[\log \mathcal{N}(\mathrm{conv}(\Omega), \|\cdot\|, \epsilon)]^{1/2} \lesssim \Big(\log \frac{R}{\epsilon}\Big)^{3/2} T_2(\|\cdot\|) \max_{\epsilon \le \epsilon' \le R} \epsilon' [\log \mathcal{N}(\Omega, \|\cdot\|, \epsilon')]^{1/2} \tag{5.5}$$

Proof. It suffices to prove the result for $R = 1$, the general case then follows by considering $R^{-1}\Omega$. For each positive integer k, let $\Omega_k \subset \Omega$ be a finite set satisfying

$$|\Omega_k| \le \mathcal{N}(\Omega, \|\cdot\|, 2^{-k})$$

and

$$\inf_{y \in \Omega_k} \|x - y\| < 2^{-k} \text{ for all } x \in \Omega$$

Given $x \in \Omega$, denote $x_k \in \Omega_k$ a sequence s.t. $\|x - x_k\| < 2^{-k}$. Then, setting $y_0 = x_0$, $y_k = x_k - x_{k-1}$, we obtain

$$\Big\| x - \sum_{\epsilon/2 < 2^{-k} \le 1} y_k \Big\| < \epsilon \tag{5.6}$$

$$y_k \in \Omega_k - \Omega_{k-1} \text{ and } \|y_k\| < 2^{-k} + 2^{-k+1} = 3 \cdot 2^{-k} \tag{5.7}$$

It follows from Eq. (5.6) that

$$\mathrm{conv}(\Omega) \subset \sum_{\epsilon/2 < 2^{-k} \le 1} \mathrm{conv}(\tilde{\Omega}_k) + B_{\|\cdot\|}(0, \epsilon)$$

with

$$\tilde{\Omega}_k = \Big\{ y \in \Omega_k - \Omega_{k-1} : \|y\| < 3 \cdot 2^{-k} \Big\} \tag{5.8}$$

By Maurey's lemma, given $z \in \mathrm{conv}(\tilde{\Omega}_k)$ and a positive integer ℓ_k, there are points $y_1, \ldots, y_\ell \in \tilde{\Omega}_k$ such that

$$\Big\| z - \frac{1}{\ell_k}(y_1 + \ldots + y_{\ell_k}) \Big\| \lesssim T_2(\|\cdot\|) \frac{1}{\sqrt{\ell_k}} \cdot 3 \cdot 2^{-k} \tag{5.9}$$

Taking $\ell_k \simeq T_2(\|\cdot\|)^2 4^{-k} \epsilon^{-2} \log^2(1/\epsilon)$, we obtain a point $z_k \in \Omega'_k = \frac{1}{\ell_k}(\tilde{\Omega}_k + \ldots + \tilde{\Omega}_k)$ such that

$$\|z - z_k\| < \frac{\epsilon}{2\log(1/\epsilon)}$$

Moreover

$$\log |\Omega'_k| \le \ell_k \log |\tilde{\Omega}_k| \lesssim T_2(\|\cdot\|)^2 4^{-k} \epsilon^{-2} \log^2(1/\epsilon) \log \mathcal{N}(\Omega, \|\cdot\|, 2^{-k})$$

Since $|\{k \ : \ \epsilon/2 < 2^{-k} \le 1\}| \le \log(2/\epsilon)$, we obtain from the preceding,

$$\log \mathcal{N}(\mathrm{conv}(\Omega), \|\cdot\|, \epsilon) \lesssim \log^2(1/\epsilon) T_2(\|\cdot\|)^2 \epsilon^{-2} \sum_{\epsilon/2 < 2^{-k} \le 1} 4^{-k} \log \mathcal{N}(\Omega, \|\cdot\|, 2^{-k})$$

and Eq. (5.5) follows. This proves the claim. $\square$

Observe that

$$\begin{aligned} d^*_\delta(T) &= \sup_{x \in T} \max_{1 \le i \le m} \|P_{J_i} x\|_2 \\ &= \sup_{x \in T} \sup_{y \in B_{\ell_2^n}} \max_{1 \le i \le m} \langle x, P_{J_i} y \rangle = d_{|||\cdot|||_T}\Big(\bigcup_{i=1}^m B_{J_i}\Big) \le d_{\|\cdot\|}\Big(\bigcup_{i=1}^m B_{J_i}\Big) \end{aligned}$$

Write $d_{\|\cdot\|} := d_{\|\cdot\|}(\cup_{i=1}^m B_{J_i})$ for brevity. Applying Claim 9 with $\Omega = \cup_{i=1}^m B_{J_i}$, $\epsilon = t\sqrt{s}$, and the dominating semi-norm $\|\cdot\| \ge |||\cdot|||_T$ to the right hand side of Eq. (5.4), we find

$$t\sqrt{s}\Big[\log \mathcal{N}(\tilde{T}, \|\cdot\|_\delta, t)\Big]^{1/2} \le \Big(\log \frac{d_{\|\cdot\|}}{t\sqrt{s}}\Big)^2 (\log m)^{1/2} T_2(\|\cdot\|) \cdot$$

$$\Big[\max_{t\sqrt{s}\le t'\le d_{\|\cdot\|}} t'\Big[\log\mathcal{N}\Big(\bigcup_{i=1}^{m} B_{J_i},\|\cdot\|,t'\Big)\Big]^{1/2}\Big]. \tag{5.10}$$

Using the dual Sudakov inequality (Lemma 29 of Appendix A) we arrive at

$$\begin{aligned}\Big[\log\mathcal{N}(\tilde{T},\|\cdot\|_\delta,t)\Big]^{1/2} \le & \frac{1}{t\sqrt{s}}\Big(\log\frac{d_{\|\cdot\|}}{t\sqrt{s}}\Big)^2(\log m)^{1/2}T_2(\|\cdot\|)\cdot \\ & \Big[\max_{1\le i\le m}\mathbb{E}_g\Big\|\sum_{j\in J_i} g_j e_j\Big\| + d_{\|\cdot\|}(\log m)^{1/2}\Big]. \end{aligned}\tag{5.11}$$

We now apply Eq. (5.11) for $(sn)^{-1/2}d_{\|\cdot\|}\le t< s^{-1/2}d_{\|\cdot\|}$ and the estimate (cf. Eq. (A.6))

$$\mathcal{N}(T,\|\cdot\|_\delta,t)\le\mathcal{N}(d_\delta(T)B_{\|\cdot\|_\delta},\|\cdot\|_\delta,t)\le\Big(1+\frac{2d_\delta^*(T)}{t\sqrt{s}}\Big)^n\le\Big(1+\frac{2d_{\|\cdot\|}}{t\sqrt{s}}\Big)^n$$

for $0<t<(sn)^{-1/2}d_{\|\cdot\|}$, respectively, in Eq. (2.2). A straightforward computation, similar to Eq. (4.7), yields the result. $\square$

5.1. **Application to k-sparse vectors.** Let $\|x\|_0$ denote the number of non-zero entries of x and set

$$S_{n,k}=\{x\in\mathbb{R}^n:\|x\|_2=1,\|x\|_0\le k\}.$$

Theorem 10. *Let $A\in\mathbb{R}^{n\times n}$ be orthogonal and denote*

$$T=A(S_{n,k})$$

Let $\eta\le 1/m$. If

$$\max|A_{ij}|<(k\log n)^{-1/2} \tag{5.12}$$

and

$$\begin{aligned} m &\gtrsim k(\log n)^8(\log m)/\varepsilon^2 \\ s &\gtrsim (\log n)^7(\log m)(\log\eta^{-1})/\varepsilon^2, \end{aligned}\tag{5.13}$$

then with probability at least $1-\eta$ we have

$$(1-\varepsilon)\|x\|_2^2\le\|\Phi x\|_2^2\le(1+\varepsilon)\|x\|_2^2 \qquad \text{for all } x\in T.$$

Proof. We apply Eq. (2.10) and estimate $(\mathbb{E}_\delta\,\gamma_2^{2p}(T,\|\cdot\|_\delta))^{1/p}$ and $\mathbb{E}_\delta\,d_\delta^{2p}(T)$ for $p\ge\log m$. We trivially estimate $\mathbb{E}_\delta\,d_\delta^{2p}(T)\le 1/s$. To bound the γ_2-functional, note that $T\subset K$, where

$$K=B_{\ell_2^n}\cap\sqrt{k}A(B_{\ell_1^n}). \tag{5.14}$$

The polar body of K is given by

$$K^\circ=B_{\ell_2^n}+\frac{1}{\sqrt{k}}A(B_{\ell_\infty^n})$$

and one can readily calculate that $T_2(\|\!|\cdot|\!\|_K)\lesssim\sqrt{\log n}$ and $d_{\|\!|\cdot|\!\|_K}(\cup_{i=1}^m B_{J_i})\le 1$. We apply Lemma 7 with $\|\cdot\|=\|\!|\cdot|\!\|_K$. Note that for every $1\le i\le m$,

$$\mathbb{E}_g\Big\|\!\Big|\sum_j g_j\delta_{ij}e_k\Big|\!\Big\|_K\le\sqrt{k}\,\mathbb{E}_g\max_{1\le\ell\le n}\Big|\sum_j g_j\delta_{ij}A_{j\ell}\Big|\le\sqrt{k\log n}\max_{1\le\ell\le n}\Big(\sum_j\delta_{ij}A_{j\ell}^2\Big)^{1/2}$$

and therefore the lemma implies that

$$\begin{aligned}&\gamma_2(T,\|\cdot\|_\delta)\\ &\quad\le\frac{1}{\sqrt{s}}(\log n)^{7/2}\log m+\sqrt{\frac{k}{s}}(\log n)^4(\log m)^{1/2}\max_{1\le i\le m,1\le\ell\le n}\Big(\sum_j\delta_{ij}A_{j\ell}^2\Big)^{1/2}.\end{aligned}$$

As a consequence, as $p \geq \log m$,

$$(\mathbb{E}_{\delta} \gamma_2^{2p}(T, \|\cdot\|_{\delta}))^{1/p} \\ \leq \frac{1}{s}(\log n)^7(\log m)^2 + \frac{k}{s}(\log n)^8(\log m) \max_{1\leq i\leq m} \Big(\mathbb{E}_{\delta} \max_{1\leq \ell\leq n} \Big(\sum_j \delta_{ij} A_{j\ell}^2\Big)^p\Big)^{1/p}. \tag{5.15}$$

Since $\mathbb{E}\,\delta_{ij} = s/m$ and A is orthogonal, we obtain using symmetrization and Khintchine's inequality,

$$\begin{aligned}
&\Big(\mathbb{E}_{\delta} \max_{1\leq \ell\leq n} \Big(\sum_j \delta_{ij} A_{j\ell}^2\Big)^p\Big)^{1/p} \\
&\quad \leq \frac{s}{m} + \Big(\mathbb{E}_r \mathbb{E}_{\delta} \max_{1\leq \ell\leq n} \Big(\sum_j r_j \delta_{ij} A_{j\ell}^2\Big)^p\Big)^{1/p} \\
&\quad \leq \frac{s}{m} + \sqrt{p}\Big(\mathbb{E}_{\delta} \max_{1\leq \ell\leq n} \Big(\sum_j \delta_{ij} A_{j\ell}^4\Big)^{p/2}\Big)^{1/p} \\
&\quad \leq \frac{s}{m} + \sqrt{p}\max_{j,\ell}|A_{j\ell}|\Big(\mathbb{E}_{\delta} \max_{1\leq \ell\leq n} \Big(\sum_j \delta_{ij} A_{j\ell}^2\Big)^p\Big)^{1/(2p)}.
\end{aligned}$$

By solving this quadratic inequality, we find

$$\Big(\mathbb{E}_{\delta} \max_{1\leq \ell\leq n} \Big(\sum_j \delta_{ij} A_{j\ell}^2\Big)^p\Big)^{1/p} \lesssim \frac{s}{m} + p\max_{j,\ell} A_{j\ell}^2 \leq \frac{s}{m} + \frac{p}{k\log n}.$$

Combine this bound with Eq. (5.15) and Eq. (2.10) to arrive at

$$\begin{aligned}
&\Big(\mathbb{E}_{\delta,\sigma} \sup_{x\in T}\Big|\|\Phi x\|_2^2 - \|x\|_2^2\Big|^p\Big)^{1/p} \\
&\lesssim \frac{1}{s}(\log n)^7(\log m)^2 + \frac{k}{m}(\log n)^8(\log m) + \frac{p}{s}(\log n)^7(\log m) \\
&\quad + \Big(\frac{1}{s}(\log n)^7(\log m)^2 + \frac{k}{m}(\log n)^8(\log m) + \frac{p}{s}(\log n)^7(\log m)\Big)^{1/2} \\
&\qquad\qquad + \sqrt{\frac{p}{s}} + \frac{p}{s}.
\end{aligned}$$

Since $p \geq \log m$ was arbitrary, the result now follows from Lemma 27. □

6. Sketching constrained least squares programs

Let us now apply the previous results to constrained least squares minimization. We refer to Appendix B for any unexplained terminology. Consider $A \in \mathbb{R}^{n\times d}$ and a sketching matrix $\Phi \in \mathbb{R}^{m\times n}$. Define $f(x) = \|Ax - b\|_2^2$ and $g(x) = \|\Phi Ax - \Phi b\|_2^2$. Let $\mathcal{C} \subset \mathbb{R}^d$ be any closed convex set. Let x_* be a minimizer of the constrained least squares program

$$\min f(x) \qquad \text{subject to} \qquad x \in \mathcal{C} \tag{6.1}$$

and let $\hat{x}$ be a minimizer of the associated sketched program

$$\min g(x) \qquad \text{subject to} \qquad x \in \mathcal{C}. \tag{6.2}$$

Before giving our analysis, we define two quantities introduced in [PW14]. Given $\mathcal{K} \subset \mathbb{R}^d$ and $u \in S^{n-1}$ we set

$$\begin{aligned}
Z_1(A,\Phi,\mathcal{K}) &= \inf_{v\in A\mathcal{K}\cap S^{n-1}} \|\Phi v\|_2^2 \\
Z_2(A,\Phi,\mathcal{K},u) &= \sup_{v\in A\mathcal{K}\cap S^{n-1}} |\langle \Phi u, \Phi v\rangle - \langle u, v\rangle|.
\end{aligned}$$

We denote the tangent cone of $\mathcal{C}$ at a point x by $T_{\mathcal{C}}(x)$ (see Appendix B for a definition). The first statement in the following lemma is [PW14, Lemma 1]. For the convenience of the reader we provide a proof, which is in essence the same as in [PW14], with the second case (when x_* is a global minimizer) being a slight modification of the proof there. In what follows we assume $Ax_* \neq b$, since otherwise $f(\hat{x}) = f(x_*) = 0$.

Lemma 11. *Define* $u = (Ax_* - b)/\|Ax_* - b\|_2$ *and set* $Z_1 = Z_1(A, \Phi, T_{\mathcal{C}}(x_*))$ *and* $Z_2 = Z_2(A, \Phi, T_{\mathcal{C}}(x_*), u)$. *Then,*

$$f(\hat{x}) \leq \Big(1 + \frac{Z_2}{Z_1}\Big)^2 f(x_*).$$

If x_* *is a global minimizer of* f, *then*

$$f(\hat{x}) \leq \Big(1 + \frac{Z_2^2}{Z_1^2}\Big) f(x_*).$$

Proof. Set $e = \hat{x} - x_* \in T_{\mathcal{C}}(x_*)$ and $w = b - Ax_*$. By optimality of x_*, for any $x \in \mathcal{C}$,

$$\langle \nabla f(x_*), x - x_* \rangle = \langle Ax_* - b, A(x - x_*) \rangle \geq 0.$$

In particular, taking $x = \hat{x}$ gives $\langle w, Ae \rangle \leq 0$. As a consequence,

$$\Big\langle \frac{w}{\|w\|_2}, \frac{\Phi^*\Phi Ae}{\|Ae\|_2} \Big\rangle \leq \Big\langle \frac{w}{\|w\|_2}, \frac{\Phi^*\Phi Ae}{\|Ae\|_2} \Big\rangle - \Big\langle \frac{w}{\|w\|_2}, \frac{Ae}{\|Ae\|_2} \Big\rangle \leq Z_2. \tag{6.3}$$

By optimality of $\hat{x}$, for any $x \in \mathcal{C}$,

$$\langle \nabla g(\hat{x}), x - \hat{x} \rangle = \langle \Phi A\hat{x} - \Phi b, \Phi A(x - \hat{x}) \rangle \geq 0.$$

Taking $x = x_*$ yields

$$0 \geq \langle \Phi A\hat{x} - \Phi b, \Phi Ae \rangle = \langle -\Phi w, \Phi Ae \rangle + \|\Phi Ae\|_2^2.$$

Therefore, by (6.3),

$$Z_1 \|Ae\|_2^2 \leq \|\Phi Ae\|_2^2 \leq \langle \Phi w, \Phi Ae \rangle \leq Z_2 \|w\|_2 \|Ae\|_2.$$

In other words,

$$\|Ae\|_2 \leq \frac{Z_2}{Z_1} \|w\|_2$$

and by Cauchy-Schwarz,

$$\|A\hat{x} - b\|_2^2 = \|w\|_2^2 + \|Ae\|_2^2 - 2\langle w, Ae \rangle \leq \|w\|_2^2 \Big(1 + \frac{Z_2}{Z_1}\Big)^2.$$

If x_* is a global minimizer of f, then $\nabla f(x_*) = 0$ and therefore $\langle w, Ae \rangle = 0$ (a similar observation to [Sar06, Theorem 12]). This yields the second assertion. □

Clearly, if Φ satisfies (1.2) for $T = AT_{\mathcal{C}}(x_*) \cap S^{n-1}$ then $Z_1 \geq 1 - \varepsilon$. We do not immediately obtain an upper bound for Z_2, however, as u is in general not in $AT_{\mathcal{C}}(x_*) \cap S^{n-1}$. We mend this using the observation in Lemma 13. For its proof we recall a chaining result from [Dir15]. Let (T, ρ) be a semi-metric space. Recall that a real-valued process $(X_t)_{t \in T}$ is called subgaussian if for all $s, t \in T$,

$$\mathbb{P}(|X_t - X_s| \geq w\rho(t, s)) \leq 2\exp(-w^2) \qquad (w \geq 0).$$

The following result is a special case of [Dir15, Theorem 3.2].

Lemma 12. *If* $(X_t)_{t \in T}$ *is subgaussian, then for any* $t_0 \in T$ *and* $1 \leq p < \infty$,

$$\Big(\mathbb{E} \sup_{t \in T} |X_t - X_{t_0}|^p\Big)^{1/p} \lesssim \gamma_2(T, \rho) + \sqrt{p}\, d_\rho(T).$$

In particular, if T contains n elements, then

$$\Big(\mathbb{E}\sup_{t\in T}|X_t - X_{t_0}|^p\Big)^{1/p} \lesssim (\sqrt{p} + (\log n)^{1/2})d_\rho(T).$$

Lemma 13. *Fix $u \in B_{\ell_2^n}$, $T \subset \mathbb{R}^n$ and let Φ be the SJLT. Set*

$$Z = \sup_{v\in T}|\langle \Phi u, \Phi v\rangle - \langle u, v\rangle|.$$

For any $p \geq 1$,

$$(\mathop{\mathbb{E}}_{\delta,\sigma} Z^p)^{1/p} \lesssim \Big(\sqrt{\frac{p}{s}} + 1\Big)\Big((\mathop{\mathbb{E}}_{\delta}\gamma_2^p(T, \|\cdot\|_\delta))^{1/p} + \sqrt{p}(\mathop{\mathbb{E}}_{\delta} d_\delta^p(T))^{1/p}\Big).$$

Proof. With the notation (2.6) we have $\Phi x = A_{\delta,x}\sigma$ for any $x \in \mathbb{R}^n$. Since $\mathbb{E}_\sigma \Phi^*\Phi = I$ we find

$$Z = \sup_{v\in T}|\langle u, (\Phi^*\Phi - I)v\rangle| = \sup_{v\in T}|\sigma^* A_{\delta,u}^* A_{\delta,v}\sigma - \mathop{\mathbb{E}}_{\sigma}(\sigma^* A_{\delta,u}^* A_{\delta,v}\sigma)|.$$

Let σ' be an independent copy of σ. By decoupling (see (A.5)),

$$(\mathop{\mathbb{E}}_{\sigma} Z^p)^{1/p} \leq 4(\mathop{\mathbb{E}}_{\sigma,\sigma'}\sup_{v\in T}|\sigma^* A_{\delta,u}^* A_{\delta,v}\sigma'|^p)^{1/p}.$$

For any $v_1, v_2 \in T$, Hoeffding's inequality implies that for all $w \geq 0$

$$\begin{aligned}\mathbb{P}_{\sigma'}(|\sigma^* A_{\delta,u}^* A_{\delta,v_1}\sigma' - \sigma^* A_{\delta,u}^* A_{\delta,v_2}\sigma'|& \\ \geq \sqrt{w}\|\sigma^* A_{\delta,u}^*\|_2\ \|A_{\delta,v_1} - A_{\delta,v_2}\|_{\ell_2\to\ell_2}) &\leq 2e^{-w}\end{aligned}$$

and therefore $v \mapsto \sigma^* A_{\delta,u}^* A_{\delta,v}\sigma'$ is a subgaussian process. By Lemma 12 (and (2.8)),

$$\begin{aligned}&(\mathop{\mathbb{E}}_{\sigma'}\sup_{v\in T}|\sigma^* A_{\delta,u}^* A_{\delta,v}\sigma'|^p)^{1/p} \\ &\qquad\lesssim \|A_{\delta,u}\sigma\|_2\gamma_2(T, \|\cdot\|_\delta) + \sqrt{p}\|A_{\delta,u}\sigma\|_2 d_\delta(T).\end{aligned}$$

Taking the $L_p(\Omega_\sigma)$-norm on both sides and using Lemma 28 of Appendix A we find

$$\begin{aligned}(\mathop{\mathbb{E}}_{\sigma} Z^p)^{1/p} &\lesssim (\mathop{\mathbb{E}}_{\sigma}\|A_{\delta,u}\sigma\|_2^p)^{1/p}\Big(\gamma_2(T, \|\cdot\|_\delta) + \sqrt{p}d_\delta(T)\Big) \\ &\lesssim (\sqrt{p}\|A_{\delta,u}\| + \|A_{\delta,u}\|_F)\Big(\gamma_2(T, \|\cdot\|_\delta) + \sqrt{p}d_\delta(T)\Big) \\ &\leq \Big(\sqrt{\frac{p}{s}} + 1\Big)\Big(\gamma_2(T, \|\cdot\|_\delta) + \sqrt{p}d_\delta(T)\Big).\end{aligned}$$

Now take the $L_p(\Omega_\delta)$-norm on both sides to obtain the result. □

6.1. **Unconstrained case.** We first consider unconstrained least squares minimization, i.e., the constraint set is $\mathcal{C} = \mathbb{R}^d$.

Theorem 14. *Let $\mathrm{Col}(A)$ be the column space of A and let*

$$\mu(A) = \max_{1\leq j\leq n}\|P_{\mathrm{Col}(A)}e_j\|_2$$

be its largest leverage score. Set $\mathcal{C} = \mathbb{R}^d$ and let x_* and $\hat{x}$ be minimizers of (6.1) and (6.2), respectively. Let $r(A)$ be the rank of A. Suppose that

$$\begin{aligned}m &\gtrsim \varepsilon^{-1}((\log\eta^{-1})(r(A) + \log m)(\log m)^2 + (\log\eta^{-1})^2 r(A)) \\ s &\gtrsim \varepsilon^{-1}\mu(A)^2((\log\eta^{-1})^2 + (\log\eta^{-1})(\log m)^3) \\ &\qquad + \varepsilon^{-1/2}\mu(A)((\log\eta^{-1})^{3/2} + (\log\eta^{-1})(\log m)^{3/2})\end{aligned}$$

and $\eta \leq \frac{1}{m}$. Then, with probability at least $1 - \eta$,

$$f(\hat{x}) \leq (1 - \varepsilon)^{-2} f(x_*). \tag{6.4}$$

Proof. Set $Z_1 = Z_1(A, \Phi, T_{\mathcal{C}}(x_*))$ and $Z_2 = Z_2(A, \Phi, T_{\mathcal{C}}(x_*), u)$. Observe that $T_{\mathcal{C}}(x_*) = \mathbb{R}^d$. Applying Theorem 5 for the $r(A)$-dimensional subspace $E = \mathrm{Col}(A)$ immediately yields that $Z_1 \geq 1 - \sqrt{\varepsilon}$ with probability at least $1 - \eta$. Setting $T = AT_{\mathcal{C}}(x_*) \cap S^{n-1} = \mathrm{Col}(A) \cap S^{n-1}$ in Lemma 13 and subsequently using (4.2) and (4.3) yields, for any $p \geq \log(m)$,

$$\begin{aligned}(\mathop{\mathbb{E}}_{\delta,\sigma} Z_2^p)^{1/p} &\lesssim \Big(\sqrt{\frac{p}{s}} + 1\Big)\Big((\mathop{\mathbb{E}}_{\delta} \gamma_2^p(T, \|\cdot\|_\delta))^{1/p} + \sqrt{p}(\mathop{\mathbb{E}}_{\delta} d_\delta^p(T))^{1/p}\Big) \\ &\lesssim \Big(\sqrt{\frac{p}{s}} + 1\Big)\Big[\frac{(\sqrt{r(A)} + \log^{1/2}(m))\log(m)}{\sqrt{m}} + \sqrt{\frac{pr(A)}{m}} \\ &\qquad\qquad + \frac{\sqrt{p}\log^{3/2}(m)}{\sqrt{s}}\mu(A) + \frac{p}{\sqrt{s}}\mu(A)\Big].\end{aligned}$$

Applying Lemma 27 (and setting $w = \log(\eta^{-1})$ in (A.1)) shows that $Z_2 \leq \sqrt{\varepsilon}$ with probability at least $1 - \eta$. The second statement in Lemma 11 now implies that with probability at least $1 - 2\eta$,

$$f(\hat{x}) \leq \Big(1 + \frac{\varepsilon}{(1 - \sqrt{\varepsilon})^2}\Big) f(x_*) \leq (1 - \varepsilon)^{-2} f(x_*).$$

To see the last inequality, note that it is equivalent to

$$-\varepsilon + 4\varepsilon^{3/2} - 3\varepsilon^2 - 2\varepsilon^{5/2} + 2\varepsilon^3 \leq 0.$$

Clearly this is satisfied for ε small enough (since then $-\varepsilon$ is the leading term). In fact, one can verify by calculus that this holds for $\varepsilon \leq 1/10$, which we may assume without loss of generality. □

If Φ is a full random sign matrix (i.e., $s = m$) then it follows from our proof that (6.4) holds with probability at least $1 - \eta$ if

$$m \gtrsim \varepsilon^{-1}(r(A) + \log(\eta^{-1})).$$

This bound on m is new, and was also recently shown in [Woo14]. Previous works [Sar06, KN14] allowed either $m \gtrsim \varepsilon^{-2}(r(A) + \log(\eta^{-1}))$ or $m \gtrsim \varepsilon^{-1} r(A) \log(\eta^{-1})$. Theorem 14 substantially improves s while maintaining m up to logarithmic factors.

6.2. $\ell_{2,1}$**-constrained case.** Throughout this section, we set $d = bD$. For $x = (x_{B_1}, \ldots, x_{B_b}) \in \mathbb{R}^d$ consisting of b blocks, each of dimension D, we define its $\ell_{2,1}$-norm by

$$\|x\|_{2,1} := \|x\|_{\ell_1^b(\ell_2^D)} = \sum_{\ell=1}^{b} \|x_{B_\ell}\|_2. \tag{6.5}$$

The corresponding dual norm is

$$\|x\|_{2,\infty} := \|x\|_{\ell_\infty^b(\ell_2^D)} = \max_{1 \leq \ell \leq b} \|x_{B_\ell}\|_2.$$

In this section we study the effect of sketching on the $\ell_{2,1}$-constrained least squares program

$$\min \|Ax - b\|_2^2 \qquad \text{subject to} \qquad \|x\|_{2,1} \leq R,$$

which is Eq. (6.1) for $\mathcal{C} = \{x \in \mathbb{R}^d : \|x\|_{2,1} \leq R\}$. In the statistics literature, this is called the *group Lasso* (with non-overlapping groups of equal size). The $\ell_{2,1}$-constraint encourages a block sparse solution, i.e., a solution which has few blocks containing non-zero entries. We refer to e.g. [BvdG11] for more information. In the special case $D = 1$ the program reduces to

$$\min \|Ax - b\|_2^2 \qquad \text{subject to} \qquad \|x\|_1 \leq R,$$

which is the well-known *Lasso* [Tib96].

To formulate our results we consider two norms on $\mathbb{R}^{n\times d}$, given by

$$|||A||| := \max_{1\leq \ell\leq b}\Big(\sum_{j=1}^n\sum_{k\in B_\ell}|A_{jk}|^2\Big)^{1/2} \tag{6.6}$$

and

$$\|A\|_{\ell_{2,1}\to\ell_\infty} = \max_{1\leq j\leq n}\max_{1\leq \ell\leq b}\|(A_{jk})_{k\in B_\ell}\|_2.$$

Lemma 15. *Let $A\in\mathbb{R}^{n\times d}$ and consider any set $\mathcal{K}\subset\mathbb{R}^d$. If $p\geq \log m$, then*

$$\begin{aligned}&(\mathbb{E}_\delta\gamma_2^{2p}(A\mathcal{K}\cap S^{n-1},\|\cdot\|_\delta))^{1/p}\\ &\qquad\lesssim \alpha\Big[\sup_{x\in\mathcal{K}\,:\,\|Ax\|_2=1}\|x\|_{2,1}^2\Big]\Big(\frac{1}{s}(\log b+p)\|A\|^2_{\ell_{2,1}\to\ell_\infty}+\frac{1}{m}|||A|||^2\Big).\end{aligned}$$

where $\alpha=(\log n)^6(\log m)^2(\log b)^2$.

Proof. We deduce the assertion from Lemma 7. By Hölder's inequality,

$$\sup_{y\in A\mathcal{K}\cap S^{n-1}}\langle x,y\rangle = \sup_{v\in\mathcal{K}\,:\,\|Av\|_2=1}\langle A^*x,v\rangle \leq \|A^*x\|_{2,\infty}\sup_{v\in\mathcal{K}\,:\,\|Av\|_2=1}\|v\|_{2,1}.$$

We define $\|x\|$ to be the expression on the right hand side. Observe that for any sequence $(x_k)_{k\geq1}$ in $\mathbb{R}^n$,

$$\begin{aligned}\mathbb{E}_g\Big\|\sum_{k\geq1}g_kA^*x_k\Big\|_{2,\infty} &= \mathbb{E}_g\max_{1\leq\ell\leq b}\Big\|\sum_{k\geq1}g_k(A^*x_k)_{B_\ell}\Big\|_2\\ &\leq (\log b)^{1/2}\max_{1\leq\ell\leq b}\Big(\sum_{k\geq1}\|(A^*x_k)_{B_\ell}\|_2^2\Big)^{1/2}\\ &\leq (\log b)^{1/2}\Big(\sum_{k\geq1}\|A^*x_k\|^2_{2,\infty}\Big)^{1/2}\end{aligned}$$

and therefore $T_2(\|\cdot\|)\leq(\log b)^{1/2}$. Similarly,

$$\begin{aligned}\mathbb{E}_g\Big\|\sum_{j=1}^n g_j\delta_{ij}A^*e_j\Big\|_{2,\infty} &= \mathbb{E}_g\max_{1\leq\ell\leq b}\Big\|\Big(\sum_{j=1}^n g_j\delta_{ij}A_{jk}\Big)_{k\in B_\ell}\Big\|_2\\ &\leq (\log b)^{1/2}\max_{1\leq\ell\leq b}\Big(\sum_{j=1}^n\delta_{ij}\|(A_{jk})_{k\in B_\ell}\|_2^2\Big)^{1/2}.\end{aligned} \tag{6.7}$$

Finally,

$$d_{\|\cdot\|}\Big(\bigcup_{i=1}^m B_{J_i}\Big) = \sup_{v\in\mathcal{K}\,:\,\|Av\|_2=1}\|v\|_{2,1}\max_{1\leq i\leq m}\sup_{x\in B_{J_i}}\|A^*x\|_{2,\infty}$$

and for any $1\leq i\leq m$ and $x\in B_{J_i}$,

$$\begin{aligned}\|A^*x\|_{2,\infty} &= \max_{1\leq\ell\leq b}\Big\|\sum_{j=1}^n x_j\delta_{ij}(A_{jk})_{k\in B_\ell}\Big\|_2\\ &\leq \max_{1\leq\ell\leq b}\sum_{j=1}^n|x_j|\ \delta_{ij}\|(A_{jk})_{k\in B_\ell}\|_2 \leq \max_{1\leq\ell\leq b}\Big(\sum_{j=1}^n\delta_{ij}\|(A_{jk})_{k\in B_\ell}\|_2^2\Big)^{1/2}.\end{aligned}$$

We now apply Lemma 7 for $T=A\mathcal{K}\cap S^{n-1}$ and subsequently take $L_p(\Omega_\delta)$-norms to conclude that for any $p\geq\log m$,

$$(\mathbb{E}\,\gamma_2^{2p}(A\mathcal{K}\cap S^{n-1},\|\cdot\|_\delta))^{1/p}$$

$$\lesssim \frac{\alpha}{s}\Big[\sup_{x\in\mathcal{K}\,:\,\|Ax\|_2=1}\|x\|_{2,1}^2\Big]\max_{1\le i\le m}\Big(\mathop{\mathbb{E}}_{\delta}\max_{1\le\ell\le b}\Big(\sum_{j=1}^n \delta_{ij}\|(A_{jk})_{k\in B_\ell}\|_2^2\Big)^p\Big)^{1/p}. \tag{6.8}$$

Since $\mathbb{E}\,\delta_{ij}=\frac{s}{m}$ and $(\delta_{ij})_{j=1}^n$ is independent for any fixed i, we obtain by symmetrization (see Eq. (A.4)) for (r_j) a vector of independent Rademachers

$$\begin{aligned}&\Big(\mathop{\mathbb{E}}_{\delta}\max_{1\le\ell\le b}\Big(\sum_{j=1}^n \delta_{ij}\|(A_{jk})_{k\in B_\ell}\|_2^2\Big)^p\Big)^{1/p}\\ &\qquad\le \frac{s}{m}\max_{1\le\ell\le b}\sum_{j=1}^n\|(A_{jk})_{k\in B_\ell}\|_2^2+2\Big(\mathop{\mathbb{E}}_{\delta}\mathop{\mathbb{E}}_{r}\max_{1\le\ell\le b}\Big(\sum_{j=1}^n r_j\delta_{ij}\|(A_{jk})_{k\in B_\ell}\|_2^2\Big)^p\Big)^{1/p}.\end{aligned}$$

By Khintchine's inequality,

$$\begin{aligned}&\Big(\mathop{\mathbb{E}}_{\delta}\mathop{\mathbb{E}}_{r}\max_{1\le\ell\le b}\Big(\sum_{j=1}^n r_j\delta_{ij}\|(A_{jk})_{k\in B_\ell}\|_2^2\Big)^p\Big)^{1/p}\\ &\qquad\lesssim((\log b)^{1/2}+p^{1/2})\Big(\mathop{\mathbb{E}}_{\delta}\max_{1\le\ell\le b}\Big(\sum_{j=1}^n \delta_{ij}\|(A_{jk})_{k\in B_\ell}\|_2^4\Big)^{p/2}\Big)^{1/p}\\ &\qquad\le((\log b)^{1/2}+p^{1/2})\Big(\mathop{\mathbb{E}}_{\delta}\max_{1\le\ell\le b}\Big(\sum_{j=1}^n \delta_{ij}\|(A_{jk})_{k\in B_\ell}\|_2^2\Big)^{p}\Big)^{1/(2p)}\\ &\qquad\qquad\times\max_{1\le\ell\le b}\max_{1\le j\le n}\|(A_{jk})_{k\in B_\ell}\|_2.\end{aligned}$$

Putting the last two displays together we find a quadratic inequality. By solving it, we arrive at

$$\begin{aligned}&\Big(\mathop{\mathbb{E}}_{\delta}\max_{1\le\ell\le b}\Big(\sum_{j=1}^n \delta_{ij}\|(A_{jk})_{k\in B_\ell}\|_2^2\Big)^p\Big)^{1/p}\\ &\qquad\lesssim \frac{s}{m}\max_{1\le\ell\le b}\sum_{j}\|(A_{jk})_{k\in B_\ell}\|_2^2+(\log b+p)\max_{1\le\ell\le b}\max_{1\le j\le n}\|(A_{jk})_{k\in B_\ell}\|_2^2.\end{aligned}\tag{6.9}$$

Combining this estimate with (6.8) yields the assertion. □

Theorem 16. *Let $\mathcal{C}\subset\mathbb{R}^d$ be a closed convex set and let $\alpha=(\log n)^6(\log m)^2(\log b)^2$. Let x_* and $\hat{x}$ be minimizers of (6.1) and (6.2), respectively. Set*

$$d_{2,1}=\sup_{x\in T_{\mathcal{C}}(x_*)\,:\,\|Ax\|_{2,1}=1}\|x\|_{2,1}.$$

Suppose that

$$\begin{aligned}m&\gtrsim\varepsilon^{-2}\log(\eta^{-1})(\alpha+\log(\eta^{-1})\log b)\|\!|A|\!\|^2 d_{2,1}^2,\\ s&\gtrsim\varepsilon^{-2}\log(\eta^{-1})(\log b+\log(\eta^{-1}))(\alpha+\log(\eta^{-1})\log b)\|A\|_{\ell_{2,1}\to\ell_\infty}^2 d_{2,1}^2,\end{aligned}$$

and $\eta\le\frac{1}{m}$. Then, with probability at least $1-\eta$,

$$f(\hat{x})\le(1-\varepsilon)^{-2}f(x_*).$$

Proof. Set $T=AT_{\mathcal{C}}(x_*)\cap S^{n-1}$, $Z_1=Z_1(A,\Phi,T_{\mathcal{C}}(x_*))$ and $Z_2=Z_2(A,\Phi,T_{\mathcal{C}}(x_*),u)$. We apply Eq. (5.3), Eq. (6.7) and Eq. (6.9) to find for any $p\ge\log m$,

$$\begin{aligned}&(\mathbb{E}\,d_\delta^{2p}(T))^{1/p}\\ &\qquad\le\frac{1}{s}\Big[\sup_{x\in T_{\mathcal{C}}(x_*)\,:\,\|Ax\|_{2,1}=1}\|x\|_{2,1}^2\Big]\Big(\mathop{\mathbb{E}}_{\delta}\Big(\max_{1\le i\le m}\mathop{\mathbb{E}}_{g}\Big\|\sum_{j=1}^n g_j\delta_{ij}A^*e_j\Big\|_{2,\infty}\Big)^{2p}\Big)^{1/p}\end{aligned}$$

$$\leq \frac{1}{s}(\log b)\Big[\sup_{x\in T_{\mathcal{C}}(x_*)\,:\,\|Ax\|_{2,1}=1}\|x\|_{2,1}^2\Big]\Big(\mathbb{E}_\delta \max_{1\leq \ell\leq b}\Big(\sum_{j=1}^n \delta_{ij}\|(A_{jk})_{k\in B_\ell}\|_2^2\Big)^p\Big)^{1/p}$$

$$\leq (\log b)\Big[\sup_{x\in T_{\mathcal{C}}(x_*)\,:\,\|Ax\|_{2,1}=1}\|x\|_{2,1}^2\Big]\Big(\frac{1}{m}\|\!|A|\!\|^2 + \frac{1}{s}(\log b + p)\|A\|_{\ell_{2,1}\to\ell_\infty}^2\Big). \tag{6.10}$$

By Eq. (2.10), Lemma 15 and Eq. (6.10)

$$\begin{aligned}
&\Big(\mathbb{E}_{\delta,\sigma}\sup_{x\in T}\Big|\|\Phi x\|_2^2 - \|x\|_2^2\Big|^p\Big)^{1/p}\\
&\lesssim (\mathbb{E}_\delta \gamma_2^{2p}(T,\|\cdot\|_\delta))^{1/p} + (\mathbb{E}_\delta \gamma_2^p(T,\|\cdot\|_\delta))^{1/p} + \sqrt{p}(\mathbb{E}\, d_\delta^p(T))^{1/p} + p(\mathbb{E}\, d_\delta^{2p}(T))^{1/p}\\
&\lesssim \frac{\alpha E}{s}(\log b + p) + \frac{\alpha F}{m} + \Big(\frac{\alpha E}{s}(\log b + p) + \frac{\alpha F}{m}\Big)^{1/2}\\
&\qquad + \frac{Ep\log b}{s}(\log b + p) + \frac{Fp\log b}{m} + \Big(\frac{Ep\log b}{s}(\log b + p) + \frac{Fp\log b}{m}\Big)^{1/2},
\end{aligned}$$

where we write

$$E = \|A\|_{\ell_{2,1}\to\ell_\infty}^2 d_{2,1}^2,\ \ F = \|\!|A|\!\|^2 d_{2,1}^2.$$

Similarly, using Lemma 13 (with $Z(u) := Z_2(A,\Phi,T_{\mathcal{C}}(x_*),u)$) we find

$$\begin{aligned}
(\mathbb{E}_{\delta,\sigma} Z_2^p)^{1/p} &\lesssim \Big(\sqrt{\frac{p}{s}}+1\Big)\Big((\mathbb{E}_\delta \gamma_2^p(T,\|\cdot\|_\delta))^{1/p} + \sqrt{p}(\mathbb{E}_\delta d_\delta^p(T))^{1/p}\Big)\\
&\lesssim \Big(\sqrt{\frac{p}{s}}+1\Big)\Big[\Big(\frac{\alpha E}{s}(\log b + p) + \frac{\alpha F}{m}\Big)^{1/2}\\
&\qquad\qquad + \Big(\frac{Ep\log b}{s}(\log b + p) + \frac{Fp\log b}{m}\Big)^{1/2}\Big].
\end{aligned}$$

Now apply Lemma 27 (with $w = \log(\eta^{-1})$) to conclude that with probability at least $1-\eta$ we have $Z_2 \leq \varepsilon$ and $Z_1 \geq 1-\varepsilon$ under the assumptions on m and s. Lemma 11 now implies that

$$f(\hat{x}) \leq \Big(1 + \frac{\varepsilon}{1-\varepsilon}\Big)^2 f(x_*) = (1-\varepsilon)^{-2} f(x_*).$$

□

Corollary 17. *Set*

$$\mathcal{C} = \{x\in\mathbb{R}^d\ :\ \|x\|_{2,1}\leq R\}.$$

Suppose that x_* is k-block sparse and $\|x_*\|_{2,1} = R$. Define

$$\sigma_{\min,k} = \inf_{\|y\|_2=1,\ \|y\|_{2,1}\leq 2\sqrt{k}} \|Ay\|_2.$$

Set $\alpha = (\log n)^6(\log m)^2(\log b)^2$. Assume that

$$m \gtrsim \varepsilon^{-2}\log(\eta^{-1})(\alpha + \log(\eta^{-1})\log b)\|\!|A|\!\|^2 k\sigma_{\min,k}^{-2},$$

$$s \gtrsim \varepsilon^{-2}\log(\eta^{-1})(\log(\eta^{-1}) + \log b)(\alpha + \log(\eta^{-1})\log b)\|A\|_{\ell_{2,1}\to\ell_\infty}^2 k\sigma_{\min,k}^{-2},$$

and $\eta \leq \frac{1}{m}$. Then with probability at least $1-\eta$

$$f(\hat{x}) \leq (1-\varepsilon)^{-2} f(x^*).$$

Proof. As calculated in Example 33 of Appendix B, every $x \in T_{\mathcal{C}}(x_*)$ satisfies $\|x\|_{2,1}^2 \leq 4k\|x\|_2^2$. If moreover $\|Ax\|_2 = 1$, then

$$\sigma_{\min,k}^2 \leq \|x\|_2^{-2} \leq 4k\|x\|_{2,1}^{-2}.$$

In other words,

$$\sup_{x\in T_{\mathcal{C}}(x_*)\,:\,\|Ax\|_2=1}\|x\|_{2,1}^2 \leq 4k\sigma_{\min,k}^{-2}.$$

□

For a dense random sign matrix Φ, it follows from [PW14, Theorem 1] that the result in Corollary 17 holds under the condition

$$m \gtrsim \varepsilon^{-2}(\|\!|A|\!\|^2 k\sigma_{\min,k}^{-2} + \log(\eta^{-1})).$$

Thus, our result shows that one can substantially increase the sparsity in the matrix Φ, while maintaining the embedding dimension m (up to logarithmic factors).

Let us now specialize our results to the case $D = 1$, which corresponds to the Lasso. In this case, if we let A_k denote the columns of A,

$$\|\!|A|\!\| = \max_{1\le k\le d} \|A_k\|_2,$$

the maximum column of A. We can alternatively interpret this as the norm $\|A\|_{\ell_1\to\ell_2}$. Moreover, the norm $\|A\|_{\ell_{2,1}\to\ell_\infty}$ reduces to

$$\|A\|_{\ell_1\to\ell_\infty} = \max_{1\le j\le n}\max_{1\le k\le d} |A_{jk}|,$$

the largest entry of the matrix. The first norm can be significantly larger than the second. For example, if A is filled with ±1 entries, then

$$\|A\|_{\ell_1\to\ell_2} = \sqrt{n}, \qquad \|A\|_{\ell_1\to\ell_\infty} = 1.$$

A similar result for the fast J-L transform, but with a worse dependence of m on the matrix A, was obtained in [PW14]. For completeness, we will show in Appendix C that the fast J-L transform in fact satisfies similar results as Theorem 16 and Corollary 17, with the essentially the same dependence of m on A.

7. Proof of the main theorem

After having seen instantiations of various subsets of our ideas for specific applications (linear subspaces, $\|\!|\cdot|\!\|_T$ of small type-2 constant, and closed convex cones), we now prove the main theorem of this work, Theorem 3. Our starting point is the observation that inequality Eq. (5.4), i.e.,

$$\begin{aligned}&\log\mathcal{N}(\tilde{T},\|\cdot\|_\delta,t)\\&\quad\lesssim \Big(\log\frac{d_\delta(T)}{t}\Big)(\log m)\log\mathcal{N}\Big(\mathrm{conv}\Big(\bigcup_{i=1}^m B_{J_i}\Big),\|\!|\cdot|\!\|_T,\frac{1}{8}\sqrt{s}t\Big)\end{aligned} \tag{7.1}$$

holds in full generality. We will need the following replacement of Claim 9.

Lemma 18. *Let $\epsilon > 0$. Then*

$$\begin{aligned}&\log\mathcal{N}\Big(\mathrm{conv}\Big(\bigcup_{i=1}^m B_{J_i}\Big),\|\!|\cdot|\!\|_T,\epsilon\Big)\\&\quad\lesssim \frac{1}{\epsilon^2}\log m + (\log 1/\epsilon)\max_{k\lesssim\frac{1}{\epsilon^2}}\max_{|A|=k}\log\mathcal{N}\Big(\frac{1}{k}\sum_{i\in A}B_{J_i},\|\!|\cdot|\!\|_T,\epsilon\Big)\end{aligned} \tag{7.2}$$

Proof. Set $\rho_T(x,y) = \|\!|x-y|\!\|_T$ and let $\rho_{\ell_2}(x,y) = \|x-y\|_2$. By Maurey's lemma (Lemma 31 for $\|\cdot\|_2$), given $x\in\mathrm{conv}(B_{J_i}: 1\le i\le m)$, there is a $k\lesssim 1/\epsilon^2$ and a set $A\subset\{1,\ldots,m\}$, $|A|=k$, such that

$$\rho_T(x,\mathrm{conv}(B_{J_i}; i\in A)) \le \rho_{\ell_2}(x,\mathrm{conv}(B_{J_i}; i\in A)) \le \epsilon \tag{7.3}$$

Now, let us take an element $y\in\mathrm{conv}(B_{J_i}; i\in A)$, $|A|\lesssim 1/\epsilon^2$. Thus

$$y = \sum_{i=1}^k \lambda_i y_i$$

with $y_i \in B_{J_i}$, $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$, $k \lesssim 1/\epsilon^2$. Firstly, we may dismiss the coefficients $\lambda_i < \epsilon^3$. Indeed, let $S = \{i \ : \ \lambda_i < \epsilon^3\}$ and set $\hat{y} = \sum_{i\in S} \lambda_i y_i$. Then,

$$\|\hat{y}\|_2 \leq \sum_{i\in S} \lambda_i \leq k\epsilon^3 \lesssim \epsilon.$$

Consider now the λ_i with $\epsilon^3 \leq \lambda_i \leq 1$. For $\ell = 0, 1, \ldots, \ell_*$, $\ell_* \simeq \log(1/\epsilon)$, denote

$$A_\ell = \Big\{1 \leq i \leq k \ : \ \frac{1}{2^\ell} \geq \lambda_i > \frac{1}{2^{\ell+1}}\Big\}$$

and write

$$y = \sum_{\ell=0}^{\ell_*} 2^{-\ell}|A_\ell| \cdot \Big(\frac{1}{|A_\ell|}\sum_{i\in A_\ell} y_i'\Big) + \hat{y} \tag{7.4}$$

where $y_i' = \lambda_i 2^\ell y_i \in B_{J_i}$. Note that

$$\sum_{\ell=0}^{\ell_*} 2^{-\ell}|A_\ell| < 2\sum_i \lambda_i \leq 2 \tag{7.5}$$

and

$$\frac{1}{|A_\ell|}\sum_{i\in A_\ell} y_i' \in \frac{1}{k_\ell}\sum_{i\in A_\ell} B_{J_i}$$

where $k_\ell = |A_\ell| \lesssim 1/\epsilon^2$. Take finite sets $\xi_\ell \subset \frac{1}{k_\ell}\sum_{i\in A_\ell} B_{J_i}$ such that

$$\rho_T(z, \xi_\ell) < \epsilon \text{ for all } z \in \frac{1}{k_\ell}\sum_{i\in A_\ell} B_{J_i} \tag{7.6}$$

$$|\xi_\ell| = \mathcal{N}\Big(\frac{1}{k_\ell}\sum_{i\in A_\ell} B_{J_i}, \|\!|\!| \cdot |\!|\!\|_T, \epsilon\Big) \tag{7.7}$$

Let $z_\ell \in \xi_\ell$ satisfy

$$\Big|\!\Big|\!\Big| \frac{1}{|A_\ell|}\sum_{i\in A_\ell} y_i' - z_\ell \Big|\!\Big|\!\Big|_T < \epsilon$$

and set

$$z = \sum_{\ell=0}^{\ell_*} 2^{-\ell}|A_\ell| z_\ell. \tag{7.8}$$

By (7.4), (7.5)

$$\|\!|\!| y - z |\!|\!\|_T \lesssim \sum_{\ell=0}^{\ell_*} 2^{-\ell}|A_\ell|\epsilon + \epsilon \lesssim \epsilon$$

To summarize, we can find an ϵ-net for $\mathrm{conv}(\cup_{1\leq i\leq m} B_{J_i})$ with respect to $\|\!|\!| \cdot |\!|\!\|_T$ as follows. For every $A \subset [m]$ with $|A| = k$, $k \lesssim 1/\epsilon^2$, we select $\xi_1^A, \ldots, \xi_{\ell_*}^A$ as above. Then,

$$\bigcup_{k\in[m] \ : \ k\lesssim 1/\epsilon^2} \ \bigcup_{A\subset[m], |A|=k} \ \sum_{\ell=0}^{\ell_*} 2^{-\ell}|A_\ell|\xi_\ell^A$$

is an ϵ-net of cardinality at most

$$\begin{aligned}
&\max_{k\lesssim 1/\epsilon^2} \frac{1}{\epsilon^2} m \binom{m}{k} \prod_{\ell=0}^{\ell_*} \mathcal{N}\Big(\frac{1}{k_\ell}\sum_{i\in A_\ell} B_{J_i}, \|\!|\!| \cdot |\!|\!\|_T, \epsilon\Big) \\
&\quad \lesssim \max_{k\lesssim 1/\epsilon^2} \frac{1}{\epsilon^2} m \Big(\frac{em}{k}\Big)^k \Big[\max_{k\leq 1/\epsilon^2} \max_{|A|=k} \mathcal{N}\Big(\frac{1}{k}\sum_{i\in A} B_{J_i}, \|\!|\!| \cdot |\!|\!\|_T, \epsilon\Big)\Big]^{\log(1/\epsilon)}.
\end{aligned}$$

This yields the result. □

Next, we analyze further the set $(1/k)\sum_{i\in A} B_{J_i}$ for some $k \lesssim 1/\epsilon^2$ ($\epsilon > 0$ will be fixed later). The elements of $(1/k)\sum_{i\in A} B_{J_i}$ are of the form

$$y = \frac{1}{k}\sum_{j=1}^{n}\Big(\sum_{i\in A}\delta_{ij}x_j^{(i)}\Big)e_j$$

with

$$\sum_{j\in J_i}|x_j^{(i)}|^2 \le 1, \qquad \text{for all } i.$$

Therefore, by Cauchy-Schwarz,

$$\begin{aligned}\|y\|_2 &= \frac{1}{k}\Big(\sum_{j=1}^{n}\Big|\sum_{i\in A}\delta_{ij}x_j^{(i)}\Big|^2\Big)^{1/2}\\ &\le \frac{1}{k}\Big[\sum_{j=1}^{n}\Big(\sum_{i\in A}\delta_{ij}\Big)\sum_{i\in A}|x_j^{(i)}|^2\Big]^{1/2}\end{aligned}$$

Define for $\alpha = 1,\dots,\log(\min\{k,s\})$ the set

$$U_\alpha = U_\alpha(\delta) = \Big\{1\le j\le n : 2^\alpha \le \sum_{i\in A}\delta_{ij} < 2^{\alpha+1}\Big\} \tag{7.9}$$

and define

$$U_0 = U_0(\delta) = \Big\{1\le j\le n : \sum_{i\in A}\delta_{ij} < 2\Big\}.$$

Estimate for fixed j

$$\tau_{k,\alpha} \overset{\text{def}}{=} \mathop{\mathbb{P}}_{\delta}\Big(2^\alpha \le \sum_{i\in A}\delta_{ij} < 2^{\alpha+1}\Big) \le \begin{cases} 1, \text{ if } 2^\alpha \le \frac{2esk}{m}\\ \min\Big\{2^{-\alpha}\frac{sk}{m}, 2^{-2^\alpha}\Big\}, \text{ if } 2^\alpha > \frac{2esk}{m},\end{cases} \tag{7.10}$$

where the first term in the min is a consequence of Markov's inequality, and the second term follows by using that

$$\tau_{k,\alpha} \le \mathbb{P}\Big(\sum_{j=1}^{k}\zeta_j \ge 2^\alpha\Big),$$

whenever $\zeta_1,\dots,\zeta_k$ are i.i.d. Bernoulli random variables with mean s/m (see (7.19) for bounding Bernoulli moments). Hence, U_α is a random set of intensity $\tau_{k,\alpha}$, i.e., $\mathbb{P}_\delta(j\in U_\alpha(\delta)) = \tau_{k,\alpha}$. Write according to the preceding

$$y = \sum_\alpha y_\alpha, \text{ with } y_\alpha = \sum_{j\in U_\alpha} y_j e_j$$

and

$$\|y_\alpha\|_2 \lesssim \frac{1}{\sqrt{k}}2^{\alpha/2}$$

Hence, denoting $B_{U_\alpha} := \{\sum_{j\in U_\alpha} x_j e_j \ : \ \sum_{j\in U_\alpha}|x_j|^2 \le 1\}$,

$$\frac{1}{k}\sum_{i\in A} B_{J_i} \subset \sum_\alpha \frac{1}{\sqrt{k}}2^{\alpha/2}B_{U_\alpha}.$$

Therefore, we deduce that

$$\begin{aligned}\log\mathcal{N}\Big(\frac{1}{k}\sum_{i\in A}B_{J_i}, |||\cdot|||_T, \epsilon\Big) &\lesssim \sum_\alpha \log\mathcal{N}\Big(\frac{1}{\sqrt{k}}2^{\alpha/2}B_{U_\alpha}, |||\cdot|||_T, \frac{\epsilon}{\log m}\Big)\\ &= \sum_\alpha \log\mathcal{N}\Big(B_{U_\alpha}, |||\cdot|||_T, \sqrt{\frac{k}{2^\alpha}}\frac{\epsilon}{\log m}\Big)\end{aligned} \tag{7.11}$$

By the dual Sudakov inequality (Lemma 29 of Appendix A),

$$t\Big[\log\mathcal{N}(B_{U_\alpha},|||\cdot|||_T,t)\Big]^{1/2} \lesssim \mathop{\mathbb{E}}_g \Big|\Big|\Big|\sum_{j\in U_\alpha} g_j e_j\Big|\Big|\Big|_T$$

It follows that

$$\Big[\log\mathcal{N}\Big(\frac{1}{k}\sum_{i\in A} B_{J_i},|||\cdot|||_T,\epsilon\Big)\Big]^{1/2} \lesssim \sum_\alpha \Big(\frac{2^\alpha}{k}\Big)^{1/2}\Big(\frac{\log m}{\epsilon}\Big)\mathop{\mathbb{E}}_g \Big|\Big|\Big|\sum_{j\in U_\alpha} g_j e_j\Big|\Big|\Big|_T$$

and, by Lemma 18

$$\begin{aligned}&\Big[\log\mathcal{N}\Big(\mathrm{conv}\Big(\bigcup_{i=1}^m B_{J_i}\Big),|||\cdot|||_T,\epsilon\Big)\Big]^{1/2}\\ &\quad\le \frac{1}{\epsilon}(\log m)^{1/2} + \frac{\log m}{\epsilon}\Big(\log\frac{1}{\epsilon}\Big)^{1/2}\max_{\substack{k\lesssim\frac{1}{\epsilon^2}\\ |A|=k}}\Big[\sum_\alpha\sqrt{\frac{2^\alpha}{k}}\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j\in U_\alpha} g_j e_j\Big|\Big|\Big|_T\Big] \end{aligned} \tag{7.12}$$

Applying Eq. (7.12) to Eq. (7.1), taking $\epsilon \simeq \sqrt{s}t$, gives

$$\begin{aligned} t\log^{1/2}\mathcal{N}(\tilde{T},\|\cdot\|_\delta,t) \lesssim{}& \frac{1}{\sqrt{s}}\Big(\log\frac{1}{\sqrt{s}t}\Big)^{1/2}\log m\\ &+\frac{1}{\sqrt{s}}(\log m)^{3/2}\Big(\log\frac{1}{\sqrt{s}t}\Big)\cdot\\ &\max_{\substack{|A|=k\\ k\lesssim\frac{1}{\epsilon^2}}}\Big\{\sum_\alpha\sqrt{\frac{2^\alpha}{k}}\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j\in U_\alpha} g_j e_j\Big|\Big|\Big|_T\Big\}. \end{aligned}$$

Using this bound for large t and the elementary bound in Eq. (A.6) for small t in Eq. (2.2), we obtain

$$\begin{aligned} \gamma_2(T,\|\cdot\|_\delta) \lesssim{}& \frac{1}{\sqrt{s}}(\log n)^{3/2}\log m\\ &+\frac{1}{\sqrt{s}}(\log m)^{3/2}(\log n)^2\cdot\\ &\max_{k\le m}\max_{|A|=k}\Big\{\sum_{\alpha,2^\alpha\le k}\sqrt{\frac{2^\alpha}{k}}\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j\in U_\alpha} g_j e_j\Big|\Big|\Big|_T\Big\} \end{aligned} \tag{7.13}$$

We split the sum over α into three different parts. Firstly,

$$\begin{aligned}\max_{k\le m}\max_{|A|=k}\sum_{\alpha,2^\alpha\le 2esk/m}\sqrt{\frac{2^\alpha}{k}}\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j\in U_\alpha} g_j e_j\Big|\Big|\Big|_T &\lesssim \sqrt{\frac{s}{m}}\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j=1}^n g_j e_j\Big|\Big|\Big|_T\\ &\lesssim \sqrt{\frac{s}{m}}\gamma_2(T,\|\cdot\|_2). \end{aligned}\tag{7.14}$$

Next, by setting

$$U_A = \Big\{j\in[n]\ :\ \frac{2esk}{m}\le\sum_{i\in A}\delta_{ij}\Big\},$$

we obtain

$$\begin{aligned}&\max_{k\le m}\max_{|A|=k}\sum_{\alpha,2esk/m<2^\alpha\le 10\log n}\sqrt{\frac{2^\alpha}{k}}\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j\in U_\alpha} g_j e_j\Big|\Big|\Big|_T\\ &\quad\lesssim(\log n)^{1/2}\max_{k\le m}\max_{|A|=k}\frac{1}{\sqrt{k}}\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j\in U_A} g_j e_j\Big|\Big|\Big|_T\end{aligned}$$

$$\le (\log n)^{1/2}\Big[\max_{m/s\le k\le m}\max_{|A|=k}\frac{1}{\sqrt{k}}\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j\in U_A} g_j e_j\Big|\Big|\Big|_T + \max_{k\le m/s}\max_{|A|=k}\frac{1}{\sqrt{k}}\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j\in U_A} g_j e_j\Big|\Big|\Big|_T\Big].$$

Note that the first term on the right hand side is bounded by

$$\max_{m/s\le k\le m}\max_{|A|=k}\frac{1}{\sqrt{k}}\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j\in U_A} g_j e_j\Big|\Big|\Big|_T \lesssim \sqrt{\frac{s}{m}}\gamma_2(T,\|\cdot\|_2). \tag{7.15}$$

To bound the second term, we take expectations with respect to δ and find for $p_k = k\log\left(\frac{m}{sk}\right)$

$$\mathop{\mathbb{E}}_\delta\Big[\max_{k\le m/s}\max_{|A|=k}\frac{1}{\sqrt{k}}\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j\in U_A} g_j e_j\Big|\Big|\Big|_T\Big] \lesssim \max_{k\le m/s}\max_{|A|=k}\frac{1}{\sqrt{k}}\Big(\mathop{\mathbb{E}}_\delta\Big(\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j\in U_A} g_j e_j\Big|\Big|\Big|_T\Big)^{p_k}\Big)^{1/p_k}. \tag{7.16}$$

By Eq. (7.10), for any $A\subset[m]$ with $|A|=k$, U_A is a random set of intensity at most $esk/m = sq/(m\log s)$, where we set $q = ek(\log s)$. If we now let $\eta_1,\ldots,\eta_n$ be i.i.d. $\{0,1\}$-valued random variables with expectation $qs/(m\log s)$, then the random set U defined by $\mathbb{P}(j\in U) = \mathbb{P}(\eta_j = 1)$ has a higher intensity than U_A and therefore,

$$\mathop{\mathbb{E}}_\delta\Big[\max_{k\le m/s}\max_{|A|=k}\frac{1}{\sqrt{k}}\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j\in U_A} g_j e_j\Big|\Big|\Big|_T\Big] \lesssim (\log s)^{1/2}\max_{q\le\frac{m}{s}\log s}\frac{1}{\sqrt{q}}\Big(\mathop{\mathbb{E}}_\eta\Big(\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j=1}^n \eta_j g_j e_j\Big|\Big|\Big|_T\Big)^q\Big)^{1/q}. \tag{7.17}$$

Finally, consider the α with $2^\alpha > 10\log n$. Since $|||\cdot|||_T\le\|\cdot\|_2$,

$$\max_{k\le m}\max_{|A|=k}\sum_{\alpha,10\log n<2^\alpha\le k}\sqrt{\frac{2^\alpha}{k}}\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j\in U_\alpha} g_j e_j\Big|\Big|\Big|_T \le \max_{k\le m}\max_{|A|=k}\sum_{\alpha,10\log n<2^\alpha\le k}\sqrt{\frac{2^\alpha}{k}}|U_\alpha|^{1/2}. \tag{7.18}$$

By Eq. (7.10), the intensity of U_α is bounded by 2^{-2^α} and therefore,

$$\mathop{\mathbb{E}}_\delta\Big[\max_{k\le m}\max_{|A|=k}\sum_{\alpha,10\log n<2^\alpha\le k}\sqrt{\frac{2^\alpha}{k}}\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j\in U_\alpha} g_j e_j\Big|\Big|\Big|_T\Big] \le \max_{k\le m}\max_{|A|=k}\sum_{\alpha,10\log n<2^\alpha\le k}\sqrt{\frac{2^\alpha}{k}}\Big(\mathbb{E}\Big|\sum_{j=1}^n\zeta_j\Big|^{q_k}\Big)^{1/(2q_k)},$$

where $q_k\simeq k\log(m/k)$ and the ζ_j are i.i.d. $\{0,1\}$-valued with mean 2^{-2^α}. Since $\sum_j\zeta_j$ is binomially distributed, we find using that $2^\alpha > 10\log n$,

$$\Big\|\sum_j\zeta_j\Big\|_{L_\zeta^{q_k}} \le \Big(\sum_{t=1}^n t^{q_k}e^{-\frac{9}{10}2^\alpha t}\Big)^{1/q_k} \lesssim n^{1/q_k}2^{-\alpha}q_k \lesssim 2^{-\alpha}k(\log m) \tag{7.19}$$

and therefore

$$\mathop{\mathbb{E}}_\delta\Big[\max_{k\le m}\max_{|A|=k}\sum_{\alpha,10\log n<2^\alpha\le k}\sqrt{\frac{2^\alpha}{k}}\mathop{\mathbb{E}}_g\Big|\Big|\Big|\sum_{j\in U_\alpha} g_j e_j\Big|\Big|\Big|_T\Big] \lesssim (\log m)^{3/2}. \tag{7.20}$$

By combining Eq. (7.13), Eq. (7.14), Eq. (7.15), Eq. (7.17), and Eq. (7.20) we obtain

$$\begin{aligned}
&\mathbb{E}\,\gamma_2(T, \|\cdot\|_\delta) \\
&\quad\lesssim \frac{(\log m)^{3/2}(\log n)^{5/2}\gamma_2(T, \|\cdot\|_2)}{\sqrt{m}} + \frac{1}{\sqrt{s}}\Bigg\{(\log m)^3(\log n)^2 + \\
&\qquad (\log m)^2(\log n)^{5/2} \max_{q\leq \frac{m}{s}\log s} \frac{1}{\sqrt{q}}\Big(\mathop{\mathbb{E}}_{\eta}\Big(\mathop{\mathbb{E}}_{g}\Big\|\Big|\sum_{j=1}^n \eta_j g_j e_j\Big\|\Big|_T\Big)^q\Big)^{1/q}\Bigg\},
\end{aligned} \tag{7.21}$$

with η_j i.i.d. Bernoulli with mean $qs/(m\log s)$.

From our proof it is clear that $\mathbb{E}\,\gamma_2^2(T, \|\cdot\|_\delta)$ can be bounded by the square of the right hand side of Eq. (7.21). Using Eq. (2.10) (for $p = 1$), we conclude that

$$\mathbb{E}\sup_{x\in T}\Big|\|\Phi x\|_2^2 - 1\Big| < \varepsilon$$

provided that m, s satisfy

$$m \gtrsim (\log m)^3(\log n)^5\frac{\gamma_2^2(T, \|\cdot\|_2)}{\varepsilon^2} \tag{7.22}$$

$$s \gtrsim (\log m)^6(\log n)^4\frac{1}{\varepsilon^2} \tag{7.23}$$

and

$$(7.24) = (\log m)^2(\log n)^{5/2} \max_{q\leq \frac{m}{s}\log s}\Big\{\frac{1}{\sqrt{qs}}\Big(\mathop{\mathbb{E}}_{\eta}\Big(\mathop{\mathbb{E}}_{g}\Big\|\Big|\sum_{j=1}^n \eta_j g_j e_j\Big\|\Big|_T\Big)^q\Big)^{1/q}\Big\} < \varepsilon. \tag{7.24}$$

This completes the proof.

8. Example applications of main theorem

In this section we use our main theorem to give explicit conditions under which

$$\mathbb{E}\sup_{x\in T}\Big|\|\Phi x\|_2^2 - 1\Big| < \varepsilon$$

for several interesting sets $T \subset S^{n-1}$. This amounts to computing an upper bound for the parameters $\gamma_2(T, \|\cdot\|_2)$ and

$$\kappa(T) = \max_{q\leq \frac{m}{s}\log s}\Big\{\frac{1}{\sqrt{qs}}\Big(\mathop{\mathbb{E}}_{\eta}\Big(\mathop{\mathbb{E}}_{g}\Big\|\Big|\sum_{j=1}^n \eta_j g_j e_j\Big\|\Big|_T\Big)^q\Big)^{1/q}\Big\}. \tag{8.1}$$

We focus on the latter two and refer to [Dir14] for details on how to estimate $\gamma_2(T, \|\cdot\|_2)$. Note, however, that $\gamma_2(T, \|\cdot\|_2) \lesssim (m\log s)^{1/2}\kappa(T)$. Indeed, take $q = \frac{m}{s}\log s$ in Eq. (8.1) and note that the η_j are identically equal to 1 in this case. This gives

$$\kappa(T) \geq (m\log s)^{-1/2}g(T) \simeq (m\log s)^{-1/2}\gamma_2(T, \|\cdot\|_2).$$

Thus, if we ignore logarithmic factors, it suffices to bound $\kappa(T)$.

8.1. **Linear subspace.** In the application from Section 4 with T the unit ball of a d-dimensional linear subspace E of ℓ_2^n, we have $\gamma_2(T, \|\cdot\|_2) \simeq \sqrt{d}$ and

$$(7.24) \lesssim (\log m)^2 (\log n)^{5/2} \frac{1}{\sqrt{s}} \max_{q \le \frac{m}{s} \log s} \frac{1}{\sqrt{q}} \Big\| \sum_j \|P_E e_j\|_2^2 \eta_j \Big\|_{L^q_\eta}^{1/2} \tag{8.2}$$

with $\eta_j \in \{0,1\}$ i.i.d. of mean $qs/(m \log s)$. Using Khintchine's inequality,

$$\Big\| \sum_j \eta_j \|P_E e_j\|_2^2 \Big\|_{L^q_\eta} < \frac{dqs}{m \log s} + q \max_j \|P_E e_j\|_2^2 \tag{8.3}$$

and therefore

$$(7.24) \lesssim (\log m)^2 (\log n)^{5/2} \Big(\sqrt{\frac{d}{m}} + \frac{1}{\sqrt{s}} \max_j \|P_E e_j\|_2 \Big) \tag{8.4}$$

Eq. (7.22) and Eq. (7.23) then give conditions

$$m \gtrsim d (\log n)^5 (\log m)^4 / \varepsilon^2 \tag{8.5}$$

$$s \gtrsim (\log n)^4 (\log m)^6 / \varepsilon^2 + (\log n)^5 (\log m)^4 \max_j \|P_E e_j\|_2^2 / \varepsilon^2 \tag{8.6}$$

8.2. **k-sparse vectors.** Consider next the application from Section 5.1, replacing T by K given by Eq. (5.14). Thus $\gamma_2(K) \lesssim \sqrt{k \log n}$ and Eq. (7.24) is bounded by

$$\sqrt{k} (\log m)^2 (\log n)^3 \frac{1}{\sqrt{s}} \max_{q \le \frac{m}{s} \log s} \Big\{ \frac{1}{\sqrt{q}} \Big\| \max_i \Big(\sum_j \eta_j |A_{ij}|^2 \Big) \Big\|_{L^q_\eta}^{1/2} \Big\} \tag{8.7}$$

and

$$\Big\| \max_i (\sum_j \eta_j |A_{ij}|^2) \Big\|_{L^q_\eta} \lesssim \frac{qs}{m \log s} + q \max_{i,j} |A_{ij}|^2 \tag{8.8}$$

Hence

$$(7.24) \lesssim (\log m)^2 (\log n)^3 \sqrt{\frac{k}{m}} + (\log m)^2 (\log n)^3 \sqrt{\frac{k}{s}} \max_{i,j} |A_{ij}| \tag{8.9}$$

We then arrive at the conditions

$$m \gtrsim k (\log n)^6 (\log m)^4 / \varepsilon^2 \tag{8.10}$$

$$s \gtrsim (\log n)^4 (\log m)^6 / \varepsilon^2 \tag{8.11}$$

provided that

$$\max_{i,j} |A_{ij}| < k^{-1/2} (\log n)^{-1} \tag{8.12}$$

(compare with Eq. (5.13)-(5.12)).

8.3. **Flat vectors.** Let $T \subseteq S^{n-1}$ be finite with $\|x\|_\infty \le \alpha$ for all $x \in T$. Then by a similar calculation as in (4.10),

$$\Big(\mathbb{E}_\eta \Big(\mathbb{E}_g \sup_{x \in T} \Big| \sum_{i=1}^n \eta_i g_i x_i \Big| \Big)^q \Big)^{1/q} \lesssim \sqrt{\log |T|} \cdot \Big\| \sup_{x \in T} \sum_{i=1}^n \eta_i x_i^2 \Big\|_{L^q_\eta}^{1/2} \tag{8.13}$$

$$\lesssim \sqrt{\frac{qs \log |T|}{m}} + (\alpha \log |T|) \sqrt{q} \tag{8.14}$$

Since $\gamma_2(T, \|\cdot\|_2) \lesssim \sqrt{\log |T|}$ we find the conditions

$$m \gtrsim (\log |T|)(\log m)^4 (\log n)^5 / \varepsilon^2 \tag{8.15}$$

$$s \gtrsim (\alpha \log |T|)^2 (\log m)^4 (\log n)^5 / \varepsilon^2 + (\log m)^6 (\log n)^4 / \varepsilon^2, \tag{8.16}$$

which is qualitatively similar to [Mat08, Theorem 4.1].

8.4. **Finite collection of subspaces.** Let Θ be a finite collection of d-dimensional subspaces $E \subset \mathbb{R}^n$ with $|\Theta| = N$. Define

$$T = \bigcup_{E\in\Theta} \{x \in E : \|x\|_2 = 1\}.$$

In this case, $\gamma_2(T, \|\cdot\|_2) \lesssim \sqrt{d} + \sqrt{\log N}$. For the duration of the next two sections, we define

$$\alpha = \sup_{E\in\Theta} \max_j \|P_E e_j\|_2.$$

Recall that $\max_j \|P_E e_j\|_2$ is referred to as the *incoherence* of E, and thus $\sqrt{d/n} \leq \alpha \leq 1$ (cf. Remark 6) is the maximum incoherence in Θ.

8.4.1. *Collection of incoherent subspaces.* To estimate $\kappa(T)$, consider the collection $\mathcal{A}$ of operators $A = \sum_j (P_E e_j) \otimes e_j$, $E \in \Theta$. Fix η and define $R_\eta x = \sum_j \eta_j x_j e_j$. Then applying [KMR14, Theorem 3.5] to $\mathcal{A}R_\eta$,

$$\begin{aligned} \Big\|\max_{E\in\Theta} \|\sum_j \eta_j g_j P_E e_j\|\Big\|_{L^1_g} &= \Big\|\max_{A\in\mathcal{A}} \|AR_\eta g\|\Big\|_{L^1_g} \\ &\lesssim d_F(\mathcal{A}R_\eta) + \gamma_2(\mathcal{A}R_\eta, \|\cdot\|). \end{aligned} \tag{8.17}$$

Clearly

$$\|A\| \leq 1,\ \|AR_\eta\|_F = \Big(\sum_j \eta_j \|P_E e_j\|_2^2\Big)^{1/2}.$$

By [RV07] (see the formulation in [Tro08, Proposition 7]),

$$\Big(\mathbb{E}_\eta \|AR_\eta\|^p\Big)^{1/p} \leq 3\sqrt{p}\Big(\mathbb{E}\|AR_\eta\|_{1,2}^p\Big)^{1/p} + \sqrt{\varrho}\cdot\|A\|, \tag{8.18}$$

where $\mathbb{E}\,\eta_j = \varrho$ and we assume $p \geq 2\log n$. Here $\|\cdot\|_{1,2}$ is the $\ell_1^n \to \ell_n^2$ norm, hence

$$\|AR_\eta\|_{1,2} = \max_{1\leq k\leq n} \|AR_\eta e_k\|_2 = \max_{1\leq k\leq n} \eta_k \|P_E e_k\|_2 \leq \alpha.$$

Taking $\varrho = \frac{qs}{m\log s}$, it follows that

$$\Big(\mathbb{E}_\eta \|AR_\eta\|^p\Big)^{1/p} \lesssim \alpha\sqrt{p} + \Big(\frac{qs}{m\log s}\Big)^{1/2} \tag{8.19}$$

To estimate $\kappa(T)$, we need to bound

$$\|(8.17)\|_{L^q_\eta} \leq \underbrace{\Big\|\max_E \Big(\sum_j \eta_j\|P_E e_j\|_2^2\Big)\Big\|_{L^q_\eta}^{1/2}}_{\sqrt{(8.20)}} \tag{8.20}$$

$$+ \underbrace{\|\gamma_2(\mathcal{A}R_\eta)\|_{L^q_\eta}}_{(8.21)} \tag{8.21}$$

First, by Eq. (8.3) and denoting $q_1 = q + \log N$,

$$(8.20) \lesssim \varrho d + q_1 \max_j \|P_E e_j\|_2^2 = \frac{dqs}{m\log s} + (q + \log N)\alpha^2 \tag{8.22}$$

Estimate trivially

$$\gamma_2(\mathcal{A}R_\eta) \leq \sqrt{\log N}\cdot\max_{A\in\mathcal{A}} \|AR_\eta\|$$

and hence, applying Eq. (8.19) with $p = q + 2\log n + \log N$

$$\begin{aligned} (8.21) &\lesssim \sqrt{\log N}\cdot\Big(\max_A \mathbb{E}\|AR_\eta\|^p\Big)^{1/p} \\ &\lesssim \sqrt{\log N}\cdot\Big\{(q + \log n + \log N)\alpha^2 + \frac{qs}{m\log s}\Big\}^{1/2} \end{aligned} \tag{8.23}$$

Collecting our estimates we find

$$\begin{aligned}\kappa(T) &\lesssim \max_{q\le \frac{m}{s}\log s} \frac{1}{\sqrt{qs}}\Big\{\Big(\frac{dqs}{m\log s}\Big)^{1/2} + \alpha\sqrt{q} \\ &\qquad + \sqrt{\log N}\cdot\Big[(\sqrt{q}+\sqrt{\log n}+\sqrt{\log N})\alpha + \Big(\frac{qs}{m\log s}\Big)^{1/2}\Big]\Big\} \\ &\lesssim \Big(\frac{d+\log N}{m\log s}\Big)^{1/2} + \frac{(\sqrt{\log n}\sqrt{\log N}+\log N)\alpha}{\sqrt{s}} \qquad (8.24)\end{aligned}$$

Hence in this application (assuming $\log N \ge \log n$),

$$(7.24) < (\log m)^2(\log n)^{5/2}\Big\{\Big(\frac{d+\log N}{m}\Big)^{1/2} + \frac{\alpha\log N}{\sqrt{s}}\Big\} \qquad (8.25)$$

and the conditions on m, s become

$$m \gtrsim (\log m)^4(\log n)^5(d+\log N)\varepsilon^{-2} \qquad (8.26)$$

$$s \gtrsim ((\log m)^6(\log n)^4 + (\alpha\log N)^2(\log m)^4(\log n)^5)\varepsilon^{-2} \qquad (8.27)$$

Notice that s depends *only* on $|\Theta|$ and not on the dimension d. Thus this bound is of interest when $\log|\Theta|$ is small compared with d.

8.4.2. *Collection of coherent subspaces.* We can also obtain a bound on s that does not improve for small α, but has linear dependence on $\log N$. Here we will not rely on [RV07]. As described in Section 1, this setting has applications to model-based compressed sensing. For example, for approximate recovery of block-sparse signals using the notation of Section 1, our bounds will show that a measurement matrix Φ may have $m \lesssim_* kb + k\log(n/k)$, $s \lesssim_* k\log(n/k)$ and allow for efficient recovery. This is non-trivial if the number of blocks satisfies $b \gtrsim_* \log(n/k)$.

One may indeed trivially bound

$$\gamma_2(\mathcal{A}R_\eta) \lesssim \sqrt{\log N}$$

since certainly $\|AR_\eta\| \le \|A\| \le 1$. Hence

$$(8.21) \lesssim \sqrt{\log N}$$

leading to the following bound on $\kappa(T)$

$$\max_{q\le\frac{m}{s}\log s}\frac{1}{\sqrt{qs}}\Big\{\sqrt{\frac{dqs}{m\log s}} + (\sqrt{q}+\sqrt{\log N})\alpha + \sqrt{\log N}\Big\} \lesssim \sqrt{\frac{d}{m}} + \sqrt{\frac{\log N}{s}}$$

Instead of (8.26), (8.27), one may impose the conditions

$$m \gtrsim (\log m)^4(\log n)^5\frac{d}{\varepsilon^2} + (\log m)^3(\log n)^5\frac{\log N}{\varepsilon^2} \qquad (8.28)$$

$$s \gtrsim (\log m)^6(\log n)^4\frac{1}{\varepsilon^2} + (\log m)^4(\log n)^5\frac{\log N}{\varepsilon^2} \qquad (8.29)$$

We remark that previous work [NN13a] which achieved $m \approx d/\varepsilon^2$ for small s had worse dependence on $\log N$: in particular $s \gtrsim (\log N)^3, m \gtrsim (\log N)^6$. In fact, Conjecture 14 of [NN13a] if true would imply that the correct dependence on N in both m and s should be $\log N$ (which is optimal due to known lower bounds for the Johnson-Lindenstrauss lemma, i.e., the special case $d = 1$). We thus have shown that this implication is indeed true.

8.5. **Possibly infinite collection of subspaces.** Assume next Θ is an arbitrary family of linear d-dimensional subspaces $E \subset \mathbb{R}^n$, equipped with the Finsler metric

$$\rho_{\mathrm{Fin}}(E, E') = \|P_E - P'_E\|.$$

Let

$$T = \bigcup_{E \in \Theta} \{x \in E : \|x\|_2 = 1\}$$

for which (cf. [Dir14])

$$\gamma_2(T, \|\cdot\|_2) \lesssim \sqrt{d} + \gamma_2(\Theta, \rho_{\mathrm{Fin}}). \tag{8.30}$$

Fix some parameter $\varepsilon_0 > 0$ and let $\Theta_1 \subset \Theta$ be a finite subset such that

$$|\Theta_1| \leq \mathcal{N}(\Theta, \rho_{\mathrm{Fin}}, \varepsilon_0) \tag{8.31}$$

$$\rho_{\mathrm{Fin}}(E, \Theta_1) \leq \varepsilon_0 \text{ for any } E \in \Theta \tag{8.32}$$

Let further

$$T_1 = \bigcup_{E' \in \Theta_1} \{x \in E' : \|x\|_2 = 1\}$$

Let $x \in T, x \in E, E \in \Theta$ and $\rho_{\mathrm{Fin}}(E, E') \leq \varepsilon_0$ for some $E' \in \Theta_1$. Hence

$$x = P_{E'}x + (P_E x - P_{E'}x) = x_1 + x_2 \tag{8.33}$$

where $x_1 \in T_1$ and $x_2 \in B_E + B_{E'}$, $\|x_2\|_2 \leq \varepsilon_0$. Hence $x_2 \in T_2$ with

$$T_2 = \bigcup_{E, F \in \Theta} \{x \in B_E + B_F : \|x\|_2 \leq \varepsilon_0\}$$

For $t < \varepsilon_0$, we estimate $\mathcal{N}(T_2, \|\cdot\|_2, t)$. Let $\Theta_t \subset \Theta$ satisfy

$$|\Theta_t| \leq \mathcal{N}\Big(\Theta, \rho_{\mathrm{Fin}}, \frac{t}{4}\Big) \tag{8.34}$$

$$\rho_{\mathrm{Fin}}(E, \Theta_t) \leq \frac{t}{4} \text{ for all } E \in \Theta \tag{8.35}$$

By Eq. (A.6), for each $E' \in \Theta_t$ we can find $\xi_{E'} \subset B_{E'}$ such that

$$\log |\xi_{E'}| \lesssim d \log \frac{1}{t} \tag{8.36}$$

$$\rho_{\ell_2}(x, \xi_{E'}) \leq t \text{ for all } x \in B_{E'} \tag{8.37}$$

Denote

$$\xi_t = \bigcup_{E', F' \in \Theta_t} (\xi_{E'} - \xi_{F'})$$

for which by construction

$$\log |\xi_t| \lesssim \log \mathcal{N}\Big(\Theta, \rho_{\mathrm{Fin}}, \frac{t}{4}\Big) + d \log \frac{1}{t}$$

Also, for $x \in T_2$, $x = y + z \in B_E + B_F$ and $E', F' \in \Theta_t$ satisfying

$$\rho_{\mathrm{Fin}}(E, E') \leq \frac{t}{4}, \rho_{\mathrm{Fin}}(F, F') \leq \frac{t}{4},$$

we have

$$\|x - (P_{E'}y + P_{F'}z)\|_2 \leq \frac{t}{2},$$

while

$$\rho_{\ell_2}(P_{E'}y, \xi_{E'}) \leq \frac{t}{4}, \qquad \rho_{\ell_2}(P_{F'}z, \xi_{F'}) \leq \frac{t}{4}$$

Therefore

$$\rho_{\ell_2}(x, \xi_t) \leq t$$

and we get for $t < \varepsilon_0$ (otherwise (8.38) $= 0$)

$$\log \mathcal{N}(T_2, \|\cdot\|_2, t) \lesssim \log \mathcal{N}\Big(\Theta, \rho_{\mathrm{Fin}}, \frac{t}{4}\Big) + d \log \frac{1}{t} \tag{8.38}$$

Using the decomposition Eq. (8.33) and the bound Eq. (8.25) for the contribution of T_1, we find

$$\kappa(T) \lesssim \Big\{ \Big(\frac{d + \log \mathcal{N}(\Theta, \rho_{\mathrm{Fin}}, \varepsilon_0)}{m} \Big)^{1/2} + \frac{\alpha \log \mathcal{N}(\Theta, \rho_{\mathrm{Fin}}, \varepsilon_0)}{\sqrt{s}} \Big\} \tag{8.39}$$

$$+ \max_{q \leq \frac{m}{s} \log s} \Big\{ \frac{1}{\sqrt{qs}} \Big(\mathop{\mathbb{E}}_{\eta} \Big(\mathop{\mathbb{E}}_{g} \sup_{x \in T_2} \Big| \sum_{j=1}^n \eta_j g_j x_j \Big| \Big)^q \Big)^{1/q} \Big\} \tag{8.40}$$

with $(\eta_j) \in \{0, 1\}$ i.i.d. of mean $\frac{qs}{m \log s}$.

Estimate by the contraction principle [Kah68], the Gaussian concentration for Lipschitz functions (Eq. (4.1)), and Dudley's inequality Eq. (2.2)[Dud67]

$$\begin{aligned}
\frac{1}{\sqrt{q}} \Big\| \sup_{x \in T_2} \Big| \sum_{j=1}^n \eta_j g_j x_j \Big| \Big\|_{L_g^1, L_\eta^q} &\leq \frac{1}{\sqrt{q}} \Big\| \sup_{x \in T_2} \Big| \sum_{j=1}^n g_j x_j \Big| \Big\|_{L_g^q} \\
&\lesssim \Big\| \sup_{x \in T_2} \Big| \sum_{j=1}^n g_j x_j \Big| \Big\|_{L_g^1} \\
&\lesssim \int_0^{\varepsilon_0} [\log \mathcal{N}(T_2, \|\cdot\|_2, t)]^{1/2} dt \\
&\lesssim \int_0^{\varepsilon_0} [\log \mathcal{N}(\Theta, \rho_{\mathrm{Fin}}, t)]^{1/2} dt + \sqrt{d} \varepsilon_0 \sqrt{\log \frac{1}{\varepsilon_0}},
\end{aligned} \tag{8.41}$$

where in the final step we used Eq. (8.38).

Summarizing, the conditions on m and s are as follows (for any $\varepsilon_0 > 0$)

$$\begin{aligned}
m &\gtrsim \varepsilon^{-2} (\log m)^3 (\log n)^5 \gamma_2^2(\Theta, \rho_{\mathrm{Fin}}) \\
&\quad + \varepsilon^{-2} (\log m)^4 (\log n)^5 [d + \log \mathcal{N}(\Theta, \rho_{\mathrm{Fin}}, \varepsilon_0)]
\end{aligned} \tag{8.42}$$

$$\begin{aligned}
s &\gtrsim \varepsilon^{-2} (\log m)^6 (\log n)^4 + \varepsilon^{-2} (\log m)^4 (\log n)^5 [\alpha \log \mathcal{N}(\Theta, \rho_{\mathrm{Fin}}, \varepsilon_0)]^2 \\
&\quad + \varepsilon^{-2} (\log m)^4 (\log n)^5 \varepsilon_0^2 \Big(\log \frac{1}{\varepsilon_0} \Big) d \\
&\quad + \varepsilon^{-2} (\log m)^4 (\log n)^5 \Big[\int_0^{\varepsilon_0} [\log \mathcal{N}(\Theta, \rho_{\mathrm{Fin}}, t)]^{1/2} dt \Big]^2
\end{aligned} \tag{8.43}$$

If $|\Theta| = N < \infty$, then $\log \mathcal{N}(\Theta, \rho_{\mathrm{Fin}}, t) \leq \log N$ and (8.42), (8.43) turn into (8.26), (8.27) for $\varepsilon_0 \to 0$.

8.6. **Manifolds.** Let $\mathcal{M} \subset \mathbb{R}^n$ be a d-dimensional manifold obtained as the image $\mathcal{M} = F(B_{\ell_2^d})$, for a smooth map $F : B_{\ell_2^d} \to \mathbb{R}^n$. More precisely, we assume that $\|F(x) - F(y)\|_2 \simeq \|x - y\|_2$ and the *Gauss map*, which sends $x \in \mathcal{M}$ to E_x, the tangent plane at x, is Lipschitz from the geodesic distance $\rho_{\mathcal{M}}$ to ρ_{Fin}. Following [Dir14], we want to ensure that the sparse matrix Φ satisfies

$$(1 - \varepsilon)|\gamma| \leq |\Phi(\gamma)| \leq (1 + \varepsilon)|\gamma| \tag{8.44}$$

for any C^1-curve $\gamma \subset \mathcal{M}$, where $|\cdot|$ denotes curve length. Note that Eq. (8.44) is equivalent to requiring

$$1 - \varepsilon \leq \|\Phi(v)\|_2 \leq 1 + \varepsilon \tag{8.45}$$

for any tangent vector v of $\mathcal{M}$ at a point $x \in \mathcal{M}$. Denote by

$$\Theta = \{E_x : x \in \mathcal{M}\}$$

the tangent bundle of $\mathcal{M}$, to which we apply the estimates on m, s obtained above. By assumption,

$$\rho_{\mathrm{Fin}}(E_x, E_y) \lesssim \rho_{\mathcal{M}}(x, y) \simeq \|F^{-1}(x) - F^{-1}(y)\|_2,$$

so that for $0 \le t \le 1/2$ by Eq. (A.6),

$$\log \mathcal{N}(\Theta, \rho_{\mathrm{Fin}}, t) \lesssim \log \mathcal{N}(\mathcal{M}, \|\cdot\|_2, ct) \simeq \log \mathcal{N}(B_{\ell_2^d}, \|\cdot\|_2, t) \lesssim d \log \frac{1}{t}.$$

In this application

$$\alpha = \max_{x\in\mathcal{M}} \max_{\substack{v\in E_x\\ \|v\|_2=1}} \max_{1\le j\le n} |\langle v, e_j\rangle| \tag{8.46}$$

We assume

$$\alpha \ll \frac{1}{\sqrt{d}} \tag{8.47}$$

to make the below of interest (otherwise apply the result of [Dir14]). Taking $\varepsilon_0 = \alpha\sqrt{d}$ in (8.42), (8.43), it follows that Eq. (8.44) may be ensured under parameter conditions

$$m \gtrsim \varepsilon^{-2}(\log m)^4(\log n)^5 d \cdot \log\Big(\frac{1}{\alpha\sqrt{d}}\Big) \tag{8.48}$$

$$s \gtrsim \varepsilon^{-2}(\log m)^6(\log n)^4 + \varepsilon^{-2}(\log m)^4(\log n)^7(\alpha d)^2 \tag{8.49}$$

Thus for $\alpha = o(1/\sqrt{d})$, the condition on s becomes non-trivial. Recall from Remark 6 that $\alpha \ge \sqrt{d/n}$ and therefore the $\log(1/(\alpha\sqrt{d}))$ term in Eq. (8.48) is at most $\log n$.

Remark 19. Returning to the assumptions on F, consider $DF : B_{\ell_2^d} \to \mathcal{L}(\mathbb{R}^d, \mathbb{R}^n)$, the space of linear operators from $\mathbb{R}^d$ to $\mathbb{R}^n$. The first statement means that uniformly for $x \in B_{\ell_2^d}$,

$$c^{-1}\|\xi\|_2 \le \|DF(x)\xi\|_2 \le c\|\xi\|_2 \text{ for } \xi \in \mathbb{R}^d \tag{8.50}$$

The second statement follows from requiring

$$\|DF(x) - DF(y)\|_{2\to 2} \le c\|x-y\|_2 \text{ for } x, y \in B_{\ell_2^d} \tag{8.51}$$

Indeed, since $E_x = DF(x)(\mathbb{R}^d), E_y = DF(y)(\mathbb{R}^d)$, it follows from Eq. (8.51)

$$\inf_{\substack{u\in E_x\\ \|u\|_2=1}} \rho_{\ell_2^n}(u, E_y) \lesssim \|x-y\|_2$$

implying

$$\rho_{\mathrm{Fin}}(E_x, E_y) = \|P_{E_x} - P_{E_y}\|_{2\to 2} \lesssim \|x-y\|_2.$$

Remark 20. If $\alpha \ge 1/\sqrt{d}$, then necessarily $s \gg d$ in Eq. (8.49) (up to polylogarithmic factors). Thus the manifold case is very different from the case of linear subspaces. We sketch a construction to demonstrate this.

Let $n > d^{10}$. Denote by $0 \le \varphi \le 1$ a smooth bump function on $\mathbb{R}$ such that

$$\varphi(t) = t \text{ for } \frac{1}{4} \le t \le \frac{1}{2},\ \mathrm{supp}(\varphi) \subset [0,1] \tag{8.52}$$

By the lower bound in Eq. (A.6), there is a collection $\{a_\beta\}_{1\le\beta\le 2^d}$ of 2^d points in $B_{\ell_2^d}(0,1/2) \overset{\mathrm{def}}{=} B_{1/2} \subset \mathbb{R}^d$ such that

$$\|a_{\beta'} - a_\beta\|_2 > \frac{1}{10} \text{ for } \beta' \ne \beta$$

and let $(\eta_\beta)_{1\le\beta\le 2^d}$ be any collection of unit vectors in $\mathbb{R}^n$. Consider the map $f:\mathbb{R}^d\to\mathbb{R}^n$ defined by

$$f(x)=\sum_{\beta=1}^{2^d}\varphi(10^4\|x-a_\beta\|_2^2)\eta_\beta \tag{8.53}$$

Thus by construction, the summands in Eq. (8.53) are disjointly supported functions of x. Clearly

$$Df(x)\xi=2\cdot 10^4\sum_\beta \varphi'(10^4\|x-a_\beta\|_2^2)\,\langle x-a_\beta,\xi\rangle\,\eta_\beta \tag{8.54}$$

implying that

$$\|Df(x)\|_{2\to 2}\le C$$

and

$$\|Df(x)-Df(y)\|_{2\to 2}\le C\|x-y\|_2 \tag{8.55}$$

Next, let $\theta_1,\dots,\theta_d\in\mathbb{R}^n$ be orthogonal vectors such that

$$\|\theta_j\|_2=1,\|\theta_j\|_\infty=\frac{1}{\sqrt{n}} \tag{8.56}$$

and define the map

$$F:B_{\ell_2^d}\subset\mathbb{R}^d\to\mathbb{R}^{2n}=\mathbb{R}^n\times\mathbb{R}^n$$
$$F(x)=\Big(\sum_{j=1}^d x_j\theta_j, f(x)\Big)$$

In view of Eq. (8.55), F satisfies conditions (8.50),(8.51). Also, by (8.54),(8.56)

$$\alpha\lesssim\sup_{x\in B_{\ell_2^d}}\max_{\|\xi\|_2=1}\|DF(x)\xi\|_\infty\lesssim\sqrt{\frac{d}{n}}+\max_\beta\|\eta_\beta\|_\infty \tag{8.57}$$

Let $\mathcal{M}=F(B_{\ell_2^d})$. Fix some $1\le\beta\le 2^d$ and let for $1/200\le t\le 1/(100\sqrt{2})$

$$\gamma(t)=F(a_\beta+te_1)=\Big(\sum_{j=1}^d a_{\beta,j}\theta_j+t\theta_1,10^4t^2\eta_\beta\Big)$$

where we used Eq. (8.52). Thus γ is a C^1-curve in $\mathcal{M}$ and

$$\gamma'\Big(\frac{1}{200}\Big)=(\theta_1,100\eta_\beta) \tag{8.58}$$

Let Φ be a sparse $m\times 2n$ matrix for which Eq. (8.44) holds. Then Φ has to satisfy in particular

$$(1-\varepsilon)(1+10^4)^{1/2}\le\|\Phi(\theta_1,100\eta_\beta)\|_2\le(1+\varepsilon)(1+10^4)^{1/2}$$

and hence

$$\|\Phi(\theta_1,100\eta_\beta)\|_2<10^3 \text{ for all } 1\le\beta\le 2^d \tag{8.59}$$

Choose k such that

$$2^d>\binom{n}{k}2^k,\ \text{i.e. } k\simeq\frac{d}{\log n}$$

and let (η_β) be the collection of $2^k\binom{n}{k}$ vectors in $\mathbb{R}^n$ of the form

$$\eta=\frac{1}{\sqrt{k}}\sum_{j\in I}\pm e_j \text{ with } I\subset\{1,\dots,n\},|I|=k \tag{8.60}$$

By Eq. (8.59), Φ needs to satisfy

$$\|\Phi(0,\eta)\|_2 \le 20 \tag{8.61}$$

for all vectors η of the form Eq. (8.60). But if $m < n/d$, Eq. (8.61) implies that $s \ge k/400 \gtrsim d/\log n$. On the other hand, Eq. (8.57) gives

$$\alpha \lesssim \sqrt{\frac{d}{n}} + \frac{1}{\sqrt{k}} \simeq \sqrt{\frac{\log n}{d}}$$

8.6.1. *Geodesic distances.* In this section we show that not only do sparse maps preserve curve lengths on manifolds, but in fact they preserve geodesic distances.

Lemma 21. *Fix $x \in \mathbb{R}^n$ such that $|x_j| \ge 1$ for $j \in J \subset \{1,\dots,n\}, |J| = r$. Then*

$$\Big|\Big\{1 \le i \le m : \sum \Phi_{ij}^2 x_j^2 \ge \frac{1}{s}\Big\}\Big| > \min\Big\{\frac{sr}{3}, cm\Big\} \stackrel{\text{def}}{=} r_1 \tag{8.62}$$

with probability (with respect to Φ) at least

$$1 - 2^{-sr} \tag{8.63}$$

Proof. Fix $I \subset \{1,\dots,m\}$, $|I| \le r_1$ and assume $\sum_j \Phi_{ij}^2 x_j^2 < 1/s$ for $i \notin I$. This means that for any $j \in J$, $S = \stackrel{\text{def}}{=} \{i; \Phi_{ij} \ne 0\} \subset I$. The probability (with respect to Φ) that $\{i; \Phi_{ij} \ne 0\} \subset I$ (with j fixed) is (by Eq. (2.3))

$$\begin{aligned}
\sum_{\substack{K \subset I \\ |K|=s}} \mathbb{P}_{\Phi}(S = K) &= \sum_{\substack{K \subset I \\ |K|=s}} \mathbb{E} \prod_{i \in K} \delta_{ij} \\
&\le \sum_{\substack{K \subset I \\ |K|=s}} \Big(\frac{s}{m}\Big)^s \\
&\le \binom{|I|}{s} \cdot \Big(\frac{s}{m}\Big)^s \\
&\le \Big(\frac{e \cdot |I|}{m}\Big)^s
\end{aligned}$$

Since for different j the events are independent, it follows the probability that

$$\Big\{1 \le i \le m; \sum \Phi_{ij}^2 x_j^2 \ge \frac{1}{s}\Big\} \subset I$$

is bounded by $((e|I|/m)^s)^r \le (er_1/m)^{sr}$. Taking a union bound over all $I \subset \{1,\dots,m\}, |I| \le r_1$ gives by the choice of r_1

$$\binom{m}{|I|} \cdot \Big(\frac{er_1}{m}\Big)^{sr} \le \Big(\frac{em}{r_1}\Big)^{r_1} \Big(\frac{er_1}{m}\Big)^{sr} \le \Big(\frac{e^2 r_1}{m}\Big)^{2sr/3} < 2^{-sr}$$

□

For the following lemma recall that for $a \in \mathbb{R}^n$ and (σ_j) a Rademacher vector

$$\mathbb{P}_{\sigma}\Big(\Big|\sum_j a_j \sigma_j\Big| < \frac{1}{2}\|a\|_2\Big) < \frac{4}{5}. \tag{8.64}$$

This is a consequence of the Paley-Zygmund inequality, see e.g. [Ber97, Theorem 3.6].

Lemma 22. *Let $x \in \mathbb{R}^n$ satisfy the assumption of Lemma 21. Then*

$$\mathbb{P}_{\Phi}\Big(\|\Phi x\|_2 < \frac{1}{2\sqrt{s}}\Big) < 2^{-cr_1} \tag{8.65}$$

Proof. We may assume that $\sum \Phi_{ij}^2 x_j^2 \geq 1/s$ for i in a subset $I \subset \{1, \dots, m\}$, $|I| \geq r_1$. Exploiting the random signs of Φ_{ij}, we find by Eq. (8.64) for each $i \in I$

$$\mathbb{P}\Big(\Big|\sum_j \Phi_{ij} x_j\Big| < \frac{1}{2\sqrt{s}}\Big) < \frac{4}{5} \tag{8.66}$$

If $\|\Phi x\|_2 < 1/(2\sqrt{s})$, then $|\sum_j \Phi_{ij} x_j| < 1/(2\sqrt{s})$ for all i, in particular for all $i \in I$. By Eq. (8.66) the probability for this event is at most

$$\Big(\frac{4}{5}\Big)^{|I|} < 2^{-cr_1},$$

proving Eq. (8.65). □

Corollary 23. *Let $\xi \subset \mathbb{R}^n$ be a set of vectors x with following properties:*

(a) *Each $x \in \xi$ has a decomposition $x = x' + x''$, $x' \in \xi'$ and there is a set $J = J_{x'} \subset \{1, \dots, n\}$, $|J| \geq r$ so that $|x'_j| \geq 1$ for $j \in J$. Moreover*

$$\|x''\|_2 < \frac{1}{10n} \tag{8.67}$$

(b)

$$|\xi'| < 2^{c \min\{sr, m\}} \tag{8.68}$$

Then

$$\|\Phi x\|_2 > \frac{1}{4\sqrt{s}} \text{ for all } x \in \xi \tag{8.69}$$

with probability at least $1 - 2^{-c\min\{sr,m\}}$.

Proof. Write

$$\|\Phi x\|_2 \geq \|\Phi x'\|_2 - \|\Phi x''\|_2 \geq \|\Phi x'\|_2 - \sqrt{n} \cdot \|x''\|_2 \geq \|\Phi x'\|_2 - \frac{1}{10\sqrt{n}}.$$

since clearly $\|\Phi\|_2 \leq \sqrt{n}$. Next, Lemma 22 and a union bound will ensure that $\|\Phi x'\|_2 \geq 1/(2\sqrt{s})$ for all $x' \in \xi'$ with the desired probability. □

Lemma 24. *Let $s \geq c(\log n)^2$. Let $\xi \subset \mathbb{R}^n$ be a finite set of unit vectors satisfying*

$$\log |\xi| < cm \tag{8.70}$$

Then with probability at least $1 - e^{-cs}$ for some constant $c > 0$, Φ satisfies

$$\|\Phi x\| > e^{-c(\log n)^2} \text{ for } x \in \xi. \tag{8.71}$$

Proof. Given $x \in \mathbb{R}^n$, let x^* be the decreasing rearrangement of $|x_i|$. Let $K = c\log n$. We partition ξ as

$$\xi = \bigcup_\ell (\xi_\ell \backslash (\xi_0 \cup \xi_1 \cup \ldots \cup \xi_{\ell-1}))$$

where $\xi_{-1} = \emptyset$, and for $\ell \geq 0$ satisfying $2^\ell K^2 < m$,

$$\xi_\ell = \{x \in \xi : x^*_{m/(2^\ell K^2)} > n^{-K+2\ell}\}$$

For the vectors $x \in \xi_0$, apply Corollary 23 with $x = x', r = m/K^2$ (after rescaling by n^{-K}). Since $csr > cm > \log|\xi| \geq \log|\xi_0|$, (8.68) holds and by Eq. (8.69), we ensure that $\|\Phi x\|_2 > n^{-K}/(4\sqrt{s})$ for all $x \in \xi_0$.

Consider the set $\xi'_\ell = \xi_\ell \backslash (\xi_0 \cup \xi_1 \cup \ldots \cup \xi_{\ell-1})$. Thus each $x \in \xi'_\ell$ has a decomposition

$$x = y + z$$

where y is obtained by considering the $m/(2^{\ell-1}K^2)$ largest coordinates of x and $\|z\|_\infty \leq x^*_{m/(2^{\ell-1}K^2)} \leq n^{-K+2\ell-2}$ since $x \notin \xi_{\ell-1}$.

Note also that

$$y \in \bigcup_{\substack{S\subset\{1,\dots,n\}\\ |S|=\frac{m}{2^{\ell-1}K^2}}} B_{X_S}, \text{ with } X_S = \{e_j : j \in S\}.$$

Let $\mathfrak{f}_S \subset B_{X_S}$ be a finite subset such that

$$\begin{aligned} &\operatorname{dist}(y, \mathfrak{f}_S) < n^{-K} \text{ for all } y \in B_{X_S} \\ &\log|\mathfrak{f}_S| \lesssim \frac{m}{2^{\ell-1}K^2}\log n^K = \frac{m\log n}{2^{\ell-1}K} \end{aligned} \tag{8.72}$$

Hence $y = x' + w$ with

$$x' \in \bigcup_S \mathfrak{f}_S,\ \|w\|_2 < n^{-K} \tag{8.73}$$

Apply Corollary 23 to the set $n^{K-2\ell}\xi'_\ell$, considering the decomposition

$$n^{K-2\ell}x = n^{K-2\ell}x' + n^{K-2\ell}(w+z)$$

satisfying by Eq. (8.73)

$$\begin{aligned} \|n^{K-2\ell}(w+z)\|_2 &< n^{-2\ell} + n^{K-2\ell}\|z\|_\infty \\ &< n^{-2\ell} + n^{-2} < \frac{1}{10n} \end{aligned}$$

which is condition (8.67).

Also $n^{K-2\ell}x' \in \xi'$, where by Eq. (8.72) (and choice of s, K)

$$\log|\xi'| \le \log\binom{n}{m/(2^{\ell-1}K^2)} + \frac{m\log n}{2^{\ell-1}K} \lesssim \frac{m\log n}{2^{\ell-1}K} < c\min\Big\{\frac{sm}{2^\ell K^2}, m\Big\},$$

which is condition (8.68) with $r = m/(2^\ell K^2)$.

Let J_x be the set of $r = m/(2^\ell K^2)$ largest coordinates of x (which are also the r largest coordinates of y). Hence, for $j \in J_x$

$$\begin{aligned} n^{K-2\ell}|x'_j| &\ge n^{K-2\ell}|y_j| - n^{K-2\ell}|w_j| \\ &= n^{K-2\ell}|x_j| - n^{K-2\ell}|w_j| \\ &> n^{K-2\ell}n^{-K+2\ell} - n^{K-2\ell}n^{-K} > \frac{1}{2} \end{aligned}$$

since $x \in \xi_\ell$. By Eq. (8.69)

$$\|\Phi x\| > n^{-K+2\ell}\cdot\frac{1}{4\sqrt{s}} \text{ for all } x \in \xi_\ell$$

with probability at least $1 - e^{-c\min\{m,sr\}} \ge 1 - e^{-cs}$, since $m/(2^\ell K^2) \ge 1$. ☐

Let $\mathcal{M} \subset \mathbb{R}^n$ be a d-dimensional manifold obtained as a bi-Lipschitz image of the unit ball $B_{\ell_2^d}$, i.e. $F : B_{\ell_2^d} \to \mathcal{M}$ satisfies

$$\|F(x) - F(y)\|_2 \simeq \|x - y\|_2. \tag{8.74}$$

We assume moreover F is smooth, more specifically $DF : B_{\ell_2^d} \to \mathcal{L}(\mathbb{R}^d, \mathbb{R}^n)$ (i.e. linear maps from $\mathbb{R}^d$ to $\mathbb{R}^n$ under operator norm) is Lipschitz, i.e.,

$$\|DF(x) - DF(y)\|_{\ell_2^d\to\ell_2^n} \lesssim \|x - y\|_2 \tag{8.75}$$

Lemma 25. *Let $\mathcal{M}$ be as above. Assume*

$$\begin{aligned} m &\gtrsim d(\log n)^2 \\ s &\gtrsim (\log n)^2 \end{aligned} \tag{8.76}$$

Then with probability at least $1-e^{-cs}$ *for some constant* $c>0$, $\Phi|_{\mathcal{M}}$ *is bi-Lipschitz, and more specifically*

$$\|\Phi(x)-\Phi(y)\|_2 \geq e^{-c(\log n)^2}\cdot\|x-y\|_2 \text{ for } x,y\in\mathcal{M} \tag{8.77}$$

Proof. We treat separately the pairs $x,y\in\mathcal{M}$ which are at a "large" and "small" distance from each other. Fix $\varepsilon_1>0$ and $\varepsilon_2>\varepsilon_1$ to be specified later. Let $A_{\varepsilon_1}\subset\mathcal{M}$ be an ε_1-net for $\mathcal{M}$. By Eq. (8.74) and Eq. (A.6), we can assume that

$$\log|A_{\varepsilon_1}|\lesssim d\log(1/\varepsilon_1).$$

Assume that $x,y\in\mathcal{M}$, $\|x-y\|_2>\varepsilon_2$. Take $x_1,y_1\in A_{\varepsilon_1}$ s.t. $\|x-x_1\|_2<\varepsilon_1$, $\|y-y_1\|_2<\varepsilon_1$. Since Φ is linear

$$\|\Phi(x)-\Phi(y)\|_2\geq\|\Phi(x_1-y_1)\|_2-2\sqrt{n}\varepsilon_1$$

Apply Lemma 24 to the set

$$\xi=\Big\{\frac{x_1-y_1}{\|x_1-y_1\|_2}:x_1,y_1\in A_{\varepsilon_1}\Big\}.$$

Assuming

$$m\gtrsim d\log(1/\varepsilon_1)>c\log|A_{\varepsilon_1}| \tag{8.78}$$

we ensure Φ to satisfy

$$\Big\|\Phi\Big(\frac{x_1-y_1}{\|x_1-y_1\|_2}\Big)\Big\|_2>e^{-c(\log n)^2}\text{ for }x_1,y_1\in A_{\varepsilon_1}.$$

Therefore, if $x,y\in\mathcal{M}$, $\|x-y\|_2>\varepsilon_2$

$$\begin{aligned}\|\Phi(x)-\Phi(y)\|_2&\geq e^{-c(\log n)^2}\cdot\|x_1-y_1\|_2-2\sqrt{n}\frac{\varepsilon_1}{\varepsilon_2}\cdot\|x-y\|_2\\&\geq\frac{1}{2}e^{-c(\log n)^2}\cdot\|x-y\|_2\end{aligned} \tag{8.79}$$

choosing $\varepsilon_2=5\sqrt{n}e^{c(\log n)^2}\varepsilon_1$. This takes care of large distances.

In order to deal with small distances, we first ensure that

$$\|\Phi(v)\|>e^{-c(\log n)^2}\text{ for any unit tangent vector }v\text{ of }\mathcal{M} \tag{8.80}$$

Consider

$$\mathcal{T}_{\varepsilon_1}=\bigcup_{x_1\in A_{\varepsilon_1}}\Big\{v\in T_{x_1}:\|v\|=1\Big\} \tag{8.81}$$

Under the assumption (8.78) on m (and making an ε_1-discretization of each $\{v\in T_{x_1}:\|v\|_2=1\}$), another application of Lemma 24 will ensure that (8.80) holds for all $v\in\mathcal{T}_{\varepsilon_1}$. Next, if $x\in\mathcal{M}$, $x_1\in A_{\varepsilon_1}$, $\|x-x_1\|_2<\varepsilon_1$, it follows from Eq. (8.75) that $\|DF(x)-DF(x_1)\|_{\ell_2\to\ell_2}\lesssim\varepsilon_1$ and hence $\rho_{\mathrm{Fin}}(T_x,T_{x_1})\lesssim\varepsilon_1$. Therefore, if $v\in T_x$, $\|v\|_2=1$, there is $v_1\in\mathcal{T}_{\varepsilon_1}$ s.t. $\|v-v_1\|_2\lesssim\varepsilon_1$. Therefore

$$\|\Phi(v)\|_2\geq\|\Phi(v_1)\|_2-\sqrt{n}\varepsilon_1\geq e^{-c(\log n)^2}-\sqrt{n}\varepsilon_1>\frac{1}{2}e^{-c(\log n)^2}$$

since $\varepsilon_1<n^{-1/2}e^{-c(\log n)^2}/2$. Thus Φ satisfies (8.80).

Assume $x,y\in\mathcal{M}$, $\|x-y\|_2<\varepsilon_2$. Let $u=F^{-1}(x),w=F^{-1}(y)\in B_{\ell_2^d}$. By Eq. (8.74), $\|u-w\|_2\simeq\|x-y\|_2$. Write

$$\begin{aligned}y-x=F(w)-F(u)&=\int_0^1\partial_tF(tw+(1-t)u)dt\\&=\int_0^1 DF(tw+(1-t)u)(w-u)dt.\end{aligned}$$

Hence, again invoking Eq. (8.75),

$$\|F(w)-F(u)-DF(u)(w-u)\|_2$$

$$\le \|w-u\|_2 \cdot \sup_t \|DF(tw+(1-t)u) - DF(u)\|_{\ell_2\to\ell_2}$$
$$\lesssim \|u-w\|_2^2. \tag{8.82}$$

Denote
$$v = DF(u)\Big(\frac{w-u}{\|w-u\|_2}\Big) \in T_u.$$

By Eq. (8.82) and using $\|\Phi\| \le \sqrt{n}$,
$$\|y-x-\|u-w\|_2 v\|_2 < c\varepsilon_2\|x-y\|_2$$
$$\|\Phi(y)-\Phi(x)-\|u-w\|_2\Phi(v)\|_2 < c\sqrt{n}\varepsilon_2\|x-y\|_2$$

and by (8.80)
$$\begin{aligned}\|\Phi(x)-\Phi(y)\|_2 &\ge \|u-w\|_2 e^{-c(\log n)^2} - c\sqrt{n}\varepsilon_2\|x-y\|_2\\ &\ge \Big(ce^{-c(\log n)^2} - c\sqrt{n}\varepsilon_2\Big)\|x-y\|_2\\ &\gtrsim e^{-c(\log n)^2}\|x-y\|_2\end{aligned}$$

provided
$$\varepsilon_2 < c\frac{1}{\sqrt{n}}e^{-c(\log n)^2} \tag{8.83}$$

Thus we may take $\log(1/\varepsilon_1), \log(1/\varepsilon_2) \simeq (\log n)^2$ and condition (8.78) will hold for
$$m \gtrsim d(\log n)^2.$$

$\square$

Theorem 26. *Let $\mathcal{M}$ be as above and m, s satisfy* (8.76)*. Assume moreover that m, s satisfy the appropriate conditions to ensure that*
$$1-\varepsilon \le \|\Phi(v)\|_2 \le 1+\varepsilon \tag{8.84}$$
for all unit tangent vectors v of $\mathcal{M}$. Then with probability at least $1-e^{-cs}$ for some constant $c > 0$, Φ preserves geodesic distances up to factor $1+\varepsilon$.

Proof. By (8.84), $(1-\varepsilon)|\gamma| < |\Phi(\gamma)| < (1+\varepsilon)|\gamma|$ for any C^1-curve γ in $\mathcal{M}$. Let $\rho_{\mathcal{M}}$ refer to the geodesic distance in $\mathcal{M}$. Clearly
$$\rho_{\Phi(\mathcal{M})}(\Phi(x),\Phi(y)) \le (1+\varepsilon)\rho_{\mathcal{M}}(x,y) \tag{8.85}$$
from the above. We need to show the reverse inequality. Let γ_1 be a C^1-curve in $\Phi(\mathcal{M})$ joining $\Phi(x), \Phi(y)$ such that $\rho_{\Phi(\mathcal{M})}(\Phi(x),\Phi(y)) = |\gamma_1|$. Since Φ satisfies (8.77), $\Phi|_{\mathcal{M}} : \mathcal{M} \to \Phi(\mathcal{M})$ is bi-Lipschitz and hence a diffeomorphism (since Φ is also smooth). It follows that $\gamma = \Phi^{-1}\gamma_1$ is a C^1-curve joining x and y and
$$\rho_{\mathcal{M}}(x,y) \le |\gamma| \le (1+\varepsilon)|\Phi(\gamma)| = (1+\varepsilon)|\gamma_1| = (1+\varepsilon)\rho_{\Phi(\mathcal{M})}(\Phi(x),\Phi(y)).$$

$\square$

9. Discussion

We have provided a general theorem which captures sparse dimensionality reduction in Euclidean space and qualitatively unifies much of what we know in specific applications. There is still much room though for quantitative improvement. We here list some known quantitative shortcomings of our bounds and discuss some avenues for improvement in future work.

First, our dependence on $1/\varepsilon$ in s in all our theorems is, up to logarithmic factors, quadratic. Meanwhile the works [KN14, NN13a] show that the correct dependence in the case of small m should be linear. Part of the reason for this discrepancy may be our use of the chaining result of [KMR14]. Specifically, chaining is a technique that in general converts tail bounds into bounds on the expected supremum of stochastic processes (see [Tal05] for details). Perhaps one could generalize the tail bound of [KN14] to give a broad understanding of the decay behavior for the error random variable in the SJLT as a function of s, m then feed such a quantity into a chaining argument to improve our use of [KMR14].

Another place where we lost logarithmic factors is in our use of the duality of entropy numbers [BPSTJ89, Proposition 4] in Eq. (5.4). It is believed that in general if K, D are symmetric convex bodies and $\mathcal{N}(K, D)$ is the number of translations of D needed to cover K then

$$\mathcal{N}(D^\circ, aK^\circ) \lesssim \mathcal{N}(K, D) \lesssim \mathcal{N}(D^\circ, a^{-1}K^\circ)$$

for some universal constant $a > 0$. This is known as Pietsch's duality conjecture [Pie72, p. 38], and unfortunately resolving it has been a challenging open problem in convex geometry for over 40 years.

We have also lost logarithmic factors in our use of dual Sudakov minoration, which bounds $\sup_{t>0} t[\log \mathcal{N}(B_E, \|\cdot\|_X, t)]^{1/2}$ by a constant times $\mathbb{E}_g \|P_E g\|_X$ for any norm $\|\cdot\|_X$ even though what we actually wish to bound is the γ_2 functional. Passing from this functional via Eq. (2.2) to $\sup_{t>0} t[\log \mathcal{N}(B_E, \|\cdot\|_X, t)]^{1/2}$ costs us logarithmic factors. The majorizing measures theory (see [Tal05]) shows that the loss can be avoided when $\|\cdot\|_X$ is the ℓ_2 norm; it would be interesting to explore how the loss can be avoided in our case.

Finally, note that the SJLT Φ considered in this work is a randomly signed adjacency matrix of a random bipartite graph with m vertices in the left bipartition, n in the right, and with all right vertices having equal degree s. Sasha Sodin has asked in personal communication whether taking a random signing of the adjacency matrix of a random biregular graph (degree s on the right and degree ns/m on the left) can yield improved bounds on s, m for some T of interest. We leave this as an interesting open question.

Acknowledgement. We would like to thank Huy L. Nguyễn for pointing out the second statement in Lemma 11. Part of this work was done during a visit of the second-named author to Harvard University. He wishes to thank the Theory of Computation group for its hospitality.

References

[ABTZ13] Haim Avron, Christos Boutsidis, Sivan Toledo, and Anastasios Zouzias. Efficient dimensionality reduction for canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

[AC09] Nir Ailon and Bernard Chazelle. The Fast Johnson–Lindenstrauss Transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009.

[Ach03] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

[ADR14] Ulaş Ayaz, Sjoerd Dirksen, and Holger Rauhut. Uniform recovery of fusion frame structured sparse signals. *CoRR*, abs/1407.7680, 2014.

[AHK06] Sanjeev Arora, Elad Hazan, and Satyen Kale. A fast random sampling algorithm for sparsifying matrices. In *APPROX-RANDOM*, pages 272–279, 2006.

[AL09] Nir Ailon and Edo Liberty. Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete Comput. Geom.*, 42(4):615–630, 2009.

[AL13] Nir Ailon and Edo Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. *ACM Transactions on Algorithms*, 9(3):21, 2013.

[Alo03] Noga Alon. Problems and results in extremal combinatorics–i. *Discrete Mathematics*, 273(1-3):31–53, 2003.

[AMT10] Haim Avron, Petar Maymounkov, and Sivan Toledo. Blendenpik: Supercharging LAPACK's least-squares solver. *SIAM J. Scientific Computing*, 32(3):1217–1236, 2010.

[AN13] Alexandr Andoni and Huy L. Nguyễn. Eigenvalues of a matrix in the streaming model. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1729–1737, 2013.

[BB05] Maria-Florina Balcan and Avrim Blum. A pac-style model for learning from labeled and unlabeled data. In *Learning Theory, 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005, Proceedings*, pages 111–126, 2005.

[BBDS12] Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. The johnson-lindenstrauss transform itself preserves differential privacy. In *53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2012, New Brunswick, NJ, USA, October 20-23, 2012*, pages 410–419, 2012.

[BBV06] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. Kernels as features: On kernels, margins, and low-dimensional mappings. *Machine Learning*, 65(1):79–94, 2006.

[BCDH10] Richard G. Baraniuk, Volkan Cevher, Marco F. Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Trans. Inf. Theory*, 56:1982–2001, 2010.

[BCIS05] Mihai Badoiu, Julia Chuzhoy, Piotr Indyk, and Anastasios Sidiropoulos. Low-distortion embeddings of general metrics into the line. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing, Baltimore, MD, USA, May 22-24, 2005*, pages 225–233, 2005.

[BD08] Thomas Blumensath and Mike E. Davies. Iterative hard thresholding for compressed sensing. *J. Fourier Anal. Appl.*, 14:629–654, 2008.

[BDF+11] Jean Bourgain, Stephen Dilworth, Kevin Ford, Sergei Konyagin, and Denka Kutzarova. Explicit constructions of RIP matrices and related problems. *Duke J. Math.*, 159(1):145–185, 2011.

[Ber97] Bonnie Berger. The fourth moment method. *SIAM J. Comput.*, 26(4):1188–1207, 1997.

[BLM89] Jean Bourgain, Joram Lindenstrauss, and Vitali D. Milman. Approximation of zonoids by zonotopes. *Acta Math.*, 162:73–141, 1989.

[BOR10] Vladimir Braverman, Rafail Ostrovsky, and Yuval Rabani. Rademacher chaos, random Eulerian graphs and the sparse Johnson-Lindenstrauss transform. *CoRR*, abs/1011.2590, 2010.

[BPSTJ89] Jean Bourgain, Alain Pajor, Stanisław J. Szarek, and Nicole Tomczak-Jaegermann. On the duality problem for entropy numbers of operators. *Geometric Aspects of Functional Analysis*, 1376:50–163, 1989.

[BT02] Jeremy Buhler and Martin Tompa. Finding motifs using random projections. *Journal of Computational Biology*, 9(2):225–242, 2002.

[BvdG11] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer, Heidelberg, 2011.

[BW09] Richard G. Baraniuk and Michael B. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, 2009.

[BZMD11] Christos Boutsidis, Anastasios Zouzias, Michael W. Mahoney, and Petros Drineas. Stochastic dimensionality reduction for k-means clustering. *CoRR*, abs/1110.2897, 2011.

[Can08] Emmanuel Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendes Mathematique*, 346 (9-10): 589–592, 2008.

[Car85] Bernd Carl. Inequalities of Bernstein-Jackson-type and the degree of compactness operators in Banach spaces. *Ann. Inst. Fourier (Grenoble)*, 35(3):79–118, 1985.

[CCFC04] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3–15, 2004.

[CDMI+13] Kenneth L. Clarkson, Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, Xiangrui Meng, and David P. Woodruff. The Fast Cauchy Transform and faster robust linear regression. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 466–477, 2013.

[CEM+14] Michael Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Mădălina Persu. Dimensionality reduction for k-means clustering and low rank approximation. *CoRR*, abs/1410.6801, 2014.

[CGV13] Mahdi Cheraghchi, Venkatesan Guruswami, and Ameya Velingker. Restricted isometry of Fourier matrices and list decodability of random linear codes. *SIAM J. Comput.*, 42(5):1888–1914, 2013.

[Cla08] Kenneth L. Clarkson. Tighter bounds for random projections of manifolds. In *Proceedings of the 24th ACM Symposium on Computational Geometry, College Park, MD, USA, June 9-11, 2008*, pages 39–48, 2008.

[CM12] Pedro Contreras and Fionn Murtagh. Fast, linear time hierarchical clustering using the Baire metric. *J. Classification*, 29(2):118–143, 2012.

[Coh14] Michael B. Cohen. Personal communication, 2014.

[CT65] James W. Cooley and John M. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comput.*, 19:297–301, 1965.

[CT05] Emmanuel Candès and Terence Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12):4203–4215, 2005.

[CT06] Emmanuel J. Candès and Terence Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52:5406–5425, 2006.

[CW09] Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pages 205–214, 2009.

[CW13] Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the 45th ACM Symposium on Theory of Computing (STOC)*, pages 81–90, 2013.

[DG13] David L. Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci.*, 100(10):5591–5596, 2013.

[Dir14] Sjoerd Dirksen. Dimensionality reduction with subgaussian matrices: a unified theory. *CoRR*, abs/1402.3973, 2014.

[Dir15] Sjoerd Dirksen. Tail bounds via generic chaining. *Electron. J. Probab.*, 20:no. 53, 1–29, 2015.

[DKS10] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)*, pages 341–350, 2010.

[DMIMW12] Petros Drineas, Malik Magdon-Ismail, Michael Mahoney, and David Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3475–3506, 2012.

[Don06] David Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

[Dud67] Richard M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Functional Analysis*, 1:290–330, 1967.

[EW13] Armin Eftekhari and Michael B Wakin. New analysis of manifold embeddings and signal recovery from compressive measurements. *CoRR*, abs/1306.4748, 2013.

[Fer75] Xavier Fernique. Regularité des trajectoires des fonctions aléatoires gaussiennes. *Lecture Notes in Math.*, 480:1–96, 1975.

[Fig76] T. Figiel. On the moduli of convexity and smoothness. *Studia Math.*, 56(2):121–155, 1976.

[FR13] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhaüser, Boston, 2013.

[GMPTJ07] O. Guédon, S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Subspaces and orthogonal decompositions generated by bounded orthogonal systems. *Positivity*, 11(2):269–283, 2007.

[GMPTJ08] O. Guédon, S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Majorizing measures and proportional subsets of bounded orthonormal systems. *Rev. Mat. Iberoam.*, 24(3):1075–1095, 2008.

[Gor88] Yehoram Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. *Geometric Aspects of Functional Analysis*, pages 84–106, 1988.

[HIM12] Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani. Approximate nearest neighbor: Towards removing the curse of dimensionality. *Theory of Computing*, 8(1):321–350, 2012.

[HUL01] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer-Verlag, Berlin, 2001.

[HWB07] C. Hegde, M. Wakin, and R. Baraniuk. Random projections for manifold learning. In *Advances in neural information processing systems*, pages 641–648, 2007.

[Ind01] Piotr Indyk. Algorithmic applications of low-distortion geometric embeddings. In *Proceedings of the 42nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 10–33, 2001.

[IR13] Piotr Indyk and Ilya Razenshteyn. On model-based RIP-1 matrices. In *Proceedings of the 40th International Colloquium on Automata, Languages and Programming (ICALP)*, pages 564–575, 2013.

[JL84] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[Kah68] Jean-Pierre Kahane. *Some Random Series of Functions*. Heath Math. Monographs. Cambridge University Press, 1968.

[KM05] Bo'az Klartag and Shahar Mendelson. Empirical processes and random projections. *J. Funct. Anal.*, 225(1):229–245, 2005.

[KMR14] F. Krahmer, S. Mendelson, and Holger Rauhut. Suprema of chaos processes and the restricted isometry property. *Comm. Pure Appl. Math.*, 67(11):1877–1904, 2014.

[KN10] Daniel M. Kane and Jelani Nelson. A derandomized sparse Johnson-Lindenstrauss transform. *CoRR*, abs/1006.3585, 2010.

[KN14] Daniel M. Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss transforms. *Journal of the ACM*, 61(1):4, 2014.

[KW11] Felix Krahmer and Rachel Ward. New and improved Johnson-Lindenstrauss embeddings via the Restricted Isometry Property. *SIAM J. Math. Anal.*, 43(3):1269–1281, 2011.

[KZM02] Geza Kovács, Shay Zucker, and Tsevi Mazeh. A box-fitting algorithm in the search for periodic transits. *Astronomy and Astrophysics*, 391:369–377, 2002.

[LDFU13] Yichao Lu, Paramveer Dhillon, Dean Foster, and Lyle Ungar. Faster ridge regression via the subsampled randomized Hadamard transform. In *Proceedings of the 26th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2013.

[LN14] Kasper Green Larsen and Jelani Nelson. The Johnson-Lindenstrauss lemma is optimal for linear dimensionality reduction, 2014. Manuscript.

[LP86] François Lust-Piquard. Inégalités de Khintchine dans C_p $(1 < p < \infty)$. *C. R. Math. Acad. Sci. Paris*, 303(7):289–292, 1986.

[LPP91] François Lust-Piquard and Gilles Pisier. Noncommutative Khintchine and Paley inequalities. *Ark. Mat.*, 29(2):241–260, 1991.

[LS13] Yin Tat Lee and Aaron Sidford. Matching the universal barrier without paying the costs : Solving linear programs with õ(sqrt(rank)) linear system solves. *CoRR*, abs/1312.6677, 2013.

[Mat08] Jirí Matousek. On variants of the Johnson-Lindenstrauss lemma. *Random Struct. Algorithms*, 33(2):142–156, 2008.

[MM13] Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the 45th ACM Symposium on Theory of Computing (STOC)*, pages 91–100, 2013.

[MPTJ07] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.*, 17(4):1248–1282, 2007.

[MV97] Reinhold Meise and Dietmar Vogt. *Introduction to Functional Analysis*. Oxford Graduate Texts in Mathematics (Book 2). Oxford University Press, 1997.

[NN13a] Jelani Nelson and Huy L. Nguyễn. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 117–126, 2013.

[NN13b] Jelani Nelson and Huy L. Nguyễn. Sparsity lower bounds for dimensionality-reducing maps. In *Proceedings of the 45th ACM Symposium on Theory of Computing (STOC)*, pages 101–110, 2013.

[NPW14] Jelani Nelson, Eric Price, and Mary Wootters. New constructions of RIP matrices with fast multiplication and fewer rows. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2014.

[NT09] Deanna Needell and Joel A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.*, 26:301–332, 2009.

[PBMID13] Saurabh Paul, Christos Boutsidis, Malik Magdon-Ismail, and Petros Drineas. Random projections for support vector machines. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 498–506, 2013.

[Pie72] Albrecht Pietsch. *Theorie der Operatorenideale (Zusammenfassung)*. Friedrich-Schiller-Universität Jena, 1972.

[Pis86] Gilles Pisier. Probabilistic methods in the geometry of Banach spaces. *Probability and Analysis, Lecture Notes in Math.*, 1206:167–241, 1986.

[Pis81] Gilles Pisier. Remarques sur un résultat non public de B. Maurey. *Séminaire d'analyse fonctionnelle*, Exp. V., 1980/81.

[PTJ86] Alain Pajor and Nicole Tomczak-Jaegermann. Subspaces of small codimension of finite dimensional Banach spaces. *Proc. Amer. Math. Soc.*, 97:637–642, 1986.

[PW14] M. Pilanci and M. Wainwright. Randomized sketches of convex programs with sharp guarantees. *arXiv*, abs/1404.7203, 2014.

[Roc70] R. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[RS00] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[RV07] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM*, 54(4), 2007.

[RV08] Mark Rudelson and Roman Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

[Sar06] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

[SS11] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM J. Comput.*, 40(6):1913–1926, 2011.

[Tal05] Michel Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Verlag, 2005.

[TdSL00] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[Tro08] Joel A. Tropp. The random paving property for uniformly bounded matrices. *Studia Math.*, 185:67–82, 2008.

[Tro11] Joel A. Tropp. Improved analysis of the subsampled randomized hadamard transform. *Adv. Adapt. Data Anal.*, 3(1–2):115–126, 2011.

[VJ11] Andrew Vanderburg and John Asher Johnson. A technique for extracting highly precise photometry for the two-wheeled Kepler mission. *CoRR*, abs/1408.3853, 2011.

[WDL+09] Kilian Q. Weinberger, Anirban Dasgupta, John Langford, Alexander J. Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 1113–1120, 2009.

[Woo14] David P. Woodruff. Personal communication, 2014.

[WZ13] David P. Woodruff and Qin Zhang. Subspace embeddings and ℓ_p regression using exponential random variables. In *Proceedings of the 26th Conference on Learning Theory (COLT)*, 2013.

Appendix A. Tools from probability theory

We collect some useful tools from probability theory that are used throughout the text. For further reference we record an easy consequence of Markov's inequality.

Lemma 27. *If ξ is a real-valued random variable satisfying*

$$(\mathbb{E}\,|\xi|^p)^{1/p} \leq a_1p^2 + a_2p^{3/2} + a_3p + a_4p^{1/2} + a_5, \qquad \text{for all } p \geq p_0,$$

for some $0 \leq a_1, a_2, a_3, a_4, a_5 < \infty$, then

$$\mathbb{P}(|\xi| \geq e(a_1w^2 + a_2w^{3/2} + a_3w + a_4\sqrt{w} + a_5)) \leq \exp(-w) \qquad (w \geq p_0). \tag{A.1}$$

Let us call $\sigma = (\sigma_i)_{i=1}^n$ a Rademacher vector if its entries are independent Rademacher random variables. Khintchine's inequality states that for any $1 \leq p < \infty$

$$(\mathbb{E}_\sigma\,|\langle \sigma, x\rangle|^p)^{1/p} \leq C_p\|x\|_2 \qquad (x \in \mathbb{R}^n), \tag{A.2}$$

where $C_p \leq \sqrt{p}$. The noncommutative Khintchine inequality [LP86, LPP91] says that for any $A_1, \ldots, A_n \in \mathbb{R}^{m\times n}$ and $1 \leq p < \infty$

$$\Big(\mathbb{E}\Big\|\sum_{i=1}^n \sigma_iA_i\Big\|_{S_p}^p\Big)^{1/p} \leq C\sqrt{p}\max\Big\{\Big\|\Big(\sum_i A_i^*A_i\Big)^{1/2}\Big\|_{S_p}, \Big\|\Big(\sum_i A_iA_i^*\Big)^{1/2}\Big\|_{S_p}\Big\},$$

where $\|\cdot\|_{S_p}$ is the p-th Schatten norm. In particular, if $p \geq \max\{\log m, \log n\}$, then

$$\Big(\mathbb{E}\Big\|\sum_{i=1}^n \sigma_iA_i\Big\|^p\Big)^{1/p} \leq C\sqrt{p}\max\Big\{\Big\|\Big(\sum_i A_i^*A_i\Big)^{1/2}\Big\|, \Big\|\Big(\sum_i A_iA_i^*\Big)^{1/2}\Big\|\Big\} \tag{A.3}$$

as $\|\cdot\| \leq \|\cdot\|_{S_p} \leq e\|\cdot\|$ in this case. Khintchine's inequality and its noncommutative version are frequently used in combination with symmetrization: if $\zeta_1, \ldots, \zeta_n$ are X-valued random variables and σ is a Rademacher vector, then for any $1 \leq p < \infty$,

$$\Big(\mathbb{E}_\zeta\Big\|\sum_{i=1}^n \zeta_i - \mathbb{E}\,\zeta_i\Big\|_X^p\Big)^{1/p} \leq \Big(\mathbb{E}_\zeta\mathbb{E}_\sigma\Big\|\sum_{i=1}^n \sigma_i\zeta_i\Big\|_X^p\Big)^{1/p}. \tag{A.4}$$

The following decoupling inequality is elementary to prove (see e.g. [FR13, Theorem 8.11]). Let $\mathcal{M} \subset \mathbb{R}^{n\times n}$, let σ be an n-dimensional Rademacher vector and σ' an independent copy. Then, for any $1 \leq p < \infty$

$$(\mathbb{E}_\sigma \sup_{M\in\mathcal{M}} |\sigma^*M\sigma - \mathbb{E}(\sigma^*M\sigma)|^p)^{1/p} \leq 4(\mathbb{E}_{\sigma,\sigma'} \sup_{M\in\mathcal{M}} |(\sigma')^*M\sigma|^p)^{1/p}. \tag{A.5}$$

A special case of this bound, combined with Khintchine's inequality, implies the following.

Lemma 28. *Let $A \in \mathbb{R}^{m\times n}$ and let σ be an n-dimensional Rademacher vector. For any $p \geq 1$,*

$$(\mathbb{E}\,\|A\sigma\|_2^p)^{1/p} \leq \|A\|_F + 2\sqrt{2p}\|A\|.$$

Proof. By decoupling,

$$\begin{aligned}(\mathbb{E}\,\|A\sigma\|_2^p)^{2/p} &\leq (\mathbb{E}_\sigma\,|\sigma^*A^*A\sigma - \mathbb{E}\,\sigma^*A^*A\sigma|^{p/2})^{2/p} + \mathbb{E}(\sigma^*A^*A\sigma)\\ &\leq 4(\mathbb{E}_{\sigma,\sigma'}\,|(\sigma')^*A^*A\sigma|^{p/2})^{2/p} + \|A\|_F^2\end{aligned}$$

and therefore Khintchine's inequality implies that

$$\begin{aligned}(\mathbb{E}\,\|A\sigma\|_2^p)^{2/p} &\le 4\sqrt{p/2}(\mathbb{E}_\sigma\,\|A^*A\sigma\|_2^{p/2})^{2/p} + \|A\|_F^2\\ &\le 2\sqrt{2p}\|A\|(\mathbb{E}_\sigma\,\|A\sigma\|_2^p)^{1/p} + \|A\|_F^2.\end{aligned}$$

Solving this quadratic inequality yields the claim. □

To conclude we collect some tools to estimate covering numbers. Given two closed sets $U, V \subset \mathbb{R}^n$, we let the *covering number* $\mathcal{N}(U, V)$ be the minimal number of translates of V needed to cover U. If V is closed, convex and symmetric, then we can associate with it a semi-norm

$$\|x\|_V = \inf\{t > 0 \;:\; x \in tV\}.$$

In this case it is customary to write

$$\mathcal{N}(U, \|\cdot\|_V, \varepsilon) := \mathcal{N}(U, \varepsilon V) \qquad (\varepsilon > 0).$$

Conversely, if $\|\cdot\|$ is a semi-norm, then $V = \{x \in \mathbb{R}^n \;:\; \|x\| \le 1\}$ is closed, convex and symmetric, and $\|\cdot\|_V = \|\cdot\|$.

As a first tool we state two elementary bounds that follow from volumetric comparision. If $\|\cdot\|$ is any semi-norm on $\mathbb{R}^n$ and $B_{\|\cdot\|}$ is the associated unit ball,

$$\Big(\frac{1}{\varepsilon}\Big)^n \le \mathcal{N}(B_{\|\cdot\|}, \|\cdot\|, \varepsilon) \le \Big(1 + \frac{2}{\varepsilon}\Big)^n \qquad (0 < \varepsilon \le 1). \tag{A.6}$$

The following is known as Sudakov minoration or the dual Sudakov inequality [BLM89, Proposition 4.2], [PTJ86]. Eq. (B.1) in Appendix B gives a definition of the polar V°.

Lemma 29. *Let $V \subset \mathbb{R}^n$ be closed, convex and symmetric and let g be a standard Gaussian vector. Then,*

$$\sup_{\varepsilon>0} \varepsilon[\log \mathcal{N}(B_{\ell_2^n}, \|\cdot\|_V, \varepsilon)]^{1/2} \lesssim \mathbb{E} \sup_{x \in V^\circ} \langle g, x\rangle.$$

We will also use the following duality for covering numbers from [BPSTJ89, Proposition 4].

Lemma 30. *Let $U, V \subset \mathbb{R}^n$ be closed, bounded, convex and symmetric. For every $\theta \ge \varepsilon$,*

$$\mathcal{N}(U, \|\cdot\|_V, \varepsilon) \le \mathcal{N}(U, \|\cdot\|_V, \theta)[\mathcal{N}(V^\circ, \|\cdot\|_{U^\circ}, \varepsilon/8)]^r,$$

with $r \le (2^7 T_2(\|\cdot\|_V))^2(1 + \log(\theta/\varepsilon))$.

Finally, we state Maurey's lemma [Pis81, Car85] and its usual proof.

Lemma 31. *Let $\|\cdot\|$ be a semi-norm and $T_2(\|\cdot\|)$ be its type-2 constant. Let Ω be a set of points x each with $\|x\| \le 1$. Then for any $z \in \mathrm{conv}(\Omega)$ and any integer $\ell > 0$ there exist $y_1, \ldots, y_\ell \in \Omega$ with*

$$\Big\|z - \frac{1}{\ell}\sum_{i=1}^{\ell} y_i\Big\| \le \frac{T_2(\|\cdot\|)}{\sqrt{\ell}}$$

Proof. Write $z = \sum_i \lambda_i x_i$ for $0 < \lambda_i \le 1$, $\sum_i \lambda_i = 1$, $x_i \in \Omega$. Let $y = (y_1, \ldots, y_\ell)$ have i.i.d. entries from the distribution in which $y_j = x_j$ with probability λ_j. Let σ be a Rademacher vector and let g be a Gaussian vector. By symmetrization,

$$\begin{aligned}\mathbb{E}_y \Big\|z - \frac{1}{\ell}\sum_{i=1}^{\ell} y_i\Big\| &\le \frac{2}{\ell}\,\mathbb{E}_{\sigma,y}\Big\|\sum_{i=1}^{\ell}\sigma_i y_i\Big\|\\ &= \frac{2}{\ell}\,\mathbb{E}_{\sigma,y}\Big\|\mathbb{E}_g\sum_{i=1}^{\ell}\sigma_i|g_i|y_i\Big\| \lesssim \frac{1}{\ell}\,\mathbb{E}_{y,g}\Big\|\sum_{i=1}^{\ell} g_i y_i\Big\| \le \frac{T_2(\|\cdot\|)}{\sqrt{\ell}}.\end{aligned}$$

□

Appendix B. Tools from convex analysis

In this appendix we recall some basic facts from convex analysis. More details can be found in e.g. [HUL01, Roc70]. For any set $S \subset \mathbb{R}^n$ we let $\mathrm{conv}(S)$ denote the closed convex hull of S, i.e., the closure of the set of all convex combinations of elements in S. The polar of a set S is

$$S^\circ = \{y \in \mathbb{R}^n \ : \ \langle x, y\rangle \leq 1 \text{ for all } x \in S\}, \tag{B.1}$$

which is always closed and convex. A set $\mathcal{K}$ is called a cone if $\alpha\mathcal{K} \subset \mathcal{K}$ for all $\alpha > 0$. It is called a convex cone if it is, in addition, convex. We use $\mathrm{cone}(S)$ to denote the closed convex cone generated by a set S. The polar $\mathcal{K}^\circ$ of a convex cone $\mathcal{K}$ can be written as

$$\mathcal{K}^\circ = \{y \in \mathbb{R}^n \ : \ \langle x, y\rangle \leq 0 \text{ for all } x \in \mathcal{K}\}$$

and is always a closed convex cone. The bipolar theorem states that $\mathcal{K}^{\circ\circ}$ is equal to the closure $\overline{\mathcal{K}}$ of $\mathcal{K}$ (cf. [MV97, Theorem 6.11]). Let $\mathcal{C}$ be a closed convex set in $\mathbb{R}^n$. The tangent cone $T_{\mathcal{C}}(x)$ to $\mathcal{C}$ at $x \in \mathcal{C}$ is the closed cone generated by $\mathcal{C} - \{x\}$, i.e.,

$$T_{\mathcal{C}}(x) = \mathrm{cone}\{d \in \mathbb{R}^n \ : \ d = y - x, \ y \in \mathcal{C}\}.$$

Clearly, if x is in the interior of $\mathcal{C}$, then $T_{\mathcal{C}}(x) = \mathbb{R}^n$. The normal cone $N_{\mathcal{C}}(x)$ to $\mathcal{C}$ at $x \in \mathcal{C}$ is given by

$$N_{\mathcal{C}}(x) = \{s \in \mathbb{R}^n \ : \ \langle s, y - x\rangle \leq 0 \ \forall y \in \mathcal{C}\}.$$

It is easy to see that $(N_{\mathcal{C}}(x))^\circ = T_{\mathcal{C}}(x)$. Since $T_{\mathcal{C}}(x)$ is closed, the bipolar theorem implies that $(T_{\mathcal{C}}(x))^\circ = N_{\mathcal{C}}(x)$.

If $f : \mathbb{R}^n \to \mathbb{R}$ is any function, then a vector $\xi \in \mathbb{R}^n$ is called a subgradient of f at $x \in \mathrm{dom}(f)$ if for any $y \in \mathrm{dom}(f)$

$$f(y) \geq f(x) + \langle \xi, y - x\rangle. \tag{B.2}$$

The set of all subgradients of f at x is called the subdifferential of f at x and denoted by $\partial f(x)$. If $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a proper convex function (i.e., $f(x) < \infty$ for some $x \in \mathbb{R}^n$), then the descent cone $D(f, x)$ of f at $x \in \mathbb{R}^n$ is defined by

$$D(f, x) = \bigcup_{t>0} \{d \in \mathbb{R}^n \ : \ f(x + td) \leq f(x)\}.$$

The descent cone is always a convex cone, but it may not be closed.

Theorem 32. *[Roc70, Theorem 23.7] Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper convex function. Suppose that the subdifferential $\partial f(x)$ is nonempty and does not contain 0. Then,*

$$D(f, x)^\circ = \mathrm{cone}(\partial f(x)) = \overline{\bigcup_{t \geq 0} t\partial f(x)}.$$

Let $\|\cdot\|$ be any norm on $\mathbb{R}^n$ and set

$$\mathcal{C} = \{x \in \mathbb{R}^n \ : \ \|x\| \leq R\}.$$

Then, for any $x \in \mathcal{C}$ with $\|x\| = R$, $T_{\mathcal{C}}(x)$ is equal to the descent cone of $\|\cdot\|$ at x. By Theorem 32 and the bipolar theorem this implies that

$$T_{\mathcal{C}}(x) = [\mathrm{cone}(\partial\|\cdot\|(x))]^\circ. \tag{B.3}$$

Let $\|\cdot\|_*$ denote the dual norm of $\|\cdot\|$, i.e.,

$$\|y\|_* = \sup_{\|x\| \leq 1} \langle x, y\rangle.$$

It is easy to verify from the definition (B.2) that

$$\partial\|\cdot\|(x) = \begin{cases} \{y \in \mathbb{R}^n \ : \ \|y\|_* \leq 1\}, & \text{if } x = 0, \\ \{y \in \mathbb{R}^n \ : \ \|y\|_* \leq 1 \text{ and } \langle y, x\rangle = \|x\|\}, & \text{otherwise.} \end{cases}$$

Using these tools, we can readily calculate the tangent cone to an $\ell_{2,1}$-ball.

Example 33. *Consider the $\ell_{2,1}$-norm defined in (6.5) and its dual norm, the $\ell_{2,\infty}$-norm. Set*

$$\mathcal{C} = \{x \in \mathbb{R}^{bD} \ : \ \|x\|_{2,1} \leq R\}.$$

Suppose that $x \neq 0$ and let $S \subset [b]$ be the indices corresponding to nonzero blocks of x. For $z \in \mathbb{R}^{bD}$, let z_S be the vector

$$(z_S)_{B_i} = \begin{cases} z_{B_i} & \text{if } i \in S \\ 0 & \text{if } i \in S^c. \end{cases}$$

Then,

$$\begin{aligned} \partial\|\cdot\|_{2,1}(x) &= \{z \in \mathbb{R}^{bD} \ : \ \|z\|_{2,\infty} \leq 1 \text{ and } \langle z, x\rangle = \|x\|_{2,1}\} \\ &= \{z \in \mathbb{R}^{bD} \ : \ \|z_{S^c}\|_{2,\infty} \leq 1 \text{ and } z_{B_i} = x_{B_i}/\|x_{B_i}\|_2 \text{ for all } i \in S\}. \end{aligned}$$

If $x \in \mathbb{R}^{bD}$ satisfies $\|x\|_{2,1} = R$, then by (B.3),

$$\begin{aligned} T_{\mathcal{C}}(x) &= \{y \in \mathbb{R}^{bD} \ : \ \langle z, y\rangle \leq 0 \text{ for all } z \in \partial\|\cdot\|_{2,1}(x)\} \\ &= \Big\{y \in \mathbb{R}^{bD} \ : \ \sum_{i\in S} \Big\langle \frac{x_{B_i}}{\|x_{B_i}\|_2}, y_{B_i}\Big\rangle + \|y_{S^c}\|_{2,1} \leq 0\Big\}. \end{aligned}$$

In particular, any $y \in T_{\mathcal{C}}(x)$ satisfies

$$\|y\|_{2,1} = \|y_S\|_{2,1} + \|y_{S^c}\|_{2,1} \leq 2\|y_S\|_{2,1} \leq 2\sqrt{|S|}\ \|y\|_2.$$

Appendix C. Sketching least squares programs with an FJLT

In this appendix we study sketching of least squares programs using a fast Johnson-Lindenstrauss transform (FJLT). We show in Theorem 40 that for $\ell_{2,1}$-constrained least squares minimization, one can achieve the same sketching dimension as in the sparse case (cf. Section 6.2).

We first recall the definition of the FJLT. Let F be the discrete Fourier transform. Let $\theta_1, \ldots, \theta_n : \Omega_\theta \to \{0, 1\}$ be independent random selectors satisfying $\mathbb{P}_\theta(\theta_i = 1) = m/n$. Let $\sigma_1, \ldots, \sigma_n : \Omega_\sigma \to \{-1, 1\}$ be independent Rademacher random variables. The FJLT is defined by $\Psi = \Theta F D_\sigma$, where $\Theta = \sqrt{n/m}\ \mathrm{diag}((\theta_i)_{i=1}^n)$ and $D_\sigma = \mathrm{diag}((\sigma_i)_{i=1}^n)$. Here $\mathrm{diag}((x_i))$ denotes the diagonal matrix with the elements of the sequence (x_i) on its diagonal. To prove Theorem 40 we apply Lemma 11 using a suitable upper bound for Z_2 and lower bound for Z_1. To obtain the latter, we use the following chaining estimate.

Let T be some index set. For a given set $(x_{t,i})_{t\in T, 1\leq i\leq n}$ in $\mathbb{R}$, we define a semi-metric ρ_x on T by

$$\rho_x(t, s) = \max_{1\leq i\leq n} |x_{t,i} - x_{s,i}|$$

and the denote the associated radius by

$$d_x(T) = \sup_{t\in T} \max_{1\leq i\leq n} |x_{t,i}|.$$

For every $1 \leq i \leq n$ we fix a probability space $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ and let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the associated product space. The following observation was proven in the special case $p = 1$ in [GMPTJ07, Theorem 1.2].

Lemma 34. *Fix $1 \leq p < \infty$. For every $t \in T$ and $1 \leq i \leq n$ let $X_{t,i} \in L_{2p}(\Omega_i)$. Then,*

$$\begin{aligned}
&\Big(\mathbb{E} \sup_{t\in T} \Big|\sum_{i=1}^n X_{t,i}^2 - \mathbb{E}\, X_{t,i}^2\Big|^p\Big)^{1/p} \\
&\quad \lesssim (\mathbb{E}\, \gamma_2^{2p}(T,\rho_X))^{1/p} + \sup_{t\in T}\Big(\sum_{i=1}^n \mathbb{E}\, X_{t,i}^2\Big)^{1/2} (\mathbb{E}\, \gamma_2^p(T,\rho_X))^{1/p} \\
&\qquad + \sqrt{p} \sup_{t\in T}\Big(\sum_{i=1}^n \mathbb{E}\, X_{t,i}^2\Big)^{1/2} (\mathbb{E}\, d_X^p(T))^{1/p} + p(\mathbb{E}\, d_X^{2p}(T))^{1/p}.
\end{aligned}$$

Proof. Let $(r_i)_{i\geq 1}$ be a Rademacher sequence. By symmetrization,

$$\Big(\mathbb{E} \sup_{t\in T} \Big|\sum_{i=1}^n X_{t,i}^2 - \mathbb{E}\, X_{t,i}^2\Big|^p\Big)^{1/p} \leq 2\Big(\mathbb{E}\, \mathbb{E}_r \sup_{t\in T} \Big|\sum_{i=1}^n r_i X_{t,i}^2\Big|^p\Big)^{1/p}. \tag{C.1}$$

By Hoeffding's inequality, we have for any $s, t \in T$,

$$\mathbb{P}_r\Big(\sum_{i=1}^n r_i(X_{t,i}^2 - X_{s,i}^2) \geq u\Big(\sum_{i=1}^n (X_{t,i}^2 - X_{s,i}^2)^2\Big)^{1/2}\Big) \leq \exp(-u^2/2).$$

Since

$$\Big(\sum_{i=1}^n (X_{t,i}^2 - X_{s,i}^2)^2\Big)^{1/2} \leq \sqrt{2} \sup_{t\in T}\Big(\sum_{i=1}^n X_{t,i}^2\Big)^{1/2} \rho_X(t,s)$$

we conclude that $(\sum_{i=1}^n r_i X_{t,i}^2)_{t\in T}$ is subgaussian with respect to the semi-metric

$$\rho_*(s,t) = \sqrt{2} \sup_{t\in T}\Big(\sum_{i=1}^n X_{t,i}^2\Big)^{1/2} \rho_X(s,t).$$

By Lemma 12,

$$\begin{aligned}
&\Big(\mathbb{E}_r \sup_{t\in T} \Big|\sum_{i=1}^n r_i X_{t,i}^2\Big|^p\Big)^{1/p} \\
&\quad \lesssim \sup_{t\in T}\Big(\sum_{i=1}^n X_{t,i}^2\Big)^{1/2} \gamma_2(T,\rho_X) + \sqrt{p} \sup_{t\in T}\Big(\sum_{i=1}^n X_{t,i}^2\Big)^{1/2} d_X(T) \\
&\quad \leq \sup_{t\in T}\Big|\sum_{i=1}^n X_{t,i}^2 - \mathbb{E}\, X_{t,i}^2\Big|^{1/2} \Big(\gamma_2(T,\rho_X) + \sqrt{p} d_X(T)\Big) \\
&\qquad + \sup_{t\in T}\Big(\sum_{i=1}^n \mathbb{E}\, X_{t,i}^2\Big)^{1/2} \Big(\gamma_2(T,\rho_X) + \sqrt{p} d_X(T)\Big).
\end{aligned}$$

Taking L_p-norms on both sides, using (C.1) and applying Hölder's inequality yields

$$\begin{aligned}
&\Big(\mathbb{E} \sup_{t\in T} \Big|\sum_{i=1}^n X_{t,i}^2 - \mathbb{E}\, X_{t,i}^2\Big|^p\Big)^{1/p} \\
&\lesssim \Big(\mathbb{E} \sup_{t\in T} \Big|\sum_{i=1}^n X_{t,i}^2 - \mathbb{E}\, X_{t,i}^2\Big|^p\Big)^{1/2p} \Big((\mathbb{E}\, \gamma_2^{2p}(T,\rho_X))^{1/2p} + \sqrt{p}(\mathbb{E}\, d_X^{2p}(T))^{1/2p}\Big) \\
&\quad + \sup_{t\in T}\Big(\sum_{i=1}^n \mathbb{E}\, X_{t,i}^2\Big)^{1/2} \Big((\mathbb{E}\, \gamma_2^p(T,\rho_X))^{1/p} + \sqrt{p}(\mathbb{E}\, d_X^p(T))^{1/p}\Big).
\end{aligned}$$

Solving this quadratic inequality yields the result. □

To estimate the γ_2-functional occuring in Lemma 34 we use a covering number estimate from [GMPTJ08]. Recall the following definitions. Let E be a Banach space and let E^* denote its dual space. The modulus of convexity of E is defined by

$$\delta_E(\varepsilon) = \inf\Big\{1 - \frac{1}{2}\|x+y\|,\ \|x\|=1,\ \|y\|=1,\ \|x-y\|>\varepsilon\Big\} \qquad (0 \leq \varepsilon \leq 2).$$

We say that E is uniformly convex if $\delta_E(\varepsilon) > 0$ for all $\varepsilon > 0$ and that E is uniformly convex of power type 2 with constant λ if $\delta_E(\varepsilon) \geq \varepsilon^2/(8\lambda^2)$ for all $\varepsilon > 0$. The following observation is due to Figiel [Fig76, Proposition 24].

Lemma 35. *Suppose that E is a p-convex and q-concave Banach lattice for some $1 < p \leq q < \infty$. Set $r = \max\{2, q\}$ and $K = \max\{2, \frac{2}{\sqrt{p-1}}\}$. Then, $\delta_E(\varepsilon) \geq r^{-1}K^{-r}\varepsilon^r$ for all $0 \leq \varepsilon \leq 2$.*

Using the Hölder-Minkowski inequalities one readily checks that $\ell_{2,p} = \ell_p(\ell_2)$ is p-convex and 2-concave if $1 < p \leq 2$. Therefore, $\delta_{\ell_{2,p}}(\varepsilon) \geq \frac{1}{8}(p-1)\varepsilon^2$.

Lemma 36. *[GMPTJ08, Lemma 1] Let E be uniformly convex of power type 2 with constant λ. Let $T_2(E^*)$ be the type 2 constant of E^*. Consider $v_1, \ldots, v_N \in E^*$ and define an associated semi-metric on E by*

$$\rho_v(x,y) = \max_{1\leq i\leq N} |\langle v_i, x-y\rangle|. \tag{C.2}$$

Set $\nu = \max_{1\leq i\leq N}\|v_i\|_{E^*}$ and let $U \subset B_E$. Then, for all $t > 0$,

$$\log^{1/2}(2\mathcal{N}(U,\rho_v,t)) \lesssim \nu\lambda^2 T_2(E^*)\log^{1/2}(N)t^{-1}. \tag{C.3}$$

We can now estimate the parameter Z_1.

Lemma 37. *Set $d = bD$. Let Ψ be the FJLT, fix $A \in \mathbb{R}^{n\times d}$ and let $\mathcal{K} \subset \mathbb{R}^d$. Consider the norm $\|\!|A|\!\|$ defined in (6.6) and set*

$$\beta = (\log^2(\eta^{-1}) + \log(b) + \log(n))\log(n)\log^3(b)\log^2(d). \tag{C.4}$$

If

$$m \gtrsim \beta\varepsilon^{-2}\|\!|A|\!\|^2\Big[\sup_{x\in\mathcal{K}\ :\|Ax\|_2=1}\|x\|_{2,1}^2\Big], \tag{C.5}$$

then with probability at least $1-\eta$ we have

$$(1-\varepsilon)\|x\|_2^2 \leq \|\Psi x\|_2^2 \leq (1+\varepsilon)\|x\|_2^2, \qquad \text{for all } x \in A\mathcal{K}\cap S^{n-1}.$$

In particular, $Z_1(A,\Psi,\mathcal{K}) \geq 1-\varepsilon$.

Proof. Let F_i denote the i-th row of F. Since

$$\mathop{\mathbb{E}}_{\theta}\|\Psi x\|_2^2 = \|D_\sigma x\|_2^2 = \|x\|_2^2, \qquad \text{for all } x \in \mathbb{R}^n,$$

we have

$$\begin{aligned}
&\sup_{x\in A\mathcal{K}\cap S^{n-1}}\Big|\|\Psi x\|_2^2 - \|x\|_2^2\Big| \\
&\quad = \sup_{x\in A\mathcal{K}\cap S^{n-1}}\Big|\sum_{i=1}^n \theta_i\Big\langle\sqrt{\frac{n}{m}}D_\sigma F_i, x\Big\rangle^2 - \mathop{\mathbb{E}}_{\theta}\theta_i\Big\langle\sqrt{\frac{n}{m}}D_\sigma F_i, x\Big\rangle^2\Big| \\
&\quad = \sup_{x\in\mathcal{K}\cap A^{-1}(S^{n-1})}\Big|\sum_{i=1}^n \theta_i\Big\langle\sqrt{\frac{n}{m}}A^*D_\sigma F_i, x\Big\rangle^2 - \mathop{\mathbb{E}}_{\theta}\theta_i\Big\langle\sqrt{\frac{n}{m}}A^*D_\sigma F_i, x\Big\rangle^2\Big|.
\end{aligned}$$

We apply Lemma 34 (with $X_{x,i} := \langle\theta_i\sqrt{\frac{n}{m}}A^*D_\sigma F_i, x\rangle$) to find

$$\Big(\mathop{\mathbb{E}}_{\theta}\sup_{x\in A\mathcal{K}\cap S^{n-1}}\Big|\|\Psi x\|_2^2 - \|x\|_2^2\Big|^p\Big)^{1/p}$$

$$\leq (\mathbb{E}_\theta \gamma_2^{2p}(\mathcal{K}\cap A^{-1}(S^{n-1}),\rho_X))^{1/p} + (\mathbb{E}_\theta \gamma_2^{p}(\mathcal{K}\cap A^{-1}(S^{n-1}),\rho_X))^{1/p}$$
$$+ \sqrt{p}(\mathbb{E}_\theta d_X^p(\mathcal{K}\cap A^{-1}(S^{n-1}))^{1/p} + p(\mathbb{E}_\theta d_X^{2p}(\mathcal{K}\cap A^{-1}(S^{n-1}))^{1/p}$$
$$\leq \gamma_2^2(\mathcal{K}\cap A^{-1}(S^{n-1}),\rho_v) + \gamma_2(\mathcal{K}\cap A^{-1}(S^{n-1}),\rho_v)$$
$$+ \sqrt{p}d_v(\mathcal{K}\cap A^{-1}(S^{n-1})) + pd_v^2(\mathcal{K}\cap A^{-1}(S^{n-1})),$$

where we have set $v_i = \sqrt{\frac{n}{m}}A^*D_\sigma F_i$, defined ρ_v as in (C.2) and used that $\rho_X \leq \rho_v$ uniformly. Set $d_{2,1} = d_{\|\cdot\|_{2,1}}(\mathcal{K}\cap A^{-1}(S^{n-1}))$ and $\nu := \max_{1\leq i\leq n}\|v_i\|_{2,\infty}$. Clearly

$$d_v(\mathcal{K}\cap A^{-1}(S^{n-1})) \leq d_{2,1}\nu. \tag{C.6}$$

We estimate the γ_2-functional by an entropy integral

$$\gamma_2(\mathcal{K}\cap A^{-1}(S^{n-1}),\rho_v)$$
$$\lesssim \int_0^{t^*} \mathcal{N}(\mathcal{K}\cap A^{-1}(S^{n-1}),\rho_v,t) + \int_{t_*}^\infty \mathcal{N}(\mathcal{K}\cap A^{-1}(S^{n-1}),\rho_v,t)\ dt.$$

The first integral we estimate using the volumetric bound

$$\mathcal{N}(\mathcal{K}\cap A^{-1}(S^{n-1}),\rho_v,t) \leq \mathcal{N}(\mathcal{K}\cap A^{-1}(S^{n-1}),\nu\|\cdot\|_{2,1},t) \leq \Big(1+\frac{2\nu d_{2,1}}{t}\Big)^d.$$

For the second integral we set $p = \log(b)/(\log(b)-1)$ and apply Lemma 36 with $E = \ell_{2,p}^d$ and $E^* = \ell_{2,p'}^d$, where $p' = \log(b)$. Note that $\|\cdot\|_{2,\infty} \leq \|\cdot\|_{2,p'} \leq e\|\cdot\|_{2,\infty}$ and therefore $T_2(E^*) \leq e\log^{1/2}(b)$. Also, by the remark after Lemma 35 we have $\lambda^2 = (p-1)^{-1} = \log(b)-1$. By (C.3),

$$\log^{1/2}(\mathcal{N}(B_{\ell_{2,1}^d},\rho_v,t)) \leq \log^{1/2}(\mathcal{N}(B_{\ell_{2,p}^d},\rho_v,t)) \lesssim \nu\log^{3/2}(b)\log^{1/2}(n)t^{-1}. \tag{C.7}$$

Since $\mathcal{K}\cap A^{-1}(S^{n-1}) \subset d_{2,1}B_{\ell_{2,1}^d}$, we arrive at

$$\gamma_2(\mathcal{K}\cap A^{-1}(S^{n-1}),\rho_v)$$
$$\lesssim \sqrt{d}\int_0^{t_*}\log^{1/2}\Big(1+\frac{2\nu d_{2,1}}{t}\Big)\ dt + \int_{t_*}^{\nu d_{2,1}} \nu d_{2,1}\log^{1/2}(n)\log^{3/2}(b)t^{-1}\ dt$$
$$\leq \sqrt{d}t_*\log^{1/2}(e+2et_*^{-1}\nu d_{2,1}) + \nu d_{2,1}\log^{1/2}(n)\log^{3/2}(b)\log(t_*^{-1}\nu d_{2,1}).$$

Take $t_* = d^{-1/2}\nu d_{2,1}$ to obtain

$$\gamma_2(\mathcal{K}\cap A^{-1}(S^{n-1}),\rho_v)$$
$$\lesssim \nu d_{2,1}\log^{1/2}(e+2e\sqrt{d}) + \nu d_{2,1}\log^{1/2}(n)\log^{3/2}(b)\log(\sqrt{d})$$
$$\lesssim \nu d_{2,1}\log^{1/2}(n)\log^{3/2}(b)\log(d). \tag{C.8}$$

In conclusion,

$$\Big(\mathbb{E}_\theta \sup_{x\in A\mathcal{K}\cap S^{n-1}}\Big|\|\Psi x\|_2^2 - \|x\|_2^2\Big|^p\Big)^{1/p}$$
$$\lesssim \nu^2 d_{2,1}^2\log(n)\log^3(b)\log^2(d) + \nu d_{2,1}\log^{1/2}(n)\log^{3/2}(b)\log(d)$$
$$+ \sqrt{p}\nu d_{2,1} + p\nu^2 d_{2,1}^2. \tag{C.9}$$

Since $|\sqrt{n}F_{ij}| \leq 1$, Khintchine's inequality implies that

$$(\mathbb{E}_\sigma \nu^p)^{1/p}$$
$$= \frac{1}{\sqrt{m}}(\mathbb{E}_\sigma \max_{1\leq i\leq n}\sqrt{n}\|A^*D_\sigma F_i\|_{2,\infty}^p)^{1/p}$$
$$= \frac{1}{\sqrt{m}}\Big(\mathbb{E}_\sigma \max_{1\leq i\leq n}\max_{1\leq \ell\leq b}\Big(\sum_{k\in B_\ell}\Big|\sum_{j=1}^n \sigma_j A_{jk}\sqrt{n}F_{ij}\Big|^2\Big)^{p/2}\Big)^{1/p}$$

$$\leq \frac{1}{\sqrt{m}}(\sqrt{p}+\log^{1/2}(b)+\log^{1/2}(n)) \max_{1\leq i\leq n}\max_{1\leq \ell\leq b}\Big(\sum_{k\in B_\ell}\sum_{j=1}^n |A_{jk}\sqrt{n}F_{ij}|^2\Big)^{1/2}$$

$$\leq \frac{1}{\sqrt{m}}(\sqrt{p}+\log^{1/2}(b)+\log^{1/2}(n)) \max_{1\leq \ell\leq b}\Big(\sum_{k\in B_\ell}\sum_{j=1}^n |A_{jk}|^2\Big)^{1/2}. \tag{C.10}$$

Taking $L_p(\Omega_\sigma)$-norms in (C.9) and using (C.10), we conclude that

$$\begin{aligned}
&\Big(\mathbb{E}_{\theta,\sigma}\sup_{x\in A\mathcal{K}\cap S^{n-1}}\Big|\|\Psi x\|_2^2-\|x\|_2^2\Big|^p\Big)^{1/p}\\
&\quad\lesssim \frac{1}{m}(\sqrt{p}+\log^{1/2}(b)+\log^{1/2}(n))^2|\!|\!|A|\!|\!|^2 d_{2,1}^2\log(n)\log^3(b)\log^2(d)\\
&\qquad+\frac{1}{\sqrt{m}}(\sqrt{p}+\log^{1/2}(b)+\log^{1/2}(n))|\!|\!|A|\!|\!| d_{2,1}\log^{1/2}(n)\log^{3/2}(b)\log(d)\\
&\qquad+\sqrt{p}\frac{1}{\sqrt{m}}(\sqrt{p}+\log^{1/2}(b)+\log^{1/2}(n))|\!|\!|A|\!|\!| d_{2,1}\\
&\qquad+p\frac{1}{m}(\sqrt{p}+\log^{1/2}(b)+\log^{1/2}(n))^2|\!|\!|A|\!|\!|^2 d_{2,1}^2.
\end{aligned}$$

The result now follows from Lemma 27 and taking $w=\log(\eta^{-1})$ in (A.1). □

To prove an upper bound for Z_2 we use the following variation of Lemma 34. Note that the element u below does not need to be in the index set T.

Lemma 38. *For every $t\in T$ and $1\leq i\leq n$ let $X_{t,i}\in L_{2p}(\Omega_i)$. Fix also $X_{u,i}\in L_p(\Omega_i)$. For any $1\leq p<\infty$,*

$$\begin{aligned}
&\Big(\mathbb{E}\sup_{t\in T}\Big|\sum_{i=1}^n X_{t,i}X_{u,i}-\mathbb{E}(X_{t,i}X_{u,i})\Big|^p\Big)^{1/p}\\
&\quad\lesssim\Big(\sqrt{p}\Big(\mathbb{E}\max_{1\leq i\leq n}|X_{u,i}|^{2p}\Big)^{1/(2p)}+\Big(\sum_{i=1}^n\mathbb{E}X_{u,i}^2\Big)^{1/2}\Big)\\
&\qquad\qquad * \Big((\mathbb{E}\gamma_2^{2p}(T,\rho_X))^{1/(2p)}+\sqrt{p}(\mathbb{E}d_X^{2p}(T))^{1/(2p)}\Big).
\end{aligned}\tag{C.11}$$

Proof. Let (r_i) be a Rademacher sequence. By Hoeffding's inequality, we have for any $s,t\in T$ and $w\geq 0$,

$$\mathbb{P}_r\Big(\sum_{i=1}^n r_i(X_{t,i}X_{u,i}-X_{s,i}X_{u,i})\geq w\Big(\sum_{i=1}^n(X_{t,i}X_{u,i}-X_{s,i}X_{u,i})^2\Big)^{1/2}\Big)\leq e^{-w^2/2}.$$

Since

$$\Big(\sum_{i=1}^n(X_{t,i}X_{u,i}-X_{s,i}X_{u,i})^2\Big)^{1/2}\leq\Big(\sum_{i=1}^n X_{u,i}^2\Big)^{1/2}\rho_X(t,s),$$

we conclude that $(\sum_{i=1}^n r_iX_{t,i}X_{u,i})_{t\in T}$ is subgaussian with respect to the semi-metric

$$\rho_*(s,t)=\Big(\sum_{i=1}^n X_{u,i}^2\Big)^{1/2}\rho_X(s,t).$$

By Lemma 12,

$$\Big(\mathbb{E}_r\sup_{t\in T}\Big|\sum_{i=1}^n r_iX_{t,i}X_{u,i}\Big|^p\Big)^{1/p}\lesssim\Big(\sum_{i=1}^n X_{u,i}^2\Big)^{1/2}\gamma_2(T,\rho_X)+\sqrt{p}d_X(T).$$

Using symmetrization (A.4), this implies that

$$\Big(\mathbb{E}\sup_{t\in T}\Big|\sum_{i=1}^n X_{t,i}X_{u,i}-\mathbb{E}(X_{t,i}X_{u,i})\Big|^p\Big)^{1/p}$$

$$\le 2\Big(\mathbb{E}\,\mathbb{E}_r \sup_{t\in T}\Big|\sum_{i=1}^n r_i X_{t,i}X_{u,i}\Big|^p\Big)^{1/p}$$
$$\lesssim \Big(\mathbb{E}\Big(\sum_{i=1}^n X_{u,i}^2\Big)^p\Big)^{1/(2p)}\Big((\mathbb{E}\,\gamma_2^{2p}(T,\rho_X))^{1/(2p)} + \sqrt{p}(\mathbb{E}\,d_X^{2p}(T))^{1/(2p)}\Big).$$

By symmetrization and Khintchine's inequality,

$$\begin{aligned}
&\Big(\mathbb{E}\Big(\sum_{i=1}^n X_{u,i}^2\Big)^p\Big)^{1/p}\\
&\quad\le \Big(\mathbb{E}\Big|\sum_{i=1}^n X_{u,i}^2 - \mathbb{E}\,X_{u,i}^2\Big|^p\Big)^{1/p} + \sum_{i=1}^n \mathbb{E}\,X_{u,i}^2\\
&\quad\le 2\Big(\mathbb{E}\,\mathbb{E}_r\Big|\sum_{i=1}^n r_i X_{u,i}^2\Big|^p\Big)^{1/p} + \sum_{i=1}^n \mathbb{E}\,X_{u,i}^2\\
&\quad\le 2\sqrt{p}\Big(\mathbb{E}\Big|\sum_{i=1}^n X_{u,i}^4\Big|^{p/2}\Big)^{1/p} + \sum_{i=1}^n \mathbb{E}\,X_{u,i}^2\\
&\quad\le 2\sqrt{p}\Big(\mathbb{E}\max_{1\le i\le n}|X_{u,i}|^{2p}\Big)^{1/(2p)}\Big(\mathbb{E}\Big(\sum_{i=1}^n X_{u,i}^2\Big)^p\Big)^{1/(2p)} + \sum_{i=1}^n \mathbb{E}\,X_{u,i}^2.
\end{aligned}$$

Solving this quadratic inequality yields

$$\Big(\mathbb{E}\Big(\sum_{i=1}^n X_{u,i}^2\Big)^p\Big)^{1/(2p)} \le 2\sqrt{p}\Big(\mathbb{E}\max_{1\le i\le n}|X_{u,i}|^{2p}\Big)^{1/(2p)} + \Big(\sum_{i=1}^n \mathbb{E}\,X_{u,i}^2\Big)^{1/2}.$$

□

Lemma 39. *Let Ψ be the FJLT, let $A\in\mathbb{R}^{n\times d}$, $\mathcal{K}\subset\mathbb{R}^d$ and $u\in S^{n-1}$. Let $|||A|||$ be as in (6.6) and β as in (C.4). If*

$$m \gtrsim \beta\varepsilon^{-2}|||A|||^2\Big[\sup_{x\in\mathcal{K}\,:\|Ax\|_2=1}\|x\|_{2,1}^2\Big],$$

then $Z_2(A,\Psi,\mathcal{K},u)\le\varepsilon$ with probability at least $1-\eta$.

Proof. If Ψ_i denotes the i-th row of Ψ, then we can write

$$Z_2(A,\Psi,\mathcal{K},u) = \sup_{x\in\mathcal{K}\cap A^{-1}(S^{n-1})}\Big|\sum_{i=1}^n\langle A^*\Psi_i,x\rangle\langle\Psi_i,u\rangle - \mathbb{E}(\langle A^*\Psi_i,x\rangle\langle\Psi_i,u\rangle)\Big|.$$

Set $v_i=\sqrt{\frac{n}{m}}A^*D_\sigma F_i$ and let ρ_v be the semi-metric in (C.2). We apply Lemma 38 with $X_{x,i}:=\langle\theta_i\sqrt{\frac{n}{m}}A^*D_\sigma F_i,x\rangle$ and $X_{u,i}:=\langle\theta_i\sqrt{\frac{n}{m}}D_\sigma F_i,u\rangle$. By (C.6) and (C.8) we know that

$$\begin{aligned}
(\mathbb{E}_\theta\, d_X^{2p}(T))^{1/(2p)} &\le d_{2,1}\nu\\
(\mathbb{E}_\theta\,\gamma_2^{2p}(T,\rho_X))^{1/(2p)} &\lesssim d_{2,1}\nu\log^{1/2}(n)\log^{3/2}(b)\log(d).
\end{aligned}$$

Moreover,

$$\sum_{i=1}^n\mathbb{E}_\theta X_{u,i}^2 = \|D_\sigma u\|_2^2 = 1, \qquad (\mathbb{E}_\theta\max_{1\le i\le n}|X_{u,i}|^{2p})^{1/(2p)} \le \max_{1\le i\le n}\sqrt{\frac{n}{m}}|\langle F_i,D_\sigma u\rangle|.$$

Applying these estimates in (C.11) and taking $L_p(\Omega_\sigma)$-norms yields

$$\begin{aligned}
(\mathbb{E}_{\theta,\sigma} Z_2^p)^{1/p} \lesssim{}& \Big(\sqrt{\frac{p}{m}}\Big(\mathbb{E}_\sigma\max_{1\le i\le n}\sqrt{n}|\langle F_i,D_\sigma u\rangle|^{2p}\Big)^{1/(2p)}+1\Big)\\
&* (\mathbb{E}_\sigma\,\nu^{2p})^{1/(2p)}d_{2,1}\Big(\log^{1/2}(n)\log^{3/2}(b)\log(d)+\sqrt{p}\Big).
\end{aligned}$$

By Khintchine's inequality,

$$\Big(\mathbb{E}_{\sigma} \max_{1\le i\le n} \sqrt{n}|\langle F_i, D_\sigma u\rangle|^{2p}\Big)^{1/(2p)} \lesssim (\sqrt{p}+\log^{1/2}(n)) \max_{1\le i\le n}\Big(\sum_{j=1}^n nF_{ij}^2u_j^2\Big)^{1/2}$$
$$\lesssim \sqrt{p}+\log^{1/2}(n).$$

and by (C.10)

$$(\mathbb{E}\,\nu^{2p})^{1/(2p)} \lesssim \frac{1}{\sqrt{m}}(\sqrt{p}+\log^{1/2}(b)+\log^{1/2}(n))|||A|||.$$

Combining these estimates and using Lemma 27 yields the result. □

Combining Lemmas 11, 37, and 39 yields the following result.

Theorem 40. *Set $d = bD$. Let Ψ be the FJLT, $A \in \mathbb{R}^{n\times d}$ and let $\mathcal{C}$ be a closed convex set in $\mathbb{R}^d$. Set β as in (C.4). Let x_* and $\hat{x}$ be the minimizers of (6.1) and (6.2), respectively. If*

$$m \gtrsim \beta\varepsilon^{-2}|||A|||^2\Big[\sup_{x\in T_{\mathcal{C}}(x_*)\ :\ \|Ax\|_2=1}\|x\|_{2,1}^2\Big],$$

then, with probability at least $1-\eta$,

$$f(\hat{x}) \le (1-\varepsilon)^{-2}f(x_*).$$

The proof of Corollary 17 immediately yields the following consequence.

Corollary 41. *Set $d = bD$. Let Ψ be the FJLT, $A \in \mathbb{R}^{n\times d}$ and let $\mathcal{C} = \{x \in \mathbb{R}^d : \|x\|_{2,1} \le R\}$. Define*

$$\sigma_{\min,k} = \inf_{\|y\|_2=1,\ \|y\|_{2,1}\le 2\sqrt{k}}\|Ay\|_2.$$

Suppose that x_* is k-block sparse and $\|x_*\|_{2,1} = R$. If

$$m \gtrsim \beta\varepsilon^{-2}|||A|||^2k\sigma_{\min,k}^{-2},$$

then, with probability at least $1-\eta$,

$$f(\hat{x}) \le (1-\varepsilon)^{-2}f(x_*).$$

Observe that the condition on m in Theorem 40 and Corollary 41 is, up to different log-factors, the same as the condition for the SJLT in Theorem 16 and Corollary 17.

In the special case $D = 1$, which corresponds to the Lasso, the result in Corollary 41 gives a qualitative improvement over [PW14, Corollary 3]. Recall from the discussion following Corollary 17 that in this case

$$|||A||| = \max_{1\le k\le d}\|A_k\|_2, \qquad \sigma_{\min,k} = \inf_{\|y\|_2=1,\ \|y\|_1\le 2\sqrt{k}}\|Ay\|_2.$$

In [PW14, Corollary 3] the condition

$$m \gtrsim \varepsilon^{-2}\log(\eta^{-1}) + \varepsilon^{-2}\min\Big\{\log^2(d)\Big(|||A|||^2k\sigma_{\min,k}^{-2}\Big)^2, k\log(d)\log^4(n)\frac{\sigma_{\max,k}^4}{\sigma_{\min,k}^4}\Big\},$$

was obtained, where $\sigma_{\max,k} = \sup_{\|y\|_2=1,\ \|y\|_1\leq 2\sqrt{k}} \|Ay\|_2$. In terms of the dependence on A this bound is worse than our result. We note, however, that the bound contains fewer log-factors and in particular the dependence on η is better.

Institute for Advanced Study, Princeton, NJ 08540
E-mail address: bourgain@math.ias.edu

RWTH Aachen University, 52062 Aachen, Germany
E-mail address: dirksen@mathc.rwth-aachen.de

Harvard University, Cambridge, MA 02138
E-mail address: minilek@seas.harvard.edu