Practical Procedures for Dimension Reduction in l_1

Ping Li PINGLI@STAT.STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305, USA

Trevor J. Hastie HASTIE@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA 94305, USA

Kenneth W. Church CHURCH@MICROSOFT.COM
Microsoft Research
Microsoft Corporation
Redmond, WA 98052, USA

Editor:

Abstract

We show that an analog of the Johnson-Lindestrauss (JL) lemma for dimension reduction in l_1 can be established using linear projections and nonlinear estimators. Previous studies have proved that no JL lemma exists for l_1 using linear estimators. We develop two nonlinear estimators including a strictly unbiased estimator and an improved estimator based on the maximum likelihood. While the maximum likelihood estimator (MLE) does not have a closed-form density function, we propose highly accurate closed-form approximations.

Sampling is also effective for dimension reduction in l_1. We apply a sketch-based sampling technique for l_1 dimension reduction, which is a combination of sketching and sampling and is particularly advantageous when the data are sparse.

Our results will be useful for applications concerning pairwise l_1 distances, including clustering, nearest neighbor searching, as well as approximating l_1 kernels for (e.g.) support vector machines (SVM).

Keywords: Dimension reduction, l_1 norm, Random projections, Sampling, JL bound

1. Introduction

This paper focuses on dimension reduction in l_1. Many known results for l_1 dimension reduction are negative, mainly because they were restricted to linear estimators. Indyk (2000, 2001) proposed *Cauchy random projections*, utilizing the "1-stable" property of the Cauchy distribution (Zolotarev, 1986). Unlike *normal random projections* for approximating l_2 distances (Vempala, 2004; Achlioptas, 2003), there are no "simple linear estimators" for recovering the original l_1 distances, as the Cauchy distribution does not even have a finite first moment. More recently, (Charikar and Sahai, 2002; Brinkman and Charikar, 2003; Lee and Naor, 2004; Brinkman and Charikar, 2005) proved that there is no analog of Johnson-Lindestrauss (JL) embedding lemma for l_1, if we are restricted to linear estimators. As a consequence, the distortions for l_1 dimension reduction based on linear projections and linear estimators are large.

In this study, we show that l_1 dimension reduction is not as pessimistic if the goal is to recover the original l_1 distances from projections or sampling. For Cauchy random projections, we propose two nonlinear estimators based on statistical estimation theory, an unbiased estimator and a maximum likelihood estimator (MLE). The unbiased estimator enables us to prove an analog of the JL lemma for l_1. The maximum likelihood estimator further improves the unbiased estimator.

Sampling is another option for dimension reduction in l_1. In fact, we can estimate distances in any norm (e.g., l_1 or l_2) from random samples by a simple scaling. The key behind the JL bound is that the data have exponential tails. Therefore, if the data are generated from some common thin-tailed distributions such as normal or gamma, the JL bound does exist using sampling. In severely heavy-tailed data, however, sampling may cause quite large errors.

For sparse boolean data, Li and Church (Li and Church, 2005a,b) developed a sketch-based sampling algorithm for dimension reduction. This general technique can be considered a combination of sampling and sketching (hashing), as it generates random samples online from sketches of the data. We show that it is straightforward to extend the algorithm to estimating l_1 distances in general real-valued data. This technique is advantageous when the data are highly sparse, often the case in large-scale data mining applications.

1.1 Motivations for Dimension Reduction

In modern learning and data mining applications, very large datasets are often encountered. We denote a data matrix by $\mathbf{A} \in \mathbb{R}^{n \times D}$, i.e., n data points in D dimensions. Both n and D can be very large, for example, $\mathbf{A}$ can be the *term-by-document matrix* at Web scale, or the market basket data for Amazon or Wal-Mart. Many operations on the data matrix are based solely on the pairwise distances, such as multidimensional scaling, clustering, or nearest neighbor searching, as well as SVM kernels. Computing all pairwise distances for $\mathbf{A}$ costs $O(n^2D)$, often infeasible; hence dimension reductions would be desirable. In addition, after reducing the dimensions, the data may be small enough to reside in the main memory hence many tasks (e.g., information retrieval, database query optimizations) can be efficiently conducted in the memory.

A popular technique called *random projections* multiplies the original data matrix $\mathbf{A}$ by a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ to generate a small projected data matrix $\mathbf{B} = \mathbf{AR} \in \mathbb{R}^{n \times k}$, with $k \ll \min(n, D)$. For dimension reduction in l_2, we often let $\mathbf{R}$ consist of i.i.d. standard normal $N(0, 1)$ entries; and hence we name it *normal random projections*. The well-known Johnson-Lindestrauss (JL) lemma (Johnson and Lindenstrauss, 1984) says that we need only $k = O\left(\log(n)/\epsilon^2\right)$ so that the l_2 distance between any pair of data points can be estimated within $(1 \pm \epsilon)$ fraction of the original l_2 distance, with a high probability.

Dimension reduction techniques are often crucial in machine learning, for example, the various projection pursuit and spectral projection algorithms (Friedman, 1987; Vempala and Wang, 2002; Kannan et al., 2005; Achlioptas and McSherry, 2005). Dasgupta (1999) applied random projections for learning mixtures of Gaussians. For training on large datasets using kernel machines (e.g., SVM), computing the Gram matrix (e.g., $\mathbf{AA}^{\mathrm{T}}$) and kernels could be very time-consuming and various projections or sampling methods have been suggested for speeding up the computations, e.g., (Smola and Schölkopf, 2000; Williams and Seeger, 2000; Achlioptas et al., 2001; Fine and Scheinberg, 2002; Drineas and Mahoney, 2005a,b; Keerthi et al., 2006).

1.2 Motivations for l_1 Dimension Reduction

It is well-known that the l_1 distance is far more robust than the l_2 distance against "outliers," i.e., exceptionally large-valued data points. There are numerous success stories on using the l_1 distance, for example, Lasso (Tibshirani, 1996), 1-norm SVM (Zhu et al., 2003).

There are also l_1 norm SVM kernels such as the Laplacian radial basis kernel.[1] As reported in (Chapelle et al., 1999; Ferecatu et al., 2004), in some applications (e.g., histogram-based image classification), the l_1 norm kernels perform better. Hein et al. (2004); Hein and Bousquet (2005) proposed various kernels on probability measures including the total variation kernel. In a discrete form, the total variation distance is the l_1 distance between two probability measures. The technique for dimension reduction in l_1, therefore, will also be applicable for efficiently computing the SVM kernels involving the l_1 norm.

1.3 Paper Organization

The rest of the paper is presented as follows. Section 2 briefly reviews the Cauchy random projections, for which simple linear estimators are provably bad. Section 3 presents our recommended procedure for Cauchy random projections before giving any detailed proofs. Section 4 presents some nonlinear estimators. In particular, we show that an analog of the JL lemma exits for l_1 using a strictly unbiased estimator. Section 5 is devoted to developing a bias-corrected maximum likelihood estimator, whose first four moments are derived. The distribution of the estimator, however, can not be exactly characterized. Hence in Section 6, we propose some approximations. The inverse Gaussian approximation is remarkably accurate. In Section 7, we introduce a sketch-based sampling approach for dimension reduction in l_1. Section 8 concludes the paper.

2. Introduction to Cauchy Random Projections

We give a brief introduction to Cauchy random projections for l_1 dimension reduction. The idea was originated from Indyk (2000, 2001).

We denote the original data matrix by $\mathbf{A} \in \mathbb{R}^{n \times D}$, i.e., n data points in D dimensions. Let $\{u_i^{\mathrm{T}}\}_{i=1}^n \in \mathbb{R}^D$ be the ith row of $\mathbf{A}$. Let $\mathbf{R} \in \mathbb{R}^{D \times k}$ be a random matrix whose entries are i.i.d. standard Cauchy random variables. Denote the entries of $\mathbf{R}$ by $\{r_{ji}\}_{j=1}^{D}{}_{i=1}^{k}$, i.e., $r_{ji} \sim C(0,1)$. Let $\mathbf{B} = \mathbf{AR} \in \mathbb{R}^{n \times k}$ and $\{v_i^{\mathrm{T}}\}_{i=1}^n \in \mathbf{R}^k$ be the ith row of $\mathbf{B}$, i.e., $v_i = \mathbf{R}^{\mathrm{T}} u_i$, $v_{i,j} = \mathbf{R}_j^{\mathrm{T}} u_i$, i.i.d. $j = 1$ to k, where $\mathbf{R}_j$ is the jth column of $\mathbf{R}$.

For simplicity, we focus on the leading two rows, $u_1\ u_2$ and $v_1\ v_2$. Because

$$v_{1,i} - v_{2,i} = \sum_{j=1}^{D} r_{ji} \left(u_{1,j} - u_{2,j} \right), \qquad i = 1, 2, ..., k \tag{1}$$

by the 1-stability of Cauchy (Zolotarev, 1986), we know that

$$v_{1,i} - v_{2,i} \sim C\left(0, \sum_{j=1}^{D} |u_{1,j} - u_{2,j}| \right). \tag{2}$$

1. The Laplacian radial basis kernel is of the form $\exp\left(-|u_i - u_j|\right)$, where $|u_i - u_j|$ is the l_1 distance between two vectors u_i and u_j.

That is, the projected data $v_{1,i} - v_{2,i}$ are also Cauchy random variables with the scale parameter being the l_1 distance, $|u_1 - u_2| = \sum_{j=1}^{D} |u_{1,j} - u_{2,j}|$, in the original space.

Recall that a Cauchy random variable $z \sim C(0, \gamma)$ has the density

$$f(z) = \frac{\gamma}{\pi} \frac{1}{z^2 + \gamma^2}, \qquad \gamma > 0, \quad -\infty < z < \infty \tag{3}$$

The easiest way to see the 1-stability is via the characteristic function,

$$\mathrm{E}\left(\exp(\sqrt{-1} z_1 t)\right) = \exp\left(-\gamma |t|\right), \tag{4}$$

$$\mathrm{E}\left(\exp\left(\sqrt{-1} t \sum_{j=1}^{D} c_j z_j\right)\right) = \exp\left(-\gamma \sum_{j=1}^{D} |c_j| t\right), \tag{5}$$

for z_1, z_2, ..., z_D, i.i.d. $C(0, \gamma)$, and any constants c_1, c_2, ..., c_D.

For simplicity, we denote

$$d = |u_1 - u_2| = \sum_{j=1}^{D} |u_{1,j} - u_{2,j}|, \tag{6}$$

$$x_i = v_{1,i} - v_{2,i}, \quad i = 1, 2, ..., k \tag{7}$$

x_1, x_2, ..., x_k are i.i.d. $C(0, d)$.

The task is to estimate d from x_i's. The simple linear estimator in terms of $\sum_{i=1}^{k} |x_i|$ is provably bad (Charikar and Sahai, 2002; Brinkman and Charikar, 2003; Lee and Naor, 2004; Brinkman and Charikar, 2005). Our work resorts to nonlinear estimators,[2] including a strictly unbiased nonlinear estimator in Section 4 and a maximum likelihood estimator in Section 5. But first, we present our recommended procedure for Cauchy random projections so that the casual readers may obtain enough information without the need of knowing the detailed proofs.

3. Our Recommended Procedures for Cauchy Random Projections

We outline the steps we recommend for Cauchy random projections. Recall that we assume a data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ and we would like to store a small projected data matrix $\mathbf{B} \in \mathbb{R}^{n \times k}$, with $k \ll \min(n, D)$. The basic steps for Cauchy random projections are:

1. Determine k.
2. Generate a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ consisting of i.i.d. standard Cauchy random variables.
3. Let $\mathbf{B} = \mathbf{AR}$ be the projected data matrix.
4. Estimate the original l_1 distances from $\mathbf{B}$ whenever needed.

Next, we present our method for estimating the l_1 distances from Cauchy random projections, followed by our recommended approach for determining k in a principled way.

2. For certain applications, nonlinear estimators are not applicable. We thank Moses Charikar and Assaf Naor for pointing this out in private communications. Also, we thank Assaf Naor for pointing out that the nonlinear estimators we provide can not be norms, unlike linear estimators. Of course, these restrictions do not affect many applications in learning and data mining, which only concern the pairwise distances, regardless whether the distances are computed by linear estimators or nonlinear estimators.

3.1 Estimating the Original l_1 Distances

We illustrate the estimation method using the leading two rows of the data. Recall that the leading two rows in the projected data matrix $\mathbf{B} \in \mathbb{R}^{n \times k}$ are denoted by v_1 and v_2; while the corresponding two rows in the original data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ are denoted by u_1 and u_2.

The original l_1 distance $d = |u_1 - u_2|$ is estimated by

$$\hat{d}_1 = \hat{d}\left(1 - \frac{1}{k}\right), \tag{8}$$

where $\hat{d}$ solves the nonlinear MLE equation

$$-\frac{k}{d} + \sum_{i=1}^{k} \frac{2d}{(v_{1,i} - v_{2,i})^2 + d^2} = 0, \tag{9}$$

by iterative methods, starting with the following initial guess

$$\hat{d}_{gm} = \cos^k\left(\frac{\pi}{2k}\right) \exp\left(\frac{1}{k}\sum_{i=1}^{k} \log(|v_{1,i} - v_{2,i}|)\right). \tag{10}$$

For numerical stability, we suggest normalizing $v_{1,i} - v_{2,i}$ by $\hat{d}_{gm}$ before iteratively solving the nonlinear equation.

3.2 Determining k

The users should provide three parameters: ϵ, ξ, and ν such that

With a probability at least $1 - \xi$, except $1/\nu$ fraction, any distance will be approximated within $(1 \pm \epsilon)$ fraction of the truth.

- $0 < \epsilon < 1$ is the *accuracy*, i.e., within $(1 \pm \epsilon)$ fraction of the truth.
- ξ is the *confidence*, or the *level of significance*. In most fields, ξ is routinely chosen to be 0.01 or 0.05, but sometimes $\xi = \frac{1}{4}$ or $\frac{1}{3}$ is also used.
- ν is the *correction factor*, taking into account the multiple comparison effect. This factor is a bit controversial. The standard JL-type of argument adopts the most stringent criterion by letting $\nu = \frac{n^2}{2}$, known as the Bonferroni correction. Alternatively, if we allow at most $1/\nu$ fraction (among all $\frac{n^2}{2}$ pairs) to be arbitrarily distorted, one can take (e.g.,) $\nu = 10^6$ (one out of a million) or $\nu = 100$ (one out of a hundred). This approach is accepted in multiple hypothesis testing (Lehmann and Romano, 2005) and is probably reasonable in many applications (Abraham et al., 2005).[3]

3. In l_2 norm embedding with *normal random projections*, k computed from the JL bound can easily exceed 1000 (in fact sometimes may exceed n), while in applied publications, k is often quite small. For example, (Bingham and Mannila, 2001) reported that in their experiments $k \approx 50$ seemed enough while the JL bound gave $k = 1600$. As surveyed in (Indyk, 2000), (Kaski, 1998) considered $k = 90$ was good enough in their application.

If $10^{-4} > \xi/\nu \geq 10^{-10}$, we suggest

$$k = \frac{4.4\left(\log 2p - \log \xi\right)}{\epsilon^2/(1+\epsilon)}, \tag{11}$$

which is unnecessarily conservative when $\xi/\nu \geq 10^{-4}$.

If $\xi/\nu \geq 10^{-4}$, we suggest (iteratively) solving for k such that

$$\Phi\left(-\epsilon\sqrt{\frac{\alpha}{1-\epsilon}}\right) + \Phi\left(-\epsilon\sqrt{\frac{\alpha}{1+\epsilon}}\right) - e^{2\alpha}\Phi\left(-(2+\epsilon)\sqrt{\frac{\alpha}{1+\epsilon}}\right) + e^{2\alpha}\Phi\left(-(2-\epsilon)\sqrt{\frac{\alpha}{1-\epsilon}}\right) = \xi/\nu, \tag{12}$$

where $\alpha = \frac{1}{2/k+3/k^2}$. $\Phi(.)$ is the standard normal cumulative density function (CDF).

We expect that the bound (11) should still hold at least for $\xi/\nu \geq 10^{-12}$, although we could not verify this. Alternatively, if the users would like some bounds that hold for arbitrarily small ξ/ν, they may consider those rather crude (but true for any ξ/ν) tail bounds presented in Section 4.

4. Simple Nonlinear Estimators

We present some nonlinear estimators of the Cauchy scale parameter. Since the more accurate maximum likelihood estimator (MLE) to be discussed in Section 5 requires some iterative procedure for solving a nonlinear equation, these simpler estimators may give excellent initial values as well as the normalizing constant for numerical stability. Also, we show that an analog of the JL lemma for l_1 can be established using simple nonlinear estimators.

Lemma 1 derives two biased nonlinear estimators based on the fractional moment, i.e., $\mathrm{E}(|x|^\lambda)$ ($|\lambda| < 1$) and the logarithmic moment, i.e, $\mathrm{E}\left(\log(|x|)\right)$, respectively, where $x \sim C(0,d)$ is a Cauchy random variable with the scale parameter d. These two estimators are useful for deriving a better estimator and proving tail bounds.

Lemma 1 *Assume $x \sim C(0,d)$. Then*[4]

$$E\left(|x|^\lambda\right) = \frac{d^\lambda}{\cos(\lambda\pi/2)}, \qquad |\lambda| < 1 \tag{13}$$

$$E\left(\log(|x|)\right) = \log(d), \tag{14}$$

$$Var\left(\log(|x|)\right) = \frac{\pi^2}{4}, \tag{15}$$

4. One can also infer $\mathrm{E}(|x|^\lambda)$ from the monograph on stable distributions (Zolotarev, 1986), e.g., equations (2.1.8)-(2.1.12). Note that those formulas were derived for $\lambda > 0$, while we need to study the case $\lambda < 0$ for proving tail bounds later. Also note that there are a couple of obvious typos in those formulas. In (2.1.9), the term "t^{s-1}" inside the integral should be "t^{-s-1}". In (2.1.9)-(2.1.12), there should be a multiplicative term $\frac{1}{\pi}$. With these typos in mind, we can infer from (2.1.12) that $\mathrm{E}(|x|^\lambda) = \frac{2}{\pi}d^\lambda\Gamma(1-\lambda)\Gamma(\lambda)\sin(\lambda\pi/2)$, which agrees with our results, because $\Gamma(1-\lambda)\Gamma(\lambda) = \frac{\pi}{\sin(\pi\lambda)}$ by the property of the gamma function $\Gamma(.)$.

from which we can derive two biased estimators of the Cauchy scale parameter d:

$$\hat{d}_\lambda = \left(\frac{1}{k} \sum_{i=1}^{k} |x_i|^\lambda \cos(\lambda\pi/2) \right)^{1/\lambda}, \quad |\lambda| < 1, \tag{16}$$

$$\hat{d}_{log} = \exp\left(\frac{1}{k} \sum_{i=1}^{k} \log(|x_i|) \right), \tag{17}$$

with variances

$$Var\left(\hat{d}_\lambda\right) = \frac{d^2}{k} \frac{\sin^2(\lambda\pi/2)}{\lambda^2 \cos(\lambda\pi)} + O\left(\frac{1}{k^2}\right), \quad |\lambda| < 1/2 \tag{18}$$

$$Var\left(\hat{d}_{log}\right) = \frac{\pi^2 d^2}{4k} + O\left(\frac{1}{k^2}\right). \tag{19}$$

The term $\frac{\sin^2(\lambda\pi/2)}{\lambda^2 \cos(\lambda\pi)}$ decreases monotonically with decreasing $|\lambda|$, reaching a limit

$$\lim_{\lambda \to 0} \frac{\sin^2(\lambda\pi/2)}{\lambda^2 \cos(\lambda\pi)} = \frac{\pi^2}{4}. \tag{20}$$

In other words, the variance of $\hat{d}_\lambda$ converges to that of $\hat{d}_{log}$ as $|\lambda|$ approaches zero.

Therefore, $\hat{d}_{log}$ is preferred as it has smaller variance than $\hat{d}_\lambda$. Note that $\hat{d}_{log}$ can be written as the *geometric mean*:

$$\hat{d}_{log} = \prod_{i=1}^{k} |x_i|^{1/k}. \tag{21}$$

We can evaluate the moment $\mathrm{E}\left(\hat{d}_{log}\right)^t$ for $|t| < k$. Although $\hat{d}_{log}$, strictly speaking, does not have exponential tails, it does have high-order polynomial tails. Given k and ϵ, we can in principle always find an exponential upper bound for $\mathbf{Pr}\left(|\hat{d}_{log} - d| \geq \epsilon d\right)$ because polynomial tails can be bounded by exponential tails in a finite range (in particular, $0 \leq \epsilon < 1$).

In fact, the bias of $\hat{d}_{log}$ can be removed. The properties of the bias-removed estimator, denoted by $\hat{d}_{gm}$, are addressed in the next lemma.

Lemma 2 *Denoted by $\hat{d}_{gm}$, the estimator*

$$\hat{d}_{gm} = \cos^k\left(\frac{\pi}{2k}\right) \prod_{i=1}^{k} |x_i|^{1/k}, \quad k > 1 \tag{22}$$

is unbiased, with the variance (valid when $k > 2$)

$$Var\left(\hat{d}_{gm}\right) = d^2 \left(\frac{\cos^{2k}\left(\frac{\pi}{2k}\right)}{\cos^k\left(\frac{\pi}{k}\right)} - 1 \right) \tag{23}$$

$$= \frac{d^2}{k}\frac{\pi^2}{4} + \frac{\pi^4}{32}\frac{d^2}{k^2} + O\left(\frac{1}{k^3}\right) = 2.5\frac{d^2}{k} + 3.0\frac{d^2}{k^2} + O\left(\frac{1}{k^3}\right). \tag{24}$$

The third and fourth central moments are (for $k > 3$ and $k > 4$, respectively)

$$E\left(\hat{d}_{gm} - E\left(\hat{d}_{gm}\right)\right)^3 = \frac{3\pi^4}{16}\frac{d^3}{k^2} + O\left(\frac{1}{k^3}\right) = 18.3\frac{d^3}{k^2} + O\left(\frac{1}{k^3}\right) \tag{25}$$

$$E\left(\hat{d}_{gm} - E\left(\hat{d}_{gm}\right)\right)^4 = \frac{3\pi^4}{16}\frac{d^4}{k^2} + O\left(\frac{1}{k^3}\right) = 18.3\frac{d^4}{k^2} + O\left(\frac{1}{k^3}\right) \tag{26}$$

We can also study the tail properties of $\hat{d}_{gm}$, which is crucial for proving the JL lemma.

Lemma 3 *The tail probability of $\hat{d}_{gm}$ can be bounded by*

$$\mathbf{Pr}\left(\hat{d}_{gm} \geq (1+\epsilon)d\right) \leq \frac{\cos^{kt_1^*}\left(\frac{\pi}{2k}\right)}{\cos^k\left(\frac{\pi t_1^*}{2k}\right)(1+\epsilon)^{t_1^*}}, \qquad \epsilon > 0 \tag{27}$$

where

$$t_1^* = \frac{2k}{\pi}\tan^{-1}\left(\left(\log(1+\epsilon) - k\log\cos\left(\frac{\pi}{2k}\right)\right)\frac{2}{\pi}\right). \tag{28}$$

And

$$\mathbf{Pr}\left(\hat{d}_{gm} \leq (1-\epsilon)d\right) \leq \frac{(1-\epsilon)^{t_2^*}}{\cos^k\left(\frac{\pi t_2^*}{2k}\right)\cos^{kt_2^*}\left(\frac{\pi}{2k}\right)}, \qquad 0 \leq \epsilon < 1, \quad k > \frac{\pi^2}{8\epsilon} \tag{29}$$

where

$$t_2^* = \frac{2k}{\pi}\tan^{-1}\left(\left(-\log(1-\epsilon) + k\log\cos\left(\frac{\pi}{2k}\right)\right)\frac{2}{\pi}\right). \tag{30}$$

Furthermore, for $0 \leq \epsilon < 1$, the following exponential tail bounds hold:

$$\mathbf{Pr}\left(\hat{d}_{gm} \geq (1+\epsilon)d\right) \leq \exp\left(-k\frac{\epsilon^2}{8(1+\epsilon)}\right) \tag{31}$$

$$\mathbf{Pr}\left(\hat{d}_{gm} \leq (1-\epsilon)d\right) \leq \exp\left(-k\frac{\epsilon^2}{20}\right), \qquad k > \frac{\pi^2}{4\epsilon} \tag{32}$$

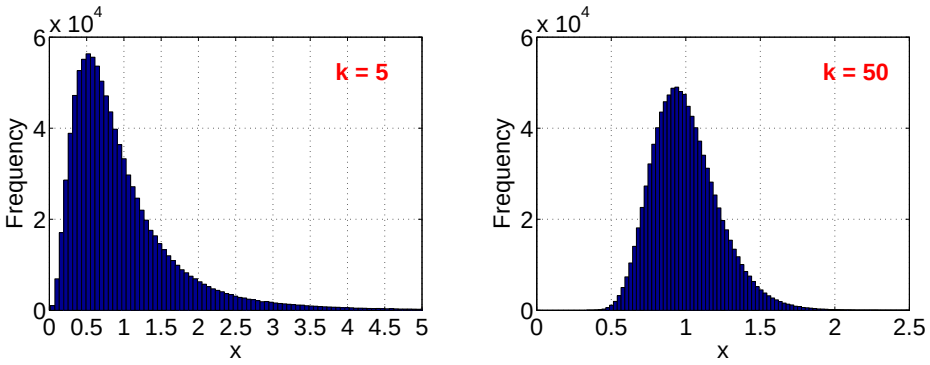
The above three Lemmas are proved in Appendix A, B, C, respectively.

A JL bound for l_1 follows immediately from the exponential tail bounds (31) and (32), by letting (31) + (32) $\leq \xi/\nu$ with $\nu = \frac{n^2}{2}$ and ξ being the specified confidence level (e.g., 0.05).

The "restriction" of $0 \leq \epsilon < 1$ is usually not an issue because we are mostly interested in small ϵ, e.g., $\epsilon \leq 0.5$. The additional restriction of $k > \frac{\pi^2}{4\epsilon}$ in (32) is not problem either. (Recall in the l_2 JL lemma, $k = O\left(\frac{\log(n)}{\epsilon^2}\right)$.)

We shall mention that the exponential tail bounds (31) and (32) are by no means "tight." We do not seek better exponential bounds because we believe they are only for theoretical demonstrations, since we actually recommend a maximum likelihood estimator (MLE) as opposed to $\hat{d}_{gm}$. Of course, $\hat{d}_{gm}$ would be an excellent starting point when iteratively solving for the MLE.

Figure 1 presents the histograms of $\hat{d}_{gm}$ for $d = 1$, with $k = 5$ and $k = 50$, obtained from simulations. The histograms reveal some characteristics shared by the MLE we will soon discuss:



(a) $k = 5$ (b) $k = 50$

Figure 1: Histograms of $\hat{d}_{gm}$, obtained from 10^6 simulations. At least in the range not too far from the mean, the distribution of $\hat{d}_{gm}$ resembles a gamma; and also resembles a normal when k is large enough.

- Supported on $[0, \infty)$, $\hat{d}_{gm}$ is positively skewed.
- The distribution of $\hat{d}_{gm}$ is still "heavy-tailed." However, in the region not too far from the mean, the distribution of $\hat{d}_{gm}$ may be well captured by a gamma (or a generalized gamma) distribution. For large k, even a normal approximation may suffice.

There are other nonlinear estimators for the Cauchy scale parameter. See (Zolotarev, 1986, Chapter 4) and the references therein. One particularly simple and popular estimator is based on the order statistics (sample quantiles) (Fama and Roll, 1968, 1971). We present an improved version of the quantile estimator given by (McCulloch, 1986):

$$\hat{d}_{or} = \frac{\hat{x}_{.75} - \hat{x}_{.25}}{2.0}, \tag{33}$$

where $\hat{x}_{.75}$ and $\hat{x}_{.25}$ are the .75 and .25 sample quantiles, respectively.

Figure 2 plots of the ratio of the mean square errors (MSE) $\frac{\text{Mse}(\hat{d}_{or})}{\text{Mse}(\hat{d}_{gm})}$, indicating that the quantile-based estimator $\hat{d}_{or}$ has considerably larger errors than the unbiased estimator $\hat{d}_{gm}$ we derive, especially when $k < 50$. For example, when $k = 9, 21, 49, 101, 197$, the ratio $\frac{\text{Mse}(\hat{d}_{or})}{\text{Mse}(\hat{d}_{gm})}$ = 2.24, 1.33, 1.13, 1.07, 1.04, respectively.

5. A Maximum Likelihood Estimator

This section presents a maximum likelihood estimator (MLE) for estimating l_1 distances from Cauchy random projections and recommends some practical criteria for choosing the sample size. Our main contribution includes the higher-order analysis for the bias and moments and accurate closed-from approximations to the distribution of the MLE.

The method of maximum likelihood has been widely used and its theoretical properties had been thoroughly studied. For example, the authors' recent work (Li et al., 2006a) applied the maximum

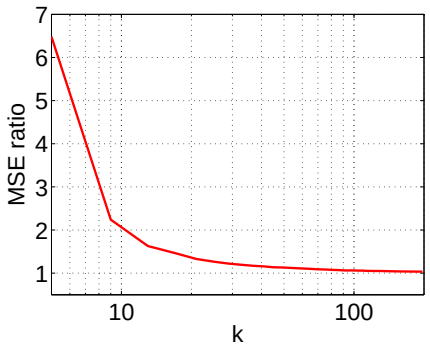


Figure 2: The ratios of the mean square errors (MSE): $\frac{\text{Mse}(\hat{d}_{or})}{\text{Mse}(\hat{d}_{gm})}$, from 10^5 simulations (for each k). The quantile-based estimator $\hat{d}_{or}$ always has larger errors than the unbiased estimator $\hat{d}_{gm}$ we derive. To help $\hat{d}_{or}$ achieve its best performance, the sample size k is chosen to be 5, 9, 13, ..., 197, spaced at 4. For example, when $k = 5$, $\hat{x}_{.25}$ is the 2nd smallest sample while $\hat{x}_{.75}$ is the 4th smallest sample. If $k - 1$ were not divisible by 4, we would have to either discard up to 3 samples or do some kind of smoothing; otherwise the estimation errors of $\hat{d}_{or}$ would be very large.

likelihood method to *normal random projections* and provided an improved estimator of the l_2 distance (hence in a sense, improved the l_2 JL bound).

The Cauchy distribution is frequently mentioned as a "challenging" example because of the "multiple roots" problem when estimating the location parameter (Barnett, 1966; Haas et al., 1970). In our case, since the location parameter is always zero, much of difficulty is avoided.

Recall that after applying Cauchy random projections k times on two data vectors $u_1, u_2 \in \mathbb{R}^D$, we generate random variables x_1, x_2, ..., x_k, i.i.d. Cauchy $C(0, d)$, where $d = |u_1 - u_2|$ is the l_1 distance in the original space. The $\log$ joint likelihood of x_i is

$$L(x_1, x_2, ...x_k; d) = k \log(d) - k \log(\pi) - \sum_{i=1}^{k} \log(x_i^2 + d^2), \tag{34}$$

whose first and second derivatives (w.r.t. d) are

$$L'(d) = \frac{k}{d} - \sum_{i=1}^{k} \frac{2d}{x_i^2 + d^2} \tag{35}$$

$$L''(d) = -\frac{k}{d^2} - \sum_{i=1}^{k} \frac{2x_i^2 - 2d^2}{(x_i^2 + d^2)^2} = -\frac{L'(d)}{d} - 4\sum_{i=1}^{k} \frac{x_i^2}{(x_i^2 + d^2)^2} \tag{36}$$

The maximum likelihood estimator of d, denoted by $\hat{d}$, is the solution to $L'(d) = 0$, i.e.,

$$-\frac{k}{d} + \sum_{i=1}^{k} \frac{2d}{x_i^2 + d^2} = 0. \tag{37}$$

Because $L''(\hat{d}) \leq 0$, we know that $\hat{d}$ indeed maximizes the joint likelihood and $\hat{d}$ is the only solution to the MLE equation (37). Solving (37) numerically is not difficult (e.g., a few iterations using the Newton's method). For even better accuracy, we recommend the following bias-corrected estimator:

$$\hat{d}_1 = \hat{d}\left(1 - \frac{1}{k}\right), \tag{38}$$

where $\hat{d}$ is the MLE solution to (37).

The following lemma concerns the asymptotic moments of $\hat{d}$ and $\hat{d}_1$, proved in Appendix D.

Lemma 4 *Both $\hat{d}$ and $\hat{d}_1$ are asymptotically unbiased and normal. The first four moments of $\hat{d}$ are:*

$$E\left(\hat{d} - d\right) = \frac{d}{k} + O\left(\frac{1}{k^2}\right) \tag{39}$$

$$Var\left(\hat{d}\right) = \frac{2d^2}{k} + \frac{7d^2}{k^2} + O\left(\frac{1}{k^3}\right) \tag{40}$$

$$E\left(\hat{d} - E(\hat{d})\right)^3 = \frac{12d^3}{k^2} + O\left(\frac{1}{k^3}\right) \tag{41}$$

$$E\left(\hat{d} - E(\hat{d})\right)^4 = \frac{12d^4}{k^2} + \frac{222d^4}{k^3} + O\left(\frac{1}{k^4}\right) \tag{42}$$

The first four moments of $\hat{d}_1$ are:

$$E\left(\hat{d}_1 - d\right) = E\left(\hat{d}\left(1 - \frac{1}{k}\right) - d\right) = O\left(\frac{1}{k^2}\right) \tag{43}$$

$$Var\left(\hat{d}_1\right) = \frac{2d^2}{k} + \frac{3d^2}{k^2} + O\left(\frac{1}{k^3}\right) \tag{44}$$

$$E\left(\hat{d}_1 - E(\hat{d}_1)\right)^3 = \frac{12d^3}{k^2} + O\left(\frac{1}{k^3}\right) \tag{45}$$

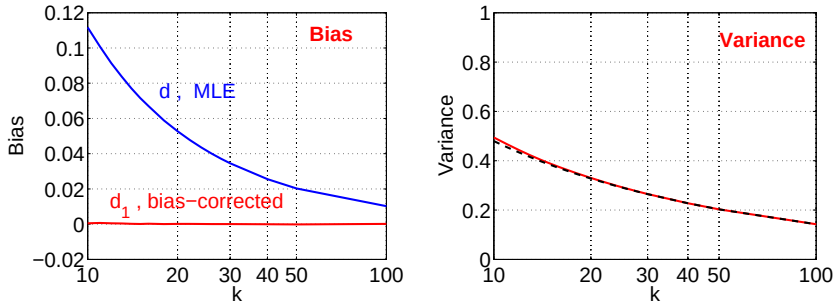
$$E\left(\hat{d}_1 - E(\hat{d}_1)\right)^4 = \frac{12d^4}{k^2} + \frac{186d^4}{k^3} + O\left(\frac{1}{k^4}\right) \tag{46}$$

The order $O\left(\frac{1}{k}\right)$ term of the variance, i.e., $\frac{2d^2}{k}$, is known, e.g., (Haas et al., 1970). We derive the bias-corrected estimator, $\hat{d}_1$, and the higher order moments using stochastic Taylor expansions (Bartlett, 1953; Shenton and Bowman, 1963; Ferrari et al., 1996; Cysneiros et al., 2001). The reason we need to correct the bias is because $\frac{1}{k}$ is obvious enough when $k < 20$. Even with the bias-correction, the $O\left(\frac{1}{k^2}\right)$ term of Var $\left(\hat{d}_1\right)$ is still $\frac{3}{2k}$ of the $O\left(\frac{1}{k}\right)$ term.

5.1 An Experimental Example

The proposed maximum likelihood estimator is tested on real Web crawl data provided by MSN. The data is a term-by-document matrix with $D = 2^{16}$ Web pages, i.e., each row represents the numbers of occurrences of one word in document 1 to document D. We conduct Cauchy random

projections on the words and estimate the original l_1 distances between words, by the estimators ($\hat{d}$ and $\hat{d}_1$) we have derived. In this experiment, we compare the empirical and (asymptotic) theoretical moments. Figure 3 presents the results for one pair of words. It indicates that the bias correction is effective and these (asymptotic) formulas for the first four moments of $\hat{d}_1$ are accurate, especially when $k \geq 20$.



(a) $\mathrm{E}(\hat{d}-d)/d$ v.s. $\mathrm{E}(\hat{d}_1-d)/d$

(b) $\left(\mathrm{E}(\hat{d}_1-\mathrm{E}(\hat{d}_1))^2/d^2\right)^{1/2}$

(c) $\left(\mathrm{E}(\hat{d}_1-\mathrm{E}(\hat{d}_1))^3/d^3\right)^{1/3}$

(d) $\left(\mathrm{E}(\hat{d}_1-\mathrm{E}(\hat{d}_1))^4/d^4\right)^{1/4}$

Figure 3: One pair of words are selected from an MSN term-by-document matrix with $D = 2^{16}$ Web pages. We conduct Cauchy random projections and estimate the l_1 distance using the maximum likelihood estimator $\hat{d}$ and the bias-corrected version $\hat{d}_1$, with 5×10^5 simulations (for each k). Panel (a) plots the biases for $\hat{d}/d$ and $\hat{d}_1/d$, indicating that the bias correction is indeed effective. Panel (b), (c), and (d) plot the variance, third moment, and fourth moment of $\hat{d}_1$, respectively. The dashed curves are the theoretical asymptotic moments. The plots are convincing that with $k \geq 20$, these formulas for moments are accurate (especially the variance).

5.2 How to Select k?

How to select k? This is a sample size selection problem. In the special case in which the exact distributions are known, we can select k by controlling the tail probability. A famous example is the

Johnson-Lindestrauss (JL) bound in *normal random projections* for l_2 (Johnson and Lindenstrauss, 1984; Frankl and Maehara, 1987; Indyk and Motwani, 1998; Arriaga and Vempala, 1999; Dasgupta and Gupta, 2003; Achlioptas, 2003).

In general, we can determine k from

$$\mathbf{Pr}\left(|\hat{d}_1 - d| > \epsilon d\right) \leq \xi/\nu, \tag{47}$$

where ξ is the level of significance (e.g, 0.05) and ν is the correction factor taking into account the effect of simultaneous multiple comparisons. That is, given ξ, ϵ, and ν, we choose k so that

With a probability at least $1-\xi$, *except* $1/\nu$ *fraction, any distance will be approximated within* $(1 \pm \epsilon)$ *fraction of the truth.*

The standard JL-type argument uses $\nu = \frac{n^2}{2} \approx \frac{n(n-1)}{2}$, the total number of distance pairs for n data points. Alternatively, we may allow at most $\frac{1}{\nu}$ fraction (e.g., $\nu = 100$, or $\nu = 10^6$) of the distances (among $\frac{n^2}{2}$ pairs) to be arbitrarily distorted.

We still need the distribution of $\hat{d}_1$ to compute k. It is in general not easy to characterize the distribution of an MLE exactly;[5] and asymptotic analysis is usually the standard approach. Note that, in our case, the distribution of $\frac{\hat{d}}{d}$ (and $\frac{\hat{d}_1}{d}$) is only a function of k, as shown in (Antle and Bain, 1969; Haas et al., 1970). Hence, we can, without loss of generality, take $d = 1$, and simulate the distribution of $\hat{d}_1$ at an affordable cost (at least for $\xi/\nu \geq 10^{-10}$). For example, we could tabulate the probability of $\hat{d}_1$ for the reasonable range of k values (e.g., $k \leq 1000$). Then given ϵ, ξ and ν, we can find the smallest k that satisfies $\mathbf{Pr}\left(|\hat{d}_1 - d| > \epsilon\right) \leq \xi/\nu$.

The asymptotic analysis does have the advantage due to its simplicity. In fact, most of the widely used approaches for selecting sample sizes are based on the asymptotic confidence intervals (Agresti, 2002; Lehmann and Romano, 2005), e.g., asymptotic normality or asymptotic chi-squared. In machine learning, the popular *Akaike Information Criterion* (AIC) for feature selection, is also based on asymptotic properties (Akaike, 1973).

The standard asymptotic approach for MLE is to assume that $\hat{d}_1$ is "exactly" normal with mean d and variance $\frac{2d^2}{k} + \frac{3d^2}{k^2}$. We can use the normal quantiles to determine k, or use the normal upper bound

$$\mathbf{Pr}\left(|\hat{d}_1 - d| > \epsilon d\right) \leq 2\exp\left(-\frac{\epsilon^2}{2\left(\frac{2}{k} + \frac{3}{k^2}\right)}\right) \leq \xi/\nu. \tag{48}$$

The problem with the normal approximation is that it is only accurate for a certain tail probability range (e.g., $\xi/\nu > 10^{-2} \sim 10^{-3}$). We will have to seek better alternatives when ξ/ν is chosen to be very small. For example, if we take $\nu = \frac{n^2}{2}$, ξ/ν will be less than 10^{-10} when $n > 10^5$.

In the next section, we use gamma and *generalized gamma* distributions to better approximate $\hat{d}_1$. One of the authors' prior work (Li et al., 2006b) applied these approximations to model the performance measure distribution in some wireless communication channels and produced accurate results in evaluating the error probabilities.

5. In fact, conditional on the observations $x_1, x_2, ..., x_k$, the distribution of $\hat{d}$ can be exactly characterized (Fisher, 1934). (Lawless, 1972) studied the conditional confi dence interval of the MLE. Later, (Hinkley, 1978) proposed the normal approximation to the exact conditional confi dence interval and showed that it was superior to the unconditional normality approximation. Unfortunately, we can not take advantage of the conditional analysis because our goal is to determine the sample size k before seeing any samples. In other words, we have to rely on the unconditional analysis by considering all possible samples.

6. Distribution Approximations

We propose a couple of approximations to the distribution of $\hat{d}_1$, the bias-corrected MLE for estimating the l_1 distance.

Instead of assuming normality, we can basically use any other (asymptotically equivalent) distributions with the same first two moments. For example, we can approximate $\hat{d}_1$ by a gamma random variable, which has the same support as $\hat{d}_1$, i.e., $[0, \infty)$. Besides, the odd central moments of a gamma distribution are non-zero, which to an extent, speeds up the rate of convergence because $\hat{d}_1$ also has non-zero odd moments. In contrast, a normal distribution always has zero odd central moments and is supported on $(-\infty, \infty)$.

We can further improve the gamma approximation by considering a *generalized gamma* distribution, which allows us to match the first three moments of $\hat{d}_1$. Interestingly, in this case, the generalized gamma approximation turns out to be an inverse Gaussian distribution, which has a closed-form probability density. More interestingly, this inverse Gaussian distribution automatically matches (at least) the first four (instead of three) moments, and exhibits remarkable accuracy even in the very small tail probability area.

6.1 Gamma Approximation

The gamma approximation is an obvious improvement over the normal approximation. Note that for dimension reduction in l_2 using normal random projections, the resultant estimator of the l_2 distance has a chi-squared distribution (e.g., (Vempala, 2004, Lemma 1.3)), which is a special of gamma.

A gamma distribution, $G(\alpha, \beta)$, has two parameters α and β, which can be determined by matching the first two moments of $\hat{d}_1$. That is, we assume that $\hat{d}_1 \sim G(\alpha, \beta)$, with

$$\alpha\beta = d, \qquad \alpha\beta^2 = \frac{2d^2}{k} + \frac{3d^2}{k^2}, \tag{49}$$

i.e.,

$$\alpha = \frac{1}{\frac{2}{k} + \frac{3}{k^2}}, \qquad \beta = \frac{2d}{k} + \frac{3d}{k^2}. \tag{50}$$

We could use the gamma distribution quantiles to determine the k directly. Or we can the following gamma Chernoff bounds:

$$\mathbf{Pr}\left(\hat{d}_1 \geq (1+\epsilon)d\right) \leq \exp\left(-\alpha\left(\epsilon - \log(1+\epsilon)\right)\right), \quad \epsilon \geq 0 \tag{51}$$

$$\mathbf{Pr}\left(\hat{d}_1 \leq (1-\epsilon)d\right) \leq \exp\left(-\alpha\left(-\epsilon - \log(1-\epsilon)\right)\right), \quad 0 \leq \epsilon < 1 \tag{52}$$

For simplicity, we could replace α by $\frac{k}{2.2}$ (assuming $k > 15$). If one prefers a "symmetric" upper bound, we can let

$$\mathbf{Pr}\left(|\hat{d}_1 - d| > \epsilon d\right) \leq 2\exp\left(-\alpha\left(\epsilon - \log(1+\epsilon)\right)\right), \quad 0 \leq \epsilon < 1, \tag{53}$$

because $\epsilon - \log(1+\epsilon) \leq -\epsilon - \log(1-\epsilon)$.

A JL-type of bound follows by restricting that $\mathbf{Pr}\left(|\hat{d}_1 - d| > \epsilon d\right) \leq \xi/\nu$, with $\nu = \frac{n^2}{2}$:

$$k \geq \frac{2.2\left(2\log n - \log \xi\right)}{\epsilon - \log(1+\epsilon)}. \tag{54}$$

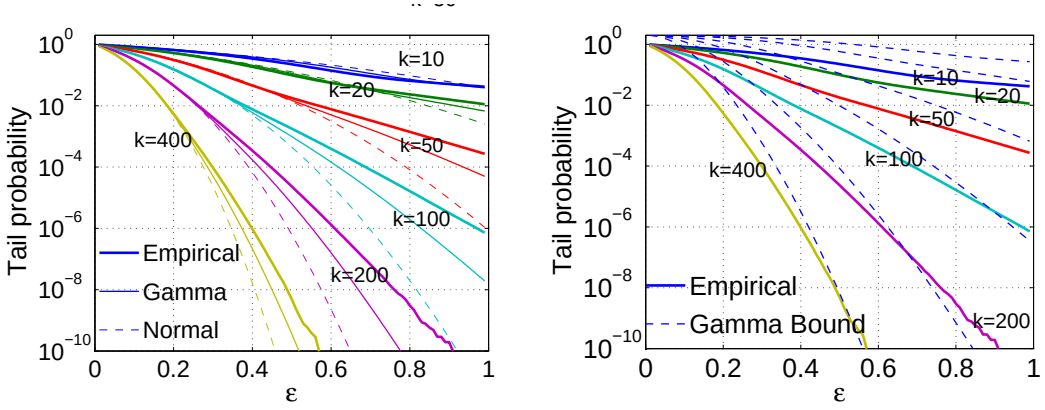
For general ν, we have

$$k \geq \frac{2.2\left(\log 2\nu - \log \xi\right)}{\epsilon - \log(1+\epsilon)}. \tag{55}$$

6.1.1 SIMULATIONS

We conduct extensive simulations on the maximum likelihood estimator $\hat{d}_1$, for verifying various approximate distributions we propose. Here, we first compare simulations with the gamma approximation, as shown in Figure 4.

Recall that the distribution of $\hat{d}_1/d$ is only a function of k. Therefore, we can, without loss of generality, take $d = 1$. We directly simulate standard Cauchy random variables and empirically compute $\mathbf{Pr}\left(|\hat{d}_1 - d| > \epsilon d\right)$ for k = 10, 20, 50, 100, 200, and 400, with $0 < \epsilon < 1$.



(a) (b)

Figure 4: We consider k = 10, 20, 50, 100, 200, and 400. For each k, we simulate standard Cauchy samples, from which we estimate the Cauchy parameter by the MLE $\hat{d}_1$ and compute the tail probabilities. Panel (a) compares the empirical tail probabilities (thick solid) with the gamma tail probabilities (thin solid), indicating that the gamma distribution is better than the normal (dashed) for approximating the distribution of $\hat{d}_1$. Panel (b) compares the empirical tail probabilities with the gamma upper bound (51)+(52). In the range of the tail probabilities $\geq 10^{-5} \sim 10^{-6}$, the gamma upper bounds hold.

Figure 4(a) shows that both the gamma and normal approximations are fairly accurate when the tail probability $\geq 10^{-2} \sim 10^{-3}$; and the gamma approximation is obviously better.

Figure 4(b) compares the empirical tail probabilities with the gamma Chernoff upper bound (51)+(52), indicating that these bounds are reliable, when the tail probability $\geq 10^{-5} \sim 10^{-6}$.

6.2 Inverse Gaussian (Generalized Gamma) Approximation

The distribution of $\hat{d}_1$ can be very well approximated by an inverse Gaussian distribution, which is a special case of the three-parameter generalized gamma distribution (Hougaard, 1986; Gerber, 1991), denoted by $GG(\alpha, \beta, \eta)$. Note that the usual gamma distribution is a special case with $\eta = 1$.

If $z \sim GG(\alpha, \beta, \eta)$, then the first three moments are

$$\mathrm{E}(z) = \alpha\beta, \quad \mathrm{Var}(z) = \alpha\beta^2, \quad \mathrm{E}\left(z - \mathrm{E}(z)\right)^3 = \alpha\beta^3(1+\eta). \tag{56}$$

We can approximate the distribution of $\hat{d}_1$ by matching the first three moments, i.e.,

$$\alpha\beta = d, \quad \alpha\beta^2 = \frac{2d^2}{k} + \frac{3d^2}{k^2}, \quad \alpha\beta^3(1+\eta) = \frac{12d^3}{k^2}, \tag{57}$$

from which we obtain

$$\alpha = \frac{1}{\frac{2}{k} + \frac{3}{k^2}}, \quad \beta = \frac{2d}{k} + \frac{3d}{k^2}, \quad \eta = 2 + O\left(\frac{1}{k}\right). \tag{58}$$

Taking only the leading term for η, the generalized gamma approximation for $\hat{d}_1$ would be

$$GG\left(\frac{1}{\frac{2}{k} + \frac{3}{k^2}}, \frac{2d}{k} + \frac{3d}{k^2}, 2\right). \tag{59}$$

In general, a generalized gamma distribution does not have a closed-form density function although it always has a closed-from moment generating function. In our case, since $\eta = 2$, (59) is in fact the inverse Gaussian distribution, which has a closed-form density function. That is, if we assume $\hat{d}_1 \sim IG(\alpha, \beta)$, i.e., an inverse Gaussian distribution with parameters α and β defined in (58), then the moment generating function (MGF), the probability density function (PDF), and cumulative density function (CDF) would be (Seshadri, 1993, Chapter 2) (Tweedie, 1957a,b)[6]

$$\mathrm{E}\left(\exp(\hat{d}_1 t)\right) = \exp\left(\alpha\left(1 - (1-2\beta t)^{1/2}\right)\right), \tag{60}$$

$$\mathbf{Pr}(\hat{d}_1 = y) = \frac{\alpha\sqrt{\beta}}{\sqrt{2\pi}} y^{-\frac{3}{2}} \exp\left(-\frac{(y/\beta - \alpha)^2}{2y/\beta}\right) = \sqrt{\frac{\alpha d}{2\pi}} y^{-\frac{3}{2}} \exp\left(-\frac{(y-d)^2}{2y\beta}\right), \tag{61}$$

$$\begin{aligned}\mathbf{Pr}\left(\hat{d}_1 \le y\right) &= \Phi\left(\sqrt{\frac{\alpha^2\beta}{y}}\left(\frac{y}{\alpha\beta} - 1\right)\right) + e^{2\alpha}\Phi\left(-\sqrt{\frac{\alpha^2\beta}{y}}\left(\frac{y}{\alpha\beta} + 1\right)\right) \\ &= \Phi\left(\sqrt{\frac{\alpha d}{y}}\left(\frac{y}{d} - 1\right)\right) + e^{2\alpha}\Phi\left(-\sqrt{\frac{\alpha d}{y}}\left(\frac{y}{d} + 1\right)\right),\end{aligned} \tag{62}$$

where $\Phi(.)$ is the standard normal CDF, i.e., $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$.

6. The inverse Gaussian distribution was first noted as the distribution of the first passage time of the Brownian motion with a positive drift. It has many interesting properties such as infinitely divisible. Two monographs (Chhikara and Folks, 1989; Seshadri, 1993) are devoted entirely to the inverse Gaussian distributions. For a quick reference, one can check *http://mathworld.wolfram.com/InverseGaussianDistribution.html*.

It is interesting that this inverse Gaussian approximation of $\hat{d}_1$ actually matches the fourth moment of $\hat{d}_1$, in addition to the first three. Assuming $\hat{d}_1 \sim IG(\alpha, \beta)$, then the fourth central moment should be

$$\begin{aligned} \mathrm{E}\left(\hat{d}_1 - \mathrm{E}\left(\hat{d}_1\right)\right)^4 &= 15\alpha\beta^4 + 3\left(\alpha\beta^2\right)^2 \\ &= 15d\left(\frac{2d}{k} + \frac{3d}{k^2}\right)^3 + 3\left(\frac{2d^2}{k} + \frac{3d^2}{k^2}\right)^2 \\ &= \frac{12d^4}{k^2} + \frac{156d^4}{k^3} + O\left(\frac{1}{k^4}\right). \end{aligned} \tag{63}$$

That is, the inverse Gaussian approximation matches not only the leading term, $\frac{12d^4}{k^2}$, but also almost the higher order term, i.e., $\frac{186d^4}{k^3}$, of the true asymptotic fourth moment of $\hat{d}_1$.

Assuming $\hat{d}_1 \sim IG(\alpha, \beta)$, the tail probability of $\hat{d}_1$ can then be expressed conveniently as

$$\mathbf{Pr}\left(\hat{d}_1 \geq (1+\epsilon)d\right) = \Phi\left(-\epsilon\sqrt{\frac{\alpha}{1+\epsilon}}\right) - e^{2\alpha}\Phi\left(-(2+\epsilon)\sqrt{\frac{\alpha}{1+\epsilon}}\right), \quad \epsilon \geq 0 \tag{64}$$

$$\mathbf{Pr}\left(\hat{d}_1 \leq (1-\epsilon)d\right) = \Phi\left(-\epsilon\sqrt{\frac{\alpha}{1-\epsilon}}\right) + e^{2\alpha}\Phi\left(-(2-\epsilon)\sqrt{\frac{\alpha}{1-\epsilon}}\right), \quad 0 \leq \epsilon < 1. \tag{65}$$

We can then select the sample size k by controlling the tail probabilities. We also obtain the Chernoff bounds

$$\mathbf{Pr}\left(\hat{d}_1 \geq (1+\epsilon)d\right) \leq \exp\left(-\frac{\alpha\epsilon^2}{2(1+\epsilon)}\right), \quad \epsilon \geq 0 \tag{66}$$

$$\mathbf{Pr}\left(\hat{d}_1 \leq (1-\epsilon)d\right) \leq \exp\left(-\frac{\alpha\epsilon^2}{2(1-\epsilon)}\right), \quad 0 \leq \epsilon < 1. \tag{67}$$

Again, we can replace α by $\frac{k}{2.2}$ assuming $k > 15$. A symmetric upper bound would be

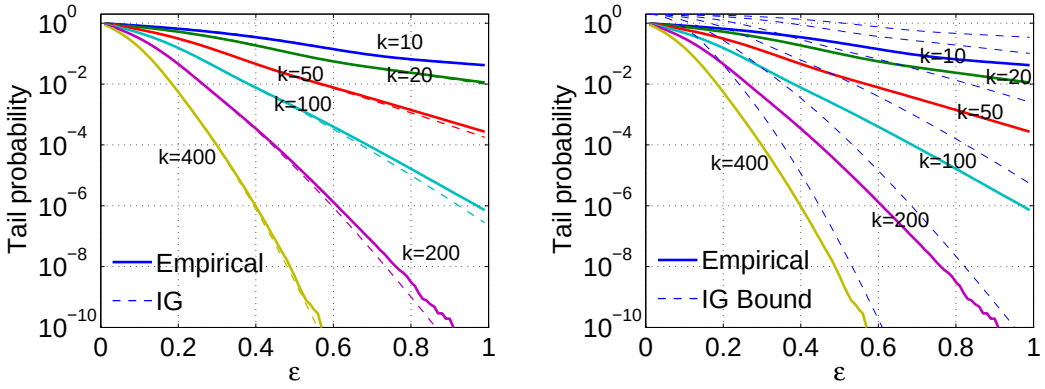
$$\mathbf{Pr}\left(|\hat{d}_1 - d| \geq \epsilon d\right) \leq 2\exp\left(-\frac{\alpha\epsilon^2}{2(1+\epsilon)}\right), \quad 0 \leq \epsilon < 1 \tag{68}$$

A JL-type of bound follows by restricting that $\mathbf{Pr}\left(|\hat{d}_1 - d| > \epsilon d\right) \leq \xi/\nu$, with $\nu = \frac{n^2}{2}$:

$$k \geq \frac{4.4\left(2\log n - \log\xi\right)}{\epsilon^2/(1+\epsilon)}, \qquad (\text{for} \quad k > 15) \tag{69}$$

For general ν, we have

$$k \geq \frac{4.4\left(\log 2\nu - \log\xi\right)}{\epsilon^2/(1+\epsilon)}. \tag{70}$$



(a) (b)

Figure 5: We compare the inverse Gaussian approximation with the same simulations as presented in Figure 4. Panel (a) compares the empirical tail probabilities with the inverse Gaussian tail probabilities, indicating that the approximation is highly accurate. Panel (b) compares the empirical tail probabilities with the inverse Gaussian upper bound (67)+(66). The upper bounds are all above the corresponding empirical curves, indicating that our proposed bounds are reliable at least in our simulation range.

6.2.1 SIMULATIONS

Figure 5 compares the inverse Gaussian approximation with the same simulations as presented in Figure 4. The results indicate that the inverse Gaussian approximation is highly accurate. When the tail probability $\geq 10^{-4} \sim 10^{-6}$, we can treat the inverse Gaussian as the exact distribution. The Chernoff upper bounds for the inverse Gaussian are always reliable in our simulation range.

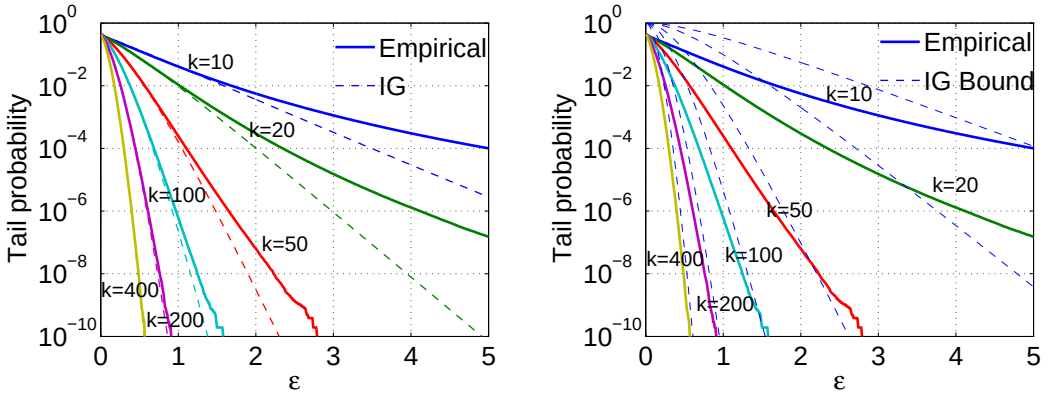
Although we are unable to simulate even smaller tail probabilities, we expect that the Chernoff upper bounds should at least hold in the $\geq 10^{-12}$ tail probability range. Note that the convenient bound (70) is considerably more conservative than the Chernoff bounds (67)+(66); and we expect that (70) should hold in even smaller tail probability range.

6.2.2 WHAT IF $\epsilon > 1$?

In the problem setting of interest in this study, we always restrict $0 \leq \epsilon < 1$, because in most applications we hope ϵ to be as small as possible (e.g., $\epsilon \leq 0.2 \sim 0.5$). For theoretical interest, Figure 6 plots $\mathbf{Pr}\left(\hat{d}_1 - d \geq \epsilon d\right)$, illustrating that when $\epsilon > 1.5 \sim 2$, the inverse Gaussian distribution no longer accurately approximates the distribution of $\hat{d}_1$.

7. A Sketch-based Sampling Algorithm for l_1 Dimension Reduction

We introduce another dimension reduction technique based on sampling. In particular, we extend Li and Church's sketching-based sampling algorithm (Li and Church, 2005a,b) for estimating l_1 distances in general real-valued data (the original algorithm was for boolean data). Their algorithm



(a) (b)

Figure 6: What if $\epsilon > 1$? Only the right tail, i.e., $\mathbf{Pr}\left(\hat{d}_1 - d \geq \epsilon d\right)$ is presented. Panel (a) shows the inverse Gaussian approximation deviates significantly from the simulations when $\epsilon > 1.5$. Panel (b) shows that the inverse Gaussian upper bound (67) is no longer reliable when $\epsilon > 2$.

is advantageous when the data are highly sparse, which is often the case in large-scale data mining applications, for example, the term-by-document matrix (Dhillon and Modha, 2001), or the market-basket data (Aggarwal and Wolf, 1999; Strehl and Ghosh, 2000).

In boolean data, this sketching algorithm improves Broder's (one permutation version) min-wise sketches (Broder, 1997; Broder et al., 1998, 2000; Charikar, 2002; Broder et al., 2003) by offering more flexibility and roughly halving the estimation variance. Note that, in boolean data, the l_1 distance coincides with the l_2 distance.

We first briefly describe our sampling procedure. We start with constructing random samples from a data matrix as shown in Figure 7. Then we show how to generate equivalent random samples using sketches in Figure 8.

In Figure 7, assuming that the column IDs are uniform at random (we will soon discuss how to achieve this), we can simply take the first D_s columns from the data matrix of D columns to construct a random sample.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
u_1	0	1	0	2	0	1	0	0	1	2	1	0	1	0	2
u_2	1	3	0	0	1	2	0	1	0	0	3	0	0	2	1

Figure 7: A data matrix with $D = 15$. If the column IDs are random, the first $D_s = 10$ columns constitute a random sample. u_i denotes the ith row.

For sparse data, we only need to store the non-zero elements in the form of tuples "ID (Value)." This structure is often called "postings" (or inverted index) in information retrieval. We denote the

postings by P_i for each row u_i. Figure 8(a) shows the postings for the same data matrix in Figure 7. The tuples are sorted ascending by the IDs.

P_1:	2 (1)	4 (2)	6 (1)	9 (1)	10 (2)	11 (1)	13 (1)	15 (2)
P_2:	1 (1)	2 (3)	5 (1)	6 (2)	8 (1)	11 (3)	14 (2)	15 (1)

(a)

K_1:	2 (1)	4 (2)	6 (1)	9 (1)	10 (2)	
K_2:	1 (1)	2 (3)	5 (1)	6 (2)	8 (1)	11 (3)

(b)

Figure 8: (a) Postings consist of tuples in the form "ID (Value)," where "ID" is the column ID of the entry in the original data matrix and "Value" is the value of that entry. (b) Sketches are simply the first few entries of postings. In this example, K_1 and K_2, are the first $k_1 = 5$ and $k_2 = 6$ elements of P_1 and P_2, respectively. Let $D_s = \min(10, 11) = 10$. After excluding 11(3) in K_2, we obtain exactly the same samples as if we sampled the first $D_s = 10$ columns in the original data matrix.

We sample directly from beginning of the postings as shown in Figure 8(b). We call the samples "sketches." A sketch, K_i, of postings P_i, is the first k_i entries (i.e., the smallest k_i IDs) of P_i. If we exclude all elements in sketches whose IDs are larger than

$$D_s = \min\left(\max(\text{ID}(K_1)), \max(\text{ID}(K_2))\right),$$

we obtain exactly the same samples as if we directly sampled the first D_s columns from the data matrix in Figure 7, as far as the two data points u_1 and u_2 are concerned. This way, we can convert sketches into random samples by conditioning on D_s, which we do not know beforehand. When estimating pairwise l_1 distances for all n data points, we will have $\frac{n(n-1)}{2}$ different values of D_s.

In order for our algorithm to work, we have to make sure that the columns are random. This can be achieved in various ways, e.g., hashing (Rabin, 1981; Broder, 1997). For simplicity, we apply a random permutation, denoted by Π, on the column IDs, i.e.,

$$\Pi : \Omega \to \Omega, \qquad \Omega = \{1, 2, 3, ..., D\}. \tag{71}$$

Let $\Pi(P_i)$ denote the postings P_i after permutation. We have to scan $\Pi(P_i)$ to find the k_i smallest to construct sketch K_i.

For estimating l_1 distances, our algorithm consists of the following steps:

- Construct sketches for all data points.
- Construct equivalent random samples from sketches online, after computing D_s for each pair of points (i.e., D_s is different pairwise). Compute the sample l_1 distances.
- Estimate the original l_1 distances by multiplying the sample l_1 distances with $\frac{D}{D_s}$.

The sampling procedure is highly flexible because we can adjust the sketch size (i.e., k_i) according to the sparsity of each data vector. For extremely sparse data points, we can probably afford to take the whole posting list as sketch. Note that the sketches can be reused for estimating distances in any other norms. This is is advantage when applications require estimating both l_1 and l_2 distances.

We can study some theoretical properties of sketches. Denote the estimated l_1 distance between u_1 and u_2 using sketches by $\hat{d}_{sk}$. It is easy to show that it is unbiased,

$$\mathrm{E}\left(\hat{d}_{sk}\right) = \mathrm{E}\left(\mathrm{E}\left(\hat{d}_{sk}|D_s\right)\right) = d, \tag{72}$$

with the variance

$$\begin{aligned}
\mathrm{Var}\left(\hat{d}_{sk}\right) &= \mathrm{E}\left(\mathrm{Var}\left(\hat{d}_{sk}|D_s\right)\right) + \mathrm{Var}\left(\mathrm{E}(\hat{d}_{sk}|D_s)\right) \\
&= \mathrm{E}\left(\frac{D^2}{D_s^2} D_s \left(\frac{\sum_{j=1}^D |u_{1,j}-u_{2,j}|^2}{D} - \left(\frac{\sum_{j=1}^D |u_{1,j}-u_{2,j}|}{D}\right)^2\right)\right) + 0 \\
&= \mathrm{E}\left(\frac{D}{D_s}\right)\left(d^{(2)} - \frac{d^2}{D}\right),
\end{aligned}$$

where $d^{(2)} = \sum_{j=1}^D |u_{1,j}-u_{2,j}|^2$ is the (squared) l_2 distance between u_1 and u_2.

Apparently, $\mathrm{Var}\left(\hat{d}_{sk}\right)$ is data-dependent. Note that $\mathrm{Var}\left(\hat{d}_{sk}\right)$ contains the l_2 distance $d^{(2)}$, which is related to the second moment of the data. Therefore, in severely heavy-tailed data, $\mathrm{Var}\left(\hat{d}_{sk}\right)$ can be large. On the other hand, $\mathrm{E}\left(\frac{D}{D_s}\right)$ is closely related to the sparsity of the data and may compensate the errors due to heavy-tailedness.

Denote f_1 and f_2 the numbers of non-zero elements in u_1 and u_2, respectively, and k_1 and k_2 the sketch sizes for u_1 and u_2, respectively. The intuitive approximation $\mathrm{E}\left(\frac{D}{D_s}\right) \approx \max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right)$ is quite accurate (especially when $k_1, k_2 \geq 20$) and is asymptotically exact.[7] Although it is apparently advantageous to take k_1 and k_2 to be proportional to f_1 and f_2, respectively, we assume that $k_1 = k_2 = k$, i.e., $\mathrm{E}\left(\frac{D}{D_s}\right) \approx \frac{\max(f_1,f_2)}{k}$, to simplify the discussion.

We can compare $\hat{d}_{sk}$ with the maximum likelihood estimator $\hat{d}_1$ we have discussed in terms of their variances. Here, we only consider the leading term of $\mathrm{Var}(\hat{d}_1)$, i.e.,

$$\mathrm{Var}\left(\hat{d}_{sk}\right) = \frac{\max(f_1,f_2)}{k}\left(d^{(2)} - \frac{d^2}{D}\right) = \frac{\max(f_1,f_2)}{D}\frac{1}{k}\left(Dd^{(2)} - d^2\right), \tag{73}$$

$$\mathrm{Var}\left(\hat{d}_1\right) = \frac{2d^2}{k}. \tag{74}$$

$\frac{\max(f_1,f_2)}{D} \leq 1$ is the sparsity factor, which can be very low in some applications. For example, in a term-by-document matrix, the sparsity can often be lower than 1% (Dhillon and Modha, 2001; Lin and Gunopulos, 2003). Therefore, this sketch-based sampling algorithm tends to have smaller variance in highly sparse data. On the other hand, the Cauchy-Schwartz inequality says $d^2 \leq Dd^{(2)}$. Particularly, in heavy-tailed data, d^2 can be much smaller than $Dd^{(2)}$, i.e., $d^2 \ll Dd^{(2)}$; hence the variance of $\hat{d}_{sk}$ tends to be larger.

7. It is easy to show that $\mathrm{E}\left(\frac{\max(\mathrm{ID}(\mathrm{K}_1))}{D}\right) = \frac{k_1}{f_1+1}$, $\mathrm{Var}\left(\frac{\max(\mathrm{ID}(\mathrm{K}_1))}{D}\right) = O\left(\frac{1}{k_1}\right)\mathrm{E}^2\left(\frac{\max(\mathrm{ID}(\mathrm{K}_1))}{D}\right)$, hence the approximation $\mathrm{E}\left(\frac{D_s}{D}\right) \approx \min\left(\frac{k_1}{f_1}, \frac{k_2}{f_2}\right)$ is very accurate and asymptotically exact (as k_1, k_2 become large). The reciprocal $\mathrm{E}\left(\frac{D}{D_s}\right) \approx \max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right)$ is also asymptotically exact and is still quite accurate.

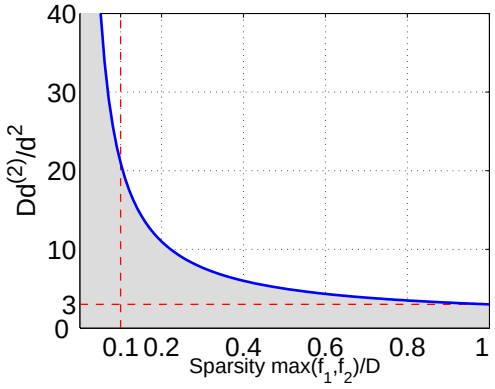


Figure 9: In the shaded area, the variance of sketches is smaller than the variance of Cauchy random projections. The horizontal axis is the sparsity factor $\frac{\max(f_1,f_x)}{D}$; and the vertical axis is $\frac{Dd^{(2)}}{d^2}$. When $\frac{Dd^{(2)}}{d^2} < 3$, the variance of sketches is always smaller.

Therefore, the data tail-heaviness and the data sparsity are two competing factors controlling which method may have smaller variance. If the data are located in the shaded area in Figure 9, then the variance of sketches is smaller. In particular, the variance of sketches is smaller when

$$\frac{Dd^{(2)}}{d^2} < 3, \quad \text{or} \tag{75}$$

$$\frac{Dd^{(2)}}{d^2} < 21 \quad \text{and} \quad \frac{\max(f_1, f_2)}{D} < 0.1 \tag{76}$$

7.1 Analysis Based on Moments

In this subsection, we show that, roughly speaking, as long as the original data have reasonably bounded second moments, sampling methods could be competitive in estimating l_1 distances.

We have shown that that sketches will have smaller variance than Cauchy random projections if

$$\frac{\max(f_1, f_2)}{D}\left(Dd^{(2)} - d^2\right) < 2d^2. \tag{77}$$

Because the data dimension D is large, by the law of large numbers, we know that

$$\frac{d}{D} = \frac{1}{D}\sum_{j=1}^{D}|u_{1,j} - u_{2,j}| \to \mathrm{E}\left(|u_{1,j} - u_{2,j}|\right) = \mathrm{E}(z) \tag{78}$$

$$\frac{d^{(2)}}{D} = \frac{1}{D}\sum_{j=1}^{D}|u_{1,j} - u_{2,j}|^2 \to \mathrm{E}\left(|u_{1,j} - u_{2,j}|^2\right) = \mathrm{E}(z^2), \tag{79}$$

$$\frac{Dd^{(2)}}{d^2} = \frac{d^{(2)}/D}{d^2/D^2} \to \frac{\mathrm{E}(z^2)}{\mathrm{E}^2(z)} \tag{80}$$

where $z = |u_{1,j} - u_{2,j}|$ denotes a random sample. Therefore, for convenience, we can analyze the relative variances of $\hat{d}_{sk}$ and $\hat{d}_1$ using the moments of z.

In order to continue the analysis, we make some assumptions on the data (i.e., z) distribution. This can at least give us some insightful information. We consider a Pareto distribution, which is popular for analyzing heavy-tailed data.[8]

Assume $z \sim$ Pareto(θ). Then $\mathrm{E}(z^m) = \frac{\theta}{\theta - m}$, which exists if $\theta > m$. It is easy to show that sketches have smaller variances if

$$\theta > 1 + \sqrt{1 + \frac{\max(f_1, f_2)}{2D}}. \tag{81}$$

Even in the extreme case where $\max(f_1, f_2) = D$, we only need $\theta > 1 + \sqrt{1.5} = 2.2247$. Note that while often heavy-tailed data are modeled to have $\theta < 2$, there are also many heavy-tailed data that exhibit $\theta > 2$, e.g., see (Newman, 2005).

This kind of analysis is, of course, very idealistic. For example, when the sparsity is as small as 10^{-3}, we are almost certain that sketches will have smaller errors, regardless whether the data are modeled to have bounded second moment or not.

7.2 An Example

We take two rows out of a chunk of the MSN Web crawl data and compute the sparsity, d, $d^{(2)}$, $\frac{Dd^{(2)}}{d^2}$, as well as the sample kurtosis in Table 1. For comparison, as the original data are clearly heavy-tailed (as indicated by the sample kurtosis value), we also quantize the data to be binary, since the binary quantization is an important term-weighting approach in practice.

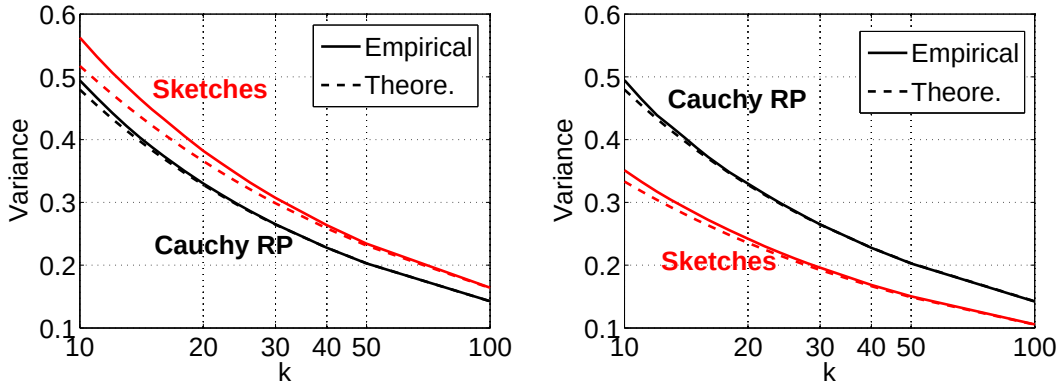
Table 1: Two words (rows) are taken from a MSN term-by-document matrix with $D = 2^{16} = 65536$. The values for sparsity, d, $d^{(2)}$, $\frac{Dd^{(2)}}{d^2}$, and sample kurtosis are given in the table. For comparison, we also quantized the data to be binary and compute the same statistics.

	$\frac{\max(f_1,f_2)}{D}$	d	$d^{(2)}$	$\frac{Dd^{(2)}}{d^2}$	Kurtosis
Original	0.4226	$104,752$	$1,227,692$	7.3324	216.81
Binary	0.4226	18,077	18,077	3.6254	0.4335

We estimate the l_1 distance using both sketches and Cauchy random projections. We present the empirical variances along with the theoretical variances in Figure 10. In this example, sketches have larger variances than Cauchy random projections in the original (heavy-tailed) data. After the binary quantization, sketches have (considerably) smaller variances. The figure also indicates that the theoretical variances of sketches are quite close to the empirical values. The errors are mainly due to the approximation $\mathrm{E}\left(\frac{D}{D_s}\right) \approx \frac{\max(f_1,f_2)}{k}$.

Note that in this case the sparsity factor, 0.4226, happens to be not small; hence the advantage of our sketch-based sampling algorithm is not fully represented.

8. For a quick reference to the Pareto distribution, see *http://en.wikipedia.org/wiki/Pareto_distribution*.



(a) Original (b) Binary

Figure 10: A pair of words are selected from a chunk of MSN Web crawl term-by-document matrix. We conduct both Cauchy random projections and sketches for estimating the original l_1 distance at the same sample size k. Panels (a) presents the comparisons using the original (heavy-tailed) data, indicating that sketches have larger variances than the Cauchy random projections. Panel (b) presents the results using the (binary) quantized data, indicating that sketches have considerably smaller variances in this case. The vertical axis is the square root of variance normalized by the original l_1 distance.

8. Conclusion

Dimension reduction in l_1 norm had been considered a hard problem. We show that if we do not restrict ourselves to linear estimators, we can estimate the original l_1 distances accurately using linear projections and various nonlinear estimators. For many applications in information retrieval, machine learning, and data mining, since we only need the pairwise distances, these nonlinear estimators should be useful.

For Cauchy random projections, we prove an analog of the JL lemma for l_1 using a strictly unbiased nonlinear estimator. For better accuracy, we recommend a maximum likelihood estimator (MLE) and provide convenient and reliable closed-form approximations to the tail probabilities and tail bounds. The extra computational cost is to solve a univariate nonlinear MLE equation, which does not seem to be a computational burden. In our experiments, we solve the MLE equation by the Newton's method. Usually after about five iterations, the results are accurate enough.

Sampling is effective for dimension reduction in l_1. The heavy-tailedness of the data does not affect the performance of sampling very strongly in l_1. The sketch-based sampling algorithm is particularly advantageous when the data are highly sparse, often the case in many important applications. The "disadvantage" of the sampling approach is that the performance is not guaranteed unless we make assumptions on the data. In this sense, sampling methods are not as "theoretically appealing" as random projections; although this is probably not much a practical concern.

Acknowledgment

We highly appreciate Moses Charikar and Assaf Naor for the detailed comments on l_1 embedding and the JL bound. We thank Christopher Burges for reading the draft and thank Dimitris Achlioptas and Tim Roughgarden for helpful references. We also thank Tze L. Lai, Art B. Owen, John Platt, Joseph Romano and Guenther Walther for helpful conversations. Finally, we thank Silvia Ferrari and Gauss Cordeiro for clarifying some parts of their papers.

Trevor Hastie was partially supported by grant DMS-0505676 from the National Science Foundation, and grant 2R01 CA 72028-07 from the National Institutes of Health.

References

Ittai Abraham, Yair Bartal, Hubert T.-H. Chan, Kedar Dhamdhere, Anupam Gupta, Jon M. Kleinberg, Ofer Neiman, and Aleksandrs Slivkins. Metric embeddings with relaxed guarantees. In *Proc. of FOCS*, pages 83–100, Pittsburgh, PA, 2005.

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

Dimitris Achlioptas and Frank McSherry. On spectral learning for mixture distributions. In *Proc. of COLT*, pages 458–469, Bertinoro, Italy, 2005.

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

Charu C. Aggarwal and Joel L. Wolf. A new method for similarity indexing of market basket data. In *Proc. of SIGMOD*, pages 407–418, Philadelphia, PA, 1999.

Alan Agresti. *Categorical Data Analysis*. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2002.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory, 2nd*, pages 267–281, 1973.

Charles Antle and Lee Bain. A property of maximum likelihood estimators of location and scale parameters. *SIAM Review*, 11(2):251–253, 1969.

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.

V. D. Barnett. Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots. *Biometrika*, 53(1/2):151–165, 1966.

M. S. Bartlett. Approximate confi dence intervals, II. *Biometrika*, 40(3/4):306–317, 1953.

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.

Bo Brinkman and Mose Charikar. On the impossibility of dimension reduction in l_1. In *Proc. of FOCS*, pages 514–523, Cambridge, MA, 2003.

Bo Brinkman and Mose Charikar. On the impossibility of dimension reduction in l_1. *Journal of ACM*, 52(2): 766–788, 2005.

Andrei Z. Broder. On the resemblance and containment of documents. In *Proc. of the Compression and Complexity of Sequences*, pages 21–29, Positano, Italy, 1997.

Andrei Z. Broder, Moses Charikar, Alan M. Frieze, and Michael Mitzenmacher. Min-wise independent permutations (extended abstract). In *Proc. of STOC*, pages 327–336, Dallas, TX, 1998.

Andrei Z. Broder, Moses Charikar, Alan M. Frieze, and Michael Mitzenmacher. Min-wise independent permutations. *Journal of Computer Systems and Sciences*, 60(3):630–659, 2000.

Andrei Z. Broder, Moses Charikar, and Michael Mitzenmacher. A derandomization using min-wise independent permutations. *Journal of Discrete Algorithms*, 1(1):11–20, 2003.

Olivier Chapelle, Patrick Haffner, and Vladimir N. Vapnik. Support vector machines for histogram-based image classifi cation. *IEEE Trans. Neural Networks*, 10(5):1055–1064, 1999.

Moses Charikar and Amit Sahai. Dimension reduction in the l_1 norm. In *Proc. of FOCS*, pages 551–560, Vancouver, BC, Canada, 2002.

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

Raj S. Chhikara and J. Leroy Folks. *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*. Marcel Dekker, Inc, New York, 1989.

Francisco Jose De. A. Cysneiros, Sylvio Jose P. dos Santos, and Gass M. Cordeiro. Skewness and kurtosis for maximum likelihood estimator in one-parameter exponential family models. *Brazilian Journal of Probability and Statistics*, 15(1):85–105, 2001.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *Proc. of FOCS*, pages 634–644, New York, 1999.

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

Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1-2):143–175, 2001.

Petros Drineas and Michael W. Mahoney. Approximating a gram matrix for improved kernel-based learning. In *Proc. of COLT*, pages 323–337, Bertinoro, Italy, 2005a.

Petros Drineas and Michael W. Mahoney. On the nystrom method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6(Dec):2153–2175, 2005b.

Eugene F. Fama and Richard Roll. Some properties of symmetric stable distributions. *Journal of the American Statistical Association*, 63(323):817–836, 1968.

Eugene F. Fama and Richard Roll. Parameter estimates for symmetric stable distributions. *Journal of the American Statistical Association*, 66(334):331–338, 1971.

Marin Ferecatu, Michel Crucianu, and Nozha Boujemaa. Retrieval of diffi cult image classes using svd-based relevance feedback. In *Prof. of Multimedia Information Retrieval*, pages 23–30, New York, NY, 2004.

Silvia L. P. Ferrari, Denise A. Botter, Gauss M. Cordeiro, and Francisco Cribari-Neto. Second and third order bias reduction for one-parameter family models. *Stat. and Prob. Letters*, 30:339–345, 1996.

Shai Fine and Katya Scheinberg. Effi cient svm training using low-rank kernel representations. *Journal of Machine Learning Research*, 2(Dec):243–264, 2002.

R. A. Fisher. Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London*, 144(852):285–307, 1934.

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.

Jerome H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82 (397):249–266, 1987.

Hans U. Gerber. From the generalized gamma to the generalized negative binomia l distribution. *Insurance:Mathematics and Economics*, 10(4):303–309, 1991.

I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, New York, fi fth edition, 1994.

Gerald Haas, Lee Bain, and Charles Antle. Inferences for the cauchy distribution based on maximum likelihood estimation. *Biometrika*, 57(2):403–408, 1970.

Matthias Hein and Olivier Bousquet. Hilbertian metrics and positive defi nite kernels on probability measures. In *Proc. of AISTATS*, pages 136–143, Barbados, 2005.

Matthias Hein, Thomas Navin Lal, and Olivier Bousquet. Hilbertian metrics on probability measures and their application in svm?s. In *DAGM-Symposium*, pages 270–277, Tübingen, Germany, 2004.

David V. Hinkley. Likelihood inference about location and scale parameters. *Biometrika*, 65(2):253–261, 1978.

P. Hougaard. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2):387–396, 1986.

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

Piotr Indyk. Algorithmic applications of low-distortion geometric embeddings. In *Proc. of FOCS*, pages 10–33, Las Vegas, NV, 2001.

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

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

Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. In *Proc. of COLT*, pages 447–457, Bertinoro, Italy, 2005.

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

S. Sathiya Keerthi, Olivier Chapelle, and Dennis DeCoste. Building support vector machines with reduced classifi er complexity. *Journal of Machine Learning Research*, 8:1–22, 2006.

J. F. Lawless. Conditional confi dence interval procedures for the location and scale parameters of the cauchy and logistic distributions. *Biometrika*, 59(2):377–386, 1972.

James R. Lee and Assaf Naor. Embedding the diamond graph in l_p and dimension reduction in l_1. *Geometric And Functional Analysis*, 14(4):745–747, 2004.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypothesis*. Springer, New York, NY, third edition, 2005.

Ping Li and Kenneth W. Church. Using sketches to estimate associations. In *Proc. of HLT/EMNLP*, pages 708–715, Vancouver, BC, Canada, 2005a.

Ping Li and Kenneth W. Church. Using sketches to estimate two-way and multi-way associations. Technical Report TR-2005-115, Microsoft Research, (A shorter version is available at www.stanford.edu/~pingli98/publications/Report_Sketch.pdf), Redmond, WA, September 2005b.

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

Ping Li, Debashis Paul, Ravi Narasimhan, and John Cioffi . On the distribution of SINR for the MMSE MIMO receiver and performance analysis. *IEEE Trans. Inform. Theory*, 52(1):271–286, 2006b.

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

J. Huston McCulloch. Simple consistent estimators of stable distribution parameters. *Communications on Statistics-Simulation*, 15(4):1109–1136, 1986.

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

Michael O. Rabin. Fingerprinting by random polynomials. Technical Report TR-15-81, Center for Research in Computing Technology, Cambridge, MA, 1981.

V. Seshadri. *The Inverse Gaussian Distribution: A Case Study in Exponential Families*. Oxford University Press Inc., New York, 1993.

L. R. Shenton and K. Bowman. Higher moments of a maximum-likelihood estimate. *Journal of Royal Statistical Society B*, 25(2):305–317, 1963.

Alex J. Smola and Bernhard Schölkopf. Sparse greedy matrix approximation for machine learning. In *Proc. of ICML*, pages 911–918, Standord, CA, 2000.

Alexander Strehl and Joydeep Ghosh. A scalable approach to balanced, high-dimensional clustering of market-baskets. In *Proc. of HiPC*, pages 525–536, Bangalore, India, 2000.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B*, 58(1):267–288, 1996.

M. C. K. Tweedie. Statistical properties of inverse gaussian distributions. I. *The Annals of Mathematical Statistics*, 28(2):362–377, 1957a.

M. C. K. Tweedie. Statistical properties of inverse gaussian distributions. II. *The Annals of Mathematical Statistics*, 28(3):696–705, 1957b.

Santosh Vempala and Grant Wang. A spectral algorithm for learning mixtures of distributions. In *Proc. of FOCS*, pages 113–122, Vancouver, BC, Canada, 2002.

Santosh S. Vempala. *The Random Projection Method*. American Mathematical Society, Providence, RI, 2004.

Christopher K. I. Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Proc. of NIPS*, pages 682–688, Denver, CO, 2000.

Ji Zhu, Saharon Rosset, Trevor Hastie, and Robert Tibshirani. 1-norm support vector machines. In *NIPS*, 2003.

V. M. Zolotarev. *One-dimensional Stable Distributions*. American Mathematical Society, Providence, RI, 1986.

Appendix A. Proof of Lemma 1

Assume $x \sim C(0, d)$, a Cauchy random variable with the scale parameter d. Then the probability density function (PDF) of $|x|$ is

$$\mathbf{Pr}(|x| = y) = \frac{2d}{\pi}\frac{1}{d^2 + y^2}, \qquad y > 0 \tag{82}$$

The first moment of $\log(|x|)$ would be

$$\begin{aligned}
\mathrm{E}\left(\log(|x|)\right) &= \frac{2d}{\pi}\int_0^\infty \frac{\log(y)}{y^2 + d^2}dy \\
&= \frac{1}{\pi}\int_0^\infty \frac{\log(d)y^{-1/2}}{y+1} + \frac{1/2\log(y)y^{-1/2}}{y+1}dy \\
&= \log(d),
\end{aligned} \tag{83}$$

with the help of the integral tables(Gradshteyn and Ryzhik, 1994, 3.221.1, 4.251.1).

Therefore, given i.i.d. samples x_1, x_2, ..., x_k of $C(0, d)$, a simple nonlinear estimator of d would be

$$\hat{d}_{log} = \exp\left(\frac{1}{k}\sum_{i=1}^{k}\log(|x_i|)\right). \tag{84}$$

We can derive another nonlinear estimator by evaluating $\mathrm{E}\left(|x|^\lambda\right)$ for some $|\lambda| < 1$. Again, using the integral tables(Gradshteyn and Ryzhik, 1994, 3.221.1), we obtain

$$\begin{aligned}
\mathrm{E}\left(|x|^\lambda\right) &= \frac{2d}{\pi}\int_0^\infty \frac{y^\lambda}{y^2 + d^2}dy \\
&= \frac{d^\lambda}{\pi}\int_0^\infty \frac{y^{\frac{\lambda-1}{2}}}{y+1}dy \\
&= \frac{d^\lambda}{\cos(\lambda\pi/2)},
\end{aligned} \tag{85}$$

from which a nonlinear estimator follows immediately

$$\hat{d}_\lambda = \left(\frac{1}{k}\sum_{i=1}^{k}|x_i|^\lambda\cos(\lambda\pi/2)\right)^{1/\lambda}, \qquad |\lambda| < 1 \tag{86}$$

Both nonlinear estimators $\hat{d}_{log}$ and $\hat{d}_\lambda$ are biased. The leading terms of their variances can be obtained by the *Delta method*, i.e., the second order stochastic Taylor expansion.

With the help of (Gradshteyn and Ryzhik, 1994, 4.261.10), we obtain

$$\mathrm{E}\left(\log^2(|x|)\right) = \log^2(d) + \frac{\pi^2}{4}, \quad \text{i.e.,} \quad \mathrm{Var}\left(\log^2(|x|)\right) = \frac{\pi^2}{4}. \tag{87}$$

By the *Delta Method*, the asymptotic variance of $\hat{d}_{log}$ would be

$$\mathrm{Var}\left(\hat{d}_{log}\right) = \frac{\pi^2}{4}\exp^2\left(\log(d)\right) + O\left(\frac{1}{k^2}\right) = \frac{\pi^2 d^2}{4k} + O\left(\frac{1}{k^2}\right). \tag{88}$$

Similarly, the asymptotic variance of $\hat{d}_\lambda$ is

$$\mathrm{Var}\left(\hat{d}_\lambda\right) = \frac{d^2}{k}\frac{\sin^2(\lambda\pi/2)}{\lambda^2\cos(\lambda\pi)} + O\left(\frac{1}{k^2}\right), \quad |\lambda| < 1/2 \tag{89}$$

Note that $\mathrm{Var}\left(\hat{d}_\lambda\right) \to \infty$ as $|\lambda| \to \frac{1}{2}$; and $\mathrm{Var}\left(\hat{d}_\lambda\right)$ converges to $\mathrm{Var}\left(\hat{d}_{log}\right)$ as $\lambda \to 0$, because

$$\lim_{\lambda\to 0}\frac{\sin^2(\lambda\pi/2)}{\lambda^2\cos(\lambda\pi)} = \frac{\pi^2}{4}. \tag{90}$$

This completes the proof of Lemma 1.

Appendix B. Proof of Lemma 2

Assume that x_1, x_2, ..., x_k, are i.i.d. $C(0, d)$. The estimator, $\hat{d}_{gm}$, expressed as

$$\hat{d}_{gm} = \cos^k\left(\frac{\pi}{2k}\right)\prod_{i=1}^{k}|x_i|^{1/k}, \tag{91}$$

is unbiased, because

$$\begin{aligned}
\mathrm{E}\left(\hat{d}_{gm}\right) &= \cos^k\left(\frac{\pi}{2k}\right)\prod_{i=1}^{k}\mathrm{E}\left(|x_i|^{1/k}\right) \\
&= \cos^k\left(\frac{\pi}{2k}\right)\prod_{i=1}^{k}\left(\frac{d^{1/k}}{\cos\left(\frac{\pi}{2k}\right)}\right) \\
&= d,
\end{aligned} \tag{92}$$

using the results in Lemma 1.

The variance is

$$\begin{aligned}
\mathrm{Var}\left(\hat{d}_{gm}\right) &= \cos^{2k}\left(\frac{\pi}{2k}\right)\prod_{i=1}^{k}\mathrm{E}\left(|x_i|^{2/k}\right) - d^2 \\
&= d^2\left(\frac{\cos^{2k}\left(\frac{\pi}{2k}\right)}{\cos^k\left(\frac{\pi}{k}\right)} - 1\right) \qquad (93) \\
&= \frac{d^2}{k}\frac{\pi^2}{4} + \frac{d^2}{k^2}\frac{\pi^4}{32} + O\left(\frac{1}{k^3}\right), \qquad (94)
\end{aligned}$$

because

$$\begin{aligned}\frac{\cos^{2k}\left(\frac{\pi}{2k}\right)}{\cos^{k}\left(\frac{\pi}{k}\right)} &= \left(\frac{1}{2}+\frac{1}{2}\left(\frac{1}{\cos(\pi/k)}\right)\right)^k \\ &= \left(1+\frac{1}{4}\frac{\pi^2}{k^2}+\frac{5}{48}\frac{\pi^4}{k^4}+O\left(\frac{1}{k^6}\right)\right)^k \\ &= 1+k\left(\frac{1}{4}\frac{\pi^2}{k^2}+\frac{5}{48}\frac{\pi^4}{k^4}\right)+\frac{k(k-1)}{2}\left(\frac{1}{4}\frac{\pi^2}{k^2}+\frac{5}{48}\frac{\pi^4}{k^4}\right)^2+... \\ &= 1+\frac{\pi^2}{4}\frac{1}{k}+\frac{\pi^4}{32}\frac{1}{k^2}+O\left(\frac{1}{k^3}\right) \end{aligned} \tag{95}$$

We can similarly show the third and fourth central moments:

$$\mathrm{E}\left(\hat{d}_{gm}-\mathrm{E}\left(\hat{d}_{gm}\right)\right)^3 = \frac{3\pi^4}{16}\frac{d^3}{k^2}+O\left(\frac{1}{k^3}\right) \tag{96}$$

$$\mathrm{E}\left(\hat{d}_{gm}-\mathrm{E}\left(\hat{d}_{gm}\right)\right)^4 = \frac{3\pi^4}{16}\frac{d^4}{k^2}+O\left(\frac{1}{k^3}\right), \tag{97}$$

but we skip the messy algebra.

Therefore, we have completed the proof of Lemma 2.

Appendix C. Proof of Lemma 3

This section proves the tail bounds for $\hat{d}_{gm}$. Note that $\hat{d}_{gm}$ does not have a moment generating function because $\mathrm{E}\left(\hat{d}_{gm}\right)^t$ does not exist if $t \geq k$. However, we can still use the Markov moment bound for the tail probability.

For any $\epsilon > 0$ and $0 < t < k$, the Markov inequality says

$$\mathbf{Pr}\left(\hat{d}_{gm} \geq (1+\epsilon)d\right) \leq \frac{\mathrm{E}\left(\hat{d}_{gm}\right)^t}{(1+\epsilon)^t d^t} = \frac{\cos^{kt}\left(\frac{\pi}{2k}\right)}{\cos^{k}\left(\frac{\pi t}{2k}\right)(1+\epsilon)^t}, \tag{98}$$

which can be minimized by choosing the optimum $t = t_1^*$, where

$$t_1^* = \frac{2k}{\pi}\tan^{-1}\left(\left(\log(1+\epsilon)-k\log\cos\left(\frac{\pi}{2k}\right)\right)\frac{2}{\pi}\right). \tag{99}$$

We need to make sure that $0 < t_1^* < k$. $t_1^* > 0$ because $\log\cos(.) \leq 0$; and $t_1^* < k$ because $\tan^{-1}(.) \leq \frac{\pi}{2}$, with equality held only when $k \to \infty$.

For $0 \leq \epsilon < 1$, we can prove an exponential bound for $\mathbf{Pr}\left(\hat{d}_{gm} \geq (1+\epsilon)d\right)$. First of all, note that we do not have to choose the optimum $t = t_1^*$. Using Taylor expansion, for small ϵ, t_1^* can be well approximated by

$$t_1^* \approx \frac{4k\epsilon}{\pi^2}+\frac{1}{2} \approx \frac{4k\epsilon}{\pi^2} = t_1^{**}. \tag{100}$$

Therefore, taking $t = t_1^{**} = \frac{4k\epsilon}{\pi^2}$, the tail bound becomes

$$\begin{aligned}
\mathbf{Pr}\left(\hat{d}_{gm} \geq (1+\epsilon)d\right) &\leq \frac{\cos^{kt_1^{**}}\left(\frac{\pi}{2k}\right)}{\cos^k\left(\frac{\pi t_1^{**}}{2k}\right)(1+\epsilon)^{t_1^{**}}} \\
&= \left(\frac{\cos^{t_1^{**}}\left(\frac{\pi}{2k}\right)}{\cos\left(\frac{2\epsilon}{\pi}\right)(1+\epsilon)^{4\epsilon/\pi^2}}\right)^k \\
&\leq \left(\frac{1}{\cos\left(\frac{2\epsilon}{\pi}\right)(1+\epsilon)^{4\epsilon/\pi^2}}\right)^k \\
&= \exp\left(-k\log\left(\cos\left(\frac{2\epsilon}{\pi}\right)(1+\epsilon)^{4\epsilon/\pi^2}\right)\right) \\
&\leq \exp\left(-k\frac{\epsilon^2}{8(1+\epsilon)}\right), \quad 0 \leq \epsilon < 1
\end{aligned} \tag{101}$$

where the last step can be proved by the trick that, if $f'(x) \geq 0$, for $x \geq 0$, then $f(x) \geq f(0)$. We need to use this trick recursively for a couple of times.

Now we can show that other tail bound $\mathbf{Pr}\left(\hat{d}_{gm} \leq (1-\epsilon)d\right)$, which appears simpler (because we know the probability is 0, if $\epsilon \geq 1$) but is in fact more troublesome as we can no longer apply the Markov inequality directly.

$$\begin{aligned}
&\mathbf{Pr}\left(\hat{d}_{gm} \leq (1-\epsilon)d\right) = \mathbf{Pr}\left(\cos\left(\frac{\pi}{2k}\right)^k \prod_{i=1}^k |x_i|^{1/k} \leq (1-\epsilon)d\right) \\
=&\mathbf{Pr}\left(\sum_{i=1}^k \log\left(|x_i|^{1/k}\right) \leq \log\left(\frac{(1-\epsilon)d}{\cos^k\left(\frac{\pi}{2k}\right)}\right)\right) \\
=&\mathbf{Pr}\left(\exp\left(\sum_{i=1}^k \log\left(|x_i|^{-t/k}\right)\right) \geq \exp\left(-t\log\left(\frac{(1-\epsilon)d}{\cos^k\left(\frac{\pi}{2k}\right)}\right)\right)\right), \quad 0 < t < k \\
\leq &\left(\frac{(1-\epsilon)}{\cos^k\left(\frac{\pi}{2k}\right)}\right)^t \frac{1}{\cos^k\left(\frac{\pi t}{2k}\right)}, \quad \text{(Chernoff bound)}
\end{aligned} \tag{102}$$

which is minimized at $t = t_2^*$

$$t_2^* = \frac{2k}{\pi}\tan^{-1}\left(\left(-\log(1-\epsilon) + k\log\cos\left(\frac{\pi}{2k}\right)\right)\frac{2}{\pi}\right), \tag{103}$$

provided $k > \frac{\pi^2}{8\epsilon}$, otherwise t_2^* may be less than 0.

Again, t_2^* can be replaced by its approximation

$$t_2^* \approx t_2^{**} = \frac{4k\epsilon}{\pi^2}, \tag{104}$$

provided $k > \frac{\pi^2}{4\epsilon}$, otherwise the probability upper bound may exceed one.

After some more manipulations, we obtain an exponential upper bound for $0 < \epsilon < 1$:

$$\mathbf{Pr}\left(\hat{d}_{gm} \leq (1-\epsilon)d\right) \leq \exp\left(-k\frac{\epsilon^2}{20}\right), \quad k > \frac{\pi^2}{4\epsilon} \tag{105}$$

This completes the proof of Lemma 3.

Appendix D. Proof of Lemma 4

Recall we have x_1, x_2, ..., x_k, i.i.d. $C(0,d)$, whose $\log$ likelihood function, denoted by $l(x;d)$, and the first three derivatives are

$$l(x;d) = \log(d) - \log(\pi) - \log(x^2+d^2), \tag{106}$$

$$l'(d) = \frac{1}{d} - \frac{2d}{x^2+d^2} \tag{107}$$

$$l''(d) = -\frac{1}{d^2} - \frac{2x^2-2d^2}{(x^2+d^2)^2} \tag{108}$$

$$l'''(d) = \frac{2}{d^3} + \frac{4d}{(x^2+d^2)^2} + \frac{8d(x^2-d^2)}{(x^2+d^2)^3} \tag{109}$$

The MLE $\hat{d}$ is asymptotically normal with mean d and variance $\frac{1}{k\mathrm{I}(d)}$, where $\mathrm{I}(d)$, the expected Fisher Information, is

$$\mathrm{I} = \mathrm{I}(d) = \mathrm{E}\left(-l''(d)\right) = \frac{1}{d^2} + 2\mathrm{E}\left(\frac{x^2-d^2}{(x^2+d^2)^2}\right) = \frac{1}{2d^2}, \tag{110}$$

because

$$\begin{aligned}
\mathrm{E}\left(\frac{x^2-d^2}{(x^2+d^2)^2}\right) &= \frac{d}{\pi}\int_{-\infty}^{\infty}\frac{x^2-d^2}{(x^2+d^2)^3}dx \\
&= \frac{d}{\pi}\int_{-\pi/2}^{\pi/2}\frac{d^2(\tan^2(t)-1)}{d^6/\cos^6(t)}\frac{d}{\cos^2(t)}dt \\
&= \frac{1}{d^2\pi}\int_{-\pi/2}^{\pi/2}\cos^2(t) - 2\cos^4(t)dt \\
&= \frac{1}{d^2\pi}\left(\frac{\pi}{2} - 2\frac{3}{8}\pi\right) = -\frac{1}{4d^2}
\end{aligned} \tag{111}$$

Therefore, we obtain

$$\mathrm{Var}\left(\hat{d}\right) = \frac{2d^2}{k} + O\left(\frac{1}{k^2}\right). \tag{112}$$

General formulas for the bias and higher moments of the MLE were derived about half century ago (Bartlett, 1953; Shenton and Bowman, 1963). We need to evaluate the expressions in (Shenton

and Bowman, 1963, 16a-16d), involving quite tedious algebra:

$$\mathrm{E}\left(\hat{d}\right) = d - \frac{[12]}{2k\mathrm{I}^2} + O\left(\frac{1}{k^2}\right) \tag{113}$$

$$\mathrm{Var}\left(\hat{d}\right) = \frac{1}{k\mathrm{I}} + \frac{1}{k^2}\left(-\frac{1}{\mathrm{I}} + \frac{[1^4] - [1^2 2] - [13]}{\mathrm{I}^3} + \frac{3.5[12]^2 - [1^3]^2}{\mathrm{I}^4}\right) + O\left(\frac{1}{k^3}\right) \tag{114}$$

$$\mathrm{E}\left(\hat{d} - \mathrm{E}\left(\hat{d}\right)\right)^3 = \frac{[1^3] - 3[12]}{k^2\mathrm{I}^2} + O\left(\frac{1}{k^3}\right) \tag{115}$$

$$\begin{aligned}\mathrm{E}\left(\hat{d} - \mathrm{E}\left(\hat{d}\right)\right)^4 = {} & \frac{3}{k^2\mathrm{I}^2} + \frac{1}{k^3}\left(-\frac{9}{\mathrm{I}^2} + \frac{7[1^4] - 6[1^2 2] - 10[13]}{\mathrm{I}^4}\right) \\ & + \frac{1}{k^3}\left(\frac{-6[1^3]^2 - 12[1^3][12] + 45[12]^2}{\mathrm{I}^5}\right) + O\left(\frac{1}{k^4}\right),\end{aligned} \tag{116}$$

where, after reformatting

$$\begin{aligned}&[12] = \mathrm{E}(l')^3 + \mathrm{E}(l'l''), \qquad [1^4] = \mathrm{E}(l')^4, \qquad [1^2 2] = \mathrm{E}(l''(l')^2) + \mathrm{E}(l')^4, \\ &[13] = \mathrm{E}(l')^4 + 3\mathrm{E}(l''(l')^2) + \mathrm{E}(l'l'''), \qquad [1^3] = \mathrm{E}(l')^3\end{aligned} \tag{117}$$

We will neglect most of the algebra. To help readers verifying the results, the following formula we derive may be useful:

$$\mathrm{E}\left(\frac{1}{x^2 + d^2}\right)^m = \frac{1 \times 3 \times 5 \times ... \times (2m-1)}{2 \times 4 \times 6 \times ... \times (2m)} \frac{1}{d^{2m}}, \qquad m = 1, 2, 3, ... \tag{118}$$

Without giving the details, we report

$$\begin{aligned}&\mathrm{E}\left(l'\right)^3 = 0, \qquad \mathrm{E}\left(l'l''\right) = -\frac{1}{2}\frac{1}{d^3}, \qquad \mathrm{E}\left(l'\right)^4 = \frac{3}{8}\frac{1}{d^4}, \\ &\mathrm{E}(l''(l')^2) = -\frac{1}{8}\frac{1}{d^4}, \qquad \mathrm{E}\left(l'l'''\right) = \frac{3}{4}\frac{1}{d^4}\end{aligned} \tag{119}$$

Hence

$$[12] = -\frac{1}{2}\frac{1}{d^3}, \qquad [1^4] = \frac{3}{8}\frac{1}{d^4}, \qquad [1^2 2] = \frac{1}{4}\frac{1}{d^4}, \qquad [13] = \frac{3}{4}\frac{1}{d^4}, \qquad [1^3] = 0 \tag{120}$$

Thus, we obtain

$$\mathrm{E}\left(\hat{d}\right) = d + \frac{d}{k} + O\left(\frac{1}{k^2}\right) \tag{121}$$

$$\mathrm{Var}\left(\hat{d}\right) = \frac{2d^2}{k} + \frac{7d^2}{k^2} + O\left(\frac{1}{k^3}\right) \tag{122}$$

$$\mathrm{E}\left(\hat{d} - \mathrm{E}\left(\hat{d}\right)\right)^3 = \frac{12d^3}{k^2} + O\left(\frac{1}{k^3}\right) \tag{123}$$

$$\mathrm{E}\left(\hat{d} - \mathrm{E}\left(\hat{d}\right)\right)^4 = \frac{12d^4}{k^2} + \frac{222d^4}{k^3} + O\left(\frac{1}{k^4}\right). \tag{124}$$

Because the raw MLE $\hat{d}$ has $O\left(\frac{1}{k}\right)$ bias, we recommend the the bias-corrected estimator

$$\hat{d}_1 = d\left(1 - \frac{1}{k}\right), \tag{125}$$

known as the "Bartlett correction." We obtain the moments of $\hat{d}_1$ as

$$\mathrm{E}\left(\hat{d}_1\right) = d + O\left(\frac{1}{k^2}\right) \tag{126}$$

$$\mathrm{Var}\left(\hat{d}_1\right) = \frac{2d^2}{k} + \frac{3d^2}{k^2} + O\left(\frac{1}{k^3}\right) \tag{127}$$

$$\mathrm{E}\left(\hat{d}_1 - \mathrm{E}\left(\hat{d}_1\right)\right)^3 = \frac{12d^3}{k^2} + O\left(\frac{1}{k^3}\right) \tag{128}$$

$$\mathrm{E}\left(\hat{d}_1 - \mathrm{E}\left(\hat{d}_1\right)\right)^4 = \frac{12d^4}{k^2} + \frac{186d^4}{k^3} + O\left(\frac{1}{k^4}\right), \tag{129}$$

by straightforward algebra. First, it is obvious that

$$\mathrm{E}\left(\hat{d} - d\right)^2 = \frac{2d^2}{k} + \frac{8d^2}{k^2} + O\left(\frac{1}{k^3}\right). \tag{130}$$

Then

$$\begin{aligned}
\mathrm{Var}\left(\hat{d}_1\right) &= \mathrm{E}\left(\hat{d}_1 - \mathrm{E}(\hat{d}_1)\right)^2 \\
&= \mathrm{E}\left(\hat{d}\left(1 - \frac{1}{k}\right) - d + O\left(\frac{1}{k^2}\right)\right)^2 \\
&= \mathrm{E}\left(\left(\hat{d} - d\right)\left(1 - \frac{1}{k}\right) - \frac{d}{k} + O\left(\frac{1}{k^2}\right)\right)^2 \\
&= \mathrm{E}\left(\hat{d} - d\right)^2\left(1 - \frac{2}{k}\right) + \frac{d^2}{k^2} - 2\frac{d}{k}\left(1 - \frac{1}{k}\right) + O\left(\frac{1}{k^3}\right) \\
&= \frac{2d^2}{k} + \frac{3d^2}{k^2} + O\left(\frac{1}{k^3}\right).
\end{aligned} \tag{131}$$

We can evaluate the higher central moments of $\hat{d}_1$ similarly, but we skip the messy algebra. Therefore, we have completed the proof for Lemma 4.