Journal of Machine Learning Research 8 (2007) 2497-2532 Submitted 5/06; Revised 10/06; Published 10/07

Nonlinear Estimators and Tail Bounds for Dimension Reduction in l_1 Using Cauchy Random Projections

Ping Li PINGLI@CORNELL.EDU
Department of Statistical Science
Faculty of Computing and Information Science
Cornell University
Ithaca, NY 14853, 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: Sam Roweis

Abstract

For[1] dimension reduction in the l_1 norm, the method of *Cauchy random projections* multiplies the original data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ with a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ ($k \ll D$) whose entries are i.i.d. samples of the standard Cauchy $C(0,1)$. Because of the impossibility result, one can not hope to recover the pairwise l_1 distances in $\mathbf{A}$ from $\mathbf{B} = \mathbf{A} \times \mathbf{R} \in \mathbb{R}^{n \times k}$, using linear estimators without incurring large errors. However, nonlinear estimators are still useful for certain applications in data stream computations, information retrieval, learning, and data mining.

We study three types of nonlinear estimators: the *sample median* estimators, the *geometric mean* estimators, and the *maximum likelihood* estimators (MLE). We derive tail bounds for the *geometric mean* estimators and establish that $k = O\left(\frac{\log n}{\varepsilon^2}\right)$ suffices with the constants explicitly given. Asymptotically (as $k \to \infty$), both the *sample median* and the *geometric mean* estimators are about 80% efficient compared to the MLE. We analyze the moments of the MLE and propose approximating its distribution of by an inverse Gaussian.

Keywords: dimension reduction, l_1 norm, Johnson-Lindenstrauss (JL) lemma, Cauchy random projections

1. Introduction

There has been considerable interest in the l_1 norm in statistics and machine learning, as it is now well-known that the l_1 distance is far more robust than the l_2 distance against "outliers" (Huber, 1981). It is sometimes a good practice to replace the l_2 norm minimization with the l_1 norm minimization, for example, the Least Absolute Deviation (LAD) Boost (Friedman, 2001). Chapelle et al. (1999) demonstrated that using the l_1 (Laplacian) radial basis kernel produced better classification

1. A preliminary version appeared in COLT 2007 (Li et al., 2007b).

results than the usual l_2 (Gaussian) radial basis kernel, in their histogram-based image classification project using support vector machines (SVM). Recently, it also becomes popular to use the l_1 norm for variable (feature) selection; success stories include LASSO (Tibshirani, 1996), LARS (Efron et al., 2004) and 1-norm SVM (Zhu et al., 2003).

This paper focuses on dimension reduction in the l_1 norm, in particular, on the method based on *Cauchy random projections*, which is a special case of *linear (stable) random projections* (Johnson and Schechtman, 1982; Indyk, 2000, 2006; Li, 2008).

The idea of *linear random projections* is to multiply the original data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ with a random projection matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$, resulting in a projected matrix $\mathbf{B} = \mathbf{AR} \in \mathbb{R}^{n \times k}$. We would like k to be as small as possible. If $k \ll D$, then it should be much more efficient to compute certain summary statistics (e.g., pairwise distances) from $\mathbf{B}$ as opposed to $\mathbf{A}$. Moreover, $\mathbf{B}$ may be small enough to reside in physical memory while $\mathbf{A}$ is often too large to fit in the main memory.

The choice of the random projection matrix $\mathbf{R}$ depends on which norm we would like to work with. For dimension reduction in l_p ($0 < p \leq 2$), it is common practice to construct $\mathbf{R}$ from i.i.d. samples of p-stable distributions (Johnson and Schechtman, 1982; Indyk, 2000, 2006; Li, 2008). In the stable distribution family (Zolotarev, 1986), normal is 2-stable and Cauchy is 1-stable. Thus, we will call random projections for l_2 and l_1, *normal random projections* and *Cauchy random projections*, respectively.

In *normal random projections* (Vempala, 2004), we can estimate the original pairwise l_2 distances in $\mathbf{A}$ directly using the corresponding l_2 distances in $\mathbf{B}$ (up to a normalizing constant). Furthermore, the Johnson-Lindenstrauss (JL) Lemma (Johnson and Lindenstrauss, 1984) provides the performance guarantee. We will review *normal random projections* in more detail in Section 2.

For *Cauchy random projections*, however, one shall not use the l_1 distance in $\mathbf{B}$ to approximate the original l_1 distance in $\mathbf{A}$, as the Cauchy distribution does not even have a finite first moment. The impossibility results (Brinkman and Charikar, 2003; Lee and Naor, 2004; Brinkman and Charikar, 2005) have proved that one can not hope to recover the l_1 distance using linear projections and linear estimators (e.g., sample mean), without incurring large errors. Fortunately, the impossibility results do not rule out nonlinear estimators, which may be still useful in certain applications in data stream computations, information retrieval, learning, and data mining.

In this paper, we study three types of nonlinear estimators: the *sample median* estimators, the *geometric mean* estimators, and the *maximum likelihood* estimators (MLE). The *sample median* and the *geometric mean* estimators are asymptotically (as $k \to \infty$) equivalent (i.e., both are about 80% efficient as the MLE), but the latter is more accurate at small sample size k. Furthermore, we derive explicit tail bounds for the *geometric mean* estimators and establish an analog of the JL Lemma for dimension reduction in l_1.

This analog of the JL Lemma for l_1 is weaker than the classical JL Lemma for l_2, as the geometric mean is not convex and hence is not a metric. Many efficient algorithms, such as some sub-linear time (using super-linear memory) nearest neighbor algorithms (Shakhnarovich et al., 2005), rely on metric properties (e.g., the triangle inequality). Nevertheless, nonlinear estimators may be still useful in important scenarios.

- *Estimating l_1 distances online*
  The original data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ requires $O(nD)$ storage space; and hence it is often too large for physical memory. The storage cost of materializing all pairwise distances is $O(n^2)$, which may be also too large for the memory. For example, in information retrieval, n could be the total number of word types or documents at Web scale. To avoid page faults, it may

be more efficient to estimate the distances *on the fly* from the projected data matrix $\mathbf{B}$ in the memory.

- *Computing all pairwise l_1 distances*
  In distance-based clustering, classification, and kernels (e.g., for SVM), we need to compute all pairwise distances in $\mathbf{A}$, at the cost of time $O(n^2D)$, which can be prohibitive, especially when $\mathbf{A}$ does not fit in the memory. Using *Cauchy random projections*, the cost is reduced to $O(nDk+n^2k)$.

- *Linear scan nearest neighbor searching*
  Nearest neighbor searching is notorious for being inefficient, especially when the data matrix $\mathbf{A}$ is too large for the memory. Searching for the nearest neighbors from the projected data matrix $\mathbf{B}$ (which is in the memory) becomes much more efficient, even by linear scans. The cost of searching for the nearest neighbor for one data point is reduced from $O(nD)$ to $O(nk)$.

- *Data stream computations*.
  Massive data streams come from Internet routers, phone switches, atmospheric observations, sensor networks, highway traffic, finance data, and more (Henzinger et al., 1999; Feigenbaum et al., 1999; Indyk, 2000; Babcock et al., 2002; Cormode et al., 2002). Unlike in the traditional databases, it is not common to store massive data streams; and hence the processing is often done *on the fly*. In data stream computations, Cauchy random projections can be used for (A): approximating the l_1 frequency moments for individual streams; (B): approximating the l_1 differences between a pair of streams.

We briefly comment on *random coordinate sampling*, another strategy for dimension reduction. One can randomly sample k columns from $\mathbf{A} \in \mathbb{R}^{n\times D}$ and estimate the summary statistics (including l_1 and l_2 distances). Despite its simplicity, this strategy has two major drawbacks. First, in heavy-tailed data, one may have to choose k very large in order to achieve a sufficient accuracy. Second, large data sets are often highly sparse, for example, text data (Dhillon and Modha, 2001) and market-basket data (Aggarwal and Wolf, 1999; Strehl and Ghosh, 2000). For sparse data, Li and Church (2005, 2007); Li et al. (2007a) provided an alternative coordinate sampling strategy, called *Conditional Random Sampling (CRS)*. For non-sparse data, however, methods based on *linear (stable) random projections* are superior.

The rest of the paper is organized as follows. Section 2 reviews *linear random projections*. Section 3 summarizes the main results for three types of nonlinear estimators. Section 4 presents the *sample median* estimators. Section 5 concerns the *geometric mean* estimators. Section 6 is devoted to the *maximum likelihood* estimators. Section 7 concludes the paper.

2. Introduction to Linear (Stable) Random Projections

We give a review on *linear random projections*, including *normal* and *Cauchy random projections*.

Denote the original data matrix by $\mathbf{A} \in \mathbb{R}^{n\times D}$, that is, 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 projection matrix and denote the entries of $\mathbf{R}$ by $\{r_{ij}\}_{i=1}^{D}{}_{j=1}^{k}$. The projected data matrix $\mathbf{B} = \mathbf{AR} \in \mathbb{R}^{n\times k}$. Let $\{v_i^{\mathrm{T}}\}_{i=1}^n \in \mathbb{R}^k$ be the ith row of $\mathbf{B}$, that is, $v_i = \mathbf{R}^{\mathrm{T}}u_i$.

For simplicity, we focus on the leading two rows, u_1 and u_2, in $\mathbf{A}$, and the leading two rows, v_1 and v_2, in $\mathbf{B}$. Define $\{x_j\}_{j=1}^k$ to be

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

If we sample r_{ij} i.i.d. from a *p-stable distribution* (Zolotarev, 1986), then x_j's are also i.i.d. samples of a p-stable distribution with a different scale parameter. In the family of stable distributions, normal $(p = 2)$ and Cauchy $(p = 1)$ are two important special cases.

2.1 Normal Random Projections

When r_{ij} is sampled from the standard normal, that is, $r_{ij} \sim N(0,1)$, i.i.d., then

$$x_j = v_{1,j} - v_{2,j} = \sum_{i=1}^{D} r_{ij}\left(u_{1,i} - u_{2,i}\right) \sim N\left(0, \sum_{i=1}^{D} |u_{1,i} - u_{2,i}|^2\right), \quad j = 1, 2, ..., k,$$

because a weighted sum of normals is also normal.

Denote the squared l_2 distance between u_1 and u_2 by

$$d_{l_2} = \|u_1 - u_2\|_2^2 = \sum_{i=1}^{D} |u_{1,i} - u_{2,i}|^2.$$

We can estimate d_{l_2} from the sample squared l_2 distance (i.e., sample mean):

$$\hat{d}_{l_2} = \frac{1}{k}\sum_{j=1}^{k} x_j^2.$$

Note that $k\hat{d}_{l_2}/d_{l_2}$ follows a Chi-square distribution with k degrees of freedom, χ_k^2. Therefore, it is easy to prove the following Lemma about the tail bounds:

Lemma 1

$$\mathbf{Pr}(\hat{d}_{l_2} - d_{l_2} \geq \varepsilon d_{l_2}) \leq \exp\left(-\frac{k}{2}\left(\varepsilon - \log(1+\varepsilon)\right)\right) = \exp\left(-k\frac{\varepsilon^2}{G_R}\right), \qquad \varepsilon > 0,$$

$$\mathbf{Pr}(\hat{d}_{l_2} - d_{l_2} \leq -\varepsilon d_{l_2}) \leq \exp\left(-\frac{k}{2}\left(-\varepsilon - \log(1-\varepsilon)\right)\right) = \exp\left(-k\frac{\varepsilon^2}{G_L}\right), \qquad 0 < \varepsilon < 1,$$

where the constants

$$G_R = \frac{2\varepsilon^2}{\varepsilon - \log(1+\varepsilon)} \leq \frac{4}{1 - \frac{2}{3}\varepsilon},$$

$$G_L = \frac{2\varepsilon^2}{-\varepsilon - \log(1-\varepsilon)} \leq \frac{4}{1 + \frac{2}{3}\varepsilon} \leq \frac{4}{1 - \frac{2}{3}\varepsilon}.$$

Proof *Using the standard Chernoff inequality (Chernoff, 1952),*

$$\begin{aligned}\mathbf{Pr}(\hat{d}_{l_2} - d_{l_2} \geq \varepsilon d_{l_2}) =& \mathbf{Pr}\left(k\hat{d}_{l_2}/d_{l_2} \geq k(1+\varepsilon)\right) \\ \leq& \frac{E\left(\exp(k\hat{d}_{l_2}/d_{l_2}t)\right)}{\exp((1+\varepsilon)kt)} \qquad (t>0) \\ =& \exp\left(-\frac{k}{2}\left(\log(1-2t)+2(1+\varepsilon)t\right)\right),\end{aligned}$$

which is minimized at $t = \frac{\varepsilon}{2(1+\varepsilon)}$. *Thus, for any* $\varepsilon > 0$

$$\mathbf{Pr}(\hat{d}_{l_2} - d_{l_2} > \varepsilon d_{l_2}) \leq \exp\left(-\frac{k}{2}\left(\varepsilon - \log(1+\varepsilon)\right)\right).$$

We can similarly prove the other tail bound for $\mathbf{Pr}(\hat{d}_{l_2} - d_{l_2} \leq -\varepsilon d_{l_2})$. ■

For convenience, sometimes we would like to write the tail bounds in a symmetric form

$$\mathbf{Pr}\left(\left|\hat{d}_{l_2} - d_{l_2}\right| \geq \varepsilon d_{l_2}\right) \leq 2\exp\left(-k\frac{\varepsilon^2}{G}\right), \quad 0 < \varepsilon < 1,$$

and we know that it suffices to let $G = \max\{G_R, G_L\} \leq \frac{4}{1-\frac{2}{3}\varepsilon}$.

Since there are in total $\frac{n(n-1)}{2} < \frac{n^2}{2}$ pairs among n data points, we would like to bound the tail probabilities simultaneously for all pairs. By the Bonferroni union bound, it suffices if

$$\frac{n^2}{2}\mathbf{Pr}\left(\left|\hat{d}_{l_2} - d_{l_2}\right| \geq \varepsilon d_{l_2}\right) \leq \delta,$$

that is, it suffices if

$$\frac{n^2}{2}2\exp\left(-k\frac{\varepsilon^2}{G}\right) \leq \delta \Longrightarrow k \geq G\frac{2\log n - \log\delta}{\varepsilon^2}.$$

Therefore, we obtain one version of the Johnson-Lindenstrauss (JL) Lemma:

Lemma 2 *If* $k \geq G\frac{2\log n - \log\delta}{\varepsilon^2}$, *where* $G = \frac{4}{1-\frac{2}{3}\varepsilon}$, *then with probability at least* $1-\delta$, *the squared* l_2 *distance between any pair of data points (among n data points) can be approximated within a* $1 \pm \varepsilon$ *factor* $(0 < \varepsilon < 1)$, *using the squared* l_2 *distance of the projected data after normal random projections.*

Many versions of the JL Lemma have been proved (Johnson and Lindenstrauss, 1984; Frankl and Maehara, 1987; Indyk and Motwani, 1998; Arriaga and Vempala, 1999; Dasgupta and Gupta, 2003; Indyk, 2000, 2001; Achlioptas, 2003; Arriaga and Vempala, 2006; Ailon and Chazelle, 2006).

Note that we do not have to use $r_{ij} \sim N(0,1)$ for dimension reduction in l_2. For example, we can sample r_{ij} from the following *sparse projection distribution*:

$$r_{ij} = \sqrt{s} \times \begin{cases} 1 & \text{with prob. } \frac{1}{2s} \\ 0 & \text{with prob. } 1 - \frac{1}{s} \\ -1 & \text{with prob. } \frac{1}{2s} \end{cases}. \tag{1}$$

When $1 \leq s \leq 3$, Achlioptas (2001, 2003) proved the JL Lemma for the above sparse projection. Recently, Li et al. (2006b) proposed *very sparse random projections* using $s \gg 3$ in (1), based on two practical considerations:

- D should be very large, otherwise there would be no need for dimension reduction.
- The original l_2 distance should make engineering sense, in that the second (or higher) moments should be bounded (otherwise various *term-weighting* schemes will be applied).

Based on these two practical assumptions, the projected data are asymptotically normal at a fast rate of convergence when $s = \sqrt{D}$ and the data have bounded third moments. Of course, *very sparse random projections* do not have worst case performance guarantees.

2.2 Cauchy Random Projections

In *Cauchy random projections*, we sample r_{ij} i.i.d. from the standard Cauchy distribution, that is, $r_{ij} \sim C(0,1)$. By the 1-stability of Cauchy (Zolotarev, 1986), we know that

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

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

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

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

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

$$\mathrm{E}\left(\exp(\sqrt{-1} z_1 t)\right) = \exp\left(-\gamma |t|\right),$$
$$\mathrm{E}\left(\exp\left(\sqrt{-1} t \sum_{i=1}^{D} c_i z_i\right)\right) = \exp\left(-\gamma \sum_{i=1}^{D} |c_i| t\right),$$

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

Therefore, in *Cauchy random projections*, the problem boils down to estimating the Cauchy scale parameter of $C(0,d)$ from k i.i.d. samples $x_j \sim C(0,d)$. Unlike in *normal random projections*, we can no longer estimate d from the sample mean (i.e., $\frac{1}{k}\sum_{j=1}^{k} |x_j|$) because $\mathrm{E}(x_j) = \infty$.

3. Main Results

Although the impossibility results (Lee and Naor, 2004; Brinkman and Charikar, 2005) have ruled out accurate estimators that are also metrics, there is enough information to recover d from k samples $\{x_j\}_{j=1}^{k}$, with high accuracy.

We analyze three types of nonlinear estimators: the *sample median* estimators, the *geometric mean* estimators, and the *maximum likelihood* estimators.

3.1 The Sample Median Estimators

The *sample median* estimator,

$$\hat{d}_{me} = \text{median}(|x_j|, j = 1, 2, ..., k),$$

is simple and computationally convenient. We recommend the bias-corrected version:

$$\hat{d}_{me,c} = \frac{\hat{d}_{me}}{b_{me}},$$

where

$$b_{me} = \int_0^1 \frac{(2m+1)!}{(m!)^2} \tan\left(\frac{\pi}{2}t\right)\left(t-t^2\right)^m dt, \quad k = 2m+1.$$

Here, for convenience, we only consider $k = 2m+1$, m = 1, 2, 3, ...
Some properties of $\hat{d}_{me,c}$:

- $\mathrm{E}\left(\hat{d}_{me,c}\right) = d$, that is, $\hat{d}_{me,c}$ is unbiased.
- When $k \geq 5$, the variance of $\hat{d}_{me,c}$ is

$$\begin{aligned}\mathrm{Var}\left(\hat{d}_{me,c}\right) =& d^2\left(\frac{(m!)^2}{(2m+1)!}\frac{\int_0^1 \tan^2\left(\frac{\pi}{2}t\right)\left(t-t^2\right)^m dt}{\left(\int_0^1 \tan\left(\frac{\pi}{2}t\right)\left(t-t^2\right)^m dt\right)^2} - 1\right), \quad k \geq 5\\ =& \frac{\pi^2}{4k}d^2 + O\left(\frac{1}{k^2}\right).\end{aligned}$$

- $b_{me} \geq 1$ and $b_{me} \to 1$ monotonically with increasing k.

3.2 The Geometric Mean Estimators

The *geometric mean* estimator

$$\hat{d}_{gm} = \prod_{j=1}^{k} |x_j|^{1/k}$$

has tail bounds

$$\begin{aligned}&\mathbf{Pr}\left(\hat{d}_{gm} \geq (1+\varepsilon)d\right) \leq U_{R,gm} = \exp\left(-k\frac{\varepsilon^2}{G_{R,gm}}\right), \quad \varepsilon > 0\\ &\mathbf{Pr}\left(\hat{d}_{gm} \leq (1-\varepsilon)d\right) \leq U_{L,gm} = \exp\left(-k\frac{\varepsilon^2}{G_{L,gm}}\right), \quad 0 < \varepsilon < 1\end{aligned}$$

where

$$\begin{aligned}G_{R,gm} &= \frac{\varepsilon^2}{\left(-\frac{1}{2}\log\left(1+\left(\frac{2}{\pi}\log(1+\varepsilon)\right)^2\right) + \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1+\varepsilon)\right)\log(1+\varepsilon)\right)},\\ G_{L,gm} &= \frac{\varepsilon^2}{\left(-\frac{1}{2}\log\left(1+\left(\frac{2}{\pi}\log(1-\varepsilon)\right)^2\right) + \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1-\varepsilon)\right)\log(1-\varepsilon)\right)}.\end{aligned}$$

Moreover, for small ε, we obtain the following convenient approximations:

$$G_{R,gm} = \frac{\pi^2}{2}\left(1+\varepsilon+\left(\frac{1}{12}+\frac{2}{3\pi^2}\right)\varepsilon^2+...\right),$$
$$G_{L,gm} = \frac{\pi^2}{2}\left(1-\varepsilon+\left(\frac{1}{12}+\frac{2}{3\pi^2}\right)\varepsilon^2+...\right).$$

Consequently, we establish an analog of the Johnson-Lindenstrauss (JL) Lemma for dimension reduction in l_1:

If $k \geq G_{gm}\frac{(2\log n-\log\delta)}{\varepsilon^2}$, then with probability at least $1-\delta$, one can recover the original l_1 distance between any pair of data points (among all n data points) within a $1\pm\varepsilon$ factor ($0<\varepsilon<1$), using $\hat{d}_{gm}$. The constant G_{gm} can be specified from $G_{R,gm}$ and $G_{L,gm}$: $G_{gm}=\max\{G_{R,gm},G_{L,gm}\}$.

To remove the bias and also reduce the variance, we recommend the bias-corrected *geometric mean* estimator:

$$\hat{d}_{gm,c} = \cos^k\left(\frac{\pi}{2k}\right)\prod_{j=1}^{k}|x_j|^{1/k},$$

which is unbiased and has variance

$$\mathrm{Var}\left(\hat{d}_{gm,c}\right) = d^2\left(\frac{\cos^{2k}\left(\frac{\pi}{2k}\right)}{\cos^k\left(\frac{\pi}{k}\right)}-1\right) = \frac{\pi^2}{4}\frac{d^2}{k}+\frac{\pi^4}{32}\frac{d^2}{k^2}+O\left(\frac{1}{k^3}\right).$$

We also derive tail bounds for $\hat{d}_{gm,c}$:

$$\mathbf{Pr}\left(\hat{d}_{gm,c}\geq(1+\varepsilon)d\right)\leq U_{R,gm,c}, \qquad \varepsilon>0$$
$$\mathbf{Pr}\left(\hat{d}_{gm,c}\leq(1-\varepsilon)d\right)\leq U_{L,gm,c}, \qquad 0<\varepsilon<1,$$

and show that, compared with $\hat{d}_{gm}$, the ratios of the tail bounds

$$\rho_{R,k}=\frac{U_{R,gm,c}}{U_{R,gm}}\rightarrow\rho_{R,\infty}=\frac{1}{(1+\varepsilon)^{C_1}}\exp\left(-\frac{\pi^2}{8}A_1+\frac{\pi}{2}C_1\tan\left(\frac{\pi^2}{2}A_1\right)\right),$$
$$\rho_{L,k}=\frac{U_{L,gm,c}}{U_{L,gm}}\rightarrow\rho_{L,\infty}=\frac{1}{(1-\varepsilon)^{C_2}}\exp\left(\frac{\pi^2}{8}A_2-\frac{\pi}{2}C_2\tan\left(\frac{\pi^2}{2}A_2\right)\right),$$

as $k\rightarrow\infty$, where A_1, C_1, A_2, and C_2 are only functions of ε.

3.3 The Maximum Likelihood Estimators

Denoted by $\hat{d}_{MLE,c}$, the bias-corrected maximum likelihood estimator (MLE) is

$$\hat{d}_{MLE,c}=\hat{d}_{MLE}\left(1-\frac{1}{k}\right),$$

where $\hat{d}_{MLE}$ solves a nonlinear MLE equation

$$-\frac{k}{\hat{d}_{MLE}}+\sum_{j=1}^{k}\frac{2\hat{d}_{MLE}}{x_j^2+\hat{d}_{MLE}^2}=0.$$

Some properties of $\hat{d}_{MLE,c}$:

- It is nearly unbiased, $\mathrm{E}\left(\hat{d}_{MLE,c}\right) = d + O\left(\frac{1}{k^2}\right)$.
- Its asymptotic variance is

$$\mathrm{Var}\left(\hat{d}_{MLE,c}\right) = \frac{2d^2}{k} + \frac{3d^2}{k^2} + O\left(\frac{1}{k^3}\right),$$

  that is, $\frac{\mathrm{Var}\left(\hat{d}_{MLE,c}\right)}{\mathrm{Var}\left(\hat{d}_{me,c}\right)} \to \frac{8}{\pi^2}$, $\frac{\mathrm{Var}\left(\hat{d}_{MLE,c}\right)}{\mathrm{Var}\left(\hat{d}_{gm,c}\right)} \to \frac{8}{\pi^2}$, as $k \to \infty$. ($\frac{8}{\pi^2} \approx 80\%$)
- Its distribution can be accurately approximated by an inverse Gaussian, at least in the small deviation range, which suggests the following approximate tail bound

$$\mathbf{Pr}\left(|\hat{d}_{MLE,c} - d| \geq \varepsilon d\right) \overset{\sim}{\leq} 2\exp\left(-\frac{\varepsilon^2/(1+\varepsilon)}{2\left(\frac{2}{k} + \frac{3}{k^2}\right)}\right), \quad 0 < \varepsilon < 1,$$

  which is verified by simulations for the tail probability $\geq 10^{-10}$ range.

4. The Sample Median Estimators

Recall in Cauchy random projections, $\mathbf{B} = \mathbf{AR}$, we denote the leading two rows in $\mathbf{A}$ by $u_1, u_2 \in \mathbb{R}^D$, and the leading two rows in $\mathbf{B}$ by $v_1, v_2 \in \mathbb{R}^k$. Our goal is to estimate the l_1 distance $d = |u_1 - u_2| = \sum_{i=1}^D |u_{1,i} - u_{2,i}|$ from $\{x_j\}_{j=1}^k$, $x_j = v_{1,j} - v_{2,j} \sim C(0,d)$, i.i.d.

A widely-used estimator in statistics is based on the sample inter-quantiles (Fama and Roll, 1968, 1971; McCulloch, 1986). For the symmetric Cauchy, the (absolute) *sample median* estimator

$$\hat{d}_{me} = \mathrm{median}\{|x_j|, j = 1, 2, ..., k\}$$

is convenient because the population median of absolute Cauchy is exactly d (Indyk, 2006).

It is well-known in statistics that $\hat{d}_{me}$, is asymptotically unbiased and normal; see Lemma 3. For small samples (e.g., $k \leq 20$), however, $\hat{d}_{me}$ is severely biased.

Lemma 3 *The* sample median *estimator, $\hat{d}_{me}$, is asymptotically unbiased and normal*

$$\sqrt{k}\left(\hat{d}_{me} - d\right) \overset{D}{\Longrightarrow} N\left(0, \frac{\pi^2}{4}d^2\right).$$

When $k = 2m+1$, m = 1, 2, 3, ..., the r^{th} moment of $\hat{d}_{me}$ can be represented as

$$E\left(\hat{d}_{me}\right)^r = d^r\left(\int_0^1 \frac{(2m+1)!}{(m!)^2}\tan^r\left(\frac{\pi}{2}t\right)\left(t - t^2\right)^m dt\right), \quad m \geq r \tag{2}$$

If $m < r$, then $E\left(\hat{d}_{me}\right)^r = \infty$.

Proof *Let $f(z;d)$ and $F(z;d)$ be the probability density and cumulative density respectively for $|C(0,d)|$:*

$$f(z;d) = \frac{2d}{\pi}\frac{1}{z^2 + d^2}, \quad F(z;d) = \frac{2}{\pi}\tan^{-1}\left(\frac{z}{d}\right), \quad z \geq 0.$$

The inverse of $F(z;d)$ *is* $F^{-1}(q;d) = d\tan\left(\frac{\pi}{2}q\right)$. *Here, we take* $q = 0.5$, *to consider the sample median. By the asymptotic normality of sample quantiles (David, 1981, Theorem 9.2), we know that*

$$\sqrt{k}\left(\hat{d}_{me} - d\right) \xrightarrow{D} N\left(0, \frac{\frac{1}{2}\frac{1}{2}}{\left(f\left(d\tan\left(\frac{\pi}{2}\frac{1}{2}\right);d\right)\times\tan\left(\frac{\pi}{2}\frac{1}{2}\right)\right)^2} = \frac{\pi^2}{4}d^2\right),$$

that is, $\hat{d}_{me}$ *is asymptotically unbiased and normal with the variance* $Var\left(\hat{d}_{me}\right) = \frac{\pi^2}{4k}d^2 + O\left(\frac{1}{k^2}\right)$.

For convenience, we assume $k = 2m+1$. *Again, by properties of sample quantile (David, 1981, Chapter 2.1), the probability density of* $\hat{d}_{me}$ *is*

$$f_{\hat{d}_{me}}(z) = \frac{(2m+1)!}{(m!)^2}\left(F(z;d)(1-F(z;d))\right)^m f(z;d),$$

from which we can write down the r^{th} *moment of* $\hat{d}_{me}$ *in (2), after some change of variables.* ■

Once we know $\mathrm{E}\left(\hat{d}_{me}\right)$, we can design an unbiased estimator as described in Lemma 4.

Lemma 4 *The estimator,*

$$\hat{d}_{me,c} = \frac{\hat{d}_{me}}{b_{me}},$$

is unbiased, that is, $E\left(\hat{d}_{me,c}\right) = d$, *where the bias-correction factor* b_{me} *is*

$$b_{me} = \frac{E\left(\hat{d}_{me}\right)}{d} = \int_0^1 \frac{(2m+1)!}{(m!)^2}\tan\left(\frac{\pi}{2}t\right)\left(t-t^2\right)^m dt, \qquad (k = 2m+1). \tag{3}$$

The variance of $\hat{d}_{me,c}$ *is*

$$Var\left(\hat{d}_{me,c}\right) = d^2\left(\frac{(m!)^2}{(2m+1)!}\frac{\int_0^1\tan^2\left(\frac{\pi}{2}t\right)\left(t-t^2\right)^m dt}{\left(\int_0^1\tan\left(\frac{\pi}{2}t\right)\left(t-t^2\right)^m dt\right)^2} - 1\right), \quad k = 2m+1 \geq 5.$$

$\hat{d}_{gm,c}$ *and* $\hat{d}_{gm}$ *are asymptotically equivalent, that is,*

$$\sqrt{k}\left(\hat{d}_{me,c} - d\right) \xrightarrow{D} N\left(0, \frac{\pi^2}{4}d^2\right).$$

The bias-correction factor b_{me} *is monotonically decreasing with increasing* m, *and*

$$b_{me} \geq 1, \qquad \lim_{m\to\infty} b_{me} = 1.$$

Proof *Most of the results follow directly from Lemma 3. Here we only show* b_{me} *decreases monotonically and* $b_{me} \to 1$ *as* $m \to \infty$.

Note that $\frac{(2m+1)!}{(m!)^2}\left(t-t^2\right)^m$, $0 \leq t \leq 1$, *is the probability density of a Beta distribution* $Beta(m+1, m+1)$, *whose rth moment is* $E\left(z^r\right) = \frac{(2m+1)!(m+r)!}{(2m+1+r)!m!}$.

By Taylor expansions (Gradshteyn and Ryzhik, 1994, 1.411.6),

$$\tan\left(\frac{\pi}{2}t\right) = \sum_{j=1}^{\infty} \frac{2^{2j}\left(2^{2j}-1\right)}{(2j)!}|B_{2j}|\left(\frac{\pi}{2}\right)^{2j-1} t^{2j-1},$$

where B_{2j} is the Bernoulli number *(Gradshteyn and Ryzhik, 1994, 9.61).*

Therefore,

$$b_{me} = \sum_{j=1}^{\infty} \frac{2^{2j}\left(2^{2j}-1\right)}{(2j)!}|B_{2j}|\left(\frac{\pi}{2}\right)^{2j-1} \frac{(2m+1)!(m+2j-1)!}{(2m+2j)!m!}.$$

It is easy to show that $\frac{(2m+1)!(m+2j-1)!}{(2m+2j)!m!}$ decreases monotonically with increasing m and it converges to $\left(\frac{1}{2}\right)^{2j-1}$. Thus, b_{me} also decreases monotonically with increasing m.

From the Taylor expansion of $\tan(t)$*, we know that*

$$b_{me} \to \sum_{j=1}^{\infty} \frac{2^{2j}\left(2^{2j}-1\right)}{(2j)!}|B_{2j}|\left(\frac{\pi}{2}\right)^{2j-1}\left(\frac{1}{2}\right)^{2j-1} = \tan\left(\frac{\pi}{2}\frac{1}{2}\right) = 1.$$

■

It is well-known that bias-corrections are not always beneficial because of the bias-variance trade-off phenomenon. In our case, because the correction factor $b_{me} \geq 1$ always, the bias-correction not only removes the bias of $\hat{d}_{me}$ but also reduces the variance of $\hat{d}_{me}$.

The bias-correction factor b_{me} can be numerically evaluated and tabulated, at least for small k. Figure 1 plots b_{me} as a function of k, indicating that $\hat{d}_{me}$ is severely biased when $k \leq 20$. When $k > 50$, the bias becomes negligible.

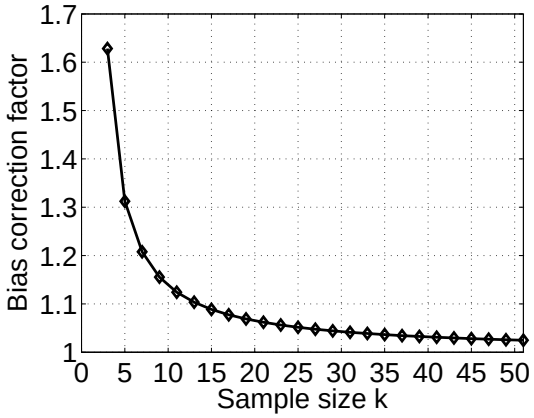


Figure 1: The bias correction factor, b_{me} in (3), as a function of $k = 2m+1$. After $k > 50$, the bias is negligible. Note that $b_{me} = \infty$ when $k = 1$.

5. The Geometric Mean Estimators

This section derives estimators based on the *geometric mean*, which are more accurate than the *sample median* estimators. The *geometric mean* estimators allow us to derive tail bounds in explicit forms and (consequently) establish an analog of the Johnson-Lindenstrauss (JL) Lemma for dimension reduction in the l_1 norm.

Lemma 5 *Assume $x \sim C(0,d)$. Then*

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

Proof *Using the integral table (Gradshteyn and Ryzhik, 1994, 3.221.1, page 337),*

$$E\left(|x|^\lambda\right) = \frac{2d}{\pi}\int_0^\infty \frac{y^\lambda}{y^2+d^2}dy = \frac{d^\lambda}{\pi}\int_0^\infty \frac{y^{\frac{\lambda-1}{2}}}{y+1}dy = \frac{d^\lambda}{\cos(\lambda\pi/2)}.$$

■

From Lemma 5, by taking $\lambda = \frac{1}{k}$, we obtain an unbiased estimator, $\hat{d}_{gm,c}$, based on the bias-corrected *geometric mean* in the next lemma, which is proved in Appendix A.

Lemma 6

$$\hat{d}_{gm,c} = \cos^k\left(\frac{\pi}{2k}\right)\prod_{j=1}^k |x_j|^{1/k}, \quad k > 1 \tag{4}$$

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

$$Var\left(\hat{d}_{gm,c}\right) = d^2\left(\frac{\cos^{2k}\left(\frac{\pi}{2k}\right)}{\cos^k\left(\frac{\pi}{k}\right)} - 1\right) = \frac{d^2}{k}\frac{\pi^2}{4} + \frac{\pi^4}{32}\frac{d^2}{k^2} + O\left(\frac{1}{k^3}\right).$$

The third and fourth central moments are

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

The higher (third or fourth) moments may be useful for approximating the distribution of $\hat{d}_{gm,c}$. In Section 6, we will show how to approximate the distribution of the maximum likelihood estimator by matching the first four moments (in the leading terms). We could apply the similar technique to approximate $\hat{d}_{gm,c}$. Fortunately, we do not have to do so because we are able to derive the exact tail bounds for $\hat{d}_{gm,c}$ in Lemma 9.

Note that in (4), as $k \to \infty$, the bias-correction term converges to 1 quickly:

$$\cos^k\left(\frac{\pi}{2k}\right) = \left(1 - \frac{1}{2}\left(\frac{\pi}{2k}\right)^2 + ...\right)^k = 1 - \frac{k}{2}\left(\frac{\pi}{2k}\right)^2 + ... = 1 - \frac{\pi^2}{8}\frac{1}{k} + ... \to 1.$$

When k is not too small (e.g., $k > 50$), the *geometric mean* estimator without bias-correction,

$$\hat{d}_{gm} = \prod_{j=1}^{k} |x_j|^{1/k}, \quad k > 1,$$

gives similar results as $\hat{d}_{gm,c}$. As shown in Figure 2, the ratios of the mean square errors (MSE)

$$\frac{\text{MSE}\left(\hat{d}_{gm}\right)}{\text{MSE}\left(\hat{d}_{gm,c}\right)} = \frac{\frac{1}{\cos^k\left(\frac{\pi}{k}\right)} - \frac{2}{\cos^k\left(\frac{\pi}{2k}\right)} + 1}{\frac{\cos^{2k}\left(\frac{\pi}{2k}\right)}{\cos^k\left(\frac{\pi}{k}\right)} - 1} \tag{5}$$

demonstrate that the two *geometric mean* estimators are similar when $k > 50$, in terms of the MSE.

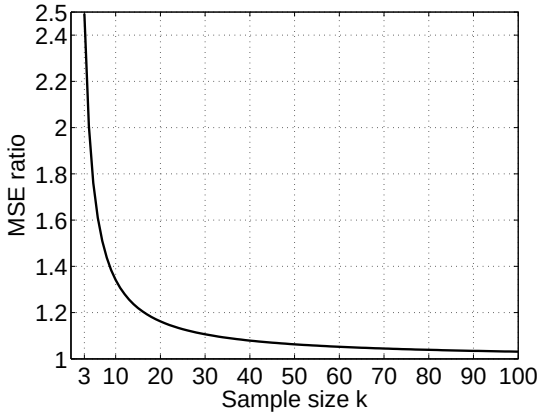


Figure 2: The ratios of the mean square errors (MSE) $\frac{\text{MSE}\left(\hat{d}_{gm}\right)}{\text{MSE}\left(\hat{d}_{gm,c}\right)}$ in (5) indicate that the difference between $\hat{d}_{gm}$ and $\hat{d}_{gm,c}$ becomes negligible when $k > 50$.

One advantage of $\hat{d}_{gm}$ is the convenience for deriving tail bounds. Thus, before presenting Lemma 9 for $\hat{d}_{gm,c}$, we prove tail bounds for $\hat{d}_{gm}$ in Lemma 7 (proved in Appendix B).

Lemma 7

$$\mathbf{Pr}\left(\hat{d}_{gm} \geq (1+\varepsilon)d\right) \leq U_{R,gm} = \exp\left(-k\frac{\varepsilon^2}{G_{R,gm}}\right), \quad \varepsilon > 0$$

$$\mathbf{Pr}\left(\hat{d}_{gm} \leq (1-\varepsilon)d\right) \leq U_{L,gm} = \exp\left(-k\frac{\varepsilon^2}{G_{L,gm}}\right), \quad 0 < \varepsilon < 1$$

where

$$G_{R,gm} = \frac{\varepsilon^2}{\left(-\frac{1}{2}\log\left(1+\left(\frac{2}{\pi}\log(1+\varepsilon)\right)^2\right) + \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1+\varepsilon)\right)\log(1+\varepsilon)\right)}, \tag{6}$$

$$G_{L,gm} = \frac{\varepsilon^2}{\left(-\frac{1}{2}\log\left(1+\left(\frac{2}{\pi}\log(1-\varepsilon)\right)^2\right) + \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1-\varepsilon)\right)\log(1-\varepsilon)\right)}. \tag{7}$$

Moreover, for small ε*, we have the following convenient approximations:*

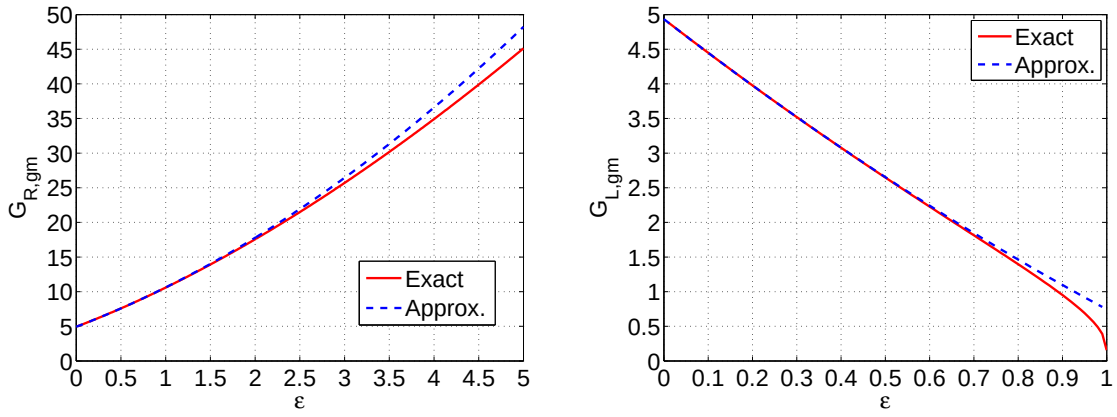
$$G_{R,gm} = \frac{\pi^2}{2}\left(1+\varepsilon+\left(\frac{1}{12}+\frac{2}{3\pi^2}\right)\varepsilon^2+...\right), \tag{8}$$

$$G_{L,gm} = \frac{\pi^2}{2}\left(1-\varepsilon+\left(\frac{1}{12}+\frac{2}{3\pi^2}\right)\varepsilon^2+...\right). \tag{9}$$

Consequently, as $\varepsilon \to 0+$*, we know*

$$G_{R,gm} \to \frac{\pi^2}{2}, \qquad G_{L,gm} \to \frac{\pi^2}{2}.$$

Figure 3 plots the constants $G_{R,gm}$ and $G_{L,gm}$ in (6) and (7), along with their convenient approximations (8) and (9). For $G_{R,gm}$, the exact and approximate expressions are indistinguishable when $\varepsilon < 2$. For $G_{L,gm}$, the exact and approximate expressions are indistinguishable when $\varepsilon < 0.7$. The plots also suggest that the approximations, (8) and (9), are upper bounds of the exact constants, (6) and (7), respectively.



(a) $G_{R,gm}$ (b) $G_{L,gm}$

Figure 3: We plot the constants $G_{R,gm}$ and $G_{L,gm}$ in (6) and (7), along with their convenient approximations (8) and (9).

Consequently, Lemma 7 establishes an analog of the Johnson-Lindenstrauss (JL) Lemma for dimension reduction in l_1:

Lemma 8 *If* $k \geq G_{gm}\frac{(2\log n - \log\delta)}{\varepsilon^2}$*, then with probability at least* $1-\delta$*, one can recover the original* l_1 *distance between any pair of data points (among all* n *data points) within a* $1\pm\varepsilon$ *factor (*$0<\varepsilon<1$*), using* $\hat{d}_{gm}$*. It suffices to specify the constant* $G_{gm} = \max\{G_{R,gm}, G_{L,gm}\}$.

Similarly, we derive tail bounds for the unbiased *geometric mean* estimator $\hat{d}_{gm,c}$, in Lemma 9, which is proved in Appendix C.

Lemma 9

$$\mathbf{Pr}\left(\hat{d}_{gm,c} \geq (1+\varepsilon)d\right) \leq U_{R,gm,c} = \frac{\cos^{kt_1^*}\left(\frac{\pi}{2k}\right)}{\cos^k\left(\frac{\pi t_1^*}{2k}\right)(1+\varepsilon)^{t_1^*}}, \qquad \varepsilon > 0$$

where

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

$$\mathbf{Pr}\left(\hat{d}_{gm,c} \leq (1-\varepsilon)d\right) \leq U_{L,gm,c} = \frac{(1-\varepsilon)^{t_2^*}}{\cos^k\left(\frac{\pi t_2^*}{2k}\right)\cos^{kt_2^*}\left(\frac{\pi}{2k}\right)}, \qquad 0<\varepsilon<1, \quad k \geq \frac{\pi^2}{8\varepsilon}$$

where

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

As $k \to \infty$, for any fixed ε, we have

$$\rho_{R,k} = \frac{U_{R,gm,c}}{U_{R,gm}} \to \rho_{R,\infty} = \frac{1}{(1+\varepsilon)^{C_1}}\exp\left(-\frac{\pi^2}{8}A_1 + \frac{\pi}{2}C_1\tan\left(\frac{\pi^2}{2}A_1\right)\right), \tag{10}$$

$$\rho_{L,k} = \frac{U_{L,gm,c}}{U_{L,gm}} \to \rho_{L,\infty} = \frac{1}{(1-\varepsilon)^{C_2}}\exp\left(\frac{\pi^2}{8}A_2 - \frac{\pi}{2}C_2\tan\left(\frac{\pi^2}{2}A_2\right)\right), \tag{11}$$

where $U_{R,gm}$ and $U_{L,gm}$ are upper bounds for $\hat{d}_{gm}$ as derived in Lemma 7, and

$$A_1 = \frac{2}{\pi}\left(\tan^{-1}\left(\log(1+\varepsilon)\frac{2}{\pi}\right)\right), \quad C_1 = \frac{1/2}{1+\left(\log(1+\varepsilon)\frac{2}{\pi}\right)^2},$$

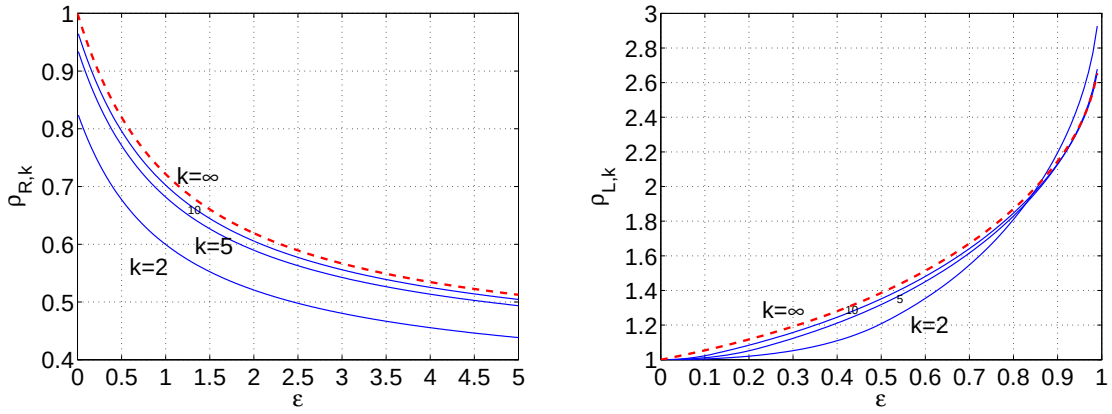
$$A_2 = \frac{2}{\pi}\left(\tan^{-1}\left(-\log(1-\varepsilon)\frac{2}{\pi}\right)\right), \quad C_2 = \frac{1/2}{1+\left(\log(1-\varepsilon)\frac{2}{\pi}\right)^2}.$$

Figure 4 plots the tail bound ratios $\rho_{R,k}$ and $\rho_{L,k}$ as defined in Lemma 9, indicating that the asymptotic expressions $\rho_{R,\infty}$ and $\rho_{L,\infty}$ are in fact very accurate even for small k (e.g., $k = 10$).

Figure 4 illustrates that introducing the bias-correction term in $\hat{d}_{gm,c}$ reduces the right tail bound but amplifies the left tail bound. Because the left tail bound is usually much smaller than the right tail bound, we expect that overall the bias-correction should be beneficial, as shown in Figure 5, which plots the overall ratio of tail bounds:

$$\rho_k = \frac{U_{R,gm,c} + U_{L,gm,c}}{U_{R,gm} + U_{L,gm}} = \frac{\frac{\cos^{kt_1^*}\left(\frac{\pi}{2k}\right)}{\cos^k\left(\frac{\pi t_1^*}{2k}\right)(1+\varepsilon)^{t_1^*}} + \frac{(1-\varepsilon)^{t_2^*}}{\cos^k\left(\frac{\pi t_2^*}{2k}\right)\cos^{kt_2^*}\left(\frac{\pi}{2k}\right)}}{\exp\left(-k\frac{\varepsilon^2}{G_{R,gm}}\right) + \exp\left(-k\frac{\varepsilon^2}{G_{L,gm}}\right)}. \tag{12}$$

Finally, Figure 6 compares $\hat{d}_{gm,c}$ with the sample median estimators $\hat{d}_{me}$ and $\hat{d}_{me,c}$, in terms of the mean square errors. $\hat{d}_{gm,c}$ is considerably more accurate than $\hat{d}_{me}$ at small k. The bias correction significantly reduces the mean square errors of $\hat{d}_{me}$.



(a) Right bound ratios $\rho_{R,k}$ (b) Left bound ratios $\rho_{L,k}$

Figure 4: Tail bound ratios $\rho_{R,k}$ and $\rho_{L,k}$ as defined in (10) and (11) for $k = 2, 5, 10$, along with the asymptotic expressions $\rho_{R,\infty}$ and $\rho_{L,\infty}$. The dashed curves correspond to $k = \infty$.

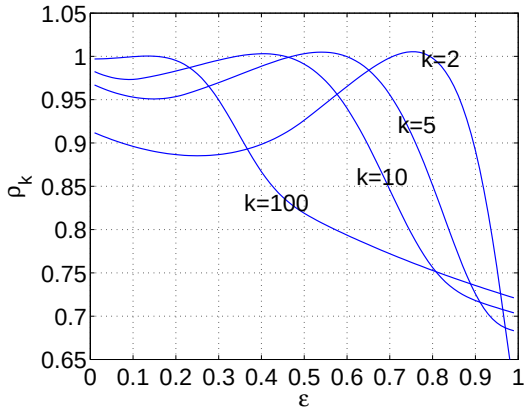


Figure 5: The overall ratios of tail bounds, ρ_k as defined in (12) are almost always below one, demonstrating that the bias-corrected estimator $\hat{d}_{gm,c}$ may exhibit better overall tail behavior than the biased estimator $\hat{d}_{gm}$.

6. The Maximum Likelihood Estimators

This section analyzes the maximum likelihood estimators (MLE), which are asymptotically optimum (in terms of the variance). In comparisons, the *sample median* and *geometric mean* estimators are not optimum. Our contribution in this section 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 is widely used. For example, Li et al. (2006a) applied the maximum likelihood method to *normal random projections* and provided an improved estimator of the l_2 distance by taking advantage of the marginal information.

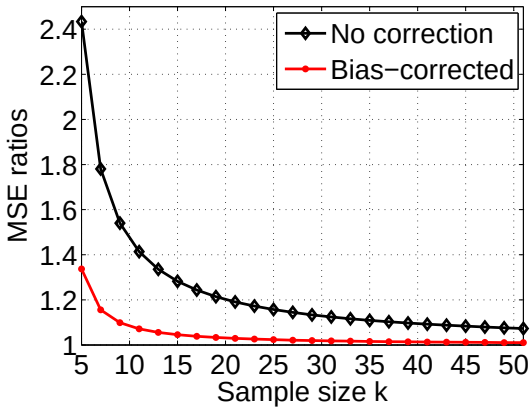


Figure 6: The ratios of the mean square errors (MSE), $\frac{\text{MSE}(\hat{d}_{me})}{\text{MSE}(\hat{d}_{gm,c})}$ and $\frac{\text{MSE}(\hat{d}_{me,c})}{\text{MSE}(\hat{d}_{gm,c})}$, demonstrate that the bias-corrected geometric mean estimator $\hat{d}_{gm,c}$ is considerably more accurate than the sample median estimator $\hat{d}_{me}$. The bias correction on $\hat{d}_{me}$ considerably reduces the MSE. Note that when $k = 3$, the ratios are ∞.

Recall our goal is to estimate d from k i.i.d. samples $x_j \sim C(0,d), j = 1,2,...,k$. The log joint likelihood of $\{x_j\}_{j=1}^k$ is

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

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

$$L'(d) = \frac{k}{d} - \sum_{j=1}^{k} \frac{2d}{x_j^2 + d^2},$$

$$L''(d) = -\frac{k}{d^2} - \sum_{j=1}^{k} \frac{2x_j^2 - 2d^2}{(x_j^2 + d^2)^2} = -\frac{L'(d)}{d} - 4\sum_{j=1}^{k} \frac{x_j^2}{(x_j^2 + d^2)^2}.$$

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

$$-\frac{k}{\hat{d}_{MLE}} + \sum_{j=1}^{k} \frac{2\hat{d}_{MLE}}{x_j^2 + \hat{d}_{MLE}^2} = 0. \tag{13}$$

Because $L''(\hat{d}_{MLE}) \leq 0$, $\hat{d}_{MLE}$ indeed maximizes the joint likelihood and is the only solution to the MLE equation (13). Solving (13) numerically is not difficult (e.g., a few iterations using the Newton's method). For a better accuracy, we recommend the following bias-corrected estimator:

$$\hat{d}_{MLE,c} = \hat{d}_{MLE}\left(1 - \frac{1}{k}\right).$$

Lemma 10 concerns the asymptotic moments of $\hat{d}_{MLE}$ and $\hat{d}_{MLE,c}$, proved in Appendix D.

Lemma 10 *Both $\hat{d}_{MLE}$ and $\hat{d}_{MLE,c}$ are asymptotically unbiased and normal. The first four moments of $\hat{d}_{MLE}$ are*

$$E\left(\hat{d}_{MLE}-d\right)=\frac{d}{k}+O\left(\frac{1}{k^2}\right)$$
$$Var\left(\hat{d}_{MLE}\right)=\frac{2d^2}{k}+\frac{7d^2}{k^2}+O\left(\frac{1}{k^3}\right)$$
$$E\left(\hat{d}_{MLE}-E(\hat{d}_{MLE})\right)^3=\frac{12d^3}{k^2}+O\left(\frac{1}{k^3}\right)$$
$$E\left(\hat{d}_{MLE}-E(\hat{d}_{MLE})\right)^4=\frac{12d^4}{k^2}+\frac{222d^4}{k^3}+O\left(\frac{1}{k^4}\right).$$

The first four moments of $\hat{d}_{MLE,c}$ are

$$E\left(\hat{d}_{MLE,c}-d\right)=O\left(\frac{1}{k^2}\right)$$
$$Var\left(\hat{d}_{MLE,c}\right)=\frac{2d^2}{k}+\frac{3d^2}{k^2}+O\left(\frac{1}{k^3}\right)$$
$$E\left(\hat{d}_{MLE,c}-E(\hat{d}_{MLE,c})\right)^3=\frac{12d^3}{k^2}+O\left(\frac{1}{k^3}\right)$$
$$E\left(\hat{d}_{MLE,c}-E(\hat{d}_{MLE,c})\right)^4=\frac{12d^4}{k^2}+\frac{186d^4}{k^3}+O\left(\frac{1}{k^4}\right).$$

The order $O\left(\frac{1}{k}\right)$ term of the variance, that is, $\frac{2d^2}{k}$, is well-known (Haas et al., 1970). We derive the bias-corrected estimator, $\hat{d}_{MLE,c}$, and the higher order moments using stochastic Taylor expansions (Bartlett, 1953; Shenton and Bowman, 1963; Ferrari et al., 1996; Cysneiros et al., 2001).

We will propose an inverse Gaussian distribution to approximate the distribution of $\hat{d}_{MLE,c}$, by matching the first four moments (at least in the leading terms).

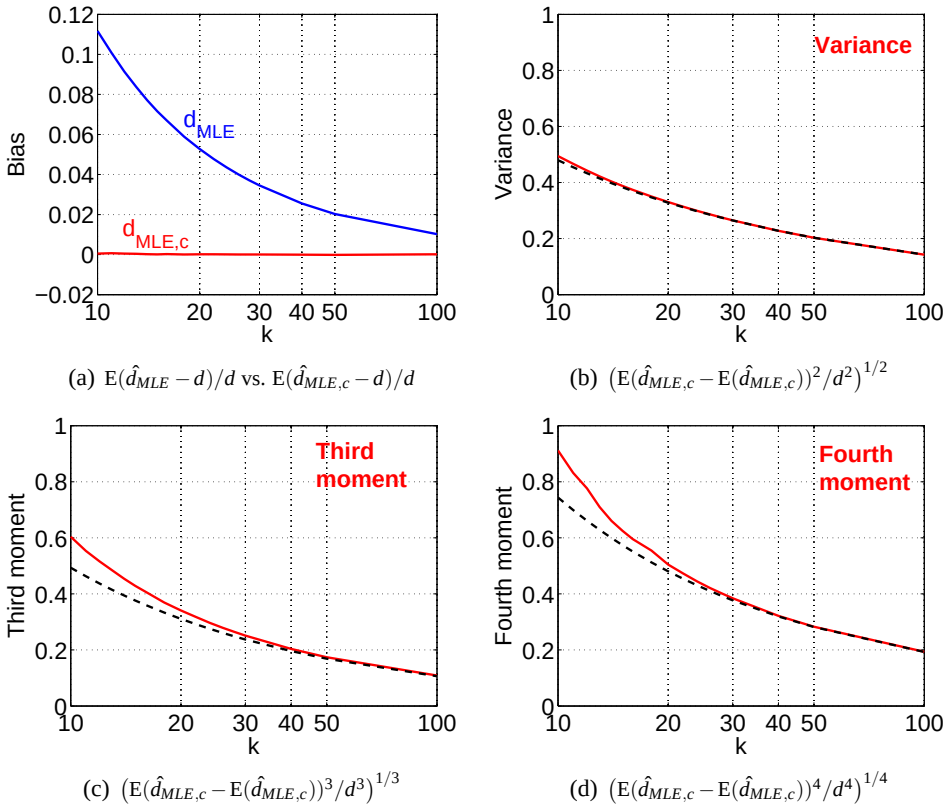
6.1 A Numerical Example

The maximum likelihood estimators are tested on some Microsoft Web crawl data, a term-by-document matrix with $D=2^{16}$ Web pages. We conduct Cauchy random projections and estimate the l_1 distances between words. In this experiment, we compare the empirical and (asymptotic) theoretical moments, using one pair of words. Figure 7 illustrates that the bias correction is effective and these (asymptotic) formulas for the first four moments of $\hat{d}_{MLE,c}$ in Lemma 10 are accurate, especially when $k \geq 20$.

6.2 Approximate Distributions

Theoretical analysis on the exact distribution of a maximum likelihood estimator is difficult. It is common practice to assume normality, which, however, is inaccurate.[2] The *Edgeworth expansion*

2. The simple normal approximation can be improved by taking advantage of the conditional density on the ancillary configuration statistic, based on the observations x_1, x_2, ..., x_k (Fisher, 1934; Lawless, 1972; Hinkley, 1978).



(a) $\mathrm{E}(\hat{d}_{MLE}-d)/d$ vs. $\mathrm{E}(\hat{d}_{MLE,c}-d)/d$

(b) $\left(\mathrm{E}(\hat{d}_{MLE,c}-\mathrm{E}(\hat{d}_{MLE,c}))^2/d^2\right)^{1/2}$

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

(d) $\left(\mathrm{E}(\hat{d}_{MLE,c}-\mathrm{E}(\hat{d}_{MLE,c}))^4/d^4\right)^{1/4}$

Figure 7: One pair of words are selected from a Microsoft term-by-document matrix with $D = 2^{16}$ Web pages. We conduct Cauchy random projections and estimate the l_1 distance between one pair of words using the maximum likelihood estimator $\hat{d}_{MLE}$ and the bias-corrected version $\hat{d}_{MLE,c}$. Panel (a) plots the biases of $\hat{d}_{MLE}$ and $\hat{d}_{MLE,c}$, indicating that the bias correction is effective. Panels (b), (c), and (d) plot the variance, third moment, and fourth moment of $\hat{d}_{MLE,c}$, respectively. The dashed curves are the theoretical asymptotic moments. When $k \geq 20$, the theoretical asymptotic formulas for moments are accurate.

improves the normal approximation by matching higher moments (Feller, 1971; Bhattacharya and Ghosh, 1978; Severini, 2000), which however, has some well-known drawbacks. The resultant expressions are quite sophisticated and are not accurate at the tails. It is possible that the approximate probability has values below zero. Also, Edgeworth expansions consider the support to be $(-\infty, \infty)$, while $\hat{d}_{MLE,c}$ is non-negative.

The *saddle-point approximation* (Jensen, 1995) in general improves Edgeworth expansions, often considerably. Unfortunately, we can not apply the saddle-point approximation in our case (at least not directly), because it requires a bounded moment generating function.

We propose approximating the distributions of $\hat{d}_{MLE,c}$ directly using some well-studied common distributions. We will first consider a gamma distribution with the same first two (asymptotic) moments of $\hat{d}_{MLE,c}$. That is, the gamma distribution will be asymptotically equivalent to the normal approximation. While a normal has zero third central moment, a gamma has nonzero third central

moment. This, to an extent, speeds up the rate of convergence. Another important reason why a gamma is more accurate is because it has the same support as $\hat{d}_{MLE,c}$, that is, $[0,\infty)$.

We will furthermore consider a *generalized gamma* distribution, which allows us to match the first three (asymptotic) moments of $\hat{d}_{MLE,c}$. 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 also matches the fourth central moment of $\hat{d}_{MLE,c}$ in the $O\left(\frac{1}{k^2}\right)$ term and almost in the $O\left(\frac{1}{k^3}\right)$ term. By simulations, the inverse Gaussian approximation is highly accurate.

Note that, since we are interested in the very small (e.g., 10^{-10}) tail probability range, $O\left(k^{-3/2}\right)$ is not too meaningful. For example, $k^{-3/2}=10^{-3}$ if $k=100$. Therefore, we will have to rely on simulations to assess the accuracy of the approximations. On the other hand, an upper bound may hold exactly (verified by simulations) even if it is based on an approximate distribution.

As the related work, Li et al. (2006c) applied gamma and generalized gamma approximations to model the performance measure distribution in some wireless communication channels using random matrix theory and produced accurate results in evaluating the error probabilities.

6.2.1 THE GAMMA APPROXIMATION

The gamma approximation is an obvious improvement over the normal approximation.[3] A gamma distribution, $G(\alpha,\beta)$, has two parameters, α and β, which can be determined by matching the first two (asymptotic) moments of $\hat{d}_{MLE,c}$. That is, we assume that $\hat{d}_{MLE,c}\sim G(\alpha,\beta)$, with

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

Assuming a gamma distribution, it is easy to obtain the following Chernoff bounds:

$$\mathbf{Pr}\left(\hat{d}_{MLE,c}\geq(1+\varepsilon)d\right)\overset{\sim}{\leq}\exp\left(-\alpha\left(\varepsilon-\log(1+\varepsilon)\right)\right),\qquad \varepsilon>0 \tag{14}$$

$$\mathbf{Pr}\left(\hat{d}_{MLE,c}\leq(1-\varepsilon)d\right)\overset{\sim}{\leq}\exp\left(-\alpha\left(-\varepsilon-\log(1-\varepsilon)\right)\right),\qquad 0<\varepsilon<1, \tag{15}$$

where we use $\overset{\sim}{\leq}$ to indicate that these inequalities are based on an approximate distribution.

Note that the distribution of $\hat{d}_{MLE}/d$ (and hence $\hat{d}_{MLE,c}/d$) is only a function of k (Antle and Bain, 1969; Haas et al., 1970). Therefore, we can evaluate the accuracy of the gamma approximation by simulations with $d=1$, as presented in Figure 8.

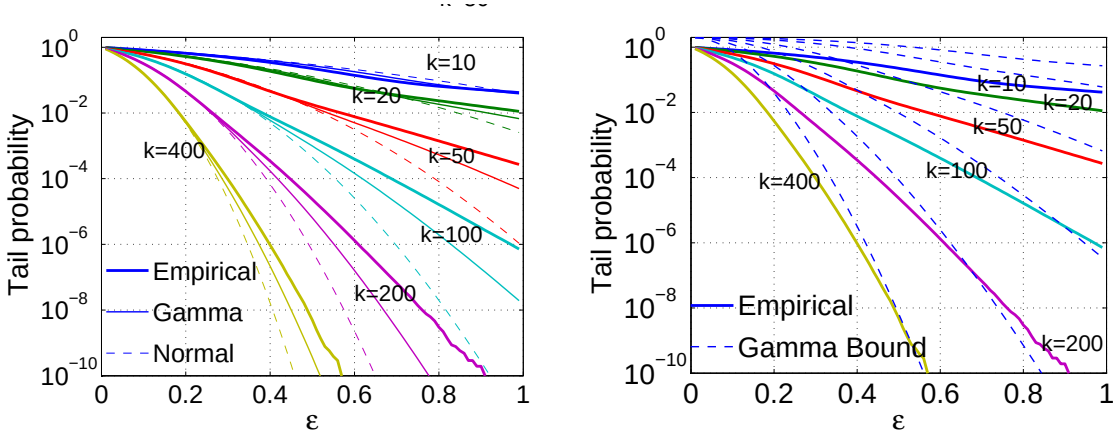
Figure 8(a) shows that both the gamma and normal approximations are fairly accurate when the tail probability $\geq 10^{-2}\sim 10^{-3}$; and the gamma approximation is obviously better.

Figure 8(b) compares the empirical tail probabilities with the gamma Chernoff upper bound (14)+(15), indicating that these bounds are reliable, when the tail probability $\geq 10^{-5}\sim 10^{-6}$.

6.2.2 THE INVERSE GAUSSIAN (GENERALIZED GAMMA) APPROXIMATION

The distribution of $\hat{d}_{MLE,c}$ can be well approximated by an inverse Gaussian distribution, which is a special case of the three-parameter generalized gamma distribution (Hougaard, 1986; Gerber, 1991;

3. Recall that, in *normal random projections* for dimension reduction in l_2 (see Lemma 1), the resultant estimator of the squared l_2 distance has a Chi-squared distribution, which is a special case of gamma.



(a) (b)

Figure 8: 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 $\hat{d}_{MLE,c}$ 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 $\hat{d}_{MLE,c}$. Panel (b) compares the empirical tail probabilities with the gamma upper bound (14)+(15).

Li et al., 2006c), 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

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

We can approximate the distribution of $\hat{d}_{MLE,c}$ by matching the first three moments, that is,

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

from which we obtain

$$\alpha = \frac{1}{\frac{2}{k} + \frac{3}{k^2}}, \quad \beta = \frac{2d}{k} + \frac{3d}{k^2}, \quad \eta = 2 + O\left(\frac{1}{k}\right). \tag{16}$$

Taking only the leading term for η, the generalized gamma approximation of $\hat{d}_{MLE,c}$ would be

$$GG\left(\frac{1}{\frac{2}{k} + \frac{3}{k^2}}, \frac{2d}{k} + \frac{3d}{k^2}, 2\right). \tag{17}$$

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, (17) is actually an inverse Gaussian (IG) distribution, which has a closed-form density function. Assuming $\hat{d}_{MLE,c} \sim IG(\alpha,\beta)$, with parameters α and β defined in (16), the moment generating function (MGF),

the probability density function (PDF), and cumulative density function (CDF) would be (Seshadri, 1993, Chapter 2) (Tweedie, 1957a,b)[4]

$$\begin{aligned}
&\mathrm{E}\left(\exp(\hat{d}_{MLE,c}t)\right) \cong \exp\left(\alpha\left(1-(1-2\beta t)^{1/2}\right)\right),\\
&f_{\hat{d}_{MLE,c}}(y) \cong \frac{\alpha\sqrt{\beta}}{\sqrt{2\pi}}y^{-\frac{3}{2}}\exp\left(-\frac{(y/\beta-\alpha)^2}{2y/\beta}\right) = \sqrt{\frac{\alpha d}{2\pi}}y^{-\frac{3}{2}}\exp\left(-\frac{(y-d)^2}{2y\beta}\right),\\
&\mathbf{Pr}\left(\hat{d}_{MLE,c}\le y\right) \cong \Phi\left(\sqrt{\frac{\alpha^2\beta}{y}}\left(\frac{y}{\alpha\beta}-1\right)\right)+e^{2\alpha}\Phi\left(-\sqrt{\frac{\alpha^2\beta}{y}}\left(\frac{y}{\alpha\beta}+1\right)\right)\\
&\qquad = \Phi\left(\sqrt{\frac{\alpha d}{y}}\left(\frac{y}{d}-1\right)\right)+e^{2\alpha}\Phi\left(-\sqrt{\frac{\alpha d}{y}}\left(\frac{y}{d}+1\right)\right),
\end{aligned}$$

where $\Phi(.)$ is the standard normal CDF, that is, $\Phi(z)=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$. Here we use $\cong$ to indicate that these equalities are based on an approximate distribution.

Assuming $\hat{d}_{MLE,c}\sim IG(\alpha,\beta)$, the fourth central moment should be

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

Lemma 10 has shown the true asymptotic fourth central moment:

$$\mathrm{E}\left(\hat{d}_{MLE,c}-\mathrm{E}\left(\hat{d}_{MLE,c}\right)\right)^4 = \frac{12d^4}{k^2}+\frac{186d^4}{k^3}+O\left(\frac{1}{k^4}\right).$$

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

Assuming $\hat{d}_{MLE,c}\sim IG(\alpha,\beta)$, the tail probability of $\hat{d}_{MLE,c}$ can be expressed as

$$\begin{aligned}
&\mathbf{Pr}\left(\hat{d}_{MLE,c}\ge(1+\varepsilon)d\right) \cong \Phi\left(-\varepsilon\sqrt{\frac{\alpha}{1+\varepsilon}}\right)-e^{2\alpha}\Phi\left(-(2+\varepsilon)\sqrt{\frac{\alpha}{1+\varepsilon}}\right),\quad \varepsilon>0\\
&\mathbf{Pr}\left(\hat{d}_{MLE,c}\le(1-\varepsilon)d\right) \cong \Phi\left(-\varepsilon\sqrt{\frac{\alpha}{1-\varepsilon}}\right)+e^{2\alpha}\Phi\left(-(2-\varepsilon)\sqrt{\frac{\alpha}{1-\varepsilon}}\right),\quad 0<\varepsilon<1.
\end{aligned}$$

Assuming $\hat{d}_{MLE,c}\sim IG(\alpha,\beta)$, it is easy to show the following Chernoff bounds:

$$\mathbf{Pr}\left(\hat{d}_{MLE,c}\ge(1+\varepsilon)d\right) \lesssim \exp\left(-\frac{\alpha\varepsilon^2}{2(1+\varepsilon)}\right),\quad \varepsilon>0 \tag{18}$$

$$\mathbf{Pr}\left(\hat{d}_{MLE,c}\le(1-\varepsilon)d\right) \lesssim \exp\left(-\frac{\alpha\varepsilon^2}{2(1-\varepsilon)}\right),\quad 0<\varepsilon<1. \tag{19}$$

4. 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 divisibility. 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*.

To see (18), assume $z \sim IG(\alpha, \beta)$. Then, using the Chernoff inequality:

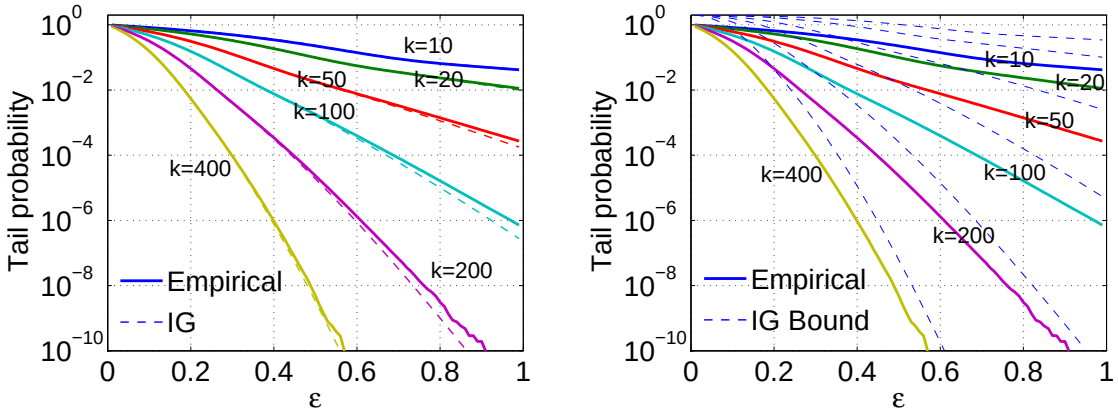
$$\begin{aligned}\mathbf{Pr}\left(z \geq (1+\varepsilon)d\right) \leq & \mathrm{E}\left(zt\right)\exp\left(-(1+\varepsilon)dt\right) \\ = & \exp\left(\alpha\left(1-(1-2\beta t)^{1/2}\right)-(1+\varepsilon)dt\right),\end{aligned}$$

whose minimum is $\exp\left(-\frac{\alpha\varepsilon^2}{2(1+\varepsilon)}\right)$, attained at $t=\left(1-\frac{1}{(1+\varepsilon)^2}\right)\frac{1}{2\beta}$. We can similarly show (19).

Combining (18) and (19) yields a symmetric approximate bound

$$\mathbf{Pr}\left(\left|\hat{d}_{MLE,c}-d\right| \geq \varepsilon d\right) \overset{\sim}{\leq} 2\exp\left(-\frac{\varepsilon^2/(1+\varepsilon)}{2\left(\frac{2}{k}+\frac{3}{k^2}\right)}\right), \quad 0<\varepsilon<1.$$

Figure 9 compares the inverse Gaussian approximation with the same simulations as presented in Figure 8, indicating 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 of $\hat{d}_{MLE,c}$. The Chernoff upper bounds for the inverse Gaussian are always reliable in our simulation range (the tail probability $\geq 10^{-10}$).



(a) (b)

Figure 9: We compare the inverse Gaussian approximation with the same simulations as presented in Figure 8. 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 (18)+(19). The upper bounds are all above the corresponding empirical curves, indicating that our proposed bounds are reliable at least in our simulation range.

7. Conclusion

In machine learning, it is well-known that the l_1 distance is far more robust than the l_2 distance against "outliers." Dimension reduction in the l_1 norm, however, has been proved *impossible* if we use *linear random projections* and *linear estimators*. In this study, we analyze three types of

nonlinear estimators for *Cauchy random projections*: the *sample median* estimators, the *geometric mean* estimators, and the *maximum likelihood* estimators. Our theoretical analysis has shown that these nonlinear estimators can accurately recover the original l_1 distance, even though none of them can be a metric.

The *sample median* estimators and the *geometric mean* estimators are asymptotically equivalent but the latter are more accurate at small sample size. We have derived explicit tail bounds for the *geometric mean* estimators in exponential forms. Using these tail bounds, we have established an analog of the Johnson-Lindenstrauss (JL) Lemma for dimension reduction in l_1, which is weaker than the classical JL Lemma for dimension reduction in l_2.

We conduct theoretic analysis on the *maximum likelihood* estimators (MLE), which are "asymptotically optimum." Both the *sample median* and *geometric mean* estimators are about 80% efficient as the MLE. We propose approximating the distribution of the MLE by an inverse Gaussian, which has the same support and matches the leading terms of the first four moments of the MLE. Approximate tail bounds have been provided based on the inverse Gaussian approximation. Verified by simulations, these approximate tail bounds hold at least in the $\geq 10^{-10}$ tail probability range.

Although these nonlinear estimators are not metrics, they are still useful for certain applications in, for example, data stream computations, information retrieval, learning and data mining, whenever the goal is to compute the l_1 distances efficiently using a small storage space in a single pass of the data.

Li (2008) generalized the *geometric mean* estimators to the stable distribution family, for dimension reduction in the l_p norm ($0 < p \leq 2$). Li (2008) also proposed the *harmonic mean* estimator for $p \to 0+$, which is far more accurate than the *geometric mean* estimator.[5] In addition, Li (2007) suggested *very sparse stable random projections* by replacing the stable distribution with a mixture of a symmetric Pareto distribution and point mass at the origin, for considerably simplifying the sampling procedure (to generate the projection matrix) and for achieving a significant cost reduction of matrix multiplication operations.

The general method of *linear (stable) random projections* is an appealing paradigm for applications involving massive, high-dimensional, non-sparse, and heavy-tailed data. If there is prior information that the data are highly sparse (e.g., text data), other alternative dimension reduction methods may be more suitable; for example, the new technique called *Conditional Random Sampling (CRS)* (Li and Church, 2005, 2007; Li et al., 2007a) was particularly designed for approximating distances (and other summary statistics) in highly sparse data.

Acknowledgments

We thank Piotr Indyk, Assaf Naor, Art Owen, and Anand Vidyashankar. Ping Li was partially supported by NSF Grant DMS-0505676 and Cornell University junior faculty startup fund. Trevor Hastie was partially supported by NSF Grant DMS-0505676 and NIH Grant 2R01 CA 72028-07.

5. Stable random projections with very small p ($p \to 0+$) have been applied to approximating the Hamming distances (Cormode et al., 2002, 2003) and the max-dominance norm (Cormode and Muthukrishnan, 2003).

Appendix A. Proof of Lemma 6

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

$$\hat{d}_{gm,c} = \cos^k\left(\frac{\pi}{2k}\right)\prod_{j=1}^{k}|x_j|^{1/k},$$

is unbiased, because, from Lemma 5,

$$\mathrm{E}\left(\hat{d}_{gm,c}\right) = \cos^k\left(\frac{\pi}{2k}\right)\prod_{j=1}^{k}\mathrm{E}\left(|x_j|^{1/k}\right) = \cos^k\left(\frac{\pi}{2k}\right)\prod_{j=1}^{k}\left(\frac{d^{1/k}}{\cos\left(\frac{\pi}{2k}\right)}\right) = d.$$

The variance is

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

because

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

Some more algebra can similarly show the third and fourth central moments:

$$\begin{aligned}\mathrm{E}\left(\hat{d}_{gm,c} - \mathrm{E}\left(\hat{d}_{gm,c}\right)\right)^3 &= \frac{3\pi^4}{16}\frac{d^3}{k^2} + O\left(\frac{1}{k^3}\right),\\ \mathrm{E}\left(\hat{d}_{gm,c} - \mathrm{E}\left(\hat{d}_{gm,c}\right)\right)^4 &= \frac{3\pi^4}{16}\frac{d^4}{k^2} + O\left(\frac{1}{k^3}\right).\end{aligned}$$

Appendix B. Proof of Lemma 7

We will use the Markov moment bound, because $\hat{d}_{gm,c}$ does not have a moment generating function ($\mathrm{E}\left(\hat{d}_{gm,c}\right)^t = \infty$ if $t \geq k$). In fact, even when the Chernoff bound is applicable, for any positive random variable, the Markov moment bound is always sharper than the Chernoff bound (Philips and Nelson, 1995; Lugosi, 2004).

By the Markov moment bound, for any $\varepsilon > 0$ and $0 < t < k$,

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

whose minimum is attained at $t = k\frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1+\varepsilon)\right)$. Thus

$$\begin{aligned}
&\mathbf{Pr}\left(\hat{d}_{gm} \geq (1+\varepsilon)d\right) \\
&\leq \exp\left(-k\left(\log\left(\cos\left(\tan^{-1}\left(\frac{2}{\pi}\log(1+\varepsilon)\right)\right)\right) + \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1+\varepsilon)\right)\log(1+\varepsilon)\right)\right) \\
&= \exp\left(-k\left(-\frac{1}{2}\log\left(1+\left(\frac{2}{\pi}\log(1+\varepsilon)\right)^2\right) + \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1+\varepsilon)\right)\log(1+\varepsilon)\right)\right) \\
&= \exp\left(-k\frac{\varepsilon^2}{G_{R,gm}}\right),
\end{aligned}$$

where

$$G_{R,gm} = \frac{\varepsilon^2}{\left(-\frac{1}{2}\log\left(1+\left(\frac{2}{\pi}\log(1+\varepsilon)\right)^2\right) + \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1+\varepsilon)\right)\log(1+\varepsilon)\right)}.$$

Again, by the Markov moment bound, for any $0 < \varepsilon < 1$,

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

whose minimum is attained at $t = -k\frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1-\varepsilon)\right)$. Thus

$$\begin{aligned}
&\mathbf{Pr}\left(\hat{d}_{gm} \leq (1-\varepsilon)d\right) \\
&\leq \exp\left(-k\left(\log\left(\cos\left(\tan^{-1}\left(\frac{2}{\pi}\log(1-\varepsilon)\right)\right)\right) + \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1-\varepsilon)\right)\log(1-\varepsilon)\right)\right) \\
&= \exp\left(-k\left(-\frac{1}{2}\log\left(1+\left(\frac{2}{\pi}\log(1-\varepsilon)\right)^2\right) + \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1-\varepsilon)\right)\log(1-\varepsilon)\right)\right) \\
&= \exp\left(-k\frac{\varepsilon^2}{G_{L,gm}}\right),
\end{aligned}$$

where

$$G_{L,gm} = \frac{\varepsilon^2}{\left(-\frac{1}{2}\log\left(1+\left(\frac{2}{\pi}\log(1-\varepsilon)\right)^2\right) + \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1-\varepsilon)\right)\log(1-\varepsilon)\right)}.$$

Finally, we derive convenient approximations for $G_{G,gm}$ and $G_{L,gm}$, for small ε (e.g., $\varepsilon < 1$). Recall that, for $|x| < 1$, we have

$$\begin{aligned}
\log(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + ... \\
\tan^{-1}(x) &= x - \frac{x^3}{3} + ...
\end{aligned}$$

Thus for small ε, we have

$$
\begin{aligned}
&G_{R,gm} \\
=&\frac{\varepsilon^2}{\left(-\frac{1}{2}\log\left(1+\left(\frac{2}{\pi}\log(1+\varepsilon)\right)^2\right)+\frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1+\varepsilon)\right)\log(1+\varepsilon)\right)} \\
=&\frac{\varepsilon^2}{-\frac{1}{2}\left(\left(\frac{2}{\pi}\log(1+\varepsilon)\right)^2-\frac{1}{2}\left(\frac{2}{\pi}\log(1+\varepsilon)\right)^4+...\right)+\frac{2}{\pi}\left(\frac{2}{\pi}\log(1+\varepsilon)-\frac{1}{3}\left(\frac{2}{\pi}\log(1+\varepsilon)\right)^3+...\right)\log(1+\varepsilon)} \\
=&\frac{\frac{\pi^2}{2}\varepsilon^2}{\log^2(1+\varepsilon)\left(1-\frac{2}{3\pi^2}\log^2(1+\varepsilon)+...\right)}=\frac{\frac{\pi^2}{2}\varepsilon^2}{\left(\varepsilon-\frac{\varepsilon^2}{2}+\frac{\varepsilon^3}{3}+...\right)^2\left(1-\frac{2}{3\pi^2}\varepsilon^2+...\right)} \\
=&\frac{\pi^2}{2}\left(1-\frac{\varepsilon}{2}+\frac{\varepsilon^2}{3}+...\right)^{-2}\left(1-\frac{2}{3\pi^2}\varepsilon^2+...\right)^{-1} \\
=&\frac{\pi^2}{2}\left(1+\varepsilon-\frac{2}{3}\varepsilon^2+\frac{(-2)(-3)}{2}\left(-\frac{\varepsilon}{2}+\frac{\varepsilon^2}{3}+...\right)^2+...\right)\left(1+\frac{2}{3\pi^2}\varepsilon^2+...\right) \\
=&\frac{\pi^2}{2}\left(1+\varepsilon+\left(\frac{1}{12}+\frac{2}{3\pi^2}\right)\varepsilon^2+...\right).
\end{aligned}
$$

Similarly, for small ε, we have

$$
\begin{aligned}
&G_{L,gm} \\
=&\frac{\varepsilon^2}{\left(-\frac{1}{2}\log\left(1+\left(\frac{2}{\pi}\log(1-\varepsilon)\right)^2\right)+\frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1-\varepsilon)\right)\log(1-\varepsilon)\right)} \\
=&\frac{\varepsilon^2}{-\frac{1}{2}\left(\left(\frac{2}{\pi}\log(1-\varepsilon)\right)^2-\frac{1}{2}\left(\frac{2}{\pi}\log(1-\varepsilon)\right)^4+...\right)+\frac{2}{\pi}\left(\frac{2}{\pi}\log(1-\varepsilon)-\frac{1}{3}\left(\frac{2}{\pi}\log(1-\varepsilon)\right)^3+...\right)\log(1-\varepsilon)} \\
=&\frac{\frac{\pi^2}{2}\varepsilon^2}{\log^2(1-\varepsilon)\left(1-\frac{2}{3\pi^2}\log^2(1-\varepsilon)+...\right)}=\frac{\frac{\pi^2}{2}\varepsilon^2}{\left(-\varepsilon-\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}+...\right)^2\left(1-\frac{2}{3\pi^2}\varepsilon^2+...\right)} \\
=&\frac{\pi^2}{2}\left(1+\frac{\varepsilon}{2}+\frac{\varepsilon^2}{3}+...\right)^{-2}\left(1-\frac{2}{3\pi^2}\varepsilon^2+...\right)^{-1} \\
=&\frac{\pi^2}{2}\left(1-\varepsilon-\frac{2}{3}\varepsilon^2+\frac{(-2)(-3)}{2}\left(\frac{\varepsilon}{2}+\frac{\varepsilon^2}{3}+...\right)^2+...\right)\left(1+\frac{2}{3\pi^2}\varepsilon^2+...\right) \\
=&\frac{\pi^2}{2}\left(1-\varepsilon+\left(\frac{1}{12}+\frac{2}{3\pi^2}\right)\varepsilon^2+...\right).
\end{aligned}
$$

Appendix C. Proof of Lemma 9

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

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

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

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

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

Now we show the other tail bound $\mathbf{Pr}\left(\hat{d}_{gm,c} \le (1-\varepsilon)d\right)$. Let $0 < t < k$.

$$\begin{aligned}\mathbf{Pr}\left(\hat{d}_{gm,c} \le (1-\varepsilon)d\right) =& \mathbf{Pr}\left(\cos\left(\frac{\pi}{2k}\right)^k \prod_{j=1}^k |x_j|^{1/k} \le (1-\varepsilon)d\right)\\ =& \mathbf{Pr}\left(\prod_{j=1}^k |x_j|^{-t/k} \ge \left(\frac{(1-\varepsilon)d}{\cos^k\left(\frac{\pi}{2k}\right)}\right)^{-t}\right) \le \left(\frac{(1-\varepsilon)}{\cos^k\left(\frac{\pi}{2k}\right)}\right)^t \frac{1}{\cos^k\left(\frac{\pi t}{2k}\right)},\end{aligned}$$

which is minimized at $t = t_2^*$

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

provided $k \ge \frac{\pi^2}{8\varepsilon}$, otherwise t_2^* may be less than 0. To see this, in order for $t_2^* \ge 0$, we must have

$$\log(1-\varepsilon) \le k\log\cos\left(\frac{\pi}{2k}\right), \text{ i.e., } 1-\varepsilon \le \cos^k\left(\frac{\pi}{2k}\right).$$

Because

$$\cos^k\left(\frac{\pi}{2k}\right) \ge \left(1 - \frac{1}{2}\left(\frac{\pi}{2k}\right)^2\right)^k \ge 1 - \frac{\pi^2}{8k},$$

it suffices if $1-\varepsilon \le 1 - \frac{\pi^2}{8k}$, that is, $k \ge \frac{\pi^2}{8\varepsilon}$.

Now we prove the asymptotic (as $k \to \infty$) expressions for the ratios of tail bounds, another way to compare $\hat{d}_{gm}$ and $\hat{d}_{gm,c}$.

First, we consider the right tail bounds. For large k, the optimal $t = t_1^*$ can be approximated as

$$\begin{aligned}t_1^* =& \frac{2k}{\pi}\tan^{-1}\left(\left(\log(1+\varepsilon) - k\log\cos\left(\frac{\pi}{2k}\right)\right)\frac{2}{\pi}\right)\\ \sim& \frac{2k}{\pi}\tan^{-1}\left(\log(1+\varepsilon)\frac{2}{\pi} - k\frac{2}{\pi}\log\left(1 - \frac{\pi^2}{8k^2}\right)\right)\\ \sim& \frac{2k}{\pi}\tan^{-1}\left(\log(1+\varepsilon)\frac{2}{\pi} + \frac{\pi}{4k}\right)\\ \sim& \frac{2k}{\pi}\left(\tan^{-1}\left(\log(1+\varepsilon)\frac{2}{\pi}\right) + \frac{\pi}{4k}\frac{1}{1+\left(\log(1+\varepsilon)\frac{2}{\pi}\right)^2}\right) \qquad \text{(Taylor expansion)}\\ \sim& \frac{2k}{\pi}\left(\tan^{-1}\left(\log(1+\varepsilon)\frac{2}{\pi}\right)\right) + \frac{1/2}{1+\left(\log(1+\varepsilon)\frac{2}{\pi}\right)^2}\\ =& kA_1 + C_1,\end{aligned}$$

where

$$A_1 = \frac{2}{\pi}\left(\tan^{-1}\left(\log(1+\varepsilon)\frac{2}{\pi}\right)\right), \quad C_1 = \frac{1/2}{1+\left(\log(1+\varepsilon)\frac{2}{\pi}\right)^2}.$$

Note that, in the asymptotic decomposition, $t_1^* \sim kA_1 + C_1$, the term kA_1 is the optimal "t" in proving the right tail bound for $\hat{d}_{gm}$. Thus to study the asymptotic ratio of the right tail bounds, we only need to keep track of the additional terms in

$$\frac{\cos^{kt_1^*}\left(\frac{\pi}{2k}\right)}{\cos^k\left(\frac{\pi t_1^*}{2k}\right)(1+\varepsilon)^{t_1^*}}.$$

Because

$$\cos^{kt_1^*}\left(\frac{\pi}{2k}\right) \sim \left(1-\frac{\pi^2}{8k^2}\right)^{k^2A_1+kC_1} \sim \exp\left(-\frac{\pi^2}{8}A_1\right),$$

and

$$\begin{aligned}
\cos^k\left(\frac{\pi t_1^*}{2k}\right) \sim & \cos^k\left(\frac{\pi A_1}{2}+\frac{\pi C_1}{2k}\right) \\
\sim & \cos^k\left(\frac{\pi A_1}{2}\right)\left(1-\frac{\pi C_1}{2k}\tan\left(\frac{\pi}{2}A_1\right)\right)^k \quad \text{(Taylor expansion)} \\
\sim & \cos^k\left(\frac{\pi A_1}{2}\right)\exp\left(-\frac{\pi C_1}{2}\tan\left(\frac{\pi}{2}A_1\right)\right),
\end{aligned}$$

we know the ratio of the right tail bounds

$$\rho_{R,k} = \frac{\frac{\cos^{kt_1^*}\left(\frac{\pi}{2k}\right)}{\cos^k\left(\frac{\pi t_1^*}{2k}\right)(1+\varepsilon)^{t_1^*}}}{\exp\left(-k\frac{\varepsilon^2}{G_{R,gm}}\right)} \to \rho_{R,\infty} = \frac{1}{(1+\varepsilon)^{C_1}}\exp\left(-\frac{\pi^2}{8}A_1+\frac{\pi}{2}C_1\tan\left(\frac{\pi^2}{2}A_1\right)\right).$$

Next, we consider the left tail bound. First, we obtain an asymptotic decomposition of t_2^*:

$$\begin{aligned}
t_2^* = & \frac{2k}{\pi}\tan^{-1}\left(\left(-\log(1-\varepsilon)+k\log\cos\left(\frac{\pi}{2k}\right)\right)\frac{2}{\pi}\right) \\
\sim & \frac{2k}{\pi}\tan^{-1}\left(-\log(1-\varepsilon)\frac{2}{\pi}-\frac{\pi}{4k}\right) \\
\sim & \frac{2k}{\pi}\left(\tan^{-1}\left(-\log(1-\varepsilon)\frac{2}{\pi}\right)\right)-\frac{1/2}{1+\left(\log(1-\varepsilon)\frac{2}{\pi}\right)^2} \quad \text{(Taylor expansion)} \\
= & kA_2 - C_2,
\end{aligned}$$

where

$$A_2 = \frac{2}{\pi}\left(\tan^{-1}\left(-\log(1-\varepsilon)\frac{2}{\pi}\right)\right), \quad C_2 = \frac{1/2}{1+\left(\log(1-\varepsilon)\frac{2}{\pi}\right)^2}.$$

Again, in the above asymptotic decomposition, $t_1^* \sim kA_2 - C_2$, the term kA_2 is the optimal "t" in proving the left tail bound for $\hat{d}_{gm}$. Thus to study the asymptotic ratio of the left tail bounds, we only need to keep track of the additional terms in

$$\frac{(1-\varepsilon)^{t_2^*}}{\cos^{kt_2^*}\left(\frac{\pi}{2k}\right)} \frac{1}{\cos^k\left(\frac{\pi t_2^*}{2k}\right)}.$$

Because

$$\cos^{kt_2^*}\left(\frac{\pi}{2k}\right) \sim \left(1-\frac{\pi^2}{8k^2}\right)^{k^2A_2-kC_2} \sim \exp\left(-\frac{\pi^2}{8}A_2\right),$$

and

$$\begin{aligned}\cos^k\left(\frac{\pi t_2^*}{2k}\right) \sim & \cos^k\left(\frac{\pi A_2}{2}-\frac{\pi C_2}{2k}\right)\\ \sim & \cos^k\left(\frac{\pi A_2}{2}\right)\left(1+\frac{\pi C_2}{2k}\tan\left(\frac{\pi}{2}A_2\right)\right)^k\\ \sim & \cos^k\left(\frac{\pi A_2}{2}\right)\exp\left(\frac{\pi C_2}{2}\tan\left(\frac{\pi}{2}A_2\right)\right),\end{aligned}$$

we know the ratio of the left tail bounds

$$\rho_{L,k} = \frac{\frac{(1-\varepsilon)^{t_2^*}}{\cos^{kt_2^*}\left(\frac{\pi}{2k}\right)}\frac{1}{\cos^k\left(\frac{\pi t_2^*}{2k}\right)}}{\exp\left(-k\frac{\varepsilon^2}{G_{L,gm}}\right)} \to \rho_{L,\infty} = \frac{1}{(1-\varepsilon)^{C_2}}\exp\left(\frac{\pi^2}{8}A_2 - \frac{\pi}{2}C_2\tan\left(\frac{\pi^2}{2}A_2\right)\right).$$

Appendix D. Proof of Lemma 10

Assume $x \sim C(0,d)$. The log likelihood $l(x;d)$ and its first three derivatives are

$$\begin{aligned}l(x;d) &= \log(d) - \log(\pi) - \log(x^2+d^2),\\ l'(d) &= \frac{1}{d} - \frac{2d}{x^2+d^2},\\ l''(d) &= -\frac{1}{d^2} - \frac{2x^2-2d^2}{(x^2+d^2)^2},\\ l'''(d) &= \frac{2}{d^3} + \frac{4d}{(x^2+d^2)^2} + \frac{8d(x^2-d^2)}{(x^2+d^2)^3}.\end{aligned}$$

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

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

because

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

Therefore, we obtain

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

General formulas for the bias and higher moments of the MLE are available in Bartlett (1953) and Shenton and Bowman (1963). We need to evaluate the expressions in (Shenton and Bowman, 1963, 16a-16d), involving tedious algebra:

$$\begin{aligned}
&\mathrm{E}\left(\hat{d}_{MLE}\right) = d-\frac{[12]}{2k\mathrm{I}^2}+O\left(\frac{1}{k^2}\right) \\
&\mathrm{Var}\left(\hat{d}_{MLE}\right) = \frac{1}{k\mathrm{I}}+\frac{1}{k^2}\left(-\frac{1}{\mathrm{I}}+\frac{[1^4]-[1^22]-[13]}{\mathrm{I}^3}+\frac{3.5[12]^2-[1^3]^2}{\mathrm{I}^4}\right)+O\left(\frac{1}{k^3}\right) \\
&\mathrm{E}\left(\hat{d}_{MLE}-\mathrm{E}\left(\hat{d}_{MLE}\right)\right)^3 = \frac{[1^3]-3[12]}{k^2\mathrm{I}^3}+O\left(\frac{1}{k^3}\right) \\
&\mathrm{E}\left(\hat{d}_{MLE}-\mathrm{E}\left(\hat{d}_{MLE}\right)\right)^4 = \frac{3}{k^2\mathrm{I}^2}+\frac{1}{k^3}\left(-\frac{9}{\mathrm{I}^2}+\frac{7[1^4]-6[1^22]-10[13]}{\mathrm{I}^4}\right) \\
&\qquad\qquad +\frac{1}{k^3}\left(\frac{-6[1^3]^2-12[1^3][12]+45[12]^2}{\mathrm{I}^5}\right)+O\left(\frac{1}{k^4}\right),
\end{aligned}$$

where, after re-formatting,

$$\begin{aligned}
&[12] = \mathrm{E}(l')^3+\mathrm{E}(l'l''), \qquad [1^4] = \mathrm{E}(l')^4, \qquad [1^22] = \mathrm{E}(l''(l')^2)+\mathrm{E}(l')^4, \\
&[13] = \mathrm{E}(l')^4+3\mathrm{E}(l''(l')^2)+\mathrm{E}(l'l'''), \qquad [1^3] = \mathrm{E}(l')^3.
\end{aligned}$$

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

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

Without giving the detail, we report

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

Hence

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

Thus, we obtain

$$\begin{aligned}
&\mathrm{E}\left(\hat{d}_{MLE}\right) = d + \frac{d}{k} + O\left(\frac{1}{k^2}\right) \\
&\mathrm{Var}\left(\hat{d}_{MLE}\right) = \frac{2d^2}{k} + \frac{7d^2}{k^2} + O\left(\frac{1}{k^3}\right) \\
&\mathrm{E}\left(\hat{d}_{MLE} - \mathrm{E}\left(\hat{d}_{MLE}\right)\right)^3 = \frac{12d^3}{k^2} + O\left(\frac{1}{k^3}\right) \\
&\mathrm{E}\left(\hat{d}_{MLE} - \mathrm{E}\left(\hat{d}_{MLE}\right)\right)^4 = \frac{12d^4}{k^2} + \frac{222d^4}{k^3} + O\left(\frac{1}{k^4}\right).
\end{aligned}$$

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

$$\hat{d}_{MLE,c} = \hat{d}_{MLE}\left(1 - \frac{1}{k}\right),$$

whose first four moments are, after some algebra,

$$\begin{aligned}
&\mathrm{E}\left(\hat{d}_{MLE,c}\right) = d + O\left(\frac{1}{k^2}\right) \\
&\mathrm{Var}\left(\hat{d}_{MLE,c}\right) = \frac{2d^2}{k} + \frac{3d^2}{k^2} + O\left(\frac{1}{k^3}\right) \\
&\mathrm{E}\left(\hat{d}_{MLE,c} - \mathrm{E}\left(\hat{d}_{MLE,c}\right)\right)^3 = \frac{12d^3}{k^2} + O\left(\frac{1}{k^3}\right) \\
&\mathrm{E}\left(\hat{d}_{MLE,c} - \mathrm{E}\left(\hat{d}_{MLE,c}\right)\right)^4 = \frac{12d^4}{k^2} + \frac{186d^4}{k^3} + O\left(\frac{1}{k^4}\right).
\end{aligned}$$

References

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

Dimitris Achlioptas. Database-friendly random projections. In *PODS*, pages 274–281, Santa Barbara, CA, 2001.

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

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *STOC*, pages 557–563, Seattle, WA, 2006.

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. *Machine Learning*, 63(2):161–182, 2006.

Rosa Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *FOCS*, pages 616–623, New York, 1999.

Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom. Models and issues in data stream systems. In *PODS*, pages 1–16, Madison, WI, 2002.

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

Rabi N. Bhattacharya and Jayanta K. Ghosh. On the validity of the formal edgeworth expansion. *The Annals of Statistics*, 6(2):434–451, 1978.

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

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

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

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

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

Graham Cormode and S. Muthukrishnan. Estimating dominance norms of multiple data streams. In *ESA*, pages 148–160, 2003.

Graham Cormode, Mayur Datar, Piotr Indyk, and S. Muthukrishnan. Comparing data streams using hamming norms (how to zero in). In *VLDB*, pages 335–345, Hong Kong, China, 2002.

Graham Cormode, Mayur Datar, Piotr Indyk, and S. Muthukrishnan. Comparing data streams using hamming norms (how to zero in). *IEEE Transactions on Knowledge and Data Engineering*, 15 (3):529–540, 2003.

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

Herbert A. David. *Order Statistics*. John Wiley & Sons, Inc., New York, NY, second edition, 1981.

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

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

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.

Joan Feigenbaum, Sampath Kannan, Martin Strauss, and Mahesh Viswanathan. An approximate l_1-difference algorithm for massive data streams. In *FOCS*, pages 501–511, New York, 1999.

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

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.

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

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

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

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

Izrail S. Gradshteyn and Iosif M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, New York, fifth 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.

Monika R. Henzinger, Prabhakar Raghavan, and Sridhar Rajagopalan. *Computing on Data Streams*. American Mathematical Society, Boston, MA, USA, 1999.

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

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

Peter J. Huber. *Robust Statistics*. Wiley, New York, NY, 1981.

Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of ACM*, 53(3):307–323, 2006.

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 *FOCS*, pages 10–33, Las Vegas, NV, 2001.

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

Jens Ledet Jensen. *Saddlepoint Approximations*. Oxford University Press, New York, 1995.

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

William B. Johnson and Gideon Schechtman. Embedding l_p into l_1. *Acta. Math.*, 149:71–85, 1982.

Jerry F. Lawless. Conditional confidence 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.

Ping Li. Very sparse stable random projections for dimension reduction in l_α ($0 < \alpha \leq 2$) norm. In *KDD*, San Jose, CA, 2007.

Ping Li. Estimators and tail bounds for dimension reduction in l_α ($0 < \alpha \leq 2$) using stable random projections. In *SODA*, 2008.

Ping Li and Kenneth W. Church. A sketch algorithm for estimating two-way and multi-way associations. *Computational Linguistics*, 33(3):305–354, 2007.

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

Ping Li, Trevor J. Hastie, and Kenneth W. Church. Improving random projections using marginal information. In *COLT*, pages 635–649, Pittsburgh, PA, 2006a.

Ping Li, Trevor J. Hastie, and Kenneth W. Church. Very sparse random projections. In *KDD*, pages 287–296, Philadelphia, PA, 2006b.

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, 2006c.

Ping Li, Kenneth W. Church, and Trevor J. Hastie. Conditional random sampling: A sketch-based sampling technique for sparse data. In *NIPS*, pages 873–880, Vancouver, BC, Canada, 2007a.

Ping Li, Trevor J. Hastie, and Kenneth W. Church. Nonlinear estimators and tail bounds for dimensional reduction in l_1 using Cauchy random projections. In *COLT*, 2007b.

Gabor Lugosi. Concentration-of-measure inequalities. *Lecture Notes*, 2004.

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

Thomas K. Philips and Randolph Nelson. The moment bound is tighter than Chernoff's bound for positive tail probabilities. *The American Statistician*, 49(2):175–178, 1995.

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

Thomas A. Severini. *Likelihood Methods in Statistics*. Oxford University Press, New York, 2000.

Gregory Shakhnarovich, Trevor Darrell, and Piotr Indyk, editors. *Nearest-Neighbor Methods in Learning and Vision, Theory and Practice*. The MIT Press, Cambridge, MA, 2005.

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

Alexander Strehl and Joydeep Ghosh. A scalable approach to balanced, high-dimensional clustering of market-baskets. In *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.

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

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

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

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

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