

IMPROVED ANALYSIS OF THE SUBSAMPLED RANDOMIZED HADAMARD TRANSFORM

JOEL A. TROPP

ABSTRACT. 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.

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 [JL84].

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 \mathbf{A} by a relatively small dimension-reduction matrix $\mathbf{\Omega}$ to obtain $\mathbf{Y} = \mathbf{A}\mathbf{\Omega}$. We then perform a QR factorization $\mathbf{Y} = \mathbf{Q}\mathbf{R}$ to identify the range of the reduced matrix \mathbf{Y} . For an appropriately designed dimension-reduction map, it can be shown that $\mathbf{A} \approx \mathbf{Q}\mathbf{Q}^*\mathbf{A}$ with high probability. In 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 \mathbf{A} by manipulating the low-rank approximation $\mathbf{Q}\mathbf{Q}^*\mathbf{A}$. See the recent survey [HMT11] for a comprehensive treatment of these ideas and an extensive bibliography.

In linear algebra applications, the cost of multiplication by an unstructured random matrix $\mathbf{\Omega}$, such as a Gaussian matrix, can sometimes be prohibitive. In that case, we may prefer to draw the random matrix $\mathbf{\Omega}$ from a *highly structured* distribution that allows us to multiply $\mathbf{\Omega}$ into the input matrix substantially faster. Sarlós is credited with bringing structured dimension reduction to numerical linear algebra [Sar06]; see also [WLRT08].

The *subsampled randomized Hadamard transform* (SRHT) is a type of structured dimension-reduction map that is based on the Walsh–Hadamard matrix. We will 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 [HMT11, Sec. 11] for details.

Date: 17 June 2010. Revised 5 November 2010 and 17 July 2011.

Key words and phrases. Dimension reduction, numerical linear algebra, random matrix, Walsh–Hadamard matrix. 2010 *Mathematics Subject Classification.* Primary: 15B52.

JAT is with Computing & Mathematical Sciences, MC 305-16, California Inst. Technology, Pasadena, CA 91125. E-mail: jtropp@cms.caltech.edu. Research supported by ONR award N00014-08-1-0883, DARPA award N66001-08-1-2065, and AFOSR award FA9550-09-1-0643.

This note will appear in *Advances in Adaptive Data Analysis*, special issue “Sparse Representation of Multiscale Data and Images.”

The literature already contains a number of papers, including [AC09, Lib09, NDT09, HMT11, AL11, KW11], 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 [AW02, Tro11]. 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 concrete performance guarantees.

This document is adapted from Appendix B of the technical report [HMT09], but much of the proof is new. This paper should be viewed as a codicil to the published version [HMT11] 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

$$\Phi = \sqrt{\frac{n}{\ell}} \cdot \mathbf{R} \mathbf{H} \mathbf{D}$$

where

- \mathbf{D} is a random $n \times n$ diagonal matrix whose entries are independent random signs, i.e., random variables uniformly distributed on $\{\pm 1\}$;
- \mathbf{H} is an $n \times n$ Walsh–Hadamard matrix, scaled by $n^{-1/2}$ so it is an orthogonal matrix;
- \mathbf{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 \mathbf{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; 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 [AC09]. Suppose that we want to estimate the energy (i.e., squared ℓ_2 norm) of a fixed vector \mathbf{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 \mathbf{x} has a few large components, the variance of this estimator is very high. On the other hand, when the components of \mathbf{x} have comparable magnitude, the estimator has much lower variance, so it is precise with very high probability.

The purpose of the matrix product $\mathbf{H}\mathbf{D}$ is to flatten out input vectors before we sample. To see why it achieves this goal, fix a unit vector \mathbf{x} , and examine the first component of $\mathbf{H}\mathbf{D}\mathbf{x}$.

$$(\mathbf{H}\mathbf{D}\mathbf{x})_1 = \sum_{j=1}^n h_{1j} \varepsilon_j x_j,$$

where h_{ij} are the components of the matrix \mathbf{H} . This random sum clearly has zero mean. Since the entries of \mathbf{H} have magnitude $n^{-1/2}$, the variance of the sum is n^{-1} . Hoeffding’s inequality [Hoe63] shows that

$$\mathbb{P}\{ |(\mathbf{H}\mathbf{D}\mathbf{x})_1| \geq t \} \leq 2e^{-nt^2/2}.$$

In words, the magnitude of the first component of $\mathbf{H}\mathbf{D}\mathbf{x}$ 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 $\mathbf{H}\mathbf{D}\mathbf{x}$ 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 \mathbf{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 \mathbf{D} from other subgaussian distributions. These changes complicate the analysis somewhat.

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 $p = O(e^k)$ 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 \mathbf{V} with orthonormal columns, and draw a random $\ell \times n$ SRHT matrix Φ where the embedding dimension ℓ satisfies*

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

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

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

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

The proof of Theorem 1.3 appears below. This result replaces Theorem B.4 of [HMT09]. Some precedents include [RV07, Thm. 3.1] and [Tro08, Sec. 9]; see also [NDT09]. 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 Section 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

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

coordinates is sufficient to ensure that $\Phi \mathbf{V}$ has constant condition number. See Theorem 3.2 for a more precise statement. The discussion in Section 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 j th 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.

Throughout this work, $\|\cdot\|$ refers to the ℓ_2 vector norm or the associated operator norm. We also write $\|\cdot\|_F$ for the Frobenius norm.

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

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

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 [Led96, Eqn. (1.9)]; see also [Led01, §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*

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

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

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

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 [AW02]. We derive the result using more recent ideas [Tro11], which deliver an essential improvement on the earlier work. The other key tool is a method [GN10], ultimately due to Hoeffding [Hoe63], for transferring results from the model where we sample *with* replacement to the model where we sample *without* replacement.

Theorem 2.2 (Matrix Chernoff). *Let \mathcal{X} be a finite set of positive-semidefinite matrices with dimension k , and suppose that*

$$\max_{\mathbf{X} \in \mathcal{X}} \lambda_{\max}(\mathbf{X}) \leq B.$$

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

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

Then

$$\begin{aligned} \mathbb{P}\left\{\lambda_{\min}\left(\sum_j \mathbf{X}_j\right) \leq (1-\delta)\mu_{\min}\right\} &\leq k \cdot \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/B} \quad \text{for } \delta \in [0, 1), \text{ and} \\ \mathbb{P}\left\{\lambda_{\max}\left(\sum_j \mathbf{X}_j\right) \geq (1+\delta)\mu_{\max}\right\} &\leq k \cdot \left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_{\max}/B} \quad \text{for } \delta \geq 0. \end{aligned}$$

Proof sketch. We establish only the upper bound; the lower bound is established by applying a similar method to $-\sum_j \mathbf{X}_j$. By homogeneity, take $B = 1$. We use the matrix Laplace transform method [Tro11, Prop. 3.1] to bound the probability of a large deviation:

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_j \mathbf{X}_j\right) \geq (1+t)\mu_{\max}\right\} \leq \inf_{\theta > 0} \left\{e^{-\theta(1+t)\mu_{\max}} \cdot \mathbb{E} \operatorname{tr} \exp\left(\sum_j \theta \mathbf{X}_j\right)\right\}. \quad (2.1)$$

The Laplace transform bound (2.1) is essentially due to Ahlswede and Winter [AW02].

Let $\{\mathbf{Y}_1, \dots, \mathbf{Y}_\ell\}$ be a random sample from \mathcal{X} drawn *with* replacement. Gross and Nesme [GN10] 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:

$$\mathbb{E} \operatorname{tr} \exp\left(\sum_j \theta \mathbf{X}_j\right) \leq \mathbb{E} \operatorname{tr} \exp\left(\sum_j \theta \mathbf{Y}_j\right). \quad (2.2)$$

Lemma 5.8 of [Tro11] gives a semidefinite bound for the mgf of \mathbf{Y}_j . (See also [AW02, Thm. 19].)

$$\mathbb{E} e^{\theta \mathbf{Y}_j} \preceq e^{(e^\theta - 1)(\mathbb{E} \mathbf{Y}_j)} = e^{(e^\theta - 1)(\mathbb{E} \mathbf{X}_1)} \quad \text{for } \theta \in \mathbb{R}. \quad (2.3)$$

On the right-hand side of this bound, we have exploited the fact that $\mathbf{Y}_j \sim \mathbf{Y}_1 \sim \mathbf{X}_1$. Lemma 3.4 of [Tro11] allows us to use the mgf bound (2.3) to bound the mgf of the entire sum:

$$\mathbb{E} \operatorname{tr} \exp \left(\sum_j \theta \mathbf{Y}_j \right) \leq \operatorname{tr} \exp \left((e^\theta - 1) \cdot \ell \cdot (\mathbb{E} \mathbf{X}_1) \right) \leq k \cdot \exp \left((e^\theta - 1) \cdot \mu_{\max} \right). \quad (2.4)$$

Substitute (2.4) into (2.2) and introduce the resulting inequality into the probability bound (2.1). The infimum is achieved at $\theta = \log(1 + \delta)$. \square

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 \mathbf{V} be an $n \times k$ matrix with orthonormal columns. Select a parameter ℓ that satisfies*

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

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

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

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 $\iota \leq c$. Let \mathbf{V} be an $n \times k$ matrix with orthonormal columns, where $k \geq C\iota^{-2} \log(n)$. Select a parameter ℓ that satisfies*

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

Draw an $\ell \times n$ SRHT Φ . Then

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

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

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 $\mathbf{H}\mathbf{D}$ equilibrates row norms.

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

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

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 \mathbf{W} be an $n \times k$ matrix with orthonormal columns, and define the quantity $M := n \cdot \max_{j=1, \dots, n} \|\mathbf{e}_j^* \mathbf{W}\|^2$. 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

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

with failure probability at most

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

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

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

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

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

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

except with probability $3k^{-1}$. Since $\Phi \mathbf{V} = \sqrt{n/\ell} \cdot \mathbf{R}_T \mathbf{W}$, 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 $\eta = e - 1$. 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 $\mathbf{H} \mathbf{D}$ balances row norms.

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

$$(\mathbf{H} \mathbf{D} \mathbf{V})^* (\mathbf{H} \mathbf{D} \mathbf{V}) = \mathbf{V}^* \mathbf{V} = \mathbf{I}$$

because \mathbf{D}, \mathbf{H} are orthogonal matrices.

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

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

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

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

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

$$f(\boldsymbol{\varepsilon}) = \|\mathbf{e}_j^* \mathbf{H} \mathbf{D} \mathbf{V}\|.$$

First, we bound the expectation.

$$\mathbb{E} f(\varepsilon) \leq [\mathbb{E} f(\varepsilon)^2]^{1/2} = \|\mathbf{E}\mathbf{V}\|_{\text{F}} \leq \|\mathbf{E}\| \|\mathbf{V}\|_{\text{F}} = \sqrt{\frac{k}{n}}.$$

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

$$\mathbb{P} \left\{ \|\mathbf{e}_j^*(\mathbf{H}\mathbf{D}\mathbf{V})\| \geq \sqrt{\frac{k}{n}} + \sqrt{\frac{8 \log(\beta n)}{n}} \right\} \leq e^{-8 \log(\beta n)/8} = \frac{1}{\beta n}.$$

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. \square

Now we establish the result on row sampling.

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

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

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

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 \mathbf{Y} as a sum of ℓ random matrices $\mathbf{X}_1, \dots, \mathbf{X}_\ell$ sampled without replacement from the family $\mathcal{X} := \{\mathbf{w}_j \mathbf{w}_j^* : j = 1, 2, \dots, n\}$ of positive semidefinite matrices with dimension k .

The matrix Chernoff bound, Theorem 2.2, allows us to obtain large deviation bounds for the extreme eigenvalues of \mathbf{Y} . We just need to determine the parameters involved in the statement. First, note that

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

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

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

As a consequence, we obtain

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

The lower Chernoff bound yields

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

The upper Chernoff bound yields

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

Substitute the identities

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

Simplify the formulae to complete the proof. \square

3.3. Collecting Coupons. The logarithmic factor in these results is *necessary*, as we show by example. This discussion is extracted from [HMT11, Sec. 11].

Fix an integer k , and set $n = k^2$. Form an $n \times k$ orthonormal matrix \mathbf{W} by regular decimation of the $n \times n$ identity matrix. More precisely, \mathbf{W} is the matrix whose j th 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 $\mathbf{V} = \mathbf{H}\mathbf{D}\mathbf{W}$. Observe that, up to scaling and modulation of rows, \mathbf{V} consists of k copies of a $k \times k$ Walsh–Hadamard transform stacked vertically.

Now, suppose that we apply an SRHT $\Phi = \mathbf{R}\mathbf{H}\mathbf{D}$ to the matrix \mathbf{W} . We obtain a matrix of the form $\mathbf{X} = \Phi\mathbf{W} = \mathbf{R}\mathbf{V}$, which consists of ℓ random rows sampled from \mathbf{V} . Theorem 3.1 cannot hold unless $\sigma_k(\mathbf{X}) > 0$. To ensure the latter event takes place, we must select at least one copy each of the k distinct rows of \mathbf{W} . This is the coupon collector’s problem [MR95, Sec. 3.6]. To obtain a complete set of k rows with nonnegligible 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.

REFERENCES

- [AC09] N. Ailon and B. Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 31(1):302–322, May 2009.
- [AL11] N. Ailon and E. Liberty. Almost optimal unrestricted fast Johnson–Lindenstrauss transform. In *Proc. 22nd Ann. ACM–SIAM Symp. Discrete Algorithms (SODA)*, pages 185–191, San Francisco, Jan. 2011.
- [AW02] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569–579, Mar. 2002.
- [GN10] D. Gross and V. Nesme. A note on sampling without replacement from a finite collection of matrices. Available at [arXiv:1001.2738](https://arxiv.org/abs/1001.2738), Jan. 2010.
- [HMT09] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. ACM Report 2009-05, Caltech, Pasadena, Sep. 2009.
- [HMT11] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, June 2011.
- [Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.
- [JL84] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.*, 26:189–206, 1984.
- [KW11] F. Kraahmer and R. Ward. New and improved Johnson–Lindenstrauss embeddings via the Restricted Isometry Property. *SIAM J. Math. Anal.*, 43(3):1269–1281, 2011.
- [Led96] M. Ledoux. On Talagrand’s deviation inequalities for product measures. *ESAIM Probab. Statist.*, 1:63–87, Jul. 1996.
- [Led01] M. Ledoux. *The Concentration of Measure Phenomenon*. Number 89 in MSM. American Mathematical Society, Providence, 2001.
- [Lib09] E. Liberty. *Accelerated dense random projections*. Ph.D. thesis, Computer Science Dept., Yale University, New Haven, 2009.
- [MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge Univ. Press, 1995.
- [NDT09] N. H. Nguyen, T. T. Do, and T. D. Tran. A fast and efficient algorithm for low-rank approximation of a matrix. In *STOC ’09: Proc. 41st Ann. ACM Symp. Theory of Computing*, 2009.
- [RV07] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. Assoc. Comput. Machinery*, 54(4):Art. 21, 19 pp., Jul. 2007. (electronic).
- [Sar06] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proc. 47th Ann. IEEE Symp. Foundations of Computer Science (FOCS)*, pages 143–152, 2006.
- [Tro08] J. A. Tropp. On the conditioning of random subdictionaries. *Appl. Comput. Harmon. Anal.*, 25:1–24, 2008.
- [Tro11] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 2011. To appear.
- [WLR08] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert. A fast randomized algorithm for the approximation of matrices. *Appl. Comput. Harmon. Anal.*, 25(3):335–366, 2008.