Toward a unified theory of sparse dimensionality reduction in Euclidean space

Jean Bourgain[*] Sjoerd Dirksen[†] Jelani Nelson[‡]

November 4, 2014

Abstract

Let $\Phi \in \mathbb{R}^{m \times n}$ be a sparse Johnson-Lindenstrauss transform [52] 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.

[*] IAS, Princeton, NJ. bourgain@math.ias.edu. Partially supported by NSF grant DMS-1301619.

[†] RWTH Aachen, Aachen, Germany. dirksen@mathc.rwth-aachen.de. Partially supported by SFB grant 1060 of the Deutsche Forschungsgemeinschaft (DFG).

[‡] Harvard University, Cambridge, MA. minilek@seas.harvard.edu. 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 [79], high-dimensional computational geometry [48], privacy [15], compressed sensing [25], spectral graph theory [73], interior point methods for linear programming [58], numerical linear algebra [72], computational learning theory [11, 12], manifold learning [47, 29], motif-finding in computational biology [22], astronomy [35], 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 [79]. In astronomy a data point could be a star, represented as a vector of light intensities measured over various points sampled in time [54, 78]. Dimensionality reduction techniques in such applications provide the following benefits: (1) smaller storage consumption, (2) speedup during data analysis, (3) cheaper signal acquisition, and (4) cheaper transmission of data across computing clusters.

Typically such methods reduce the dimension while preserving point geometry, e.g. inter-point distances and angles. That is, one has $X \subset \mathbb{R}^n$ with n very large, and we would like a dimensionality-reducing map $f : X \to \mathbb{R}^m$, $m \ll n$, for some norm $\| \cdot \|$ such that $\forall x, y \in X,\ (1-\varepsilon)\|x-y\| \le \|f(x)-f(y)\| \le (1+\varepsilon)\|x-y\|$. A powerful tool for achieving this is the *Johnson-Lindenstrauss (JL) lemma* [50].

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 the above distance preservation for $\| \cdot \| = \| \cdot \|_2$.*

This bound on m is nearly tight: for any $n \ge 1$ Alon exhibited a point set X, $|X| = n+1$, such that any such JL map f must have $m = \Omega(\varepsilon^{-2} \frac{\log n}{\log(1/\varepsilon)})$ [5]. 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 [57]. 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, move beyond worst case analysis and be as efficient as possible *for our particular X*.

Unfortunately the previous question seems difficult. For the related question of computing the optimal distortion for embedding X into a line (i.e. $m = 1$), it is NP-hard to approximate the optimal distortion even up to a $|X|^{\Omega(1)}$ factor [10]. 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, f is required to be differentiable [79]. For several applications it is also crucial f be linear, e.g. in numerical linear algebra [72] and compressed sensing [25, 40]. In one-pass streaming applications [31] and data structural problems such as nearest neighbor search [46], it is further required that f even be chosen randomly without knowing X and still works with good probability. In streaming this is because X is not known up front. In data structure applications f must preserve distances to some future query points, which are not known when 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 [46] or Rademacher [1] 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 in X. Thus the JL guarantee is equivalent to

$$\sup_{x\in T} \left| \|\Phi x\|^2 - 1 \right| < \varepsilon. \tag{1.1}$$

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

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

Instance-wise understanding for achieving Eq. (1.2) was given by Gordon [45], who proved a random gaussian matrix satisfies Eq. (1.2) for $m \gtrsim (g^2(T) + 1)/\varepsilon^2$, where we write $A \gtrsim B$ if $A \geq CB$ for a universal constant C. Letting g be a standard n-dimensional Gaussian, the parameter $g(T)$ is defined as the *gaussian mean width* $g(T) \stackrel{\text{def}}{=} \mathbb{E}_g \sup_{x\in T}\langle g, x\rangle$. One thinks 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 [25] or subspace embeddings from numerical linear algebra [72], the best bound on m is a corollary of Gordon's theorem. Later works extended Gordon's theorem to Φ having Rademacher entries [53, 61, 39].

Although Gordon's theorem gives a good understanding for m, 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 [72]) multiplying a dense unstructured Φ times the input is slower than solving the exact solution of the original, high-dimensional problem! In compressed sensing, certain iterative recovery algorithms such as CoSamp [63] and Iterative Hard Thresholding [16] involve repeated multiplications by Φ and Φ^*, the conjugate transpose of Φ, and thus Φ supporting fast matrix-vector multiplication are desirable in such applications as well.

The first work to provide Φ with small m supporting faster multiplication is the FJLT of [2] for finite T, achieving $m = O(\varepsilon^{-2}\log|T|)$ but with the time to multiply Φx being $O(n\log n + m^3)$. Improvements to the $O(m^3)$ term are in [3, 4, 56, 66]. In these works Φ is the product of some sparse matrices and Fourier matrices, with the speed coming from the Fast Fourier Transform (FFT) [36]. This FFT-based approach can also be used to obtain fast RIP matrices for compressed sensing [26, 71, 28] and fast subspace embeddings for numerical linear algebra applications [72] (see also [77, 59] for refined analyses for the latter).

Another line of work, initiated in [1] and greatly advanced in [37], sought speedup by sparsifying Φ. If Φ has at most s non-zeroes per column, then Φx can be computed in time $s \cdot \|x\|_0$. After some initial improvements [51, 21], the best known s to date for JL with $m \lesssim \varepsilon^{-2}\log|T|$ is the sparse Johnson-Lindenstrauss Transform (SJLT) of [52], achieving $s \lesssim \varepsilon^{-1}\log|T| \lesssim \varepsilon m$. Furthermore, an example T exists requiring this bound on s up to $O(\log(1/\varepsilon))$ for any linear JL map [65]. Note though, that this is again 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 being the unit norm vectors in a d-dimensional subspace [32, 62, 64], or all elements of T having small ℓ_∞ norm [60]. However, our understanding for general T is non-existent. This brings us to the main question addressed in this work, where $S^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ and we assume $T \subseteq S^{n-1}$ and Φ is the SJLT.

Question 2. What relationship must s, m satisfy, in terms of the geometry of T, to ensure (1.2)?

We also note that while the FFT-based and sparse Φ approaches may seem orthogonal, they are not, as pointed out before [2, 60, 66]. The FJLT sets $\Phi = SP$ where P is some random preconditioning matrix that makes T "nice" with high probability, and S is, similarly to the SJLT, a random *sparse* matrix.

The analog of Question 2 for a standard gaussian matrix depends only on the ℓ_2-metric structure of T, since both ℓ_2-distances and gaussian matrices are invariant under orthogonal transformations. However a resolution of Question 2 cannot solely depend on the ℓ_2-metric structure of T since Φ must be sparse *in a particular basis*. Thus an answer to Question 2 must be more nuanced.

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 qualitatively either (1) 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 [13], manifold learning [75, 41], and constrained least squares problems such as the Lasso [76].

Theorem 3 (Main Theorem)**.** *Let $T \subset S^{n-1}$ and Φ be an SJLT with column sparsity s. Define $\kappa(T) \overset{\text{def}}{=} \kappa_{s,m}(T) = \max_{q \le \frac{m}{s}\log s}\{\frac{1}{\sqrt{qs}}(\mathbb{E}_\eta(\mathbb{E}_g \sup_{x\in T}|\sum_{j=1}^n \eta_j g_j x_j|)^q)^{1/q}\}$ where (g_j) are i.i.d. standard gaussian and (η_j) i.i.d. Bernoulli with mean $qs/(m\log s)$. If*

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

Then (1.2) *holds as long as s, m furthermore satisfy the condition* $(\log m)^2(\log n)^{5/2}\kappa(T) < \varepsilon$.

The complexity parameter $\kappa(T)$ may seem daunting at first, but Section 7 shows that it can be controlled quite easily for all the T we have come across in applications.

1.1 Applications

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

Finite $|T|$: Here $|T| < \infty$, for which the SJLT satisfies Eq. (1.2) with $s \lesssim \varepsilon^{-1}\log|T|$, $m \lesssim \varepsilon^{-2}\log|T|$ [52]. If also $T \subset B_{\ell_\infty^n}(\alpha)$, i.e. $\|x\|_\infty \le \alpha$ for all $x \in T$, [60] 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 a d-dimensional linear subspace $E \subset \mathbb{R}^n$. Here achieving Eq. (1.2) with $m \lesssim d/\varepsilon^2$ is possible [7, 32]. A distribution satisfying Eq. (1.2) for any d-dimensional subspace is known as an *oblivious subspace embedding (OSE)*. [72] pioneered the use of OSE's for fast 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 [42], k-means clustering [20, 33], canonical correlation analysis [8], support vector machines [68], ℓ_p regression [30, 80], ridge regression [59], streaming approximation of eigenvalues [6], and speeding up interior point methods for linear programming [58]. In many 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 [32] showed $s = 1$ with small m is sufficient, yielding algorithms for least squares regression and low-rank approximation with runtimes linear in the number of non-zero entries in A for sufficiently lopsided rectangular matrices.

Our theorem implies $s \underset{*}{<} 1$ and $m \underset{*}{<} d$ suffices, which is correct. Furthermore, a subset of our techniques reveal that if the maximum leverage score, or *incoherence*, $\alpha = \max_{1 \le i \le n} \|P_E e_i\|_2$ is at most $poly(\varepsilon / \log n)$, then $s = 1$ suffices. 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 $A \in \mathbb{R}^{n \times d}$ $(n \gg d)$, $b \in \mathbb{R}^n$, 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 [76], 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 x_* (see Appendix B for a definition). Suppose we wish to accelerate approximately solving the constrained least squares problem by instead computing $\tilde{x}_*$, the minimizer of $\|\Phi Ax - \Phi b\|_2^2$ subject to $x \in \mathcal{C}$. The work [70] showed that to guarantee $\|A\tilde{x}_* - b\|_2^2 \le (1 + \varepsilon)\|Ax_* - b\|_2^2$, it suffices that Φ satisfy two conditions, one of which is Eq. (1.1) for $T = AT_{\mathcal{C}}(x_*) \cap S^{n-1}$. [70] then analyzed dense random matrices for sketching constrained least squares problems. For example, for the Lasso if we are promised the optimal solution x_* is k-sparse, [70] shows $m \gtrsim \varepsilon^{-2} \max_{j=1,\ldots,d} \|A_j\|_2^2 \sigma_{\min,k}^{-2} k \log d$ suffices for A_j the jth column of A, and where $\sigma_{\min,k}$ is 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 full version). For example for the Lasso, we show that again it suffices that $m \underset{*}{>} \max_{j=1,\ldots,d} \|A_j\|_2^2 \sigma_{\min,k}^{-2} k \log d$, but for s we only need $s \underset{*}{>} \max_{\substack{1 \le i \le n \\ 1 \le j \le d}} A_{i,j}^2 \sigma_{\min,k}^{-2} k$. That is, the sparsity of Φ need only depend on the largest entry in A as opposed to the largest column norm in A, which 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 $\{e_{i_j}\}_{j \in \theta}$ (so $d = k$). Then T is simply the set of all k-sparse unit vectors of unit Euclidean norm: $S_{n,k} \overset{\text{def}}{=} \{x \in \mathbb{R}^n : \|x\|_2 = 1, \|x\|_0 \le k\}$ for $\|\cdot\|_0$ denoting support size. Φ satisfying (1.1) is then referred to as having the *restricted isometry property (RIP) of order* k with restricted isometry constant $\varepsilon_k = \varepsilon$ [25]. 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 [25, 24]. Unfortunately it is known for $\varepsilon = \Theta(1)$ that *any* RIP Φ with such small m must have $s \gtrsim m$ [65]. Related examples are 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 [13] (so that $|\Theta| < \binom{n}{k}$).

Our theorem also implies RIP matrices with $s, m \underset{*}{<} k \log(n/k)$. More generally, when a dictionary D is involved such that the subspaces $\text{span}(\{De_{i_j}\}_{j \in \theta})$ are all α-incoherent (as defined above), the sparsity can be improved to $s \underset{*}{<} 1 + (\alpha k \log(n/k))^2$. That is, for 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 $m \underset{*}{<} d + \log|\Theta|, s \underset{*}{<} \log|\Theta|$ or $m \underset{*}{<} d + \log|\Theta|, s \underset{*}{<} 1 + (\alpha \log|\Theta|)^2$. Previous work required m to depend on the *product* of d and $(\log|\Theta|)^c$

instead of the *sum* [64], including a nice recent improvement by Cohen [34], and is thus unsuitable for this application (RIP matrices with $\underset{*}{<}k^2$ rows are already attainable via simpler methods using incoherence; e.g. see [17, Proposition 1]). Iterative recovery algorithms such as CoSamp can also be used in model-based sparse recovery [13], 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 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" (all coordinates 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 $s \underset{*}{<} \log|\Theta| = \log(\binom{n/b}{k/b}) \lesssim (k/b)\log(n/k)$. 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 $m \lesssim k + (k/b)\log(n/k)$ [9]). 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, [49] 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 $\max_{i,j}|H_{i,j}| = O(1/\sqrt{n})$ (e.g. the Fourier matrix). Work of [26, 71, 28] shows Φ can then be a sampling matrix (one non-zero per row) with $m \lesssim \varepsilon^{-2}k\log(n/k)(\log k)^3$. Since randomly flipping the signs of every column in an RIP matrix yields JL [56], this also gives a good implementation of an FJLT. Our theorem recovers a similar statement, but using the SJLT instead of a sampling matrix and with $m \underset{*}{<} k$ and $s \underset{*}{<} 1$ for orthogonal H satisfying the weaker requirement $\max_{i,j}|H_{i,j}| = O(1/\sqrt{k})$.

Smooth manifolds: Suppose we are given several images of a human face with varying lighting and angle of rotation, or many sample handwritten images of letters. Though these inputs are high-dimensional (n being the number of pixels), we imagine all 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, 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 [75]. For human faces, [75] shows that different images of a human face can be well-represented by a 3-dimensional manifold, with parameters being brightness and two angles of rotation.

Baraniuk and Wakin [14] proposed using dimensionality reduction to first map $\mathcal{M}$ to $\Phi\mathcal{M}$, then learn the parameters of interest in the reduced space (for improved speed). Later sharper analyses are in [29, 43, 39]. Of interest are both (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 have their preimage (in $\mathcal{M}$) be a C^1 curve in $\mathcal{M}$. Then by (1) and (2), *geodesic* distances are preserved between $\mathcal{M}$ and $\Phi\mathcal{M}$.

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 the unit ball of X). We assume $\|F(x) - F(y)\|_2 \simeq \|x - y\|_2$ (where $A \simeq B$ denotes 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 $\rho_{\mathcal{M}}$ to ρ_{Fin}. Here $\rho_{\mathcal{M}}$ is geodesic distance on $\mathcal{M}$, and $\rho_{Fin}(E_x, E_y) = \|P_{E_x} - P_{E_y}\|_{\ell_2^n \to \ell_2^n}$ is the *Finsler distance*, for P_E the orthogonal projection onto E.

We want Φ satisfying $(1-\varepsilon)|\gamma| \le |\Phi(\gamma)| \le (1+\varepsilon)|\gamma|$ for all C^1 curves $\gamma \subset \mathcal{M}$, for $|\cdot|$ being curve length. It then suffices Φ satisfy Eq. (1.1) for $T = \bigcup_{x\in\mathcal{M}} E_x \cap S^{n-1}$ [39], an infinite union of subspaces. [39] showed it suffices $s = m \gtrsim d/\varepsilon^2$ with a *dense* matrix of subgaussian entries. For F as given above,

set T to preserve	**our m**	**our s**	**previous m**	**previous s**	**ref**
$\lvert T\rvert < \infty$	$\log \lvert T\rvert$	$\log \lvert T\rvert$	$\log \lvert T\rvert$	$\log \lvert T\rvert$	[50]
$\lvert T\rvert < \infty, \forall x \in T \\|x\\|_\infty \leq \alpha$	$\log \lvert T\rvert$	$\lceil \alpha \log \lvert T\rvert \rceil^2$	$\log \lvert T\rvert$	$\lceil \alpha \log \lvert T\rvert \rceil^2$	[60]
$E, \dim(E) \leq d$	d	1	d	1	[64]
$S_{n,k}$	$k \log(n/k)$	$k \log(n/k)$	$k \log(n/k)$	$k \log(n/k)$	[25]
$HS_{n,k}$	$k \log(n/k)$	1	$k \log(n/k)$	1	[71]
tangent cone for Lasso	$\max_j \frac{\\|A_j\\|_2^2}{\sigma_{min,k}^2} k$	$\max_{i,j} \frac{A_{i,j}^2}{\sigma_{min,k}^2} k$	same as here	$s = m$	[70]*
$\lvert\Theta\rvert < \infty$ $\forall E \in \Theta, \dim(E) \leq d$	$d + \log \lvert\Theta\rvert$	$\log \lvert\Theta\rvert$	$d \cdot (\log \lvert\Theta\rvert)^6$	$(\log \lvert\Theta\rvert)^3$	[64]
$\lvert\Theta\rvert < \infty$ $\forall E \in \Theta, \dim(E) \leq d$ $\max_{\substack{1 \leq j \leq n \\ E \in \Theta}} \\|P_E e_j\\|_2 \leq \alpha$	$d + \log \lvert\Theta\rvert$	$\lceil \alpha \log \lvert\Theta\rvert \rceil^2$	—	—	—
$\lvert\Theta\rvert$ infinite $\forall E \in \Theta, \dim(E) \leq d$	see appendix	see appendix (non-trivial)	similar to this work	m	[39]
$\mathcal{M}$ a smooth manifold	d	$1 + (\alpha\sqrt{d})^2$	d	d	[39]

Figure 1: The m, s that suffice when using the SJLT with various T via our main theorem, compared with best bounds from previous work. All bounds hide $\mathrm{poly}(\varepsilon^{-1} \log n)$ factors. Some cells are blank due to no non-trivial results being previously known. For the Lasso, we assume k is the sparsity of the true optimum.

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 $s \lesssim_* 1 + (\alpha\sqrt{d})^2$, where α is the largest incoherence of any tangent space E_x for $x \in \mathcal{M}$. Thus we have non-trivial sparsity with $m \lesssim_* d$ for $\alpha \ll 1/\sqrt{d}$. Furthermore, we show that this is *optimal* by constructing a manifold with maximum incoherence of a tangent space $1/\sqrt{d}$ such that preserving curve lengths with $m \gtrsim_* d$ requires $s \gtrsim_* d$ (see full version). We also show Φ is a manifold embedding with large probability if the weaker condition $m \gtrsim_* d, s \gtrsim_* 1$ is satisfied, implying that the SJLT also preserves geodesics.

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 section in the full version.

Due to space constraints, many results are stated without proof; proofs are contained in the attached full version. The full version also has several appendices to help the reader, by reviewing probabilistic tools and introductory convex analysis used in this work. We also show in Appendix C an analysis of using the FJLT for sketching constrained least squares, providing some improvements to [70].

2 Preliminaries

We fix some notation. Denote $[t] = \{1, \ldots, t\}$. To any $S \subset \mathbb{R}^n$ we associate a semi-norm $|||z|||_S \overset{\text{def}}{=} \sup_{x \in S} |\langle z, x\rangle|$. We use e_i to denote a standard basis vector. If $\eta = (\eta_i)_{i \geq 1}$ 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_η^p to denote the associated expected value and L^p-space, respectively. If ζ is another sequence of r.v.'s, then $\\| \cdot \\|_{L_\eta^p, L_\zeta^q}$ means that we first take the L_η^p-norm and afterwards the L_ζ^q-norm. We reserve the symbol g for standard gaussian vectors. For $A \in \mathbb{R}^{m \times n}$, $\\|A\\|$ and $\\|A\\|_{\ell_2^n \to \ell_2^m}$ are both the operator norm; $\mathrm{Tr}(\cdot)$ is trace; $\\| \cdot \\|_F$ is Frobenius norm.

In the remainder, we reserve the letter ρ to denote (semi-)metrics. If ρ is a (semi-)norm $\\| \cdot \\|_X$, 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 (1.2).

We make use of chaining results in the remainder, so we define some relevant quantities. For a (semi-)metric ρ on $\mathbb{R}^n$, Talagrand's γ_2-functional is defined by $\gamma_2(S,\rho) = \inf_{\{S_r\}_{r=0}^{\infty}} \sup_{x\in S} \sum_{r=0}^{\infty} 2^{r/2} \cdot \rho(x, S_r)$, 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 [44, 74]. 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 called Dudley's inequality (see [74]): $gamma_2(S,\rho) \lesssim \int_0^{\infty} \log^{1/2} \mathcal{N}(S, \rho, u)du$. 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 is a (semi-)norm $\|\cdot\|_X$, we write $\mathcal{N}(S, \|\cdot\|_X, u)$ instead of $\mathcal{N}(S, \rho_X, u)$.

Let us now introduce the SJLT. Let $\sigma_{ij} \in \{-1, 1\}$ be independent Rademachers (i.e. uniformly random signs). We consider random variables $\delta_{ij} \in \{0, 1\}$ satisfying:

- $\forall j$ the δ_{ij} are negatively correlated, i.e. $\mathbb{E} \prod_{t=1}^{k} \delta_{i_t,j} \le (s/m)^k$ for any k distinct indices i_t;
- For any fixed j there are exactly s nonzero $\delta_{i,j}$, i.e., $\sum_{i=1}^{m} \delta_{ij} = s$;
- The vectors $(\delta_{i,j})_{i=1}^{m}$ 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 SJLT is defined by $\Phi_{i,j} = (1/\sqrt{s})\sigma_{i,j}\delta_{i,j}$. [52] 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 $\delta_{i,j}$. The other example is essentially the CountSketch of [27]; for details see the full version.

In the following we will be interested in estimating $\sup_{x\in T} |\|\Phi x\|_2^2 - 1|$. For this purpose, we use the following bound [55] (see [38, Theorem 6.5] for the refinement stated here).

Theorem 4. *For $\mathcal{A} \subset \mathbb{R}^{m\times n}$ and (σ_j) independent Rademachers, $p \ge 1$, $(\mathbb{E}_\sigma \sup_{A\in\mathcal{A}} |\|A\sigma\|_2^2 - \mathbb{E}\|A\sigma\|_2^2|^p)^{1/p}$ is $\lesssim \gamma_2^2(\mathcal{A}, \|\cdot\|_{\ell_2\to\ell_2}) + d_F(\mathcal{A})\gamma_2(\mathcal{A}, \|\cdot\|_{\ell_2\to\ell_2}) + \sqrt{p}d_F(\mathcal{A})d_{\ell_2\to\ell_2}(\mathcal{A}) + pd_{\ell_2\to\ell_2}^2(\mathcal{A})$.*

For any $x \in \mathbb{R}^n$ we can write $\Phi x = A_{\delta,x}\sigma$, where

$$A_{\delta,x} := \frac{1}{\sqrt{s}} \sum_{i=1}^{m} \sum_{j=1}^{n} \delta_{ij} x_j e_i \otimes e_{ij} = \frac{1}{\sqrt{s}} \begin{bmatrix} -x^{(\delta_{1,\cdot})}- & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & -x^{(\delta_{m,\cdot})}- \end{bmatrix}. \tag{2.1}$$

for $x_j^{(\delta_{i,})} = \delta_{i,j}x_j$. 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 $\|x\|_\delta = (1/\sqrt{s}) \max_{1\le i\le m} (\sum_{j=1}^{n} \delta_{ij} x_j^2)^{1/2}$. Then for any $x, y \in T$ that $\|A_{\delta,x} - A_{\delta,y}\| = \|x - y\|_\delta$ and $\|A_{\delta,x} - A_{\delta,y}\|_F = \|x - y\|_2$. Therefore, by Theorem 4 and taking $L_p(\Omega_\delta)$-norms on both sides,

$$\Big(\mathop{\mathbb{E}}_{\Phi} \sup_{x\in T} \Big|\|\Phi x\|_2^2 - \|x\|_2^2\Big|^p\Big)^{1/p} \lesssim (\mathop{\mathbb{E}}_{\delta} \gamma_2^{2p}(T, \|\cdot\|_\delta))^{1/p} + (\mathop{\mathbb{E}}_{\delta} \gamma_2^{p}(T, \|\cdot\|_\delta))^{1/p} + \sqrt{p}(\mathbb{E}\, d_\delta^p(T))^{1/p} + p(\mathbb{E}\, d_\delta^{2p}(T))^{1/p}. \tag{2.2}$$

Thus, to bound the LHS of Eq. (1.2), it suffices to estimate $\mathbb{E}_\delta \gamma_2^2(T, \|\cdot\|_\delta)$ and $\mathbb{E}_\delta d_\delta^2(T)$. 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 $p \ge 1$, yield in addition a high probability bound. Unless stated otherwise, Φ always denotes the SJLT with column sparsity s.

3 Overview of proof of main theorem

Here we give an overview of the proof of Theorem 3. To illustrate the ideas, it is first simplest to consider the case of T being the set of all unit vectors in a d-dimensional linear subspace $E \subset \mathbb{R}^n$. By Eq. (2.2) for $p = 1$ we have to bound for example $\mathbb{E}_\delta\, \gamma_2(T, \|\cdot\|_\delta)$. Standard estimates give $\gamma_2(T, \|\cdot\|_\delta) \le \gamma_2(B_E, \|\cdot\|_\delta) \ll \sup_{t>0} t[\log \mathcal{N}(B_E, \|\cdot\|_\delta, t)]^{1/2}$ for B_E the unit ball of E. Let $U \in \mathbb{R}^{n\times d}$ have columns forming an orthonormal basis for E. Dual Sudakov minoration [18, Proposition 4.2], [67] states $\sup_{t>0} t[\log \mathcal{N}(B_E, \|\cdot\|_\delta, t)]^{1/2} \le \mathbb{E}_g \|Ug\|_\delta$ for a gaussian vector g. Then bounding $\mathbb{E}_g \|Ug\|_\delta$ is a probability exercise.

Unfortunately, dual Sudakov is specific to unit balls of subspaces and has no analog for general T. For general T we use a statement about the duality of entropy numbers [19, Proposition 4]. This states 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, and D° is the polar body; see Appendix B for a definition). Although it has been a conjecture for over 40 years as to whether this holds in general [69, p. 38], [19, Proposition 4] shows 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 (the norm whose unit vectors are those on D's boundary). In our case, this lets us relate $\log \mathcal{N}(\tilde{T}, \|\cdot\|_\delta, t)$ with $\{\log \mathcal{N}(\mathrm{conv}(B_{J_i}, |||\cdot|||_T, \sqrt{s}t/8)\}_{i=1}^m$, losing small factors, for $\tilde{T}$ the convex hull of $T \cup -T$, and B_{J_i} the unit ball of $\mathrm{span}\{e_j : \delta_{i,j} = 1\}$. 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 $\log \mathcal{N}(\mathrm{conv}(B_{J_i}), |||\cdot|||_T, \epsilon)$ 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 $k \lesssim 1/\epsilon^2$. For a fixed A, we bucket $j \in [n]$ according to $\sum_{i\in A} \delta_{i,j}$ and define $U_\alpha = \{j \in [n] : \sum_{i\in A} \delta_{i,j} \simeq 2^\alpha\}$. Abusing notation, we also let U_α denote the coordinate subspace spanned by $j \in U_\alpha$. This leads to (see Eq. (6.4)) the inequality $\log \mathcal{N}(\frac{1}{k}\sum_{i\in A} B_{J_i}, |||\cdot|||_T, \epsilon) \lesssim \sum_\alpha \log \mathcal{N}(B_{U_\alpha}, |||\cdot|||_T, \sqrt{\frac{k}{2^\alpha}\frac{epsilon}{\log m}})$.

Finally we are in a position to apply dual Sudakov minoration to the right hand side of the above, 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 called the *incoherence* $\mu(E)$ of E.

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

$$(\mathbb{E}\, \gamma_2^{2p}(T, \|\cdot\|_\delta))^{1/p} \lesssim \epsilon^2 + \frac{(d + \log m)\log^2(d/\epsilon)}{m} + \frac{p\log^2(d/\epsilon)\log m}{s}\mu(E)^2 \tag{4.1}$$

and

$$(\mathbb{E}\, d_\delta^{2p}(T))^{1/p} \lesssim \frac{d}{m} + \frac{p}{s}\mu(E)^2. \tag{4.2}$$

As a consequence, Eq. (1.1) *holds with probability at least* $1 - \eta$ *if* $\eta \le 1/m$ *and*

$$\begin{aligned} m &\gtrsim ((d + \log m)\min\{\log^2(d/\varepsilon), \log^2(m)\} + d\log(\eta^{-1}))/\varepsilon^2 \\ s &\gtrsim (\log(m)\log(\eta^{-1})\min\{\log^2(d/\varepsilon), \log^2(m)\} + \log^2(\eta^{-1}))\mu(E)^2/\varepsilon^2 \end{aligned} \tag{4.3}$$

Proof. By dual Sudakov minoration (see full version), $\log \mathcal{N}(B_E, \|\cdot\|_\delta, t) \lesssim (1/t)\,\mathbb{E}_g \|Ug\|_\delta$ for all $t > 0$, with $U \in \mathbb{R}^{n\times d}$ having columns an orthonormal basis for E and g gaussian. Let $U^{(i)}$ be U but where each row j is multiplied by $\delta_{i,j}$. A simple calculation with $\ell = \log m$ using gaussian concentration of Lipschitz functions (see full version) implies $\mathbb{E}_g \|Ug\|_\delta \lesssim \frac{1}{\sqrt{s}}(\max_{1\le i\le m} \|U^{(i)}\|_F + \sqrt{\ell}\cdot \max_{1\le i\le m} \|U^{(i)}\|_{\ell_2^d\to\ell_2^n})$.

Using Dudley's integral estimate, the full version shows for any $\epsilon > 0$ and $t^* = (\epsilon/d)/\log(d/\epsilon)$,

$$\begin{aligned}\gamma_2(T, \|\cdot\|_\delta) &\lesssim \int_0^{t^*} \sqrt{d}\cdot \left[\log\left(2+\frac{1}{t\sqrt{s}}\right)\right]^{1/2} dt + \int_{t^*}^{1/\sqrt{s}} \frac{\mathbb{E}_g \|Ug\|_\delta}{t} dt\\ &\lesssim \sqrt{d}t_*\left[\log\left(\frac{1}{t_*\sqrt{s}}\right)\right]^{1/2} + \mathop{\mathbb{E}}_g \|Ug\|_\delta \log\left(\frac{1}{t_*\sqrt{s}}\right)\\ &\lesssim \epsilon + \frac{\log(d/\epsilon)}{\sqrt{s}}\cdot\left[\max_{1\le i\le m}\left[\sum_{j=1}^n \delta_{i,j}\|P_E e_j\|_2^2\right]^{1/2} + \sqrt{\log m}\max_{1\le i\le m}\|U^{(i)}\|_{\ell_2^d\to\ell_2^n}\right] \qquad (4.4)\end{aligned}$$

As a consequence,

$$(\mathop{\mathbb{E}}_\delta \gamma_2^{2p}(T,\|\cdot\|_\delta))^{1/p} \lesssim \epsilon^2 + \frac{\log^2(d/\epsilon)}{s}\left[\left(\mathop{\mathbb{E}}_\delta \max_{1\le i\le m}\left[\sum_{j=1}^n \delta_{i,j}\|P_E e_j\|_2^2\right]^p\right)^{1/p} + \log(m)\left(\mathop{\mathbb{E}}_\delta\max_{1\le i\le m}\|U^{(i)}\|_{\ell_2^d\to\ell_2^n}^{2p}\right)^{1/p}\right] \qquad (4.5)$$

The first sum inside brackets is treated essentially by a standard Chernoff-type argument. The second summand is bounded by the non-commutative Khintchine inequality, leading to the bound in Theorem 5. $(\mathbb{E}\, d_\delta^{2p}(T))^{1/p}$ is bounded by a similar but simpler argument; see full version. □

Theorem 5 recovers a result similar to the main result of [64] but via a different method, less logarithmic factors in m, better dependence on $1/\eta$, and the revelation that s can be smaller if $\mu(E)$ is small (note if $\|P_E e_j\|_2 \ll (\log d\cdot \log m)^{-1}$ for all j, we may take $s = 1$). Our dependence on $1/\varepsilon$ in s is quadratic instead of the linear dependence in [64], though in most applications $\varepsilon = \Theta(1)$. Also note if $d \gtrsim \log n$, then a random d-dim. subspace E has $\mu(E) \lesssim \sqrt{d/n}$ by the JL lemma.

5 Sketching constrained least squares programs

Consider $A \in \mathbb{R}^{n\times d}$, with $n \ge d$, 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. Define the minimizers of the constrained least squares programs x_* to be $\mathrm{argmin} f(x)$ subject to $x \in \mathcal{C}$ and $\hat{x}$ to be $\mathrm{argmin} g(x)$ subject to $x \in \mathcal{C}$. We define two quantities introduced in [70]. Given $\mathcal{K} \subset \mathbb{R}^d$ and $u \in S^{n-1}$ we set $Z_1(A,\Phi,\mathcal{K}) = \inf_{v\in A\mathcal{K}\cap S^{n-1}} \|\Phi v\|_2^2$ and $Z_2(A,\Phi,\mathcal{K},u) = \sup_{v\in A\mathcal{K}\cap S^{n-1}} |\langle \Phi u, \Phi v\rangle - \langle u, v\rangle|$. 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 [70, Lemma 1]. The second statement (when x_* is the global minimizer) follows by a slight modification of their proof.

Lemma 6. *Define $u = (Ax_* - b)/\|Ax_* - b\|_2$. For $Z_1 = Z_1(A,\Phi,T_{\mathcal{C}}(x_*))$ and $Z_2 = Z_2(A,\Phi,T_{\mathcal{C}}(x_*),u)$. Then $f(\hat{x}) \le (1 + \frac{Z_2}{Z_1})^2 f(x_*)$. If x_* is the global minimizer of f, then $f(\hat{x}) \le (1 + \frac{Z_2^2}{Z_1^2}) f(x_*)$.*

We give a proof of the above in the full version. Clearly, if Φ satisfies (1.1) for $T = AT_{\mathcal{C}}(x_*) \cap S^{n-1}$ then $Z_1 \ge 1 - \varepsilon$. We do not immediately obtain an upper bound for Z_2, however, as u is in general not in $AT_{\mathcal{C}}(x_*) \cap S^{n-1}$. Nevertheless, we show the following in the full version.

Lemma 7. *Fix $u \in B_{\ell_2^n}$, $T \subset \mathbb{R}^n$ and let Φ be the SJLT. Set $Z = \sup_{v\in T} |\langle \Phi u, \Phi v\rangle - \langle u, v\rangle|$. Then for any $p \ge 1$, $(\mathbb{E}_{\delta,\sigma} Z^p)^{1/p} \lesssim (\sqrt{p/s} + 1)((\mathbb{E}_\delta \gamma_2^p(T, \|\cdot\|_\delta))^{1/p} + \sqrt{p}(\mathbb{E}_\delta d_\delta^p(T))^{1/p})$.*

5.1 $\ell_{2,1}$-constrained case

In the full version we discuss any convex set $\mathcal{C}$. For illustration, here we discuss a special case. 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 $\|x\|_{2,1} := \|x\|_{\ell_1^b(\ell_2^D)} = \sum_{\ell=1}^b \|x_{B_\ell}\|_2$. We study the effect of sketching on the problem

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

which is corresponds to $\mathcal{C} = \{x \in \mathbb{R}^d : \|x\|_{2,1} \leq R\}$. 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. [23] for more information. In the special case $D = 1$ the program reduces to $\min \|Ax - b\|_2^2$ subject to $\|x\|_1 \leq R$, which is the well-known *Lasso* [76]. To formulate our results we consider two norms on $\mathbb{R}^{n \times d}$, given by $|||A||| := \max_{1 \leq \ell \leq b} (\sum_{j=1}^n \sum_{k \in B_\ell} |A_{jk}|^2)^{1/2}$ and $\|A\|_{\ell_{2,1} \to \ell_\infty} = \max_{1 \leq j \leq n} \max_{1 \leq \ell \leq b} \|(A_{jk})_{k \in B_\ell}\|_2$. In the case $D = 1$, $|||A|||$ is the maximum Euclidean norm of a column of A, and $\|A\|_{\ell_{2,1} \to \ell_\infty}$ is the maximum magnitude of any entry of A. In the full version, we use the previous section to show the following, which is qualitatively similar to [70] but allows for a much sparser Φ (i.e. our s depends on the maximum entry of A as opposed to the maximum column norm when $D = 1$ for Lasso).

Theorem 8. *Let Φ be the SJLT. Set $\mathcal{C} = \{x \in \mathbb{R}^d : \|x\|_{2,1} \leq R\}$. Suppose x_* is k-block sparse, $\|x_*\|_{2,1} = R$. Define $\sigma_{\min,k} = \inf_{\|y\|_2=1,\ \|y\|_{2,1} \leq 2\sqrt{k}} \|Ay\|_2$ and $\alpha = (\log n)^6 (\log m)(\log b)^2$. Assume*

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

and $\eta \leq \frac{1}{m}$. Then with probability at least $1 - \eta$, $f(\hat{x}) \leq (1 - \varepsilon)^{-2} f(x^*)$.

6 Proof sketch of the main theorem

In the full version we show following inequality and lemma:

$$\log \mathcal{N}(\tilde{T}, \|\cdot\|_\delta, t) \lesssim \Big(\log \frac{1}{\sqrt{st}} \Big) (\log m) \log \mathcal{N}\Big(\mathrm{conv}\Big(\bigcup_{i=1}^m B_{J_i} \Big), ||| \cdot |||_T, \frac{1}{8}\sqrt{st} \Big) \tag{6.1}$$

Lemma 9. *Let $\epsilon > 0$. Then*

$$\log \mathcal{N}\Big(\mathrm{conv}\Big(\bigcup_{i=1}^m B_{J_i} \Big), ||| \cdot |||_T, \epsilon \Big) \lesssim \frac{1}{\epsilon^2} \log m + (\log 1/\epsilon) \max_{k \lesssim \frac{1}{\epsilon^2}} \max_{|A|=k} \log \mathcal{N}\Big(\frac{1}{k} \sum_{i \in A} B_{J_i}, ||| \cdot |||_T, \epsilon \Big) \tag{6.2}$$

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 $y = (1/k) \sum_{j=1}^n (\sum_{i \in A} \delta_{i,j} x_j^{(i)}) e_j$ with $\sum_{j \in J_i} |x_j^{(i)}|^2 \leq 1$ for all i. Therefore, by Cauchy-Schwarz,

$$\|y\|_2 = \frac{1}{k} \Big(\sum_{j=1}^n \Big| \sum_{i \in A} \delta_{i,j} x_j^{(i)} \Big|^2 \Big)^{1/2} \leq \frac{1}{k} \Big[\sum_{j=1}^n \Big(\sum_{i \in A} \delta_{i,j} \Big) \sum_{i \in A} |x_j^{(i)}|^2 \Big]^{1/2}$$

Define for $\alpha = 1, \dots, \log(\min\{k, s\})$ the set $U_\alpha = U_\alpha(\delta) = \{1 \le j \le n : 2^\alpha \le \sum_{i \in A} \delta_{i,j} < 2^{\alpha+1}\}$ and $U_0 = U_0(\delta) = \{1 \le j \le n : \sum_{i\in A} \delta_{i,j} < 2\}$. We show in the full version that for any fixed j,

$$\tau_{k,\alpha} \overset{\text{def}}{=} \underset{\delta}{\mathbb{P}}\left(2^\alpha \le \sum_{i \in A} \delta_{i,j} < 2^{\alpha+1}\right) \le \begin{cases} 1, \text{ if } 2^\alpha \le \frac{2esk}{m} \\ \min\left\{2^{-\alpha}\frac{sk}{m}, 2^{-2^\alpha}\right\}, \text{ if } 2^\alpha > \frac{2esk}{m}. \end{cases} \tag{6.3}$$

Write according to the preceding $y = \sum_\alpha y_\alpha$ 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 $B_{U_\alpha} := \{\sum_{j\in U_\alpha} x_j e_j \,:\, \sum_{j \in U_\alpha} |x_j|^2 \le 1\}$, we have $\frac{1}{k}\sum_{i\in A} B_{J_i} \subset \sum_\alpha \frac{1}{\sqrt{k}} 2^{\alpha/2} B_{U_\alpha}$. Therefore,

$$\log \mathcal{N}\Big(\frac{1}{k}\sum_{i \in A} B_{J_i}, |||\cdot|||_T, \epsilon\Big) \lesssim \sum_\alpha \log \mathcal{N}\Big(\frac{1}{\sqrt{k}} 2^{\alpha/2} B_{U_\alpha}, |||\cdot|||_T, \frac{\epsilon}{\log m}\Big) = \sum_\alpha \log \mathcal{N}\Big(B_{U_\alpha}, |||\cdot|||_T, \sqrt{\frac{k}{2^\alpha}}\frac{\epsilon}{\log m}\Big) \tag{6.4}$$

Now we are in familiar territory: on the RHS we would like to bound the covering number of the unit ball of a subspace under some norm (namely the subspace B_{U_α}). Then proceeding as in Section 4 using dual Sudakov minoration (see full version for details), we show

$$\Big[\log \mathcal{N}\Big(\text{conv}\Big(\bigcup_{i=1}^m B_{J_i}\Big), |||\cdot|||_T, \epsilon\Big)\Big]^{1/2} \le \frac{1}{\epsilon}(\log m)^{1/2} + \frac{\log m}{\epsilon}\Big(\log\frac{1}{\epsilon}\Big)^{1/2} \max_{\substack{k \lesssim \frac{1}{\epsilon^2} \\ |A| = k}} \Big[\sum_\alpha \sqrt{\frac{2^\alpha}{k}} \underset{g}{\mathbb{E}} \Big|\Big|\Big| \sum_{j \in U_\alpha} g_j e_j \Big|\Big|\Big|_T\Big] \tag{6.5}$$

We then show in the full version using standard arguments involving Dudley's inequality that

$$\gamma_2(T, \|\cdot\|_\delta) \lesssim \frac{1}{\sqrt{s}}(\log n)^{3/2} \log m + \frac{1}{\sqrt{s}}(\log m)^{3/2}(\log n)^2 \cdot \max_{k \le m} \max_{|A| = k} \Big\{\sum_{\alpha, 2^\alpha \le k} \sqrt{\frac{2^\alpha}{k}} \underset{g}{\mathbb{E}} \Big|\Big|\Big| \sum_{j \in U_\alpha} g_j e_j \Big|\Big|\Big|_T\Big\} \tag{6.6}$$

The above then just amounts to estimating the random variable $\mathbb{E}_g \,|||\sum_{j\in U_\alpha} g_j e_j|||_T$ for various α. We do this in the full version by splitting α into three regions and applying various probabilistic arguments for each case (see full version). We then conclude by bounding the above when taking expectation over δ and show (1.2) holds for $m \gtrsim (\log m)^3 (\log n)^5 \frac{\gamma_2^2(T)}{\varepsilon^2}$ and $s \gtrsim (\log m)^6 (\log n)^4 \frac{1}{\varepsilon^2}$ and $(\log m)^2 (\log n)^{5/2} \kappa(T) < \varepsilon$ for $\kappa_T \overset{\text{def}}{=} \max_{q \le \frac{m}{s}\log s}\{\frac{1}{\sqrt{qs}}(\mathbb{E}_\eta(\mathbb{E}_g \,|||\sum_{j=1}^n \eta_j g_j e_j|||_T)^q)^{1/q}\}$.

7 Example applications of main theorem

Understanding applications amounts to upper bounding $\gamma_2(T, \|\cdot\|_2)$ and $\kappa(T)$. Note however $\gamma_2(T, \|\cdot\|_2) \lesssim (\log s)^{1/2}\kappa(T)$. Indeed, take $q = \frac{m}{s}\log s$ in the definition of $\kappa(T)$; then $\eta_j = 1$. This gives $\kappa(T) \ge (\log s)^{-1/2} g(T) \simeq (\log s)^{-1/2}\gamma_2(T, \|\cdot\|_2)$. Thus, ignoring log factors, it suffices to bound $\kappa(T)$.

In the full version we show how to bound $\kappa(T)$ for several examples of T, including: finite T, flat T (i.e. all $x \in T$ have small $\|x\|_\infty$), T the set of k sparse vectors in a dictionary that is a bounded orthonormal system, finite unions of subspaces (both the general and the incoherent cases), infinite unions of subspaces, and manifolds. In the full version we also construct a manifold for which the maximum incoherence of any tangent space is approximately $1/\sqrt{d}$ such that Φ with $m \underset{*}{<} d$ requires $s \underset{*}{>} d$ (in contrast, we show that max incoherence $o(1/\sqrt{d})$ allows for non-trivially small $s \underset{*}{<} 1 + (\alpha\sqrt{d})^2$). We also show that under the weak conditions $m \underset{*}{>} d$ and $s \underset{*}{>} 1$, the SJLT even preserves geodesic distances.

References

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

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

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

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

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

[6] 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.

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

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

[17] 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.

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

[19] 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.

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

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

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

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

[24] Emmanuel Candès. The restricted isometry property and its implications for compressed sensing. *C. R. Acad. Sci. Paris*, Ser. I(346):589–592, 2008.

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

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

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

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

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

[37] 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.

[38] Sjoerd Dirksen. Tail bounds via generic chaining. *arXiv*, abs/1309.3522, 2013.

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

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

[41] 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.

[42] 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.

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

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

[45] 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.

[46] 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.

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

[48] 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.

[49] 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.

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

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

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

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

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

[55] 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.

[56] 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.

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

[58] 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.

[59] 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.

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

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

[62] 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.

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

[64] 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.

[65] 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.

[66] 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.

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

[68] 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.

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

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

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

[72] 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.

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

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

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

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

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

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

[79] 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.

[80] 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.

FULL VERSION STARTS ON NEXT PAGE.