An Almost Optimal Unrestricted Fast Johnson-Lindenstrauss Transform

Nir Ailon* Edo Liberty†

May 30, 2010

Abstract

The problems of random projections and sparse reconstruction have much in common and individually received much attention. Surprisingly, until now they progressed in parallel and remained mostly separate. Here, we employ new tools from probability in Banach spaces that were successfully used in the context of sparse reconstruction to advance on an open problem in random pojection. In particular, we generalize and use an intricate result by Rudelson and Vershynin for sparse reconstruction which uses Dudley's theorem for bounding Gaussian processes. Our main result states that any set of $N = \exp(\tilde{O}(n))$ real vectors in n dimensional space can be linearly mapped to a space of dimension $k = O(\log N \operatorname{polylog}(n))$, while (1) preserving the pairwise distances among the vectors to within any constant distortion and (2) being able to apply the transformation in time $O(n \log n)$ on each vector. This improves on the best known $N = \exp(\tilde{O}(n^{1/2}))$ achieved by Ailon and Liberty and $N = \exp(\tilde{O}(n^{1/3}))$ by Ailon and Chazelle. The dependence in the distortion constant however is believed to be suboptimal and subject to further investigation. For constant distortion, this settles the open question posed by these authors up to a $\operatorname{polylog}(n)$ factor while considerably simplifying their constructions.

1 Introduction

Designing computationally efficient transformations that reduce dimensionality of data while approximately preserving its metric information lies at the heart of many problems. While in compressed sensing such techniques are sought for sparse data in a real or complex metric space (with respect to some basis), in random projections, following the seminal work of Johnson and Lindenstrauss, one seeks to reduce dimension of any set of finite data.[1] In both applications, random matrices of a suitable size [1][2][3][4] result in optimal construction [5] in the parameters n (the original dimension), k (the target dimension), N (the number of input vectors) and δ (the distortion). However, these constructions' resulting running time complexity, measured as number of operations needed in order to map a vector, is suboptimal. A major open question is that of designing such matrix distributions that can be applied efficiently to any vector, with optimal dependence in the parameters n, k, N and δ. Applications for such transformations were found e.g. in designing fast approximation algorithms for solving large scale linear algebraic operations (e.g. [6, 7]).

Although random projection and compressed sensing have much in common, they have mostly progressed in parallel. Here we combine recent work on bounds for sparse reconstruction to improve bounds of Ailon and Chazelle [8, 9] and Ailon and Liberty [10] on fast random projections, also known as Fast Johnson-Lindenstrauss transformations. The new bounds allow obtaining the well known Fast Johnson-Lindenstrauss Transform for finite sets of bounded cardinality $N = \exp(\tilde{O}(n))$ where n is the original dimension. The best known so far was obtained by Ailon and Liberty for sets of size up to $N = \exp\{\tilde{O}(n^{1/2})\}$.[2] The latter improved on Ailon and Chazelle's original bound of $N = \exp\{O(n^{1/3})\}$, which initiated the construction of Fast Johnson-Lindenstrauss Transforms. We also mention Dasgupta et al.'s work [11] on construction of Johnson-Linenstrauss random matrices which can be more efficiently applied to sparse vectors, with applications in the streaming model, and Ailon et al's work [12] on design of Johnson-Lindenstrauss matrices that run in linear time under certain assumptions on various norms of the input vectors. [3]

*Technion, Haifa, Israel, nailon@gmail.com

†Yahoo! Research, Haifa, Israel, edo@yahoo-inc.com

[1]The term "random projections" describes Johnson and Lindenstrauss's original construction. It became synonymous with the process of approximate metric preserving dimension reduction using randomized linear mappings. However, these linear mappings need not be (and indeed are usually not) projections in the linear algebraic sense of the word.

[2]The notation $\tilde{O}(\cdot)$ suppresses arbitrarily small polynomial coefficients and polylogarithmic factors.

[3]In previous work by Ailon and Chazelle [8, 9] and Ailon and Liberty [10], a different notation was used. The number of vectors was n, the original dimension was d and the distortion parameter was ε. Here, we chose to follow the notation used by Rudelson and Vershynin [13], since our construction and analysis closely

The transformation we derive here is a composition of two random matrices: A random sign matrix and a random selection of a suitable number k of rows from a Fourier matrix, where $k = O(\delta^{-4}(\log N)\,\mathrm{polylog}(n))$, and δ is the tolerated distortion level. The result, for constant δ, is believed to be suboptimal within the $\mathrm{polylog}(n)$ factor in the target dimension k. The running time of performing the transformation on a vector is dominated by the $O(n \log n)$ of the Fast Fourier Transform, and is believed to be optimal. The possibility of obtaining such a running time for fixed distortion was left as an open problem in Ailon and Chazelle and Ailon and Liberty's work, and here we resolve it up to a factor of $\mathrm{polylog}(n)$. The dependence on the constant δ is also believed to be suboptimal, and the "correct" dependence should be δ^{-2}. The question of improving this dependence is left as an open problem.

The use of a combination of random sign matrices and various forms of subsampled Fourier matrices was also used for random projections in the work of Ailon and Chazelle [8] and later Ailon and Liberty [10], as well as that of Matousek [14].[4] Here we obtain improved analysis using recent work by Rudelson and Vershynin for sparse reconstruction [13].

1.1 Restricted Isometry An underlying idea common to both random projections and sparse reconstruction is the preservation of metric information under a dimension reducing transformation. In sparse reconstruction theory, this property is known as *restricted isometry* [15][16]. A matrix Φ is a restricted isometry with sparseness paramater r if for some $\delta > 0$,

(1.1)

$$\forall \ r\text{-sparse } y \in \mathbb{R}^n \quad (1-\delta)\|y\|_2^2 \leq \|\Phi y\|_2^2 \leq (1+\delta)\|y\|_2^2 \ .$$

By r-sparse y we mean vectors in $\mathbb{R}^n$ with all but at most r coordinates zero. It was shown in [15] that the restricted isometry property is sufficient for the purpose of perfect reconstruction of sparse vectors, *compressed sensing* being one of the prominent applications.

In [13], Rudelson and Vershynin construct a distribution over $k \times n$ matrices Φ such that, with high probability, Φ has the restricted isometry property with sparseness parameter r and arbitrarily small $\delta > 0$.[5] In their analysis, $k = O(\delta^{-2} r \log(n) \cdot \log^2(r) \log(r \log n))$ and Φ can be applied (to a given vector x) in running time $O(n \log n)$. Assuming r polynomial in n, this takes the simpler form of $k = O(\delta^{-2} r \log^4 n)$.[6] In fact, Φ is (up to a constant) nothing other than a random choice of k rows from the (unnormalized) Hadamard matrix, defined as $\Psi_{\omega,t} = (-1)^{\langle \omega, t \rangle}$, where $\langle \cdot, \cdot \rangle$ is the dot product over the binary field, n is assumed to be a power of 2 and ω, t are thought of as $\log n$ dimensional vectors over the binary field in an obvious way.[7] As a corollary of the result, one obtains a universal matrix for reconstructing sparse signals, which can be applied to a vector in time $O(n \log n)$. The conjecture is that the same distribution with $k = O(\delta^{-2} r \log n)$ should work as well, but this is a major open question beyond the scope of this work. For an excellent survey explaining how restricted isometry can be used for sparse reconstruction, and why designing such matrices with good computational properties is important we refer the readers to [17] and to references therein.

Independently, Ailon and Chazelle [8] and Ailon and Liberty [10] were interested in constructing a distribution of $k \times n$ matrices Φ such that for any set $Y \subseteq \mathbb{R}^n$ of cardinality N, one gets

$$\forall \ y \in Y \ \ (1-\delta)\|y\|_2^2 \leq \|\Phi y\|_2^2 \leq (1+\delta)\|y\|_2^2, \tag{1.2}$$

with constant probability. Additionally, the number of steps required for applying Φ on any given x is $O(n \log n)$. In their result k was taken as $O(\delta^{-2} \log N)$, which is also essentially the best possible [5]. Unfortunately, both results break down when $k = \Omega(n^{1/2})$.[8] Assuming the tolerance parameter δ fixed, this limitation can be rephrased as follows: The techniques fail when the number of vectors N is in $\exp\{\Omega(n^{1/2})\}$.

In both Ailon and Chazelle [8] and Ailon and Liberty's [10] results, as well as in previous work [1][2][3][14][4] the bounds (1.2) are obtained by proving strong tail bounds on the distribution of the estimator $\|\Phi y\|_2$, and then applying a simple union bound on the finite collection Y. It is worth a moment's thought to realize that Ailon and Chazelle's result as well as that of Ailon and Liberty can be used for restricted isometry as well. Indeed, a simple epsilon-net argument for the set of r-sparse vectors can turn that set into a finite set of $\exp\{O(r \log n)\}$ vectors, on which a union bound can be applied. However, the current limitation of random projections mentioned above will limit r to

follow their techniques.

[4] In fact, in [10] the combination of random signs and Fourier is applied iteratively many times.

[5] Their analysis is done over the complex field, but we restrict the discussion to the reals here.

[6] In their work, the dependence of k on δ is not analyzed because δ is assumed to be fixed (for sparse signal reconstruction purposes, this dependence is not important). It is not hard to derive the quadratic dependence of k in δ^{-1} from their work.

[7] Rudelson and Vershynin use the complex Discrete Fourier Transform matrix, but their analysis does not change when using the Hadamard matrix.

[8] Ailon and Chazelle [8] and Ailon and Liberty [10] used d to denote the data dimension, n its cardinality and ε the sought distortion bound. Here we follow Rudelson and Vershynin's convention using n to denote the dimension and δ the distortion bound. We now use N to denote the data cardinality.

be in $n^{O(1/2-\mu)}$ (for arbitrarily small μ). Interestingly, Rudelson and Vershynin's result does not break down for r polynomial in n. A careful inspection of their techniques reveals that instead of union bounding on a finite set of strongly concentrated random variables, they use a result due to Dudley to bound extreme values of Gaussian processes. Can this idea be used to improve [8] and [10]? Intuitively there is no reason why a result which is designed for preserving the metric of sparse vectors should help with preserving the metric of any finite set of vectors. It turns out, luckily, that such a reduction can be done, though not in an immediate way.

1.2 Our Result A suitable generalization of Rudelson and Vershynin's result (Section 2), combined with Ailon and Chazelle [8] and Ailon and Liberty's [10] method of random sign matrix preconditioning achieves our main result (Theorem 3.1) which can be formulated as follows: Assume we have a set of N column vectors in an n dimensional real space. Fix an error parameter δ and pick (1) a $k \times n$ matrix Φ, with $k = O(\delta^{-4}\log(N)\log^4 n)$, drawing each row uniformly at random from the $n \times n$ Hadamard matrix, and (2) an $n \times n$ diagonal matrix D with each diagonal element drawn uniformly from the set $\{-1, 1\}$. Multiplying any vector in our set by ΦD requires $O(n \log n)$ operations, and with high (constant) probability uniformly preserves the N vector norms by a relative error of δ.

1.3 Notation In what follows, we fix N to denote the cardinality of a set Y of vectors in $\mathbb{R}^n$, where n is fixed. We also fix a distortion parameter $\delta \in (0, 1/2]$, and define k to be an integer in $\Theta(\delta^{-4}(\log N)(\log^4 n))$.

Now let Φ be a random $k \times n$ matrix obtained as follows: Pick k random rows, with repetition, from the unnormalized $n \times n$ Hadamard matrix (the Euclidean norm of each column in the resulting matrix Φ is $\sqrt{k}$). Let Ω denote the probability space for the choice of Φ.

Let b denote a uniformly chosen vector in $\{-1, 1\}^n$, and let Γ denote the probability space on the choice of b. For a vector $y \in \mathbb{R}^n$, we denote by D_y the diagonal $n \times n$ matrix with the coordinates of y on the diagonal. For a real matrix, $\|\cdot\|$ denotes its spectral norm and $(\cdot)^t$ its transpose. For a set $T \subseteq \{1, \ldots n\}$, we let Id_T denote the diagonal matrix with $\mathrm{Id}_T(i, i) = 1$ if $i \in T$, and 0 otherwise. For a vector $y \in \mathbb{R}^n$, let $\mathrm{supp}(y)$ denote the support of y, namely, its set of nonzero coordinates. For a number $p \geq 1$, let $B_p \subseteq \mathbb{R}^n$ denote the set of vectors $y \in \mathbb{R}^n$ with $\|y\|_p \leq 1$ and αB_p as the set of vectors $y \in \mathbb{R}^n$ for which $\|y\|_p \leq \alpha$.

2 Restricted isometry result generalization

We follow the main path of Rudelson et al. in [13] to prove a more general formulation of their main theorem which is more suitable for us here.

THEOREM 2.1. *[Derived from Rudelson and Vershynin[13]] Let $\alpha > 0$ be any real number. Define E_α as*

$$E_\alpha = E_\Omega\left[\sup_{y \in B_2 \cap \alpha B_\infty}\left\|D_y^2 - \frac{1}{k}D_y\Phi^t\Phi D_y\right\|\right] . \tag{2.3}$$

Then for some global $C_1 > 0$,

$$E_\alpha \leq \frac{C_1 \log^{3/2}(n)\log^{1/2}(k)}{\sqrt{k}}(E_\alpha + \alpha^2)^{1/2} . \tag{2.4}$$

In particular, if $\frac{(\log^{3/2} n)(\log^{1/2} k)}{\sqrt{k}} = O(\alpha)$, then

$$E_\alpha = O\left(\frac{\alpha(\log^{3/2} n)(\log^{1/2} k)}{\sqrt{k}}\right) . \tag{2.5}$$

The proof we present is an adaptation of the proof of Theorem 3.6 in [13] to a more general setting. In fact, the latter theorem [13] can be obtained as an easy consequence of theorem 2.1 by replacing $\sup_{y \in B_2 \cap \alpha B_\infty}$ in (2.3) by $\sup_{y \in \frac{1}{\sqrt{r}} Y_r}$ where $Y_r \subseteq \mathbb{R}^n$ is defined as the set of vectors with at most r coordinates equalling 1 and the remaining coordinates zero. Indeed, $\frac{1}{\sqrt{r}} Y_r \subseteq B_2 \cap r^{-1/2} B_\infty$. We can therefore conclude that for $\alpha = \frac{1}{\sqrt{r}}$, by definition,

$$E_\Omega\left[\sup_{y \in \frac{1}{\sqrt{r}} Y_r}\left\|D_y^2 - \frac{1}{k}D_y\Phi^t\Phi D_y\right\|\right] \leq E_\alpha .$$

If we also assume that $k = \Theta(r \log^4 n)$, then (2.5) will hold, from which we conclude that

$$E_\Omega\left[\sup_{y \in \frac{1}{\sqrt{r}} Y_r}\left\|D_y^2 - \frac{1}{k}D_y\Phi^t\Phi D_y\right\|\right] \leq O\left(\frac{(\log^{3/2} n)(\log^{1/2} k)}{\sqrt{rk}}\right) . \tag{2.6}$$

Now we notice that $D_y = \frac{1}{\sqrt{r}}\mathrm{Id}_{\mathrm{supp}\, y}$, where for a set of indexes T the diagonal matrix Id_T (as defined in [13]) has 1 in diagonal position i if and only if $i \in T$. Using this observation and multiplying (2.6) by r we conclude that

$$E_\Omega\left[\sup_{|T| \leq r}\left\|\mathrm{Id}_T - \frac{1}{k}\mathrm{Id}_T\,\Phi^t\Phi\,\mathrm{Id}_T\right\|\right] \leq O\left(\frac{\sqrt{r}(\log^{3/2} n)(\log^{1/2} k)}{\sqrt{k}}\right) ,$$

which is exactly the main result of Rudelson and Vershynin in [13] for restricted isometry.

The proof of Theorem 2.1 below points out the necessary changes to the proof of Theorem 3.6 in [13]. The difference between the theorems is that in our case, the supremum in the definition of E_α is taken not only over the set of sparse vectors, but over a richer set. It turns out however that [13] uses sparsity in a very limited way: In fact, the dominating effect of sparsity there is obtained using the fact that the L_1 norm of a sparse vector is small, compared to its L_2 norm. These arguments appear at the very end of their proof. For the sake of contributing to the self containment of the paper we walk through the main milestones of the proof of Theorem 3.6 in [13], and point out the changes necessary for our purposes. The reader is nevertheless encouraged to refer to the enlightening exposition in [13] first.

Proof. Clearly $E[\frac{1}{k} D_y \Phi^t \Phi D_y] = D_y^2$. We define new independent random i.i.d. variables $\{\epsilon_1, \ldots, \epsilon_n\}$ obtaining each the values $\{+1, -1\}$ with equal probability. Let Π denote the probability space for $\{\epsilon_1, \ldots, \epsilon_n\}$. It suffices to prove (using a symmetrization argument, see Lemma 6.3 in [18]) that

$$E_{\Omega\times\Pi}\left[\sup_{y\in B_2\cap\alpha B_\infty}\left\|\frac{1}{k}\sum_{i=1}^{k}\epsilon_i(x_iD_y)^t(x_iD_y)\right\|\right] \leq \frac{2C_1(\log^{3/2} n)(\log^{1/2} k)}{\sqrt{k}}(E_\alpha+\alpha^2)^{1/2}, \tag{2.7}$$

where x_i is the (random) i'th row of Φ. To that end, as claimed in [13] (Lemma 3.8), if we can show that for any fixed choice of Φ,

$$E_\Pi\left[\sup_{y\in B_2\cap\alpha B\infty}\left\|\sum_{i=1}^{k}\epsilon_i(x_iD_y)^t(x_iD_y)\right\|\right] \leq k_1 \sup_{y\in B_2\cap\alpha B_\infty}\left\|\sum_{i=1}^{k}(x_iD_y)^t(x_iD_y)\right\|^{1/2} \tag{2.8}$$

for some number k_1, then by taking E_Ω on both sides and using Jensen's inequality (to swap $(\cdot)^{1/2}$ on the RHS with E_Ω) and the triangle inequality, the conclusion would be that

$$E_\alpha \leq \frac{2k_1}{\sqrt{k}}\left(E_\alpha + \|D_y^2\|\right)^{1/2} \ . \tag{2.9}$$

Since $\|D_y^2\| = \|y\|_\infty^2 \leq \alpha$, we would get the stated result. It thus suffices to prove (2.8) with $k_1 = O((\log^{3/2} n)(\log^{1/2} k))$. To do so, [13] continue by replacing the k binary random variables $\epsilon_1, \ldots, \epsilon_k$ in (2.8) with k Gaussian random variables $g_1, \ldots, g_k$ using a comparison principle (inequality (4.8) in [18]), reducing the problem to that of bounding the expected extreme value of a Gaussian process. Using Dudley's inequality (Theorem 11.17 in [18]), as Rudelson and Vershynin do, one concludes that (2.8) will hold with k_1 taken as:

$$\int_0^\infty \log^{1/2}\mathcal{N}(B, \|\cdot\|_X, u)du \ , \tag{2.10}$$

where:

- For a norm $\|\cdot\|_\star$, a set S and number u, $\mathcal{N}(S, \|\cdot\|_\star, u)$ denotes the minimal number of balls of radius u in norm $\|\cdot\|_\star$ centered in points of S needed to cover the set S,
- B is defined as $\cup_{y\in B_2\cap\alpha B_\infty} B_y$, where $B_y = \{D_y z : z \in B_2\}$, and
- $\|x\|_X = \max_{i\leq k}|\langle x_i, x\rangle|$, where we remind the reader that x_i is the $i'th$ row of Φ.

Rudelson and Vershynin derive bounds on $\mathcal{N}(B_{RV}, \|\cdot\|_X, u)$ for small u and for large u separately, where in their case B_{RV} was the set of r-sparse vectors of Euclidean norm 1 (denoted by $D_2^{r,n}$ in [13]). The sparsity of the vectors in the set B_{RV} is used in both derivations, as follows:

- For large u, a containment argument is used in [13], asserting that $B_{RV} \subseteq \sqrt{r}B_1$. Note that by Cauchy Schwartz and the definition of B, $B \subseteq B_1$, hence we can also use an L_1 bound on the elements of B to bound $\mathcal{N}(B, \|\cdot\|_X, u)$. Indeed, by definition of $\mathcal{N}$, $\mathcal{N}(B, \|\cdot\|_X, u) \leq \mathcal{N}(B_1, \|\cdot\|_X, u)$. Using the probabilistic method, the details of which can be found in [13], the following bound can be obtained:
$$\mathcal{N}(B_1, \|\cdot\|_X, u) \leq (2n)^{O((\log k)/u^2)} \ .$$
- For small u, we note again that with respect to the norm $\|\cdot\|_X$, the set has diameter at most 2. Indeed, for any two points $z_1, z_2 \in B$,
$$\|z_1 - z_2\|_X = \max_{i\leq k}|\langle x_i, z_1 - z_2\rangle| \leq \max_{i\leq k}\|x_i\|_\infty\|z_1 - z_2\|_1 \leq 2 \ ,$$
the last inequality from $\|\Phi\|_\infty = 1$ together with our above assertion that $B \subseteq B_1$. A volumetric argument [19] is used to then conclude that
$$\mathcal{N}(B, \|\cdot\|_X, u) \leq (1 + O(1/u))^n \ .$$

Following Rudelson and Vershyni's final step in [13], we derive a bound for the integral $\int_0^\infty \mathcal{N}^{1/2}(B, \|\cdot\|_X, u)du$ by balancing the two bounds at $u = 1/\sqrt{n}$ as follows:

$$\int_0^\infty \log^{1/2} \mathcal{N}(B, \|\cdot\|_X, u)du \tag{2.11}$$

$$\leq \sqrt{n} \int_0^{1/\sqrt{n}} \sqrt{\log(1 + O(1/u))}du \tag{2.12}$$

$$+O(\sqrt{(\log k)(\log n)}) \int_{1/\sqrt{n}}^\infty \frac{1}{u}du \tag{2.13}$$

$$= O(\log n \sqrt{(\log n \log k)}) \ . \tag{2.14}$$

The conclusion is that we can take k_1 to be $O\left((\log n)(\sqrt{\log n})(\log k)\right) = O\left((\log^{3/2} n)(\log k)\right)$, as required.

3 Random Projections

Our main result claims that the same construction used by Rudelson et al. also gives improved bounds for random projections. In what follows, we fix r to be $\lceil \delta^{-2} \log N \rceil$ and α to be $1/\sqrt{r}$. Additionally, we assume that Φ is such that

$$\sup_{y \in B_2 \cap \alpha B_\infty} \left\| D_y^2 - \frac{1}{k} D_y \Phi^t \Phi D_y \right\| = O(\alpha^2) \ . \tag{3.15}$$

Indeed, Theorem 2.1 and the choice of our parameters guarantee that this holds with probability at least 0.99 in Ω.

THEOREM 3.1. *Let $Y \subseteq B_2$ denote a set of cardinality N, and let Φ satisfy (3.15). With probability at least 0.98 (in Γ) we have the following uniform bound for all $y \in Y$:*

$$1 - O(\delta) \leq \left\| \frac{1}{\sqrt{k}} \Phi D_y b \right\| \leq 1 + O(\delta) \ .$$

We provide some intuition for the proof. We split our input vectors Y into sums of two vectors, one of which is r-sparse and the other with ℓ_∞ norm bounded by $1/\sqrt{r}$. We use Rudelson et al.'s original result for the sparse part and our generalization of it (Theorem 2.1), together with Talagrand's measure concentration theorem for the ℓ_∞-bounded part.

Proof. Let r and α be defined as in Section 2. For each $y \in Y$ we write $y = \hat{y} + \check{y}$, where $\hat{y}$ is the restriction of y to its r largest (in absolute value) coordinates and $\check{y}$ is the restriction to its remaining coordinates. Note that $\|y\|^2 = \|\hat{y}\|^2 + \|\check{y}\|^2$ and that $\hat{y}$ is r-sparse and that $\|\check{y}\|_\infty \leq \alpha$.

$$\begin{aligned} \left\| \frac{1}{\sqrt{k}} \Phi D_y b \right\|^2 &= \left\| \frac{1}{\sqrt{k}} \Phi D_{\hat{y}} b \right\|^2 \\ &+ \left\| \frac{1}{\sqrt{k}} \Phi D_{\check{y}} b \right\|^2 + \frac{2}{k} b^t D_{\hat{y}} \Phi^t \Phi D_{\check{y}} b. \end{aligned}$$

For the first term we have $\left\| \frac{1}{\sqrt{k}} \Phi D_{\hat{y}} b \right\|^2 = \|\hat{y}\|^2 + O(\delta)$. This stems from the facts that $\hat{y}$ is r-sparse and that Φ exhibits the RIP property. This happens with probability 0.99 over Ω, see discussion of Theorem 2.1.

In what follows we will use the bound on $\|\check{y}\|_\infty$ to show that with high probability, for all $y \in Y$, $\left\| \frac{1}{\sqrt{k}} \Phi D_{\check{y}} b \right\|^2 = \|\check{y}\|^2 + O(\delta)$. A similar argument will bound the cross product $\frac{2}{k} b^t D_{\hat{y}} \Phi^t \Phi D_{\check{y}} b$. Combining the three gives the desired result that $\left\| \frac{1}{\sqrt{k}} \Phi D_y b \right\|^2 = \|y\|^2 + O(\delta)$.

We start by analyzing the measure concentration properties of $\left\| \frac{1}{\sqrt{k}} \Phi D_{\check{y}} b \right\|^2$. Let $X_{\check{y}}$ be the Rademacher random variable defined by

$$X_{\check{y}} = \left\| \frac{1}{\sqrt{k}} \Phi D_{\check{y}} b \right\| \ .$$

Let $\mu_{\check{y}}$ denote a median of $X_{\check{y}}$. By Talagrand [18], we have that for all $t > 0$,

$$\Pr[X_{\check{y}} > \mu_{\check{y}} + t] \leq \exp\{-C_2 t^2/\sigma_{\check{y}}^2\} \tag{3.16}$$

$$\Pr[X_{\check{y}} < \mu_{\check{y}} - t] \leq \exp\{-C_2 t^2/\sigma_{\check{y}}^2\} \tag{3.17}$$

for some global C_2, where $\sigma_{\check{y}} = \left\| \frac{1}{\sqrt{k}} \Phi D_{\check{y}} \right\|$. By the triangle inequality and Equation (3.15) we have $\sigma_{\check{y}}^2 = \|\frac{1}{k} D_{\check{y}} \Phi^t \Phi D_{\check{y}} - D_{\check{y}}^2 + D_{\check{y}}^2\| \leq \alpha^2 + \|D_{\check{y}}^2\|$. Clearly $\|D_{\check{y}}\| = \|\check{y}\|_\infty \leq \alpha$. Hence, $\sigma_{\check{y}}^2 = O(\alpha^2)$. From the fact that $E[X_{\check{y}}^2] = \|\check{y}\|^2$ and using Appendix A and (3.16)-(3.17) we conclude that $\|\check{y}\| - O(\sigma_{\check{y}}) \leq \mu_{\check{y}} \leq \|\check{y}\| + O(\sigma_{\check{y}})$. Hence, again using (3.16)-(3.17) and union bounding over the N vectors in Y, we conclude that with probability 0.99, uniformly for all $y \in Y$:

$$\|\check{y}\| - O(\delta) \leq \frac{1}{\sqrt{k}} \|\Phi D_{\check{y}} b\| \leq \|\check{y}\| + O(\delta) \ .$$

We now bound the cross term $Z = \frac{1}{k} b^t D_{\hat{y}} \Phi^t \Phi D_{\check{y}} b$ (y is now held fixed). By disjointness of $\mathrm{supp}(\hat{y})$ and $\mathrm{supp}(\check{y})$, $E[Z] = 0$. Decompose b into $\check{b} + \hat{b}$, where $\mathrm{supp}(\check{b}) = \mathrm{supp}(\check{y})$ and $\mathrm{supp}(\hat{b}) = \mathrm{supp}(\hat{y})$. For any fixed $\hat{b}$, the function Z is linear (and hence convex) in $\check{b}$. Also for all possible values $\hat{b}'$ of $\hat{b}$, $E[Z|\hat{b} = \hat{b}'] = 0$. Hence, again by Talagrand,

$$\Pr[Z > \mu_{\hat{b}'} + t] \leq \exp\{-C_2 t^2/\sigma_{\hat{b}'}^2\} \tag{3.18}$$

$$\Pr[Z < \mu_{\hat{b}'} - t] \leq \exp\{-C_2 t^2/\sigma_{\hat{b}'}^2\} \tag{3.19}$$

where $\mu'_{\hat{b}}$ is a median of $(Z|\hat{b} = \hat{b}')$, and $\sigma_{\hat{b}'} = \|\frac{1}{k}(\hat{b}')^t D_{\hat{y}} \Phi^t \Phi D_{\check{y}}\|$. Clearly,

$$\begin{aligned}\sigma_{\hat{b}'} &\leq \left\|\frac{1}{\sqrt{k}}(\hat{b}')^t D_{\hat{y}} \Phi^t\right\| \cdot \left\|\frac{1}{\sqrt{k}} \Phi D_{\check{y}}\right\| \\ &= O(\|\hat{y}\|\sigma_{\check{y}}) = O(\sigma_{\check{y}}) = O(\alpha) \ .\end{aligned}$$

Again using Appendix A and $E[Z|\hat{b} = \hat{b}'] = 0$ gives that $|\mu'_{\hat{b}}| = O(\alpha)$, and again we conclude using a union bound that with probability at least 0.99, uniformly for all $y \in Y$, $\left|\frac{1}{k} b^t D_{\hat{y}} \Phi^t \Phi D_{\check{y}} b\right| = O(\delta)$.

Tying it all together, we conclude that with probability at least 0.98, uniformly for all $y \in Y$,

$$\begin{aligned}\frac{1}{k}\|\Phi D_y b\|^2 &= \frac{1}{k}\|\Phi D_{\check{y}} b\|^2 \\ &\quad + \frac{1}{k}\|\Phi D_{\hat{y}} b\|^2 + 2 b^t D_{y^H} \Phi^t \Phi D_{\check{y}} b \\ &= \|y\|^2 + O(\delta) \ ,\end{aligned}$$

as required.

4 A note on running time

In [11] and [20] the authors present random operators which try to minimize the application time for *sparse* vectors. This is an important line of research given the increasing popularity of random projections for online learning and regression tasks in which the input vectors are usually not dense. We claim that a careful implementation of the operation $x \to \Phi x$ can also capitalize slightly from sparseness of input vectors. Since each entry in Φ can be computed in $O(1)$ operations a naive implementation would require $O(rk)$ operations for r-sparse vectors. This matches the running time of applying a naive dense i.i.d. matrix. Note however, that such naive constructions still require $O(dk)$ storage while Φ requires only $O(d)$. Moreover, in [10] claim that computing $x \to \Phi x$ requires $O(d \log k)$ operations by iteratively adding as subtracting sections of the input vector. If the a similar analysis is performed using sparse vector operations, the running time reduces to an expected $O(d \log(rk/d))$.

5 Conclusions

The obvious problems left open are those of (1) improving the dependence of k in δ (from δ^{-4} to δ^{-2}) and (2) removing the dependence of k in polylog(n). Other directions of research include not only reducing the computational efficiency of random dimension reduction, but also the amount of randomness needed for the construction.

Acknowledgements

We thank Emmanuel Candes for helpful discussions.

References

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

[2] P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory Series A*, 44:355–362, 1987.

[3] S. DasGupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. *Technical Report, UC Berkeley*, 99-006, 1999.

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

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

[6] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 143–152, Washington, DC, USA, 2006.

[7] Franco Woolfe, Edo Liberty, Vladimir Rokhlin, and Mark Tygert. A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25(3):335 – 366, 2008.

[8] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the 38st Annual Symposium on the Theory of Compututing (STOC)*, pages 557–563, Seattle, WA, 2006.

[9] Nir Ailon and Bernard Chazelle. Faster dimension reduction. *Commun. ACM*, 53(2):97–104, 2010.

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

[11] Dasgupta A., Kumar R., and Sarlos T. A sparse johnson-lindenstrauss transform. In *Proceedings of the 42nd ACM Symposium on Theorey of Computing (STOC)*, 2010.

[12] Edo Liberty, Nir Ailon, and Amit Singer. Dense fast random projections and lean walsh transforms. In *APPROX-RANDOM*, pages 512–522, 2008.

[13] Mark Rudelson and Roman Veshynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:10251045, 2008.

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

[15] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency infor-

mation. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[16] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[17] Alfred M. Bruckstein, David L. Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Rev.*, 51(1):34–81, 2009.

[18] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, 1991.

[19] G. Pisier. *The volume of convex bodies and Banach space geometry*. Number 94 in Cambridge Tracts in Mathematics. Cambridge University Press, 1989.

[20] Daniel M. Kane and Jelani Nelson. A derandomized sparse johnson-lindenstrauss transform.

A

FACT A.1. *For any real valued random variable Z such that for all $t > 0$*

$$\begin{aligned} \Pr[Z > \mu + t] &\leq \exp\{-ct^2/\sigma^2\} \\ \Pr[Z < \mu - t] &\leq \exp\{-ct^2/\sigma^2\} \end{aligned} \tag{A.1}$$

we have that $\sqrt{E(Z^2)} - O(\sigma) \leq \mu \leq \sqrt{E(Z^2)} + O(\sigma)$, where the big-O notation hides a dependence on the value of c.

Proof. Define the variable $Z' = (Z - \mu)/\sigma$.

$$\begin{aligned} E[Z'] \leq E[|Z'|] &\leq \sum_{i=1}^{\infty} i \Pr(i-1 \leq |Z'| \leq i) \\ &\leq \sum_{i=1}^{\infty} i \Pr(|Z'| \geq i-1) \\ &\leq 2\sum_{i=1}^{\infty} i \exp\{-c(i-1)^2\} = O(1) \ . \end{aligned}$$

Clearly, the last argument implies $E(Z) = \mu + O(\sigma)$. Similarly, we get $E[Z'^2] = O(1)$. Thus, $E[Z^2] - 2\mu E[Z] + \mu^2 = O(\sigma^2)$ and $E[Z^2] = (\mu \pm O(\sigma))^2$.