

FAST DIMENSION REDUCTION THROUGH RANDOM PERMUTATION

Lu Gan[†], Thong T. Do^{*} and Trac D. Tran^{‡*}

[†] School of Engineering and Design, Brunel University, UK

^{*} Thomson Corporate Research, USA

[‡] Dept. of Electrical and Computer Engineering, The Johns Hopkins University, USA

ABSTRACT

This paper studies permutation-based dimension reduction, which can be implemented by first scrambling the input data, then applying the FFT, DCT or Walsh-Hadamard transform and finally using either uniformly random sampling or sparse random projection. By exploiting concentration inequalities of random permutation, we show that this subclass of operators can offer (near) optimal theoretical guarantee. Besides, as random permutation of N elements can be implemented in $\mathcal{O}(N)$ time, the proposed algorithm has very low complexity. Some numerical examples are presented to demonstrate the validity of our theoretical development and their promising applications in image processing.

Index Terms— Random permutation, dimension reduction, structurally random matrix, Johnson–Lindenstrauss lemma, Stein’s method, compressed sensing;

1. INTRODUCTION

Random projection (RP) has been proposed as a simple technique for efficient acquisition and processing of large scale data sets. Simply speaking, the basic idea is to project a high dimensional signal $\mathbf{x} \in \mathcal{R}^N$ to a low dimensional one $\mathbf{y} \in \mathcal{R}^M$ through the following linear transform

$$\mathbf{y} = \Phi \mathbf{x}, \quad (1)$$

where Φ is an $M \times N$ random matrix. So far, the RP methods have led to theoretical breakthroughs in compressed sensing [1, 2], dimension reduction [3] and numerical algebra. Practically, RP methods have found many image/video processing applications such as compressive imaging [4], secure image retrieval [5] and image hashing [6].

In this paper, we focus on the study of RP methods in dimension reduction. Recall that the celebrated Johnson-Lindenstrauss (JL) lemma [3] states that any set of Q points in \mathcal{R}^N can be embedded into $M = \mathcal{O}(\epsilon^{-2} \log Q)$ -dimensional Euclidean space so that all pairwise distances are kept within an arbitrarily small factor ϵ . It is well known that when Φ is a full random matrix with i.i.d Gaussian/Bernoulli variables on its entries, optimal theoretical performance can be achieved. However, for large scale applications, full random matrices have huge memory requirement and high computational cost. To address these issues, various fast JL transforms have been developed. Sparse random projection was proposed by Achlioptas in [7], where the number of non-zero coefficients in Φ is about $\frac{1}{3}$ as that of a full random matrix. However, it was shown that the matrix can not be further sparse without incurring a penalty in dimensionality. Later, N. Ailon *et al.* [8] proposed Fast JL-Transform

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

(FJLT) using randomized FFT followed by a even more sparse random projection. In [9], J. Matousek then proved that by replacing Gaussian distribution with Bernoulli (± 1) distribution, the theoretical performance is still guaranteed. A simpler variant of FJLT was developed in [10] by replacing the sparse random projector with a deterministic 4-wise independent code matrix (e.g. BCH codes). More recently, E. Liberty introduced the Lean Walsh transform [11]. In our previous work [12], we demonstrated that by using uniform random sampling, rather than sparse random projection of FJLT, the computation complexity can be significantly reduced and the number of required projection is increased to $M = \mathcal{O}(\epsilon^{-2} \log^3 Q)$.

Note that most existing fast dimension reduction algorithms rely on pre-processing the input vector through random sign flipping. In this paper, we study fast scrambling-based dimension reduction algorithm, where Φ takes the following form

$$\Phi = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \alpha \mathbf{S} \end{bmatrix} \mathbf{F} \mathbf{P}, \quad (2)$$

in which

- \mathbf{P} is an $N \times N$ uniform random permutation operator;
- \mathbf{F} represents an $N \times N$ fast-computable unitary matrix such as the FFT, the DCT or the Walsh-Hadamard transform (WHT);
- \mathbf{S} is either a *sampling operator* or a *sparse projection matrix*. Specifically, in sampling-based postprocessing, \mathbf{S} selects exactly $M - 1$ rows out of $N - 1$ ones (except the first row) of \mathbf{F} uniformly at random. In projection-based postprocessing, \mathbf{S} takes the same form as that defined in [8], i.e.,

$$\begin{aligned} \mathbf{S}_{ij} &\sim \mathcal{N}(0, 1/q) && \text{with probability } q; \\ \mathbf{S}_{ij} &= 0 && \text{with probability } (1 - q), \end{aligned} \quad (3)$$

where $\mathcal{N}(0, 1/q)$ means a zero-mean normal distribution with variance $1/q$, in which

$$q = \min \left(\mathcal{O} \left(\frac{\log^2 Q}{N} \right), 1 \right),$$

with Q denoting the total number of points.

- α is a scaling factor used to normalize the expected vector norm of each column of Φ . In particular,

$$\alpha = \begin{cases} \sqrt{\frac{N-1}{M-1}}, & \text{Sampling-based postprocessing,} \\ \sqrt{\frac{1}{M-1}}, & \text{Projection-based postprocessing.} \end{cases} \quad (4)$$

Compared with existing work, the main difference is that our algorithm uses a random permutation operator \mathbf{P} for pre-processing, rather than a sign flipping operator [8, 12, 13]. There are several reasons to study permutation-based algorithm. First, as \mathbf{P} can be implemented with only $\mathcal{O}(N)$ time, the algorithm has very low complexity. Besides, analog video scrambling was already implemented in some commercial satellite and cable TV systems [14]. Hence, Φ given in (2) has a hardware friendly structure. Moreover, unlike random sign flipping, the operation of permutation does not change the dynamic range of the input data. Furthermore, the operation of permutation is widely used in encryption and watermarking as it offers a large key space with $N!$ choices. In this light, the proposed dimension reduction method could be combined with these security-related algorithms.

The rest of this paper is organized as follows. In Section 2, we developed new concentration inequality of random permutations using Stein's method of exchangeable pairs. Based on this inequality, we present theoretical analysis of permutation-based dimension reduction in Section 3. Simulation results are shown in Section 4, followed by conclusions and future works in Section 5.

2. CONCENTRATION INEQUALITIES OF RANDOM PERMUTATION

Recall that for FJLT in [8], the analysis relies on concentration inequalities of the sum of a sequence of independent random variables. Although some results of concentration inequalities for random permutations were developed before, they are not sharp enough. Bearing this in mind, we developed a new concentration inequality by exploiting Stein's method of exchangeable pairs [15], as presented in the following theorem:

Theorem 1. Consider two fixed vectors $\mathbf{v} = [v_1, v_2, v_3, \dots, v_N]^T$ and $\mathbf{w} = [w_1, w_2, w_3, \dots, w_N]^T$ in \mathbb{R}^N . Define a random variable U as

$$U = \sum_{i=1}^N v_i w_{\pi(i)},$$

where π is drawn from the uniform distribution over the set of all permutations of $\{1, \dots, N\}$. Then, the expectation of U is

$$\mathbb{E}(U) = \frac{S_v S_w}{N}, \quad (5)$$

in which S_v and S_w represent the sums of \mathbf{v} and \mathbf{w} , respectively, i.e., $S_v = \sum_{i=1}^N v_i$ and $S_w = \sum_{i=1}^N w_i$. Besides, the following inequality holds

$$\mathbb{P}\{|U - \mathbb{E}(U)| \geq t\} \leq 2 \exp\left(-\frac{t^2}{N\zeta^2}\right). \quad (6)$$

where

$$\zeta^2 = 4 \min(|v_{max}|^2 \sigma_w^2, |w_{max}|^2 \sigma_v^2),$$

in which v_{max} and w_{max} represent the maximum magnitude of \mathbf{v} and \mathbf{w} , respectively, i.e.,

$$v_{max} = \max_{1 \leq i \leq N} |v_i| \quad \text{and} \quad w_{max} = \max_{1 \leq i \leq N} |w_i|,$$

while σ_w^2 and σ_v^2 individually represent the variances of \mathbf{v} and \mathbf{w}

$$\sigma_w^2 = \frac{\sum_{i=1}^N w_i^2}{N} - \frac{S_w^2}{N^2} \quad \text{and} \quad \sigma_v^2 = \frac{\sum_{i=1}^N v_i^2}{N} - \frac{S_v^2}{N^2}. \quad (7)$$

The proof of the above theorem is a modification of Proposition 1.1 in [15]. Due to lack of space, details are omitted here and they will be provided in the journal version. Note that (6) resembles the classical Hoeffding inequality for sum of a sequence of independent random variables. Intuitively, the smaller the variance of \mathbf{v} (or \mathbf{w}), the better U will concentrate around its mean, as suggested by (6). In the special case when all elements in \mathbf{v} (or \mathbf{w}) are identical, i.e., $v_i = \alpha$ (or $w_i = \alpha$) for all i , U becomes a constant and the right-hand side of (6) becomes 0 as $\sigma_v^2 = 0$ (or $\sigma_w^2 = 0$).

3. PERFORMANCE ANALYSIS

This section presents theoretical analysis of permutation-based dimension reduction algorithm. To this end, let us first present some important properties of \mathbf{F} that are crucial to our development. Note that when \mathbf{F} is a fast-computable transform such as the FFT, the Type-II DCT and the WHT, it satisfies the following conditions:

$$\mathbf{F}_{1,j} = \sqrt{\frac{1}{N}} \quad \text{for } 1 \leq j \leq N; \quad (8)$$

$$\sum_{j=1}^N \mathbf{F}_{i,j} = 0 \quad \text{for } 2 \leq i \leq N, \quad (9)$$

$$\max |\mathbf{F}_{i,j}| = \sqrt{\frac{c_0}{N}} \quad \text{for some constant } c_0; \quad (10)$$

Eq. (8) implies that the first row of \mathbf{F} is used to capture the DC component of the input signal. (9) indicates that the remaining rows have zero DC leakages, which act as band-pass or high pass filters. Finally, Eq. (10) suggests that \mathbf{F} is a dense matrix with the magnitude of its elements on the same order. By exploiting these properties of \mathbf{F} and the concentration inequality of (6), we can derive the following theorem:

Theorem 2. Let Ω be an arbitrary set of Q points in \mathcal{R}^N and suppose that $Q \geq N$. For an $M \times N$ operator Φ given in (2), assume that the unitary transform matrix \mathbf{F} satisfies (8)-(10). Then, for all $\mathbf{u}, \mathbf{v} \in \Omega$, the following inequality

$$(1 - \epsilon) \|\mathbf{u} - \mathbf{v}\|^2 \leq \|\Phi \mathbf{u} - \Phi \mathbf{v}\|^2 \leq (1 + \epsilon) \|\mathbf{u} - \mathbf{v}\|^2 \quad (11)$$

holds with probability at least $1 - \frac{1}{Q}$, if

(i)

$$M = \mathcal{O}(\epsilon^{-2} \log^3 Q) \quad (12)$$

when \mathbf{S} is a sampling operator that uniformly picks up $M - 1$ samples from $N - 1$ outputs of \mathbf{F} ;

(ii)

$$M = \mathcal{O}(\epsilon^{-2} \log Q) \quad (13)$$

when \mathbf{S} is a sparse projection matrix as given in (3).

Proof. According to [10, 12], the proof of the above theorem is achieved if we can show that for any arbitrary N -dimensional vector \mathbf{x} with $\|\mathbf{x}\|_2 = 1$, the following inequality holds,

$$\mathbb{P}(|\|\Phi \mathbf{x}\|^2 - 1| > \epsilon) < \mathcal{O}\left(\frac{1}{Q^3}\right). \quad (14)$$

Also, since the DC component of \mathbf{x} is completely captured by the first output of \mathbf{F} , without loss of generality, we can assume that \mathbf{x} is a unit norm vector with zero mean.

Define $\mathbf{z} = \mathbf{F}\mathbf{P}\mathbf{x}$. Just as in [8, 12], we then need to bound $z_{max} = \|\mathbf{z}\|_\infty$, the maximum magnitude of elements in \mathbf{z} . Note that the i -th entry of \mathbf{z} can be expressed as

$$z_i = \sum_{j=1}^N \mathbf{F}_{i,j} x_{\pi(j)}.$$

The zero-mean condition of \mathbf{x} indicates that $z_1 = 0$. For other z_i ($2 \leq i \leq N$), by Theorem 1, we have

$$\mathbb{E}(z_i) = 0; \quad (15)$$

$$\mathbb{P}(|z_i| \geq t) \leq 2 \exp\left(\frac{-Nt^2}{4c_0}\right). \quad (16)$$

Here, (15) can be easily obtained from (9) or the zero-mean assumption of \mathbf{x} . (16) holds because of (10) and $\sigma_{\mathbf{x}} = \frac{1}{N}$. Then, applying the union bound for a supreme of a random sequence yields

$$\mathbb{P}(z_{max} \geq t) \leq 2(N-1) \exp\left(\frac{-Nt^2}{4c_0}\right) \leq 2N \exp\left(\frac{-Nt^2}{4c_0}\right),$$

where $z_{max} = \max_{2 \leq i \leq N} |z_i|$. Setting $t = \sqrt{\frac{4c_0 \log(2NQ^3)}{N}}$ produces

$$\mathbb{P}\left(z_{max} \geq \sqrt{\frac{4c_0 \log(2NQ^3)}{N}}\right) \leq \frac{1}{Q^3}. \quad (17)$$

The above result implies that the scrambled transform $\mathbf{F}\mathbf{P}$ spreads the AC component of the input vector so that the magnitude of z_i is bounded by $\sqrt{\frac{\mathcal{O}(\log Q)}{N}}$.

When \mathbf{S} is a sparse random projection matrix as in (3), the rest of the proof is the same as that in [8]. For sampling-based postprocessing, one can follow the method in [12] by modelling \mathbf{S} with a Bernoulli distribution and then applying Bernstein's inequality. In what follows, we provide an alternative way of proof by using the random permutation model. Note that since both \mathbf{F} and \mathbf{P} are unitary matrix, we have $\|\mathbf{z}\|_2^2 = \|\mathbf{x}\|_2^2 = 1$. As the sampling operator \mathbf{S} chooses $M-1$ outputs of \mathbf{z} uniform at random from $2, 3, \dots, N$, we can express $\|\Phi\mathbf{x}\|_2^2$ using a permutation operator \mathcal{S} as follows:

$$\|\Phi\mathbf{x}\|_2^2 = \frac{N-1}{M-1} \sum_{i=1}^{M-1} z_{\mathcal{S}(i)}^2 = \frac{N-1}{M-1} \sum_{i=1}^{M-1} \beta_i z_{\mathcal{S}(i)}^2, \quad (18)$$

in which \mathcal{S} is drawn from the uniform distribution over the set of all permutations of $\{2, \dots, N\}$ and β_i take the following form:

$$\beta_i = \begin{cases} \frac{N}{M}, & 1 \leq i \leq M-1; \\ 0, & M \leq i \leq N-1. \end{cases}$$

It can be easily shown that $S_\beta = (N-1)$ and $\sigma_\beta = \frac{N-1}{M-1} - 1$. Substituting them into (5) and (6), we arrive at

$$\begin{aligned} \mathbb{P}(|\|\Phi\mathbf{x}\|_2^2 - 1| > \varepsilon) &< 2 \exp\left(-\frac{(M-1)\varepsilon^2}{4(N-1)(N-M)z_{max}^4}\right) \\ &< 2 \exp\left(-\frac{M\varepsilon^2}{4N^2 z_{max}^4}\right). \end{aligned} \quad (19)$$

Finally, combining (17) with (19) yields

$$\begin{aligned} &\mathbb{P}(|\|\Phi\mathbf{x}\|_2^2 - 1| > \varepsilon) \\ &< 2 \exp\left(-\frac{M\varepsilon^2}{64c_0^2(\log(2NQ^3))^2}\right) + \frac{1}{Q^3} \end{aligned} \quad (20)$$

When $M \geq 64c_0^2\varepsilon^{-2}(\log Q^3)(\log 2NQ^3)^2 = \mathcal{O}(\varepsilon^{-2} \log^3 Q)$, the first term on the right side of (20) is less than $\frac{2}{Q^3}$ and thus the right side is less than $\frac{3}{Q^3}$, which completes the proof. \square

Note that (12) and (13) are identical to those obtained in [12] and [8], respectively. This suggests that the effect of random permutation is similar to that of random sign flipping. One can also notice that in sampling-based postprocessing, there is an upscaling factor of $\log^2 Q$ in M . But we want to stress here that (12) is only a theoretical bound for the worst case analysis. Our numerical experiments suggest that sampling-based postprocessing provides very similar performance as that of full random matrix at much lower implementation cost. In terms of computation complexity, as we have mentioned before, random permutation requires $\mathcal{O}(N)$ operations. When \mathbf{S} is a sampling operator, the computation of an $M \times N$ partial FFT/DCT/WHT requires $\mathcal{O}(M \log N) = \mathcal{O}(\varepsilon^{-2} \log^3 Q \log N)$ operations. Hence, the total complexity is

$$\max(\mathcal{O}(\varepsilon^{-2} \log^3 Q \log N), \mathcal{O}(N)).$$

For projection-based approach, it remains the same as that of FJLT in [8], which requires

$$\mathcal{O}(N \log N + \min(N\varepsilon^{-2} \log Q, \varepsilon^{-2} \log^3 Q))$$

operations.

4. SIMULATION RESULTS

To demonstrate the effectiveness of permutation-based dimension reduction operators, we evaluate their performance on a Corel database with 1000 images¹ through a couple of experiments as follows:

Experiment 1: In this experiment, the global HSV histogram is used as feature vector due to its popularity in image indexing and retrieval. Specifically, a $N = 128$ -dimensional vector was computed for every image by quantizing the hue, the saturation, and the intensity channels into 8, 4 and 4 levels, respectively, as suggested by [5]. The reduced dimensionality M ranges from 5 to 90. For each M , we generate Φ according to (2) where \mathbf{F} is chosen as the WHT. 3000 pairs of image feature vectors were then selected randomly and for each pair of \mathbf{u}_i and \mathbf{v}_i , the projection distortion is measured by

$$D_M(i) = \left| \frac{\|\Phi(\mathbf{u}_i - \mathbf{v}_i)\|_2}{\|\mathbf{u}_i - \mathbf{v}_i\|_2} - 1 \right|.$$

The average distortions for each M are shown in Fig. 1(a). In this figure, 'SWHT+P' and 'SWHT+S' represent scrambled WHT (SWHT) with projection-based and sampling-based post-processing, respectively. For comparison purposes, we also include the results of i.i.d Gaussian random matrix and FJLT [8].

Experiment 2: This experiment was motivated by the recent work in [6] and the experiment suggested by [16]. For each image in the database, we random select 20 subimages with size of 32×32 , yielding $Q = 20 * 1000 = 2 \times 10^4$ vectors in total. For each M , we did a similar experiment in Experiment 1 calculating the pairwise distance for 4000 pairs and the results are given in Fig. 1(b).

As one can see, in both experiments, permutation-based operators offer comparable performance to those of full random Gaussian matrix and FJLT. This suggests that the this subclass of operators is quite promising in practical large-scale applications. In fact, it could be easily combined with recent works on secure image retrieval [5]

¹ Available at <http://wang.ist.psu.edu/docs/related.shtml>

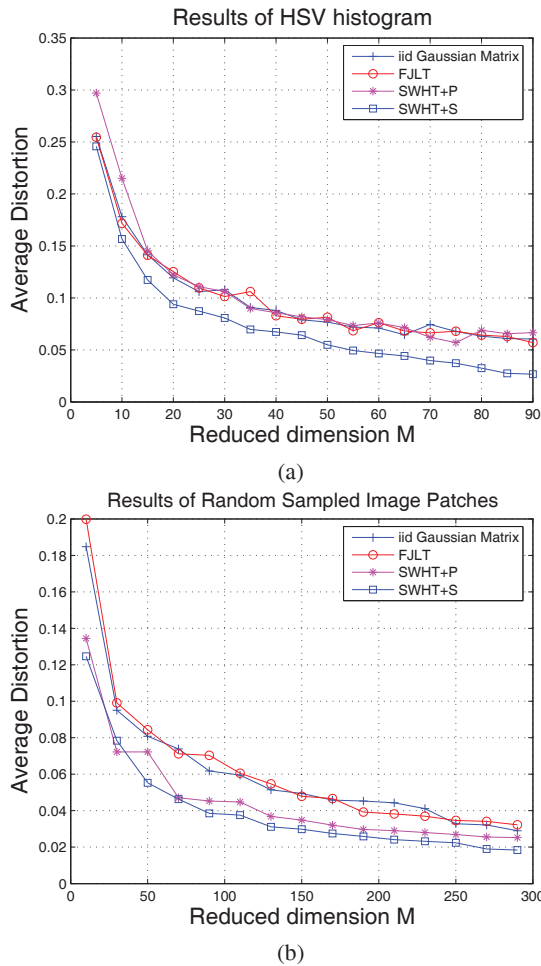


Fig. 1. Average relative distortion with different random projection methods. (a) Results of HSV histogram; (b) Results of random sampled 32×32 image patches. The projection methods include (1) i.i.d Gaussian matrix; (2) FJLT proposed in [8]; (3)SWHT+P: Our proposed scrambled WHT (SWHT) with random sparse projection; (4) SWHT+S: Our proposed SWHT with random sampling;

and image hashing [6], where random projection plays a key role. It should be pointed out that although sampling-based post-processing schemes (e.g., SWHT+S in Fig. 1) do not provide optimal theoretical performance, the simulation results suggest that they could be used as a fast dimension reduction tool for practical image processing without much performance degradation. This also implies that in Theorem 2, the power 3 of the $\log Q$ term could be very likely improved to 1.

5. CONCLUSIONS

In this paper, we propose fast dimension algorithm that can be implemented through the following steps (i) Preprocessing the input vector using random permutation; (ii) Applying a fast transform such as the FFT, DCT or the WHT; (iii) Post-processing the output vector using either random sampling or sparse random projection. Theoretically, the proposed algorithms were proved to be (sub)optimal Johnson-Lindenstrauss transforms. Simulation results are also included to demonstrate their promising applications in image processing.

6. REFERENCES

- [1] E. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. on Information Theory*, vol. 52, pp. 5406 – 5425, Dec. 2006.
- [2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1289–1306, July 2006.
- [3] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Conf. in Modern Analysis and Probability*, pp. 189–206, 1984.
- [4] L. Gan, T. T. Do, and T. D. Tran, "Fast compressive imaging using scrambled block hadamard ensemble," in *Proc. of EUSIPCO 2008*, 2008.
- [5] W. Lu, A.-L. Varna, A. Swaminathan, and M. Wu, "Secure image retrieval through feature protection," in *Proc. of ICASSP 2009*, 2009, pp. 1533–1536.
- [6] X. Lv and Z. Wang, "An extended image hashing concept: Content-based fingerprinting using FJLT," *EURASIP Journal on Information Security*, 2009.
- [7] D. Achlioptas, "Database-friendly random projections: Johnson-Lindenstrauss with binary coins," *Journal of Computer and System Sciences*, vol. 66, pp. 671–687, 2003.
- [8] 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.
- [9] J. Matousek, "On the variants of the Johnson-Lindenstrauss lemma," *Random Structure and Algorithms*, vol. 33, pp. 142–156, 2008.
- [10] N. Ailon and E. Liberty, "Fast dimension reduction using Rademacher series on BCH codes," in *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, San Francisco, California, 2008, pp. 1–9.
- [11] E. Liberty, N. Ailon, and A. Singer, "Dense fast random projections and Lean Walsh transform," in *RANDOM*, Boston, MA, Aug. 2008.
- [12] T. T. Do, L. Gan, Y. Chen, N. Nguyen, and T. D. Tran, "Fast and efficient dimension reduction using structurally random matrices," in *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, 2009, pp. 1821–1824.
- [13] N. Ailon and E. Liberty, "Fast dimension reduction using rademacher series on dual BCH codes," *To appear in Discrete and Computational Geometry*, 2008.
- [14] R. Graf and W. Sheets, *Video Scrambling and Descrambling for Satellite and Cable TV*. Newnes, 1998.
- [15] S. Chatterjee, "Stein's method for concentration inequalities," *Probability Theory Related Fields*, vol. 138, pp. 305–321, 2007.
- [16] 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.