Improving Random Projections Using Marginal Information

Ping Li[1], Trevor J. Hastie[1], and Kenneth W. Church[2]

[1] Department of Statistics, Stanford University, Stanford CA 94305, USA,
{pingli, hastie}@stat.stanford.edu
[2] Microsoft Research, One Microsoft Way, Redmond WA 98052, USA,
church@microsoft.com

Abstract. We present an improved version of random projections that takes advantage of marginal norms. Using a maximum likelihood estimator (MLE), margin-constrained random projections can improve estimation accuracy considerably. Theoretical properties of this estimator are analyzed in detail.

1 Introduction

Random projections[1] have been used in machine learning [2–6] and many other applications in data mining and information retrieval, e.g., [7–12].

One application of random projections is to compute the Gram matrix $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ efficiently, where $\mathbf{A} \in \mathbb{R}^{n \times D}$ is a collection of n data points $\in \mathbb{R}^D$. In modern applications, n and D can be very large hence computing $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ is prohibitive. The method of random projections multiplies $\mathbf{A}$ with a projection matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$, which typically consists of i.i.d. $N(0,1)$ entries.[3] Let $\mathbf{B} = \frac{1}{\sqrt{k}}\mathbf{A}\mathbf{R}$. Suppose u_i^{T} is the i^{th} row of $\mathbf{A}$, and the corresponding i^{th} row in $\mathbf{B}$ is v_i^{T}, then as shown in Lemma 1.3 of [1]

$$\mathrm{E}\left(\|v_i - v_j\|^2\right) = \|u_i - u_j\|^2, \qquad \mathrm{Var}\left(\|v_i - v_j\|^2\right) = \frac{2}{k}\|u_i - u_j\|^4. \quad (1)$$

Therefore, one can compute pairwise distances in k dimensions, as opposed to D dimensions. When $k \ll \min(n, D)$, the savings from $O(n^2 D)$ to $O(n^2 k + nDk)$ is enormous.

Random projections generate a small sketch (i.e., $\mathbf{B}$) of the original data. $\mathbf{B}$ may be small enough to reside in the main memory. Operations such as query optimization or nearest neighbor searching can then be conducted on the much smaller space in the main memory, avoiding disk IO, which can be convenient for applications in databases, information retrieval, etc.

1.1 Our Results

We improve random projections by taking advantage of marginal norms, which we might as well compute, since they are useful and no harder to compute than the random

[3] The only necessary condition for preserving pairwise distance is that $\mathbf{R}$ consists of i.i.d. entries with zero mean[2]. The case of i.i.d. $N(0,1)$ entries is the easiest to analyze.

projections. Given an $n \times D$ matrix $\mathbf{A}$, it costs just $O(nD)$ time to compute the marginal norms, considerably less than the $O(nDk)$ time required for k random projections.

We will propose an estimator based on maximum likelihood. Some maximum likelihood estimators suffer from severe bias, slow rate of convergence toward normality, multiple roots, etc. These concerns will be addressed.

Some (approximate) tail bounds will also be presented, which can improve the current well-known tail bounds and consequently also improve some Johnson and Lindenstrauss (JL) embedding bounds in a practical sense.[4]

2 Random Projections Using Marginal Norms

Recall $u_i \in \mathbb{R}^D$ denotes data vectors in the original space and $v_i = \frac{1}{\sqrt{k}}\mathbf{R}^{\mathrm{T}} u_i \in \mathbb{R}^k$ denotes vectors in the projection space, where the projection matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ consists of i.i.d $N(0,1)$ entries. We assume that the marginal norms, $\|u_i\|^2$, are known. As $\|u_1 - u_2\|^2 = \|u_1\|^2 + \|u_2\|^2 - 2u_1^{\mathrm{T}}u_2$, we only need to estimate the dot product $u_1^{\mathrm{T}}u_2$.

For convenience, we denote

$$a = u_1^{\mathrm{T}}u_2, \quad m_1 = \|u_1\|^2, \quad m_2 = \|u_2\|^2, \quad d = \|u_1 - u_2\|^2 = m_1 + m_2 - 2a.$$

The following lemma is proved in Appendix A.

Lemma 1. *Given $u_1, u_2 \in \mathbb{R}^D$, and a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ consisting of i.i.d. standard normal $N(0,1)$ entries, if we let $v_1 = \frac{1}{\sqrt{k}}\mathbf{R}^T u_1$, and $v_2 = \frac{1}{\sqrt{k}}\mathbf{R}^T u_2$, then*[5]

$$E\left(v_1^T v_2\right) = a, \ \mathit{Var}\left(v_1^T v_2\right) = \frac{1}{k}\left(m_1 m_2 + a^2\right), \ E\left(v_1^T v_2 - a\right)^3 = \frac{2a}{k^2}\left(3m_1 m_2 + a^2\right) \quad (2)$$

with the moment generating function

$$E\left(\exp(v_1^T v_2 t)\right) = \left(1 - \frac{2}{k}at - \frac{1}{k^2}\left(m_1 m_2 - a^2\right)t^2\right)^{-\frac{k}{2}}, \quad (3)$$

where $\frac{-k}{\sqrt{m_1 m_2} - a} \leq t \leq \frac{k}{\sqrt{m_1 m_2} + a}$.

The moment generating function may be useful for deriving tail bounds, from which one can hope to derive theorems similar to the JL-embedding bounds for $\|v_1 - v_2\|^2$ [13–15]. However, it is more difficult to derive practically useful tail bounds for $v_1^{\mathrm{T}}v_2$ than for $\|v_1 - v_2\|^2$. One intuitive way to see this is via the coefficients of variations:

$$\frac{\sqrt{\mathrm{Var}\left(\|v_1 - v_2\|^2\right)}}{\|u_1 - u_2\|^2} = \sqrt{\frac{2}{k}} \ \text{(constant)}, \qquad \frac{\sqrt{\mathrm{Var}\left(v_1^{\mathrm{T}}v_2\right)}}{u_1^{\mathrm{T}}u_2} \geq \sqrt{\frac{2}{k}} \ \text{(unbounded)}.$$

A straightforward unbiased estimator of the dot product $a = u_1^{\mathrm{T}}u_2$ would be

$$\hat{a}_{MF} = v_1^{\mathrm{T}}v_2, \qquad \mathrm{Var}\left(\hat{a}_{MF}\right) = \frac{1}{k}\left(m_1 m_2 + a^2\right), \quad (4)$$

[4] The JL-embedding bound[13] was originally defined much more generally than for estimating the 2-norm distances, which is the only case we consider.

[5] A recent proof by [12, Lemma 5.4] verified that $\mathrm{Var}\left(v_1^{\mathrm{T}}v_2\right) \leq \frac{2}{k}\left(\|u_1\|^2\|u_2\|^2\right) = \frac{2}{k}m_1 m_2$.

where the subscript "MF" stands for "margin-free."

It is expected that if the marginal norms, $m_1 = \|u_1\|^2$ and $m_2 = \|u_2\|^2$, are given, one can do better. For example,

$$\hat{a}_{SM} = \frac{1}{2}\left(m_1 + m_2 - \|v_1 - v_2\|^2\right), \qquad \text{Var}\left(\hat{a}_{SM}\right) = \frac{1}{2k}\left(m_1 + m_2 - 2a\right)^2, \quad (5)$$

where the subscript "SM" stands for "simple margin (method)." Unfortunately $\hat{a}_{SM}$ is not always better than $\hat{a}_{MF}$. For example, when $a = 0$, $\text{Var}\left(\hat{a}_{SM}\right) = \frac{1}{2k}(m_1+m_2)^2 \geq \text{Var}\left(\hat{a}_{MF}\right) = \frac{1}{k}(m_1 m_2)$. It is easy to show that

$$\text{Var}\left(\hat{a}_{SM}\right) \leq \text{Var}\left(\hat{a}_{MF}\right) \quad \text{ only when } a \geq (m_1 + m_2) - \sqrt{\frac{1}{2}(m_1^2 + m_2^2) + 2m_1 m_2}.$$

We propose an estimator based on maximum likelihood in the following lemma, proved in Appendix B. This estimator has smaller variance than both $\hat{a}_{MF}$ and $\hat{a}_{SM}$.

Lemma 2. *Suppose the margins, $m_1 = \|u_1\|^2$ and $m_2 = \|u_2\|^2$, are known; a maximum likelihood estimator (MLE), denoted as $\hat{a}_{MLE}$, is the solution to a cubic equation:*

$$a^3 - a^2\left(v_1^T v_2\right) + a\left(-m_1 m_2 + m_1\|v_2\|^2 + m_2\|v_1\|^2\right) - m_1 m_2 v_1^T v_2 = 0. \quad (6)$$

The variance of $\hat{a}_{MLE}$ (asymptotic, up to $O(k^{-2})$ terms) is

$$\textit{Var}\left(\hat{a}_{MLE}\right) = \frac{1}{k}\frac{\left(m_1 m_2 - a^2\right)^2}{m_1 m_2 + a^2} \leq \min\left(\textit{Var}\left(\hat{a}_{MF}\right), \textit{Var}\left(\hat{a}_{SM}\right)\right). \quad (7)$$

Figure 1 verifies the inequality in (7) by plotting $\frac{\text{Var}(\hat{a}_{MLE})}{\text{Var}(\hat{a}_{MF})}$ and $\frac{\text{Var}(\hat{a}_{MLE})}{\text{Var}(\hat{a}_{SM})}$. The improvement is quite substantial. For example, $\frac{\text{Var}(\hat{a}_{MLE})}{\text{Var}(\hat{a}_{MF})} = 0.2$ implies that in order to achieve the same mean square accuracy, the proposed MLE estimator needs only 20% of the samples required by the current margin-free (MF) estimator.

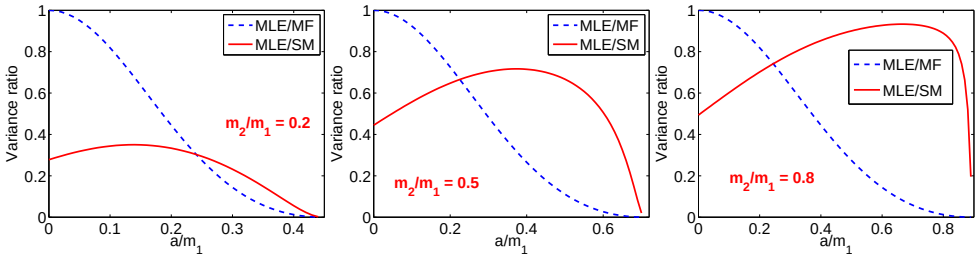
Maximum likelihood estimators can be seriously biased in some cases, but usually the bias is on the order of $O(k^{-1})$, which may be corrected by [16] "Bartlett correction." In Lemma 3 (proved in Appendix C), we are able to show that the asymptotic bias of our $\hat{a}_{MLE}$ is only $O(k^{-2})$ and therefore there is no need for bias correction. Lemma 3 also derives the asymptotic third moment of $\hat{a}_{MLE}$ as well as a more accurate variance formula up to $O(k^{-3})$ terms. The third moment is needed if we would like to model the distribution of $\hat{a}_{MLE}$ more accurately. The more accurate variance formula may be useful for small k or in the region where the $O(k^{-2})$ term in the variance is quite large.

Lemma 3. *The bias, third moment, and the variance with $O(k^{-2})$ correction for the maximum likelihood estimator, $\hat{a}_{MLE}$, derived in Lemma 2, are given by*

$$E\left(\hat{a}_{MLE} - a\right) = O(k^{-2}), \quad (8)$$

$$E\left(\left(\hat{a}_{MLE} - a\right)^3\right) = \frac{-2a(3m_1 m_2 + a^2)(m_1 m_2 - a^2)^3}{k^2(m_1 m_2 + a^2)^3} + O(k^{-3}), \quad (9)$$

$$\textit{Var}\left(\hat{a}_{MLE}\right)_2^c = \frac{1}{k}\frac{\left(m_1 m_2 - a^2\right)^2}{m_1 m_2 + a^2} + \frac{1}{k^2}\frac{4(m_1 m_2 - a^2)^4}{(m_1 m_2 + a^2)^4} m_1 m_2 + O(k^{-3}). \quad (10)$$



(a) (b) (c)

Fig. 1. The variance ratios, $\frac{\text{Var}(\hat{a}_{MLE})}{\text{Var}(\hat{a}_{MF})}$ and $\frac{\text{Var}(\hat{a}_{MLE})}{\text{Var}(\hat{a}_{SM})}$ verify that our proposed MLE has smaller variance than both the margin-free (MF) estimator and the simple margin (SM) method. Var $(\hat{a}_{MLE})$, Var $(\hat{a}_{MF})$, and Var $(\hat{a}_{SM})$ are given in (7), (4), and (5), respectively. We consider $m_2 = 0.2m_1$, $m_2 = 0.5m_1$, and $m_2 = 0.8m_1$, in panels (a), (b), and (c), respectively.

Eq. (10) indicates that when $a = 0$, the $O(k^{-2})$ term of the asymptotic variance is $\frac{4}{k}$ of the $O(k^{-1})$ term. When $k \leq 10$ and a is very small, we might want to consider using (10) instead of (7) for Var$(\hat{a}_{MLE})$. However, as we will show next, for very small k, there is also a multiple root problem in solving the cubic MLE equation (6).

Lemma 4. *The cubic MLE equation (6) in Lemma 2 admits multiple real roots with a small probability, expressed as*

$$\mathbf{Pr}\left(\textit{multiple real roots}\right) = \mathbf{Pr}\left(P^2(11 - Q^2/4 - 4Q + P^2) + (Q-1)^3 \leq 0\right), \quad (11)$$

where $P = \frac{v_1^T v_2}{\sqrt{m_1 m_2}}$, $Q = \frac{\|v_1\|^2}{m_1} + \frac{\|v_2\|^2}{m_2}$. *This probability is (crudely) bounded by*

$$\mathbf{Pr}\left(\textit{multiple real roots}\right) \leq e^{-0.0085k} + e^{-0.0966k}. \quad (12)$$

When $a = m_1 = m_2$, *this probability can be (sharply) bounded by*

$$\mathbf{Pr}\left(\textit{multiple real roots} \mid a = m_1 = m_2\right) \leq e^{-1.5328k} + e^{-0.4672k}. \quad (13)$$

Although the bound (12) is crude, the probability of admitting multiple real roots in (11) can be easily simulated. Figure 2 shows that this probability drops quickly to $< 1\%$ when $k \geq 8$.

To the best of our knowledge, there is no consensus on what is the best solution to multiple roots[17]. Because the probability of multiple roots is so small when $k \geq 8$ while in the large-scale applications we expect $k \gg 10$, we suggest not to worry about multiple roots. Also, we will only use the $O(k^{-1})$ term of Var$(\hat{a}_{MLE})$, i.e, (7).

Figure 3 presents some simulation results, using two words "THIS" and "HAVE," from some MSN Web crawl data. Here $u_{1,j}$ ($u_{2,j}$) is the number of occurrences of word "THIS" (word "HAVE") in the jth page, $j = 1$ to $D = 2^{16}$. As verified in Figure 3, due to the existence of multiple roots at small k, some small bias is observable, as well as some small discrepancies between the observed moments and the theoretical asymptotic moments. When $k \geq 8$, the asymptotic formulas for $\hat{a}_{MLE}$ are very accurate.

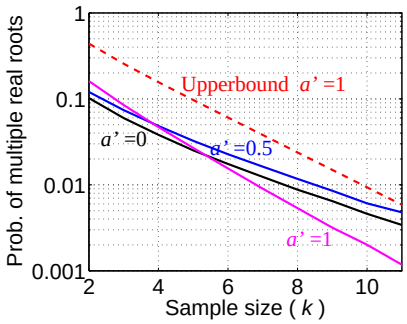


Fig. 2. Simulations show that $\mathbf{Pr}$ (multiple real roots) decreases exponentially fast with respect to increasing sample size k (notice the log scale in the vertical axis). After $k \geq 8$, the probability that the cubic MLE equation (6) admits multiple roots becomes so small ($\leq 1\%$) that it can be safely ignored in practice. Here $a' = \frac{a}{\sqrt{m_1 m_2}}$. The curve for the upper bound is given by (13).

3 Some Tail Bounds

Tails bounds are necessary for deriving JL-type bounds for determining the number of projections (i.e., k) needed in order to achieve a certain specified level of accuracy.

Recall $u_i \in \mathbb{R}^D$ denotes data vectors in the original space and $v_i \in \mathbb{R}^k$ denotes vectors in the projection space. The usual estimator for $d = \|u_1 - u_2\|^2$ is

$$\hat{d}_{MF} = \|v_1 - v_2\|^2 = d, \qquad \frac{\hat{d}_{MF}}{d/k} \sim \chi^2_k, \qquad \text{Var}\left(\hat{d}_{MF}\right) = \frac{2}{k}\|v_1 - v_2\|^4 = \frac{2d^2}{k}.$$

The well-known Chernoff chi-squared tail bound gives (for any $0 < \epsilon < 1$)[6]

$$\mathbf{Pr}\left(\left|\hat{d}_{MF} - d\right| \geq \epsilon d\right) \leq 2\exp\left(-\frac{k}{4}\epsilon^2 + \frac{k}{6}\epsilon^3\right), \tag{15}$$

from which a JL-embedding bound follows, using the Bonferroni union bound [15]:

$$\frac{n^2}{2} 2\exp\left(-\frac{k}{4}\epsilon^2 + \frac{k}{6}\epsilon^3\right) \leq n^{-\gamma} \Rightarrow k \geq k_0 = \frac{4 + 2\gamma}{\epsilon^2/2 - \epsilon^3/3}\log n, \tag{16}$$

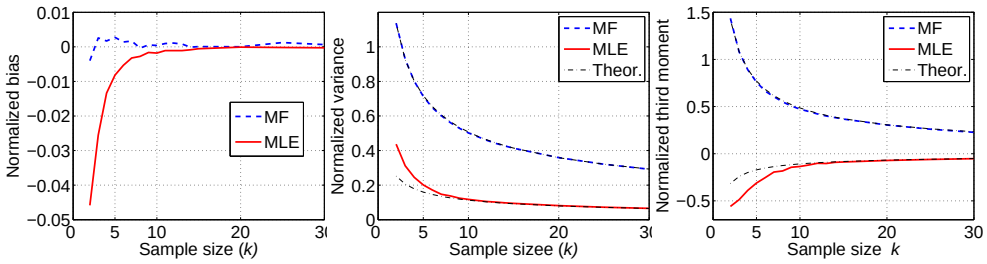
i.e., if $k \geq k_0$, then with probability at least $1 - n^{-\gamma}$, for any two rows u_i, u_j from the data matrix with n rows, we have $(1-\epsilon)\|u_i - u_j\|^2 \leq \|v_i - v_j\|^2 \leq (1+\epsilon)\|u_i - u_j\|^2$.

As mentioned in [15], the above bounds are tight. We will show that, from a practical point of view, using the marginal information can actually improve the bounds.

[6] Since we know the exact distribution in this case, we might as well compute k exactly by iteratively solving a nonlinear equation:

$$\frac{n^2}{2}\left(\mathbf{Pr}\left(\chi^2_k \geq (1+\epsilon)k\right) + \mathbf{Pr}\left(\chi^2_k \leq (1-\epsilon)k\right)\right) = \alpha \qquad \text{(e.g., } \alpha = 0.05\text{)}, \tag{14}$$

which always outputs smaller k values than the JL-bound (e.g., by about 40% when $\epsilon = 0.5$).



(a) Bias (b) Variance (c) Third moment

Fig. 3. Estimations of the dot products between two vectors "THIS" and "HAVE." (a): $\frac{\text{bias}}{a}$, (b): $\frac{\sqrt{\text{Var}(\hat{a})}}{a}$, (c): $\frac{\sqrt[3]{\text{E}(\hat{a}-a)^3}}{a}$. This experiment verifies that (A): Marginal information can improve the estimations considerably. (B): As soon as $k > 8$, $\hat{a}_{MLE}$ is essentially unbiased and the asymptotic variance and third moment match simulations remarkably well. (C): The margin-free estimator ($\hat{a}_{MF}$) is unbiased and the theoretical moments are indistinguishable from simulations.

Using $\hat{a}_{MLE}$, an MLE for $d = \|u_1 - u_2\|^2$, would be

$$\hat{d}_{MLE} = m_1 + m_2 - 2\hat{a}_{MLE}, \qquad \text{Var}\left(\hat{d}_{MLE}\right) = \frac{4}{k}\frac{(m_1 m_2 - a^2)^2}{m_1 m_2 + a^2}. \tag{17}$$

Both $\hat{a}_{MLE}$ and $\hat{d}_{MLE}$ are asymptotically normal. It is well-known that for "small deviations," (e.g., small ϵ) the asymptotic normality of MLE holds with high accuracy. We often care about the "small deviation" behavior because we would like the estimate to be close to the truth. However, when we estimate all pairwise distances simultaneously (as is considered in the JL-embedding bound), the Bonferroni union bound[7] may push the tail to the "large deviation" range hence assuming asymptotic normality could be a concern. On the other hand, the Bonferroni bound leads to larger k values; and larger k improves the accuracy of the asymptotic normality. Based on this (heuristic) argument, the asymptotic tail bounds of $\hat{a}_{MLE}$ may be still useful in practice.

3.1 Normal Approximation

Based on the asymptotic normality $\hat{a}_{MLE} \sim N(a, \text{Var}(\hat{a}_{MLE}))$, we can obtain [8]

$$\mathbf{Pr}\left(|\hat{a}_{MLE} - a| \geq \epsilon a\right) \lesssim 2\exp\left(-\frac{k\epsilon^2}{2}\frac{a^2(m_1 m_2 + a^2)}{(m_1 m_2 - a^2)^2}\right), \tag{19}$$

[7] The Bonferroni bound is well-known for being too conservative, partly because it ignores the correlations. But the major problem is that the criterion is too stringent for large n (here we actually have $\frac{n^2}{2}$ tests). A reasonable alternative is to allow a certain fraction of tests to fail [18, Chapter 9]. For example, if we allow at most $1/p$ tests to fail, we can solve for k from

$$\left(\mathbf{Pr}\left(\chi^2_k \geq (1+\epsilon)k\right) + \mathbf{Pr}\left(\chi^2_k \leq (1-\epsilon)k\right)\right) = \alpha/p \qquad \text{(e.g., } \alpha = 0.05,\ p = 100\text{)} \tag{18}$$

[8] Of course, we can also use the exact normal tail probabilities instead of the upper bounds.

where $\stackrel{\sim}{\leq}$ indicates that bound holds only asymptotically.

Similarly, the asymptotic normality $\hat{d}_{MLE} \sim N(d, \text{Var}(\hat{d}_{MLE}))$ yields

$$\mathbf{Pr}\left(\left|\hat{d}_{MLE} - d\right| \geq \epsilon d\right) \stackrel{\sim}{\leq} 2\exp\left(-\frac{k}{4}\epsilon^2 \frac{d^2}{2}\frac{m_1 m_2 + a^2}{(m_1 m_2 - a^2)^2}\right). \tag{20}$$

Note that $\frac{d^2}{2}\frac{m_1 m_2 + a^2}{(m_1 m_2 - a^2)^2} = \frac{\text{Var}(\hat{a}_{SM})}{\text{Var}(\hat{a}_{MLE})} \geq 1$ (unbounded), with equality holds when $m_1 = m_2 = a$. Therefore, as expected, we can obtain better bounds using marginal information. In practice, we have to choose some reasonable values for m_1, m_2 and a based on prior knowledge of the data, or for the regions we are most interested in.

It would be interesting to see how normal approximation on $\hat{d}_{MF}$ affects its tail bound. Assuming normality, i.e., $\hat{d}_{MF} \sim N\left(d, \frac{2d^2}{k}\right)$, we obtain

$$\mathbf{Pr}\left(\left|\hat{d}_{MF} - d\right| \geq \epsilon d\right) \stackrel{\sim}{\leq} 2\exp\left(-\frac{k}{4}\epsilon^2\right), \tag{21}$$

which agrees with the exact bound on the dominating ϵ^2 term.

When applying normal approximations, it is important to watch out for the third moments, which, to an extent, affect the rate of convergence:

$$\text{E}\left(\hat{d}_{MF} - d\right)^3 = \frac{8d^3}{k^2}, \quad \text{E}\left(\hat{d}_{MLE} - d\right)^3 = 8\frac{-2a(3m_1 m_2 + a^2)(m_1 m_2 - a^2)^3}{k^2(m_1 m_2 + a^2)^3}.$$

Some algebra can verify that

$$\left|\frac{\text{E}\left(\hat{d}_{MLE} - d\right)^3}{\text{E}\left(\hat{d}_{MF} - d\right)^3}\right| \leq \left(\frac{\text{Var}(\hat{a}_{MLE})}{\text{Var}(\hat{a}_{MF})}\right)^{\frac{3}{2}} \leq 1, \tag{22}$$

which means the third moment of $\hat{d}_{MLE}$ (and $\hat{a}_{MLE}$) is well-behaved.

3.2 Generalized Gamma Approximation

The normal approximation matches the first two (asymptotic) moments. The accuracy can be further improved by matching the third moment. For example, [19] used a generalized gamma distribution to accurately approximate the finite-dimensional behavior of the random matrix eigenvalues arising in some wireless communication channels.

For convenience, we consider $a \geq 0$ (true in most applications). Assuming $-\hat{a}_{MLE} \sim G(\alpha, \beta, \xi)$, a generalized gamma distribution with three parameters (α, β, ξ), then

$$\text{E}\left(-\hat{a}_{MLE}\right) = \alpha\beta, \quad \text{Var}\left(-\hat{a}_{MLE}\right) = \alpha\beta^2, \quad \text{E}\left(-\hat{a}_{MLE} + a\right)^3 = (\xi + 1)\alpha\beta^3, \tag{23}$$

from which we can compute (α, β, ξ):

$$\begin{aligned} \alpha &= \frac{ka^2(m_1 m_2 + a^2)}{(m_1 m_2 - a^2)^2} = k\alpha', \quad \beta = \frac{-(m_1 m_2 - a^2)^2}{k(m_1 m_2 + a^2)a} = \frac{-1}{k}\beta', \\ \xi &= \frac{2a^2(3m_1 m_2 + a^2)}{(m_1 m_2 + a^2)(m_1 m_2 - a^2)} - 1 \end{aligned} \tag{24}$$

The generalized gamma distribution does not have a closed-form density, but it does have closed-form moment generating functions [19, (69)(70)]:

$$\mathrm{E}\left(\exp\left(-\hat{a}_{MLE}t\right)\right) = \begin{cases} \exp\left(\frac{\alpha}{\xi-1}\left(1-(1-\beta\xi t)^{\frac{\xi-1}{\xi}}\right)\right) & \text{when } \xi > 1 \\ \exp\left(\frac{\alpha}{1-\xi}\left(\left(\frac{1}{1-\beta\xi t}\right)^{\frac{1-\xi}{\xi}}-1\right)\right) & \text{when } \xi < 1 \\ (1-\beta t)^{-\alpha} & \text{when } \xi = 1 \end{cases}$$

$\xi > 1$ happens when $\frac{a^2}{m_1 m_2} > \frac{\sqrt{17}-3}{4} = 0.2808$. Using the Chernoff inequality and assuming $\xi > 1$ (other cases are similar) , we obtain

$$\mathbf{Pr}\left(\hat{d}_{MLE} \geq (1+\epsilon)d\right) \lesssim \exp\left(-k\left(\left(\frac{2a}{2a-\epsilon d}\right)^{\xi-1}\left(\frac{\alpha'}{\xi-1}-\frac{a}{\beta'\xi}\right)-\frac{\alpha'}{\xi-1}+\frac{2a-\epsilon d}{2\beta'\xi}\right)\right),$$

$$\mathbf{Pr}\left(\hat{d}_{MLE} \leq (1-\epsilon)d\right) \lesssim \exp\left(-k\left(\left(\frac{2a}{2a+\epsilon d}\right)^{\xi-1}\left(\frac{\alpha'}{\xi-1}-\frac{a}{\beta'\xi}\right)-\frac{\alpha'}{\xi-1}+\frac{2a+\epsilon d}{2\beta'\xi}\right)\right).$$

4 Sign Random Projections

We give a brief introduction to "sign random projections," (i.e., only storing the signs of the projected data), and compare sign random projections with regular random projections. For each data point, sign random projections store just one bit per projection. There are efficient algorithms for computing hamming distances [14, 10, 11].

We will show that when the data are roughly uncorrelated, the variance of sign random projections is only about $\frac{\pi^2}{4} \approx 2.47$ of the variance of regular random projections, which store real numbers. With highly correlated data, however, sign random projections can be quite inefficient compared to regular random projections.

Recall $u_i \in \mathbb{R}^D$ denotes data vectors in the original space and $v_i = \frac{1}{\sqrt{k}}\mathbf{R}^{\mathrm{T}}u_i \in \mathbb{R}^k$ for vectors in the projection space. It is easy to show that[10]

$$\mathbf{Pr}\left(\text{sign}(v_{1,j}) = \text{sign}(v_{2,j})\right) = 1 - \frac{\theta}{\pi}, \qquad j = 1, 2, ..., k, \tag{25}$$

where $\theta = \cos^{-1}\left(\frac{u_1^{\mathrm{T}}u_2}{\|u_1\|\|u_2\|}\right) = \cos^{-1}\left(\frac{a}{\sqrt{m_1 m_2}}\right)$ is the angle between u_1 and u_2.

We can estimate θ as a binomial probability, whose variance would be

$$\text{Var}\left(\hat{\theta}\right) = \frac{\pi^2}{k}\left(1-\frac{\theta}{\pi}\right)\left(\frac{\theta}{\pi}\right) = \frac{\theta(\pi-\theta)}{k}. \tag{26}$$

We can also estimate $a = u_1^{\mathrm{T}}u_2$ from $\hat{\theta}$ if knowing the margins:

$$\hat{a}_{Sign} = \cos(\hat{\theta})\sqrt{m_1 m_2}. \tag{27}$$

By the Delta method, $\hat{a}_{Sign}$ is asymptotically unbiased with the asymptotic variance

$$\text{Var}\left(\hat{a}_{Sign}\right) = \text{Var}(\hat{\theta})\sin^2(\theta)m_1 m_2 = \frac{\theta(\pi-\theta)}{k}\sin^2(\theta)m_1 m_2, \tag{28}$$

provided $\sin(\theta)$ is nonzero, which is violated when $\theta = 0$ or π. In fact, when θ is close to 0 or π, due to the high nonlinearity, the asymptotic variance formula is not reliable.

Regular random projections store real numbers (32 or 64 bits). At the same number of projections (i.e., the same k) , obviously sign random projections will have larger variances. If the variance is inflated only by a factor of (e.g.,) 4, sign random projections would be preferable because we could increase k to (e.g.,) $4k$, to achieve the same accuracy while the storage cost will still be lower than regular random projections.

We compare the variance (Var $(\hat{a}_{Sign})$) of sign random projections with the variance of regular random projections considering the margins (i.e., Var $(\hat{a}_{MLE})$) by

$$V_{Sign} = \frac{\text{Var}\,(\hat{a}_{Sign})}{\text{Var}\,(\hat{a}_{MLE})} = \frac{\theta(\pi-\theta)\sin^2(\theta)m_1m_2}{\frac{(m_1m_2-a^2)^2}{m_1m_2+a^2}} = \frac{\theta(\pi-\theta)(1+\cos^2(\theta))}{\sin^2(\theta)}, \quad (29)$$

which is symmetric about $\theta = \frac{\pi}{2}$. It is easy to check (also shown in Figure 4) that V_{Sign} is monotonically decreasing in $(0, \frac{\pi}{2}]$ with minimum $\frac{\pi^2}{4} \approx 2.47$, attained at $\theta = \frac{\pi}{2}$.

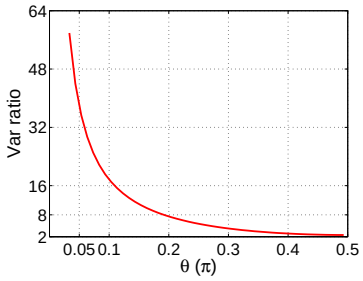


Fig. 4. The ratios of variance $V_{Sign} = \frac{\text{Var}\left(\hat{a}_{Sign}\right)}{\text{Var}(\hat{a}_{MLE})}$ decreases monotonically in $(0, \frac{\pi}{2}]$, with minimum = $\frac{\pi^2}{4} \approx 2.47$ attained at $\theta = \frac{\pi}{2}$. Note that the horizontal axis is in π.

When the data points are nearly uncorrelated (θ close to $\frac{\pi}{2}$, in fact $\theta > \frac{\pi}{5}$ could be good enough), sign random projections should have good performance. However, some applications such as duplicate detections are interested in data points that are close to each other hence sign random projections may cause relatively large errors.

5 Some Recent Progress On Random Projections

There is considerable recent interest in *sparse random projections*, proposed by Achlioptas [15]. It replaces the $N(0,1)$ entries in $\mathbf{R}$ with entries in $\sqrt{s} \times \{-1, 0, 1\}$ with probabilities $\{\frac{1}{2s}, 1-\frac{1}{s}, \frac{1}{2s}\}$, $1 \leq s \leq 3$. With $s = 3$, one can get a threefold speedup.

We[20] recently proposed *very sparse random projections* by using $s = \sqrt{D}$, to obtain a $\sqrt{D}$-fold speedup. The analysis is based on the asymptotic properties of the projected data. For example, assuming bounded third moment on the original data, the projected data converge to normal at the rate of $O\left(\frac{1}{D^{1/4}}\right)$, which is sufficiently fast

since D has to be large otherwise there would be no need of seeking approximate answers. The MLE proposed in this study is still useful in *very sparse random projections*.

The limitation of random projection is that it can not estimate multi-way distances nor can it estimate 1-norm distances. The authors' concurrent work[21] has proposed a new sketch-based sampling algorithm, which is capable of estimating two-way and multi-way distances in any norms. In particular, this algorithm provably outperforms random projections in boolean data and nearly independent data.

6 Conclusion

We propose a maximum likelihood estimator (MLE) for random projections, taking advantage of the marginal information, which can be easily computed at negligible incremental cost. This estimator has provably smaller variance than the current method; and therefore it can reduce the required number of projections.

Acknowledgment

We would like to thank Dimitris Achlioptas, Persi Diaconis, Bradley Efron, Jerome Friedman, Tze Leung Lai, Joseph Romano and Yiyuan She, for many very helpful conversations (or email communications), or pointers to relevant references.

A Proof of Lemma 1

Recall $u_1, u_2 \in \mathbb{R}^D$, $v_1 = \frac{1}{\sqrt{k}}\mathbf{R}^{\mathrm{T}}u_1$, and $v_2 = \frac{1}{\sqrt{k}}\mathbf{R}^{\mathrm{T}}u_2$, where $\mathbf{R} \in \mathbb{R}^{D\times k}$ consists of i.i.d. $N(0,1)$ entries. Note that $v_1^{\mathrm{T}}v_2 = \sum_{j=1}^{k} v_{1,j}v_{2,j} = \sum_{j=1}^{k} \frac{1}{k}u_1^{\mathrm{T}}\mathbf{R}_j\mathbf{R}_j^{\mathrm{T}}u_2$ is a sum of i.i.d. terms, where $\mathbf{R}_j$ is the j^{th} column of $\mathbf{R}$.

It is easy to show that $(v_{1,j}, v_{2,j})$ are jointly normal with zero mean and covariance Σ (denoting $m_1 = \|u_1\|^2$, $m_2 = \|u_2\|^2$, and $a = u_1^{\mathrm{T}}u_2$)

$$\begin{bmatrix} v_{1,j} \\ v_{2,j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \frac{1}{k}\begin{bmatrix} \|u_1\|^2 & u_1^{\mathrm{T}}u_2 \\ u_1^{\mathrm{T}}u_2 & \|u_2\|^2 \end{bmatrix} = \frac{1}{k}\begin{bmatrix} m_1 & a \\ a & m_2 \end{bmatrix}\right). \tag{30}$$

It is easier to work with the conditional probability:

$$v_{1,j}|v_{2,j} \sim N\left(\frac{a}{m_2}v_{2,j}, \frac{m_1m_2 - a^2}{km_2}\right), \tag{31}$$

from which we can get

$$\mathrm{E}\left(v_{1,j}v_{2,j}\right)^2 = \mathrm{E}\left(\mathrm{E}\left(v_{1,j}^2v_{2,j}^2|v_{2,j}\right)\right) = \mathrm{E}\left(v_{2,j}^2\left(\frac{m_1m_2 - a^2}{km_2} + \left(\frac{a}{m_2}v_{2,j}\right)^2\right)\right)$$

$$= \frac{m_2}{k}\frac{m_1m_2 - a^2}{km_2} + \frac{3m_2^2}{k^2}\frac{a^2}{m_2^2} = \frac{1}{k^2}\left(m_1m_2 + 2a^2\right). \tag{32}$$

Therefore,

$$\text{Var}\left(v_{1,j}v_{2,j}\right) = \frac{1}{k^2}\left(m_1m_2 + a^2\right), \qquad \text{Var}\left(v_1^{\mathrm{T}}v_2\right) = \frac{1}{k}\left(m_1m_2 + a^2\right). \tag{33}$$

The third moment can be proved similarly. In fact, one can compute any moments, using the moment generating function:

$$\begin{aligned}
&\mathrm{E}\left(\exp(v_{1,j}v_{2,j}t)\right) = \mathrm{E}\left(\mathrm{E}\left(\exp(v_{1,j}v_{2,j}t)\right)|v_{2,j}\right)\\
=&\mathrm{E}\left(\exp\left(\left(\frac{a}{m_2}v_{2,j}\right)v_{2,j}t + \left(\frac{m_1m_2 - a^2}{km_2}\right)\left(v_{2,j}t\right)^2/2\right)\right)\\
=&\mathrm{E}\left(\exp\left(v_{2,j}^2\frac{k}{m_2}\left(\frac{a}{k}t + \frac{1}{k^2}\left(m_1m_2 - a^2\right)\frac{t^2}{2}\right)\right)\right)\\
=&\left(1 - \frac{2a}{k}t - \frac{1}{k^2}\left(m_1m_2 - a^2\right)t^2\right)^{-\frac{1}{2}}.
\end{aligned} \tag{34}$$

Here, we use the fact that $\frac{v_{2,j}^2}{m_2/k} \sim \chi_1^2$, a chi-squared random variable with one degree of freedom. Note that $\mathrm{E}\left(\exp(Yt)\right) = \exp\left(\mu t + \sigma^2t^2/2\right)$ if $Y \sim N(\mu, \sigma^2)$; and $\mathrm{E}\left(\exp(Yt)\right) = (1-2t)^{-\frac{1}{2}}$ if $Y \sim \chi_1^2$. By independence, we have proved that

$$\mathrm{E}\left(\exp(v_1^{\mathrm{T}}v_2t)\right) = \left(1 - \frac{2}{k}at - \frac{1}{k^2}\left(m_1m_2 - a^2\right)t^2\right)^{-\frac{k}{2}}, \tag{35}$$

where $\frac{-k}{\sqrt{m_1m_2}-a} \le t \le \frac{k}{\sqrt{m_1m_2}+a}$. This completes the proof of Lemma 1.

B Proof of Lemma 2

From Appendix A, we can write down the joint likelihood function for $\{v_{1,j}, v_{2,j}\}_{j=1}^k$:

$$\text{lik}\left(\{v_{1,j}, v_{2,j}\}_{j=1}^k\right) \propto |\Sigma|^{-\frac{k}{2}}\exp\left(-\frac{1}{2}\sum_{j=1}^k\begin{bmatrix}v_{1,j} & v_{2,j}\end{bmatrix}\Sigma^{-1}\begin{bmatrix}v_{1,j}\\ v_{2,j}\end{bmatrix}\right). \tag{36}$$

where (assuming $m_1m_2 \neq a$ to avoid triviality)

$$|\Sigma| = \frac{1}{k^2}(m_1m_2 - a^2), \qquad \Sigma^{-1} = \frac{k}{m_1m_2 - a^2}\begin{bmatrix}m_2 & -a\\ -a & m_1\end{bmatrix},$$

which allows us to express the $\log$ likelihood function, $l(a)$, to be

$$l(a) = -\frac{k}{2}\log\left(m_1m_2 - a^2\right) - \frac{k}{2}\frac{1}{m_1m_2 - a^2}\sum_{j=1}^k\left(v_{1,j}^2m_2 - 2v_{1,j}v_{2,j}a + v_{2,j}^2m_1\right).$$

Setting $l'(a)$ to zero, we obtain $\hat{a}_{MLE}$, which is the solution to the cubic equation:

$$a^3 - a^2\left(v_1^{\mathrm{T}}v_2\right) + a\left(-m_1m_2 + m_1\|v_2\|^2 + m_2\|v_1\|^2\right) - m_1m_2v_1^{\mathrm{T}}v_2 = 0. \tag{37}$$

The well-known large sample theory says that $\hat{a}_{MLE}$ is asymptotically unbiased and converges weakly to a normal random variable $N\left(a, \text{Var}\left(\hat{a}_{MLE}\right) = \frac{1}{\text{I}(a)}\right)$, where $\text{I}(a)$, the expected Fisher Information, is $\text{I}(a) = -\text{E}\left(l''(a)\right)$. Recall $l(a)$ is the $\log$ likelihood function obtained in Appendix B. Some algebra will show that

$$\text{I}(a) = k\frac{m_1m_2 + a^2}{\left(m_1m_2 - a^2\right)^2}. \qquad \text{Var}\left(\hat{a}_{MLE}\right) = \frac{1}{k}\frac{\left(m_1m_2 - a^2\right)^2}{m_1m_2 + a^2}. \tag{38}$$

Applying the Cauchy-Schwarz inequality a couple of times can prove

$$\text{Var}\left(\hat{a}_{MLE}\right) = \frac{1}{k}\frac{\left(m_1m_2 - a^2\right)^2}{m_1m_2 + a^2} \leq \min\left(\text{Var}\left(\hat{a}_{MF}\right), \text{Var}\left(\hat{a}_{SM}\right)\right), \tag{39}$$

where $\text{Var}\left(\hat{a}_{MF}\right) = \frac{1}{k}\left(m_1m_2 + a^2\right)$, $\text{Var}\left(\hat{a}_{SM}\right) = \frac{1}{2k}\left(m_1 + m_2 - 2a\right)^2$.

C Proof of Lemma 3

We analyze the higher-order properties of $\hat{a}_{MLE}$ using stochastic Taylor expansions. We use some formulations appeared in [16, 22, 23]. The bias

$$\text{E}\left(\hat{a}_{MLE} - a\right) = -\frac{\text{E}\left(l'''(a)\right) + 2\text{I}'(a)}{2\text{I(a)}} + O(k^{-2}), \tag{40}$$

which is often called the "Bartlett correction." Some algebra can show

$$\text{I}'(a) = \frac{2ka(3m_1m_2 + a^2)}{(m_1m_2 - a^2)^3},\ \text{E}\left(l'''(a)\right) = -2\text{I}'(a),\ \text{E}\left(\hat{a}_{MLE} - a\right) = O(k^{-2}). \tag{41}$$

The third central moment

$$\begin{aligned}\text{E}\left(\hat{a}_{MLE} - a\right)^3 &= \frac{-3\text{I}'(a) - \text{E}\left(l'''(a)\right)}{\text{I}^3(a)} + O(k^{-3}) \\ &= -\frac{2a(3m_1m_2 + a^2)(m_1m_2 - a^2)^3}{k^2(m_1m_2 + a^2)^3} + O(k^{-3}).\end{aligned} \tag{42}$$

The $O(k^{-2})$ term of the variance, denoted by V_2^c, can be written as

$$\begin{aligned}V_2^c &= \frac{1}{\text{I}^3(a)}\left(\text{E}\left(l''(a)\right)^2 - \text{I}^2(a) - \frac{\partial\left(\text{E}\left(l'''(a)\right) + 2\text{I}'(a)\right)}{\partial a}\right) \\ &\quad + \frac{1}{2\text{I}^4(a)}\left(10\left(\text{I}'(a)\right)^2 - \text{E}\left(l'''(a)\right)\left(\text{E}\left(l'''(a)\right) - 4\text{I}'(a)\right)\right) \\ &= \frac{\text{E}\left(\left(l''(a)\right)^2\right) - \text{I}^2(a)}{\text{I}^3(a)} - \frac{\left(\text{I}'(a)\right)^2}{\text{I}^4(a)}, \quad (\text{as E}\left(l'''(a)\right) + 2\text{I}'(a) = 0).\end{aligned} \tag{43}$$

Computing $\mathrm{E}\left((l''(a))^2\right)$ requires some work. We can write

$$l''(a) = -\frac{k}{S^3}\left(T(4a^2+S) - S(m_1m_2+a^2) - 4aS(v_1^{\mathrm{T}}v_2)\right), \tag{44}$$

where, for simplicity, we let $S = m_1m_2 - a^2$ and $T = \|v_1\|^2 m_2 + \|v_2\|^2 m_1 - 2v_1^{\mathrm{T}}v_2 a$.

Expanding $(l''(a))^2$ generates terms involving T, T^2, $Tv_1^{\mathrm{T}}v_2$. Rewrite

$$\begin{aligned} T &= \frac{m_1m_2-a^2}{k}\left(\sum_{j=1}^{k}\frac{km_2}{m_1m_2-a^2}\left(v_{1,j}-\frac{a}{m_2}v_{2,j}\right)^2 + \sum_{j=1}^{k}v_{2,j}^2\frac{k}{m_2}\right) \\ &= \frac{m_1m_2-a^2}{k}(\eta+\zeta) \end{aligned} \tag{45}$$

Recall $v_{1,j}|v_{2,j} \sim N\left(\frac{a}{m_2}v_{2,j}, \frac{m_1m_2-a^2}{km_2}\right)$, and $v_{2,j} \sim N\left(0, \frac{m_2}{k}\right)$. Then

$$\eta\left|\{v_{1,j}\}_{j=1}^k \sim \chi_k^2, \text{ (independent of } \{v_{1,j}\}_{j=1}^k\right), \qquad \zeta = \sum_{j=1}^{k}v_{2,j}^2\frac{k}{m_2} \sim \chi_k^2, \tag{46}$$

implying that η and ζ are independent; and $\eta+\zeta \sim \chi_{2k}^2$. Thus,

$$\mathrm{E}(T) = 2(m_1m_2-a^2) = 2S, \qquad \mathrm{E}(T^2) = 4S^2(1+\frac{1}{k}). \tag{47}$$

We also need to compute $\mathrm{E}\left(Tv_1^{\mathrm{T}}v_2\right)$. Rewrite

$$Tv_1^{\mathrm{T}}v_2 = (v_1^{\mathrm{T}}v_2)\|v_1\|^2m_2 + (v_1^{\mathrm{T}}v_2)\|v_2\|^2m_1 - 2\left(v_1^{\mathrm{T}}v_2\right)^2 a. \tag{48}$$

Expand $(v_1^{\mathrm{T}}v_2)\|v_1\|^2$

$$(v_1^{\mathrm{T}}v_2)\|v_1\|^2 = \sum_{j=1}^{k}v_{1,j}v_{2,j}\sum_{j=1}^{k}v_{1,j}^2 = \sum_{j=1}^{k}v_{1,j}^3v_{2,j} + \sum_{i=1}^{k}\left(v_{1,i}^2\sum_{j\neq i}v_{1,j}v_{2,j}\right). \tag{49}$$

Again, applying the conditional probability argument, we obtain $\mathrm{E}\left(v_{1,j}^3v_{2,j}\right) = \frac{3am_1}{k^2}$, from which it follows that

$$\begin{aligned} \mathrm{E}\left((v_1^{\mathrm{T}}v_2)\|v_1\|^2\right) &= \sum_{j=1}^{k}\mathrm{E}\left(v_{1,j}^3v_{2,j}\right) + \sum_{i=1}^{k}\left(\mathrm{E}\left(v_{1,i}^2\right)\sum_{j\neq i}\mathrm{E}\left(v_{1,j}v_{2,j}\right)\right) \\ &= \frac{3am_1}{k} + k\frac{m_1}{k}\sum_{j\neq i}\frac{a}{k} = am_1\left(1+\frac{2}{k}\right). \end{aligned} \tag{50}$$

To this end, we have all the necessary components for computing $\mathrm{E}\left((l''(a))^2\right)$. After some algebra, we obtain

$$\mathrm{E}\left((l''(a))^2\right) = \frac{k^2}{S^4}\left(\left(m_1m_2+a^2\right)^2 + \frac{4}{k}\left(m_1^2m_2^2 + a^4 + 6a^2m_1m_2\right)\right), \tag{51}$$

$$V_2^c = \frac{4}{k^2}\frac{\left(m_1m_2-a^2\right)^4}{\left(m_1m_2+a^2\right)^4}m_1m_2. \tag{52}$$

We complete the proof of Lemma 3.

D Proof of Lemma 4

The cubic MLE equation derived in Lemma 2 may admit multiple roots. (Recall a cubic equation always has at least one real root.) By the well-known Cardano condition,

$$\mathbf{Pr}\left(\text{multiple real roots}\right) = \mathbf{Pr}\left(P^2(11 - Q^2/4 - 4Q + P^2) + (Q-1)^3 \le 0\right), \quad (53)$$

where $P = \frac{v_1^{\mathrm{T}} v_2}{\sqrt{m_1 m_2}}$, $Q = \frac{\|v_1\|^2}{m_1} + \frac{\|v_2\|^2}{m_2}$. We can obtain a crude upper bound using the fact that $\mathbf{Pr}(A + B \le 0) \le \mathbf{Pr}(A \le 0) + \mathbf{Pr}(B \le 0)$, i.e.,

$$\mathbf{Pr}\left(\text{multiple real roots}\right) \le \mathbf{Pr}\left(11 - Q^2/4 - 4Q \le 0\right) + \mathbf{Pr}\left(Q - 1 \le 0\right). \quad (54)$$

We will soon prove the following moment generating function

$$\mathrm{E}\left(\exp(Qt)\right) = \left(1 - \frac{4t}{k} + \frac{4t^2}{k^2}\left(\frac{m_1 m_2 - a^2}{m_1 m_2}\right)\right)^{-\frac{k}{2}}, \quad (55)$$

which enables us to prove the following upper bounds:

$$\mathbf{Pr}\left(Q - 1 \le 0\right) \le e^{-0.0966k}, \qquad \mathbf{Pr}\left(11 - Q^2/4 - 4Q \le 0\right) \le e^{-0.0085k}, \quad (56)$$

$$\mathbf{Pr}\left(\text{multiple real roots}\right) \le e^{-0.0966k} + e^{-0.0085k}, \quad (57)$$

using the standard Chernoff inequality, e.g., $\mathbf{Pr}\left(Q > z\right) = \mathbf{Pr}\left(e^{Qt} > e^{zt}\right) \le \mathrm{E}\left(e^{Qt}\right) e^{-zt}$, choosing t that minimizes the upper bound.

The upper bound (57) is very crude but nevertheless reveals that the probability of admitting multiple real roots decreases exponentially fast.

It turns out there is a simple exact solution for the special case of $a = m_1 = m_2$, i.e., $Q = 2P = \|v_1\|^2/m_1$, $kP = \frac{k\|v_1\|^2}{m_2} \sim \chi_k^2$, and a (sharp) upper bound:

$$\mathbf{Pr}\left(\text{multiple real roots}\right) = \mathbf{Pr}\left(\left(P-3\right)^2 \ge 8\right) \le e^{-1.5328k} + e^{-0.4672k}. \quad (58)$$

To complete the proof of Lemma 4, we need to outline the proof for the moment generating function $\mathrm{E}\left(\exp(Qt)\right)$. Using the conditional probability $v_{1,j}|v_{2,j}$, we know

$$\frac{km_2}{m_1 m_2 - a^2} v_{1,j}^2 | v_{2,j} \sim \chi_{1,\lambda}^2, \qquad \text{where } \lambda = \frac{ka^2}{m_2(m_1 m_2 - a^2)} v_{2,j}^2. \quad (59)$$

$\chi_{1,\lambda}^2$ denotes a non-central chi-squared random variable with one degree of freedom and non-centrality λ. If $Y \sim \chi_{1,\lambda}^2$, then $\mathrm{E}\left(\exp(Yt)\right) = \exp\left(\frac{\lambda t}{1-2t}\right)(1-2t)^{-\frac{1}{2}}$. Because

$$\mathrm{E}\left(\exp(Qt)\right) = \prod_{j=1}^{k} \mathrm{E}\left(\mathrm{E}\left(\exp\left(\frac{v_{1,j}^2}{m_1} + \frac{v_{2,j}^2}{m_2}\right) t \,\middle|\, v_{2,j}\right)\right), \quad (60)$$

we can obtain the moment generating function in (55) after some algebra.

References

1. Vempala, S.S.: The Random Projection Method. American Mathematical Society, Providence, RI (2004)
2. Arriaga, R., Vempala, S.: An algorithmic theory of learning: Robust concepts and random projection. In: Proc. of FOCS (Also to appear in Machine Learning), New York (1999) 616–623
3. Dasgupta, S.: Learning mixtures of gaussians. In: Proc. of FOCS, New York (1999) 634–644
4. Fradkin, D., Madigan, D.: Experiments with random projections for machine learning. In: Proc. of KDD, Washington, DC (2003) 517–522
5. Fern, X.Z., Brodley, C.E.: Random projection for high dimensional data clustering: A cluster ensemble approach. In: Proc. of ICML, Washington, DC (2003) 186–193
6. Balcan, M.F., Blum, A., Vempala, S.: On kernels, margins, and low-dimensional mappings. In: Proc. of ALT, Padova, Italy (2004) 194 – 205
7. Papadimitriou, C.H., Raghavan, P., Tamaki, H., Vempala, S.: Latent semantic indexing: A probabilistic analysis. In: Proc. of PODS, Seattle,WA (1998) 159–168
8. Achlioptas, D., McSherry, F., Schölkopf, B.: Sampling techniques for kernel methods. In: Proc. of NIPS, Vancouver, BC, Canada (2001) 335–342
9. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: Applications to image and text data. In: Proc. of KDD, San Francisco, CA (2001) 245–250
10. Charikar, M.S.: Similarity estimation techniques from rounding algorithms. In: Proc. of STOC, Montreal, Quebec, Canada (2002) 380–388
11. Ravichandran, D., Pantel, P., Hovy, E.: Randomized algorithms and NLP: Using locality sensitive hash function for high speed noun clustering. In: Proc. of ACL, Ann Arbor, MI (2005) 622–629
12. Liu, K., Kargupta, H., Ryan, J.: Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transactions on Knowledge and Data Engineering **18** (2006) 92–106
13. Johnson, W.B., Lindenstrauss, J.: Extensions of Lipschitz mapping into Hilbert space. Contemporary Mathematics **26** (1984) 189–206
14. Indyk, P., Motwani, R.: Approximate nearest neighbors: Towards removing the curse of dimensionality. In: Proc. of STOC, Dallas, TX (1998) 604–613
15. Achlioptas, D.: Database-friendly random projections: Johnson-Lindenstrauss with binary coins. Journal of Computer and System Sciences **66** (2003) 671–687
16. Bartlett, M.S.: Approximate confidence intervals, II. Biometrika **40** (1953) 306–317
17. Small, C.G., Wang, J., Yang, Z.: Eliminating multiple root problems in estimation. Statistical Science **15** (2000) 313–341
18. Lehmann, E.L., Romano, J.P.: Testing Statistical Hypothesis. Third edn. Springer, New York, NY (2005)
19. Li, P., Paul, D., Narasimhan, R., Cioffi, J.: On the distribution of SINR for the MMSE MIMO receiver and performance analysis. IEEE Trans. Inform. Theory **52** (2006) 271–286
20. Li, P., Hastie, T.J., Church, K.W.: Margin-constrained random projections and very sparse random projections. Technical report, Department of Statistics, Stanford University (2006)
21. Li, P., Church, K.W., Hastie, T.J.: A sketched-based sampling algorithm on sparse data. Technical report, Department of Statistics, Stanford University (2006)
22. Shenton, L.R., Bowman, K.: Higher moments of a maximum-likelihood estimate. Journal of Royal Statistical Society B **25** (1963) 305–317
23. Ferrari, S.L.P., Botter, D.A., Cordeiro, G.M., Cribari-Neto, F.: Second and third order bias reduction for one-parameter family models. Stat. and Prob. Letters **30** (1996) 339–345