Advances in Adaptive Data Analysis
Vol. 3, Nos. 1 & 2 (2011) 115–126

DOI: 10.1142/S1793536911000787

World Scientific
www.worldscientific.com

IMPROVED ANALYSIS OF THE SUBSAMPLED RANDOMIZED HADAMARD TRANSFORM

JOEL A. TROPP

Computing & Mathematical Sciences,
MC 305–16, California Institute of Technology,
Pasadena, CA 91125, USA
jtropp@cms.caltech.edu

Received 17 June 2010
Revised 5 November 2010

This paper presents an improved analysis of a structured dimension-reduction map called the subsampled randomized Hadamard transform. This argument demonstrates that the map preserves the Euclidean geometry of an entire subspace of vectors. The new proof is much simpler than previous approaches, and it offers — for the first time — optimal constants in the estimate on the number of dimensions required for the embedding.

Keywords: Dimension reduction; numerical linear algebra; random matrix; Walsh–Hadamard matrix.

Mathematics Subject Classification 2010: Primary: 15B52

1. Introduction

Dimension reduction is an elegant idea from computer science that has found applications in numerical linear algebra. Here is the basic concept: it is often more efficient to solve a computational problem presented in a high-dimensional space if we first transport the problem instance to a lower-dimensional space while preserving its essential structure. Researchers have found that randomness provides an extraordinarily effective way to construct these dimension-reduction maps. This approach is usually traced to the celebrated paper of Johnson and Lindenstrauss [1984].

In randomized algorithms for matrix approximation, the goal of dimension reduction is to find a low-dimensional subspace that captures most of the action of an input matrix. One way to accomplish this task is to multiply the input matrix $\boldsymbol{A}$ by a relatively small dimension-reduction matrix $\boldsymbol{\Omega}$ to obtain $\boldsymbol{Y} = \boldsymbol{A\Omega}$. We then perform a QR factorization $\boldsymbol{Y} = \boldsymbol{QR}$ to identify the range of the reduced matrix $\boldsymbol{Y}$. For an appropriately designed dimension-reduction map, it can be shown that $\boldsymbol{A} \approx \boldsymbol{QQ}^*\boldsymbol{A}$ with high probability. In other words, we can approximate the input matrix by compressing it to the range of the reduced matrix. It is then possible to compute standard matrix decompositions of $\boldsymbol{A}$ by manipulating the low-rank

approximation $\boldsymbol{QQ}^*\boldsymbol{A}$. See the recent survey [Halko *et al.* (2011)] for a comprehensive treatment of these ideas and an extensive bibliography.

In linear algebra applications, the cost of multiplication by an unstructured random matrix $\boldsymbol{\Omega}$, such as a Gaussian matrix, can sometimes be prohibitive. In that case, we may prefer to draw the random matrix $\boldsymbol{\Omega}$ from a *highly structured* distribution that allows us to multiply $\boldsymbol{\Omega}$ into the input matrix substantially faster. Sarlós is credited with bringing structured dimension reduction to numerical linear algebra [Sarlós (2006)]; see also [Woolfe *et al.* (2008)].

The *subsampled randomized Hadamard transform* (SRHT) is a type of structured dimension-reduction map that is based on the Walsh–Hadamard matrix. We prove that the SRHT preserves the geometry of an *entire subspace* of vectors, which is the essential ingredient required to show that the SRHT can be used in algorithms for randomized linear algebra. See the discussion in Halko *et al.* [2009, Sec. 11] for details.

The literature already contains a number of papers, including [Ailon and Chazelle (2009); Liberty (2009); Nguyen *et al.* (2009); Halko *et al.* (2011); Ailon and Liberty (2010); Krahmer and Ward (2010)], that study the behavior of the SRHT and related dimension-reduction maps. The current treatment differs in several regards. Here, the main technical difficulties are addressed using a version of the matrix Chernoff inequality [Ahslwede and Winter (2002); Tropp (2010)]. As a consequence of the simple proof schema, we are able — for the first time — to obtain optimal constants in our bounds. This improvement can be valuable in numerical applications that require guarantees concrete performance.

This document is adapted from Appendix B of the technical report [Halko *et al.* (2009)], but much of the proof is new. This paper should be viewed as a codicil to the published version [Halko *et al.* (2011)] of the technical report, which uses the results presented here.

1.1. *Construction of the SRHT matrix*

An SRHT is a wide $\ell \times n$ matrix of the form

$$\boldsymbol{\Phi} = \sqrt{\frac{n}{\ell}} \cdot \boldsymbol{RHD}$$

where

- $\boldsymbol{D}$ is a random $n \times n$ diagonal matrix whose entries are independent random signs, i.e., random variables uniformly distributed on $\{\pm 1\}$;
- $\boldsymbol{H}$ is an $n \times n$ Walsh–Hadamard matrix, scaled by $n^{-1/2}$ and so it is an orthogonal matrix; and
- $\boldsymbol{R}$ is a random $\ell \times n$ matrix that restricts an n-dimensional vector to ℓ coordinates, chosen uniformly at random.

Our analysis relies on two basic properties of the Walsh–Hadamard matrix: The derived matrix $\boldsymbol{H}$ is orthogonal, and its entries all have magnitude $\pm n^{-1/2}$. Walsh–Hadamard matrices exist for each $n = 2^p$ where $p = 1, 2, 3, \dots$.

Remark 1.1. The Walsh–Hadamard matrix is the real analog of the discrete Fourier matrix. Whereas the $n \times n$ Fourier matrix displays the characters of the cyclic group of order n, the $n \times n$ Walsh–Hadamard matrix contains the characters of the additive group $\mathbb{Z}_2^p$ where $n = 2^p$. In each case, the underlying algebraic structure allows us to multiply the matrix by a vector in about $n \log n$ arithmetic operations. Note, however, that our results are purely analytic; so, they do not rely on the algebraic properties of the Walsh–Hadamard matrix. We have chosen not to work with the Fourier matrix to avoid some small complications associated with complex random variables.

1.2. *Intuition*

The design of the SRHT may seem mysterious, but there are clear intuitions to support it [Ailon and Chazelle (2009)]. Suppose that we want to estimate the energy (i.e., squared ℓ_2 norm) of a fixed vector $\boldsymbol{x}$ by sampling ℓ of its n entries at random. On average, these random entries carry ℓ/n of the total energy. (The factor $\sqrt{n/\ell}$ reverses this scaling.) When $\boldsymbol{x}$ has a few large components, the variance of this estimator is very high. On the other hand, when the components of $\boldsymbol{x}$ have comparable magnitude, the estimator has a much lower variance, so it is precise with very high probability.

The matrix $\boldsymbol{HD}$ is orthogonal, so it preserves energy. At the same time, the matrix $\boldsymbol{HD}$ flattens out the entries of a vector to improve the performance of randomized sampling. To see how it achieves this goal, fix a unit vector $\boldsymbol{x}$, and examine the first component of $\boldsymbol{HDx}$.

$$(\boldsymbol{HDx})_1 = \sum_{j=1}^{n} h_{1j} \varepsilon_j x_j,$$

where h_{ij} are the components of the matrix $\boldsymbol{H}$. This random sum clearly has zero mean. Since the entries of $\boldsymbol{H}$ have magnitude $n^{-1/2}$, the variance of the sum is n^{-1}. Hoeffding's inequality [Hoeffding (1963)] shows that

$$\mathbb{P}\{|(\boldsymbol{HDx})_1| \geq t\} \leq 2\mathrm{e}^{-nt^2/2}.$$

In other words, the magnitude of the first component of $\boldsymbol{HDx}$ is typically about $n^{-1/2}$. The same argument applies to the remaining entries. Therefore, it is unlikely that any one of the n components of $\boldsymbol{HDx}$ is larger than $\sqrt{n^{-1} \log(n)}$.

Remark 1.2. This discussion suggests that the precise form of the SRHT is not particularly important. Indeed, we can replace $\boldsymbol{H}$ by any unitary matrix whose entries are uniformly small and which is equipped with a fast matrix–vector multiply. We can also draw the diagonal entries of $\boldsymbol{D}$ from other sub-gaussian distributions. These changes somewhat complicate the analysis.

1.3. *Main results*

In early work, dimension reduction was accomplished with uniformly random projectors or, later, with Gaussian matrices. These random maps preserve, up to a constant factor, all the pairwise distances among p points in $\mathbb{R}^n$, provided that the embedding dimension is about $\log(p)$. We can also use a Gaussian matrix to transport an (unknown) k-dimensional subspace from a high-dimensional space to a lower-dimensional space. In this case, it suffices to map the ambient space down to $O(k)$ dimensions because we can discretize a k-dimensional sphere using $p = O(\mathrm{e}^k)$ points.

It turns out that an appropriately designed SRHT also preserves the geometry of an entire subspace of vectors. For this type of structured map, the embedding requires $k \log(k)$ dimensions because of certain phenomena connected with random sampling (Section 3.3). It also becomes necessary to use more sophisticated proof techniques. We have the following result.

Theorem 1.3 (The SRHT preserves geometry). *Fix an $n \times k$ matrix $\boldsymbol{V}$ with orthonormal columns, and draw a random $\ell \times n$ SRHT matrix $\boldsymbol{\Phi}$ where the embedding dimension ℓ satisfies*

$$4[\sqrt{k} + \sqrt{8 \log(kn)}]^2 \log(k) \leq \ell \leq n.$$

Then, except with probability $O(k^{-1})$

$$0.40 \leq \sigma_k(\boldsymbol{\Phi V}) \quad \text{and} \quad \sigma_1(\boldsymbol{\Phi V}) \leq 1.48.$$

The symbol $\sigma_j(\cdot)$ denotes the jth largest singular value of a matrix.

The proof of Theorem 1.3 appears below. This result replaces Theorem B.4 of [Halko *et al.* (2009)]. Some precedents include [Rudelson and Vershynin (2007), Theorem 3.1] and [Tropp (2008), Sec. 9]; see also [Nguyen *et al.* (2009)]. Earlier results have the same structure as Theorem 1.3, but the constants are either exorbitant or absent.

Remark 1.4. In Theorem 1.3, the factor $\log(k)$ in the lower bound on ℓ cannot generally be removed. See Sec. 3.3 for an explanation and an example that demonstrates the necessity.

Remark 1.5. For large problems, we have been able to obtain optimal numerical constants. Suppose that ι is a sufficiently small positive number. If $k \gg \log(n)$,

then sampling

$$\ell \geq (1+\iota) \cdot k \log(k)$$

the coordinates is sufficient to ensure that $\boldsymbol{\Phi V}$ has a constant condition number. See Theorem 3.2 for a more precise statement. The discussion in Sec. 3.3 indicates that there are cases where it does not suffice to draw $(1-\iota) \cdot k \log(k)$ samples.

2. Technical Background

In preparation for the main argument, we present our notation and some probability inequalities. These inequalities encapsulate all the difficulty in the proof.

2.1. *Notation*

We write $\mathbf{e}_j$ for the jth standard basis vector in $\mathbb{R}^n$. The matrix $\mathbf{I}$ is a square identity matrix; we sometimes indicate the dimension with a subscript. The symbol " * " denotes transposition of vectors and matrices.

Throughout this work, $\|\cdot\|$ refers to the ℓ_2 vector norm or the associated operator norm. We also write $\|\cdot\|_{\mathrm{F}}$ for the Frobenius norm.

Given a subset T of indices in $\{1, 2, \dots, n\}$, we define the restriction operator $\boldsymbol{R}_T : \mathbb{R}^n \to \mathbb{R}^T$ via the rule

$$(\boldsymbol{R}_T \boldsymbol{x})(j) = x_j, \quad j \in T.$$

A *Rademacher random variable* takes the values ± 1 with equal probability. We reserve the letter ε for a Rademacher variable, and we often write $\boldsymbol{\varepsilon}$ for a vector whose entries are independent Rademacher variables.

2.2. *Probability inequalities*

We delegate the hard work to some probability inequalities that describe the large-deviation behavior of specific types of random variables. First, we describe a tail bound for a convex function of Rademacher variates. This result was established by Ledoux [Ledoux (1996), Eq. (1.9)]; see also [Ledoux (2001), Sec. 5.2] for a discussion of concentration in product spaces.

Proposition 2.1 (Rademacher tail bound). *Suppose f is a convex function on vectors that satisfies the Lipschitz bound*

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L \, \|\boldsymbol{x} - \boldsymbol{y}\| \quad \text{for all } \boldsymbol{x}, \boldsymbol{y}.$$

Let $\boldsymbol{\varepsilon}$ be a Rademacher vector. For all $t \geq 0$

$$\mathbb{P}\{f(\boldsymbol{\varepsilon}) \geq \mathbb{E} f(\boldsymbol{\varepsilon}) + Lt\} \leq \mathrm{e}^{-t^2/8}.$$

Next, we present a matrix analog of the well-known Chernoff inequality. The proof is based on the matrix Laplace transform method proposed in an influential paper of Ahlswede–Winter [Ahslwede and Winter (2002)]. We derive the result using more recent ideas [Tropp (2010)], which deliver an essential improvement on the earlier work. The other key tool is a method [Gross and Nesme (2010)], ultimately due to Hoeffding [Hoeffding (1963)], for transferring results from the model where we sample with replacement to the model where we sample without replacement.

Theorem 2.1. (*Matrix Chernoff*) *Let $\mathscr{X}$ be a finite set of positive-semidefinite matrices with dimension k, and suppose that*

$$\max_{\boldsymbol{X} \in \mathscr{X}} \lambda_{\max}(\boldsymbol{X}) \leq B.$$

Sample $\{\boldsymbol{X}_1, \dots, \boldsymbol{X}_\ell\}$ uniformly at random from $\mathscr{X}$ without replacement. Compute

$$\mu_{\min} := \ell \cdot \lambda_{\min}(\mathbb{E}\boldsymbol{X}_1) \quad \text{and} \quad \mu_{\max} := \ell \cdot \lambda_{\max}(\mathbb{E}\boldsymbol{X}_1).$$

Then

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_j \boldsymbol{X}_j\right) \leq (1-\delta)\mu_{\min}\right\} \leq k \cdot \left[\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/B} \quad \text{for } \delta \in [0,1), \quad \text{and}$$

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_j \boldsymbol{X}_j\right) \geq (1+\delta)\mu_{\max}\right\} \leq k \cdot \left[\frac{\mathrm{e}^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_{\max}/B} \quad \text{for } \delta \geq 0.$$

Proof sketch. We establish only the upper bound; the lower bound is established by applying a similar method to $-\sum_j \boldsymbol{X}_j$. By homogeneity, take $B = 1$. We use the matrix Laplace transform method [Tropp (2010), Proposition 3.1] to bound the probability of a large deviation:

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_j \boldsymbol{X}_j\right) \geq (1+t)\mu_{\max}\right\} \leq \inf_{\theta > 0}\left\{\mathrm{e}^{-\theta(1+t)\mu_{\max}} \cdot \mathbb{E}\,\mathrm{tr}\exp\left(\sum_j \theta \boldsymbol{X}_j\right)\right\}. \tag{1}$$

The Laplace transform bound (1) is essentially due to Ahlswede and Winter [Ahslwede and Winter (2002)].

Let $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_\ell\}$ be a random sample from $\mathscr{X}$ drawn with replacement. Gross and Nesme [Gross and Nesme (2010)] have shown that the trace of the matrix moment generating function (MGF) of a sample without replacement is dominated

by the trace of the matrix MGF of a sample with replacement:

$$\mathbb{E}\,\mathrm{tr}\,\exp\left(\sum_j \theta \boldsymbol{X}_j\right) \le \mathbb{E}\,\mathrm{tr}\,\exp\left(\sum_j \theta \boldsymbol{Y}_j\right). \tag{2}$$

Lemma 5.8 of [Tropp (2010)] gives a semidefinite bound for the MGF of $\boldsymbol{Y}_j$. (See also [Ahslwede and Winter (2002), Thm. 19].)

$$\mathbb{E}\mathrm{e}^{\theta \boldsymbol{Y}_j} \preccurlyeq \mathrm{e}^{(\mathrm{e}^\theta - 1)(\mathbb{E}\boldsymbol{Y}_j)} = \mathrm{e}^{(\mathrm{e}^\theta - 1)(\mathbb{E}\boldsymbol{X}_1)} \quad \text{for } \theta \in \mathbb{R}. \tag{3}$$

On the right-hand side of this bound, we have exploited the fact that $\boldsymbol{Y}_j \sim \boldsymbol{Y}_1 \sim \boldsymbol{X}_1$. Lemma 3.4 of [Tropp (2010)] allows us to use the MGF bound (3) to bound the MGF of the entire sum:

$$\mathbb{E}\,\mathrm{tr}\,\exp\left(\sum_j \theta \boldsymbol{Y}_j\right) \le \mathrm{tr}\,\exp\left((\mathrm{e}^\theta - 1)\cdot \ell \cdot (\mathbb{E}\boldsymbol{X}_1)\right) \le k \cdot \exp\left((\mathrm{e}^\theta - 1)\cdot \mu_{\max}\right). \tag{4}$$

Substitute Equation (4) into Equation (2) and introduce the resulting inequality into the probability bound (1). The infimum is achieved at $\theta = \log(1+\delta)$. □

3. The SRHT Preserves Geometry

In this section, we establish a slightly more specific version of our main result, Theorem 1.3.

Theorem 3.1 (The SRHT preserves geometry). *Let $\boldsymbol{V}$ be an $n \times k$ matrix with orthonormal columns. Select a parameter ℓ that satisfies*

$$4[\sqrt{k} + \sqrt{8\log(kn)}]^2 \log(k) \le \ell \le n.$$

Draw an $\ell \times n$ SRHT matrix $\boldsymbol{\Phi}$. Then, except with probability $3k^{-1}$

$$\frac{1}{\sqrt{6}} \le \sigma_k(\boldsymbol{\Phi V}) \quad \text{and} \quad \sigma_1(\boldsymbol{\Phi V}) \le \sqrt{\frac{13}{6}}.$$

The constants in the sample size ℓ specified in Theorem 3.1 are somewhat larger than we might like because we wanted to ensure that the statement is effective for reasonable values of k and n. We also have the following result of a more asymptotic flavor.

Theorem 3.2 (SRHT: Large sample bounds). *Fix a positive number $\iota \le \mathrm{c}$. Let $\boldsymbol{V}$ be an $n \times k$ matrix with orthonormal columns, where $k \ge \mathrm{C}\iota^{-2}\log(n)$. Select a parameter ℓ that satisfies*

$$(1+\iota)\cdot k\log(k) \le \ell \le n.$$

Draw an $\ell \times n$ SRHT $\boldsymbol{\Phi}$. Then

$$\iota \leq \sigma_k(\boldsymbol{\Phi}\boldsymbol{V}) \quad \text{and} \quad \sigma_1(\boldsymbol{\Phi}\boldsymbol{V}) \leq \sqrt{\mathrm{e}}$$

except with probability $\mathrm{O}(k^{-\mathrm{c}\iota})$. The numbers c and C are positive universal constants.

3.1. *Overview of Theorem 3.1*

We present the main line of reasoning here, postponing the proofs of the lemmata. The first step is to show that the matrix $\boldsymbol{HD}$ equilibrates row norms.

Lemma 3.3 (Row norms). *Let $\boldsymbol{V}$ be an $n \times k$ matrix with orthonormal columns. Then $\boldsymbol{HDV}$ is an $n \times k$ matrix with orthonormal columns, and*

$$\mathbb{P}\left\{ \max_{j=1,\dots,n} \left\| \mathbf{e}_j^*(\boldsymbol{HDV}) \right\| \geq \sqrt{\frac{k}{n}} + \sqrt{\frac{8\log(\beta n)}{n}} \right\} \leq \frac{1}{\beta}.$$

When $k \gg \log n$, the second term in the bound is negligible, in which case the row norms are essentially as small as possible. On the other hand, when $k < \log n$, small sample effects can make some row norms large.

The next result states that randomly sampling rows from a matrix with orthonormal columns results in a well-conditioned matrix. The minimum size of the sample depends primarily on the row norms, and the sampling procedure is most efficient when the matrix has uniformly small rows. We state this result in detail because it may have independent interest.

Lemma 3.4 (Row sampling). *Let $\boldsymbol{W}$ be an $n \times k$ matrix with orthonormal columns, and define the quantity $M := n \cdot \max_{j=1,\dots,n} \left\| \mathbf{e}_j^* \boldsymbol{W} \right\|^2$. For a positive parameter α, select the sample size*

$$\ell \geq \alpha M \log(k).$$

Draw a random subset T from $\{1, 2, \dots, n\}$ by sampling ℓ coordinates without replacement. Then

$$\sqrt{\frac{(1-\delta)\ell}{n}} \leq \sigma_k(\boldsymbol{R}_T\boldsymbol{W}) \quad \text{and} \quad \sigma_1(\boldsymbol{R}_T\boldsymbol{W}) \leq \sqrt{\frac{(1+\eta)\ell}{n}},$$

with failure probability at most

$$k \cdot \left[\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}} \right]^{\alpha \log k} + k \cdot \left[\frac{\mathrm{e}^{\eta}}{(1+\eta)^{(1+\eta)}} \right]^{\alpha \log k}.$$

These two lemmata allow us to establish Theorem 3.1 quickly. Recall that $\boldsymbol{V}$ is an $n \times k$ matrix with orthonormal columns, and draw an $\ell \times n$ SRHT $\boldsymbol{\Phi} = \boldsymbol{R}_T\boldsymbol{HD}$.

Define the matrix $\boldsymbol{W} = \boldsymbol{HDV}$. Lemma 3.3 with $\beta = k$ establishes that $\boldsymbol{W}$ is a matrix with orthonormal columns whose largest row norm is essentially as small as possible, so that

$$M := n \cdot \max_{j=1,\dots,n} \left\| \mathbf{e}_j^* \boldsymbol{W} \right\|^2 \leq [\sqrt{k} + \sqrt{8 \log(kn)}]^2,$$

except with probability k^{-1}. Next, we apply Lemma 3.4 with $\alpha = 4$, with $\delta = 5/6$, and with $\eta = 7/6$. A numerical reckoning shows that the additional probability of failure is at most $2k^{-1}$. Altogether, we obtain the following result. For

$$\ell \geq 4M \log(k),$$

it holds that

$$\sqrt{\frac{\ell}{6n}} \leq \sigma_k(\boldsymbol{R}_T \boldsymbol{W}) \quad \text{and} \quad \sigma_1(\boldsymbol{R}_T \boldsymbol{W}) \leq \sqrt{\frac{13\ell}{6n}}$$

except with probability $3k^{-1}$. Since $\boldsymbol{\Phi V} = \sqrt{n/\ell} \cdot \boldsymbol{R}_T \boldsymbol{W}$, we have established Theorem 3.1.

The same type of argument implies Theorem 3.2. In this case, we fix a sufficiently small positive number ι. Then, we take $\beta = k$ and $\alpha = 1 + \iota/2$ and $\delta = 1 - \iota^2$ and $\eta = \mathrm{e} - 1$. We omit the details.

Remark 3.5. We can establish a slightly stronger sample bound in Theorem 3.2 by choosing the parameter $\iota = A/\log(k)$ for a sufficiently large number A. With this selection, however, we do not obtain a constant bound for the lower singular value.

3.2. *Proofs of supporting lemmas*

It remains to check that the underlying results are true. We begin with the claim that $\boldsymbol{HD}$ balances row norms.

Proof of Lemma 3.3. The orthonormality condition on $\boldsymbol{V}$ is equivalent with the identity $\boldsymbol{V}^*\boldsymbol{V} = \mathbf{I}_k$. Therefore, $\|\boldsymbol{V}\| = 1$ and $\|\boldsymbol{V}\|_{\mathrm{F}} = \sqrt{k}$. To check that $\boldsymbol{HDV}$ is also an orthonormal matrix, simply compute that

$$(\boldsymbol{HDV})^*(\boldsymbol{HDV}) = \boldsymbol{V}^*\boldsymbol{V} = \mathbf{I}$$

because $\boldsymbol{D}, \boldsymbol{H}$ are orthogonal matrices.

Fix a row index $j \in \{1, 2, \dots, n\}$, and define the function

$$f(\boldsymbol{x}) := \left\| \mathbf{e}_j^* \boldsymbol{H} \operatorname{diag}(\boldsymbol{x}) \boldsymbol{V} \right\| =: \left\| \boldsymbol{x}^* \boldsymbol{E} \boldsymbol{V} \right\|.$$

We have written $\boldsymbol{E} := \operatorname{diag}(\mathbf{e}_j^* \boldsymbol{H})$ for the diagonal matrix constructed from the jth row of $\boldsymbol{H}$; observe that each entry of $\boldsymbol{E}$ has magnitude $n^{-1/2}$. The function f is convex, and we quickly determine its Lipschitz constant:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \le \|(\boldsymbol{x} - \boldsymbol{y})^* \boldsymbol{E} \boldsymbol{V}\| \le \|\boldsymbol{x} - \boldsymbol{y}\| \, \|\boldsymbol{E}\| \, \|\boldsymbol{V}\| = \frac{1}{\sqrt{n}} \|\boldsymbol{x} - \boldsymbol{y}\| .$$

We may use the function f to study the variation of the row norms of $\boldsymbol{HDV}$. Recall that $\boldsymbol{D} := \operatorname{diag}(\boldsymbol{\varepsilon})$ for a Rademacher vector $\boldsymbol{\varepsilon}$, and consider the random variable

$$f(\boldsymbol{\varepsilon}) = \left\| \mathbf{e}_j^* \boldsymbol{HDV} \right\| .$$

First, we bound the expectation:

$$\mathbb{E} f(\boldsymbol{\varepsilon}) \le \left[\mathbb{E} f(\boldsymbol{\varepsilon})^2 \right]^{1/2} = \|\boldsymbol{EV}\|_{\mathrm{F}} \le \|\boldsymbol{E}\| \, \|\boldsymbol{V}\|_{\mathrm{F}} = \sqrt{\frac{k}{n}}.$$

Apply the Rademacher tail bound, Proposition 2.1, with $t = \sqrt{8 \log(\beta n)}$ to reach

$$\mathbb{P} \left\{ \|\mathbf{e}_j^* (\boldsymbol{HDV})\| \ge \sqrt{\frac{k}{n}} + \sqrt{\frac{8 \log(\beta n)}{n}} \right\} \le \mathrm{e}^{-8 \log(\beta n)/8} = \frac{1}{\beta n}.$$

This estimate holds for each row index $j = 1, 2, \dots, n$. Finally, take a union bound over these n events to reach the advertised conclusion. □

Now, we establish the result on row sampling.

Proof of Lemma 3.4. Let $\boldsymbol{w}_j^*$ denote the jth row of $\boldsymbol{W}$, and define $M := n \cdot \max_j \|\boldsymbol{w}_j\|^2$.

We can control the extreme singular values of the random matrix $\boldsymbol{R}_T \boldsymbol{W}$ by bounding the extreme eigenvalues of the $k \times k$ Gram matrix

$$\boldsymbol{Y} := (\boldsymbol{R}_T \boldsymbol{W})^* (\boldsymbol{R}_T \boldsymbol{W}) = \sum_{j \in T} \boldsymbol{w}_j \boldsymbol{w}_j^*.$$

Recall that T is a random subset of $\{1, 2, \dots, n\}$ consisting of ℓ coordinates sampled without replacement. Therefore, we may just as well view $\boldsymbol{Y}$ as a sum of ℓ random matrices $\boldsymbol{X}_1, \dots, \boldsymbol{X}_\ell$ sampled without replacement from the family $\mathscr{X} := \{\boldsymbol{w}_j \boldsymbol{w}_j^* : j = 1, 2, \dots, n\}$ of positive-semidefinite matrices with dimension k.

The matrix Chernoff bound, Theorem 2.1, allows us to obtain large deviation bounds for the extreme eigenvalues of $\boldsymbol{Y}$. We just need to determine the parameters

involved in the statement. First, note that

$$\lambda_{\max}(\boldsymbol{w}_j \boldsymbol{w}_j^*) = \|\boldsymbol{w}_j\|^2 \leq \frac{M}{n} \quad \text{for } j = 1, 2, \dots, n.$$

We easily compute the expectation of the first sample using the fact that the columns of $\boldsymbol{W}$ are orthonormal:

$$\mathbb{E}\boldsymbol{X}_1 = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{w}_j \boldsymbol{w}_j^* = \frac{1}{n} \boldsymbol{W}^* \boldsymbol{W} = \frac{1}{n} \mathbf{I}.$$

As a consequence, we obtain

$$\mu_{\min} = \frac{\ell}{n} \quad \text{and} \quad \mu_{\max} = \frac{\ell}{n}.$$

The lower Chernoff bound yields

$$\mathbb{P}\left\{ \lambda_{\min}(\boldsymbol{Y}) \leq (1-\delta)\frac{\ell}{n} \right\} \leq k \cdot \left[\frac{\mathrm{e}^{-\delta}}{(1-\delta)^{1-\delta}} \right]^{\ell/M} \quad \text{for } \delta \in [0, 1).$$

The upper Chernoff bound yields

$$\mathbb{P}\left\{ \lambda_{\max}(\boldsymbol{Y}) \geq (1+\eta)\frac{\ell}{n} \right\} \leq k \cdot \left[\frac{\mathrm{e}^{\eta}}{(1+\eta)^{1+\eta}} \right]^{\ell/M} \quad \text{for } \eta \geq 0.$$

Substitute the identities

$$\lambda_{\min}(\boldsymbol{Y}) = \sigma_k(\boldsymbol{R}_T \boldsymbol{W})^2 \quad \text{and} \quad \lambda_{\max}(\boldsymbol{Y}) = \sigma_1(\boldsymbol{R}_T \boldsymbol{W})^2.$$

Simplify the formulae to complete the proof. □

3.3. *Collecting Coupons*

The logarithmic factor in these results is necessary, as we show by example. This discussion is extracted from [Halko *et al.* (2011), Sec. 11].

Fix an integer k, and set $n = k^2$. Form an $n \times k$ orthonormal matrix $\boldsymbol{W}$ by regular decimation of the $n \times n$ identity matrix. More precisely, $\boldsymbol{W}$ is the matrix whose jth row has a unit entry in column $1 + (j-1)/k$ when $j \equiv 1 \pmod{k}$ and is zero otherwise. To see why this type of matrix is inconvenient, it is helpful to consider an auxiliary matrix $\boldsymbol{V} = \boldsymbol{HDW}$. Observe that, up to scaling and modulation of rows, $\boldsymbol{V}$ consists of k copies of a $k \times k$ Walsh–Hadamard transform stacked vertically.

Now, suppose that we apply an SRHT $\boldsymbol{\Phi} = \boldsymbol{RHD}$ to the matrix $\boldsymbol{W}$. We obtain a matrix of the form $\boldsymbol{X} = \boldsymbol{\Phi W} = \boldsymbol{RV}$, which consists of ℓ random rows sampled from $\boldsymbol{V}$. Theorem 3.1 cannot hold unless $\sigma_k(\boldsymbol{X}) > 0$. To ensure the latter event takes place, we must select at least one copy each of the k distinct rows of $\boldsymbol{W}$. This is the coupon collector's problem [Motwani and Raghavan (1995), Section 3.6]. To obtain a complete set of k rows with non-negligible probability, we must sample at least $k \log(k)$ rows. The fact that we are sampling without replacement does not improve the analysis appreciably because the matrix has too many rows.

Acknowledgment

This Research was supported by ONR award N00014-08-1-0883, DARPA award N66001-08-1-2065, and AFOSR award FA9550-09-1-0643.

References

Ailon, N. and Chazelle, B. (2009). The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, **31**(1): 302–322.

Ailon, N. and Liberty, E. (2010). Almost optimal unrestricted fast Johnson–Lindenstrauss transform, Available at arXiv:1005.5513.

Ahslwede, R. and Winter, A. (2002). Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, **48**(3): 569–579.

Gross, D. and Nesme, V. (2010). A note on sampling without replacement from a finite collection of matrices, Available at arXiv:1001.2738.

Halko, N., Martinsson, P.-G. and Tropp, J. A. (2009). Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions, ACM Report 2009–05, Caltech, Pasadena.

Halko, N., Martinsson, P.-G. and Tropp, J. A. (2011). Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, **53**(2): 217–288.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, **58**: 13–30.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.*, **26**: 189–206.

Krahmer, F. and Ward, R. (2010). New and improved Johnson–Lindenstrauss embeddings via the Restricted Isometry Property, Available at arXiv:1009.0744.

Ledoux, M. (1996). On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Stat.*, **1**: 63–87.

Ledoux, M. (2001). *The Concentration of Measure Phenomenon*, Number 89 in MSM. American Mathematical Society, Providence.

Liberty, E. (2009). *Accelerated Dense Random Projections*, Ph.D. Thesis, Computer Science Department Yale University, New Haven, CT.

Motwani, R. and Raghavan, P. (1995). *Randomized Algorithms*, Cambridge University Press.

Nguyen, N. H., Do, T. T. and Tran, T. D. (2009). A fast and efficient algorithm for low-rank approximation of a matrix. *STOC '09: Proceedings of 41st Annual ACM Symposium Theory of Computing*.

Rudelson, M. and Vershynin, R. (2007). Sampling from large matrices: an approach through geometric functional analysis. *J. Assoc. Comput. Mach.*, **54**(4): Art. 21, 19pp. (electronic).

Sarlós, T. (2006). Improved approximation algorithms for large matrices via random projections. *Proceedings of 47th Annual IEEE Symposium Foundations of Computer Science (FOCS)*, pp. 143–152.

Tropp, J. A. (2008). On the conditioning of random subdictionaries. *Appl. Comput. Harmon. Anal.*, **25**: 1–24.

Tropp, J. A. (2010). User-friendly tail bounds for sums of random matrices, ACM Report 2010–01, Caltech, Pasadena.

Woolfe, F., Liberty, E., Rokhlin, V. and Tygert, M. (2008). A fast randomized algorithm for the approximation of matrices. *Appl. Comput. Harmon. Anal.*, **25**(3): 335–366.