Fast Random Projections using Lean Walsh Transforms

Abstract

We present a random projection matrix that is applicable to vectors $x \in \mathbb{R}^d$ in $O(d)$ operations if $d \geq k^{2+\delta'}$ for arbitrary small δ' and $k = O(\log(n)/\varepsilon^2)$. The projection succeeds with probability $1 - 1/n$ and it preserves lengths up to distortion ε for all vectors such that $\|x\|_\infty \leq \|x\|_2 k^{-(1+\delta)}$ for arbitrary small δ. Previous results are either not applicable in linear time or require a bound on $\|x\|_\infty$ that depends on d.

1 Introduction

We are interested in producing a random projection procedure such that applying it to a vector in $\mathbb{R}^d$ requires $O(d)$ operation. The idea is to first project an incoming vector $x \in \mathbb{R}^d$ into an intermediate dimension $\tilde{d} = d^\alpha$ $(0 < \alpha < 1)$ and then use a known construction to project it further down to the minimal Johnson Lindenstrauss (JL) dimension $k = O(\log(n)/\varepsilon^2))$. The mapping is composed of three matrices $x \mapsto RAD_s x$. The first, D_s, is a random ± 1 diagonal matrix. The second, A, is a dense matrix of size $\tilde{d} \times d$ termed the Lean Walsh Matrix. The third, R, is a $k \times \tilde{d}$ Fast Johnson Lindenstrauss matrix.

As R, we use the fast JL construction by Ailon and Liberty [2]. Their matrix, R, can be applied to $AD_s x$ in $O(\tilde{d}\log(k))$ operations as long as $\tilde{d} \geq k^{2+\delta''}$ for arbitrary small δ''. If $d > k^{(2+\delta'')/\alpha}$ this condition is met[1]. Therefore, for such d, applying R to $AD_s x$ is length preserving w.p and accomplishable in $O(\tilde{d}\log(k)) = O(d^\alpha \log(k)) = O(d)$ operations.

The main purpose of this manuscript is to produce a dense $\tilde{d} \times d$ matrix A such that $x \mapsto AD_s x$ is computable in $O(d)$ operation, and preserves length w.p for vectors such that $\|x\|_\infty \leq \|x\|_2 k^{-(1+\delta)}$ for arbitrarily small δ.

[1]Later in the manuscript we see that α can be chosen arbitrarily close to 1.

2 Norm concentration

For a matrix to exhibit the property of a random projection, it is enough for it to preserve the length of any *one* vector with very high probability:

$$\Pr\left[|\ \|AD_s x\|^2 - 1\ | \geq \varepsilon) \right] < 1/n \tag{1}$$

where A is a column normalized matrix, D_s is a random diagonal ± 1 w.p $1/2$ matrix, x is a general vector (w.l.o.g $\|x\|_2 = 1$), and n is chosen according to a desired success probability, usually polynomial in the number of projected vectors.

Notice 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 s is a vector of random ± 1. By denoting $M = AD_x$ we view the term $\|Ms\|$ as a convex function over the product space $\{1,-1\}^d$ from which the variable s is chosen. In his book, Talagrand [4] describes a strong concentration result for convex Lipschitz bounded functions over probability product spaces.

Lemma 2.1 (Talagrand [4])**.** *Consider a matrix M and a random entree-wise i.i.d ± 1 vector s. Define the random variable $Y = \|Ms\|_2$. Denote by μ the median of Y and $\sigma = \|M\|_{2\to 2}$ the spectral norm of M. The following holds:*

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

Lemma 2.1 asserts that $\|AD_x s\|$ distributes 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 (or normalized in expectancy). 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, by Jensen, $(E[Y])^2 \leq E[Y^2] = 1$. Hence $E[(Y-\mu)^2] = E[Y^2] - 2\mu E[Y] + \mu^2 \geq 1 - 2\mu + \mu^2 = (1-\mu)^2$. Combining, $|1-\mu| \leq \sqrt{32}\sigma$. By setting $\varepsilon = t + |1-\mu|$, we get:

$$\Pr[|Y - 1| > \varepsilon] < 4e^4 e^{-\varepsilon^2/8\sigma^2} \tag{3}$$

If we set $k = 16\log(n)/\varepsilon^2$ (assuming $\log(n) > 6$) the requirement of equation 1 is met for $\sigma = k^{-1/2}$. We see that a condition on $\sigma = \|AD_x\|$ is sufficient for the projection to succeed w.h.p. This naturally defines χ.

Definition 2.1. *Let $\chi \subset \mathbb{R}^d$ be such that for all $x \in \chi$ we have $\|x\|_2 = 1$ and the spectral norm of AD_x is at most $k^{-1/2}$.*

$$\chi(A,\varepsilon,n) = \left\{ x \in \mathbb{R}^d \mid \|x\| = 1, \|AD_x\|_2 \leq k^{-1/2} \right\} \tag{4}$$

where D_x is a diagonal matrix such that $D_x(i,i) = x(i)$ and $k = 16\log(n)/\varepsilon^2$.

Lemma 2.2. *For any column normalized matrix, A, and a random ± 1 diagonal matrix, D_s, the following holds:*

$$\forall x \in \chi(A, \varepsilon, n) \quad \Pr[\big|\|AD_s x\|^2 - 1\big| \geq \varepsilon] \leq 1/n \tag{5}$$

Proof. By the definition of χ (2.1) and substituting 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. Our goal is to characterize $\chi(A, \varepsilon, n)$ for a specific fixed matrix A, which we term *Lean Walsh*. It is described in the next section.

3 Lean Walsh transforms

The *Lean* Walsh Transform, similar to the Walsh Transform, is a recursive tensor product matrix defined by a constant seed matrix, A_1, and constructed recursively by using Kronecker tensor products $A_{\ell'} = A_1 \otimes A_{\ell'-1}$.

Definition 3.1. *A_1 is a Lean Walsh seed (or simply 'seed') if i) A_1 is a rectangular, possibly complex, matrix $A_1 \in \mathbb{C}^{r \times c}$, such that $r < c$; ii) A_1 is absolute valued $r^{-1/2}$ entree-wise, i.e, $|A_1(i,j)| = r^{-1/2}$; iii) the rows of A_1 are orthogonal; iv) all inner products between its different columns are bounded, in absolute value, by a constant $\rho < 1$. ρ is called the Coherence of A_1.*

Definition 3.2. *A_ℓ is a Lean Walsh transform if for all $\ell' \leq \ell$ we have $A'_\ell = A_1 \otimes A_{\ell'-1}$. Here $\otimes$ stands for the Kronecker product and A_1 is a seed as in definition 3.1.*

The following are examples of seed matrices:

$$A'_1 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}, \; A''_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 1 \\ 1 & e^{2\pi i/3} & e^{4\pi i/3} \end{pmatrix} \tag{6}$$

These examples are a part of a large family of possible seed matrices. 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 3.1. *Let F^c be the $c \times c$ Discrete Fourier matrix, not normalized such that $|F^c(i,j)| = 1$. 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 viable Lean Walsh seed with coherence $1/r$.*

Proof. The fact that $|A_1(i,j)| = 1/\sqrt{r}$ and the fact that the rows of A_1 are orthogonal are trivial. Moreover, the inner product of two different columns of A_1 must be $\rho = 1/r$ in absolute value.

$$\sum_i^c (F^c(i,j_1))^* F^c(i,j_2) = 0 \quad (7)$$

$$\frac{1}{r}\sum_i^r (F^c(i,j_1))^* F^c(i,j_2) = -\frac{1}{r}(F^c(c,j_1))^* F^c(c,j_2) \quad (8)$$

$$|\langle A_1^{(j1)}, A_1^{(j2)}\rangle| = \frac{1}{r} \quad (9)$$

□

From elementary properties of Kronecker products we characterize A_ℓ. Using the number of rows, r, number of columns, c, and the coherence, ρ, of A_1.

Fact 3.2. *The following facts hold true for A_ℓ: i) A_ℓ is of size $d^\alpha \times d$,$\alpha = \log(r)/\log(c) < 1$[2]. $c^\ell = d$ and $r^\ell = \tilde{d}$ we have that $\tilde{d} = d^\alpha$, $\alpha = \log(r)/\log(c) < 1$. ii) for all i and j, $A_\ell(i,j) \in \pm\tilde{d}^{-1/2}$ which means that A_ℓ is column normalized. iii) the rows of A_ℓ are orthogonal.*

Fact 3.3. *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 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 $T(d) = rT(d/c) + O(d)$. Due to the Master Theorem, and the fact that $r < c$, $T(d) = O(d)$, or more specifically $T(d) \le dcr/(c-r)$. □

For clarity, we demonstrate Fact 3.3 for the first example, A'_1:

$$A'_\ell z = A'_\ell \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{pmatrix} = \frac{1}{\sqrt{3}} \begin{pmatrix} A'_{\ell-1}(z_1 + z_2 - z_3 - z_4) \\ A'_{\ell-1}(z_1 - z_2 + z_3 - z_4) \\ A'_{\ell-1}(z_1 - z_2 - z_3 + z_4) \end{pmatrix} \quad (10)$$

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

[2]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 affecting the asymptotic running time

4 $\chi(A, \varepsilon, n)$ for the Lean Walsh Transform.

After describing the possible Lean Walsh Transforms we turn our attention to characterizing their "good" sets $\chi(A, \varepsilon, n)$. We claim that an ℓ_∞ bound on the x depending on k is sufficient to insure that $x \in \chi(A, \varepsilon, n)$. In order to use lemma 2.2 we upper bound $\sigma = \|AD_x\|_{2\to 2}$ for all unit length x such that $\|x\|_\infty \le 1/\sqrt{m}$ for some integer m.

$$\sigma^4 \le \|D_x A^* A D_x\|_{\text{Fro}}^2 = \sum_{i,j} B(i,j) x^2(i) x^2(j) \tag{11}$$

Here $B(i,j) = ((A^*A)(i,j))^2$. We define $y(i) = x^2(i)$. Thus $0 \le y(i) \le 1/m$ and $\sum y(i) = 1$. The term $\sum_{i,j} B(i,j) y(i) y(j)$ is a strictly convex function over the polytope defined by the constraints on y. This means that the maximal value is achieved at an extremal point of this polytope. Therefore, the maximal value is achieved for a vector y for which m of the y_i are equal to $1/m$ and the rest are zero. Let μ be the set of non zero indices in y:

$$\max_y \sum_{i,j} B(i,j) y(i) y(j) = \max_{\mu, |\mu|=m} \frac{1}{m^2} \sum_{i,j\in\mu} B(i,j) \tag{12}$$

Lemma 4.1. *Let A_ℓ be a Lean Walsh Matrix and let $B_\ell(i,j) = ((A_\ell^* A_\ell)(i,j))^2$. Also, let μ be any subset of the indices $\mu \subset \{1, \ldots, d\}$ such that $|\mu| \le m$:*

$$\sum_{i,j\in\mu} B_\ell(i,j) \le m^{2-\gamma} \tag{13}$$

for $\gamma = 1 - \log(\rho^2(c-1)+1)/\log(c)$. Here c is the number of columns and ρ is the coherence of the seed of A_ℓ.

Proof. The proof goes by induction on both the size of μ and ℓ. The assertion for the base case for which $|\mu| = 1$ and $\ell = 1$ is trivial since the largest entree in B_1 is 1. Let us denote by u the characteristic vector of μ. $u(i) = 1$ for $i \in \mu$ and zero else. Clearly $\sum_{i,j\in\mu} B_\ell(i,j) = u^T B_\ell u$. From the structure of the seed we get that B_ℓ consists of diagonal blocks of $B_{\ell-1}$ and off-diagonal blocks of $\rho^2 B_{\ell-1}$:

$$B_\ell = \begin{pmatrix} B_{\ell-1} & \rho^2 B_{\ell-1} & \cdots & \rho^2 B_{\ell-1} \\ \rho^2 B_{\ell-1} & B_{\ell-1} & \cdots & \rho^2 B_{\ell-1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^2 B_{\ell-1} & \rho^2 B_{\ell-1} & \cdots & B_{\ell-1} \end{pmatrix} \tag{14}$$

Let us now cut u into c sections of length d/c, each denoted by u_i.

$$u^T B_\ell u \quad = \quad \sum_{i=1}^{c} u_i^T B_{\ell-1} u_i + \rho^2 \sum_{i\ne j=1}^{c} u_i^T B_{\ell-1} u_j \tag{15}$$

Using the fact that $B_{\ell-1}$ is Symmetric and Positive Semi Definite

$$
\begin{aligned}
(u_i - u_j)^T B_{\ell-1}(u_i - u_j) &\geq 0 &(16)\\
\left[u_i^T B_{\ell-1} u_i + u_j^T B_{\ell-1} u_j\right]/2 &\geq u_j^T B_{\ell-1} u_i &(17)
\end{aligned}
$$

Substituting into equation 15

$$
\begin{aligned}
u^T B_\ell u &\leq \sum_{i=1}^{c} u_i^T B_{\ell-1} u_i + \frac{\rho^2}{2} \sum_{i \neq j=1}^{c} u_i^T B_{\ell-1} u_i + u_j^T B_{\ell-1} u_j &(18)\\
&= (1 + \rho^2(c-1)) \sum_{i=1}^{c} u_i^T B_{\ell-1} u_i &(19)
\end{aligned}
$$

We let m_i denote the number of non zeros in u_i and use the induction assumption that $u_i^T B_{\ell-1} u_i \leq m_i^{2-\gamma}$:

$$u^T B_\ell u \leq (1 + \rho^2(c-1)) \sum_{i}^{c} m_i^{2-\gamma} \tag{20}$$

From $\sum_{i=1}^{c} m_i = m$ we have that $\sum_i^c (m_i)^{2-\gamma} \leq c(m/c)^{2-\gamma}$:

$$u^T B_\ell u \leq (1 + \rho^2(c-1))c(m/c)^{2-\gamma} \leq m^{2-\gamma} \tag{21}$$

which holds true for $\gamma = 1 - \log(\rho^2(c-1)+1)/\log(c)$ and completes the proof. □

Lemma 4.2. *For the Lean Walsh matrix A,*

$$\left\{x \in \mathbb{R}^d \mid \|x\|_2 = 1\ ,\ \|x\|_\infty \leq k^{-1/\gamma}\right\} \subset \chi(A, \varepsilon, n) \tag{22}$$

where $\gamma = 1 - \log(\rho^2(c-1)+1)/\log(c)$. c is the number of columns and ρ is the coherence of the seed of A.

Proof. Combining the above we get

$$\sigma^4 \leq \sum_{i,j} B(i,j)x^2(i)x^2(j) \leq \frac{1}{m^2} \sum_{i,j \in \mu, |\mu|=m} B(i,j) \leq m^{-\gamma}. \tag{23}$$

By setting $m = k^{2/\gamma}$ we obtain $\sigma \leq k^{-1/2}$, which in turn, translates to the ℓ_∞ norm constraint on x, $\|x\|_\infty \leq m^{-1/2} = k^{-1/\gamma}$. □

Indeed it is well known that one can project vectors $x \in R^d$ in linear time, by sampling, if $\|x\|_\infty = O(d^{-1/2})$. However, here the constraint on the ℓ_∞ norm of x is independent of d, namely $\|x\|_\infty = O(k^{-1/\gamma})$. In other words, for large d, x might be quite sparse.

4.1 Controlling γ and α

We see that increasing γ is beneficial from the theoretical stand point since it weakens the constraint on $\|x\|_\infty$. 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 3.1. These seeds consist of r rows, $c = r + 1$ columns, and are $1/r$ coherent. Substituting into the term for γ we get

$$\gamma = 1 - \log(\rho^2(c-1)+1)/\log(c) = \log(r)/\log(c) \tag{24}$$

which approaches 1 as the size of the seed increases. More precisely, for any constant positive δ there exists a constant size seed such that $1 - \delta/2 \leq \gamma \leq 1$.

Lemma 4.3. *For any positive $\delta > 0$, there exists a Lean Walsh matrix, A, such that:*

$$\{x \in \mathbb{R}^d \mid \|x\|_2 = 1\ ,\ \|x\|_\infty \leq k^{-(1+\delta)}\} \subset \chi(A, \varepsilon, n) \tag{25}$$

Proof. Generate A from a seed described in Fact 3.1 such that $\gamma = \log(r)/\log(c) \geq 1 - \delta/2$. Substitute into Lemma 4.2 and use $1/(1-\delta/2) \leq 1 + \delta$. □

The constant α 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 to dimension k in $O(d)$ operations. This is done using the Ailon and Liberty [2] construction serving as a 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 $\tilde{d} \geq k^{2+\delta''}$ for arbitrary small δ''. We remind the reader that $\tilde{d} = d^\alpha$ and that $\alpha = \log(r)/\log(c)$. Therefore $1 - \delta/2 \leq \alpha \leq 1$ for the same choice of seed as lemma 4.3. Clearly we have that $\tilde{d} \geq d^{1-\delta/2} \geq k^{2+\delta''}$ is achievable by $d \geq k^{2+\delta'}$ for δ' arbitrary small 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.

For the purpose of this result, the seeds described in Fact 3.1 are sufficient. However, it is clear that they are not necessarily optimal. We have seen that it is beneficial to reduce the coherence. By a Humming bound on the sphere we see that this requires increasing either the number of columns or rows. On the other hand, reducing the number of rows improves the running time (possibly, by a large constant). We assert that, for a given coherence and number of columns, one should try to minimize the number of rows in the seed[3]. This means that optimal seeds constitute a spherical code or a line packing matrix. Both of the given examples are optimal in that sense.

5 $\chi(A, \varepsilon, n)$, another approach

We turn to describe another approach for characterizing $\chi(A, \varepsilon, n)$. Whereas before we considered only the ℓ_∞ norm, here we consider any norm larger then 2. This guaranties high probability for success in more situations. Although the bound we give depends polynomially on d, the dependence is week enough to improve on the results of the last section when d is a small polynomial in k.

[3]One should make sure that that the constraints on d and $\|x\|_\infty$ are acceptable for the application at hand.

We remind the reader that our objective is to show that $\sigma = \|AD_x\|_{2\to 2} \leq k^{-1/2}$. We compute the norm of $\|AD_x\|_{2\to 2}$:

$$\begin{aligned}
\|AD_x\|_{2\to 2}^2 &= \max_{y, \|y\|_2=1} \|y^T AD_x\|_2^2 & (26)\\
&= \max_{y, \|y\|_2=1} \sum_{i=1}^{d} x^2(i)(y^T A^{(i)})^2 & (27)\\
&\leq \left(\sum_{i=1}^{d} x^{2p}(i)\right)^{1/p} \left(\max_{y, \|y\|_2=1} \sum_{i=1}^{d} (y^T A^{(i)})^{2q}\right)^{1/q} & (28)\\
&= \|x\|_{2p}^2 \|A^T\|_{2\to 2q}^2 & (29)
\end{aligned}$$

Where 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 are now faced with the computing $\|A^T\|_{2\to 2q}$ in order to obtain the constraint on $\|x\|_{2p}$.

Theorem 5.1. [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 the fact that $\|A^T\|_{2\to 2} = \|A\|_{2\to 2}$ 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 $\|A^T\|_{2\to 2q} \leq d^{(1-\alpha)/2q}$.

Lemma 5.1. *For the lean Walsh transform A, we have that for any $p \geq 1$:*

$$\cup_{p\geq 1}\{x \in \mathbb{R}^d \mid \|2\| = 1, \|x\|_{2p} \leq k^{-1/2} d^{-\frac{1-\alpha}{2}(1-\frac{1}{p})}\} \subset \chi(A, \varepsilon, n) \tag{30}$$

Where $k = O(\log(n)/\varepsilon^2)$, $\alpha = \log(r)/\log(c)$, r and c are the number of rows and columns in the seed of A.

Proof. Combining the above, and using the duality of p and q, we get that:

$$\begin{aligned}
\|AD_x\|_{2\to 2} &\leq \|x\|_{2p} \|A^T\|_{2\to 2q} & (31)\\
&\leq \|x\|_{2p} d^{\frac{1-\alpha}{2q}} & (32)\\
&\leq \|x\|_{2p} d^{\frac{1-\alpha}{2}(1-\frac{1}{p})} & (33)
\end{aligned}$$

And so $\|AD_x\|_{2\to 2} \leq k^{-1/2}$ is achieved if $\|x\|_{2p} \leq k^{-1/2} d^{-\frac{1-\alpha}{2}(1-\frac{1}{p})}$, for any $p \geq 1$. $\square$

This result improves on our previous ℓ_∞ bound for the cases in which $d \leq k^{1/(1-\alpha)}$, where the parameter α can be chosen arbitrarily close to 1. However, more importantly, if allows other norm bounds as well as the ℓ_∞.

6 Conclusion and work in progress

We showed that vectors in $\mathbb{R}^d$ can be randomly projected, with distortion ε, in $O(d)$ operations if they are not pathologically sparse i.e their ℓ_∞ is at most $\mathbf{poly}(1/\log(n), \varepsilon)$. It was, and still is, our goal to produce linear time projections for *all* of $\mathbb{R}^d$. We hope that this manuscript takes a step in this direction. We hope to improve this result by randomizing A. Namely induce a random permutation on it's columns. Intuitively this should make the ℓ_∞ constraint an increasing function of d. Hopefully, for large values of d, this might eliminate the constraint altogether.

References

[1] 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.

[2] Nir Ailon and Edo Liberty. Fast dimension reduction using rademacher series on dual bch codes. In *Symposium on Discrete Algorithms (SODA), accepted*, 2008.

[3] G. Bennett. Schur multipliers. *Duke Math. J.*, 4(3):603–639, 1997.

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

A Reducing $\|x\|_\infty$

In [1] Ailon and Chazelle proposed a composition of three matrices as a projection matrix, $x \mapsto PHD_s x$. Where P is a specific random sparse matrix. In the terminology of this paper, they showed that

$$\chi(P, \varepsilon, n) = \left\{ x \in \mathbb{R}^d \mid \|x\|_2 = 1 \ , \ \|x\|_\infty \leq (\log(n)/d)^{1/2} \right\} . \tag{34}$$

Whereas we view the mapping $x \mapsto HD_s x$ (H is the Complete Walsh transform) as a preprocessing which maps all vectors in $\mathbb{R}^d$ into χ with high probability. The same idea can be applied here to give a random projection for all $\mathbb{R}^d$ in time $d \log k$. We require that $\|x\|_\infty \leq k^{-1/\gamma}$. We achieve this by cutting the vector x into blocks and performing a randomized fast orthogonal transformation on each one of them separately. More precisely:

Lemma A.1. *Let $b = k^{2/\gamma}\log(n)$ and H_b denotes the $b \times b$ Walsh Hadamard Transform. Also, let $I_{d/b}$ be a d/b identity matrix, and $D_{s'}$ a random ± 1 diagonal $d \times d$ matrix. Let $\Phi = (I_{d/b} \otimes H_b) D_{s'}$, then $\|\Phi x\|_2 = \|x\|_2$*

and:

$$\Pr[\|\Phi x\|_\infty \geq k^{-1/\gamma}] \leq 1/n \tag{35}$$

Proof. the fact that $\|\Phi x\|_2 = \|x\|_2$ is trivial since the transformation is an isometry in ℓ_2. Using the Hoeffding bound we see that w.p $1 - 1/n$, $\|\Phi x\|_\infty \leq (\log(n)/b)^{1/2}$. Substituting $b = k^{2/\gamma}\log(n)$ gives the desired result. $\square$

The time to apply $(I_{d/b} \otimes H_b)D_{s'}$ to a vector is $d\log(b) = O(d\log(k))$ since we apply d/b Walsh transforms, each in time $b\log(b)$.

It is worth mentioning that Liberty and Ailon already showed in [2] a $O(d\log(k))$ algorithm, and so this section's result is not new per se. However, it was important to see that a trivial modification to our algorithm can achieve the best know running time for the general problem.