Very Sparse Random Projections

Ping Li
Department of Statistics
Stanford University
Stanford CA 94305, USA
pingli@stat.stanford.edu

Trevor J. Hastie
Department of Statistics
Stanford University
Stanford CA 94305, USA
hastie@stanford.edu

Kenneth W. Church
Microsoft Research
Microsoft Corporation
Redmond WA 98052, USA
church@microsoft.com

ABSTRACT

There has been considerable interest in random projections, an approximate algorithm for estimating distances between pairs of points in a high-dimensional vector space. Let $\mathbf{A} \in \mathbb{R}^{n \times D}$ be our n points in D dimensions. The method multiplies $\mathbf{A}$ by a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$, reducing the D dimensions down to just k for speeding up the computation. $\mathbf{R}$ typically consists of entries of standard normal $N(0,1)$. It is well known that random projections preserve pairwise distances (in the expectation). Achlioptas proposed *sparse random projections* by replacing the $N(0,1)$ entries in $\mathbf{R}$ with entries in $\{-1, 0, 1\}$ with probabilities $\{\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\}$, achieving a threefold speedup in processing time.

We recommend using $\mathbf{R}$ of entries in $\{-1, 0, 1\}$ with probabilities $\{\frac{1}{2\sqrt{D}}, 1-\frac{1}{\sqrt{D}}, \frac{1}{2\sqrt{D}}\}$ for achieving a significant $\sqrt{D}$-fold speedup, with little loss in accuracy.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

General Terms

Algorithms, Performance, Theory

Keywords

Random projections, Sampling, Rates of convergence

1. INTRODUCTION

Random projections [1, 43] have been used in Machine Learning [2, 4, 5, 13, 14, 22], VLSI layout [42], analysis of Latent Semantic Indexing (LSI) [35], set intersections [7, 36], finding motifs in bio-sequences [6, 27], face recognition [16], privacy preserving distributed data mining [31], to name a few. The AMS sketching algorithm [3] is also one form of random projections.

We define a data matrix $\mathbf{A}$ of size $n \times D$ to be a collection of n data points $\{u_i\}_{i=1}^n \in \mathbb{R}^D$. All pairwise distances can be computed as $\mathbf{A}\mathbf{A}^{\mathrm{T}}$, at the cost of time $O(n^2 D)$, which is often prohibitive for large n and D, in modern data mining and information retrieval applications.

To speed up the computations, one can generate a random projection matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ and multiply it with the original matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ to obtain a projected data matrix

$$\mathbf{B} = \frac{1}{\sqrt{k}} \mathbf{A}\mathbf{R} \in \mathbb{R}^{n \times k}, \qquad k \ll \min(n, D). \tag{1}$$

The (much smaller) matrix $\mathbf{B}$ preserves all pairwise distances of $\mathbf{A}$ in expectations, provided that $\mathbf{R}$ consists of i.i.d. entries with zero mean and constant variance. Thus, we can achieve a substantial cost reduction for computing $\mathbf{A}\mathbf{A}^{\mathrm{T}}$, from $O(n^2 D)$ to $O(nDk + n^2 k)$.

In information retrieval, we often do not have to materialize $\mathbf{A}\mathbf{A}^{\mathrm{T}}$. Instead, databases and search engines are interested in storing the projected data $\mathbf{B}$ in main memory for efficiently responding to input queries. While the original data matrix $\mathbf{A}$ is often too large, the projected data matrix $\mathbf{B}$ can be small enough to reside in the main memory.

The entries of $\mathbf{R}$ (denoted by $\{r_{ji}\}_{j=1}^{D}{}_{i=1}^{k}$) should be i.i.d. with zero mean. In fact, this is the only necessary condition for preserving pairwise distances [4]. However, different choices of r_{ji} can change the variances (average errors) and error tail bounds. It is often convenient to let r_{ji} follow a symmetric distribution about zero with unit variance. A "simple" distribution is the standard normal[1], i.e.,

$$r_{ji} \sim N(0,1), \quad \mathrm{E}(r_{ji}) = 0, \quad \mathrm{E}\left(r_{ji}^2\right) = 1, \quad \mathrm{E}\left(r_{ji}^4\right) = 3.$$

It is "simple" in terms of theoretical analysis, but not in terms of random number generation. For example, a uniform distribution is easier to generate than normals, but the analysis is more difficult.

In this paper, when $\mathbf{R}$ consists of normal entries, we call this special case as the *conventional random projections*, about which many theoretical results are known. See the monograph by Vempala [43] for further references.

We derive some theoretical results when $\mathbf{R}$ is not restricted to normals. In particular, our results lead to significant improvements over the so-called *sparse random projections*.

1.1 Sparse Random Projections

In his novel work, Achlioptas [1] proposed using the pro-

[1]The normal distribution is *2-stable*. It is one of the few stable distributions that have closed-form density [19].

KDD'06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.

jection matrix $\mathbf{R}$ with i.i.d entries in

$$r_{ji} = \sqrt{s} \begin{cases} 1 & \text{with prob. } \frac{1}{2s} \\ 0 & \text{with prob. } 1 - \frac{1}{s} \\ -1 & \text{with prob. } \frac{1}{2s} \end{cases}, \quad (2)$$

where Achlioptas used $s = 1$ or $s = 3$. With $s = 3$, one can achieve a threefold speedup because only $\frac{1}{3}$ of the data need to be processed (hence the name *sparse random projections*). Since the multiplications with $\sqrt{s}$ can be delayed, no floating point arithmetic is needed and all computation amounts to highly optimized database aggregation operations.

This method of *sparse random projections* has gained its popularity. It was first experimentally tested on image and text data by [5] in SIGKDD 2001. Later, many more publications also adopted this method, e.g., [14, 29, 38, 41].

1.2 Very Sparse Random Projections

We show that one can use $s \gg 3$ (e.g., $s = \sqrt{D}$, or even $s = \frac{D}{\log D}$) to significantly speed up the computation.

Examining (2), we can see that *sparse random projections* are random sampling at a rate of $\frac{1}{s}$, i.e., when $s = 3$, one-third of the data are sampled. Statistical results tell us that one does not have to sample one-third ($D/3$) of the data to obtain good estimates. In fact, when the data are approximately normal, $\log D$ of the data probably suffice (i.e., $s = \frac{D}{\log D}$), because of the exponential error tail bounds, common in normal-like distributions, such as binomial, gamma, etc. For better robustness, we recommend choosing s less aggressively (e.g., $s = \sqrt{D}$).

To better understand sparse and very sparse random projections, we first give a summary of relevant results on *conventional random projections*, in the next section.

2. CONVENTIONAL RANDOM PROJECTIONS: $\mathbf{R} \sim N(0, 1)$

Conventional random projections multiply the original data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ with a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$, consisting of i.i.d. $N(0, 1)$ entries. Denote by $\{u_i\}_{i=1}^n \in \mathbb{R}^D$ the rows in $\mathbf{A}$ and by $\{v_i\}_{i=1}^n \in \mathbb{R}^k$ the rows of the projected data, i.e., $v_i = \frac{1}{\sqrt{k}} \mathbf{R}^{\mathrm{T}} u_i$. We focus on the leading two rows: u_1, u_2 and v_1, v_2. For convenience, we denote

$$m_1 = \|u_1\|^2 = \sum_{j=1}^D u_{1,j}^2, \qquad m_2 = \|u_2\|^2 = \sum_{j=1}^D u_{2,j}^2,$$

$$a = u_1^{\mathrm{T}} u_2 = \sum_{j=1}^D u_{1,j} u_{2,j}, \; d = \|u_1 - u_2\|^2 = m_1 + m_2 - 2a.$$

2.1 Moments

It is easy to show that (e.g., Lemma 1.3 of [43])

$$\mathrm{E}\left(\|v_1\|^2\right) = \|u_1\|^2 = m_1, \quad \mathrm{Var}\left(\|v_1\|^2\right)_N = \frac{2}{k} m_1^2, \quad (3)$$

$$\mathrm{E}\left(\|v_1 - v_2\|^2\right) = d, \qquad \mathrm{Var}\left(\|v_1 - v_2\|^2\right)_N = \frac{2}{k} d^2, \quad (4)$$

where the subscript "N" indicates that a "normal" projection matrix is used.

From our later results in Lemma 3 (or [28, Lemma 1]) we can derive

$$\mathrm{E}\left(v_1^{\mathrm{T}} v_2\right) = a, \qquad \mathrm{Var}\left(v_1^{\mathrm{T}} v_2\right)_N = \frac{1}{k}\left(m_1 m_2 + a^2\right). \quad (5)$$

Therefore, one can compute both pairwise 2-norm distances and inner products in k (instead of D) dimensions, achieving a huge cost reduction when $k \ll \min(n, D)$.

2.2 Distributions

It is easy to show that (e.g. Lemma 1.3 of [43])

$$\frac{v_{1,i}}{\sqrt{m_1/k}} \sim N(0, 1), \qquad \frac{\|v_1\|^2}{m_1/k} \sim \chi_k^2, \quad (6)$$

$$\frac{v_{1,i} - v_{2,i}}{\sqrt{d/k}} \sim N(0, 1), \qquad \frac{\|v_1 - v_2\|^2}{d/k} \sim \chi_k^2, \quad (7)$$

$$\begin{bmatrix} v_{1,i} \\ v_{2,i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \frac{1}{k}\begin{bmatrix} m_1 & a \\ a & m_2 \end{bmatrix}\right). \quad (8)$$

where χ_k^2 denotes a chi-squared random variable with k degrees of freedom. $v_{1,i}$ i.i.d. is any entry in $v_1 \in \mathbb{R}^k$.

Knowing the distributions of the projected data enables us to derive (sharp) error tail bounds. For example, various Johnson and Lindenstrauss (JL) embedding theorems [1,4,9, 15,20,21] have been proved for precisely determining k given some specified level of accuracy, for estimating the 2-norm distances. According to the best known result [1]:

If $k \geq k_0 = \frac{4+2\gamma}{\epsilon^2/2 - \epsilon^3/3} \log n$, then with probability at least $1 - n^{-\gamma}$, for any two rows u_i, u_j, we have

$$(1 - \epsilon)\|u_i - u_j\|^2 \leq \|v_i - v_j\|^2 \leq (1 + \epsilon)\|u_i - u_j\|^2. \quad (9)$$

Remark: (a) The JL lemma is conservative in many applications because it was derived based on Bonferroni correction for multiple comparisons. (b) It is only for the l_2 distance, while many applications care more about the inner product. As shown in (5), the variance of the inner product estimator, $\mathrm{Var}\left(v_1^{\mathrm{T}} v_2\right)_N$, is dominated by the margins (i.e., $m_1 m_2$) even when the data are uncorrelated. This is probably the weakness of random projections.

2.3 Sign Random Projections

A popular variant of *conventional random projections* is to store only the signs of the projected data, from which one can estimate the vector cosine angles, $\theta = \cos^{-1}\left(\frac{a}{\sqrt{m_1 m_2}}\right)$, by the following result [7, 17]:

$$\mathbf{Pr}\left(\mathrm{sign}(v_{1,i}) = \mathrm{sign}(v_{2,i})\right) = 1 - \frac{\theta}{\pi}, \quad (10)$$

One can also estimate a by assuming that m_1, m_2 are known, from $a = \cos(\theta)\sqrt{m_1 m_2}$, at the cost of some bias.

The advantage of sign random projections is the saving in storing the projected data because only one bit is needed for the sign. With sign random projections, we can compare vectors using hamming distances for which efficient algorithms are available [7,20,36]. See [28] for more comments on sign random projections.

3. OUR CONTRIBUTIONS

We propose *very sparse random projections* to speed up the (processing) computations by a factor of $\sqrt{D}$ or more.

- We derive exact variance formulas for $\|v_1\|^2$, $\|v_1 - v_2\|^2$, and $v_1^{\mathrm{T}} v_2$ as functions of s.[2] Under reasonable regularity conditions, they converge to the corresponding variances when $r_{ji} \sim N(0, 1)$ is used, as long as $s = o(D)$

[2] [1] proved the upper bounds for the variances of $\|v_1\|^2$ and $\|v_1 - v_2\|^2$ for $s = 1$ and $s = 3$.

(e.g., $s = \sqrt{D}$, or even $s = \frac{D}{\log D}$). When $s = \sqrt{D}$, the rate of convergence is $O\left(\frac{1}{D^{1/4}}\right)$, which is fast since D has to be large otherwise there would be no need of seeking approximate answers. This means we can achieve a $\sqrt{D}$-fold speedup with little loss in accuracy.

- We show that $v_{1,i}$, $v_{1,i} - v_{2,i}$ and $(v_{1,i}, v_{2,i})$ converge to normals at the rate $O\left(\frac{1}{D^{1/4}}\right)$ when $s = \sqrt{D}$. This allows us to apply, with a high level of accuracy, results of *conventional random projections*, e.g., the JL-embedding theorem in (9) and the sign random projections in (10). In particular, we suggest using a maximum likelihood estimator of the asymptotic (normal) distribution to estimate the inner product $a = u_1^{\mathrm{T}} u_2$, taking advantage of the marginal norms m_1, m_2.
- Our results essentially hold for any other distributions of r_{ji}. When r_{ji} is chosen to have negative kurtosis, we can achieve strictly smaller variances (errors) than *conventional random projections*.

4. MAIN RESULTS

Main results of our work are presented in this section with detailed proofs in Appendix A. For convenience, we always let $s = o(D)$ (e.g., $s = \sqrt{D}$) and assume all *fourth* moments are bounded, e.g., $\mathrm{E}(u_{1,j}^4) < \infty$, $\mathrm{E}(u_{2,j}^4) < \infty$ and $\mathrm{E}(u_{1,j}^2 u_{2,j}^2) < \infty$. In fact, analyzing the rate of convergence of asymptotic normality only requires bounded *third* moments and an even much weaker assumption is needed for ensuring asymptotic normality. Later we will discuss the possibility of relaxing this assumption of bounded moments.

4.1 Moments

The first three lemmas concern the moments (means and variances) of v_1, $v_1 - v_2$ and $v_1^{\mathrm{T}} v_2$, respectively.

LEMMA 1.

$$E\left(\|v_1\|^2\right) = \|u_1\|^2 = m_1, \tag{11}$$

$$Var\left(\|v_1\|^2\right) = \frac{1}{k}\left(2m_1^2 + (s-3)\sum_{j=1}^{D} u_{1,j}^4\right). \tag{12}$$

As $D \to \infty$,

$$\frac{(s-3)\sum_{j=1}^{D}\left(u_{1,j}\right)^4}{m_1^2} \to \frac{s-3}{D}\frac{E\left(u_{1,j}\right)^4}{E^2\left(u_{1,j}\right)^2} \to 0, \tag{13}$$

$$i.e., \quad Var\left(\|v_1\|^2\right) \overset{D}{\sim} \frac{1}{k}\left(2m_1^2\right). \tag{14}$$

$\overset{D}{\sim}$ *denotes "asymptotically equivalent" for large* D.

Note that $m_1^2 = \left(\sum_{j=1}^{D} u_{1,j}^2\right)^2 = \sum_{j=1}^{D} u_{1,j}^4 + \sum_{j \neq j'} u_{1,j}^2 u_{1,j'}^2$, with D diagonal terms and $\frac{D(D-1)}{2}$ cross-terms. When all dimensions of u_1 are roughly equally important, the cross-terms dominate. Since D is very large, the diagonal terms are negligible. However, if a few entries are extremely large compared to the majority of the entries, the cross-terms may be of the same order as the diagonal terms. Assuming bounded fourth moment prevents this from happening.

The next Lemma is strictly analogous to Lemma 1. We present them separately because Lemma 1 is more convenient to present and analyze, while Lemma 2 contains the results on the 2-norm distances, which we will use.

LEMMA 2.

$$E\left(\|v_1 - v_2\|^2\right) = \|u_1 - u_2\|^2 = d, \tag{15}$$

$$Var\left(\|v_1 - v_2\|^2\right)$$

$$= \frac{1}{k}\left(2d^2 + (s-3)\sum_{j=1}^{D}\left(u_{1,j} - u_{2,j}\right)^4\right) \tag{16}$$

$$\overset{D}{\sim} \frac{1}{k}\left(2d^2\right). \tag{17}$$

The third lemma concerns the inner product.

LEMMA 3.

$$E\left(v_1^T v_2\right) = u_1^T u_2 = a, \tag{18}$$

$$Var\left(v_1^T v_2\right)$$

$$= \frac{1}{k}\left(m_1 m_2 + a^2 + (s-3)\sum_{j=1}^{D} u_{1,j}^2 u_{2,j}^2\right). \tag{19}$$

$$\overset{D}{\sim} \frac{1}{k}\left(m_1 m_2 + a^2\right). \tag{20}$$

Therefore, *very sparse random projections* preserve pairwise distances in expectations with variances as functions of s. Compared with $\mathrm{Var}(\|v_1\|^2)_N$, $\mathrm{Var}(\|v_1 - v_2\|^2)_N$, and $\mathrm{Var}(v_1^{\mathrm{T}} v_2)_N$ in (3), (4), and (5), respectively, the extra terms all involve $(s-3)$ and are asymptotically negligible. The rate of convergence is $O\left(\sqrt{\frac{s-3}{D}}\right)$, in terms of the standard error (square root of variance). When $s = \sqrt{D}$, the rate of convergence is $O\left(\frac{1}{D^{1/4}}\right)$.

When $s < 3$, "sparse" random projections can actually achieve slightly smaller variances.

4.2 Asymptotic Distributions

The asymptotic analysis provides a feasible method to study distributions of the projected data.

The task of analyzing the distributions is easy when a normal random matrix $\mathbf{R}$ is used. The analysis for other types of random projection distributions is much more difficult (in fact, intractable). To see this, each entry $v_{1,i} = \frac{1}{\sqrt{k}}\mathbf{R}_i^{\mathrm{T}} u_1 = \frac{1}{\sqrt{k}}\sum_{j=1}^{D} r_{ji} u_{1,j}$. Other than the case $r_{ji} \sim N(0,1)$, analyzing $v_{1,i}$ and v_1 exactly is basically impossible, although in some simple cases [1] we can study the bounds of the moments and moment generating functions.

Lemma 4 and Lemma 5 present the asymptotic distributions of v_1 and $v_1 - v_2$, respectively. Again, Lemma 5 is strictly analogous to Lemma 4.

LEMMA 4. *As* $D \to \infty$,

$$\frac{v_{1,i}}{\sqrt{m_1/k}} \overset{\mathcal{L}}{\Longrightarrow} N(0,1), \qquad \frac{\|v_1\|^2}{m_1/k} \overset{\mathcal{L}}{\Longrightarrow} \chi_k^2, \tag{21}$$

with the rate of convergence

$$\left|F_{v_{1,i}}(y) - \Phi(y)\right| \leq 0.8\sqrt{s}\frac{\sum_{j=1}^{D}\left|u_{1,j}\right|^3}{m_1^{3/2}} \to 0.8\sqrt{\frac{s}{D}}\frac{E|u_{1,j}|^3}{\left(E\left(u_{1,j}^2\right)\right)^{3/2}} \to 0, \tag{22}$$

where $\overset{\mathcal{L}}{\Longrightarrow}$ *denotes "convergence in distribution;"* $F_{v_{1,i}}(y)$ *is the empirical cumulative density function (CDF) of* $v_{1,i}$ *and* $\Phi(y)$ *is the standard normal* $N(0,1)$ *CDF.*

LEMMA 5. *As $D \to \infty$,*

$$\frac{v_{1,i} - v_{2,i}}{\sqrt{d/k}} \xrightarrow{\mathcal{L}} N(0,1), \qquad \frac{\|v_1 - v_2\|^2}{d/k} \xrightarrow{\mathcal{L}} \chi^2_k, \tag{23}$$

with the rate of convergence

$$|F_{v_{1,i}-v_{2,i}}(y) - \Phi(y)| \leq 0.8\sqrt{s}\frac{\sum_{j=1}^{D}|u_{1,j} - u_{2,j}|^3}{d^{3/2}} \to 0. \tag{24}$$

The above two lemmas show that both $v_{1,i}$ and $v_{1,i} - v_{2,i}$ are approximately normal, with the rate of convergence determined by $\sqrt{s/D}$, which is $O\left(\frac{1}{D^{1/4}}\right)$ when $s = \sqrt{D}$.

The next lemma concerns the joint distribution of $(v_{1,i}, v_{2,i})$.

LEMMA 6. *As $D \to \infty$,*

$$\Sigma^{-\frac{1}{2}} \begin{bmatrix} v_{1,i} \\ v_{2,i} \end{bmatrix} \xrightarrow{\mathcal{L}} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right), \tag{25}$$

and

$$\mathbf{Pr}\left(sign(v_{1,i}) = sign(v_{2,i})\right) \to 1 - \frac{\theta}{\pi}. \tag{26}$$

where

$$\Sigma = \frac{1}{k}\begin{bmatrix} m_1 & a \\ a & m_2 \end{bmatrix}, \qquad \theta = \cos^{-1}\left(\frac{a}{\sqrt{m_1 m_2}}\right).$$

The asymptotic normality shows that we can use other random projections matrix $\mathbf{R}$ to achieve asymptotically the same performance as *conventional random projections*, which are the easiest to analyze. Since the convergence rate is so fast, we can simply apply results on *conventional random projections* such as the JL lemma and sign random projections when a non-normal projection matrix is used.[3]

4.3 A Margin-free Estimator

Recall that, because $\mathrm{E}(v_1^{\mathrm{T}} v_2) = u_1^{\mathrm{T}} u_2$, one can estimate $a = u_1^{\mathrm{T}} u_2$ without bias as $\hat{a}_{MF} = v_1^{\mathrm{T}} v_2$, with the variance

$$\mathrm{Var}\left(\hat{a}_{MF}\right) = \frac{1}{k}\left(m_1 m_2 + a^2 + (s-3)\sum_{j=1}^{D} u_{1,j}^2 u_{2,j}^2\right), \tag{27}$$

$$\mathrm{Var}\left(\hat{a}_{MF}\right)_\infty = \frac{1}{k}\left(m_1 m_2 + a^2\right), \tag{28}$$

where the subscript "MF" indicates "Margin-free," i.e., an estimator of a without using margins. $\mathrm{Var}\left(\hat{a}_{MF}\right)$ is the variance of $v_1^{\mathrm{T}} v_2$ in (19). Ignoring the asymptotically negligible part involving $s - 3$ leads to $\mathrm{Var}\left(\hat{a}_{MF}\right)_\infty$.

We will compare $\hat{a}_{MF}$ with an asymptotic maximum likelihood estimator based on the asymptotic normality.

4.4 An Asymptotic MLE Using Margins

The tractable asymptotic distributions of the projected data allow us to derive more accurate estimators using maximum likelihood.

In many situations, we can assume that the marginal norms $m_1 = \sum_{j=1}^{D} u_{1,j}^2$ and $m_2 = \sum_{j=1}^{D} u_{2,j}^2$ are known, as m_1 and m_2 can often be easily either exactly calculated or accurately estimated.[4]

The authors' very recent work [28] on *conventional random projections* shows that if we know the margins m_1 and m_2, we can estimate $a = u_1^{\mathrm{T}} u_2$ often more accurately using a maximum likelihood estimator (MLE).

The following lemma estimates $a = u_1^{\mathrm{T}} u_2$, taking advantage of knowing the margins.

LEMMA 7. *When the margins, m_1 and m_2 are known, we can use a maximum likelihood estimator (MLE) to estimate a by maximizing the joint density function of (v_1, v_2). Since $(v_{1,i}, v_{2,i})$ converges to a bivariate normal, an asymptotic MLE is the solution to a cubic equation*

$$a^3 - a^2\left(v_1^T v_2\right) + a\left(-m_1 m_2 + m_1\|v_2\|^2 + m_2\|v_1\|^2\right) - m_1 m_2 v_1^T v_2 = 0. \tag{29}$$

The asymptotic variance of this estimator, denoted by $\hat{a}_{MLE}$, is

$$Var\left(\hat{a}_{MLE}\right)_\infty = \frac{1}{k}\frac{\left(m_1 m_2 - a^2\right)^2}{m_1 m_2 + a^2} \leq Var\left(\hat{a}_{MF}\right)_\infty. \tag{30}$$

The ratio $\frac{\mathrm{Var}(\hat{a}_{MLE})_\infty}{\mathrm{Var}(\hat{a}_{MF})_\infty} = \frac{\left(m_1 m_2 - a^2\right)^2}{\left(m_1 m_2 + a^2\right)^2} = \frac{(1-cos^2(\theta))^2}{(1+cos^2(\theta))^2}$ ranges from 0 to 1, indicating possibly substantial improvements. For example, when $\cos(\theta) \approx 1$ (i.e., $a^2 \approx m_1 m_2$), the improvement will be huge. When $\cos(\theta) \approx 0$ (i.e., $a \approx 0$), we do not benefit from $\hat{a}_{MLE}$. Note that some studies (e.g., duplicate detection) are mainly interested in data points that are quite similar (i.e., $\cos(\theta)$ close to 1).

4.5 The Kurtosis of r_{ji} : $(s-3)$

We have seen that the parameter s plays an important role in the performance of *very sparse random projections*. It is interesting that $s - 3$ is exactly the kurtosis of r_{ji}:

$$\gamma_2(r_{ji}) = \frac{\mathrm{E}((r_{ji} - \mathrm{E}(r_{ji}))^4)}{\mathrm{E}^2((r_{ji} - \mathrm{E}(r_{ji}))^2)} - 3 = s - 3, \tag{31}$$

as r_{ji} has zero mean and unit variance.[5]

The kurtosis for $r_{ji} \sim N(0,1)$ is zero. If one is only interested in smaller estimation variances (ignoring the benefit of sparsity), one may choose the distribution of r_{ji} with negative kurtosis. A couple of examples are

- A continuous uniform distribution in $[-l, l]$ for any $l > 0$. It's kurtosis $= -\frac{6}{5}$.
- A discrete uniform distribution symmetric about zero, with N points. Its kurtosis $= -\frac{6}{5}\frac{N^2+1}{N^2-1}$, ranging between -2 (when $N = 2$) and $-\frac{6}{5}$ (when $N \to \infty$). The case with $N = 2$ is the same as (2) with $s = 1$.
- Discrete and continuous U-shaped distributions.

[3] In the proof of the asymptotic normality, we used $\mathrm{E}(|r_{ji}|^3)$ and $\mathrm{E}(|r_{ji}|^{2+\delta})$. They should be replaced by the corresponding moments when other projection distributions are used.

[4] Computing all marginal norms of $\mathbf{A}$ costs $O(nD)$, which is often negligible. As important summary statistics, the marginal norms may be already computed during various stage of processing, e.g., normalization and term weighting.

[5] Note that the kurtosis can not be smaller than -2 because of the Cauchy-Schwarz inequality: $\mathrm{E}^2(r_{ji}^2) \leq \mathrm{E}(r_{ji}^4)$. One may consult *http://en.wikipedia.org/wiki/Kurtosis* for references to kurtosis of various distributions.

5. HEAVY-TAIL AND TERM WEIGHTING

The *very sparse random projections* are useful even for heavy-tailed data, mainly because of *term weighting*.

We have seen that bounded *forth* and *third* moments are needed for analyzing the convergence of moments (variance) and the convergence to normality, respectively. The proof of asymptotic normality in Appendix A suggests that we only need stronger than bounded *second* moments to ensure asymptotic normality. In heavy-tailed data, however, even the second moment may not exist.

Heavy-tailed data are ubiquitous in large-scale data mining applications (especially Internet data) [25,34]. The pairwise distances computed from heavy-tailed data are usually dominated by "outliers," i.e., exceptionally large entries.

Pairwise vector distances are meaningful only when all dimensions of the data are more or less equally important. For heavy-tailed data, such as the (unweighted) term-by-document matrix, pairwise distances may be misleading. Therefore, in practice, various term weighting schemes are proposed e.g., [33, Chapter 15.2] [10, 30, 39, 45], to weight the entries instead of using the original data.

It is well-known that choosing an appropriate term weighting method is vital. For example, as shown in [23, 26], in text categorization using support vector machine (SVM), choosing an appropriate term weighting scheme is far more important than tuning kernel functions of SVM. See similar comments in [37] for the work on Naive Bayes text classifier.

We list two popular and simple weighting schemes. One variant of the *logarithmic weighting* keeps zero entries and replaces any non-zero count with 1+log(original count). Another scheme is the *square root weighting*. In the same spirit of the Box-Cox transformation [44, Chapter 6.8], these various weighting schemes significantly reduce the kurtosis (and skewness) of the data and make the data resemble normal.

Therefore, it is fair to say that assuming finite moments (third or fourth) is reasonable whenever the computed distances are meaningful.

However, there are also applications in which pairwise distances do not have to bear any clear meaning. For example, using random projections to estimate the joint sizes (set intersections). If we expect the original data are severely heavy-tailed and no term weighting will be applied, we recommend using $s = O(1)$.

Finally, we shall point out that *very sparse random projections* can be fairly robust against heavy-tailed data when $s = \sqrt{D}$. For example, instead of assuming finite fourth moments, as long as $D\frac{\sum_{j=1}^D u_{1,j}^4}{\left(\sum_{j=1}^D u_{1,j}^2\right)^2}$ grows slower than $O(\sqrt{D})$, we can still achieve the convergence of variances if $s = \sqrt{D}$, in Lemma 1. Similarly, analyzing the rate of converge to normality only requires that $\sqrt{D}\frac{\sum_{j=1}^D |u_{1,j}|^3}{\left(\sum_{j=1}^D u_{1,j}^2\right)^{3/2}}$ grows slower than $O(D^{1/4})$. An even weaker condition is needed to only ensure asymptotic normality. We provide some additional analysis on heavy-tailed data in Appendix B.

6. EXPERIMENTAL RESULTS

Some experimental results are presented as a sanity check, using one pair of words, "THIS" and "HAVE," from two rows of a term-by-document matrix provided by MSN. $D = 2^{16} = 65536$. That is, $u_{1,j}$ ($u_{2,j}$) is the number of occurrences of word "THIS" (word "HAVE") in the jth document (Web page), $j = 1$ to D. Some summary statistics are listed in Table 1.

The data are certainly heavy-tailed as the kurtoses for $u_{1,j}$ and $u_{2,j}$ are 195 and 215, respectively, far above zero. Therefore we do not expect that *very sparse random projections* with $s = \frac{D}{\log D} \approx 6000$ work well, though the results are actually not disastrous as shown in Figure 1(d).

Table 1: Some summary statistics of the word pair, "THIS" (u_1) and "HAVE" (u_2). γ_2 denotes the kurtosis. $\eta(u_{1,j}, u_{2,j}) = \frac{\mathbf{E}(u_{1,j}^2 u_{2,j}^2)}{\mathbf{E}(u_{1,j}^2)\mathbf{E}(u_{2,j}^2)+\mathbf{E}^2(u_{1,j}u_{2,j})}$, affects the convergence of Var $\left(v_1^{\mathbf{T}} v_2\right)$ (see the proof of Lemma 3). These expectations are computed empirically from the data. Two popular term weighting schemes are applied. The "square root weighting" replaces $u_{1,j}$ with $\sqrt{u_{1,j}}$ and the "logarithmic weighting" replaces any non-zero $u_{1,j}$ with $1 + \log u_{1,j}$.

	Unweighted	Square root	Logarithmic
$\gamma_2(u_{1,j})$	195.1	13.03	1.58
$\gamma_2(u_{2,j})$	214.7	17.05	4.15
$\frac{\mathrm{E}(u_{1,j}^4)}{\mathrm{E}^2(u_{1,j}^2)}$	180.2	12.97	5.31
$\frac{\mathrm{E}(u_{2,j}^4)}{\mathrm{E}^2(u_{2,j}^2)}$	205.4	18.43	8.21
$\eta(u_{1,j}, u_{2,j})$	78.0	7.62	3.34
$\cos(\theta(u_1, u_2))$	0.794	0.782	0.754

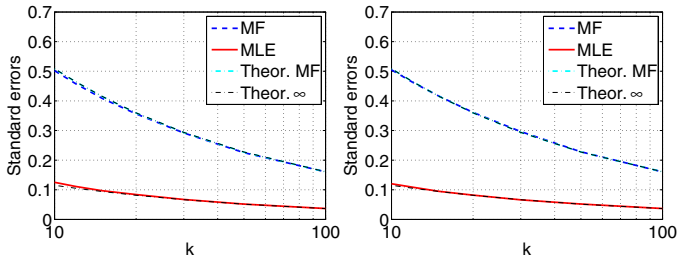
We first test random projections on the original (unweighted, heavy-tailed) data, for $s = 1, 3, 256 = \sqrt{D}$ and $6000 \approx \frac{D}{\log D}$, presented in Figure 1. We then apply square root weighting and logarithmic weighting before random projections. The results are presented in Figure 2, for $s = 256$ and $s = 6000$.

These results are consistent with what we would expect:

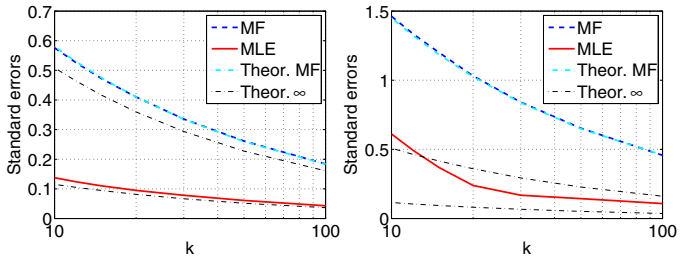
- When s is small, i.e., $O(1)$, sparse random projections perform very similarly to conventional random projections as shown in panels (a) and (b) of Figure 1 .
- With increasing s, the variances of sparse random projections increase. With $s = \frac{D}{\log D}$, the errors are large (but not disastrous), because the data are heavy-tailed. With $s = \sqrt{D}$, sparse random projections are robust.
- Since $\cos(\theta(u_1, u_2)) \approx 0.7 \sim 0.8$ in this case, marginal information can improve the estimation accuracy quite substantially. The asymptotic variances of $\hat{a}_{MLE}$ match the empirical variances of the asymptotic MLE estimator quite well, even for $s = \sqrt{D}$.
- After applying term weighting on the original data, sparse random projections are almost as accurate as conventional random projections, even for $s \approx \frac{D}{\log D}$, as shown in Figure 2.

7. CONCLUSION

We provide some new theoretical results on random projections, a randomized approximate algorithm widely used in machine learning and data mining. In particular, our theoretical results suggest that we can achieve a significant $\sqrt{D}$-fold speedup in processing time with little loss in accuracy, where D is the original data dimension. When the data

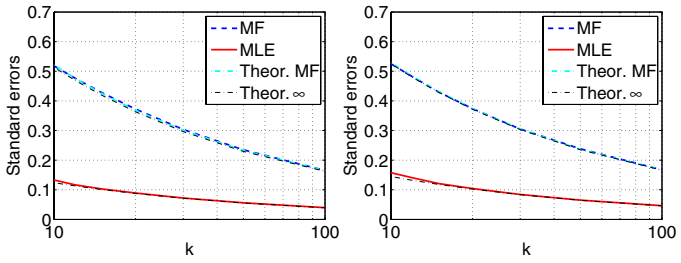


(a) $s = 1$ (b) $s = 3$

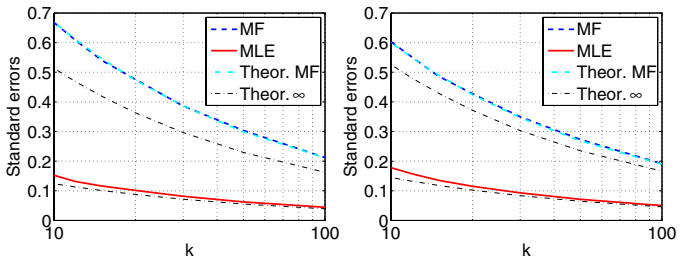


(c) $s = 256$ (d) $s = 6000$

Figure 1: Two words "THIS" (u_1) and "HAVE" (u_2) from the MSN Web crawl data are tested. $D = 2^{16}$. Sparse random projections are applied to estimated $a = u_1^T u_2$, with four values of s: 1, 3, 256 $= \sqrt{D}$ and 6000 $\approx \frac{D}{\log D}$, in panels (a), (b), (c) and (d), respectively, presented in terms of the normalized standard error, $\frac{\sqrt{\mathbf{Var}(\hat{a})}}{a}$. 10^4 simulations are conducted for each k, ranging from 10 to 100. There are five curves in each panel. The two labeled as "MF" and "Theor." overlap. "MF" stands for the empirical variance of the "Margin-free" estimator $\hat{a}_{MF}$; while "Theor. MF" for the theoretical variance of $\hat{a}_{MF}$, i.e., (27). The solid curve, labeled as "MLE," presents the empirical variance of $\hat{a}_{MLE}$, the estimator using margins as formulated in Lemma 7. There are two curves both labeled as "Theor. ∞," for the asymptotic theoretical variances of $\hat{a}_{MF}$ (the higher curve, (28)) and $\hat{a}_{MLE}$ (the lower curve, (30)).



(a) Square root ($s = 256$) (b) Logarithmic ($s = 256$)



(c) Square root ($s = 6000$) (d) Logarithmic ($s = 6000$)

Figure 2: After applying term weighting on the original data, sparse random projections are almost as accurate as conventional random projections, even for $s = 6000 \approx \frac{D}{\log D}$. Note that the legends are the same as in Figure 1.

are free of "outliers" (e.g., after careful term weighting), a cost reduction by a factor of $\frac{D}{\log D}$ is also possible.

Our proof of the asymptotic normality justifies the use of an asymptotic maximum likelihood estimator for improving the estimates when the marginal information is available.

8. ACKNOWLEDGMENT

We thank Dimitris Achlioptas for very insightful comments. We thank Xavier Gabaix and David Mason for pointers to useful references. Ping Li thanks the enjoyable and helpful conversations with Tze Leung Lai, Joseph P. Romano, and Yiyuan She. Finally, we thank the four anonymous reviewers for constructive suggestions.

9. REFERENCES

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.

[2] Dimitris Achlioptas, Frank McSherry, and Bernhard Schölkopf. Sampling techniques for kernel methods. In *Proc. of NIPS*, pages 335–342, Vancouver, BC, Canada, 2001.

[3] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proc. of STOC*, pages 20–29, Philadelphia,PA, 1996.

[4] Rosa Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *Proc. of FOCS (Also to appear in Machine Learning)*, pages 616–623, New York, 1999.

[5] Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *Proc. of KDD*, pages 245–250, San Francisco, CA, 2001.

[6] Jeremy Buhler and Martin Tompa. Finding motifs using random projections. *Journal of Computational Biology*, 9(2):225–242, 2002.

[7] Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proc. of STOC*, pages 380–388, Montreal, Quebec, Canada, 2002.

[8] G. P. Chistyakov and F. Götze. Limit distributions of studentized means. *The Annals of Probability*, 32(1A):28–77, 2004.

[9] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60 – 65, 2003.

[10] Susan T. Dumais. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, 23(2):229–236, 1991.

[11] Richard Durrett. *Probability: Theory and Examples*. Duxbury Press, Belmont, CA, second edition, 1995.

[12] William Feller. *An Introduction to Probability Theory and Its Applications (Volume II)*. John Wiley & Sons, New York, NY, second edition, 1971.

[13] Xiaoli Zhang Fern and Carla E. Brodley. Random

projection for high dimensional data clustering: A cluster ensemble approach. In *Proc. of ICML*, pages 186–193, Washington, DC, 2003.

[14] Dmitriy Fradkin and David Madigan. Experiments with random projections for machine learning. In *Proc. of KDD*, pages 517–522, Washington, DC, 2003.

[15] P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory A*, 44(3):355–362, 1987.

[16] Navin Goel, George Bebis, and Ara Nefian. Face recognition experiments with random projection. In *Proc. of SPIE*, pages 426–437, Bellingham, WA, 2005.

[17] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of ACM*, 42(6):1115–1145, 1995.

[18] F. Götze. On the rate of convergence in the multivariate CLT. *The Annals of Probability*, 19(2):724–739, 1991.

[19] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings and data stream computation. In *FOCS*, pages 189–197, Redondo Beach,CA, 2000.

[20] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proc. of STOC*, pages 604–613, Dallas, TX, 1998.

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

[22] Samuel Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proc. of IJCNN*, pages 413–418, Piscataway, NJ, 1998.

[23] Man Lan, Chew Lim Tan, Hwee-Boon Low, and Sam Yuan Sung. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In *Proc. of WWW*, pages 1032–1033, Chiba, Japan, 2005.

[24] Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, New York, NY, second edition, 1998.

[25] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the self-similar nature of Ethernet traffic. *IEEE/ACM Trans. Networking*, 2(1):1–15, 1994.

[26] Edda Leopold and Jorg Kindermann. Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, 46(1-3):423–444, 2002.

[27] Henry C.M. Leung, Francis Y.L. Chin, S.M. Yiu, Roni Rosenfeld, and W.W. Tsang. Finding motifs with insufficient number of strong binding sites. *Journal of Computational Biology*, 12(6):686–701, 2005.

[28] Ping Li, Trevor J. Hastie, and Kenneth W. Church. Improving random projections using marginal information. In *Proc. of COLT*, Pittsburgh, PA, 2006.

[29] Jessica Lin and Dimitrios Gunopulos. Dimensionality reduction by random projection and latent semantic indexing. In *Proc. of SDM*, San Francisco, CA, 2003.

[30] Bing Liu, Yiming Ma, and Philip S. Yu. Discovering unexpected information from your competitors' web sites. In *Proc. of KDD*, pages 144–153, San Francisco, CA, 2001.

[31] Kun Liu, Hillol Kargupta, and Jessica Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):92–106, 2006.

[32] B. F. Logan, C. L. Mallows, S. O. Rice, and L. A. Shepp. Limit distributions of self-normalized sums. *The Annals of Probability*, 1(5):788–809, 1973.

[33] Chris D. Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA, 1999.

[34] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):232–351, 2005.

[35] Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. In *Proc. of PODS*, pages 159–168, Seattle,WA, 1998.

[36] Deepak Ravichandran, Patrick Pantel, and Eduard Hovy. Randomized algorithms and NLP: Using locality sensitive hash function for high speed noun clustering. In *Proc. of ACL*, pages 622–629, Ann Arbor, MI, 2005.

[37] Jason D. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger. Tackling the poor assumptions of naive Bayes text classifiers. In *Proc. of ICML*, pages 616–623, Washington, DC, 2003.

[38] Ozgur D. Sahin, Aziz Gulbeden, Fatih Emekçi, Divyakant Agrawal, and Amr El Abbadi. Prism: indexing multi-dimensional data in p2p networks using reference vectors. In *Proc. of ACM Multimedia*, pages 946–955, Singapore, 2005.

[39] Gerard Salton and Chris Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.

[40] I. S. Shiganov. Refinement of the upper bound of the constant in the central limit theorem. *Journal of Mathematical Sciences*, 35(3):2545–2550, 1986.

[41] Chunqiang Tang, Sandhya Dwarkadas, and Zhichen Xu. On scaling latent semantic indexing for large peer-to-peer systems. In *Proc. of SIGIR*, pages 112–121, Sheffield, UK, 2004.

[42] Santosh Vempala. Random projection: A new approach to VLSI layout. In *Proc. of FOCS*, pages 389–395, Palo Alto, CA, 1998.

[43] Santosh Vempala. *The Random Projection Method*. American Mathematical Society, Providence, RI, 2004.

[44] William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S*. Springer-Verlag, New York, NY, fourth edition, 2002.

[45] Clement T. Yu, K. Lam, and Gerard Salton. Term weighting in information retrieval using the term precision model. *Journal of ACM*, 29(1):152–170, 1982.

APPENDIX

A. PROOFS

Let $\{u_i\}_{i=1}^n$ denote the rows of the data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$. A projection matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ consists of i.i.d. entries r_{ji}:

$$\mathbf{Pr}(r_{ji} = \sqrt{s}) = \mathbf{Pr}(r_{ji} = -\sqrt{s}) = \frac{1}{2s}, \ \mathbf{Pr}(r_{ji} = 0) = 1 - \frac{1}{s},$$

$$\mathrm{E}(r_{ji}) = 0, \quad \mathrm{E}(r_{ji}^2) = 1, \quad \mathrm{E}(r_{ji}^4) = s, \quad \mathrm{E}(|r_{ji}^3|) = \sqrt{s},$$

$$\mathrm{E}\left(r_{ji}\ r_{j'i'}\right) = 0, \quad \mathrm{E}\left(r_{ji}^2\ r_{j'i'}\right) = 0 \quad \text{when} \ \ i \neq i', \ \text{or} \ \ j \neq j'.$$

We denote the projected data vectors by $v_i = \frac{1}{\sqrt{k}} \mathbf{R}^{\mathrm{T}} u_i$. For convenience, we denote

$$m_1 = \|u_1\|^2 = \sum_{j=1}^{D} u_{1,j}^2, \qquad m_2 = \|u_2\|^2 = \sum_{j=1}^{D} u_{2,j}^2,$$

$$a = u_1^{\mathrm{T}} u_2 = \sum_{j=1}^{D} u_{1,j} u_{2,j}, \ d = \|u_1 - u_2\|^2 = m_1 + m_2 - 2a.$$

We will always assume

$$s = o(D), \quad \mathrm{E}(u_{1,j}^4) < \infty, \ \mathrm{E}(u_{2,j}^4) < \infty, \quad (\Rightarrow \mathrm{E}(u_{1,j}^2 u_{2,j}^2) < \infty).$$

By the strong law of large numbers

$$\frac{\sum_{j=1}^{D} u_{1,j}^I}{D} \to \mathrm{E}\left(u_{1,j}^I\right), \ \frac{\sum_{j=1}^{D} \left(u_{1,j} - u_{2,j}\right)^I}{D} \to \mathrm{E}\left(u_{1,j} - u_{2,j}\right)^I,$$

$$\frac{\sum_{j=1}^{D} \left(u_{1,j} u_{2,j}\right)^J}{D} \to \mathrm{E}\left(u_{1,j} u_{2,j}\right)^J, \ a.s. \qquad I = 2, 4, \qquad J = 1, 2.$$

A.1 Moments

The following expansions are useful for proving the next three lemmas.

$$m_1 m_2 = \sum_{j=1}^{D} u_{1,j}^2 \sum_{j=1}^{D} u_{2,j}^2 = \sum_{j=1}^{D} u_{1,j}^2 u_{2,j}^2 + \sum_{j \neq j'}^{D} u_{1,j}^2 u_{2,j'}^2,$$

$$m_1^2 = \left(\sum_{j=1}^{D} u_{1,j}^2 \right)^2 = \sum_{j=1}^{D} u_{1,j}^4 + 2 \sum_{j<j'} u_{1,j}^2 u_{1,j'}^2,$$

$$a^2 = \left(\sum_{j=1}^{D} u_{1,j} u_{2,j} \right)^2 = \sum_{j=1}^{D} u_{1,j}^2 u_{2,j}^2 + 2 \sum_{j<j'} u_{1,j} u_{2,j} u_{1,j'} u_{2,j'}.$$

LEMMA 1.

$$E\left(\|v_1\|^2\right) = \|u_1\|^2 = m_1,$$

$$Var\left(\|v_1\|^2\right) = \frac{1}{k}\left(2m_1^2 + (s-3)\sum_{j=1}^{D} u_{1,j}^4\right).$$

As $D \to \infty$,

$$\frac{(s-3)\sum_{j=1}^{D}(u_{1,j})^4}{m_1^2} \to \frac{s-3}{D}\frac{E(u_{1,j})^4}{E^2(u_{1,j})^2} \to 0.$$

PROOF OF LEMMA 1 . $v_1 = \frac{1}{\sqrt{k}}\mathbf{R}^T u_1$, *Let* $\mathbf{R}_i$ *be the* i^{th} *column of* $\mathbf{R}$, $1 \leq i \leq k$. *We can write the* i^{th} *element of* v_1 *to be* $v_{1,i} = \frac{1}{\sqrt{k}}\mathbf{R}_i^T u_1 = \frac{1}{\sqrt{k}}\sum_{j=1}^{D}(r_{ji})\,u_{1,j}$. *Therefore,*

$$v_{1,i}^2 = \frac{1}{k}\left(\sum_{j=1}^{D}\left(r_{ji}^2\right)u_{1,j}^2 + 2\sum_{j<j'}(r_{ji})\,u_{1,j}\,(r_{j'i})\,u_{1,j'}\right),$$

from which it follows that

$$E\left(v_{1,i}^2\right) = \frac{1}{k}\sum_{j=1}^{D} u_{1,j}^2, \quad E\left(\|v_1\|^2\right) = \sum_{j=1}^{D} u_{1,j}^2 = m_1.$$

$$v_{1,i}^4 = \frac{1}{k^2}\left(\sum_{j=1}^{D}\left(r_{ji}^2\right)u_{1,j}^2 + 2\sum_{j<j'}(r_{ji})\,u_{1,j}\,(r_{j'i})\,u_{1,j'}\right)^2$$

$$= \frac{1}{k^2}\begin{pmatrix} \sum_{j=1}^{D}\left(r_{ji}^4\right)u_{1,j}^4 + 2\sum_{j<j'}\left(r_{ji}^2\right)u_{1,j}^2\left(r_{j'i}^2\right)u_{1,j'}^2 \\ +4\left(\sum_{j<j'}(r_{ji})\,u_{1,j}\,(r_{j'i})\,u_{1,j'}\right)^2 \\ +4\left(\sum_{j=1}^{D}\left(r_{ji}^2\right)u_{1,j}^2\right)\left(\sum_{j<j'}(r_{ji})\,u_{1,j}\,(r_{j'i})\,u_{1,j'}\right) \end{pmatrix},$$

from which it follows that

$$E\left(v_{1,i}^4\right) = \frac{1}{k^2}\left(s\sum_{j=1}^{D} u_{1,j}^4 + 6\sum_{j<j'} u_{1,j}^2 u_{1,j'}^2\right),$$

$$Var\left(v_{1,i}^2\right) = \frac{1}{k^2}\left(s\sum_{j=1}^{D} u_{1,j}^4 + 6\sum_{j<j'} u_{1,j}^2 u_{1,j'}^2 - \left(\sum_{j=1}^{D} u_{1,j}^2\right)^2\right)$$

$$= \frac{1}{k^2}\left((s-1)\sum_{j=1}^{D} u_{1,j}^4 + 4\sum_{j<j'} u_{1,j}^2 u_{1,j'}^2\right)$$

$$= \frac{1}{k^2}\left(2m_1^2 + (s-3)\sum_{j=1}^{D} u_{1,j}^4\right),$$

$$Var\left(\|v_1\|^2\right) = \frac{1}{k}\left(2m_1^2 + (s-3)\sum_{j=1}^{D} u_{1,j}^4\right).$$

As $D \to \infty$,

$$\frac{(s-3)\sum_{j=1}^{D}(u_{1,j})^4}{m_1^2} = \frac{s-3}{D}\frac{\sum_{j=1}^{D}(u_{1,j})^4/D}{m_1^2/D^2} \to \frac{o(D)}{D}\frac{E(u_{1,j})^4}{E^2(u_{1,j})^2} \to 0.$$

□

LEMMA 2.

$$E\left(\|v_1 - v_2\|^2\right) = \|u_1 - u_2\|^2 = d,$$

$$Var\left(\|v_1 - v_2\|^2\right) = \frac{1}{k}\left(2d^2 + (s-3)\sum_{j=1}^{D}(u_{1,j} - u_{2,j})^4\right).$$

As $D \to \infty$,

$$\frac{(s-3)\sum_{j=1}^{D}(u_{1,j} - u_{2,j})^4}{d^2} \to \frac{s-3}{D}\frac{E(u_{1,j} - u_{2,j})^4}{E^2(u_{1,j} - u_{2,j})^2} \to 0$$

PROOF OF LEMMA 2. *The proof is analogous to the proof of Lemma 1.* □

LEMMA 3.

$$E\left(v_1^T v_2\right) = u_1^T u_2 = a,$$

$$Var\left(v_1^T v_2\right) = \frac{1}{k}\left(m_1 m_2 + a^2 + (s-3)\sum_{j=1}^{D} u_{1,j}^2 u_{2,j}^2\right).$$

As $D \to \infty$,

$$\frac{(s-3)\sum_{j=1}^{D} u_{1,j}^2 u_{2,j}^2}{m_1 m_2 + a^2} \to \frac{s-3}{D}\frac{E\left(u_{1,j}^2 u_{2,j}^2\right)}{E\left(u_{1,j}^2\right)E\left(u_{2,j}^2\right) + E^2\left(u_{1,j}u_{2,j}\right)} \to 0.$$

PROOF OF LEMMA 3.

$$v_{1,i}v_{2,i} = \frac{1}{k}\left(\sum_{j=1}^{D}\left(r_{ji}^2\right)u_{1,j}u_{2,j} + \sum_{j\neq j'}(r_{ji})\,u_{1,j}\,(r_{j'i})\,u_{2,j'}\right),$$

$$\Longrightarrow E\left(v_{1,i}v_{2,i}\right) = \frac{1}{k}\sum_{j=1}^{D} u_{1,j}u_{2,j}, \quad E\left(v_1^T v_2\right) = a.$$

$$v_{1,i}^2 v_{2,i}^2$$

$$= \frac{1}{k^2}\left(\sum_{j=1}^{D}\left(r_{ji}^2\right)u_{1,j}u_{2,j} + \sum_{j\neq j'}(r_{ji})\,u_{1,j}\,(r_{j'i})\,u_{2,j'}\right)^2$$

$$= \frac{1}{k^2}\begin{pmatrix} \sum_{j=1}^{D}\left(r_{ji}^4\right)u_{1,j}^2u_{2,j}^2 + \\ 2\sum_{j<j'}\left(r_{ji}^2\right)u_{1,j}u_{2,j}\left(r_{j'i}^2\right)u_{1,j'}u_{2,j'} + \\ \left(\sum_{j\neq j'}(r_{ji})\,u_{1,j}\,(r_{j'i})\,u_{2,j'}\right)^2 + \\ \left(\sum_{j=1}^{D}\left(r_{ji}^2\right)u_{1,j}u_{2,j}\right)\left(\sum_{j\neq j'}(r_{ji})\,u_{1,j}\,(r_{j'i})\,u_{2,j'}\right) \end{pmatrix},$$

$$\Longrightarrow$$

$$E\left(v_{1,i}^2 v_{2,i}^2\right)$$
$$=\frac{1}{k^2}\left(s\sum_{j=1}^{D}u_{1,j}^2u_{2,j}^2+4\sum_{j<j'}u_{1,j}u_{2,j}u_{1,j'}u_{2,j'}+\sum_{j\neq j'}u_{1,j}^2u_{2,j'}^2\right)$$
$$=\frac{1}{k^2}\left((s-2)\sum_{j=1}^{D}u_{1,j}^2u_{2,j}^2+\sum_{j\neq j'}u_{1,j}^2u_{2,j'}^2+2a^2\right)$$
$$=\frac{1}{k^2}\left(m_1m_2+(s-3)\sum_{j=1}^{D}u_{1,j}^2u_{2,j}^2+2a^2\right),$$

$$Var\left(v_{1,i}v_{2,i}\right)=\frac{1}{k^2}\left(m_1m_2+a^2+(s-3)\sum_{j=1}^{D}u_{1,j}^2u_{2,j}^2\right),$$

$$Var\left(v_1^Tv_2\right)=\frac{1}{k}\left(m_1m_2+a^2+(s-3)\sum_{j=1}^{D}u_{1,j}^2u_{2,j}^2\right).$$

□

A.2 Asymptotic Distributions

LEMMA 4. *As $D\to\infty$,*

$$\frac{v_{1,i}}{\sqrt{m_1/k}}\overset{\mathcal{L}}{\Longrightarrow}N(0,1),\qquad\frac{\|v_1\|^2}{m_1/k}\overset{\mathcal{L}}{\Longrightarrow}\chi_k^2,$$

with the rate of convergence

$$|F_{v_{1,i}}(y)-\Phi(y)|\leq 0.8\sqrt{s}\frac{\sum_{j=1}^{D}|u_{1,j}|^3}{m_1^{3/2}}\to 0.8\sqrt{\frac{s}{D}}\frac{E|u_{1,j}|^3}{\left(E\left(u_{1,j}^2\right)\right)^{3/2}}\to 0,$$

where $\overset{\mathcal{L}}{\Longrightarrow}$ denotes "convergence in distribution," $F_{v_{1,i}}(y)$ is the empirical cumulative density function (CDF) of $v_{1,i}$ and $\Phi(y)$ is the standard normal $N(0,1)$ CDF.

PROOF OF LEMMA 4. *The Lindeberg central limit theorem (CLT) and the Berry-Esseen theorem are needed for the proof [12, Theorems VIII.4.3 and XVI.5.2].*[6]

Write $v_{1,i}=\frac{1}{\sqrt{k}}\mathbf{R}_i^Tu_1=\sum_{j=1}^{D}\frac{1}{\sqrt{k}}\left(r_{ji}\right)u_{1,j}=\sum_{j=1}^{D}z_j$, with $z_j=\frac{1}{\sqrt{k}}\left(r_{ji}\right)u_{1,j}$. Then

$$E(z_j)=0,\quad Var(z_j)=\frac{u_{1,j}^2}{k},\quad E(|z_j|^{2+\delta})=s^{\frac{\delta}{2}}\frac{|u_{1,j}|^{2+\delta}}{k^{(2+\delta)/2}},\forall\delta>0.$$

Let $s_D^2=\sum_{j=1}^{D}Var(z_j)=\frac{\sum_{j=1}^{D}u_{1,j}^2}{k}=\frac{m_1}{k}$. Assume the Lindeberg condition

$$\frac{1}{s_D^2}\sum_{j=1}^{D}E\left(z_j^2;|z_j|\geq\epsilon s_D\right)\to 0,\ \textit{for any}\ \epsilon>0.$$

Then

$$\frac{\sum_{j=1}^{D}z_j}{s_D}=\frac{v_{1,i}}{\sqrt{m_1/k}}\overset{\mathcal{L}}{\Longrightarrow}N(0,1),$$

which immediately leads to

$$\frac{v_{1,i}^2}{m_1/k}\overset{\mathcal{L}}{\Longrightarrow}\chi_1^2,\qquad\frac{\|v_1\|^2}{m_1/k}=\sum_{i=1}^{k}\left(\frac{v_{1,i}^2}{m_1/k}\right)\overset{\mathcal{L}}{\Longrightarrow}\chi_k^2.$$

We need to go back and check the Lindeberg condition.

$$\frac{1}{s_D^2}\sum_{j=1}^{D}E\left(z_j^2;|z_j|\geq\epsilon s_D\right)\leq\frac{1}{s_D^2}\sum_{j=1}^{D}E\left(\frac{|z_j|^{2+\delta}}{(\epsilon s_D)^\delta}\right)$$
$$=\left(\frac{s}{D}\right)^{\frac{\delta}{2}}\frac{1}{\epsilon^\delta}\frac{\sum_{j=1}^{D}|u_{1,j}|^{2+\delta}/D}{\left(\sum_{j=1}^{D}u_{1,j}^2/D\right)^{(2+\delta)/2}}$$
$$\to\left(\frac{o(D)}{D}\right)^{\frac{\delta}{2}}\frac{1}{\epsilon^\delta}\frac{E|u_{1,j}|^{2+\delta}}{\left(E(u_{1,j}^2)\right)^{(2+\delta)/2}}\to 0,$$

provided $E|u_{1,j}|^{2+\delta}<\infty$, for some $\delta>0$, which is much weaker than our assumption that $E(u_{1,j}^4)<\infty$.

It remains to show the rate of convergence using the Berry-Esseen theorem. Let $\rho_D=\sum_{j=1}^{D}E|z_j|^3=\frac{s^{1/2}}{k^{3/2}}\sum_{j=1}^{D}|u_{1,j}|^3$

$$|F_{v_{1,i}}(y)-\Phi(y)|\leq 0.8\frac{\rho_D}{s_D^3}=0.8\sqrt{s}\frac{\sum_{j=1}^{D}|u_{1,j}|^3}{m_1^{3/2}}\to 0.8\sqrt{\frac{s}{D}}\frac{E|u_{1,j}|^3}{\left(E\left(u_{1,j}^2\right)\right)^{3/2}}\to 0.$$

□

LEMMA 5. *As $D\to\infty$,*

$$\frac{v_{1,i}-v_{2,i}}{\sqrt{d/k}}\overset{\mathcal{L}}{\Longrightarrow}N(0,1),\qquad\frac{\|v_1-v_2\|^2}{d/k}\overset{\mathcal{L}}{\Longrightarrow}\chi_k^2,$$

with the rate of convergence

$$|F_{v_{1,i}-v_{2,i}}(y)-\Phi(y)|\leq 0.8\sqrt{s}\frac{\sum_{j=1}^{D}|u_{1,j}-u_{2,j}|^3}{d^{3/2}}\to 0.8\sqrt{\frac{s}{D}}\frac{E|u_{1,j}-u_{2,j}|^3}{E^{\frac{3}{2}}\left(u_{1,j}-u_{2,j}\right)^2}\to 0.$$

PROOF OF LEMMA 5. *The proof is analogous to the proof of Lemma 4.* □

The next lemma concerns the joint distribution of $(v_{1,i},v_{2,i})$.

LEMMA 6. *As $D\to\infty$,*

$$\Sigma^{-\frac{1}{2}}\begin{bmatrix}v_{1,i}\\v_{2,i}\end{bmatrix}\overset{\mathcal{L}}{\Longrightarrow}N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1&0\\0&1\end{bmatrix}\right),\ \Sigma=\frac{1}{k}\begin{bmatrix}m_1&a\\a&m_2\end{bmatrix}$$

and

$$\mathbf{Pr}\left(sign(v_{1,i})=sign(v_{2,i})\right)\to 1-\frac{\theta}{\pi},\quad\theta=\cos^{-1}\left(\frac{a}{\sqrt{m_1m_2}}\right).$$

PROOF OF LEMMA 6. *We have seen that $Var(v_{1,i})=\frac{m_1}{k}$, $Var(v_{2,i})=\frac{m_2}{k}$, $E(v_{1,i}v_{2,i})=\frac{a}{k}$, i.e.,*

$$cov\left(\begin{bmatrix}v_{1,i}\\v_{2,i}\end{bmatrix}\right)=\frac{1}{k}\begin{bmatrix}m_1&a\\a&m_2\end{bmatrix}=\Sigma.$$

The Lindeberg multivariate central limit theorem [18] says

$$\Sigma^{-\frac{1}{2}}\begin{bmatrix}v_{1,i}\\v_{2,j}\end{bmatrix}\overset{\mathcal{L}}{\Longrightarrow}N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1&0\\0&1\end{bmatrix}\right).$$

[6]The best Berry-Esseen constant 0.7915 (≈ 0.8) is from [40].

The multivariate Lindeberg condition is automatically satisfied by assuming bounded third moments of $u_{1,j}$ and $u_{2,j}$. A trivial consequence of the asymptotic normality yields

$$\mathbf{Pr}\left(sign(v_{1,i}) = sign(v_{2,i})\right) \to 1 - \frac{\theta}{\pi}.$$

□

Strictly speaking, we should write $\theta = \cos^{-1}\left(\frac{\mathrm{E}(u_{1,j}u_{2,j})}{\sqrt{\mathrm{E}(u_{1,j}^2)\mathrm{E}(u_{2,j}^2)}}\right)$.

A.3 An Asymptotic MLE Using Margins

LEMMA 7. *Assuming that the margins, m_1 and m_2 are known and using the asymptotic normality of $(v_{1,i}, v_{2,i})$, we can derive an asymptotic maximum likelihood estimator (MLE), which is the solution to a cubic equation*

$$a^3 - a^2\left(v_1^T v_2\right) + a\left(-m_1 m_2 + m_1\|v_2\|^2 + m_2\|v_1\|^2\right) - m_1 m_2 v_1^T v_2 = 0,$$

Denoted by $\hat{a}_{MLE}$, the asymptotic variance of this estimator is

$$Var\left(\hat{a}_{MLE}\right)_\infty = \frac{1}{k}\frac{\left(m_1 m_2 - a^2\right)^2}{m_1 m_2 + a^2}.$$

PROOF OF LEMMA 7. *For notational convenience, we treat $(v_{1,i}, v_{2,i})$ as exactly normally distributed so that we do not need to keep track of the "convergence" notation.*

The likelihood function of $\{v_{1,i}, v_{2,i}\}_{i=1}^k$ is then

$$lik\left(\{v_{1,i}, v_{2,i}\}_{i=1}^k\right) = (2\pi)^{-\frac{k}{2}}|\Sigma|^{-\frac{k}{2}} \times \exp\left(-\frac{1}{2}\sum_{i=1}^k \left[\begin{array}{cc} v_{1,i} & v_{2,i}\end{array}\right]\Sigma^{-1}\left[\begin{array}{c} v_{1,i} \\ v_{2,i}\end{array}\right]\right).$$

where

$$\Sigma = \frac{1}{k}\left[\begin{array}{cc} m_1 & a \\ a & m_2\end{array}\right].$$

We can then express the log *likelihood function, $l(a)$, as*

$$\log lik\left(\{v_{1,i}, v_{2,i}\}_{i=1}^k\right) \propto l(a) = -\frac{k}{2}\log\left(m_1 m_2 - a^2\right) - \frac{k}{2}\frac{1}{m_1 m_2 - a^2}\sum_{i=1}^k\left(v_{1,i}^2 m_2 - 2v_{1,i}v_{2,i}a + v_{2,i}^2 m_1\right),$$

The MLE equation is the solution to $l'(a) = 0$, which is

$$a^3 - a^2\left(v_1^T v_2\right) + a\left(-m_1 m_2 + m_1\|v_2\|^2 + m_2\|v_1\|^2\right) - m_1 m_2 v_1^T v_2 = 0$$

The large sample theory [24, Theorem 6.3.10] says that $\hat{a}_{MLE}$ is asymptotically unbiased and converges in distribution to a normal random variable $N\left(a, \frac{1}{I(a)}\right)$, where $I(a)$, the expected Fisher Information, is

$$I(a) = -E\left(l''(a)\right) = k\frac{m_1 m_2 + a^2}{\left(m_1 m_2 - a^2\right)^2},$$

after some algebra.

Therefore, the asymptotic variance of $\hat{a}_{MLE}$ would be

$$Var\left(\hat{a}_{MLE}\right)_\infty = \frac{1}{k}\frac{\left(m_1 m_2 - a^2\right)^2}{m_1 m_2 + a^2}. \quad (32)$$

□

B. HEAVY-TAILED DATA

We illustrate that *very sparse random projections* are fairly robust against heavy-tailed data, by a Pareto distribution.

The assumption of finite moments has simplified the analysis of convergence a great deal. For example, assuming $(\delta+2)$th moment, $0 < \delta \leq 2$ and $s = o(D)$, we have

$$(s)^{\delta/2}\frac{\sum_{j=1}^D |u_{1,j}|^{2+\delta}}{\left(\sum_{j=1}^D (u_{1,j}^2)\right)^{1+\delta/2}} = \left(\frac{s}{D}\right)^{\delta/2}\frac{\sum_{j=1}^D |u_{1,j}|^{2+\delta}/D}{\left(\sum_{j=1}^D (u_{1,j}^2)/D\right)^{1+\delta/2}} \to \left(\frac{s}{D}\right)^{\delta/2}\frac{\mathrm{E}\left(u_{1,j}^{2+\delta}\right)}{\left(\mathrm{E}\left(u_{1,j}^2\right)\right)^{1+\delta/2}} \to 0. \quad (33)$$

Note that $\delta = 2$ corresponds to the rate of convergence for the variance in Lemma 1, and $\delta = 1$ corresponds to the rate of convergence for asymptotic normality in Lemma 4. From the proof of Lemma 4 in Appendix A, we can see that the convergence of (33) (to zero) with any $\delta > 0$ suffices for achieving asymptotic normality.

For heavy-tailed data, the fourth moment (or even the second moment) may not exist. The most common model for heavy-tailed data is the Pareto distribution with the density function[7] $f(x;\alpha) = \frac{\alpha}{x^{\alpha+1}}$, whose mth moment $= \frac{\alpha}{\alpha-m}$, only defined if $\alpha > m$. The measurements of α for many types of data are available in [34]. For example, $\alpha = 1.2$ for the word frequency, $\alpha = 2.04$ for the citations to papers, $\alpha = 2.51$ for the copies of books sold in the US, etc.

For simplicity, we assume that $2 < \alpha \leq 2 + \delta \leq 4$. Under this assumption, the asymptotic normality is guaranteed and it remains to show the rate of convergence of moments and distributions. In this case, the second moment $\mathrm{E}\left(u_{1,j}^2\right)$ exists. The sum $\sum_{j=1}^D |u_{1,j}|^{2+\delta}$ grows as $O\left(D^{(2+\delta)/\alpha}\right)$ as shown in [11, Example 2.7.4].[8] Thus, we can write

$$s^{\delta/2}\frac{\sum_{j=1}^D |u_{1,j}|^{2+\delta}}{\left(\sum_{j=1}^D (u_{1,j}^2)\right)^{1+\delta/2}} = O\left(\frac{s^{\delta/2}}{D^{1+\delta/2-\frac{2+\delta}{\alpha}}}\right) = \left\{\begin{array}{ll} O\left(\frac{s}{D^{2-4/\alpha}}\right) & \delta = 2 \\ O\left(\frac{s}{D^{3-6/\alpha}}\right)^{1/2} & \delta = 1\end{array}\right., \quad (34)$$

from which we can choose s using prior knowledge of α.

For example, suppose $\alpha = 3$ and $s = \sqrt{D}$. (34) indicates that the rate of convergence for variances would be $O(D^{1/12})$ in terms of the standard error. (34) also verifies that the rate of convergence to normality is $O(D^{1/4})$, as expected.

Of course, we could always choose s more conservatively, e.g., $s = D^{1/4}$, if we know the data are severely heavy-tailed. Since D is large, a factor of $D^{1/4}$ is still considerable.

What if $\alpha < 2$? The second moment no longer exists. The analysis will involve the so-called *self-normalizing sums* [8, 32]; but we will not delve into this topic. In fact, it is not really meaningful to compute the l_2 distances when the data do not even have bounded second moment.

[7] Note that in general, a Pareto distribution has an addition parameter x_{min}, and $f(x;\alpha,x_{min}) = \frac{\alpha x_{min}}{x^{\alpha+1}}$ with $x \geq x_{min}$. Since we are only interested in the relative ratio of moments, we can without loss of generality assume $x_{min} = 1$. Also note that in [34], their "α" is equal to our $\alpha + 1$.

[8] Note that if $x \sim$ Pareto(α), then $x^t \sim$ Pareto(α/t).