FAST AND EFFICIENT DIMENSIONALITY REDUCTION USING STRUCTURALLY RANDOM MATRICES

Thong T. Do[†]*, Lu Gan*[‡]*, Yi Chen*[†]*, Nam Nguyen*[†] *and Trac D. Tran*[†] *

[†] Department of Electrical and Computer Engineering
The Johns Hopkins University
[‡]School of Engineering and Design
Brunel University, UK

ABSTRACT

Structurally Random Matrices (SRM) are first proposed in [1] as fast and highly efficient measurement operators for large scale compressed sensing applications. Motivated by the bridge between compressed sensing and the Johnson-Lindenstrauss lemma [2] , this paper introduces a related application of SRMs regarding to realizing a fast and highly efficient embedding. In particular, it shows that a SRM is also a promising dimensionality reduction transform that preserves all pairwise distances of high dimensional vectors within an arbitrarily small factor ϵ, provided that the projection dimension is on the order of $\mathcal{O}(\epsilon^{-2}\log^3 N)$, where N denotes the number of d-dimensional vectors. In other words, SRM can be viewed as the sub-optimal Johnson-Lindenstrauss embedding that, however, owns very low computational complexity $\mathcal{O}(d\log d)$ and highly efficient implementation that uses only $\mathcal{O}(d)$ random bits, making it a promising candidate for practical, large scale applications where efficiency and speed of computation are highly critical.

Index Terms— Low-distortion embedding, Johnson-Lindenstrauss, dimensionality reduction, compressed sensing, machine learning.

1. INTRODUCTION

According to the theory of compressed sensing [3], a K-sparse signal $\mathbf{x}$ of length d which has at most K nonzero coefficients under some linear transform, can be exactly reconstructed from its random projection of a much lower dimension $\mathcal{O}(K\log d)$:

$$\mathbf{y} = \mathbf{\Phi x}$$

where $\mathbf{\Phi}$ is a random projection or a random matrix of subGaussian i.i.d entries. When an input signal $\mathbf{x}$ is large, for example, a (vectorized) megapixel image, using a random projection is clearly impractical as huge amount of computational complexity and memory buffering are needed to compute the projection $\mathbf{y}$.

Recently, Structurally Random Matrices (SRM) have been proposed [1] as a fast and highly efficient compressed sensing method that somewhat surprisingly guarantees optimal performance. A structurally random matrix $\mathbf{\Phi}$ (using the local randomizer) is a product of three matrices:

$$\mathbf{\Phi} = \sqrt{\frac{d}{M}}\mathbf{DFR} \tag{1}$$

where

- $\mathbf{R}$, the local randomizer, is a $d \times d$ random diagonal matrix whose diagonal entries R_{ii} are i.i.d Bernoulli random variables $\mathrm{P}(R_{ii} = \pm 1) = \frac{1}{2}$.
- $\mathbf{F}$ is a $d \times d$ orthonormal matrix whose absolute magnitude of all entries are on the order of $\mathcal{O}(\frac{1}{\sqrt{d}})$. In practice, only $\mathbf{F}$ with fast computation and efficient implementation such as the FFT, the DCT and the WHT... are chosen. Finally,
- $\mathbf{D}$, the uniformly random downsampler, is a matrix composed of nonzero rows of a random diagonal matrix whose diagonal entries D_{ii} are i.i.d binary random variables with $\mathrm{P}(D_{ii} = 1) = \frac{M}{d}$. On average, $\mathbf{D}$ contains M nonzero rows and thus, $\mathbf{\Phi}$ is a $M \times d$ matrix.

Algorithmically, the projection $\mathbf{y}$ can be acquired efficiently as follows: *(i)* pre-randomizing $\mathbf{x}$ randomly flipping sign of entries of $\mathbf{x}$, *(ii)* applying some fast transform to the randomized $\mathbf{x}$ and *(iii)* finally, randomly keeping M those transform coefficients.

According to the classical Johnson and Lindenstrauss (JL) lemma [4], any set of N vectors in d-dimensional Euclidean space can be embedded into $M = \mathcal{O}(\epsilon^{-2}\log N)$- dimensional Euclidean space so that all pairwise distances are preserved within an arbitrarily small factor ϵ. In other words, there exists an embedding $\mathbf{A} : \mathbf{R}^d \rightarrow \mathbf{R}^M$ such that for all pair of vectors $\mathbf{u}$ and $\mathbf{v}$ in the set of N vectors in $\mathbf{R}^d$:

$$(1-\epsilon)\|\mathbf{u}-\mathbf{v}\|^2 \leq \|\mathbf{A}(\mathbf{u}) - \mathbf{A}(\mathbf{v})\|^2 \leq (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|^2. \tag{2}$$

Such an embedding will be referred as the JL-embedding or JL-transform if $M = \mathcal{O}(\epsilon^{-2}\log N)$ and suboptimal JL-transform if $M > \mathcal{O}(\epsilon^{-2}\log N)$.

In a recent inspiring paper [2], R. Baraniuk *et al.* shows an interesting connection between compressed sensing and the JL-lemma that any distribution that yields a satisfactory JL-transform will also generate measurement ensemble satisfying Restricted Isometry Property (RIP), a sufficient condition for being the optimal measurement ensemble.

In this paper, we explore the converse relationship that SRM, the optimal measurement ensemble, is also a promising candidate of a low distortion embedding. Compared with other existing state-of-the-art JL transforms, SRM can be viewed as one of the fastest and most efficient (suboptimal) JL-transforms. Although it can only guarantee a weaker theoretical bound of dimensionality reduction $\mathcal{O}(\epsilon^{-2}\log^3 N)$, it can be easily implemented as serial operators without a need of explicitly storing the transform in memory, a unique feature that might not be available with other existing JL-transforms.

*This work has been supported in part by the National Science Foundation under Grant CCF-0728893.

2. BACKGROUND

Previously, almost all JL-transforms proposed are random matrices of i.i.d entries of some distribution such as Gaussian or Bernoulli. Since a major application of the JL lemma is in large database systems, fast and efficient implementation of JL-transform is highly critical. With a dense and random matrix, the computational complexity and the memory buffering requirement are about $\mathcal{O}(d\epsilon^{-2}\log N)$, which is expensive when ϵ is small as d is often very large (or otherwise there is no need of dimensionality reduction). One obvious solution to speed up the projection process is to use a sparse random matrix. D. Achlioptas first proposed a sparse random matrix $\mathbf{A}$ whose entries A_{ij} are i.i.d random variables with the following sparse distribution [5]:

$$A_{ij} = \begin{cases} \sqrt{3} & \text{with probability} \frac{1}{6} \\ 0 & \text{with probability} \frac{2}{3} \\ -\sqrt{3} & \text{with probability} \frac{1}{6} \end{cases} \quad (3)$$

As the number of nonzero entries of the matrix $\mathbf{A}$ is, on average, 3 times less than a dense random matrix, the speed of this sparse projection is 3 times faster than that of a dense random matrix. It is aslo shown that the matrix can not be futher sparse without incurring a penalty in the dimensionality. To speed up the projection computation process more than a constant times, in [6] N. Ailon *et al.* proposed a scheme of Fast JL-Transform (FJLT). FJLT is slightly reminiscent to our SRM as it is also a product of three matrices: $\mathbf{PFR}$, where $\mathbf{F}$ and $\mathbf{R}$ are similar to those in the SRM and $\mathbf{P}$ is the random matrix of i.i.d entries of some *sparse* distribution:

$$P_{ij} \begin{cases} \mathcal{N}(0, q^{-1}) & \text{with probability} q \\ 0 & \text{with probability} 1-q \end{cases}$$

where $q = \mathcal{O}(\frac{\log^2 N}{d})$.

Roughly speaking, since the average number of nonzero entries of the matrix $\mathbf{P}$ is just $\mathcal{O}(\log^2 N)$, FJLT is a fast scheme because there is a significant reduction of the amount of computation of $\mathbf{P}$. In [7], J. Matousek shown that it is possible to replace the Gaussian distribution $\mathcal{N}(0, q^-1)$ by Bernoulli (± 1) distribution without incurring the dimensionality penalty, further speeding up the computation. Then, in [8], D. Ailon *et al*. showed a simpler variant of FJLT by replacing a sparse random matrix $\mathbf{P}$ by a deterministic 4-wise independent code matrix (e.g. BCH codes). More recently, E. Liberty introdues the Lean Walsh transform [9]. Although all these fast transforms keep the optimality of dimension reduction, $\mathcal{O}(\epsilon^{-2}\log N)$, they all require some restriction of input vectors $\mathbf{u}$. For example, the FJLT requires $\|\mathbf{u}\|_\infty \leq \mathcal{O}((d/M)^{-1/2})$ while the BCH-code algorithm requires $\|\mathbf{u}\|_4 \leq \mathcal{O}(d^{-1/4})$. In other words, these transforms only keep the optimality of dimension reduction for a certain subset of vectors $\mathbf{u} \in \mathcal{R}^d$.

The computational complexity of FJLT's is roughly $\mathcal{O}(d\log d + \epsilon^{-2}\log^3 N)$, which is much smaller than that of a dense random embedding $\mathcal{O}(d\epsilon^{-2}\log N)$ when ϵ is relatively small. Unfortunately, although $\mathbf{P}$ is a very sparse matrix, its entries are still completely random and thus, a certain amount of space might be required to store $\mathbf{P}$ explicitly. The main idea of FJLT is that it uses a fast computatable matrix $\mathbf{FR}$ to precondition the signal before applying a very sparse matrix because in general, a sparse matrix will significantly distort a sparse signal.

Despite its independent development, our proposed SRM shares the similarly core principle in the compressed sensing field. It uses a preprocessing operator $\mathbf{FR}$ to distribute the information of input signal over all measurements, enabling us to recover the signal from a subset of measurements. It is further shown in this paper that a uniformly random subset of those measurements preserves pairwise distances of original vectors. This principle comes from the fact that energy of a randomized signal is spread out in some fixed linear transform that is illustated in Fig. 1. DCT coefficients of a randomized 512×512 Lena image (its mean is subtracted before randomization) are almost nonzero and more uniformly distributed than those of the original image, implying that energy of the randomized signal is more equally distributed among those coefficients.

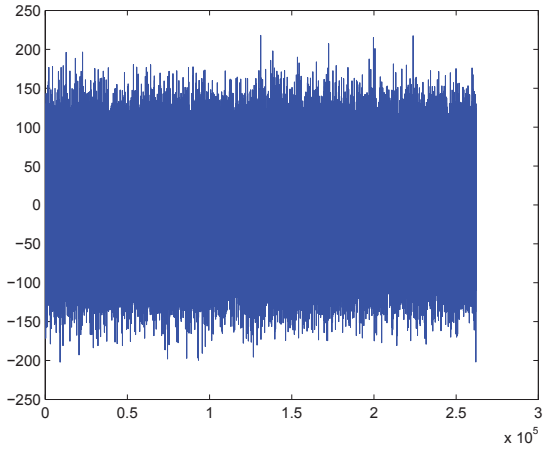

Fig. 1. 1D-DCT coefficients of a randomized zero-mean 512×512 Lena image. Randomization is done by randomly flipping sign of pixels of the image

.

Hence, rather than projecting transform coefficients onto some sparse random basis as in FJLT, SRM directly samples them in a uniformly random fashion that significantly simplify the projection process. Due to this simplification, SRM incurs some penalty in the level of dimensionality reduction, $\mathcal{O}(\epsilon^{-2}\log^3 N)$. However, this is only a theoretical bound for the worst case analysis, i.e. there is no restriction of input vectors $\mathbf{u}$. As one can clearly see in the numberical experiments in the next section, the difference of performance between SRM and completely random projection is hardly observable. In practice, SRM is more appropriate for large scale applcications that favor simple, efficient implementation and fast computation.

3. THEORETICAL ANALYSIS

Theorem 3.1. *Let $\mathcal{P}$ be an arbitrary set of N points in $\mathcal{R}^d$ and suppose that $N \geq d$. Define $\Psi = \frac{d}{M}\mathbf{DFR}$ as an $M \times d$ SRM, where $\mathbf{D}$ is the uniformly random downsampler, $\mathbf{F}$ is an orthonormal matrix with all absolute magnitude of entries on the order of $\mathcal{O}(\frac{1}{\sqrt{d}})$ and $\mathbf{R}$ is the local random randomizer as described in Section I. When $M = \mathcal{O}(\epsilon^{-2}\log^3 N)$, with probability at least $1 - \frac{1}{N}$, for all $\mathbf{u}, \mathbf{v} \in \mathcal{P}$*

$$(1-\epsilon)\|\mathbf{u}-\mathbf{v}\|^2 \leq \|\mathbf{\Phi u} - \mathbf{\Phi v}\|^2 \leq (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|^2 \quad (4)$$

Proof. The proof uses nothing more than the Hoeffding's and Bernstein's inequalities [10] of concentration. Denote $\mathbf{w} = \mathbf{u} - \mathbf{v}$. To show (4), it is sufficient to show that with probability at least $1 - \frac{1}{N}$:

$$(1-\epsilon)\|\mathbf{w}\|^2 \leq \|\mathbf{\Phi w}\|^2 \leq (1+\epsilon)\|\mathbf{w}\|^2 \quad (5)$$

for all $\binom{N}{2}$ possible values of $\mathbf{w}$. Without loss of generality, we can assume that $\|\mathbf{w}\|_2 = 1$. As shown at the following proposition, $E\{\|\mathbf{\Phi w}\|^2\} = 1$ and thus, (5) implies that $\|\mathbf{\Phi w}\|^2$ is concentrated around its expected value

Proposition 3.1. *With a vector* $\mathbf{w} \in \mathcal{R}^d$ *and* $\|\mathbf{w}\| = 1$*, denote* $\mathbf{y} = \mathbf{FRw}$ *and* $K = \max_{1\le i\le d} |y_i|^2$*. Then,*

$$P(|\|\mathbf{\Phi w}\|^2 - 1| \ge \epsilon) \le 2\exp(\frac{-\epsilon^2}{2(\frac{d^2K^2}{M} + \frac{\epsilon dK}{3M})}) \tag{6}$$

Proof. This is a simple corollary of the classical Bernstein's inequality of concentration. First, notice that $\|\mathbf{Dy}\|^2$ can be rewritten as the following:

$$\|\mathbf{\Phi w}\|^2 = \frac{d}{M}\|\mathbf{Dy}\|^2 = \frac{d}{M}\sum_{i=1}^{d}\rho_i y_i^2$$

where ρ_i are i.i.d binary random variables $P(\rho_i = 1) = \frac{M}{d}$ and $E\{\rho_i\} = \frac{M}{d}$. Due to the orthonormality of $\mathbf{FR}$:

$$\|\mathbf{y}\|^2 = \|\mathbf{FRw}\|^2 = \|\mathbf{w}\|^2 = 1$$

and thus,

$$\|\mathbf{\Phi w}\|^2 - 1 = \frac{d}{M}\sum_{i=1}^{d}(\rho_i - \frac{M}{d})y_i^2. \tag{7}$$

Notice that the right side of (7) is a sum of zero-mean independent random variables $E\{\|\mathbf{\Phi w}\|^2 - 1\} = 0$ and

$$\sigma^2 = \text{Var}\{\|\mathbf{\Phi w}\|^2 - 1\} = \frac{d}{M}(1 - \frac{M}{d})\sum_{i=1}^{d} y_i^4$$

, where the last equality is due to $\text{Var}\{\rho_i - \frac{M}{d}\} = \frac{M}{d}(1 - \frac{M}{d})$.

Also, it is easy to verify that $\sigma^2 \le \frac{d^2K^2}{M}$ and that $|(\rho_i - \frac{M}{d})y_i^2| \le \max_{1\le i\le d} y_i^2 = K$ for all $i \in \{1, 2, \ldots, d\}$. Applying the classical Bernstein's inequality of concentration for a sum of zero-mean independent random variables [10], we derive (6). □

The next proposition bounds the value of K:

Proposition 3.2. *With a vector* $\mathbf{w} \in \mathcal{R}^d$ *and* $\|\mathbf{w}\| = 1$*, denote* $\mathbf{y} = \mathbf{FRw}$*. Let c be a positive constant such that* $\max_{1\le i,j\le d} |F_{ij}| = \sqrt{\frac{c}{d}}$*. Then,*

$$P\{\max_{1\le i\le d} |\mathbf{y}_i| \ge \sqrt{\frac{2c\log(2d/\alpha)}{d}}\} \le \alpha \tag{8}$$

Proof. This is a simple corollary of the classical Hoeffding's inequality of concentration. Let F_{ik} be the k^{th} entry on the i^{th} row of the matrix $\mathbf{F}$ and R_{kk} be the k^{th} entry on the main diagonal of the diagonal matrix $\mathbf{R}$,

$$\mathbf{y}_i = \sum_{k=1}^{d} R_{kk}F_{ik}w_k = \sum_{k=1}^{d} Z_k$$

where $Z_k = R_{kk}F_{ik}w_k$ are zero-mean independent random variables, $Z_k = \pm F_{ik}w_k$. It is easy to verify $E\{y_i\} = 0$ because of $E\{Z_k\} = 0$.

Applying the Hoeffding's inequality of concentration for a sum of independent random variables $\{Z_k\}_{k=1}^{d}$ [10]

$$P(|y_i| \ge t) \le 2\exp(\frac{-t^2}{2\sum_{k=1}^{d} F_{ik}^2 w_k^2})$$

Notice that as $\|w\|_2 = 1$.

$$\sum_{k=1}^{d} F_{ik}^2 w_k^2 \le \max_{1\le k\le d} F_{ik}^2 \sum_{k=1}^{d} w_k^2 = \max_{1\le k\le d} F_{ik}^2 = \frac{c}{d}.$$

Thus,

$$P(|y_i| \ge t) \le 2\exp(\frac{-dt^2}{2c}).$$

Applying the union bound for a supreme of a random sequence

$$P(\max_{i\le i\le d} |y_i| \ge t) \le 2d\exp(\frac{-dt^2}{2c}).$$

Finally, choose $t = \sqrt{\frac{2c\log(2d/\alpha)}{d}}$, we derive the inequality (8). □

Now define the probabilistic event $\mathcal{A} = \{K \le \frac{2c\log(2d/\alpha)}{d}\}$. Conditioning on the event $\mathcal{A}$,

$$\sigma^2|\mathcal{A} \le \frac{2d^2}{M}K^2 \le \frac{(2c\log 2d/\alpha)^2}{M}.$$

For each of $\binom{N}{2}$ possible values of $\mathbf{w}$, let a probabilistic event $\mathcal{B}_\mathbf{w} = \{|\|\mathbf{\Phi w}\|^2 - 1| \ge \epsilon\}$. Also, denote the union probabilistic event $\mathcal{B} = \bigcup_\mathbf{w} \mathcal{B}_\mathbf{w}$. Note that $\mathcal{B}$ is the probabilistic event that $\mathbf{\Phi}$ does not satisfy the inequality (4). Thus, the probability of $\mathbf{\Phi}$ does not satisfy the inequality (4) is:

$$P(\mathcal{B}) \le \binom{N}{2} P(\mathcal{B}_\mathbf{w}) \le \binom{N}{2}(P(\mathcal{B}_\mathbf{w}|\mathcal{A}) + P(\overline{\mathcal{A}})). \tag{9}$$

From Proposition 3.1, we have

$$P(\mathcal{B}_\mathbf{w}|\mathcal{A}) \le 2\exp\{\frac{-\epsilon^2}{2(\frac{(2c\log 2d/\alpha)^2}{M} + \frac{\epsilon 2c\log(2d/\alpha)}{3M})}\}.$$

With $0 < \epsilon, \alpha \le 1$ and $M \le d$, $\frac{(2c\log 2d/\alpha)^2}{M} \ge \frac{\epsilon 2c\log(2d/\alpha)}{3M}$ and thus,

$$P(\mathcal{B}_\mathbf{w}|\mathcal{A}) \le 2\exp\{\frac{-\epsilon^2}{4\frac{(2c\log 2d/\alpha)^2}{M}}\}$$

Notice that from the Proposition 3.2 $P(\overline{\mathcal{A}}) \le \alpha$. Choose $\alpha = \frac{1}{N^3}$, we finally derive:

$$P(\mathcal{B}) \le 2\binom{N}{2}\exp\{\frac{-\epsilon^2}{4\frac{(2c\log 2dN^3)^2}{M}}\} + \frac{1}{2N}. \tag{10}$$

When $M \ge 32c^2\epsilon^{-2}\log N^2(\log 2dN^3)^2 = \mathcal{O}(\epsilon^{-2}\log^3 N)$, the first term in the right-hand side of (10) is less than $\frac{1}{2N}$ and thus the right-hand side is less than $\frac{1}{N}$, which implies that (4) holds with probability at least $1 - \frac{1}{N}$

□

4. SIMULATION RESULTS

To demonstrate the effectiveness of the SRM in dimensionality reduction, we conduct the experiments as described in [11]. Specifically, the dataset consists of $N = 1000$ image windows of size 50×50 (i.e., the original dimensionality $d = 2500$). These image windows are chosen randomly from thirteen 8-bit grayscale natural images[1]. The reduced dimensionality M ranges from 1 to 800. For each M, we generate a projection matrix $\mathbf{\Phi}$ that maps $\mathcal{R}^d$ to $\mathcal{R}^M$ and randomly select 100 pairs of image windows. Then, for each pair $\mathbf{u}_i$ and $\mathbf{v}_i$, we compute the projected pair $\mathbf{\Phi}\mathbf{u}_i$ and $\mathbf{\Phi}\mathbf{v}_i$. The distortion between the original and projected pairs is measured by the relative difference in the Euclidean distances between the two pairs

$$D_M(i) = \frac{|\|\mathbf{\Phi}\mathbf{u}_i - \mathbf{\Phi}\mathbf{v}_i\|_2 - \|\mathbf{u}_i - \mathbf{v}_i\|_2|}{\|\mathbf{u}_i - \mathbf{v}_i\|_2}$$

The overall distortion at dimensionality M is averaged over the 100 pairs $D(M) = 1/100 \sum_{i=1}^{100} D_M(i)$. In our experiments, we compare the performance of four projection methods. *(i)* Random projection (RP), where the entries of the projection matrix $\mathbf{\Phi}$ are i.i.d. Gaussian random variables. *(ii)* Sparse random projection (SRP), where the entries of $\mathbf{\Phi}$ are drawn according to the distribution in (3). *(iii)* Principal component analysis (PCA), where $\mathbf{\Phi}$ consists of the eigenvectors corresponding to the M largest eigenvalues of the covariance matrix of the dataset. And *(iv)* SRM. The averaged distortion D as a function of the reduced dimensionality M for each of the four projection methods is shown in Fig. 2. As one can clearly see that performances of random methods are almost the same and better than that of PCA for highly reduced dimensionality.

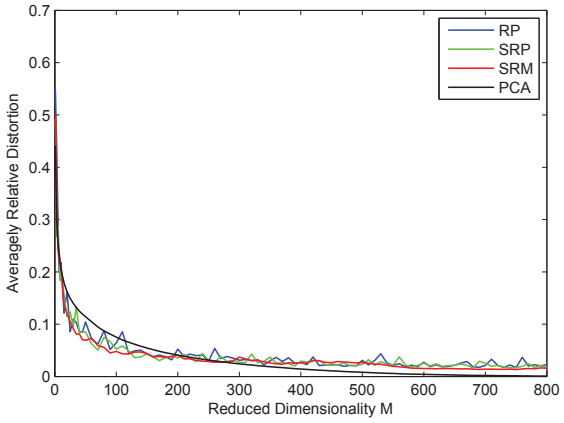

Fig. 2. Average relative distortion caused by methods of dimensionality reduction: RP, SRP, PCA and SRM.

5. CONCLUSIONS

This paper presents a theoretical analysis of SRM (using the local randomizer) as a promising dimensionality reduction transform. This novel transform has very low complexity $\mathcal{O}(d \log d)$ and $\mathcal{O}(d)$ random bits and highly efficient implementation because there is no need to store the transform explicitly in memory. Theoretically, it guarantees reduced dimensionality of $\mathcal{O}(\epsilon^{-2} \log^3 N)$ although simulation results show that its performance is completely comparable to that of a random projection. Prominent applications of SRM include speeding up similarity search algorithms based on low-distortion embedding such as the nearest neighbor approximation [6], realizing low-rank approximation of a large matrix [12], just to name a few. In addition, we also observe that SRM with the global randomizer [1] (to replace the matrix $\mathbf{R}$ by a uniformly random permutation matrix) is also a good candidate of low-distortion embedding. However, its theoretical analysis would be more involved because of the combinatorial nature of the random permutation and thus we will leave it for our future work.

[1] Available at http://www.cis.hut.fi/projects/ica/data/images/

6. REFERENCES

[1] T. T. Do, T. D. Tran, and L. Gan, "Fast compressive sampling with structurally random matrices," *Proceedings of Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, pp. 3369–3372, May 2008.

[2] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *To appear in Constructive Approximation*.

[3] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. on Information Theory*, vol. 52, pp. 489 – 509, Feb. 2006.

[4] W. B Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Conf. in Modern Analysis and Probability*, pp. 189–206, 1984.

[5] D. Achlioptas, "Database-friendly random projection: Johnson-Lindenstrauss with binary coins," *Journal of Computer and System Sciences*, vol. 66, pp. 671–687, 2003.

[6] N. Ailon and B. Chazelle, "Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform," *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, vol. 66, pp. 557 – 563, 2006.

[7] J. Matousek, "On the variants of the Johnson-Lindenstrauss lemma," *Random Structure and Algorithms*, vol. 33, pp. 142–156, 2008.

[8] N. Ailon and E. Liberty, "Fast dimension reduction using rademacher series on dual BCH codes," *To appear in Discrete and Computational Geometry*, 2008.

[9] E. Liberty, N. Ailon, and A. Singer, "Dense fast random projections and Lean Walsh transform," *In RANDOM, Boston, MA*, Aug. 2008.

[10] G. Lugosi, "Concentration-of-measure inequalities," *Lecture notes*, Feb. 2006.

[11] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data,," *In Proceedings of the seventh ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*, 2001.

[12] N. Nguyen, T. T. Do, and T. D. Tran, "A fast and efficient algorithm of low-rank approximation of a matrix,," *submitted to Symposium on Theory of Computing*, 2009.