Dense Fast Random Projections and Lean Walsh Transforms

Edo Liberty*, Nir Ailon**, and Amit Singer***

Abstract. Random projection methods give distributions over $k \times d$ matrices such that if a matrix Ψ (chosen according to the distribution) is applied to a vector $x \in \mathbb{R}^d$ the norm of the resulting vector, $\Psi x \in \mathbb{R}^k$, is up to distortion ϵ equal to the norm of x w.p. at least $1-\delta$. The Johnson Lindenstrauss lemma shows that such distributions exist over *dense* matrices for k (the target dimension) in $O(\log(1/\delta)/\varepsilon^2)$. Ailon and Chazelle and later Matousek showed that there exist entry-wise i.i.d. distributions over *sparse* matrices Ψ which give the same guaranties for vectors whose ℓ_∞ is bounded away from their ℓ_2 norm. This allows to accelerate the mapping $x \mapsto \Psi x$. We claim that setting Ψ as any column normalized *deterministic dense* matrix composed with random ± 1 diagonal matrix also exhibits this property for vectors whose ℓ_p (for any $p > 2$) is bounded away from their ℓ_2 norm. We also describe a specific tensor product matrix which we term *lean Walsh*. It is applicable to any vector in $\mathbb{R}^d$ in $O(d)$ operations and requires a weaker ℓ_∞ bound on x then the best current result, under comparable running times, using sparse matrices due to Matousek.

Key words: Random Projections, Lean Walsh Transforms, Johnson Lindenstrauss, Dimension reduction

1 Introduction

The application of various random matrices has become a common method for accelerating algorithms both in theory and in practice. These procedures are commonly referred to as *random projections*. The critical property of a $k \times d$ random projection matrix, Ψ, is that for any vector x the mapping $x \mapsto \Psi x$ is such that $(1-\varepsilon)\|x\|_2 \leq \|\Psi x\|_2 \leq (1+\varepsilon)\|x\|_2$ with probability at least $1-\delta$ for specified constants $0 < \varepsilon < 1/2$ and $0 < \delta < 1$. The name *random projections* was coined after the first construction by Johnson and Lindenstrauss in [1] who showed that such mappings exist for $k \in O(\log(1/\delta)/\varepsilon^2)$. Since Johnson and Lindenstrauss other distributions for random projection matrices have been discovered [2–6]. Their properties make random projections a key player in rank-k

* Yale University, Department of Computer Science, Supported by NGA and AFOSR.
** Google Research
*** Yale University, Department of Mathematics, Program in Applied Mathematics.
[0] Edo Liberty and Amit Singer thank the Institute for Pure and Applied Mathematics (IPAM) and its director Mark Green for their warm hospitality during the fall semester of 2007.

approximation algorithms [7–13], other algorithms in numerical linear algebra [14–16], compressed sensing [17–19], and various other applications, e.g, [20, 21].

As a remark, random projections are usually used as an approximate isometric mapping from $\mathbb{R}^d$ to $\mathbb{R}^k$ for n vectors $x_1, \ldots, x_n$. By preserving the length of all $\binom{n}{2}$ distance vectors $x = x_i - x_j$ the entire metric is preserved. Taking $\delta = \frac{1}{2}\binom{n}{2}^{-1}$ yields this w.p. at least $1/2$ due to the union bound. The resulting target dimension is $k = O(\log(n)/\varepsilon^2)$.

Considering the usefulness of random projections it is natural to ask the following question: what should be the structure of a random projection matrix, Ψ, such that mapping $x \mapsto \Psi x$ would require the least amount of computational resources? A naïve construction of a $k \times d$ unstructured matrix Ψ would result in an $O(kd)$ application cost.

In [22], Ailon and Chazelle propose the first asymptotically Fast Johnson Lindenstrauss Transform (FJLT). They give a two stage projection process. First, all input vectors are rotated, using a Fourier transform, such that their ℓ_∞ norm is bounded by $O(\sqrt{k/d})$. Then, a sparse random matrix containing only $O(k^3)$ nonzeros[1] is used to project them into $\mathbb{R}^k$. Thus, reducing the running time of dimensionality reduction from $O(kd)$ to $O(d\log(d)+k^3)$. Matousek in [6] generalized the sparse projection process and showed that if the ℓ_∞ norm of all the input vectors is bounded from above by η, they can be projected by a sparse matrix, Ψ, whose entries are nonzero with probability $\max(ck\eta^2, 1)$ for some constant c. The number of nonzeros in Ψ is therefore $O(k^2 d\eta^2)$, with high probability. The concentration analysis is done for i.i.d. entries drawn from distributions satisfying mild assumptions.

Recently, Ailon and Liberty [23] improved the running time to $O(d\log(k))$ for $k \leq d^{1/2-\zeta}$ for any arbitrarily small ζ. They replaced the sparse i.i.d. projection matrix, Ψ, with a deterministic dense code matrix, A, composed with a random ± 1 diagonal matrix[2], D_s. They showed that a careful choice of A results in AD_s being a good random projection for the set of vectors such that $\|x\|_4 \in O(d^{-1/4})$. Here, we analyze this result for general $k \times d$ deterministic matrices. Our concentration result is very much in the spirit of [23]. We claim that any column normalized matrix A can be identified with a set $\chi \subset \mathbb{R}^d$ such that for x chosen from χ, AD_s constitutes a random projection w.h.p. The set χ can be thought of as the "good" set for AD_s. We study a natural tradeoff between the possible computational efficiency of applying A and the size of χ: the smaller χ is, the faster A can be applied[3]. We examine the connection between A and χ in Section 2. The set χ should be thought of as a prior assumption on our data, which may come, for example, from a statistical model generating the data.

We propose in Section 3 a new type of fast applicable matrices and in Section 4 explore their corresponding χ. These matrices are constructed using tensor

[1] Each entry is drawn from a distribution which is gaussian with probability proportional to k^2/d, and so, for any constant probability, arbitrarily close to 1, the number of nonzeros is smaller than ck^3 for some constant c.

[2] The random isometric preprocessing is also different than that of the FJLT algorithm

[3] This, however, might require a time costly preprocessing application of Φ.

	The rectangular $k \times d$ matrix A	Application time	$x \in \chi$ if
Johnson, Lindenstrauss [1]	Random k dimensional subspace	$O(kd)$	$x \in \mathbb{R}^d$
Various Authors [2, 4–6]	Dense i.i.d. entries Gaussian or ± 1	$O(kd)$	$x \in \mathbb{R}^d$
Ailon, Chazelle [22]	Sparse Gaussian distributed entries	$O(k^3)$	$\frac{\|x\|_\infty}{\|x\|_2} = O((d/k)^{-1/2})$
Matousek [6]	Sparse sub-Gaussian symmetric i.i.d. entries	$O(k^2 d\eta^2)$	$\frac{\|x\|_\infty}{\|x\|_2} \leq \eta$
General rule (This work)	**Any deterministic matrix A**		$\frac{\|x\|_A}{\|x\|_2} = O(k^{-1/2})$
Ailon, Liberty [23]	Four-wise independent	$O(d \log k)$	$\frac{\|x\|_4}{\|x\|_2} = O(d^{-1/4})$
This work	Lean Walsh Transform	$O(d)$	$\frac{\|x\|_\infty}{\|x\|_2} = O(k^{-1/2} d^{-\zeta})$

Table 1. Types of $k \times d$ matrices and the subsets χ of $\mathbb{R}^d$ for which they constitute a random projection. The meaning of the norm $\|\cdot\|_A$ is given in Definition 2. The top two rows give random dense matrices, below are random i.i.d. sparse matrices, and the last three are *deterministic* matrices composed with random ± 1 diagonals.

products and can be applied to any vector in $\mathbb{R}^d$ in linear time, i.e., in $O(d)$. Due to the similarity in their construction to Walsh-Hadamard matrices and their rectangular shape we term them *lean Walsh Matrices*[4]. Lean Walsh matrices are of size $\widetilde{d} \times d$ where $\widetilde{d} = d^\alpha$ for some $0 < \alpha < 1$. In order to reduce the dimension to $k \leq \widetilde{d}$, $k = O(\log(1/\delta)/\varepsilon^2))$, we can compose the lean Walsh matrix, A, with a known Johnson Lindenstrauss matrix construction R. Applying R in $O(d)$ requires some relation between d, k and α as explained in subsection 4.1.

2 Norm concentration and $\chi(A, \varepsilon, \delta)$

We compose an arbitrary deterministic $\widetilde{d} \times d$ matrix A with a random sign diagonal matrix D_s and study the behavior of such matrices as random projections. In order for AD_s to exhibit the property of a random projection it is enough for it to approximately preserve the length of any single *unit* vector $x \in \mathbb{R}^d$ with high probability:

$$\Pr\left[\,|\, \|AD_s x\|_2 - 1 \,|\, \geq \varepsilon)\right] < \delta \tag{1}$$

[4] The terms *lean Walsh Transform* or simply *lean Walsh* are also used interchangeably.

Here D_s is a diagonal matrix such that $D_s(i,i)$ are random signs (i.i.d. ± 1 w.p. $1/2$ each), $0 < \delta < 1$ is a constant acceptable failure probability, and the constant $0 < \varepsilon < 1/2$ is the prescribed precision.

Note that we can replace the term $AD_s x$ with $AD_x s$ where D_x is a diagonal matrix holding on the diagonal the values of x, i.e. $D_x(i,i) = x(i)$ and similarly $s(i) = D_s(i,i)$. Denoting $M = AD_x$, we view the term $\|Ms\|_2$ as a scalar function over the hypercube $\{1,-1\}^d$, from which the variable s is uniformly chosen. This function is convex over $[-1,1]^d$ and Lipschitz bounded. Talagrand [24] proves a strong concentration result for such functions. We give a slightly restated form of his result for our case.

Lemma 1 (Talagrand [24]). *Given a matrix M and a random vector s ($s(i)$ are i.i.d. ± 1 w.p. $1/2$) define the random variable $Y = \|Ms\|_2$. Denote by μ a median of Y, and by $\sigma = \|M\|_{2\to 2}$ the spectral norm of M. Then*

$$\Pr[\,|Y-\mu| > t] \le 4e^{-t^2/8\sigma^2} \tag{2}$$

Definition 1. *$\|M\|_{p\to q}$ denoted the norm of M as an operator from ℓ_p to ℓ_q, i.e., $\|M\|_{p\to q} = \sup_{x,\,\|x\|_p=1} \|Mx\|_q$. The ordinary spectral norm of M is thus $\|M\|_{2\to 2}$.*

Lemma 1 asserts that $\|AD_x s\|$ is distributed like a (sub) Gaussian around its median, with standard deviation 2σ.

First, in order to have $E[Y^2] = 1$ it is necessary and sufficient for the columns of A to be normalized to 1 (or normalized in expectancy). To estimate a median, μ, we substitute $t^2 \to t'$ and compute:

$$\begin{aligned} E[(Y-\mu)^2] &= \int_0^\infty \Pr[(Y-\mu)^2] > t']dt' \\ &\le \int_0^\infty 4e^{-t'/(8\sigma^2)}dt' = 32\sigma^2 \end{aligned}$$

Furthermore, $(E[Y])^2 \le E[Y^2] = 1$, and so $E[(Y-\mu)^2] = E[Y^2] - 2\mu E[Y] + \mu^2 \ge 1 - 2\mu + \mu^2 = (1-\mu)^2$. Combining, $|1-\mu| \le \sqrt{32}\sigma$. We set $\varepsilon = t + |1-\mu|$:

$$\Pr[|Y-1| > \varepsilon] \le 4e^{-\varepsilon^2/32\sigma^2} \quad \text{, for } \varepsilon > 2|1-\mu| \tag{3}$$

If we set $k = 33\log(1/\delta)/\varepsilon^2$ (for $\log(1/\delta)$ larger than a sufficient constant) and set $\sigma \le k^{-1/2}$, (1) follows from (3). Moreover μ depends on ε such that the condition $\varepsilon > 2|1-\mu|$ is met for any constant ε (given $\log(1/\delta) > 4$). This can be seen by $|1-\mu| \le \sqrt{32}\sigma < \varepsilon/\sqrt{\log(1/\delta)}$. We see that $\sigma = \|AD_x\|_{2\to 2} \le k^{-1/2}$ is sufficient for the projection to succeed w.h.p. This naturally defines χ.

Definition 2. *For a given matrix $A \in \mathbb{R}^{k\times d}$ we define the vector pseudonorm of $x \in \mathbb{R}^d$ with respect to A as $\|x\|_A \equiv \|AD_x\|_{2\to 2}$ where D_x is a diagonal matrix such that $D_x(i,i) = x(i)$.* Remark: *If no column of A has norm zero $\|\cdot\|_A$ induces a proper norm on $\mathbb{R}^d$.*

Definition 3. *We define $\chi(A,\varepsilon,\delta)$ as the intersection of the Euclidian unit sphere and a ball of radius $k^{-1/2}$ in the norm $\|\cdot\|_A$*

$$\chi(A,\varepsilon,\delta) = \left\{ x \in \mathbb{S}^{d-1} \mid \|x\|_A \leq k^{-1/2} \right\} \tag{4}$$

for $k = 33\log(1/\delta)/\varepsilon^2$.

Lemma 2. *For any column normalized matrix, A, and an i.i.d. random ± 1 diagonal matrix, D_s, the following holds:*

$$\forall x \in \chi(A,\varepsilon,\delta) \quad \Pr\left[\, |\|AD_s x\|_2 - 1| \geq \varepsilon\right] \leq \delta \tag{5}$$

Proof. For any $x \in \chi$, by Definition 3, $\|x\|_A = \|AD_x\|_{2\to 2} = \sigma \leq k^{-1/2}$. The lemma follows from substituting the value of σ into Equation (3).

It is convenient to think about χ as the "good" set of vectors for which AD_s is length preserving with high probability. En route to explore $\chi(A,\varepsilon,\delta)$ for lean Walsh matrices we first turn to formally defining them.

3 Lean Walsh transforms

The *lean* Walsh Transform, similar to the Walsh Transform, is a recursive tensor product matrix. It is initialized by a constant seed matrix, A_1, and constructed recursively by using Kronecker products $A_{\ell'} = A_1 \otimes A_{\ell'-1}$. The main difference is that the lean Walsh seeds have fewer rows than columns. We formally define them as follows:

Definition 4. *A_1 is a lean Walsh seed (or simply 'seed') if: i) A_1 is a rectangular matrix $A_1 \in \mathbb{C}^{r\times c}$, such that $r < c$; ii) A_1 is absolute valued $1/\sqrt{r}$ entry-wise, i.e., $|A_1(i,j)| = r^{-1/2}$; iii) the rows of A_1 are orthogonal.*

Definition 5. *A_ℓ is a lean Walsh transform, of order ℓ, if for all $\ell' \leq \ell$ we have $A_\ell = A_1 \otimes A_{\ell'-1}$, where $\otimes$ stands for the Kronecker product and A_1 is a seed according to Definition 4.*

The following are examples of seed matrices:

$$A_1' = \tfrac{1}{\sqrt{3}} \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \quad A_1'' = \tfrac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 1 \\ 1 & e^{2\pi i/3} & e^{4\pi i/3} \end{pmatrix} \tag{6}$$

These examples are a part of a large family of possible seeds. This family includes, amongst other constructions, sub-Hadamard matrices (like A_1') or sub-Fourier matrices (like A_1''). A simple construction is given for possible larger seeds.

Fact 1 *Let F be the $c\times c$ Discrete Fourier matrix such that $F(i,j) = e^{2\pi\sqrt{-1}ij/c}$. Define A_1 to be the matrix consisting of the first $r = c-1$ rows of F normalized by $1/\sqrt{r}$. A_1 is a lean Walsh seed.*

We use elementary properties of Kronecker products to characterize A_ℓ in terms of the number of rows, r, and the number of columns, c, of its seed. The following facts hold true for A_ℓ:

Fact 2 *i) The size of A_ℓ is $d^\alpha \times d$, where $\alpha = \log(r)/\log(c) < 1$ is the skewness of A_1;*[5] *ii) for all i and j, $A_\ell(i,j) \in \pm\widetilde{d}^{-1/2}$ which means that A_ℓ is column normalized; and iii) the rows of A_ℓ are orthogonal.*

Fact 3 *The time complexity of applying A_ℓ to any vector $z \in \mathbb{R}^d$ is $O(d)$.*

Proof. Let $z = [z_1; \ldots; z_c]$ where z_i are blocks of length d/c of the vector z. Using the recursive decomposition for A_ℓ we compute $A_\ell z$ by first summing over the different z_i according to the values of A_1 and applying to each sum the matrix $A_{\ell-1}$. Denoting by $T(d)$ the time to apply A_ℓ to $z \in \mathbb{R}^d$ we get that $T(d) = rT(d/c) + rd$. A simple calculation yields $T(d) \leq dcr/(c-r)$ and thus $T(d) = O(d)$ for a constant sized seed.

For clarity, we demonstrate Fact 3 for A'_1 (Equation (6)):

$$A'_\ell z = A'_\ell \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{pmatrix} = \frac{1}{\sqrt{3}} \begin{pmatrix} A'_{\ell-1}(z_1 + z_2 - z_3 - z_4) \\ A'_{\ell-1}(z_1 - z_2 + z_3 - z_4) \\ A'_{\ell-1}(z_1 - z_2 - z_3 + z_4) \end{pmatrix} \tag{7}$$

Remark 1. For the purpose of compressed sensing, an important parameter of the projection matrix is its Coherence. The Coherence of a column normalized matrix is simply the maximal inner product between two different columns. The Coherence of a lean Walsh matrix is equal to the coherence of its seed and the seed coherence can be reduced by increasing its size. For example, the seeds described in Fact 1, of size r by $c = r + 1$, exhibit coherence of $1/r$.

In what follows we characterize $\chi(A, \varepsilon, \delta)$ for a general lean Walsh transform by the parameters of its seed. The abbreviated notation, A, stands for A_ℓ of the right size to be applied to x, i.e., $\ell = \log(d)/\log(c)$. Moreover, we freely use α to denote the skewness $\log(r)/\log(c)$ of the seed at hand.

4 An ℓ_p bound on $\|\cdot\|_A$

After describing the lean Walsh transforms we turn our attention to exploring their "good" sets χ .We remind the reader that $\|x\|_A \leq k^{-1/2}$ implies $x \in \chi$:

$$\|x\|_A^2 = \|AD_x\|_{2\to 2}^2 = \max_{y, \|y\|_2 = 1} \|y^T AD_x\|_2^2 \tag{8}$$

[5] The size of A_ℓ is $r^\ell \times c^\ell$. Since the running time is linear, we can always pad vectors to be of length c^ℓ without effecting the asymptotic running time. From this point on we assume w.l.o.g $d = c^\ell$ for some integer ℓ

$$= \max_{y,\|y\|_2=1} \sum_{i=1}^{d} x^2(i)(y^T A^{(i)})^2 \tag{9}$$

$$\leq \left(\sum_{i=1}^{d} x^{2p}(i) \right)^{1/p} \left(\max_{y,\|y\|_2=1} \sum_{i=1}^{d} (y^T A^{(i)})^{2q} \right)^{1/q} \tag{10}$$

$$= \|x\|_{2p}^2 \|A^T\|_{2\to 2q}^2 \tag{11}$$

The transition from the second to the third line follows from Hölder's inequality for dual norms p and q, satisfying $1/p + 1/q = 1$. We now compute $\|A^T\|_{2\to 2q}$.

Theorem 1. [Riesz-Thorin] *For an arbitrary matrix B, assume $\|B\|_{p_1\to r_1} \leq C_1$ and $\|B\|_{p_2\to r_2} \leq C_2$ for some norm indices p_1, r_1, p_2, r_2 such that $p_1 \leq r_1$ and $p_2 \leq r_2$. Let λ be a real number in the interval $[0,1]$, and let p, r be such that $1/p = \lambda(1/p_1) + (1-\lambda)(1/p_2)$ and $1/r = \lambda(1/r_1) + (1-\lambda)(1/r_2)$. Then $\|B\|_{p\to r} \leq C_1^{\lambda} C_2^{1-\lambda}$.*

In order to use the theorem, let us compute $\|A^T\|_{2\to 2}$ and $\|A^T\|_{2\to\infty}$. From $\|A^T\|_{2\to 2} = \|A\|_{2\to 2}$ and the orthogonality of the rows of A we get that $\|A^T\|_{2\to 2} = \sqrt{d/\widetilde{d}} = d^{(1-\alpha)/2}$. From the normalization of the columns of A we get that $\|A^T\|_{2\to\infty} = 1$. Using the theorem for $\lambda = 1/q$, for any $q \geq 1$, we obtain $\|A^T\|_{2\to 2q} \leq d^{(1-\alpha)/2q}$. It is worth noting that $\|A^T\|_{2\to 2q}$ might actually be significantly lower than the given bound. For a specific seed, A_1, one should calculate $\|A_1^T\|_{2\to 2q}$ and use $\|A_\ell^T\|_{2\to 2q} = \|A_1^T\|_{2\to 2q}^{\ell}$ to achieve a possibly lower value for $\|A^T\|_{2\to 2q}$.

Lemma 3. *For a lean Walsh transform, A, we have that for any $p > 1$ the following holds:*

$$\{x \in \mathbb{S}^{d-1} \mid \|x\|_{2p} \leq k^{-1/2} d^{-\frac{1-\alpha}{2}(1-\frac{1}{p})}\} \subset \chi(A, \varepsilon, \delta) \tag{12}$$

where $k = O(\log(1/\delta)/\varepsilon^2)$ and α is the skewness of A, $\alpha = \log(r)/\log(c)$ (r is the number of rows, and c is the number of columns in the seed of A).

Proof. We combine the above and use the duality of p and q:

$$\|x\|_A \leq \|x\|_{2p} \|A^T\|_{2\to 2q} \tag{13}$$

$$\leq \|x\|_{2p} d^{\frac{1-\alpha}{2q}} \tag{14}$$

$$\leq \|x\|_{2p} d^{\frac{1-\alpha}{2}(1-\frac{1}{p})} \tag{15}$$

The desired property, $\|x\|_A \leq k^{-1/2}$, is achieved if $\|x\|_{2p} \leq k^{-1/2} d^{-\frac{1-\alpha}{2}(1-\frac{1}{p})}$ for any $p > 1$.

Remark 2. Consider a different family of matrices containing $d/\widetilde{d}$ copies of a $\widetilde{d} \times \widetilde{d}$ identity matrices concatenated horizontally. Their spectral norm is the

same as that of lean Walsh matrices and they are clearly row orthogonal and column normalized. Considering $p \to \infty$ they require the same ℓ_∞ constraint on x as lean Walsh matrices do. However, their norm as operators from ℓ_2 to ℓ_{2q} ,for q larger than 1 ($p < \infty$), is large and fixed, whereas that of lean Walsh matrices is still arbitrarily small and controlled by the size of the their seed.

4.1 Controlling α and choosing R

We see that increasing the skewness of the seed of A, α, is beneficial from the theoretical stand point since it weakens the constraint on $\|x\|_{2p}$. However, the application oriented reader should keep in mind that this requires the use of a larger seed, which subsequently increases the constant hiding in the big O notation of the running time.

Consider the seed constructions described in Fact 1 for which $r = c-1$. Their skewness $\alpha = \log(r)/\log(c)$ approaches 1 as their size increases. Namely, for any positive constant ζ there exists a constant size seed such that $1 - 2\zeta \leq \alpha \leq 1$.

Lemma 4. *For any positive constant $\zeta > 0$ there exists a lean Walsh matrix, A, such that:*

$$\{x \in \mathbb{S}^{d-1} \mid \|x\|_\infty \leq k^{-1/2} d^{-\zeta}\} \subset \chi(A, \varepsilon, \delta) \tag{16}$$

Proof. Generate A from a seed such that its skewness $\alpha = \log(r)/\log(c) \geq 1-2\zeta$ and substitute $p = \infty$ into the statement of Lemma 3.

The skewness α also determines the minimal dimension d (relative to k) for which the projection can be completed in $O(d)$ operations. The reason being that the vectors $z = AD_s x$ must be mapped from dimension $\widetilde{d}$ ($\widetilde{d} = d^\alpha$) to dimension k in $O(d)$ operations. This can be done using Ailon and Liberty's construction [23] serving as the random projection matrix R. R is a $k \times \widetilde{d}$ Johnson Lindenstrauss projection matrix which can be applied in $\widetilde{d}\log(k)$ operations if $\widetilde{d} = d^\alpha \geq k^{2+\zeta''}$ for arbitrary small ζ''. For the same choice of a seed as in Lemma 4, the condition becomes $d \geq k^{2+\zeta''+2\zeta}$ which can be achieved by $d \geq k^{2+\zeta'}$ for arbitrary small ζ' depending on ζ and ζ''. Therefore for such values of d the matrix R exists and requires $O(d^\alpha \log(k)) = O(d)$ operations to apply.

5 Comparison to sparse projections

Sparse random ± 1 projection matrices were analyzed by Matousek in [6]. For completeness we restate his result. Theorem 4.1 in [6] (slightly rephrased to fit our notation) claims the following:

Theorem 2 (Matousek 2006 [6]). *let $\varepsilon \in (0, 1/2)$ and $\eta \in [1/\sqrt{d}, 1]$ be constant parameters. Set $q = C_0\eta^2 \log(1/\delta)$ for a sufficiently large constant C_0. Let S be a random variable such that*

$$S = \begin{cases} +\frac{1}{\sqrt{qk}} & \text{with probability } q/2 \\ -\frac{1}{\sqrt{qk}} & \text{with probability } q/2 \\ 0 & \text{with probability } 1-q\ . \end{cases} \tag{17}$$

Let k be $C_1 \log(1/\delta)/\varepsilon^2$ for a sufficiently large C_1. Draw the matrix elements of Ψ i.i.d. from S. Then:

$$\Pr[|\|\Psi x\|_2^2 - 1| > \varepsilon] \leq \delta \tag{18}$$

For any $x \in \mathbb{S}^{d-1}$ such that $\|x\|_\infty \leq \eta$.

With constant probability, the number of nonzeros in Ψ is $O(kdq) = O(k^2 d\eta^2)$ (since ε is a constant $\log(1/\delta) = O(k)$). In the terminology of this paper we say that for a sparse Ψ containing $O(k^2 d\eta^2)$ nonzeros on average (as above) $\{x \in \mathbb{S}^{d-1} \mid \|x\|_\infty \leq \eta\} \subset \chi(A, \varepsilon, \delta)$.

A lower bound on the running time of general dimensionality reduction is at least $\Omega(d)$. Our analysis shows that the problem of satisfying the condition $\Phi x \in \chi$ (via a Euclidean isometry Φ) is at least as hard. Indeed, a design of any such fast transformation, applicable in time $T(d)$, would imply a similar upper bound for general dimensionality reduction. We claim that lean Walsh matrices admit a strictly larger χ than that of sparse matrices which could be applied in the same asymptotic complexity. For $q = k^{-1}$ a sparse matrix Ψ as above contains $O(d)$ nonzeros, w.h.p., and thus can be applied in that amount of time. Due to Theorem 2 this value of q requires $\|x\|_\infty \leq O(k^{-1})$ for the length of x to be preserved w.h.p. For d polynomial in k, this is a stronger constraint on the ℓ_∞ norm of x than $\|x\|_\infty \leq O(k^{-1/2} d^{-\zeta})$ which is obtained by our analysis for lean Walsh transforms.

6 Conclusion and work in progress

We have shown that any $k \times d$ (column normalized) matrix, A, can be composed with a random diagonal matrix to constitute a random projection matrix for some part of the Euclidean space, χ. Moreover, we have given sufficient conditions, on $x \in \mathbb{R}^d$, for belonging to χ depending on different $\ell_2 \to \ell_p$ operator norms of A^T and ℓ_p norms of x. We have also seen that lean Walsh matrices exhibit both a "large" χ and a linear time computation scheme which outperforms sparse projective matrices. These properties make them good building blocks for the purpose of random projections.

However, as explained in the introduction, in order for the projection to be complete, one must design a linear time preprocessing matrix Φ which maps all vectors in $\mathbb{R}^d$ into χ (w.h.p.). Achieving such distributions for Φ would be extremely interesting from both the theoretical and practical stand point. Possible choices for Φ may include random permutations, various wavelet/wavelet-like transforms, or any other sparse orthogonal transformation.

In this framework χ was characterized by a bound over ℓ_p ($p > 2$) norms of $x \in \chi$. Understanding distributions over ℓ_2 isometries which reduce other ℓ_p norms with high probability and efficiency is an interesting problem in its own right. However, partial results hint that for lean Walsh transforms if Φ is taken to be a random permutation (which is an ℓ_p isometry for any p) then the ℓ_∞ requirement reduces to $\|x\|_\infty \leq k^{-1/2}$. Showing this however requires a different technique.

Acknowledgments

The authors would like to thank Steven Zucker, Daniel Spielman, and Yair Bartal for their insightful ideas and suggestions.

References

1. Johnson, W.B., Lindenstrauss, J.: Extensions of Lipschitz mappings into a Hilbert space. Contemporary Mathematics **26** (1984) 189–206
2. Frankl, P., Maehara, H.: The Johnson-Lindenstrauss lemma and the sphericity of some graphs. Journal of Combinatorial Theory Series A **44** (1987) 355–362
3. Indyk, P., Motwani, R.: Approximate nearest neighbors: Towards removing the curse of dimensionality. In: Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC). (1998) 604–613
4. DasGupta, S., Gupta, A.: An elementary proof of the Johnson-Lindenstrauss lemma. Technical Report, UC Berkeley **99-006** (1999)
5. Achlioptas, D.: Database-friendly random projections: Johnson-lindenstrauss with binary coins. J. Comput. Syst. Sci. **66**(4) (2003) 671–687
6. Matousek, J.: On variants of the Johnson-Lindenstrauss lemma. Private communication (2006)
7. Drineas, P., Mahoney, M.W., Muthukrishnan, S.: Sampling algorithms for ℓ_2 regression and applications. In: Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), Miami, Florida, United States (2006)
8. Sarlós, T.: Improved approximation algorithms for large matrices via random projections. In: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS), Berkeley, CA (2006)
9. Frieze, A.M., Kannan, R., Vempala, S.: Fast monte-carlo algorithms for finding low-rank approximations. In: IEEE Symposium on Foundations of Computer Science. (1998) 370–378
10. Peled, S.H.: A replacement for voronoi diagrams of near linear size. In: Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science (FOCS), Las Vegas, Nevada, USA (2001) 94–103
11. Achlioptas, McSherry: Fast computation of low rank matrix approximations. In: STOC: ACM Symposium on Theory of Computing (STOC). (2001)
12. Drineas, P., Kannan, R.: Fast monte-carlo algorithms for approximate matrix multiplication. In: IEEE Symposium on Foundations of Computer Science. (2001) 452–459
13. Liberty, E., Woolfe, F., Martinsson, P.G., Rokhlin, V., Tygert, M.: Randomized algorithms for the low-rank approximation of matrices. Proceedings of the National Academy of Sciences (2007)
14. Dasgupta, A., Drineas, P., Harb, B., Kumar, R., Mahoney, M.W.: Sampling algorithms and coresets for ℓ_p regression. Proc. of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA) (2008)
15. Drineas, P., Mahoney, M.W., Muthukrishnan, S., Sarlos, T.: Faster least squares approximation. TR arXiv:0710.1435 (submitted for publication, 2007)
16. Drineas, P., Mahoney, M., Muthukrishnan, S.: Relative-error cur matrix decompositions. TR arXiv:0708.3696 (submitted for publication, 2007)
17. Candes, E. J.; Tao, T.: Near-optimal signal recovery from random projections: Universal encoding strategies? Information Theory, IEEE Transactions on **52**(12) (Dec. 2006) 5406–5425

18. Donoho, D.L.: Compressed sensing. IEEE Transactions on Information Theory **52**(4) (2006) 1289–1306
19. Elad, M.: Optimized projections for compressed sensing. IEEE Transactions on Signal Processing **55**(12) (2007) 5695–5702
20. Paschou, P., Ziv, E., Burchard, E., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M.W., Drineas, P.: Pca-correlated snps for structure identification in worldwide human populations. PLOS Genetics, 3, pp. 1672-1686 (2007)
21. Paschou, P., Mahoney, M.W., Kidd, J., Pakstis, A., S. Gu, K.K., Drineas, P.: Intra- and inter-population genotype reconstruction from tagging snps. Genome Research, 17(1), pp. 96-107 (2007)
22. Ailon, N., Chazelle, B.: Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In: STOC '06: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing, New York, NY, USA, ACM Press (2006) 557–563
23. Ailon, N., Liberty, E.: Fast dimension reduction using rademacher series on dual bch codes. In: SODA. (2008) 1–9
24. Ledoux, M., Talagrand, M.: Probability in Banach Spaces: Isoperimetry and Processes. Springer-Verlag (1991)