Electronic Colloquium on Computational Complexity, Report No. 70 (2007)

Fast Dimension Reduction Using Rademacher Series on Dual BCH Codes

Nir Ailon* Edo Liberty†

Abstract

The Fast Johnson-Lindenstrauss Transform was recently discovered by Ailon and Chazelle as a technique for performing fast dimension reduction from ℓ_2^d to ℓ_2^k in time $O(\max\{d \log d, k^3\})$, where k is the target lower dimension. This beats the naive $O(dk)$ achieved by multiplying by random dense matrices, as noticed by several authors following the seminal result by Johnson and Lindenstrauss from the mid 80's. In this work we show how to perform a dimension reduction onto $k < d^{1/2-\delta}$ dimensions in time $O(d \log k)$ for arbitrary small δ. This beats Ailon et al's algorithm for $k > d^{1/3}$ and for $k = d^{o(1)}$. In order to achieve this, we analyze Rademacher series in Banach spaces (sums of vectors in Banach spaces with random signs) using a powerful measure concentration bound due to Talagrand. The set of vectors used is a real embedding of dual BCH code vectors over $GF(2)$. We also discuss reductions onto ℓ_1 space.

1 Introduction

Applying random matrices is by now a well known technique for reducing dimensionality of vectors in Euclidean space while preserving certain properties (most notably distance information). Beginning with the classic work of Johnson and Lindenstrauss [15] who used projections onto random subspaces, other variants of the technique using different distributions are known [1, 9, 14, 20] and have been used in many algorithms [16, 18, 3, 12, 23, 22, 11].

In all the variants of this idea, a fixed unit length vector $x \in \mathbf{R}^d$ is mapped onto $\mathbf{R}^k$ $(k < d)$ via a random linear mapping Φ from a carefully chosen distribution. A measure concentration principle is used to show that the distribution of the norm estimator error $|\|\Phi x\|_2 - 1|$ in a small fixed neighborhood of 0 is dominated by a gaussian of standard deviation $\Omega(k^{-1/2})$, uniformly for all x and independent on d. The distribution of Φ need not even be rotationally invariant. When used in an algorithm, k is often chosen as $O(\varepsilon^{-2} \log n)$ so that a union bound ensures that the error is smaller than a fixed ε simultaneously for all n vectors in some fixed input set. Noga Alon proved [2] that this choice of k is essentially optimal and cannot be significantly reduced.

It makes sense to abstract the definition of a distribution of mappings that can be used for dimension reduction in the above sense (we do this formally in Section 2). We will say that such distributions have the Johnson-Lindenstrauss (JL) property (named after the authors of the first such construction). A natural question to ask is, how much computational resources are necessary in order to apply the mapping on a vector? The resources considered here are time and randomness. Ailon and Chazelle [1] made a first nontrivial step in studying this question, and showed that

*Institute for Advanced Study, Princeton NJ
†Yale University, New Haven CT

ISSN 1433-8092

reduction from d dimensions to k dimensions can be performed in time $O(\max\{d \log d\}, k^3)$, beating the naïve $O(kd)$ for $k = o(\log d)$ or $k = O(d^{1/2})$. Similar bounds were found in [1] for reducing onto ℓ_1 (Manhattan) space, but with quadratic (not cubic) dependence on k. From recent work by Matousek [20] it can be shown, by replacing gaussian distributions with ± 1's, that Ailon and Chazelle's algorithm requires $O(d)$ random bits in the Euclidean case. For the Manhattan case, Matousek shows that $O(d)$ random bits are sufficient but the running time becomes cubic (and not quadratic) in k.

1.1 Overview of Our Results

This work contains several contributions. The first (in Section 7) is a simple trick that can be used to reduce the running time in [1] to $O(\max\{d \log k\}, k^3)$, hence making it better than the naïve algorithm for all small k. In typical applications, the running time translates to $O(d \log \log n)$, where n is the number of points we simultaneously want to reduce (assuming $n = 2^{O(d^{1/3})}$).

The main contribution (Sections 5-6) is taking care of the case, say, $k \geq d^{1/4}$. We need tools from the theory of probability and norm interpolation in Banach spaces (Section 3) as well as the theory of error correcting codes (Section 4) to construct a distribution on matrices satisfying the JL property that can be applied in time $O(d \log d)$ (note that in this case $\log d = O(\log k)$). The ideas used in our constructions are interesting in their own right, because they take advantage of advanced ideas from different classic theories. It is likely that they can be used in conjunction with very recent research on explicit embeddings of ℓ_2 in ℓ_1 [21, 13, 4] as well as research on fast approximate linear algebraic scientific computation [22, 10, 6, 7, 8].

It is illustrative to point out the an apparent weakness in [1] which was a starting point of our work. The main tool used there is to multiply the input vector x by a random sign change matrix followed by a Fourier transform, resulting in a vector y. It is claimed that $\|y\|_\infty$ is small (in other words, the "information" is spread out evenly among the coordinates). By a convexity argument the "worst case" y assuming the ℓ_∞ bound *only* is a *uniformly supported* vector, namely, a vector in which the absolute value of the coordinates in its support are all equal. Intuitively, such a vector is extremely unlikely. In this work we consider other norms.

2 Definitions and Statement of Our Theorems

We use ℓ_p^d to denote d dimensional real space equipped with the norm $\|x\| = \|x\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$, where $1 \leq p < \infty$ and $\|x\|_\infty = \max\{|x_i|\}$. The dual norm index q is defined by the solution to $1/q + 1/p = 1$. We remind the reader that $\|x\|_p = \sup_{\substack{y \in \ell_q^d \\ \|y\|=1}} x^T y$. For a real $k \times d$ matrix A, the matrix norm $\|A\|_{p_1 \to p}$ is defined as the operator norm of $A : \ell_{p_1}^d \to \ell_p^k$ or:

$$\|A\|_{p_1 \to p} = \sup_{\substack{x \in \ell_{p_1}^d \\ \|x\|=1}} \|Ax\|_p = \sup_{\substack{y \in \ell_q^k \\ \|y\|=1}} \sup_{\substack{x \in \ell_{p_1}^d \\ \|x\|=1}} y^T A x$$

In what follows we use d to denote the original dimension and $k < d$ the target (reduced) dimension. The input vector will be $x = (x_1, \ldots, x_d) \in \ell_2^d$. Since we only consider linear reductions we will assume without loss of generality that $\|x\|_2 = 1$.

Definition 2.1 *A family of distributions $\mathcal{D}(d,k)$ on $k \times d$ real matrices ($k \leq d$) has the Johnson-Lindenstrauss property with respect to a norm index p if for any unit vector $x \in \ell_2^d$ and $0 \leq \varepsilon < 1/2$,*

$$\Pr_{A \in \mathcal{D}_{d,k}} [1 - \varepsilon \leq \|Ax\|_p \leq 1 + \varepsilon] \geq 1 - c_1 e^{-c_2 k \varepsilon^2}$$

for some global $c_1, c_2 > 0$.

Note that a similar definition was given in [22]. Also note that for most dimension reduction applications $k = \Omega(\varepsilon^{-2})$, so the constant c_1 can be "swallowed" in c_2, but we prefer to keep it here for generality. In this work, we study the cases $p = 1$ (the *Manhattan* case) and $p = 2$ (the *Euclidean* case).

Recall that the Walsh-Hadamard matrix H_d is a $d \times d$ orthogonal matrix with $H_d(i,j) = 2^{-d/2}(-1)^{\langle i,j \rangle}$ for all $i, j \in [0, d-1]$, where $\langle i,j \rangle$ is dot product (over $\mathbb{F}_2$) of i, j viewed as $(\log d)$-bit vectors. The matrix encodes the Fourier transform over the binary hypercube. It is well known that $x \mapsto H_d x \in \ell_2^d$ can be computed in time $O(d \log d)$ for any $x \in \ell_2^d$, and that the mapping is isomorphic.

Definition 2.2 *A matrix $A \in \mathbf{R}^{m \times d}$ is a* code *matrix if every row of A is equal to some row of H multiplied by $\sqrt{d/m}$.*

The normalization is chosen so that columns have Euclidean norm 1.

We are now ready to state our theorems. The main contribution is in Theorem 2.4.

Theorem 2.3 *For any code matrix A of size $k \times d$ for $k < d$, the mapping $x \mapsto Ax$ can be computed in time $O(d \log k)$.*

Clearly this theorem is interesting only for $k = d^{o(1)}$, because otherwise the Walsh-Hadamard transform followed by projection onto a subset of the coordinates can do this in time $O(d \log d)$, by definition of a code matrix. As a simple corollary, the running time of the algorithms in [1] can be reduced to $O(d \log k)$ when $k < d^{1/4}$, because effectively what they do is multiply the input x (after random sign change) by a code matrix of size $O(k^3) \times d$ and then manipulate the outcome in time $O(k^3)$. We omit the details of this result and refer the reader to [1, 20].

Theorem 2.4 *Let $\delta > 0$ be some arbitrarily small constant. For any d, k satisfying $k \leq d^{1/2-\delta}$ there exists an algorithm constructing a random matrix A of size $k \times d$ with the JL property, such that the time to compute $x \mapsto Ax$ for any $x \in \mathbf{R}^d$ is $O(d \log d)$. The construction uses $O(d)$ random bits and applies to both the Euclidean and the Manhattan cases.*

Note that the running time in Theorem 2.4 can be reduced to $O(d \log k)$ (which is intersting only for $k = d^{o(1)}$). We will sketch how to do this but will omit the details because such a result is obtained from Theorem 2.3 in conjunction with [1].

3 Tools from Banach Spaces

The following is known as an interpolation theorem in the theory of Banach spaces. For a proof, refer to [5].

Theorem 3.1 Riesz-Thorin *Let A be an $m \times d$ real matrix, and assume $\|A\|_{p_1 \to r_1} \leq C_1$ and $\|A\|_{p_2 \to r_2} \leq C_2$ for some norm indices p_1, r_1, p_2, 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 $\|A\|_{p \to r} \leq C_1^{\lambda} C_2^{1-\lambda}$.*

Theorem 3.2 [Hausdorff-Young] *For norm index $1 \leq p \leq 2, \|H\|_{p \to q} \leq d^{-1/p+1/2}$, where q is the dual norm index of p.*

(The theorem is usually stated with respect to the Fourier operator for functions on the real line or on the circle, and is a simple application of Riesz-Thorin by noticing that $\|H\|_{2\to 2} = 1$ and $\|H\|_{1 \to \infty} = d^{-1/2}$.)

Let M be a real matrix $m \times d$ matrix, and let $z \in \mathbf{R}^d$ be a random vector with each z_i distributed uniformly and independently over $\{\pm 1\}$. The random vector $Mz \in \ell_p^m$ is known as a *Rademacher* random variable. A nice exposition of concentration bounds for Rademacher variables is provided in Chapter 4.7 of [17] for more general Banach spaces. For our purposes, it suffices to review the result for finite dimensional ℓ_p space. Consider the norm $Z = \|Mz\|_p$. We associate two numbers with Z,

- The deviation σ, defined as $\|M\|_{2 \to p}$, and
- a median μ of Z.

Theorem 3.3 *For any $t \geq 0$,*

$$\Pr[|Z - \mu| > t] \leq 4e^{-t^2/(8\sigma^2)} \ .$$

The theorem is a simple consequence of a theorem of Talagrand (Chapter 1, [17]) on measure concentration of functions on $\{-1,+1\}^d$ which can be extended to convex functions on ℓ_2^d with small Lipschitz norm.

4 Tools from Error Correcting Codes

Let A be a code matrix, as defined above. The columns of A can be viewed as vectors over $\mathbb{F}_2$ under the usual transformation $((+) \to 0, (-) \to 1)$. Clearly, the set of vectors thus obtained are closed under addition, and hence constitute a linear subspace of $\mathbb{F}_2^m$. Conversely, any linear subspace V of $\mathbb{F}_2^m$ of dimension ν can be encoded as an $m \times 2^\nu$ code matrix (by choosing some ordered basis of V). We will borrow well known constructions of subspaces from coding theory, hence the terminology. Incidentally, note that H_d encodes the Hadamard code, equivalent to a dual BCH code of designed distance 3.

Definition A code matrix A of size $m \times d$ is *a-wise independent* if for each $1 \leq i_1 < i_2 < \ldots < i_a \leq m$ and $[b_1, b_2, \ldots, b_a] \in \{+1, -1\}^a$, the number of columns $A^{(j)}$ for which $[A_{i_1}^{(j)}, A_{i_2}^{(j)}, \ldots, A_{i_a}^{(j)}] = m^{-1/2}[b_1, b_2, \ldots, b_a]$ is exactly $d/2^a$.

Lemma 4.1 *There exists a 4-wise independent code matrix of size $k \times f(k)$, where $f(k) = \Theta(k^2)$.*

The family of matrices is known as binary dual BCH codes of designed distance 5. Details of the construction can be found in [19].

5 Reducing to Euclidean Space for $k < d^{1/2-\delta}$

Let B be a $k \times d$ matrix with Euclidean unit length columns, and D a random $\{\pm 1\}$ diagonal matrix. Let $Y = \|BDx\|_2$. Our goal is to get a concentration bound of Y around 1. Notice that $E[Y^2] = 1$. In order to use Theorem 3.3, we let M denote the $k \times d$ matrix with the i'th column being $x_i B^{(i)}$, where $B^{(i)}$ denotes the i'th column of B. Clearly Y is distributed like the variable Z (using the notation of Section 3) for $p = 2$. We estimate the deviation σ and median μ.

$$\begin{aligned}
\sigma &= \sup_{\substack{y \in \ell_2^k \\ \|y\|=1}} \|y^T M\|_2 \\
&= \sup \left(\sum_{i=1}^{d} x_i^2 (y^T B^{(i)})^2 \right)^{1/2} \\
&\leq \|x\|_4 \sup \left(\sum_{i=1}^{d} (y^T B^{(i)})^4 \right)^{1/4} \\
&= \|x\|_4 \|B^T\|_{2\to 4} \ .
\end{aligned} \tag{1}$$

(The inequality is Cauchy-Schwartz.) Now,

$$E[(Z-\mu)^2] = \int_0^\infty \Pr[(Z-\mu)^2 > s] ds \quad \leq \quad \int_0^\infty 4e^{-s/(8\sigma^2)} ds = 32\sigma^2 \ .$$

The inequality is an application of theorem 3.3. Recall that $E[Z^2] = 1$. Also, $E[Z] = E[\sqrt{Z^2}] \leq \sqrt{E[Z^2]} = 1$ (by Jensen). Hence $E[(Z-\mu)^2] = E[Z^2] - 2\mu E[Z] + \mu^2 \geq 1 - 2\mu + \mu^2 = (1-\mu)^2$. Combining, $|1-\mu| \leq \sqrt{32}\sigma$. We conclude,

Corollary 5.1 *For any $t \geq 0$,*

$$\Pr[|Z-1| > t] \leq c_3 \exp\{-c_4 t^2/(\|x\|_4^2 \|B^T\|_{2\to 4}^2)\} \ ,$$

for some global $c_3, c_4 > 0$.

In order for the distribution of BD to have the Johnson-Lindenstrauss property, we need to have $\sigma = O(k^{-1/2})$. This requires controlling both $\|B^T\|_{2\to 4}$ and $\|x\|_4$. We first show how to design a matrix B that is both efficiently computable and has a small norm. The latter quantity is adversarial and cannot be directly contolled, but we are allowed to manipulate x by applying a (random) orthogonal matrix Φ without losing any information.

5.1 Bounding $\|B^T\|_{2\to 4}$ Using BCH Codes

Lemma 5.2 *Assume B is a $k \times d$ 4-wise independent code matrix. Then $\|B^t\|_{2\to 4} \leq (3dk^{-2})^{1/4}$.*

Proof For $y \in \ell_2^k, \|y\| = 1$,

$$\begin{aligned}
\|y^T B\|_4^4 &= d E_{j\in[d]}[(y^T B(j))^4] \\
&= dk^{-2} \sum_{i_1,i_2,i_3,i_4=1}^{k} E_{b_1,b_2,b_3,b_4}[y_{i_1} y_{i_2} y_{i_3} y_{i_4} b_1 b_2 b_3 b_4] \\
&= dk^{-2}(3\|y\|_2^4 - 2\|y\|_4^4) \leq 3dk^{-2} \ ,
\end{aligned} \tag{2}$$

where b_1, b_2, b_3, b_4 are random $\{+1, -1\}$ variables. We now use the BCH codes. Let B_k denote the $k \times f(k)$ matrix from the Lemma 4.1 (we assume here that $k = 2^a - 1$ for some integer a; This is harmless because otherwise we can reduce onto some $k' = 2^a - 1$ such that $k/2 \leq k' \leq k$ and pad the output with $k - k'$ zeros). In order to construct a matrix B of size $k \times d$ for $k < d^{1/2-\delta}$, we first make sure that d is divisible by $f(k)$ (by at most multiplying d by a constant factor and padding with zeros), and then define B to be $d/f(k)$ copies of B_k side by side. Clearly B remains 4-wise independent. Note that B may no longer be a code matrix, but $x \mapsto Bx$ is computable in time $O(d \log k)$.

5.2 Controlling $\|x\|_4$ for $k < d^{1/2-\delta}$

We define a randomized orthogonal transformation Φ that is computable in $O(d \log d)$ time and succeeds with probability $1 - O(e^{-k})$ for all $k < d^{1/2-\delta}$ for arbitrarily small δ. Success means that $\|\Phi x\|_4 = O(d^{-1/4})$. (Note: Both big-$O$'s hide factors depending on δ). Note that this construction gives a running time of $O(d \log d)$, which is $O(d \log k)$ by our assumption of $k \geq d^{1/4}$. We briefly show later how to do this for any small k with running time $O(d \log k)$.

The basic building block is the product HD', where $H = H_d$ is the Walsh-Hadamard matrix and D' is a diagonal matrix with random i.i.d. uniform $\{\pm 1\}$ on the diagonal. Note that this random transormation was the main ingredient in [1]. Let $H^{(i)}$ denote the i'th column of H.

We are interested in the random variable $X = \|HD'x\|_4$. Using the notation of Section 3, we define M as the $d \times d$ matrix with the i'th column $x_i H^{(i)}$, we let $p = 4$ ($q = 4/3$), and consider the corresponding variable Z from Section 3, which is distributed like X. By the definition of the deviation σ,

$$\begin{aligned}
\sigma &= \|M\|_{2\to 4} = \|M^T\|_{4/3\to 2} \\
&= \sup_{\substack{y \in \ell_{4/3}^k \\ \|y\|_{4/3}=1}} \left(\sum_i x_i^2 (y^T H^{(i)})^2 \right)^{1/2} \\
&\leq \left(\sum x_i^4 \right)^{1/4} \sup \left(\sum_i (y^T H^{(i)})^4) \right)^{1/4} \\
&= \|x\|_4 \|H\|_{\frac{4}{3}\to 4} \ .
\end{aligned} \tag{3}$$

By the Hausdorff-Young theorem, $\|H\|_{\frac{4}{3}\to 4} \leq d^{-1/4}$. Hence, $\sigma \leq \|x\|_4 d^{-1/4}$. We now get by Theorem 3.3 that for all $t \geq 0$,

$$\Pr[|\|HD'x\|_4 - \mu| > t] \leq 4e^{-t^2/(8\|x\|_4^2 d^{-1/2})} \ , \tag{4}$$

where μ is a median of Z.

Claim 5.3 $\mu = O(d^{-1/4})$.

Proof To see the claim, notice that for each separate coordinate $E[(HD'x)_i^4] = O(d^{-2})$ and then use linearity of expectation to get $E[\|HD'x\|_4^4] = O(d^{-1})$. By Jensen inequality, $E[\|HD'x\|_4^b] \leq$

$E[\|HD'x\|_4^4]^{b/4} = O(d^{-b/4})$ for $b = 1, 2, 3$. Now

$$E[(\|HD'x\|_4 - \mu)^4] = \int_0^\infty \Pr[(\|HD'x\|_4 - \mu)^4 > s]ds \leq \int_0^\infty 4e^{-s^{1/2}/(8\|x\|_4^2 d^{-1/2})}ds$$
$$= O(d^{-1}) \ .$$

This implies that $\gamma_1 d^{-1} - \gamma_2 d^{-3/4}\mu + \gamma_3 d^{-2/4}\mu^2 - \gamma_4 d^{-1/4}\mu^3 + \mu^4 \leq \gamma_5 d^{-1}$, where $\gamma_i = O(1)$ for $i = 1, 2, 3, 4, 5$. Clearly this implies the statement of the claim.

Let c_9 be such that $\mu_4 \leq c_9 d^{-1/4}$. We weaken inequality (4) using the last claim to obtain the following convenient form:

$$\Pr[\|HD'x\|_4 > c_9 d^{-1/4} + t] \leq 4e^{-t^2/(8\|x\|_4^2 d^{-1/2})} \ . \tag{5}$$

In order to get a desired failure probability of $O(e^{-k})$ set $t = c_8 k^{1/2}\|x\|_4 d^{-1/4}$. For $k < d^{1/2-\delta}$ this gives $t < c_8 d^{-\delta/2}\|x\|_4$. In other words, with probability $1 - O(e^{-k})$ we get

$$\|HD'x\|_4 \leq c_9 d^{-1/4} + c_8 d^{-\delta/2}\|x\|_4 \ .$$

Now compose this r times: Take independent random diagonal $\{\pm 1\}$ matrices $D' = D^{(1)}, D^{(2)}, \ldots, D^{(r)}$ and define $\Phi_d^{(r)} = HD^{(r)}HD^{(r-1)} \cdots HD^{(1)}$. Using a union bound on the conditional failure probabilities, we easily get:

Lemma 5.4 [ℓ_4 **reduction for** $k < d^{1/2-\delta}$] *With probability* $1 - O(e^{-k})$

$$\|\Phi^{(r)}x\|_4 = O(d^{-1/4})$$

for $r = \lceil 1/2\delta \rceil$.

Note that the constant hiding in the ℓ_4 bound is exponential in $1/\delta$.

Combining the above, the random transformation $A = BD\Phi^{(r)}$ has the JL Euclidean property for $k < d^{1/2-\delta}$, and can be applied to a vector in time $O(d \log d)$. This proves the Euclidean case of Theorem 2.4.

5.3 Reducing the Running Time to $O(d \log k)$

Although the case $k < d^{1/4}$ is taken care of by Theorem 2.3, we briefly show how to do this using the new tools. Recall that in the construction of B we placed $d/f(k)$ copies of the same code B_k side by side. It turns out that we can apply this "decomposition" of coordinates to $\Phi^{(r)}$. Indeed, let $I_j \subseteq [d]$ denote the j'th block of $f(k)$ consecutive coordinates. For a vector $y \in \ell_p^d$, let $y_{I_j} \in \ell_p^{f(k)}$ denote the projection of y onto the set of coordinates I_j. Now, instead of using $\Phi^{(r)} = \Phi_d^{(r)}$ as above, we use a block-diagonal $d \times d$ matrix comprised of $d/f(k)$ $f(k) \times f(k)$ blocks drawn from the same distribution as $\Phi_{f(k)}^{(r)}$. Clearly the running time of the block-diagonal matrix is $O(d \log k)$ (by applying the Walsh transform independently on each block).

In order to see why this still works, one needs to repeat the above proofs with the norm $\|\cdot\|_{(4,2)}$ defined as $\|x\|_{(4,2)} = \left(\sum_{j=1}^{d/f(k)} \|x_{I_j}\|_4^2\right)^{1/2}$. We omit the details.

6 Reducing to Manhattan Space for $k < d^{1/2-\delta}$

We sketch this simpler case. As we did for the Euclidean case, we start by studying the random variable $W \in \ell_1^k$ defined as $W = \|BDx\|_1$ for a $k \times d$ real matrix B and a random $\{\pm k^{-1/2}\}$-diagonal matrix. Setting M to be the $k \times d$ matrix with the i'th column being $x_i B^{(i)}$ and $p = 1$, we see using the notation in Section 3 that W is distributed like Z. We compute σ and μ.

$$\begin{aligned} \sigma = \sup_{\substack{y \in \ell_\infty^k \\ \|y\|=1}} \|y^T M\|_2 &= \sup \left(\sum_{i=1}^{d} x_i^2 (y^T B^{(i)})^2 \right)^{1/2} \\ &\leq \sup \|x\|_4 \|y^T B^{(i)}\|_4 = \|x\|_4 \|B^T\|_{\infty \to 4} \end{aligned} \tag{6}$$

Using the the tools developed in the Euclidean case, we can reduce $\|x\|_4$ to $O(d^{-1/4})$ with probability $1 - O(e^{-k})$ using $\Phi_r(d)$, in time $O(d \log d)$. Also we already know that $\|B^T\|_{2\to 4} = O(d^{1/4} k^{-1/2})$ if B is comprised of $k \times f(k)$ dual BCH codes (of designed distance 5) matrices side by side (assume $f(k)$ divides d). Since $\|y\|_\infty \geq k^{-1/2}\|y\|_2$, we conclude that $\|B^T\|_{\infty \to 4} = O(d^{1/4})$. Combining, we get $\sigma = O(1)$.

We now estimate the median μ of Z. Let P be any coordinate of BDx. Clearly $E[P^2] = 1/k$. Note that P is a Rademacher random variable in one dimension with $\sigma' = k^{-1/2}$. Denote the median of $|P|$ by μ'. Then by Theorem 3.3, $\Pr[||P| - t| > \mu'] \leq 4e^{-t^2 k/8}$. By a simple integral evaluating $E[P^2]$ we can solve for $\mu = k^{-1/2}(1 \pm O(k^{-1/2}))$. Then by another integral we can evaluate $E[|P|] = k^{-1/2}(1 \pm O(k^{-1/2}))$. By linearity of expectation, $E[Z] = k^{1/2}(1 \pm O(k^{-1/2}))$. Again by integration we get $\mu = k^{1/2}(1 \pm O(k^{-1/2}))$. We conclude that

$$\Pr[|Z - k^{1/2}| > t] \leq c_{10} e^{-c_{11} t^2} \ ,$$

for some global $c_{10}, c_{11} > 0$, and therefore the (normalized) $k^{-1/2} BD\Phi^{(r)} x$ has the Johnson Lindenstrauss property for the Manhattan case for $k < d^{1/2-\delta}$ and $r = O(1/\delta)$ (proving the Manhattan part of Theorem 2.4).
Reducing the running time from $O(d \log d)$ to $O(d \log k)$ is done similarly to the Euclidean case.

7 Trimmed Walsh-Hadamard transform

We prove Theorem 2.3. For simplicity, let $H = H_d$. It is well known that computing the Walsh-Hadamard transform $H\mathbf{x}$ requires $O(d \log d)$ operations. It turns out that it is possible to compute $PH\mathbf{x}$ with $O(d \log k)$ operation, as long as the matrix P contains at most k nonzeros. This will imply Theorem 2.3, because code matrices of size $k \times d$ are a product of PH_d, where P contains k rows with exactly one nonzero in each row. To see this we remind the reader sthat the Walsh-Hadamard matrix (up to normalization) can be recursively described as

$$H_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{7}$$

$$H_q = H_1 \otimes H_{q/2} \tag{8}$$

Where $\otimes$ is the Kronecker product.

We define $\mathbf{x}_1$ and $\mathbf{x}_2$ to be the first and second halves of $\mathbf{x}$. Similarly, we define P_1 and P_2 as the left and right halves of P respectively.

$$PH_q\mathbf{x} = \begin{pmatrix} P_1 & P_2 \end{pmatrix} \begin{pmatrix} H_{q/2} & H_{q/2} \\ H_{q/2} & -H_{q/2} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = P_1 H_{q/2}(\mathbf{x}_1 + \mathbf{x}_2) + P_2 H_{q/2}(\mathbf{x}_1 - \mathbf{x}_2) \quad (9)$$

P_1 and P_2 contain k_1 and k_2 nonzeros respectively, $k_1 + k_2 = k$, giving the recurrence relation $T(d,k) = T(d/2,k_1) + T(d/2,k_2) + d$ for the running time. The base cases are $T(d,0) = 0$ and $T(d,1) = d$. We use induction to show that $T(d,k) \leq 2d\log(k+1)$.

$$\begin{aligned} T(d,k) &= T(d/2,k_1) + T(d/2,k_2) + d \\ &\leq d\log(2(k_1+1)(k_2+1)) \\ &\leq d\log((k_1+k_2+1)^2) \text{ for any } \quad k_1+k_2 = k \geq 1 \\ &\leq 2d\log(k+1) \end{aligned}$$

The last sequence of inequalities together with the base cases clearly also give an algorithm and prove Theorem 2.3.

Since in [1] both Hadamard and Fourier transforms were considered we shortly describe also a simple trimmed fourier transform. In order to compute k coefficients from a d dimensional fourier transform on a vector $\mathbf{x}$, we divide $\mathbf{x}$ into L blocks of size d/L and begin with the first step of the cooley tukey algorithm which performs d/L FFT's of size L between the blocks (and multiplies them by twiddle factors). In the second step, instead of computing FFT's inside each block, each coefficient is computed directly, by summation, inside it's block. These two steps require $(d/L) \cdot L\log(L)$ and kd/L operations respectively. By choosing $k/\log(k) \leq L \leq k$ we achieve a running time of $O(d\log(k))$.

8 Future work

- *L*ower bounds. A lower on the running time of applying a random matrix with a JL property on a vector will be extremely interesting. Any nontrivial (superlinear) bound for the case $k = d^{\Omega(1)}$ will imply a lower bound on the time to compute the Fourier transform, because the bottleneck of our constructions is a Fourier transform.
- *G*oing beyond $k = d^{1/2-\delta}$. As part of our work in progress, we are trying to push the result to higher values of the target dimension k (the goal is a running time of $O(d\log d)$). We conjecture that this is possible for $k = d^{1-\delta}$, and have partial results in this direction. A more ambitious goal is $k = \Omega(d)$.

References

[1] 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.

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

[3] R. I. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *FOCS '99: Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, page 616, Washington, DC, USA, 1999. IEEE Computer Society.

[4] S. Artstein-Avidan and V. Milman. Logarithmic reduction of the level of randomness in some probabilistic geometric constructions. *SIAM Journal on Computing*, 1(34):67–88, 2004.

[5] J. Bergh and J. Lofstrom. *Interpolation Spaces.* Springer-Verlag, 1976.

[6] 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.

[7] P. Drineas, R. Kannan, and M. Mahoney. Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix, 2004.

[8] P. Drineas, R. Kannan, and M. Mahoney. Fast monte carlo algorithms for matrices iii: Computing a compressed approximate matrix decomposition, 2004.

[9] 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.

[10] 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.

[11] 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.

[12] P. Indyk. On approximate nearest neighbors in non-Euclidean spaces. In *Proceedings of the 39th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 148–155, 1998.

[13] P. Indyk. Uncertainty principles, extractors, and explicit embeddings of l2 into l1. In *Proceedings of the 39th Annual ACM Symposium on the Theory of Computing*, 2007.

[14] 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.

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

[16] E. Kushilevitz, R. Ostrovsky, and Y. Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM Journal on Computing*, 30(2):457–474, 2000.

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

[18] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988.

[19] F. MacWilliams and N. Sloane. *The Theory of Error Correcting Codes*. North-Holland, 1983.

[20] J. Matousek. On variants of the Johnson-Lindenstrauss lemma. *Private communication*, 2006.

[21] A. A. Razborov. Expander codes and somewhat Euclidean sections in ℓ_1^n. *ECCC*, 2007.

[22] 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.

[23] S. Vempala. *The Random Projection Method*. DIMACS Series in Discrete Mathematics and Theoretical Computer Science. 2004.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de/