Dense Fast Random Projections and Lean Walsh Transforms

Edo Liberty* Nir Ailon† 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 finite set of vectors $x_i \in \mathbb{R}^d$ the resulting vectors $\Psi x_i \in \mathbb{R}^k$ approximately preserve the original metric with constant probability. First, we show that any matrix (composed with a random ± 1 diagonal matrix) is a good random projector for a subset of vectors in $\mathbb{R}^d$. Second, we describe a family of tensor product matrices which we term *Lean Walsh*. We show that using Lean Walsh matrices as random projections outperforms, in terms or running time, the best known current result (due to Matousek) under comparable assumptions.

Introduction

Random Projection is a technique for reducing the dimensionality of vectors by multiplying them with a random matrix. The critical property of a $k \times d$ random projection matrix, Ψ, is that for any vector x the mapping $x \mapsto \Psi x$ preserves its length, up to distortion ε, with probability at least $1 - \delta$ for any $0 < \varepsilon < 1/2$ and $\delta > 0$. Here, the target dimension k is much smaller than the original dimension d. The name *random projections* was coined after the first construction by Johnson and Lindenstrauss in [1] who showed that such mappings exist for $k \in O(\log(1/\delta)/\varepsilon^2)$. Other constructions of random projection matrices have been discovered since [2, 3, 4, 5, 6]. Their properties make random projections a key player in rank-k approximation algorithms [7, 8, 9, 10, 11, 12, 13, 14], other algorithms in numerical linear algebra [15, 16, 17], compressed sensing [18, 19, 20], and various other applications, e.g, [21, 22].

Considering the usefulness of random projections it is important to understand their computational efficiency under different circumstances. Using a dense $k \times d$ unstructured matrix Ψ with random i.i.d. entries such as in [1, 2, 3, 4, 5, 6] would guarantee the desired low distortion property for the entire space of

*Yahoo! Research, Yale University at the time supported by NGA and AFOSR.

†Technion Israel Institute of Technology.

‡Princeton University. Yale University at the time.

[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.

input vectors but requires an $O(kd)$ application time.[1] Ailon and Chazelle in [24] showed that if the ℓ_∞ norm of the input vectors is bounded by $O(\sqrt{k/d})$ then a sparse random matrix containing only $O(k^3)$ non-zeros can be used to project them onto $\mathbb{R}^k$.[2] Thus, reducing the running time of random projection from $O(kd)$ to $O(d\log(d)+k^3)$.[3] Matousek in [6] generalized the sparse projection process and showed that if the ℓ_∞ norm of the input vector is bounded from above by any value η, it can be projected by a sparse matrix, Ψ, whose entries are non-zero with probability $\min(ck\eta^2,1)$ for some constant c. The expected number of nonzeros in Ψ and the consequent cted application complexity are therefore $O(k^2d\eta^2)$. The concentration analysis, for both cases, requires that the entries $\Psi(i,j)$ be i.i.d. However, they can be drawn from any distribution satisfying very mild assumptions. Recently, Ailon and Liberty [25] improved the overall running time of random projection to $O(d\log(k))$ for $k \leq d^{1/2-\delta}$ for any arbitrarily small δ. They replaced the sparse i.i.d. projection matrix, Ψ, with a deterministic dense code matrix, A, composed with a random ± 1 diagonal matrix, D_s.[4] In their analysis they show that $\Psi = AD_s$ is a good random projection for vectors x with bounded ℓ_4 norm.

In this work we extend and simplify the result in [25]. Here we compose any column normalized fixed matrix A with a random diagonal matrix D_s and give a sufficient condition on the ℓ_p norm of input vectors x for which the composition AD_s is a good random projection.

To show this, in Section 1.1, we restate a version of Talagrand's inequality for Rademacher random variables to characterize the concentration of high dimensional random walks. In Section 1.2 we show the connection between random walks and random projections and give the relation between the fixed projection matrix A and the set for which it is a good random projection. These good sets are given as bounds on a special vector norm which depends on the matrix and is denoted by $\|x\|_A$. Since this special norm can be rather complex, we give in Section 1.3 a simplified bound on $\|x\|_A$ in terms of $\|x\|_{2p}$ and $\|A^T\|_{2\to 2q}$ for dual norms p and q. The notation $\|A^T\|_{2\to 2q}$ stands for the operator norm of A^T from ℓ_2 to ℓ_{2q}.

The second section of the paper is dedicated to introducing a new family of fast applicable matrices and exploring their random projection properties. Due to their construction similarity to Walsh-Hadamard matrices and their rectangular shape we term them *Lean Walsh Matrices*[5]. They are constructed using tensor products and can be applied to any vector in $\mathbb{R}^d$ in linear time, i.e., in $O(d)$ operations. The motivation behind this investigation is the fact that $O(d)$ lower bounds the running time for general random projections.[6]

[1] The result of Achlioptas [5] can be improved to $O(kd/\log(d))$ using methods for fast application of binary matrices [23].

[2] Each entry is either zero or drawn from a Gaussian distribution with probability proportional to k^2/d. The expected number of non-zeros in the matrix is thus $O(k^3)$.

[3] $O(d\log(d))$ is required for a preprocessing stage in which the ℓ_∞ norm is reduced.

[4] The random isometric preprocessing is also different from that of Ailon and Chazelle's FJLT algorithm

[5] The terms *Lean Walsh Transform* or simply *Lean Walsh* are also used interchangeably.

[6] If no restriction is put on input vectors each entry must be read.

	The rectangular $k \times d$ matrix A	Application time	$x \in \chi$ if $\|x\|_2 = 1$ and:
Johnson, Lindenstrauss [1]	k rows of a random unitary matrix	$O(kd)$	
Various Authors [2, 4, 5, 6]	i.i.d. random entries	$O(kd)$	
Ailon, Chazelle [24]	Sparse Gaussian entries	$O(k^3)$	$\|x\|_\infty = O((d/k)^{-1/2})$
Matousek [6]	Sparse ± 1 entries	$O(k^2 d\eta^2)$	$\|x\|_\infty \leq \eta$
Ailon, Liberty [25]	4-wise independent Code matrix	$O(d \log k)$	$\|x\|_4 = O(d^{-1/4})$
This work	Any deterministic matrix	?	$\|x\|_A = O(k^{-1/2})$
This work	Lean Walsh Transform	$O(d)$	$\|x\|_\infty = O(k^{-1/2} d^{-\delta})$

Table 1: Types of matrix distributions and the subsets χ of $\mathbb{R}^d$ for which they constitute a random projection. Here $k = O(\log(1/\delta)/\varepsilon^2)$ where ε is the prescribed precision and $1-\delta$ is the success probability. The meaning of the norm $\|\cdot\|_A$ is given in Definition 1.1.

Thus, it is natural to ask: what is the largest possible subset of $\mathbb{R}^d$ that can be projected with low distortion and with high probability in $O(d)$ time? Using the mathematical framework presented in Section 1 we show in Section 2.3 that a construction using Lean Walsh matrices succeeds for a larger portion of $\mathbb{R}^d$ than that of sparse i.i.d. matrix distributions.

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

1.1 High dimensional random walks

We will see in the next chapter that random projections from dimension d to dimension k can be viewed as d-step random walks in dimension k, i.e., each step is a vector $M^{(i)}$ and the projection norm is the final displacement $\|\sum M^{(i)} s(i)\|$. The variable $\sum M^{(i)} s(i)$ is called a Rademacher random variable. A thorough investigation of concentration bounds for Rademacher random variables in Banach spaces can be found in [26]. For convenience and completeness we derive in this section a more limited result for finite dimensional vectors using a powerful theorem of Talagrand (Chapter 1, [26]) on measure concentration of functions on $\{-1, +1\}^d$ extendable to convex functions on $[-1, +1]^d$ with bounded Lipschitz norm.

Lemma 1.1. *For any matrix M and a random vector s, $s(i) \in \{+1, -1\}$ w.p. $1/2$ each. We have that:*

$$\Pr[\,|\,\|Ms\|_2 - \|M\|_F\,| > t] \leq 16 e^{-\frac{t^2}{32\|M\|_2^2}} \tag{1}$$

where $\|\cdot\|_2$ denotes the 2 norm for vectors and the spectral norm for matrices and $\|\cdot\|_F$ denotes the Frobenius norm $\|M\|_F = [\sum_{i,j} M(i,j)^2]^{1/2}$.

To prove this we use a result by Talagrand.

Lemma 1.2 (Talagrand [26]). *Given a Lipschitz bounded convex function f over the solid hypercube and a point s chosen uniformly from the hypercube let Y denote the random variable $f(s)$ and let μ be its median. Also denote by $\sigma = \|f\|_{Lip}$ the Lipschitz constant of f. Then:*

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

We set $f(s) = \|Ms\|_2$. This function is convex over $[-1,1]^d$ and Lipschitz bounded $\sigma = \|f\|_{Lip} = \|M\|_2$, the spectral norm of M. To estimate the median, μ, we substitute $t^2 \to t'$ and compute:

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

Furthermore, $(E[Y])^2 \leq E[Y^2] = \|M\|_F^2$, and so $E[(Y-\mu)^2] = E[Y^2] - 2\mu E[Y] + \mu^2 \geq \|M\|_F^2 - 2\mu\|M\|_F + \mu^2 = (1-\mu)^2$. Combining, $|\|M\|_F - \mu| \leq \sqrt{32}\sigma$. Using this fact we get $\Pr[|Y-\mu| > t] \geq \Pr[|Y - \|M\|_F| > t + |\|M\|_F - \mu|] \geq \Pr[|Y - \|M\|_F| > t + \sqrt{32}\sigma]$ Changing of variables $t' = t + \sqrt{32}$ and $t'' = t'/\sqrt{8}\sigma$ and using that $e^{-(t''-2)^2} < 4e^{-t''^2/4}$ the lemma follows.

1.2 Random projections

In this section we compose an arbitrary deterministic $\tilde{d} \times d$ 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 preserve the length of any single *unit* vector $x \in \mathbb{R}^d$, with arbitrary low probability δ of failure:

$$\Pr\left[\, |\, \|AD_s x\|_2 - 1\, | \geq \varepsilon)\right] < \delta \tag{3}$$

Here D_s, like before, is a diagonal matrix such that $D_s(i,i) = s(i)$ are random signs (i.i.d. $\{+1,-1\}$ w.p. $1/2$ each) and δ is chosen according to a desired success probability. Setting $M = AD_x$ where D_x is a diagonal matrix holding on its diagonal $D_x(i,i)$ the values of $x(i)$ we have $\|Ms\| = \|\sum_{i=1}^d A^{(i)} x_i s_i\| = \|AD_s x\|$.

To get the concentration result in Equation 3 from Lemma 1.1 two conditions must hold. First, that $\|M\|_F = 1$. This is true for any unit vector x and matrix A whose columns have norm 1 since $\sum_{i,j} M(i,j)^2 = \sum_j x(j)^2 \sum_i A(i,j)^2$. From this point on, we will limit ourselves to such matrices and refer to them as being column normalized. Second, setting $k = \frac{32\log(16/\delta)}{\varepsilon^2}$, we must have that $\|M\|_2 \leq k^{-1/2}$. In terms of the matrix A, we get that only two conditions must hold for AD_s to be a good random projection for x: (a) it needs to be column normalized, and (b) $\|AD_x\|_2 \leq k^{-1/2}$. This gives a simple characterization of a "good" set for A.

Definition 1.1. *For a given matrix, A, we define the vector pseudonorm of x with respect to A as $\|x\|_A \equiv \|AD_x\|_2$ where D_x is a diagonal matrix such that $D_x(i,i) = x(i)$.* Remark: *since we only consider column normalized matrices, no column of A has norm zero and so, in our case, $\|x\|_A$ is a proper norm.*

Definition 1.2. *We define $\chi(A, \varepsilon, \delta)$ as the intersection of the Euclidian unit sphere and a ball of radius $k^{-1/2}$ in the $\|\cdot\|_A$ norm.*

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

for $k = 32 \log(16/\delta)/\varepsilon^2$.

Lemma 1.3. *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. The lemma follows by substituting: $\|M\|_2 = \|AD_x\|_2 = \|x\|_A \leq k^{-1/2}$ and $\|M\|_F = \|AD_x\|_F = \sqrt{\sum_j x(j)^2 \sum_i A(i,j)^2} = 1$ into (1). □

Since $\|\cdot\|_A$ can be a rather complex norm, below we bound it using more conventional norms.

1.3 An ℓ_p bound on $\|\cdot\|_A$

In this section, for the sake of clarity, we change our notation for the spectral norm of matrices to $\|A\|_{2\to 2}$. This is to point out that it is the norm of A as an operator from ℓ_2 to ℓ_2. More generally, $\|A\|_{a\to b} = \max_{\|x\|_a = 1} \|Ax\|_b$.

Lemma 1.4. *Let p and q be dual norm indices such that $1/p + 1/q = 1$.*

$$\|x\|_A \leq \|x\|_{2p} \|A^T\|_{2\to 2q} \tag{6}$$

Proof.

$$\begin{aligned}
\|x\|_A^2 &= \|AD_x\|_{2\to 2}^2 = \max_{y, \|y\|_2 = 1} \|y^T AD_x\|_2^2 && (7)\\
&= \max_{y, \|y\|_2 = 1} \sum_{i=1}^{d} x^2(i) (y^T A^{(i)})^2 && (8)\\
&\leq \left(\sum_{i=1}^{d} x^{2p}(i) \right)^{1/p} \left(\max_{y, \|y\|_2 = 1} \sum_{i=1}^{d} (y^T A^{(i)})^{2q} \right)^{1/q} && (9)\\
&= \|x\|_{2p}^2 \|A^T\|_{2\to 2q}^2 && (10)
\end{aligned}$$

The transition from (8) to (9) follows from Hölder's inequality for dual norms p and q, satisfying $1/p + 1/q = 1$. □

It is important to state that this bound is not tight and might indeed be too crude for some applications. For example, $\|\cdot\|_p$ is invariant to permuting the vector whereas the $\|\cdot\|_A$ norm is highly influenced by it.

2 Lean Walsh transforms

The *Lean* Walsh Transform, similar to the Walsh Hadamard 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, unlike regular Hadamard matrices, the Lean Walsh seeds have fewer rows than columns. We formally define them as follows:

Definition 2.1. *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 column normalized, $\forall\, j\ \sum_i A_1(i,j)^2 = 1$; iii) the rows of A_1 are orthogonal to each other and equinormed: $\forall\, i\ \sum_j A_1(i,j)^2 = c/r$.*

Definition 2.2. *A_ℓ is a Lean Walsh transform, of order ℓ, if for all $\ell' \le \ell$ 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 2.1.*

The following are examples of seed matrices:

$$A'_1 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \quad A''_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 1 \\ 1 & e^{2\pi i/3} & e^{4\pi i/3} \end{pmatrix} \tag{11}$$

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). In general, any selection of r rows from a $c \times c$ Hadamard or Fourier transform (normalized by $1/\sqrt{r}$) would serve as a possible seed. Elementary properties of Kronecker products give us some properties of A_ℓ.

Fact 2.1. *i) A_ℓ is of size $d^\alpha \times d$, where $\ell = \log(d)/\log(c)$ and $\alpha = \log(r)/\log(c) < 1$ is the skewness of A_1 ii) A_ℓ is column normalized; and iii) the rows of A_ℓ are orthogonal to each other and equinormed $\forall\, i\ \sum_j A_\ell(i,j)^2 = d/d^\alpha$.*[7]

Fact 2.2. *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 sections 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 $T(d) = rT(d/c) + rd$. Due to the Master Theorem, and the fact that $r < c$ we have that $T(d) = O(d)$. More precisely, $T(d) \le dcr/(c-r)$. □

[7] 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 having an effect on the asymptotic running time. From this point on we assume w.l.o.g. $d = c^\ell$ for some integer ℓ

For clarity, we demonstrate Fact 2.2 for A'_1 (equation (11)):

$$A'_\ell z = A'_\ell \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{pmatrix} = \frac{1}{\sqrt{3}} \begin{pmatrix} A'_{\ell-1}(z_1 + z_2 - z_3 - z_4) \\ A'_{\ell-1}(z_1 - z_2 + z_3 - z_4) \\ A'_{\ell-1}(z_1 - z_2 - z_3 + z_4) \end{pmatrix} \tag{12}$$

Remark 2.1. *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, seeds which are $r = c - 1$ rows of $c \times c$ Hadamard of Fourier matrix have coherence $1/r$.*

In what follows we characterize $\chi(A, \varepsilon, \delta)$ for a general Lean Walsh transform by the parameters of its seed. The omitted notation, A, stands for A_ℓ of the right size to be applied to x, i.e., $\ell = \log(d)/\log(c)$. Also, we freely use α to denote the skewness $\log(r)/\log(c)$ of the seed at hand and $\tilde{d} = d^\alpha$ the number of rows in A_ℓ. Note that $\tilde{d}$ is not (in general) equal to k the target dimension. We assume that $k \leq \tilde{d}$ and use a second random projection, R, to move from dimension $\tilde{d}$ to k. This is explained in Section 2.2.

2.1 $\|A^T\|_{2\to 2q}$ for Lean Walsh matrices

We first use Riesz-Thorin's theorem to bound $\|A^T\|_{2\to 2q}$ of any Lean Walsh matrix.

Theorem 2.1. [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 $\|A^T\|_{2\to 2} = \sqrt{d/\tilde{d}} = d^{(1-\alpha)/2}$. 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 $\|A^T\|_{2\to 2q} \leq d^{(1-\alpha)/2q}$.

Remark 2.2. *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 2.1. *For a Lean Walsh transform, A, we have that for any $p > 1$ the following holds:*

$$\{x \in \mathbb{S}^{d-1} \mid \|x\|_{2p} \leq k^{-1/2} d^{-\frac{1-\alpha}{2}(1-\frac{1}{p})}\} \subset \chi(A, \varepsilon, \delta) \tag{13}$$

where $k = 32\log(16/\delta)/\varepsilon^2$ 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:

$$\begin{aligned} \|x\|_A &\leq \|x\|_{2p} \|A^T\|_{2\to 2q} && (14) \\ &\leq \|x\|_{2p} d^{\frac{1-\alpha}{2q}} && (15) \\ &\leq \|x\|_{2p} d^{\frac{1-\alpha}{2}(1-\frac{1}{p})} && (16) \end{aligned}$$

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 2.3. *For $p \to \infty$, only the spectral norm of A comes into play. In this case a simple matrix containing d/k copies of $k \times k$ identity matrices is sufficient. The spectral norm of such matrices is the same as that of Lean Walsh matrices and they are clearly row orthogonal and column normalized. However, their norm as operators from ℓ_2 to ℓ_{2q}, for q larger than 1, is large and fixed, whereas that of Lean Walsh matrices is still arbitrarily small and is controlled by the size of the their seed.*

2.2 Controlling α and choosing R

The target dimension k is, in general, smaller than $\tilde{d}$ and so a second $k \times \tilde{d}$ random projection matrix R is required. Since the skewness α controls both $\tilde{d}$ and χ it is important to chose it carefully such that the entire projection can be performed in $O(d)$ operations and χ is as large as possible.

First we see that increasing α is beneficial from the theoretical stand point since it weakens the constraint on $\|x\|_p$. From an application standpoint, 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 above for which $r = c - 1$. Their skewness $\alpha = \log(r)/\log(c)$ approaches 1 as their size increases. Namely, for any positive constant δ there exists a constant size seed such that $1 - 2\delta \leq \alpha \leq 1$.

Lemma 2.2. *For any positive constant $\delta > 0$ there exists a Lean Walsh matrix A such that:*

$$\{x \in \mathbb{S}^{d-1} \mid \|x\|_\infty \leq k^{-1/2} d^{-\delta}\} \subset \chi(A, \varepsilon, \delta) \tag{17}$$

Proof. Generate A from a seed such that its skewness $\alpha = \log(r)/\log(c) \geq 1 - 2\delta$ and substitute $p = \infty$ into the statement of Lemma 2.1. □

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 $\tilde{d}$ ($\tilde{d} = d^\alpha$) to dimension k in $O(d)$ operations. This can be done using the Ailon and Liberty [25] construction

serving as the random projection matrix R. R is a $k \times \tilde{d}$ Johnson Lindenstrauss projection matrix which can be applied in $\tilde{d}\log(k)$ operations if $\tilde{d} = d^{\alpha} \geq k^{2+\delta''}$ for arbitrary small δ''. For the same choice of a seed as in lemma 2.2, the condition becomes $d \geq k^{2+\delta''+2\delta}$ which can be achieved by $d \geq k^{2+\delta'}$ for arbitrary small δ' depending on δ and δ''. Therefore for such values of d the matrix R exists and requires $O(d^{\alpha}\log(k)) = O(d)$ operations to apply.

2.3 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 2.2 (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*

$$S = \begin{cases} +\frac{1}{\sqrt{qk}} & \text{with probability } q/2 \\ -\frac{1}{\sqrt{qk}} & \text{with probability } q/2 \\ 0 & \text{with probability } 1-q \end{cases} \tag{18}$$

Let k be $C_1\log(1/\delta)/\varepsilon^2$ for a sufficiently large C_1. Let the matrix A contain in each entry an independent copy of S then:

$$\Pr[|\|Ax\|_2^2 - 1| > \varepsilon] \leq \delta \tag{19}$$

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

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

Since the running time of general dimensionality reduction is at least $O(d)$ and at most $O(d\log(k))$ (due to [25]), matrices that can be applied within this time range are especially interesting. We claim that although Lean Walsh matrices can be applied in linear time, their ℓ_{∞} requirement is weaker than that achieved by sparse projection matrices. This holds even if the latter are allowed an $O(d\log(k))$ running time. For $q = k^{-1}\log(k)$ (in equation (18)) a matrix A as above contains $O(d\log(k))$ nonzeros w.h.p. and thus can be applied in that amount of time. Due to theorem 2.2 this value of q requires $\|x\|_{\infty} \leq O(k^{-1}\sqrt{\log k})$ for the length of x to be preserved w.h.p. For $d = poly(k)$ this is a stronger constraint on the ℓ_{∞} norm of x than $\|x\|_{\infty} \leq O(k^{-1/2}d^{-\delta})$ which is required by Lean Walsh transforms.

3 Conclusion and future work

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 Euclidian 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 enjoy both a "large" χ and a linear time computation scheme which outperforms sparse projection matrices. These properties make them good building blocks for the purpose of random projection.

In [24, 25] the projection included a preprocessing stage which, in the language of this paper, map vectors from $\mathbb{S}^{d-1}$ to χ with high probability. Designing such mappings for Lean Walsh matrices which require only $O(d)$ operations would give an optimal $O(d)$ random projection algorithm. Possible choices for these may combine random permutations, various wavelet/wavelet-like transforms, or any other sparse orthogonal transformation.

After the publication of this paper, a new result by Dasgupta, Kumar and Sarlos [27] was accepted to publication. Their result corrects and extends [28]. The authors generate an implicit random matrix using a hash function which is closely related to combining the construction from remark 2.3 with a random permutation preprocessor.

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

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] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC)*, pages 604–613, 1998.

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

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

[6] J. Matousek. On variants of the Johnson-Lindenstrauss lemma. *Preprint*, 2006.

[7] P. Drineas, M. W. Mahoney, and S.M. Muthukrishnan. 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] T. Sarlós. 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] A. M. Frieze, R. Kannan, and S. Vempala. Fast monte-carlo algorithms for finding low-rank approximations. In *IEEE Symposium on Foundations of Computer Science*, pages 370–378, 1998.

[10] S. Har-Peled. A replacement for Voronoi diagrams of near linear size. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 94–103, Las Vegas, Nevada, USA, 2001.

[11] D. Achlioptas and F. McSherry. Fast computation of low rank matrix approximations. In *STOC: ACM Symposium on Theory of Computing (STOC)*, 2001.

[12] P. Drineas and R. Kannan. Fast monte-carlo algorithms for approximate matrix multiplication. In *IEEE Symposium on Foundations of Computer Science*, pages 452–459, 2001.

[13] P.G. Martinsson, V. Rokhlin, and M. Tygert. A randomized algorithm for the approximation of matrices. *In review. Yale CS research report YALEU/DCS/RR-1361.*, 2007.

[14] E. Liberty, F. Woolfe, P. G. Martinsson, V. Rokhlin, and M. Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences (PNAS)*, Dec 2007.

[15] A. Dasgupta, P. Drineas, B. Harb, R. Kumar, and M. W. Mahoney. Sampling algorithms and coresets for ℓ_p regression. *Proc. of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2008.

[16] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlos. Faster least squares approximation. *TR arXiv:0710.1435*, submitted for publication, 2007.

[17] P. Drineas, M.W. Mahoney, and S. Muthukrishnan. Relative-error cur matrix decompositions. *TR arXiv:0708.3696*, submitted for publication, 2007.

[18] E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, Dec. 2006.

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

[20] M. Elad. Optimized projections for compressed sensing. *IEEE Transactions on Signal Processing*, 55(12):5695–5702, 2007.

[21] P. Paschou, E. Ziv, E. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas. Pca-correlated snps for structure identification in worldwide human populations. *PLOS Genetics, 3, pp. 1672-1686*, 2007.

[22] P. Paschou, M. W. Mahoney, J. Kidd, A. Pakstis, K. Kidd S. Gu, and P. Drineas. Intra- and inter-population genotype reconstruction from tagging snps. *Genome Research, 17(1), pp. 96-107*, 2007.

[23] E. Liberty and S.W. Zucker. The mailman algorithm: A note on matrix–vector multiplication. *Inf. Process. Lett.*, 109(3):179–182, 2009.

[24] N. Ailon and B. 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.

[25] N. Ailon and E. Liberty. Fast dimension reduction using Rademacher series on dual BCH codes. In *Symposium on Discrete Algorithms (SODA), accepted*, 2008.

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

[27] A. Dasgupta, R. Kumar, and T. Sarlos. A Sparse Johnson-Lindenstrauss Transform. 2010.

[28] K. Q. Weinberger, A. Dasgupta, J. Langford, A. J. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *ICML*, page 140, 2009.