STABLE RANDOM PROJECTIONS AND CONDITIONAL RANDOM SAMPLING, TWO SAMPLING TECHNIQUES FOR MODERN MASSIVE DATASETS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Ping Li

August 2007

© Copyright by Ping Li 2007
All Rights Reserved

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Trevor J. Hastie Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Art Owen

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Robert Tibshriani

Approved for the University Committee on Graduate Studies.

Abstract

The ubiquitous phenomenon of massive datasets, such as massive Web data and massive data streams, in modern applications has brought in both challenges and opportunities to scientists and engineers. For example, simple operations like computing $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ become difficult when the data matrix $\mathbf{A} \in \mathbf{R}^{n \times D}$ grows to the Web scale. Various sampling techniques have been developed to represent (store) massive data using small (memory) space and to recover the original summary statistics (e.g., pair-wise distances) from compact representations of the data. Two sampling techniques are elaborated in this thesis. The method of *stable random projections* in general works very well in heavy-tailed data and the method of *Conditional Random Sampling (CRS)* is highly effective in sparse data.

The method of *stable random projections* multiplies the original data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ by a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ ($k \ll D$), resulting in $\mathbf{B} = \mathbf{A}\mathbf{R} \in \mathbb{R}^{n \times k}$. The entries of $\mathbf{R}$ are typically sampled i.i.d. from a symmetric α-stable distribution ($0 < \alpha \leq 2$); and we can estimate the original l_α properties of $\mathbf{A}$ from $\mathbf{B}$. In the l_2 case, the advantage of *stable random projections* is highlighted by the celebrated Johnson-Lindenstrauss (JL) Lemma, which says it suffices to let $k = O\left(\frac{\log n}{\epsilon^2}\right)$ so that any pair-wise l_2 distance in $\mathbf{A}$ can be estimated within a $1 \pm \epsilon$ factor of the truth.

In this thesis, we will prove an analog of the JL Lemma for general $0 < \alpha \leq 2$. The method of *stable random projections* boils down to a statistical estimation problem, for estimating the scale parameter of a symmetric α-stable distribution. This problem is interesting because we seek estimators that are both statistically accurate and computationally efficient. We study and compare various estimators including the arithmetic mean, the geometric mean, the harmonic mean, the fractional power, and

the maximum likelihood estimators.

This thesis also addresses several special cases of *stable random projections*. For the l_2 norm case (i.e., *normal random projections*), we propose improving the estimates by taking advantage of the marginal information. Also for the l_2 case, one can sample the projection matrix $\mathbf{R}$ from a much simpler *sub-Gaussian* distribution instead of normal. Under reasonable regularity conditions, a special *sub-Gaussian* distribution in $[-1,\ 0,\ 1]$ with probabilities $\left\{\frac{1}{s}, 1-\frac{1}{2s}, \frac{1}{s}\right\}$ and very large s (i.e., *very sparse random projections*) can work just as well as *normal random projections*. For the l_1 case, i.e., *Cauchy random projections*, the estimation task is also particularly interesting. For example, the maximum likelihood estimator (MLE) is computationally feasible in the l_1 case and we propose using an *inverse Gaussian* distribution to accurately model the distribution of the MLE.

The method of *stable random projections* does not take advantage of data sparsity, while large-scale datasets are often highly sparse. *Conditional Random Sampling (CRS)* utilizes data sparsity by sampling only from non-zero elements of the data. In the estimation stage, CRS re-constructs equivalent (conditional) *random samples* on the fly, on a pair-wise or group-wise basis, with the (equivalent) sample size retrospectively determined. Theoretical analysis will be conditional on the sample size.

In many practical scenarios, CRS often outperforms *stable random projections*, theoretically or empirically. In addition, CRS is a general-purpose technique while the method of *stable random projections* is associated with a particular l_α norm. These two methods can be complementary to each other, for solving large-scale scientific & engineering problems, in search engines & information retrieval, databases, modern data streaming systems, numerical linear algebra, and many machine learning and data mining tasks involving massive computations of distances.

Acknowledgments

I became fascinated in Statistics while I was working towards master degrees in both Computer Science and Electrical Engineering at Stanford. Before I was admitted to the Ph.D. program in Statistics, I had taken courses with Professor Trevor Hastie (later, my thesis adviser), Professor Jerome Friedman, and Professor Bradley Efron. Their profound knowledge and passion in Statistics, Science, and Engineering had influenced me deeply since then.

During my most enjoyable four years as a Ph.D. student in Statistics, Trevor has provided tremendous support for my research. I have also benefited greatly from his open-mindedness and his passionate encouragement on my inter-disciplinary research.

Sampling has become a rapid-growing research area in many engineering fields because of the ever-increasing vast amount of data nowadays available, for example, the Internet data. I began my interest in sampling techniques when I was a summer intern at Microsoft Research. Dr. Ken Church, my intern mentor, originally planned that I work on a nice project in Chinese/English Machine Translation; but he was completely supportive after I found out that I would be more interested in developing a sampling algorithm for sparse data, which eventually evolved into a general algorithm called *Conditional Random Sampling (CRS)*. I thank Ken dearly for being my intern mentor for two summers and for his persistent helps and suggestions after the summers.

I would like to thank Professor Art Owen and Professor Robert Tibshirani for serving my thesis reading committee. I thank Professor Persi Diaconis and Professor Tim Roughgarden for serving my oral exam committee. Also, I thank Susan Holmes for listening to my oral exam talk.

I certainly owe the gratitude to many residents (faculty, staff, and students) in Sequoia Hall, where I have stayed for my most memorable four years.

Finally, I would like to thank my co-authors in three papers[128, 129, 118], in Medical Imaging & Computer-Aided Diagnostics, Random Matrix & Information Theory & Wireless Communications, as well as in Machine Learning for relevance ranking of Web pages in Internet search. These papers were published during my pursuing the Ph.D. in Statistics, but I could not put the materials into this dissertation, which is devoted entirely to sampling.

Contents

Abstract **iv**

Acknowledgments **vi**

1 Introduction **1**

1.1 Massive Data 1

1.1.1 Massive Web Data 2

1.1.2 Massive Data Streams 4

1.2 Sampling Massive Data and the Challenges 4

1.2.1 Advantages of Random Coordinate Sampling 6

1.2.2 Disadvantages of Random Coordinate Sampling 6

1.3 Stable Random Projections 7

1.4 Conditional Random Sampling (CRS) 8

1.5 Applications 9

1.5.1 Association Rule Mining 9

1.5.2 All Pair-Wise Associations (Distances) 10

1.5.3 Estimating Distances Online 10

1.5.4 Database Query Optimization 10

1.5.5 (Sub-linear) Nearest Neighbor Searching 11

1.6 Organization 13

1.7 Notation 13

2 Introduction to Stable Random Projections **14**

2.1 The Fundamental Problem in Stable Random Projections 15

2.1.1 Stable Distributions 15
2.1.2 The Statistical Estimation Problem 16
2.2 Applications 18
2.2.1 Data Stream Computations 18
2.2.2 Comparing Data Streams Using Hamming Norms 19
2.3 The Choice of the Norm (α) 19

3 Normal Random Projections **21**
3.1 Basic Properties 22
3.2 Improving Normal Random Projections Using Marginal Information . 24
3.2.1 A Numerical Example 25
3.2.2 Higher-Order Analysis of the MLE 26
3.2.3 The Problem of Multiple (Real) Roots 27
3.3 Sign Random Projections 29
3.4 Summary 31
3.5 Proofs 31
3.5.1 Proof of Lemma 1 31
3.5.2 Proof of Lemma 4 32
3.5.3 Proof of Lemma 5 33
3.5.4 Proof of Lemma 6 36

4 Sub-Gaussian and Very Sparse Random Projections **38**
4.1 Sub-Gaussian Random Projections 39
4.1.1 Sub-Gaussian Distributions 39
4.1.2 Analyzing the Sparse Projection Distribution 41
4.1.3 Tail Bounds for Sub-Gaussian Random Projections 42
4.2 Very Sparse Random Projections 44
4.2.1 Variances and Asymptotic Distribution 45
4.2.2 Tail Bounds 47
4.2.3 Very Sparse Random Projections for Classifying Microarray Data 49
4.3 Summary 50

4.4 Proofs . . . 51
4.4.1 Proof of Lemma 8 . . . 51
4.4.2 Proof of Lemma 9 . . . 52

5 Cauchy Random Projections for l_1 54
5.1 Summary of Main Results . . . 55
5.2 The Sample Median Estimators . . . 56
5.3 The Geometric Mean Estimators . . . 60
5.4 The Sample Median Estimators vs. the Geometric Mean Estimators . 64
5.4.1 Mean Square Errors . . . 64
5.4.2 Tail Bounds . . . 65
5.4.3 Simulated Tail Probabilities . . . 66
5.5 The Maximum Likelihood Estimators . . . 67
5.5.1 A Numerical Example . . . 69
5.5.2 Approximate Distributions . . . 69
5.6 Summary . . . 75
5.7 Proofs . . . 76
5.7.1 Proof of Lemma 11 . . . 76
5.7.2 Proof of Lemma 14 . . . 78
5.7.3 Proof of Lemma 15 . . . 79
5.7.4 Proof of Lemma 16 . . . 80
5.7.5 Proof of Lemma 19 . . . 83

6 Stable Random Projections for l_α 86
6.1 Main Results . . . 86
6.2 The Geometric Mean Estimators . . . 88
6.2.1 The Moments and Tail bounds of $\hat{d}_{(\alpha),gm}$ . . . 90
6.2.2 The Biased Geometric Mean Estimator . . . 93
6.3 Estimators for $\alpha \to 0+$ . . . 94
6.3.1 The (Unbiased) Geometric Mean Estimator . . . 96
6.3.2 The Sample Median Estimator . . . 96
6.3.3 The Harmonic Mean Estimator . . . 98

6.3.4 Comparisons with Normal Random Projections in Boolean (0/1) Data . . . 100
6.4 The Harmonic Mean Estimator for Small α . . . 100
6.5 The Fractional Power Estimator . . . 101
6.5.1 Special cases . . . 104
6.5.2 Theoretical Properties . . . 105
6.5.3 Comparing Variances at Finite Samples . . . 105
6.6 Asymptotic (Cramér-Rao) Efficiencies . . . 106
6.7 Very Sparse Stable Random Projections . . . 107
6.7.1 Our Solution: Very Sparse Stable Random Projections . . . 108
6.7.2 Classifying Cancers in Microarray Data . . . 110
6.8 Summary . . . 110
6.9 Proofs . . . 111
6.9.1 Proof of Lemma 21 . . . 111
6.9.2 Proof of Lemma 22 . . . 116
6.9.3 Proof of Lemma 24 . . . 119
6.9.4 Proof of Lemma 25 . . . 122
6.9.5 Proof of Lemma 26 . . . 125
6.9.6 Proof of Lemma 27 . . . 126
6.9.7 Proof of Lemma 28 . . . 131

7 Conditional Random Sampling 133
7.1 The Procedures of CRS . . . 134
7.1.1 The Sampling/Sketching Procedure . . . 134
7.1.2 The Estimation Procedure . . . 137
7.2 Theoretical Properties of D_s . . . 137
7.2.1 Sample-Without-Replacement . . . 141
7.3 Theoretical Variance Analysis of CRS . . . 141
7.4 Improving CRS Using Marginal Information . . . 142
7.4.1 Integer-valued Data (Histograms) . . . 143
7.4.2 Real-valued Data . . . 147

7.5 Theoretical Comparisons of CRS With Random Projections 148
7.5.1 Boolean (0/1) data 148
7.5.2 Nearly Independent Data 149
7.5.3 Comparing the Computational Efficiency 150
7.6 Empirical Evaluations 150
7.7 Summary 153
7.8 Covariance Matrix in Estimating Contingency Tables with Marginal Constraints 157
7.8.1 An Example with $I = 2, J = 2$ 159
7.8.2 An Example with $I = 1, J = 1$ (0/1 Data) 160
7.8.3 The Variance of $\hat{a}_{MLE,c}$ 160

8 Conditional Random Sampling in Boolean Data 161
8.1 An Example of CRS in Boolean Data 161
8.2 Estimators for Two-Way Associations 162
8.2.1 The Independence Baseline 164
8.2.2 The Margin-free Baseline 164
8.2.3 The Exact MLE with Margin Constraints 165
8.2.4 The "Sample-with-replacement" Simplification 166
8.2.5 A Convenient Practical Quadratic Approximation 167
8.2.6 The Variance and Bias 168
8.3 Evaluation of Two-Way Associations 169
8.3.1 Results from Small Dataset Experiment 171
8.3.2 Results from Large Dataset Experiment 172
8.3.3 Rank Retrieval by Cosine 172
8.4 Multi-way Associations 174
8.4.1 Multi-way Sketches 175
8.4.2 Baseline Independence Estimator 177
8.4.3 Baseline Margin-free Estimator 178
8.4.4 The MLE 178
8.4.5 The Covariance Matrix 179

8.4.6 Empirical Evaluation 182
8.5 Comparison with Broder's Sketches 183
8.5.1 Broder's Minwise Sketch 185
8.5.2 Broder's Original Sketch 185
8.5.3 Why Our Algorithm Improves Broders's Sketch 187
8.5.4 Comparison of Variances 187
8.5.5 Empirical Evaluations 190
8.6 Summary 190

9 Conclusions **193**

Bibliography **196**

List of Tables

1.1 Page hits for two high frequency words and a few low frequency words. 2

1.2 Joint frequencies ought to decrease monotonically as we add terms to the query, but estimates produced by search engines sometimes violate this invariant. 3

1.3 Page hits returned by Google for four words and their two-way, three-way, and four-way combinations. 11

6.1 The optimal λ^*, the variance factor $\frac{1}{\lambda^{*2}}\left(\frac{\frac{2}{\pi}\Gamma(1-2\lambda^*)\Gamma(2\lambda^*\alpha)\sin(\pi\lambda^*\alpha)}{\left[\frac{2}{\pi}\Gamma(1-\lambda^*)\Gamma(\lambda^*\alpha)\sin\left(\frac{\pi}{2}\lambda^*\alpha\right)\right]^2}-1\right)$, and the maximum order (γ_m) of bounded moments are tabulated. For example, when $\alpha = 1.90$, $\mathrm{E}\left(\hat{d}_{(1.90),fp}\right)^{\gamma} < \infty$ if $\gamma < \gamma_m = 3.5$. 104

7.1 The two word vectors "THIS" and "HAVE" are quantized. (a) Exp #1: 5 bins numbered from 0 to 4. (b) Exp #2: 3 bins. 146

7.2 For each dataset, we compute the overall sparsity, median kurtosis, and median skewness. Our data are highly heavy-tailed and severely skewed. Recall normal has zero kurtosis and zero skewness. 151

8.1 Small dataset: co-occurrences and margins for the population. The task is to estimate these values, which are referred to as the gold standard. 170

8.2 The same four words as in Table 8.1 are used for evaluating multi-way associations. There are four three-way combinations and one four-way combination. 182

List of Figures

1.1 *Stable random projections*: $\mathbf{B} = \mathbf{A} \times \mathbf{R}$. $\mathbf{A}$ is the original data matrix. 7

1.2 The sampling procedure of *Conditional Random Sampling (CRS)*. We perform a random permutation on the columns of a sparse data matrix (a) to obtain (b). The inverted index (c) only stores non-zero entries (locations and values). Sketches (d) are the front of the inverted index. 8

2.1 The method of *stable random projections* multiplies the data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ by a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$, to obtain a projected matrix $\mathbf{B} = \mathbf{AR} \in \mathbb{R}^{n \times k}$. 14

3.1 The asymptotic variance ratios, $\frac{\mathrm{Var}(\hat{a}_{MLE})}{\mathrm{Var}(\hat{a}_{MF})}$ and $\frac{\mathrm{Var}(\hat{a}_{MLE})}{\mathrm{Var}(\hat{a}_{SM})}$ verify that the MLE has smaller variance. $\mathrm{Var}\left(\hat{a}_{MLE}\right)$, $\mathrm{Var}\left(\hat{a}_{MF}\right)$, and $\mathrm{Var}\left(\hat{a}_{SM}\right)$ are given in (3.15), (3.6), and (3.13), respectively. We consider $m_2 = 0.2m_1$, $m_2 = 0.5m_1$, $m_2 = 0.8m_1$, and $m_2 = 1.0m_1$. 26

3.2 Estimation of the inner products between "THIS" and "HAVE." (a): $\frac{\mathrm{bias}}{a}$, (b): $\frac{\sqrt{\mathrm{Var}(\hat{a})}}{a}$, (c): $\frac{\sqrt[3]{\mathrm{E}(\hat{a}-a)^3}}{a}$. 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 well. (C): Simulations for the margin-free estimator match the theoretical results. 27

3.3 Simulations show that $\mathbf{Pr}$ (multiple real roots) decreases exponentially fast (notice the log scale in the vertical axis). Here $a' = \frac{a}{\sqrt{m_1 m_2}}$. The curve for the upper bound is given by (3.22). 28

3.4 The ratio of variance $V_{Sign} = \frac{\mathrm{Var}\left(\hat{a}_{Sign}\right)}{\mathrm{Var}(\hat{a}_{MLE})}$ decreases monotonically in $(0, \frac{\pi}{2}]$, with minimum $= \frac{\pi^2}{4}$ attained at $\theta = \frac{\pi}{2}$. 30

4.1 $\tau^2(s,t) = \frac{2 \log \mathrm{E}(\exp(xt))}{t^2}$, where $x = r_{ij}$ is defined in (4.1). $\tau^2(s) = \max\ \tau^2(s,t)$ for $t > 0$. Panel (a) plots $\tau^2(s,t)$ for some range of t and $s = 1, 3, 5, 10$. The peak on each curve corresponds to $\tau^2(s)$. Panel (b) plots $\tau^2(s)$ for $1 \leq s \leq 25$ along with its asymptote $\tau^2(s) \sim \frac{s}{2 \log(2s)}$. . 42

4.2 (a): k inflation factor, $\frac{k^{\tau^2(s)}}{k^1}$ in (4.15). The dashed line with slope 1/3 indicates that we can achieve "effective speedup" more than 3-fold if the curve is below this line. (b): The "effective speedup," defined as $s/\frac{k^{\tau^2(s)}}{k^1}$. Curves above the dashed line indicate the range of s where $s > 3$ may be a better choice than $s = 3$. 45

4.3 The "k inflation factor" as defined in (4.23). We simulate the expectation in (4.23) by first conditioning on the data (10^6 simulations for every D) and then generating the data 15 times for the unconditional expectation. We consider both normal and 3-Pareto. The inflation factor is close to 1. Each curve stands for a particular ϵ, from 0.1 to 1.0. 49

4.4 Using *very sparse random projections* with s from 1 to 1000, after $k > 100$, we can achieve similar performance, in terms of classification errors with a 5-nearest neighbor classifier, as using the original correlation distances. In (a), the average classification errors (over 1000 runs) indicate that when $s = 1$ to 100, the classification accuracies are similar. In (b) we plot the standard errors, which have very similar trend as the mean. The dashed curve stands for $s = 1$. 50

5.1 The bias correction factor, b_{me} in (5.22), as a function of $k = 2m + 1$. After $k > 50$, the bias is negligible. 61

5.2 The ratios of the MSE, $\frac{\mathrm{MSE}(\hat{d}_{me})}{\mathrm{MSE}(\hat{d}_{gm,c})}$ and $\frac{\mathrm{MSE}(\hat{d}_{me,c})}{\mathrm{MSE}(\hat{d}_{gm,c})}$, 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. When $k = 3$, the ratios are ∞. 64

5.3 Tail bound constants for the sample median estimator and the geometric mean estimator, in Lemma 11 and Lemma 15 respectively. Smaller constants imply smaller required sample sizes. 65

5.4 We simulate the tail probabilities with $k = 11, 21, 51, 101, 201$, and 401, for both $\hat{d}_{me,c}$ and $\hat{d}_{gm,c}$. 67

5.5 We 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. 70

5.6 We consider k from 10 to 400. For each k, we simulate standard Cauchy samples, from which we estimate the Cauchy parameter by the MLE $\hat{d}_{MLE,c}$. 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 (5.61)+(5.62). 72

5.7 We compare the inverse Gaussian approximation with the same simulations in Figure 5.6. Panel (a) indicates that empirical tail probabilities can be highly accurately approximated by the inverse Gaussian tail probabilities. Panel (b) compares the empirical tail probabilities with the inverse Gaussian upper bound (5.73)+(5.74), indicating that they are reliable at least in our simulation range. 75

5.8 (a): $\frac{\epsilon^2/(1+\epsilon)}{\log\left(\cos\left(\frac{2\epsilon}{\pi}\right)\right)+\frac{4\epsilon}{\pi^2}\log(1+\epsilon)}$. (b): $\frac{\epsilon^2/(1+\epsilon)}{\log\left(\cos\frac{2\epsilon}{\pi}\right)-\frac{4\epsilon}{\pi^2}\log(1-\epsilon)+\frac{8}{3}\log\left(\cos\frac{\epsilon}{3\pi}\right)}$. It suffices to use a constant 8 in (5.76) and (5.77). The optimal constant will be different for different ϵ. For example, if $\epsilon = 0.2$, we could replace the constant 8 by a constant 5. 82

6.1 We plot the correction factor, $\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^k$, in (6.12) as a function of the sample size k, along with the asymptotic expression $\exp\left(-\gamma_e(\alpha-1)\right)$ (dashed horizontal lines). When $k > 50 \sim 100$, the asymptotic expression is accurate. 91

6.2 We plot $G_{R,\alpha,\epsilon}$ and $G_{L,\alpha,\epsilon,k_0}$ in Lemma 22. The constants are reasonably small. We choose $\alpha = 0.001, 0.01$, and α from 0.1 to 2.0 spaced at 0.1. 93

6.3 The MSE ratios $\frac{\mathrm{MSE}\left(\hat{d}_{(\alpha),gm}\right)}{\mathrm{MSE}\left(\hat{d}_{(\alpha),gm,b}\right)}$. (a): when $\alpha \geq 1$, the ratios are always below 1. α from 1.0 to 2.0, spaced at 0.1. (b): When $0.4 \leq \alpha \leq 1$, the ratios (solid curves) are all above 1.0 and follow the obvious pattern. The ratios for $\alpha = 0.35, 0.30, 0.25, 0.20$, and 0.15 (dashed curves) follow a reverse pattern. Overall, when $0.25 < \alpha < 1$, the MSE ratios are larger than 1, i.e., $\hat{d}_{(\alpha),gm,b}$ should be preferred. 95

6.4 Ratios of the MSE $\frac{\mathrm{MSE}\left(\hat{h}_{me}\right)}{\mathrm{MSE}\left(\hat{h}_{gm}\right)}$. The horizontal (dashed) line is the theoretical asymptotic value (i.e., 1.27). The solid curve is obtained by analytically integrating (6.35). Note that (6.35) is numerically unstable when $k > 25$. 97

6.5 The solid curve represents the theoretical MSE ratios $\frac{\mathrm{MSE}(h_{gm})}{\mathrm{MSE}(h_{mle,c})}$, where we take $\mathrm{MSE}(h_{mle,c}) = h^2\left(\frac{1}{k}+\frac{2}{k^2}\right)$ to generate the curve. The dashed curve replaces $\mathrm{MSE}(h_{mle,c})$ with simulations. The harmonic mean estimator $h_{mle,c}$ considerably outperforms the geometric mean estimator h_{gm}. The plot also indicates that the asymptotic variance formula in Lemma 24 is reliable when $k > 15$. 99

6.6 We plot the MSE ratios to verify that, when $0 < \alpha \leq 0.344$, the harmonic mean estimator $\hat{d}_{(\alpha),hm}$ should be preferred over the geometric mean estimators. In both panels, the dashed curves are the asymptotic MSE ratios, i.e., $\left(\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2} - 1\right) / \left(\frac{\pi^2}{12}(\alpha^2+2)\right)$. To show the non-asymptotic behavior, we also plot the MSE ratios at finite k ($k = 5, 10, 50, 100$, and 500). Panel (a) plots $\frac{\mathrm{MSE}\left(\hat{d}_{(\alpha),hm}\right)}{\mathrm{MSE}\left(\hat{d}_{(\alpha),gm}\right)}$, and Panel (b) plots $\frac{\mathrm{MSE}\left(\hat{d}_{(\alpha),hm}\right)}{\mathrm{MSE}\left(\hat{d}_{(\alpha),gm,b}\right)}$, where we simulate MSE $\left(\hat{d}_{(\alpha),hm}\right)$. We can visualize from these curves that for finite k, the break-even points of α (i.e., the MSE ratio = 1) are actually slightly larger than 0.344. 102

6.7 Panel (a) plots the variance factor $g(\lambda;\alpha)$ as functions of λ for different values of α. We can see that $g(\lambda;\alpha)$ is a convex function of λ and the optimal solution (lowest points on the curves) are between -1 and 0.5 ($\alpha < 2$). Note that there is a discontinuity between $\alpha \to 2-$ and $\alpha = 2$. Panel (b) plots the optimal λ^* as a function of α. Since $\alpha = 2$ is not included, we only see $\lambda^* < 0.5$ in the figure. 103

6.8 We simulated the MSEs for various estimators, 10^6 times for every α and k. The MSEs for the *geometric mean* estimators were computed exactly. The *harmonic mean* estimator was used for $\alpha \leq 0.344$, the *biased geometric mean* for $0.344 < \alpha \leq 1$, and the *unbiased geometric mean* for $1 < \alpha < 2$. The *fractional power* has good accuracy except when k is small and α is close to 2. 106

6.9 The asymptotic Cramér-Rao efficiencies of various estimators for $0 < \alpha < 2$. The asymptotic variance of the *sample median* estimator $\hat{d}_{(\alpha),me}$ is computed from known statistical theory for sample quantiles. We can see that the *fractional power* estimator $\hat{d}_{(\alpha),fp}$ is close to be optimal in a wide range of α; and it always outperforms both the *geometric mean* and the *harmonic mean* estimators. Note that since we only consider $\alpha < 2$, the efficiency of $\hat{d}_{(\alpha),fp}$ does not achieve 100% when $\alpha \to 2-$. 107

6.10 We apply *Cauchy random projections* and *very sparse stable random projections* ($\beta = 0.1, 0.01, 0.001$) and classify the microarray specimens using a 5-nearest neighbor classifier and l_1 distances. Panel (a) show that *Cauchy random projections* can achieve almost the same classification accuracy with 100 projections. Panel (a) also shows that *very sparse stable random projections* with $\beta = 0.1$ and 0.01 perform almost indistinguishably from *Cauchy random projections*. Each curve is averaged over 1000 runs. Panel (b) plots the standard errors, indicating that the classification accuracy does not fluctuate much. 111

7.1 The *sketching* procedure of *Conditional Random Sampling (CRS)*. . . 134

7.2 (a): A data matrix with two rows and $D = 15$. If the column IDs are random, the first $D_s = 10$ columns constitute a random sample. u_i denotes the ith row. (b): Inverted index consists of tuples "ID (Value)." (c): Sketches are the first k_i entries of the inverted index sorted ascending by IDs. In this example, $k_1 = 5$, $k_2 = 6$, $D_s = \min(10, 11) = 10$. Excluding 11(3) in K_2, we obtain the same samples as if we directly sampled the first $D_s = 10$ columns in the data matrix. 136

7.3 The ratios $\mathrm{E}\left(\frac{D}{D_s}\right) / \frac{\max(f_1+1, f_2+1)}{k}$ show that the errors from using (7.10) are usually within 5% when $k \geq 20$. $D = 10^6$, $f_1 = \alpha D$, and $f_2 = \beta f_1$. In each panel, the five curves correspond to different values of $\#\{j : u_{1,j} > 0 \ \& \ u_{2,j} > 0, 1 \leq j \leq D\}$, which measure the correlations and are set to be γf_2, with $\gamma = 0.01, 0.1, 0.5, 0.8, 1.0$, respectively. It is clear that γ does not affect D_s strongly. In each panel, we only simulated 10^3 permutations for each k; and hence the curves are not very smooth. 140

7.4 (a): A data matrix with binned (integer) data, $D = 15$. The entries of $u_1 \in \{0, 1, 2\}$ and $u_2 \in \{0, 1, 2, 3\}$. We construct a 3×4 contingency table for u_1 and u_2 in (b). For example, in three columns ($j = 3$, $j = 7$, and $j = 12$), we have $u_{1,j} = u_{2,j} = 0$, hence the (0,0) entry in the table is 3. Suppose the column IDs of the data matrix are random, we can construct a random sample by taking the first $D_s = 10$ columns, represented by a sample contingency table in (c). 144

7.5 The inner product a (after quantization) between "THIS" and "HAVE" is estimated by both $\hat{a}_{MF,c}$ and $\hat{a}_{MLE,c}$. Results are reported in $\frac{\sqrt{\mathrm{Var}(\hat{a})}}{a}$. The two thin dashed lines both labeled "theore." are theoretical variances, which match the empirical values well especially when sketch sizes $\geq 10 \sim 20$. In this case, marginal histograms help considerably. . 146

7.6 The variance ratios, $\frac{\mathrm{Var}(\hat{a}_{MF})}{\mathrm{Var}\left(\hat{a}_{NRP,MF}\right)}$, show that CRS has smaller variances than random projections, when no marginal information is used. We let $f_1 \geq f_2$ and $f_2 = \alpha f_1$ with $\alpha = 0.2, 0.5, 0.8, 1.0$. For each α, we plot from $f_1 = 0.05D$ to $f_1 = 0.95D$ spaced at $0.05D$. 149

7.7 The ratios, $\frac{\mathrm{Var}\left(\hat{a}_{0/1,MLE}\right)}{\mathrm{Var}\left(\hat{a}_{NRP,MLE}\right)}$, show that CRS usually has smaller variances than random projections, except when $f_1 \approx f_2 \approx a$. 150

7.8 NSF data. CRS is overwhelmingly better than random projections. Upper four panels: ratios (CRS over random projections) of the median absolute errors; values < 1 indicate that CRS does better. Bottom four panels: percentage of pairs for which CRS has smaller errors than random projections; values > 0.5 indicate that CRS does better. Dashed curves correspond to fixed sample sizes while solid curves indicate that we (crudely) adjust sketch sizes according to data sparsity. 153

7.9 NEWSGROUP data. The results are similar to Figure 7.8 for the NSF data. CRS is overwhelmingly better than random projections in approximating inner products and l_2 distances (using margins).CRS is also significantly better in approximating l_1 distances. In this case, it is more obvious that adjusting sketch sizes helps. 154

7.10 COREL image data. This dataset is more heavy-tailed than NSF and NEWSGROUP datasets but not as sparse. CRS is still much better than random projections in approximating inner products as well as l_2 distances (using margins). We observe that at large sample sizes, using margins (i.e., assuming normality) may cause quite noticeable biases. This, however, is a good bias-variance trade-off. 154

7.11 DEXTER data. The data are more heavy-tailed than NSF, NEWSGROUP, and COREL data. CRS still outperforms random projections in approximating inner products, l_1, and l_2 distances (using margins). 155

7.12 MSN data (original). The data are extremely heavy-tailed. Even so, *Conditional Random Sampling* is still significantly better than random projections in approximating inner products, and are also better in approximating l_2 distances using margins. 155

7.13 MSN data (square root weighting). CRS is significantly better than random projections in estimating inner products, l_1 distances, and l_2 distances (using margins). Without using margins, CRS is about the same as random projections in approximating l_2 distances, especially when the sketch sizes are adjusted. 156

7.14 MSN data (logarithmic weighting). The results are even better, compared to Figure 7.13. In particular, when estimating l_2 distances without using margins, CRS is strictly better than random projections. . 157

8.1 The original inverted index is given in (a). $D = 15$ documents in the collection. We generate a random permutation π in (b). We apply π to P_i and store the sketch $\mathrm{K}_i = \mathrm{MIN}_{k_i}(\pi(\mathrm{P}_i))$. For example, $\pi(\mathrm{P}_1) = \{11, 13, 1, 12, 15, 6, 8\}$ and we choose $k_1 = 4$; hence $\mathrm{K}_1 = \{1, 6, 8, 11\}$. We choose $k_2 = 4$, $k_3 = 4$, $k_4 = 3$, and $k_5 = 6$. 162

8.2 (a): A contingency table for word W_1 and word W_2. Cell a is the number of documents that contain both W_1 and W_2, b is the number that contain W_1 but not W_2, c is the number that contain W_2 but not W_1, and d is the number that contain neither. The margins, $f_1 = a+b$ and $f_2 = a+c$ are known as document frequencies. For consistency with the notation for multi-way associations, (a, b, c, d) are also denoted by (x_1, x_2, x_3, x_4). (b): A sample contingency table (a_s, b_s, c_s, d_s), also denoted by (s_1, s_2, s_3, s_4). 163

8.3 An example: $a_s = 20$, $b_s = 40$, $c_s = 40$, $d_s = 800$, $f_1 = f_2 = 100$, $D = 1000$. The estimated $\hat{a} = 43$ for "sample-with-replacement," and $\hat{a} = 51$ for "sample-without-replacement." (a): The likelihood profile, normalized to have a maximum = 1. (b): The log likelihood profile, normalized to have a maximum = 0. 167

8.4 Large dataset: histograms of document frequencies, *df* (left panel), and co-occurrences, a (right panel). Left panel: max document frequency *df* = 42,564, median = 1135, mean = 2135, standard deviation = 3628. Right panel: max co-occurrence a = 33,045, mean = 188, median = 74, standard deviation = 459. 170

8.5 The proposed estimator, $\hat{a}_{MLE}$, outperforms the margin-free baseline, $\hat{a}_{MF}$, in terms of $\frac{\sqrt{\text{MSE}}}{a}$. The quadratic approximation, $\hat{a}_{MLE,a}$, is close to $\hat{a}_{MLE}$. All methods are better than assuming independence (IND). 172

8.6 Smoothing improves the proposed MLE estimators but hurts the margin-free estimator in most cases. The vertical axis is the percentage of relative improvement in $\sqrt{\text{MSE}}$ of each smoothed estimator with respect to its un-smoothed version. 173

8.7 (a): The proposed MLE methods (solid lines) have smaller errors than the baselines (dashed lines). We report the mean absolute errors (normalized by the mean co-occurrences, 188), averaged over six permutations. The proposed MLE and the recommended quadratic approximation are very close. Both are well below the margin-free (MF) baseline and the independence (IND) baseline. (b): Percentage of improvement due to smoothing. Smoothing helps MLE, but hurts MF. 174

8.8 We can find many of the most obvious associations with very little work. Two sets of cosine scores were computed for the 468,028 pairs. The gold standard scores were computed over the entire dataset, whereas sample scores were computed over a sample. The plots shows the percentage of agreement between these two lists, as a function of S. As expected, agreement rates are high ($\approx 100\%$) at high sampling rates (0.5). But it is reassuring that agreement rates remain pretty high ($\approx 70\%$) even when we crank the sampling rate way down (0.003). 175

8.9 Definitions of recall and precision. L = total number of pairs. L_G = number of pairs from the top of the gold standard similarity list. L_S = number of pairs from the top of the reconstructed similarity list. . 176

8.10 Precision-recall curves in retrieving the top-1% and top-10% pairs, at different sampling rates from 0.003 to 0.5. Note that the precision is $\geq \frac{L_G}{L}$. 176

8.11 In terms of $\frac{\sqrt{\text{MSE}(x_1)}}{x_1}$, the proposed MLE is consistently better than the margin-free baseline (MF), which is better than the independence baseline (IND), for four three-way association cases. 183

8.12 The simple "add-one" smoothing improves the accuracies for the MLE. Smoothing, however, in all cases except Case 3-1 hurts the margin-free estimator. 184

8.13 Four-way associations (Case 4). (a): The MLE has smaller MSE than the margin-free (MF) baseline, which has smaller MSE than the independence baseline. (b): Smoothing considerably improve the accuracy for MLE. 185

8.14 (a): Combining the three-way, four-way, and the two-way association results for the four words in the evaluations, the average relative improvements of $\sqrt{\text{MSE}}$ suggests that the proposed MLE is consistently better the MF baseline but the improvement decreases monotonically as the order of associations increases. (b): Average $\sqrt{\text{MSE}}$ improvements due to smoothing imply that smoothing becomes more and more important as the order of association increases. 186

8.15 We plot V_B in (8.53) for the whole range of f_1, f_2, and a, assuming equal samples: $k_1 = k_2 = k$. (a), (b), (c) and (d) correspond to $f_2 = 0.2f_1$, $f_2 = 0.5f_1$, $f_2 = 0.8f_1$ and $f_2 = f_1$, respectively. Different curves are for different f_1's, ranging from $0.05D$ to $0.95D$ spaced at $0.05D$. The horizontal lines are $\frac{\max(f_1,f_2)}{f_1+f_2}$. We can see that for all cases, $V_B \leq 1$ holds. $V_B = 1$ when $f_1 = f_2 = a$, a trivial case. When a/f_2 is small, $V_B \approx \frac{\max(f_1,f_2)}{f_1+f_2}$ holds well. It is also possible that V_B is very close to zero. 189

8.16 Compared with Figure 8.15, proportional samples further reduce V_B. Note that (a) and (b) are "squashed" versions of (a) and (b), respectively, in Figure 8.15. 190

8.17 When estimating the resemblance, our algorithm gives consistently more accurate answers than Broder's sketch. In our experiments, Broder's "minwise" construction gives almost the same answers as the "original" sketch, thus only the "minwise" results are presented here. The approximate MLE again gives very close answers to the exact MLE. Also, smoothing improves at low sampling rates. 191

8.18 Compared with Broder's sketch, the relative MSE improvement should be, approximately, $\frac{\min(f_1,f_2)}{f_1+f_2}$ with equal samples, and $\frac{1}{2}$ with proportional samples, i.e., the two horizontal lines. The actual improvements could be lower or higher. The figure verifies that proportional samples can considerably improve the accuracies. 192

Chapter 1

Introduction

Data, Data, Everywhere *Megaterabyte databases are getting downright common. But with more real-time data, complex queries, and increasing numbers of sources, managing them is anything but routine.*

—by Charles Babcock, **Information Week** (From the January 9, 2006 issue).

1.1 Massive Data

The following facts are quoted from *Information Week* (Jan. 9, 2006)[1]:

- The amount of data stored by businesses nearly doubles every 12 to 18 months.
- Databases are more real-time. Wal-Mart Stores Inc. refreshes sales data hourly, adding a billion rows of data a day, allowing more complex searches. EBay Inc. lets insiders search auction data over short time periods to get deeper insight into what affects customer behavior.
- The biggest databases are run by the Stanford Linear Accelerator Center, NASA's Ames Research Center, the National Security Agency, etc, in the petabyte (1,000-terabyte, i.e., 10^{15}) range.

[1] `http://www.informationweek.com/news/showArticle.jhtml?articleID=175801775`

The ubiquitous phenomenon of massive datasets brings computational challenges in many scientific and commercial applications including astrophysics, bio-technology, demographics, finance, geographical information systems, government, medicine, telecommunications, the environment and the Internet[1].

1.1.1 Massive Web Data

How large is the Web? The page hits in Table 1.1 suggest that the state-of-the-art search engines[2] have collected roughly $D = 10^{10}$ Web pages, inferred from the page hits of two function words "A" and "THE." Table 1.1 also tells that even those so-called "rare words" have a large number of hits.

Table 1.1: Page hits for two high frequency words and a few low frequency words.

Query	Hits (MSN)	Hits (Google)
A	2,452,759,266	3,160,000,000
The	2,304,929,841	3,360,000,000
Kalevala	159,937	214,000
Griseofulvin	105,326	149,000
Saccade	38,202	147,000

How many page hits do "ordinary" words have? To address this question, we randomly picked 15 pages from a learners' dictionary [94] (which has 57,100 entries), and selected the first entry on each page. According to Google, there are 10 million pages/word (median value, aggregated over the 15 words).

How many words are there in the English language? Here, we quote *AskOxford.com*[3]:

This suggests that there are, at the very least, a quarter of a million distinct English words, excluding inflections, and words from technical and regional vocabulary not covered by the OED, or words not yet added to the published dictionary, of which perhaps 20 per cent are no longer in current use. If distinct senses were counted, the total would probably approach three quarters of a million.

[2] All experiments with MSN.com and Google were conducted in August, 2005.

[3] `http://www.askoxford.com/asktheexperts/faq/aboutenglish/numberwords`

Therefore, if we construct a "term-by-document" matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ at Web scale, it is roughly $n \approx 10^6$ and $D = 10^{10}$. Here the (i, j) entry in $\mathbf{A}$ is the number of occurrences of word i in document j.

Working with such a large-scale data matrix can be challenging. For example, the LSI (*latent semantic indexing*)[58], a popular topic model, applies a singular value decomposition (SVD) on the term-by-document matrix, which is probably computationally infeasible at Web scale.

One major issue with massive datasets is the computer memory, because the size and speed of physical memory increase at a much slower rate than the processors (CPU). This phenomenon is known as the *memory wall* [168, 139]. For example, while it might be possible to pre-compute all pair-wise co-occurrences (associations) of $\mathbf{A}$, it may be infeasible to materialize all of them in the physical memory. Moreover, sometimes multi-way associations are also sought because queries can consist of more than two words. A feasible solution is to store a "sample" of $\mathbf{A}$ and estimate any co-occurrences *on the fly*. We suspect this is the strategy adopted by modern search engines, although the actual details are of course "trade secrete."

It is desirable that the estimates be consistent. Joint frequencies ought to decrease monotonically as we add terms to the query. Table 1.2 shows that estimates produced by current search engines are not always consistent.

Table 1.2: Joint frequencies ought to decrease monotonically as we add terms to the query, but estimates produced by search engines sometimes violate this invariant.

Query	Hits (MSN)	Hits (Google)
America	150,731,182	393,000,000
America & China	15,240,116	66,000,000
America & China & Britain	235,111	6,090,000
America & China & Britain & Japan	154,444	23,300,000

Although the total number of (correctly-spelled) English words is already impressive, in many text mining applications, we will have to work with even much higher dimensions. While a document may be represented as a vector of single words (i.e.,

the *bag-of-words* model), it is usually better to model a document as a vector of *l-contiguous* words (called *l-shingles* [34]). For example, using a 3-shingle model, the sentence "it is a nice day" will be parsed to be {"it is a", "is a nice", "a nice day"}. This model dramatically increases data dimensions, because, if there are in total 10^6 single English words, a 3-shingle model will boost the dimension from 10^6 to 10^{18}.

1.1.2 Massive Data Streams

Massive data streams are fundamental in many modern data processing applications. Data streams come from Internet routers, phone switches, atmospheric observations, sensor networks, highway traffic conditions, finance data, etc.[91, 69, 96, 19, 49, 141, 5]

Unlike traditional databases, it is not common to store (rapid-arriving) massive data streams; and hence the processing is often done "on the fly." For example, sometimes we only need to "visually monitor" the data by observing the time history of certain summary statistics, e.g., sum, number of distinct items, or some l_α norm. In some applications (e.g., audio/video classification and segmentation), we will need to build statistical (learning) models for classifications (or clustering) of massive data streams; but often we can only afford one-pass of the data.

One important feature of data streams is that they are dynamic. As a popular model, a stream u consists of entries (i, u_i), $i = 1$ to D. For example, $D = 2^{64}$ when a stream represents the arrivals of IP addresses[4] The entries can arrive in any order and can be frequently updated. The dynamic nature of massive data streams makes many sampling task more challenging than in static data.

1.2 Sampling Massive Data and the Challenges

While there are many interesting and challenging problems brought by modern massive datasets, this thesis concentrates on developing *sampling techniques* for efficiently

[4]Although we often do not know the precise D value of a stream, in most applications it suffices to use a conservative upper bound. For example, $D = 2^{64}$ if the stream represents the arrivals of IP addresses. This is also one of the reasons why the data are highly sparse. Note that assuming a very large D will not affect the computation of distances and will not affect the sampling algorithms discussed in this thesis.

computing *distances* in ultra high-dimensional data, using small (memory) space.

It is often the case that, in statistical modeling and machine learning applications, we only need *distances*, especially *pair-wise distances*, as opposed to the original data. For example, computing the *Gram matrix* $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ is common in statistics and machine learning. $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ represents all pair-wise *inner products* in the data matrix $\mathbf{A}$.

Given two data points u_1, $u_2 \in \mathbb{R}^D$, their inner product (denoted by a) and l_α distance (denoted by $d_{(\alpha)}$) are[5]

$$a = u_1^{\mathrm{T}} u_2 = \sum_{i=1}^{D} u_{1,i} u_{2,i}, \tag{1.1}$$

$$d_{(\alpha)} = \sum_{i=1}^{D} |u_1 - u_2|^\alpha. \tag{1.2}$$

Note that both the inner product and distance are defined as a sum of D terms. Therefore, when the data are too massive to handle efficiently, sampling is very natural to come along as one can simply "randomly" select k (out of D) terms to approximate the sum (up to a scaling factor $\frac{D}{k}$). In terms of the data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$, *random coordinate sampling* picks k columns from the matrix uniformly at random.

Sampling is beneficial because it saves both the CPU (processors) cycles and the storage (memory) space. In modern applications, it is often the case that the savings in the storage (memory) space will be more important. In the past half a century, the computing bottleneck has been the memory, not the processors. Processor speeds keep increasing approximately 75% per year while the memory speeds increase about 7% per year[139], a phenomenon known as the "memory wall"[168, 139]. Therefore, in applications involving massive datasets, it is often the most critical task to represent the data, e.g., via sampling, in a compact format to fit the size of available memory.

[5]We define the l_α distance by $d_{(\alpha)} = \sum_{i=1}^{D} |u_1 - u_2|^\alpha$ instead of $\left(\sum_{i=1}^{D} |u_1 - u_2|^\alpha\right)^{1/\alpha}$ because the former is more convenient and more popular in applications. For example, the JL Lemma in literature is often presented in terms of the squared l_2 distance $\sum_{i=1}^{D} |u_{1,i} - u_{2,i}|^2$ instead of $\left(\sum_{i=1}^{D} |u_{1,i} - u_{2,i}|^2\right)^{1/2}$. In this thesis, we will simply call $\sum_{i=1}^{D} |u_{1,i} - u_{2,i}|^2$ the "l_2 distance" instead of the "squared l_2 distance."

1.2.1 Advantages of Random Coordinate Sampling

Random coordinate sampling is often the default option at least for two reasons.

- **Simplicity** It takes time only $O(nk)$ for sampling k columns from $\mathbf{A} \in \mathbb{R}^{n \times D}$.
- **Flexibility** The same set of samples can be used for approximating many different summary statistics including inner products, l_α distances (for any α), and multi-way associations.

1.2.2 Disadvantages of Random Coordinate Sampling

Random coordinate sampling, however, suffers from two major problems.

- It is often not accurate, because entries with large values are likely to be missed by random sampling, especially when the data are *heavy-tailed*. Real-world large-scale datasets (especially Internet data) are ubiquitously heavy-tailed and follow the "power-law" [111, 53, 66, 142]. When estimating the l_2 distances or inner products, the variances of the estimates are determined by the fourth moments of the data. In heavy-tailed data, however, sometimes even the first moment may not be meaningful (i.e., bounded)[142].
- It does not handle sparse data efficiently. Many large-scale datasets are highly sparse, for example, text data [60] and market-basket data [7, 158]. Except for a few function words such as "A" and "The," most words only appear in a very small fraction (e.g., $< 1\%$) of documents. If we sample the data by fixing a fraction of columns, it is likely we will miss most of the informative (non-zero) entries, especially those (interesting) jointly non-zero entries.

In this thesis, we will discuss two "sampling" techniques for massive datasets. The method of *stable random projections* can efficiently handle heavy-tailed data, at least for approximating the l_α distances; while the method of *Conditional Random Sampling (CRS)* is specifically designed for sampling sparse data.

1.3 Stable Random Projections

Pictured in Figure 1.1, the idea of *stable random projections* [104, 96, 99, 166, 126, 127, 116] is to multiply the data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ by a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ ($k \ll D$), resulting in a projected matrix $\mathbf{B} = \mathbf{A} \times \mathbf{R} \in \mathbb{R}^{n \times k}$. $\mathbf{B}$ is much smaller than $\mathbf{A}$ and hence it can be more easily stored (e.g., small enough for physical memory).

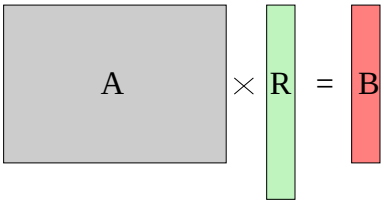


Figure 1.1: *Stable random projections*: $\mathbf{B} = \mathbf{A} \times \mathbf{R}$. $\mathbf{A}$ is the original data matrix.

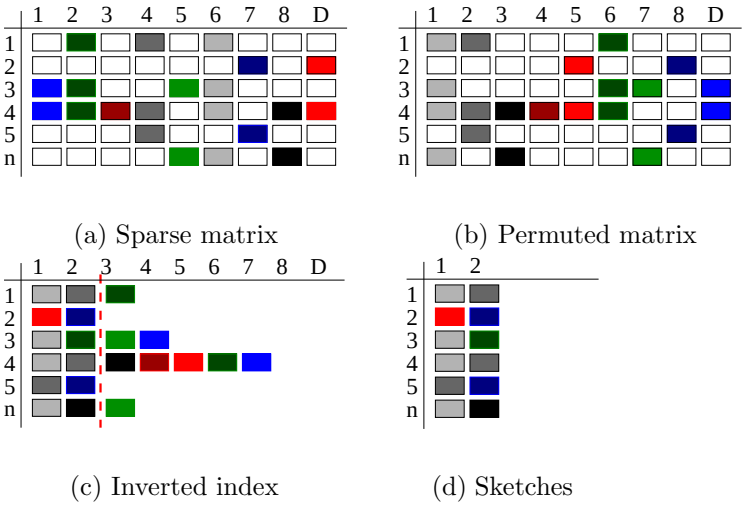
The projection matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ typically consists of i.i.d. samples from a symmetric α-stable distribution [171] (hence the name "stable random projections"). By properties of stable distributions, the projected data also follow α-stable distributions, from which we can estimate the original individual l_α norms and the pair-wise l_α distances in $\mathbf{A}$. Therefore, if all we care about are the l_α properties of the data, we can "throw away" the original data.

The success of *stable random projections* is highlighted by the Johnson-Lindenstrauss (JL) Lemma [103] for dimension reduction in l_2, which says $k = O\left(\frac{\log n}{\epsilon^2}\right)$ suffices to guarantee that any pair-wise l_2 distance among n points (in any dimensions) can be approximated within a $1 \pm \epsilon$ factor with high probability.

The JL Lemma, however, does not exist for l_α ($\alpha < 2$) if we have to use estimators that are metrics (e.g., satisfying triangle inequality). This so-called "Impossibility" result[32, 109, 33] fortunately does not rule out estimators that are not metrics. In this thesis, we will discuss a variety of non-metric estimators using the geometric mean, the harmonic mean, the fractional power, as well as the maximum likelihood.

1.4 Conditional Random Sampling (CRS)

Conditional Random Sampling (CRS), as pictured in Figure 1.2, is a technique particularly suitable for sampling sparse data. Conceptually, the first step of CRS is to apply a random permutation on columns of the data matrix $\mathbf{A}$. The *inverted index* stores only the non-zero entries of the permuted matrix. The sketches ("samples") are simply the front of the inverted index.



(a) Sparse matrix (b) Permuted matrix

(c) Inverted index (d) Sketches

Figure 1.2: The sampling procedure of *Conditional Random Sampling (CRS)*. We perform a random permutation on the columns of a sparse data matrix (a) to obtain (b). The inverted index (c) only stores non-zero entries (locations and values). Sketches (d) are the front of the inverted index.

The estimation task of *Conditional Random Sampling* is interesting because sketches are not "random samples." We will show that sketches are almost random samples on a pair-wise (or group-wise) basis and hence the estimation task is straightforward although theoretical analysis will be conditional on the "sample size," which is a random variable and differs pair-wise (or group-wise).

CRS is highly effective in sparse data because it utilizes the current commonly-used data structure, i.e., the inverted index. For those data points that are extremely sparse, we can probably afford to store the whole inverted index. Since we construct (conditional) random samples pair-wise (or group-wise), CRS shares the advantages

of regular random coordinate sampling in that the same sketches can be used for approximating many different summary statistics, including multi-way associations.

1.5 Applications

There has been considerable interest in sampling and sketching techniques[13, 160, 100, 134, 43, 3, 83, 127, 115], which are useful for numerous applications such as association rules[31, 30], clustering[159, 35, 6, 89, 90], query optimization[136, 44], duplicate detection[34, 28], and more. Sampling methods become more and more important with larger and larger collections.

Broder's sketch[34] was originally introduced to detect duplicate Web pages. Many URLs point to the same (or nearly the same) HTML blobs. Approximate answers are often good enough. We don't need to find all duplicates, but it is handy to find many of them, without spending more than it is worth on computational resources.

In information retrieval (IR) applications, physical memory is often a bottleneck, because the Web collection is too large for the memory but we want to minimize seeking data in the disk as the query response time is critical[29]. As a space saving device, dimension reduction techniques use a compact representation to produce approximate answers in the physical memory.

We mentioned page hits. If we have a two-word query, we'd like to know how many pages mention both words. We assume that pre-computing and storing page hits is infeasible, at least not for infrequent pairs of words (and multi-word sequences).

It is customary in IR to start with a large term-by-document matrix, with entry values indicating the occurrences of a term in a document. Depending on specific applications, we can construct an inverted index and store sketches either for terms (to estimate word association) or for documents (to estimate document similarity).

1.5.1 Association Rule Mining

"Market-basket" analysis and association rules[8, 10, 9] [88, Chapter 14.2] are useful tools for mining commercial databases. Commercial databases tend to be large and

sparse[7, 158]. Various sampling algorithms have been proposed[162, 45]. Sampling makes it possible to estimate association rules online, which may have some advantage in certain applications[92].

1.5.2 All Pair-Wise Associations (Distances)

In many applications including distance-based classification or clustering and bi-gram language modeling [48], we need to compute all pair-wise associations (or distances). Given a data matrix $\mathbf{A}$ of n rows and D columns, brute force computation of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ would cost $O(n^2D)$, or more efficiently, $O(n^2\bar{f})$, where $\bar{f}$ is the average number of non-zeros among all rows of $\mathbf{A}$. Brute force could be very time-consuming. In addition, when the data matrix is too large to fit in physical memory, the computation may become especially inefficient.

1.5.3 Estimating Distances Online

While the original data matrix $\mathbf{A} \in \mathbb{R}^{n\times D}$ may be too large for physical memory, storing (materializing) all pair-wise distances and associations in $\mathbf{A}$ takes space $O(n^2)$, which can be also too large for the memory, let alone multi-way associations. In many applications such as online learning, online recommendation systems, online market-basket analysis and search engines, it may be more efficient if we store the samples (sketches) in the memory and estimate any distance *on the fly* only when it is necessary.

1.5.4 Database Query Optimization

In databases, an important task is to determine the order of multi-way joins, which has a large impact on the system performance[81, Chapter 16]. Based on estimates of two-way, three-way, and even higher-order join sizes, query optimizers construct a plan to minimize a cost function (e.g., intermediate writes). Efficiency is critical as we certainly do not want to spend more time optimizing the plan than executing it.

We use an example (called *Governator*) to illustrate that estimates of two-way and multi-way association can help the query optimizer.

Table 1.3: Page hits returned by Google for four words and their two-way, three-way, and four-way combinations.

	Query	Hits (Google)
One-way	Austria	88,200,000
	Governor	37,300,000
	Schwarzenegger	4,030,000
	Terminator	3,480,000
Two-way	Governor & Schwarzenegger	1,220,000
	Governor & Austria	708,000
	Schwarzenegger & Terminator	504,000
	Terminator & Austria	171,000
	Governor & Terminator	132,000
	Schwarzenegger & Austria	120,000
Three-way	Governor & Schwarzenegger & Terminator	75,100
	Governor & Schwarzenegger & Austria	46,100
	Schwarzenegger & Terminator & Austria	16,000
	Governor & Terminator & Austria	11,500
Four-way	Governor & Schwarzenegger & Terminator & Austria	6,930

Table 1.3 shows estimates of page hits for four words and their two-way, three-way, and four-way combinations. Suppose the optimizer wants to construct a plan for the query: "Governor, Schwarzenegger, Terminator, Austria." The standard solution starts with the least frequent terms: (("Schwarzenegger" $\cap$ "Terminator") $\cap$ "Governor") $\cap$ "Austria." That plan generates 579,100 intermediate writes after the first and second joins. An improvement would be (("Schwarzenegger" $\cap$ "Austria") $\cap$ "Terminator") $\cap$ "Governor," reducing the 579,100 down to 136,000.

1.5.5 (Sub-linear) Nearest Neighbor Searching

Computing nearest neighbors is of major importance in many applications. However, due to the "curse of dimensionality," the current solution to efficiently finding nearest neighbors (even approximately) is still far from satisfactory[100, 88].

For computational considerations, there are two major issues in nearest neighbor searching. Firstly, the original data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ may be too large for physical memory; but scanning hard disks for nearest neighbors would be too slow. Secondly, searching for the nearest neighbor of one data point would cost $O(nD)$, which can be too time-consuming.

Apparently, both sample techniques presented in this thesis can save the storage and speed up the computations. For example, once the original data matrix $\mathbf{A}$ is reduced to $\mathbf{B} \in \mathbb{R}^{n \times k}$ with $k \ll D$, the storage cost is much reduced and the computational cost decreases from $O(nD)$ down to only $O(nk)$. However, $O(nk)$ is still linear (suppressing the logarithmic factor). It is often desirable to further reduce the computational cost from $O(n)$ to $O(n^\gamma)$ for some $\gamma < 1$, at least for certain applications.

Two major categories of sub-linear nearest neighbor algorithms are the *KD-Trees* (and variants) [79, 80] and *Locality-Sensitive Hashing (LSH)* [100, 56, 15]. These algorithms often work with a metric space (which satisfies the triangle inequality). For example, the l_α space is a metric when $\alpha \geq 1$. When searching for the nearest neighbors in l_α ($\alpha \geq 1$), we can (fairly easily) reduce the search space quite substantially by the triangle inequality, i.e., no need to search for all n data points.

In extremely high dimensional data, however, current sub-linear nearest neighbor algorithms, including KD-trees and LSH, are still not yet satisfactory. One of the major issues is that these techniques require *super-linear* memory in order to reduce the computational cost, which can be problematic as physical memory (instead of CPU) is often the bottleneck (again, the *memory wall*). [100] mentioned a LSH scheme which combined hashing with random projections. This scheme, unfortunately, is often impractical due to its high preprocessing cost[100].

In this thesis, the major accomplishment is to reduce the data size $\mathbf{A} \in \mathbb{R}^{n \times D}$ substantially down to $\mathbf{B} \in \mathbb{R}^{n \times k}$; and we provide accurate estimators to recover the original distances in $\mathbf{A}$ from $\mathbf{B}$. While we believe in important scenarios our results could provide a satisfactory solution to approximate nearest neighbors, developing sub-linear algorithms on top of our algorithms would be interesting future research. One major obstacle is that most of our estimators are not metric; and hence designing

clever algorithms and theoretical analysis would be difficult, although not impossible.

1.6 Organization

Chapters 2 to 6 are devoted to *stable random projections* while Chapters 7 and 8 focus on *Conditional Random Sampling (CRS)*.

Chapter 2 contains an introduction to *stable random projections*. Chapter 3 concentrates on the l_2 case, i.e., *normal random projections*. Chapter 4 discusses *sub-Gaussian random projections* and *very sparse random projections*, also for the l_2 case. Chapter 5 is devoted to the l_1 case, i.e., *Cauchy random projections*. Chapter 6 presents some general results for *stable random projections*.

Chapter 7 describes *Conditional Random Sampling* for the general scenario. Because boolean (0/1) data are particularly interesting in high-dimensional data analysis, Chapter 8 will analyze CRS for boolean data in more detail. Also, Chapter 8 will compare CRS with Broder's sketch[34], a standard technique in information retrieval.

Finally Chapter 9 concludes this thesis.

1.7 Notation

We will present a variety of methods for different situations. The notation will be consistent within each chapter; but for simplicity, we have to change the notation in different chapters.

Chapter 2

Introduction to Stable Random Projections

The method of *stable random projections*[104, 96, 99, 19, 166, 116] is a useful tool in data mining and machine learning, for efficiently computing the l_α ($0 < \alpha \leq 2$) distances in massive data (e.g., the Web or massive data streams) using a small (memory) space in one-pass of the data.

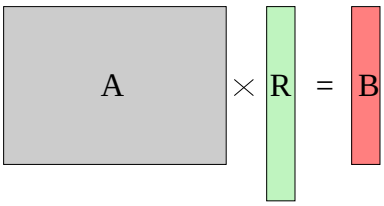


Figure 2.1: The method of *stable random projections* multiplies the data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ by a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$, to obtain a projected matrix $\mathbf{B} = \mathbf{AR} \in \mathbb{R}^{n \times k}$.

As pictured in Figure 2.1, the idea of *stable random projections* is to multiply the original data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ with a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ ($k \ll D$), resulting in a projected matrix $\mathbf{B} \in \mathbb{R}^{n \times k}$. The entries of $\mathbf{R}$ are typically sampled i.i.d. from an α-stable distribution (and hence the name *stable random projections*). Note that a 2-stable distribution corresponds to normal and a 1-stable distribution is Cauchy.

The special case, *normal random projections* (i.e., $\alpha = 2$) has been studied quite thoroughly; see the monograph[166]. Thus, a considerable portion of this thesis is

devoted to *stable random projections* with $\alpha < 2$.

After an overview of *stable random projections* for general $0 < \alpha \leq 2$ in this chapter, Chapter 3 will provide more details on the l_2 case, i.e., *normal random projections*, and present the idea of improving *normal random projections* by taking advantage of the marginal information. Chapter 4 simplifies *normal random projections* by sampling $\mathbf{R}$ from a *sparse* three-point $[-1,\ 0,\ 1]$ distribution, which is a special case of *sub-Gaussian* distribution. Chapter 5 is devoted to the l_1 norm, i.e., *Cauchy random projections*. Finally, Chapter 6 discusses the general case for $0 < \alpha \leq 2$.

2.1 The Fundamental Problem in Stable Random Projections

The fundamental issue in *stable random projections* is a statistical estimation problem.

Recall that we multiply the data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ by a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$, to obtain a (much smaller) matrix $\mathbf{B} = \mathbf{A} \times \mathbf{R} \in \mathbb{R}^{n \times k}$. The goal is to recover the original summary properties (e.g., norms and distances) of $\mathbf{A}$ from $\mathbf{B}$.

Without loss of generality, we focus on the leading two rows, $u_1,\ u_2 \in \mathbb{R}^D$ in $\mathbf{A}$ and the leading two rows, $v_1,\ v_2 \in \mathbb{R}^k$ in $\mathbf{B}$. Denote $\mathbf{R} = \{r_{ij}\}_{i=1}^{D}{}_{j=1}^{k}$. Then

$$v_{1,j} = \sum_{i=1}^{D} r_{ij} u_{1,i}, \quad v_{2,j} = \sum_{i=1}^{D} r_{ij} u_{2,i}, \quad x_j = v_{1,j} - v_{2,j} = \sum_{i=1}^{D} r_{ij} \left(u_{1,i} - u_{2,i}\right). \quad (2.1)$$

2.1.1 Stable Distributions

Typically $r_{ij} \sim S(\alpha, 1)$, i.i.d., although in later chapters we will discuss simpler alternatives. Here $S(\alpha, 1)$ stands for a symmetric α-stable random distribution[171] with the index parameter α and the scale parameter 1.

A random variable z is symmetric α-stable if its characteristic function is

$$\mathrm{E}\left(\exp\left(\sqrt{-1} z t\right)\right) = \exp\left(-d|t|^\alpha\right), \quad (2.2)$$

where $d > 0$ is the scale parameter. We write $z \sim S(\alpha, d)$, which in general does not

have a closed-form density function except for $\alpha = 2$ (normal) or $\alpha = 1$ (Cauchy).

2.1.2 The Statistical Estimation Problem

By properties of Fourier transforms, it is easy to see that the projected data also follow α-stable distributions with the scale parameters being the l_α properties (norms, distances) of the original data in $\mathbf{A}$. In particular,

$$v_{1,j} \sim S\left(\alpha, \quad \sum_{i=1}^{D} |u_{1,i}|^\alpha\right), \qquad v_{2,j} \sim S\left(\alpha, \quad \sum_{i=1}^{D} |u_{2,i}|^\alpha\right), \tag{2.3}$$

$$x_j = v_{1,j} - v_{2,j} \sim S\left(\alpha, \quad d_{(\alpha)} = \sum_{i=1}^{D} |u_{1,i} - u_{2,i}|^\alpha\right). \tag{2.4}$$

Therefore, the task boils down to estimating the scale parameter from k i.i.d. samples $x_j \sim S(\alpha, d_{(\alpha)})$. Because no closed-form density function is available except for $\alpha = 1, 2$, the estimation task is an interesting problem if we seek estimators that are both statistically accurate and computationally efficient.

A closely related task is to determine the sample size k. The standard technique is to bound the tail probability $\mathbf{Pr}\left(|\hat{d}_{(\alpha)} - d_{(\alpha)}| > \epsilon d_{(\alpha)}\right)$, where $\hat{d}_{(\alpha)}$ is an estimator of $d_{(\alpha)}$ and ϵ is the desired accuracy (typically $0 < \epsilon < 1$). Ideally we hope to show[1]

$$\mathbf{Pr}\left(|\hat{d}_{(\alpha)} - d_{(\alpha)}| > \epsilon d_{(\alpha)}\right) \leq 2\exp\left(-k\frac{\epsilon^2}{G}\right), \tag{2.5}$$

for some constant G (which can be a function of ϵ).

For a given data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$, there are in total $\frac{n(n-1)}{2} < \frac{n^2}{2}$ pair-wise distances. We usually would like to bound the tail probabilities simultaneously for

[1]Due to the central limit theorem, an estimator $\hat{d}_{(\alpha)}$ from k samples converges to normal under mild regularity conditions. By the normal tail bounds, we know, at least for certain parameters, $\mathbf{Pr}\left(|\hat{d}_{(\alpha)} - d_{(\alpha)}| \geq \epsilon d_{(\alpha)}\right) \leq 2\exp\left(-k\frac{\epsilon^2}{2V}\right)$ should hold; here $\frac{V}{k}$ is the asymptotic variance of $\hat{d}_{(\alpha)}$. Therefore, at least for a sanity check, we can verify whether the tail bound has the optimal rate by checking if $\lim_{\epsilon \to 0+} G = 2V$.

all pairs by some $0 < \delta < 1$. By the Bonferroni union bound, it suffices if

$$\frac{n^2}{2} 2 \exp\left(-k\frac{\epsilon^2}{G}\right) < \delta \Longrightarrow k > G\frac{2\log n - \log \delta}{\epsilon^2}, \tag{2.6}$$

which immediately leads to the following statement: *If* $k > G\frac{2\log n - \log \delta}{\epsilon^2} = O\left(\frac{\log n}{\epsilon^2}\right)$, *then with probability at least* $1-\delta$, *using* $\hat{d}_{(\alpha)}$ *can recover any pair-wise* l_α *distance among* n *points within a* $1 \pm \epsilon$ *factor of the truth.*

The above statement has been proved for $\alpha = 2$, known as the Johnson-Lindenstrauss (JL) Lemma[103], directly using the projected l_2 distance, i.e., $\sum_{j=1}^{k} |v_{1,j} - v_{2,j}|^2$, which is the *arithmetic mean*. Since this estimator is also a metric (e.g., satisfying the triangle inequality), the JL Lemma becomes extremely useful. In fact, many versions of the JL Lemma have been proved for the l_2 case[103, 77, 100, 17, 55, 96, 97, 3, 18, 12].

As mentioned in Chapter 1, when an estimator is a metric, we can in principle develop efficient *sub-linear* algorithms for finding (approximate) nearest neighbors, for example, using *KD-trees*[79, 80]. For *normal random projections*, since the estimator $\sum_{j=1}^{k} |v_{1,j} - v_{2,j}|^2$ is not only a metric but also the l_2 distance in the projected space, we can directly use the projected data $\mathbf{B}$ to any algorithms in l_2, e.g., linear regressions and support vector machine (SVM) with Gaussian kernels.

When $\alpha < 2$, unfortunately, there is no JL Lemma if we limit the estimators to be metrics, known as the *impossibility result* [32, 109, 33]. In this thesis, we will present an "analog" of the JL Lemma for general $0 < \alpha \leq 2$ using a *geometric mean* estimator. Because the *geometric mean* is not a metric, it becomes more difficult to develop *sub-linear* nearest neighbor algorithms to further speed up the computations. On the other hand, we expect for future research there is some hope that we can develop some *hashing* schemes [100, 56, 15] for the *geometric mean* to speed up nearest neighbor searching.

For many applications mentioned in Chapter 1 including computing all pair-wise distances, estimating distances online, and database query optimizations, we do not need estimators to be metrics. Even for finding nearest neighbors, since the bottleneck is often in physical memory (not CPU), known as the *memory wall*, linear scans (with our estimators) may already provide a satisfactory solution in some important

scenarios.

To conclude this section, we shall mention that, although the *geometric mean* estimator is convenient for theoretical analysis especially for tail bounds, other estimators, such as the *harmonic mean* and *fractional power* estimators may be more accurate, at least for some α.

2.2 Applications

There have been a lot of applications of *stable random projections*, in data streams, data mining, and machine learning, especially for the l_2 (normal) case, for example, in machine learning[17, 26, 4, 76, 72, 20], VLSI layout[165], Latent Semantic Indexing (LSI)[144, 130], set intersections[13, 43, 146], finding motifs in bio-sequences[39, 114], face recognition[84], privacy preserving distributed data mining[132], and very recently, sparse signal recovery[62, 41].

Since Chapter 1 has discussed applications of sampling techniques in general, this section will only describe applications of *stable random projections* in data streams.

2.2.1 Data Stream Computations

In data stream computations, *stable random projections* can be useful at least for (A): approximating the l_α frequency moments for individual streams; (B): approximating the l_α differences between a pair of streams; (C): approximating the number of non-zero items (the Hamming norm) in a stream using very small α[49, 50]. We will comment more on (C) in the next subsection.

One important characteristic of massive data streams is that the data do not necessarily arrive in order and they are subject to frequent updating. [96, 99] described the procedure of using Cauchy random projections to approximate the l_1 norms (or l_1 differences) in data streams. For a stream, u_1, which contains pairs (i, $u_{1,i}$), $i \in \{1, 2, ..., D\}$, [96, 99] suggested the following steps:

- Choose $k = O\left(\frac{1}{\epsilon^2}\right)$. Initialize $v_{1,j} = 0$, $j = 1, 2, ..., k$. (We will provide the precise constant for choosing k.)

- Generate a matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$, with entries r_{ij} i.i.d. samples of standard Cauchy.
- For each new pair $(i,\ u_{1,i})$, modify $v_{1,j} = v_{1,j} + r_{ij}u_{1,i}$, for each $j = 1, 2, ..., k$.
- Return median$\{|v_{1,1}|, |v_{1,2}|, ..., |v_{1,k}|\}$ as the approximate l_1 norm of u_1.

Obviously, this procedure can be extended to general $0 < \alpha \leq 2$ and we can use more accurate estimators than median.

2.2.2 Comparing Data Streams Using Hamming Norms

[49, 50] proposed approximating the Hamming norms of data streams using stable random projections with very small α, because the l_α norm approximates the Hamming norm well if $\alpha \to 0+$. The Hamming norm gives the number of non-zero items (distinct terms) present in a single stream; and it is also an important measure of (dis)similarity when applied to a pair of streams[49, 50]. Note that for static data, one could approximate the Hamming norms directly by applying 2-stable (i.e., normal) random projections on the binary-quantized (0/1) data. [49, 50] considered the dynamic setting in that the data may be subject to frequent additions/subtractions.

2.3 The Choice of the Norm (α)

The choice of α depends on the applications and datasets at hand.

When using the l_2 norm, i.e., *normal random projections*, it is implicitly assumed that the original data are "normal-like" or at least the second moments of the data are meaningful. Unfortunately, this is not always the case in practice.

Real-world large-scale datasets (especially Internet data) are ubiquitously "heavy-tailed" and follow the "power-law" [111, 53, 66, 142]. It is very often the case that one can not use the l_2 norm directly without carefully *term-weighting* the original data (e.g., heuristically taking logarithm or square root, or tf-idf)[135, 169, 150, 63, 86, 131]. The *term-weighting* procedure is often far more important than fine-tuning the machine learning parameters[112, 147, 107].

Therefore, there are practical reasons to consider the l_α norm other than l_2. Instead of weighting the data, an alternative scheme is to choose an appropriate norm such as the l_1 norm.

It is well-known that the l_1 norm is far more robust than l_2 against "outliers." In Statistics, it is a common practice to replace the l_2 norm minimization with the l_1 norm minimization (or using a combination of both l_1 and l_2 norms), for example, the Least Absolute deviation (LAD) Boost [78] and Laplacian radial basis kernel[42, 71]. Recently, it becomes popular to use the l_1 norm for variable (feature) selection; success stories include LASSO[161], LARS[65] and 1-norm SVM[170].

Other norms are also possible. In data streams, as α increases, the l_α distance attributes more significance to a large individual component; and therefore varying α provides a tunable mechanism[75]. This argument applies directly also in the machine learning content. As a concrete example, [42] proposed a family of non-Gaussian radial basis kernels for SVM in the form of $K(x, y) = \exp\left(-\rho \sum_i |x_i - y_i|^\alpha\right)$, for data points x and y in high dimensions. [42] showed that $\alpha = 0.5$ in some cases gave better results in histogram-based image classifications.

The l_α norm with $\alpha < 1$ (which is a non-metric norm)is now well-understood to be a natural measure of sparsity[61, 62]. For example, [49, 50] approximated the Hamming norm with the l_α norm using very small α. [51] adopted the similar idea to approximate the max-dominance norm in data streams using very small α.

Chapter 3

Normal Random Projections

For dimension reduction in the l_2 norm, the method of *normal random projections* multiplies the original data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ by a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ ($k \ll D$) with i.i.d. $N(0,1)$ entries, resulting in a projected data matrix $\mathbf{B} \in \mathbb{R}^{n \times k}$. The analysis for *normal random projections* is quite simple; for example, it is straightforward to derive one version of the Johnson-Lindenstrauss (JL) Lemma[103] for l_2.

We will first introduce some basic properties of *normal random projections*, and then focus on the idea of using marginal information to improve estimates. Margins (i.e., the l_2 norms of individual rows in $\mathbf{A}$) are usually already available (e.g., via data normalization). But even when they are not available, computing the l_2 norms for all rows of $\mathbf{A}$ requires only one-pass of the data at the cost of $O(nD)$, which is negligible[1] since conducting random projections $\mathbf{A} \times \mathbf{R}$ already costs $O(nDk)$.

In this chapter, we will follow the "convention" in the literature on random projections (e.g., [166]) by denoting $\mathbf{B} = \frac{1}{\sqrt{k}}\mathbf{AR}$.

The main results in this chapter appeared in a conference proceeding[124].

[1]The situation is slightly different in dynamic data streams. In streams, we are often more interested in the summary statistics of an individual stream, rather than the differences between two streams. In other words, computing the "marginal" l_2 norm is sometimes the major goal. Due to the dynamic nature of streams (e.g., frequent updating), computing the margins can be prohibitive.

3.1 Basic Properties

We assume a data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$, a projection matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ generated with i.i.d. entries in $N(0,1)$. Let $\mathbf{B} = \frac{1}{\sqrt{k}}\mathbf{AR}$. 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}. For convenience, 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 $\mathbf{B}$. We denote

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

It is easy to see that $\|v_1 - v_2\|^2$, the sample l_2 distance, and $v_1^{\mathrm{T}} v_2$, the sample inner product, are unbiased estimators of d and a, respectively. Lemma 1 characterizes the variances and the characteristic function of $v_1^{\mathrm{T}} v_2$. See the proof in Section 3.5.1.

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*

$$E\left(\|v_1 - v_2\|^2\right) = d, \quad Var\left(\|v_1 - v_2\|^2\right) = \frac{2}{k} d^2. \tag{3.1}$$

$$E\left(v_1^T v_2\right) = a, \quad Var\left(v_1^T v_2\right) = \frac{1}{k}\left(m_1 m_2 + a^2\right), \tag{3.2}$$

The third central moment of $v_1^T v_2$ is

$$E\left(v_1^T v_2 - a\right)^3 = \frac{2a}{k^2}\left(3 m_1 m_2 + a^2\right), \tag{3.3}$$

and the moment generating function of $v_1^T v_2$ is

$$E\left(\exp(v_1^T v_2 t)\right) = \left(1 - \frac{2}{k} a t - \frac{1}{k^2}\left(m_1 m_2 - a^2\right) t^2\right)^{-\frac{k}{2}}, \tag{3.4}$$

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

Therefore, straightforward unbiased estimators of the l_2 distance d and the inner

product a would be

$$\hat{d}_{MF} = \|v_1 - v_2\|^2, \qquad \mathrm{Var}\left(\hat{d}_{MF}\right) = \frac{d^2}{k}, \tag{3.5}$$

$$\hat{a}_{MF} = v_1^{\mathrm{T}} v_2, \qquad \mathrm{Var}\left(\hat{a}_{MF}\right) = \frac{1}{k}\left(m_1 m_2 + a^2\right), \tag{3.6}$$

where the subscript "MF" stands for "margin-free," indicating that the estimators do not use any marginal information $m_1 = \|u_1\|^2$ and $m_2 = \|u_2\|^2$.

Note that $k\hat{d}_{MF}/d$ follows a chi-square distribution with k degrees of freedom, χ^2_k. Therefore, it is easy to prove the following tail bounds in Lemma 2.

Lemma 2

$$\mathbf{Pr}(\hat{d}_{MF} - d > \epsilon d) \leq \exp\left(-\frac{k}{2}\left(\epsilon - \log(1+\epsilon)\right)\right), \quad \epsilon > 0 \tag{3.7}$$

$$\mathbf{Pr}(\hat{d}_{MF} - d < -\epsilon d) \leq \exp\left(-\frac{k}{2}\left(-\epsilon - \log(1-\epsilon)\right)\right), \quad 0 < \epsilon < 1 \tag{3.8}$$

Proof: *Because $k\hat{d}_{MF}/d \sim \chi^2_k$, by the Chernoff inequality[46], for any $t > 0$,*

$$\begin{aligned}
&\mathbf{Pr}(\hat{d}_{MF} - d > \epsilon d) = \mathbf{Pr}(k\hat{d}_{MF}/d > k(1+\epsilon)) \\
\leq &\frac{E\left(\exp(k\hat{d}_{MF}/dt)\right)}{\exp((1+\epsilon)kt)} = \exp\left(-\frac{k}{2}\left(\log(1-2t) + 2(1+\epsilon)t\right)\right),
\end{aligned}$$

which is minimized at $t = t_{NR} = \frac{\epsilon}{2(1+\epsilon)}$; and hence for any $\epsilon > 0$

$$\mathbf{Pr}(\hat{d}_{MF} - d > \epsilon d) \leq \exp\left(-\frac{k}{2}\left(\epsilon - \log(1+\epsilon)\right)\right) \tag{3.9}$$

We can similarly prove the other tail bound for $\mathbf{Pr}(\hat{d}_{MF} - d < -\epsilon d)$.

For convenience, it is customary to write the tail bounds in Lemma 2 in a symmetric form, i.e., $\mathbf{Pr}(|\hat{d}_{MF} - d| > \epsilon d)$. Simple inequalities on $\log(1+\epsilon)$ and $\log(1-\epsilon)$

yield

$$\mathbf{Pr}\left(\left|\hat{d}_{MF} - d\right| \geq \epsilon d\right) \leq 2\exp\left(-\frac{k}{4}\epsilon^2 + \frac{k}{6}\epsilon^3\right), \quad 0 < \epsilon < 1. \tag{3.10}$$

Since $\mathbf{A} \in \mathbb{R}^{n \times D}$ has n rows, i.e., $\frac{n(n-1)}{2}$ pairs, we need to bound the tail probabilities simultaneously for all pairs. By the Bonferroni union bound, it suffices if

$$\frac{n^2}{2}\mathbf{Pr}\left(\left|\hat{d}_{MF} - d\right| \geq \epsilon d\right) \leq \delta. \tag{3.11}$$

In other words, it suffices if

$$\frac{n^2}{2}2\exp\left(-\frac{k}{4}\epsilon^2 + \frac{k}{6}\epsilon^3\right) \leq \delta \Longrightarrow k \geq \frac{2\log n - \log\delta}{\epsilon^2/4 - \epsilon^3/6}. \tag{3.12}$$

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

Lemma 3 *If $k \geq \frac{2\log n - \log\delta}{\epsilon^2/4 - \epsilon^3/6}$, then with probability at least $1 - \delta$, the l_2 distance between any pair of data points (among n data points) can be approximated within a $1 \pm \epsilon$ factor of the truth, using the l_2 distance of the projected data after normal random projections. $0 < \delta < 1$, $0 < \epsilon < 1$.*

3.2 Improving Normal Random Projections Using Marginal Information

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 (than $\hat{d}_{MF}$ and $\hat{a}_{MF}$). For example,

$$\hat{a}_{SM} = \frac{1}{2}\left(m_1 + m_2 - \|v_1 - v_2\|^2\right), \qquad \mathrm{Var}\left(\hat{a}_{SM}\right) = \frac{1}{2k}\left(m_1 + m_2 - 2a\right)^2, \tag{3.13}$$

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

$\mathrm{Var}\left(\hat{a}_{MF}\right) = \frac{1}{k}(m_1 m_2)$. It is easy to show that

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

The estimator based on maximum likelihood can uniformly improve $\hat{a}_{MF}$, at least asymptotically (as $k \to \infty$). The following Lemma is proved in Section 3.5.2.

Lemma 4 *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:*

$$a^3 - a^2\left(v_1^T v_2\right) + a\left(-m_1 m_2 + m_1\|v_2\|^2 + m_2\|v_1\|^2\right) - m_1 m_2 v_1^T v_2 = 0. \tag{3.14}$$

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

$$Var(\hat{a}_{MLE}) = \frac{1}{k}\frac{\left(m_1 m_2 - a^2\right)^2}{m_1 m_2 + a^2} + O\left(\frac{1}{k^2}\right). \tag{3.15}$$

And

$$\frac{1}{k}\frac{\left(m_1 m_2 - a^2\right)^2}{m_1 m_2 + a^2} \leq \min\left(Var\left(\hat{a}_{MF}\right), Var\left(\hat{a}_{SM}\right)\right). \tag{3.16}$$

We notice that a special case of Lemma 4 in which $m_1 = m_2 = 1$ appeared in various places, e.g., [14, 156] [21, Example 9.2.38].

Figure 3.1 verifies the inequality in (3.15) by plotting asymptotic $\frac{\mathrm{Var}(\hat{a}_{MLE})}{\mathrm{Var}(\hat{a}_{MF})}$ and $\frac{\mathrm{Var}(\hat{a}_{MLE})}{\mathrm{Var}(\hat{a}_{SM})}$. The improvement can be substantial. For example, $\frac{\mathrm{Var}(\hat{a}_{MLE})}{\mathrm{Var}(\hat{a}_{MF})} = 0.2$ implies that in order to achieve the same mean square accuracy, the MLE estimator needs only 20% of the samples required by the margin-free (MF) estimator.

3.2.1 A Numerical Example

Figure 3.2 presents some numerical results, using two words "THIS" and "HAVE," from a chunk of MSN Web crawl data ($D = 2^{16}$). That is, $u_{1,j}$ ($u_{2,j}$) is the number of occurrences of word "THIS" (word "HAVE") in the jth document (Web page), $j = 1$ to D. The numerical results are consistent with our theoretical analysis.

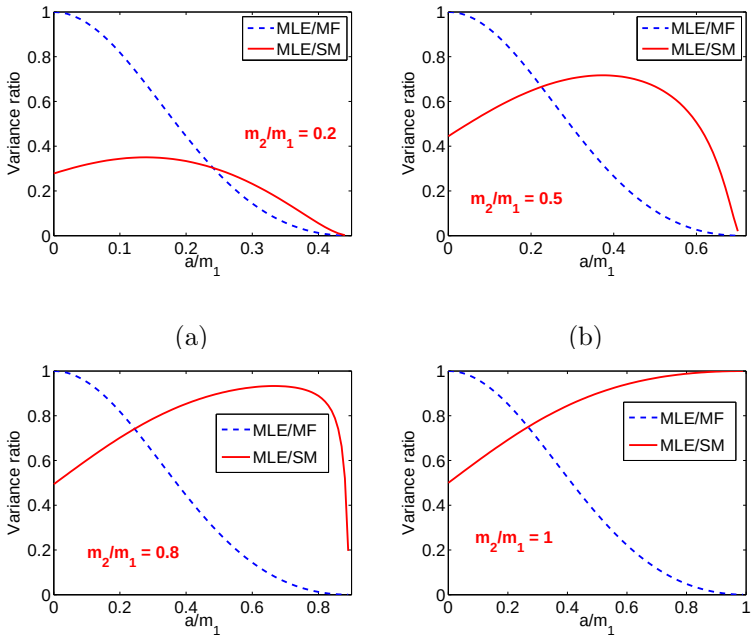
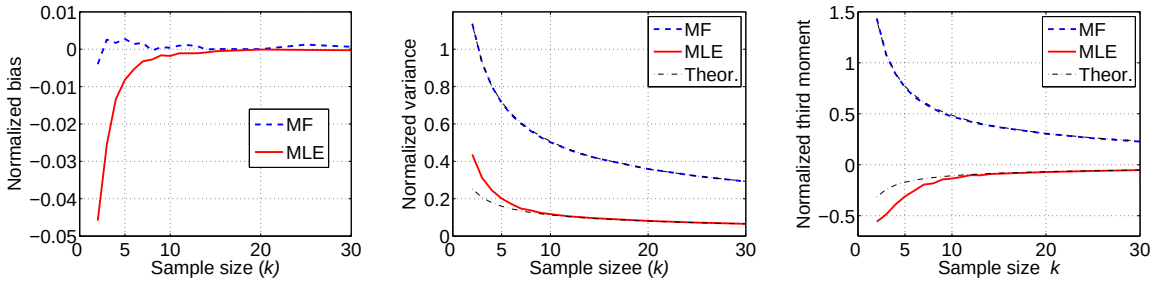


Figure 3.1: The asymptotic variance ratios, $\frac{\mathrm{Var}(\hat{a}_{MLE})}{\mathrm{Var}(\hat{a}_{MF})}$ and $\frac{\mathrm{Var}(\hat{a}_{MLE})}{\mathrm{Var}(\hat{a}_{SM})}$ verify that the MLE has smaller variance. $\mathrm{Var}\left(\hat{a}_{MLE}\right)$, $\mathrm{Var}\left(\hat{a}_{MF}\right)$, and $\mathrm{Var}\left(\hat{a}_{SM}\right)$ are given in (3.15), (3.6), and (3.13), respectively. We consider $m_2 = 0.2m_1$, $m_2 = 0.5m_1$, $m_2 = 0.8m_1$, and $m_2 = 1.0m_1$.

3.2.2 Higher-Order Analysis of the MLE

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 [23] "Bartlett correction." In Lemma 5 (proved in Section 3.5.3), we show that the asymptotic bias of $\hat{a}_{MLE}$ is only $O(k^{-2})$ and therefore there is no need for bias correction. Lemma 5 also derives the asymptotic third moment of $\hat{a}_{MLE}$ as well as a more accurate variance 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 may be useful for small k.

Lemma 5 *The bias, third moment, and the variance with $O(k^{-3})$ correction for the*



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

Figure 3.2: Estimation of the inner products between "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 well. (C): Simulations for the margin-free estimator match the theoretical results.

maximum likelihood estimator, $\hat{a}_{MLE}$, derived in Lemma 4, are given by

$$E(\hat{a}_{MLE} - a) = O(k^{-2}), \tag{3.17}$$

$$E\left((\hat{a}_{MLE} - a)^3\right) = \frac{-2a(3m_1m_2 + a^2)(m_1m_2 - a^2)^3}{k^2(m_1m_2 + a^2)^3} + O(k^{-3}), \tag{3.18}$$

$$Var(\hat{a}_{MLE}) = \frac{1}{k}\frac{(m_1m_2 - a^2)^2}{m_1m_2 + a^2} + \frac{1}{k^2}\frac{4(m_1m_2 - a^2)^4}{(m_1m_2 + a^2)^4}m_1m_2 + O(k^{-3}). \tag{3.19}$$

Eq. (3.19) 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 (3.19) instead of (3.15) for $\text{Var}(\hat{a}_{MLE})$. As we will show next, for very small k, there is also a multiple root problem in solving the cubic MLE equation (3.14).

3.2.3 The Problem of Multiple (Real) Roots

The cubic equation (3.14) always has three roots, two of which may be complex (and hence out of the scope of the consideration). When k is small, there is a certain probability of multiple real roots, as shown in Lemma 6 (proved in Section 3.5.4).

Lemma 6 *The cubic MLE equation (5.50) admits multiple real roots with probability*

$$\mathbf{Pr}\left(\textit{multiple real roots}\right) = \mathbf{Pr}\left(P^2(11 - Q^2/4 - 4Q + P^2) + (Q-1)^3 \leq 0\right), \quad (3.20)$$

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*

$$\mathbf{Pr}\left(\textit{multiple real roots}\right) \leq e^{-0.0085k} + e^{-0.0966k}. \quad (3.21)$$

When[2] $a = m_1 = m_2$, *this probability can be (sharply) bounded by*

$$\mathbf{Pr}\left(\textit{multiple real roots} \mid a = m_1 = m_2\right) \leq e^{-1.5328k} + e^{-0.4672k}. \quad (3.22)$$

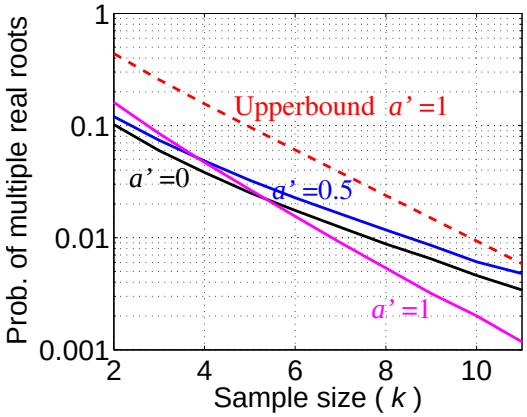


Figure 3.3: Simulations show that $\mathbf{Pr}$(multiple real roots) decreases exponentially fast (notice the log scale in the vertical axis). Here $a' = \frac{a}{\sqrt{m_1 m_2}}$. The curve for the upper bound is given by (3.22).

The bound (3.21), though in an exponential form, is crude. However, the probability of admitting multiple real roots in (3.20) can be easily simulated. Figure 3.3 shows that this probability drops quickly to $< 1\%$ when $k \geq 8$.

To the best of our knowledge, there is no consensus on how to deal with multiple real roots [22, 156]; for example, a theoretically consistent solution is not always real;

[2]For example, two documents are identical (duplicates).

and a global maximum of the likelihood is not always consistent, e.g., [106].[3]

On the other hand, the multiple-root problem is not really a big issue in our case because there are at most three real roots and we can always pick the one that maximizes the likelihood or simply use the margin-free estimate. Lemma 6 has shown that $\mathbf{Pr}\left(\text{multiple real roots}\right) \rightarrow 0$ exponentially fast, as $k \rightarrow \infty$. For large-scale applications we are interested in, the sample size k should be $\gg 10$.

3.3 Sign Random Projections

We give an introduction to *sign random projections*, which store only one bit per projection for each data point. We will show that when the data are uncorrelated, the variance of *sign random projections* is only $\frac{\pi^2}{4}$ of the variance of regular random projections, which store real numbers (32 or 64 bits). In highly correlated data, however, *sign random projections* can be 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[43]

$$\mathbf{Pr}\left(\text{sign}(v_{1,j}) = \text{sign}(v_{2,j})\right) = 1 - \frac{\theta}{\pi}, \qquad j = 1, 2, ..., k, \tag{3.23}$$

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 is

$$\text{Var}\left(\hat{\theta}\right) = \frac{\pi^2}{k}\left(1 - \frac{\theta}{\pi}\right)\left(\frac{\theta}{\pi}\right) = \frac{\theta(\pi - \theta)}{k}. \tag{3.24}$$

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

$$\hat{a}_{Sign} = \cos(\hat{\theta})\sqrt{m_1 m_2}, \tag{3.25}$$

which is biased. The severity of bias depends on the nonlinearity of $\cos(\theta)$. By the

[3]The regularity conditions in [167] to ensure that the global maximum likelihood estimate is consistent are difficult to check for models involving multiple roots.[156]

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

$$\mathrm{Var}\left(\hat{a}_{Sign}\right) = \mathrm{Var}(\hat{\theta})\sin^2(\theta)m_1m_2 = \frac{\theta(\pi-\theta)}{k}\sin^2(\theta)m_1m_2 + O\left(\frac{1}{k^2}\right), \qquad (3.26)$$

provided $\sin(\theta)$ is nonzero, which is violated when $\theta = 0$ or π.

Regular random projections store real numbers (32 or 64 bits). At the same number of projections (i.e., 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 ($\mathrm{Var}\left(\hat{a}_{Sign}\right)$) of *sign random projections* with the variance of regular random projections considering the margins (i.e., $\mathrm{Var}\left(\hat{a}_{MLE}\right)$) by

$$V_{Sign} = \frac{\mathrm{Var}\left(\hat{a}_{Sign}\right)}{\mathrm{Var}\left(\hat{a}_{MLE}\right)} = \frac{\theta(\pi-\theta)\sin^2(\theta)m_1m_2}{\frac{(m_1m_2-a^2)^2}{m_1m_2+a^2}} = \frac{\theta(\pi-\theta)(1+\cos^2(\theta))}{\sin^2(\theta)}, \qquad (3.27)$$

which is symmetric about $\theta = \frac{\pi}{2}$. It is easy to check (also see Figure 3.4) that V_{Sign} is monotonically decreasing in $(0, \frac{\pi}{2}]$ with minimum $\frac{\pi^2}{4} \approx 2.47$, attained at $\theta = \frac{\pi}{2}$.

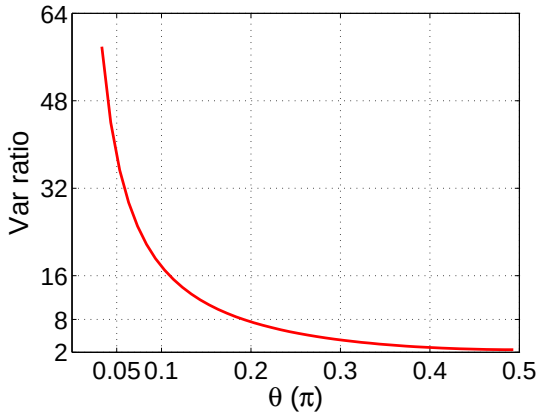


Figure 3.4: The ratio of variance $V_{Sign} = \frac{\mathrm{Var}\left(\hat{a}_{Sign}\right)}{\mathrm{Var}(\hat{a}_{MLE})}$ decreases monotonically in $(0, \frac{\pi}{2}]$, with minimum $= \frac{\pi^2}{4}$ attained at $\theta = \frac{\pi}{2}$.

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.

3.4 Summary

There have been a lot of applications of *normal random projections* for dimension reduction in the l_2 norm, highlighted by the Johnson-Lindenstrauss (JL) Lemma.

In this chapter, we show how to improve *normal random projections* by taking advantage of the margins for better estimates from the solution of a cubic MLE equation. The marginal information is particularly beneficial when the data points are highly correlated. At small samples, the MLE exhibits problems of multiple real roots and possible inflation of the variance especially in the uncorrelated situation.

When the data points are nearly uncorrelated, it is cost-effective to store only the signs instead of the projected data. When the data points are highly correlated, however, storing only the signs is not recommended.

3.5 Proofs

3.5.1 Proof of Lemma 1

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

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 $a = u_1^{\mathrm{T}} u_2$)

$$\left[\begin{array}{c} v_{1,j} \\ v_{2,j} \end{array} \right] \sim N \left(\left[\begin{array}{c} 0 \\ 0 \end{array} \right], \Sigma = \frac{1}{k} \left[\begin{array}{cc} \|u_1\|^2 & u_1^{\mathrm{T}} u_2 \\ u_1^{\mathrm{T}} u_2 & \|u_2\|^2 \end{array} \right] = \frac{1}{k} \left[\begin{array}{cc} m_1 & a \\ a & m_2 \end{array} \right] \right). \tag{3.28}$$

It is easier to work with the conditional probability:

$$v_{1,j} | v_{2,j} \sim N \left(\frac{a}{m_2} v_{2,j}, \frac{m_1 m_2 - a^2}{k m_2} \right),$$

from which we obtain

$$\begin{aligned}
&\mathrm{E}\left(v_{1,j}v_{2,j}\right)^2 = \mathrm{E}\left(\mathrm{E}\left(v_{1,j}^2 v_{2,j}^2 | v_{2,j}\right)\right) = \mathrm{E}\left(v_{2,j}^2\left(\frac{m_1m_2-a^2}{km_2} + \left(\frac{a}{m_2}v_{2,j}\right)^2\right)\right) \\
=&\frac{m_2}{k}\frac{m_1m_2-a^2}{km_2} + \frac{3m_2^2}{k^2}\frac{a^2}{m_2^2} = \frac{1}{k^2}\left(m_1m_2+2a^2\right). \\
&\Longrightarrow \mathrm{Var}\left(v_{1,j}v_{2,j}\right) = \frac{1}{k^2}\left(m_1m_2+a^2\right), \quad \mathrm{Var}\left(v_1^{\mathrm{T}}v_2\right) = \frac{1}{k}\left(m_1m_2+a^2\right).
\end{aligned}$$

The third moment can be proved similarly, or using the moment generating function:

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

Here, we use the fact that $\frac{v_{2,j}^2}{m_2/k} \sim \chi_1^2$, a chi-square 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

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

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

3.5.2 Proof of Lemma 4

From Section 3.5.1, we can write down the joint likelihood function for $\{v_{1,j}, v_{2,j}\}_{j=1}^k$:

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

where (assuming $m_1 m_2 \neq a$ to avoid triviality)

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

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

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

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

$$a^3 - a^2 \left(v_1^{\mathrm{T}} v_2\right) + a \left(-m_1 m_2 + m_1 \|v_2\|^2 + m_2 \|v_1\|^2\right) - m_1 m_2 v_1^{\mathrm{T}} v_2 = 0. \tag{3.29}$$

The large sample theory [110, Theorem 6.3.10] says that $\hat{a}_{MLE}$ is asymptotically unbiased and converges weakly to $N\left(a, \frac{1}{\mathrm{I}(a)}\right)$, where $\mathrm{I}(a)$, the expected Fisher Information, is $\mathrm{I}(a) = -\mathrm{E}\left(l''(a)\right)$. Some algebra will show that

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

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

$$\frac{1}{k} \frac{\left(m_1 m_2 - a^2\right)^2}{m_1 m_2 + a^2} \leq \min\left(\mathrm{Var}\left(\hat{a}_{MF}\right), \mathrm{Var}\left(\hat{a}_{SM}\right)\right),$$

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

3.5.3 Proof of Lemma 5

We analyze the higher-order properties of $\hat{a}_{MLE}$ using stochastic Taylor expansions. We use some formulations appeared in[23, 154, 73]. The bias

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

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

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

The third central moment

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

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

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

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

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

where we let $S = m_1m_2 - a^2$ and $T = \|v_1\|^2m_2 + \|v_2\|^2m_1 - 2v_1^{\mathrm{T}}v_2a$. Expanding $\left(l''(a)\right)^2$ generates terms involving T, T^2, $Tv_1^{\mathrm{T}}v_2$. Rewrite

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} &\mathrm{E}\left((l''(a))^2\right) = \frac{k^2}{S^4}\left(\left(m_1m_2 + a^2\right)^2 + \frac{4}{k}\left(m_1^2m_2^2 + a^4 + 6a^2m_1m_2\right)\right), \\ &V_2^c = \frac{4}{k^2}\frac{(m_1m_2 - a^2)^4}{(m_1m_2 + a^2)^4}m_1m_2. \end{aligned}$$

3.5.4 Proof of Lemma 6

The cubic MLE equation may admit multiple roots. By the Cardano condition:

$$\mathbf{Pr}\left(\text{multiple real roots}\right) = \mathbf{Pr}\left(P^2(11 - Q^2/4 - 4Q + P^2) + (Q-1)^3 \leq 0\right), \quad (3.30)$$

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 $\mathbf{Pr}(A + B \leq 0) \leq \mathbf{Pr}(A \leq 0) + \mathbf{Pr}(B \leq 0)$, i.e.,

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

We will soon prove the following moment generating function

$$\mathrm{E}\left(\exp(Qt)\right) = \left(1 - \frac{4t}{k} + \frac{4t^2}{k^2}\left(\frac{m_1 m_2 - a^2}{m_1 m_2}\right)\right)^{-\frac{k}{2}}, \quad (3.31)$$

which enables us to prove the following upper bounds:

$$\begin{aligned} &\mathbf{Pr}\left(Q - 1 \leq 0\right) \leq e^{-0.0966k}, \quad \mathbf{Pr}\left(11 - Q^2/4 - 4Q \leq 0\right) \leq e^{-0.0085k}, \\ &\mathbf{Pr}\left(\text{multiple real roots}\right) \leq e^{-0.0966k} + e^{-0.0085k}, \end{aligned} \quad (3.32)$$

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

The upper bound (3.32) 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., $Q = 2P = \|v_1\|^2/m_1$, $kP = \frac{k\|v_1\|^2}{m_2} \sim \chi^2_k$, and a (sharp) upper bound:

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

To complete the proof of Lemma 6, 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

$$\frac{km_2}{m_1m_2-a^2}v_{1,j}^2|v_{2,j} \sim \chi_{1,\lambda}^2, \quad \text{where} \;\; \lambda = \frac{ka^2}{m_2(m_1m_2-a^2)}v_{2,j}^2.$$

$\chi_{1,\lambda}^2$ denotes a non-central chi-square 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

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

we can obtain the moment generating function in (3.31) after some algebra.

Chapter 4

Sub-Gaussian and Very Sparse Random Projections

In Chapter 3, we discuss *normal random projections*, in which the projection matrix $\mathbf{R}$ is sampled from i.i.d. $N(0,1)$. This specific choice of $\mathbf{R}$, is merely for convenience of theoretical analysis. In fact, we can sample $\mathbf{R}$ from any zero-mean bounded-variance distributions for dimension reduction in the l_2 norm.

It is both theoretically appealing and computationally convenient to sample $\mathbf{R}$ from a sub-Gaussian distribution. For example, the sub-Gaussian tail bounds easily lead to a version of the Johnson-Lindenstrauss (JL) Lemma.

We will focus on a typical choice of sub-Gaussian distribution, by using $\mathbf{R}$ of entries in $\{-1, 0, 1\}$ with probabilities $\{\frac{1}{2s}, 1-\frac{1}{s}, \frac{1}{2s}\}$, where $s \geq 1$. This scheme simplifies the sampling procedure and speeds up the computations. In fact, when $s < 3$, we can achieve strictly smaller variances than using *normal random projections*.

By assuming reasonable regularity conditions, for example, the original data have bounded third moments, we can let $s \gg 3$ (even $s = \sqrt{D}$) to achieve an s-fold speedup; and hence we name this scheme *very sparse random projections*.

Part of the results in this chapter appeared in a conference proceeding[125].

4.1 Sub-Gaussian Random Projections

As in Chapter 3, we consider a data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$. We generate a random projection matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ and multiply it with $\mathbf{A}$ to obtain a projected data matrix $\mathbf{B} = \frac{1}{\sqrt{k}}\mathbf{AR} \in \mathbb{R}^{n \times k}$. We again 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 $\mathbf{B}$, and we also denote

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

We sample $\mathbf{R}$ i.i.d. from a particularly useful *sub-Gaussian* distribution ($s \geq 1$)

$$r_{ij} = \sqrt{s} \times \begin{cases} 1 & \text{with prob. } \frac{1}{2s} \\ 0 & \text{with prob. } 1 - \frac{1}{s} \\ -1 & \text{with prob. } \frac{1}{2s} \end{cases}, \tag{4.1}$$

- Sampling from (4.1) is simpler than sampling from $N(0, 1)$.
- One can achieve an s-fold speedup in the matrix multiplication $\mathbf{A} \times \mathbf{R}$ because only $\frac{1}{s}$ of the data need to be processed.
- No floating point arithmetic is needed and all computation amounts to highly optimized database aggregation operations.
- When $s < 3$, one can achieve more accurate (smaller variance) estimates.
- The cost for storing $\mathbf{R}$ is reduced from $O(Dk)$ to $O\left(Dk/s\right)$.

[2, 3] showed that when $s = 1$ and $s = 3$, one can achieve the same JL bound as *normal random projections*. We will first review properties of *sub-Gaussian distributions*, which make it convenient for tail bound analysis. In fact, sub-Gaussian analysis will show that one can let s be slightly larger than 3, even in the worst case.

4.1.1 Sub-Gaussian Distributions

We give a brief introduction to sub-Gaussian distributions. The theory of *sub-Gaussian* distributions started in 1960s. See [40] for details and references therein.

A random variable x is *sub-Gaussian* if there exists a constant $g > 0$ such that

$$\mathrm{E}\left(\exp(xt)\right) \leq \exp\left(\frac{g^2t^2}{2}\right), \ \forall t \in \mathbb{R} \tag{4.2}$$

We can find the optimal value of g^2, denoted by $T^2(x)$, from

$$T^2(x) = \sup_{t\neq 0} \frac{2\log \mathrm{E}\left(\exp(xt)\right)}{t^2}. \tag{4.3}$$

Note that $T^2(x)$ is just a notation for the optimal sub-Gaussian constant of a random variable x (not a particular realization of x).

Some basic properties of sub-Gaussian distributions:

- If x is sub-Gaussian, then $\mathrm{E}(x) = 0$ and $\mathrm{E}(x^2) \leq T^2(x)$. $T^2(cx) = c^2T^2(x)$, for any constant c. And

$$\mathbf{Pr}\left(|x| > t\right) \leq 2\exp\left(-\frac{t^2}{2T^2(x)}\right). \tag{4.4}$$

- If x_1, x_2, ..., x_D are independent sub-Gaussian, then $\sum_{i=1}^{D} x_i$ is sub-Gaussian

$$T^2\left(\sum_{i=1}^{D} x_i\right) \leq \sum_{i=1}^{D} T^2(x_i). \tag{4.5}$$

- If x is sub-Gaussian, then for all $t \in [0, 1)$,

$$\mathrm{E}\left(\exp\left(\frac{x^2t}{2T^2(x)}\right)\right) \leq (1-t)^{-\frac{1}{2}}. \tag{4.6}$$

[2, 3] also discovered (4.6) for the special distribution (4.1).

A *sub-Gaussian* random variable x is *strictly sub-Gaussian* if $\mathrm{E}(x^2) = T^2(x)$.

- If x is strictly sub-Gaussian, then $\mathrm{E}(x^3) = 0$ and the kurtosis is non-positive, i.e., $\frac{\mathrm{E}(x^4)}{\mathrm{E}^2(x^2)} - 3 \leq 0$.

- If x_1, x_2, ..., x_D are independent strictly sub-Gaussian, then $\sum_{i=1}^{D} x_i$ is also strictly sub-Gaussian,

$$T^2\left(\sum_{i=1}^{D} x_i\right) = \sum_{i=1}^{D} T^2(x_i) = \sum_{i=1}^{D} \mathrm{E}\left(x_i^2\right). \tag{4.7}$$

4.1.2 Analyzing the Sparse Projection Distribution

The sparse projection distribution defined in (4.1) is sub-Gaussian for any $s \geq 1$, and strictly sub-Gaussian for $1 \leq s \leq 3$. We show how to obtain the sub-Gaussian parameter.

Let $x = r_{ij}$ be the distribution defined in (4.1). Then for all $t \in \mathbb{R}$,

$$\begin{aligned}
\mathrm{E}\left(\exp(xt)\right) &= 1 - \frac{1}{s} + \frac{1}{2s}\left(\exp(\sqrt{s}t) + \exp(-\sqrt{s}t)\right) \\
&= 1 + \frac{t^2}{2} + \frac{t^4}{4!}s + \frac{t^6}{6!}s^2 + ... \\
&= 1 + \frac{t^2}{2} + \left(\frac{t^2}{2}\right)^2 \frac{1}{2!}\frac{s}{3} + \left(\frac{t^2}{2}\right)^3 \frac{1}{3!}\frac{s^2 2^3 3!}{6!} + ...
\end{aligned}$$

When $1 \leq s \leq 3$, $\mathrm{E}\left(\exp(xt)\right) \leq \exp\left(\frac{t^2}{2}\right) = 1 + \frac{t^2}{2} + \left(\frac{t^2}{2}\right)^2 \frac{1}{2!} + \left(\frac{t^2}{2}\right)^3 \frac{1}{3!}\frac{s^2 2^3 3!}{6!} + ...$ Also, $T^2(x) = 1 = \mathrm{E}(x^2)$. Therefore, r_{ij} is strictly sub-Gaussian when $1 \leq s \leq 3$.

For convenience, we use the following notation

$$\tau^2(s,t) = \frac{2\log \mathrm{E}\left(\exp(r_{ij}t)\right)}{t^2}, \qquad \tau^2(s) = \sup_{t \neq 0} \tau^2(s,t) = T^2(r_{ij}) \tag{4.8}$$

When $s > 3$, the optimal value, $\tau^2(s)$, can be computed from

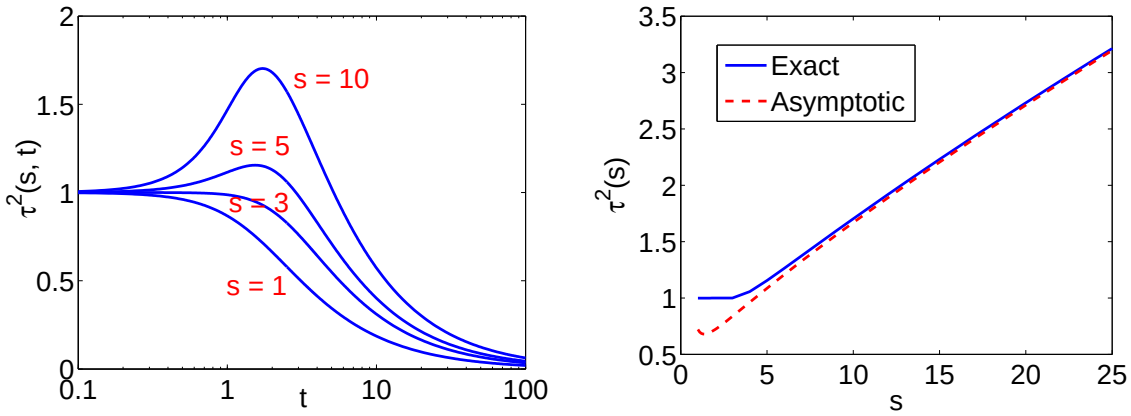
$$\begin{aligned}
\tau^2(s) &= \sup_{t \neq 0} \tau^2(s,t) = \sup_{t \neq 0} \frac{2\log \mathrm{E}\left(\exp(xt)\right)}{t^2} \\
&= \max_{t>0} \frac{2}{t^2} \log\left(1 - \frac{1}{s} + \frac{1}{2s}\left(\exp(\sqrt{s}t) + \exp(-\sqrt{s}t)\right)\right).
\end{aligned} \tag{4.9}$$

Although no simple closed-form expressions are available, the solution can be easily

obtained numerically. Asymptotically (at large s),

$$\tau^2(s) \sim \max_{t>0} \frac{2}{t^2}\left(\sqrt{s}t - \log(2s)\right) = \frac{s}{2\log(2s)}. \tag{4.10}$$

Figure 4.1(a) plots $\tau^2(s,t)$ for $s = 1$, 3, 5, and 10. Figure 4.1(b) plots $\tau^2(s)$ for $s \leq 25$, indicating that the approximation (4.10) is pretty accurate when $s > 5$.



(a) $\tau^2(s,t)$ (b) $\tau^2(s)$

Figure 4.1: $\tau^2(s,t) = \frac{2\log \mathrm{E}(\exp(xt))}{t^2}$, where $x = r_{ij}$ is defined in (4.1). $\tau^2(s) = \max\ \tau^2(s,t)$ for $t > 0$. Panel (a) plots $\tau^2(s,t)$ for some range of t and $s = 1, 3, 5, 10$. The peak on each curve corresponds to $\tau^2(s)$. Panel (b) plots $\tau^2(s)$ for $1 \leq s \leq 25$ along with its asymptote $\tau^2(s) \sim \frac{s}{2\log(2s)}$.

4.1.3 Tail Bounds for Sub-Gaussian Random Projections

In sub-Gaussian random projections, we sample r_{ij} from (4.1) and we still use the sample l_2 distance and the sample inner product to approximate the original l_2 distance d and inner product a, just like in *normal random projections*, i.e.,

$$\hat{d} = \|v_1 - v_2\|^2, \qquad \hat{a} = v_1^{\mathrm{T}} v_2, \tag{4.11}$$

where we ignore the subscript $_{MF}$, since there will be no ambiguity in this chapter.

We will defer the analysis of the variances of $\hat{d}$ and $\hat{a}$ to Section 4.2.1. Here, we analyze the tail bound, and in particular, only the right tail $\mathbf{Pr}\left(\hat{d} - d > \epsilon d\right)$ because this makes it convenient for comparisons; plus in reality we often care about the right tail bound more since the left tail bound $\mathbf{Pr}\left(\hat{d} - d < -\epsilon d\right)$ is often much smaller.

Lemma 7 *Let r_{ij} be any sub-Gaussian projection distribution with unit variance and $\tau^2 = T^2(r_{ij})$, then for any $\tau^2 \leq 1 + \epsilon$, $\epsilon > 0$, we have*

$$\mathbf{Pr}\left(\hat{d} \leq (1+\epsilon)d\right) \leq \exp\left(-\frac{k}{2}\left(\log\frac{\tau^2}{1+\epsilon} + \frac{1+\epsilon}{\tau^2} - 1\right)\right). \tag{4.12}$$

Proof:

$$x_j = v_{1,j} - v_{2,j} = \frac{1}{\sqrt{k}}\sum_{i=1}^{D} r_{ij}(u_{1,i} - u_{2,i}), \quad \hat{d} = \sum_{j=1}^{k} |x_j|^2$$

Let $\tau^2 = T^2(r_{ij})$. Then by the properties of sub-Gaussian distributions, e.g., (4.6)

$$\begin{aligned}
&T^2\left(x_j\right) \leq \frac{1}{k}\sum_{i=1}^{D}|u_{1,i} - u_{2,i}|^2\tau^2 = \frac{d}{k}\tau^2, \\
&E\left(\exp(x_j^2 t)\right) \leq \left(1 - 2\tau^2(x_j)t\right)^{-\frac{1}{2}} \leq \left(1 - 2\tau^2\frac{d}{k}t\right)^{-\frac{1}{2}} \\
&E\left(\exp(\hat{d}t)\right) = \prod_{j=1}^{k} E\left(\exp(x_j^2 t)\right) \leq \left(1 - 2\tau^2\frac{d}{k}t\right)^{-\frac{k}{2}}
\end{aligned}$$

We can then apply the Chernoff inequality:

$$\mathbf{Pr}\left(\hat{d} - d \geq \epsilon d\right) \leq \frac{E\left(\exp\left(\hat{d}t\right)\right)}{\exp\left((1+\epsilon)dt\right)}, \tag{4.13}$$

whose minimum is $\exp\left(-\frac{k}{2}\left(\log\frac{\tau^2}{1+\epsilon} + \frac{1+\epsilon}{\tau^2} - 1\right)\right)$, *attained at* $t = \left(1 - \frac{\tau^2}{1+\epsilon}\right)\frac{k}{2\tau^2 d}$, *provided* $0 \leq 1 - \frac{\tau^2}{1+\epsilon} < 1$, *i.e.*, $0 < \frac{\tau^2}{1+\epsilon} \leq 1$.

Denote $k^{\tau^2(s)}$ such that the upper bound (4.12) satisfies a specified accuracy α':

$$\exp\left(-\frac{k^{\tau^2(s)}}{2}\left(\log\frac{\tau^2}{1+\epsilon}\right)+\frac{1+\epsilon}{\tau^2}-1\right)=\alpha', \tag{4.14}$$

we can compare $k^{\tau^2(s)}$ with $k^1 = k^{\tau^2(1\leq s\leq 3)}$. We call $\frac{k^{\tau^2(s)}}{k^1}$ as "k inflation factor," which shows how much we have to increase k, as a function of s, to reach the same level of accuracy. Figure 4.2(a) plots this ratio

$$\frac{k^{\tau^2(s)}}{k^1}=\frac{\log\frac{1}{1+\epsilon}+\epsilon}{\log\frac{\tau^2}{1+\epsilon}+\frac{1+\epsilon}{\tau^2}-1}. \tag{4.15}$$

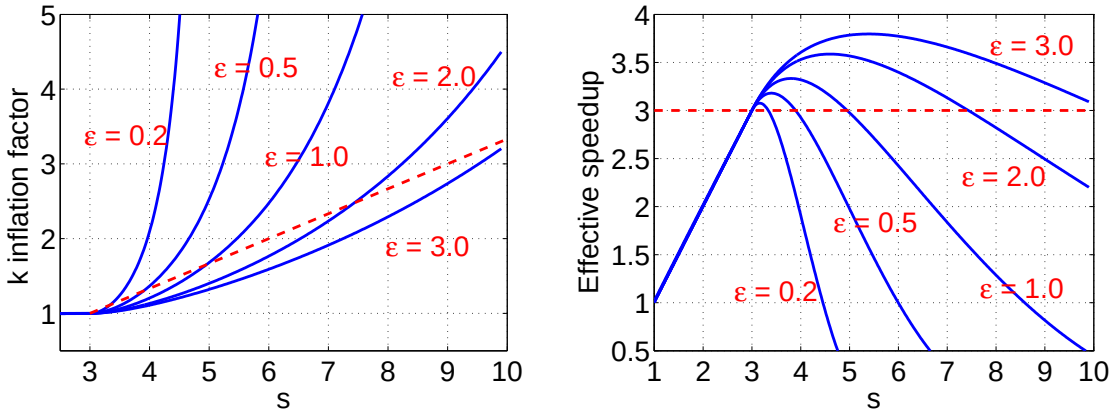
Figure 4.2(b) plots $s/\frac{k^{\tau^2(s)}}{k^1}$, reporting the "effective speedup." Computing the projection mapping costs $O\left(\frac{k}{s}nD\right)$. Increasing s speeds up the computation but may increase the required k to maintain the same accuracy. The "effective speedup" can measure the true performance. We already know that with $s = 3$, the "effective speedup" is $s = 3$. Therefore we use 3 as the baseline. Figure 4.2(b) shows that we can achieve an effective speedup more than threefold for a certain range of s, i.e., $s = 3$ is not the optimum. For any given ϵ, we can compute the "optimal" s easily.

The (annoying) restriction of $\tau^2(s) \leq 1 + \epsilon$ shows that in the worst case, there is no hope of increasing s much larger than 3 without hurting the performance.

4.2 Very Sparse Random Projections

In many applications in data mining and machine learning, the worst-case scenario (e.g., extremely "heavy-tailed") is often not interesting. Instead, it is reasonable to assume certain "regularity conditions" on the data. For example, we should at least make sure that the original data do have meaningful l_2 moments before we can talk about dimension reduction in the l_2 norm. When the data do not have meaningful l_2 moments, the data have to be preprocessed, for example, via various so-called *term-weighting* procedures.

In the mean while, the data dimension D should be large otherwise there would



(a) (b)

Figure 4.2: (a): k inflation factor, $\frac{k\tau^2(s)}{k^1}$ in (4.15). The dashed line with slope 1/3 indicates that we can achieve "effective speedup" more than 3-fold if the curve is below this line. (b): The "effective speedup," defined as $s/\frac{k\tau^2(s)}{k^1}$. Curves above the dashed line indicate the range of s where $s > 3$ may be a better choice than $s = 3$.

be no need for dimension reduction in the very beginning.

Assuming a high dimension D and data regularity conditions, the asymptotic analysis becomes useful. For example, assuming that data have bounded third moments, we will show that it is reasonable to let $s = \sqrt{D}$, a significant $\sqrt{D}$-fold speedup.

4.2.1 Variances and Asymptotic Distribution

Recall $\hat{d} = \|v_1 - v_2\|^2$ and $\hat{a} = v_1^{\mathrm{T}} v_2$ are unbiased estimators for $d = \|u_1 - u_2\|^2$ and $a = u_1^{\mathrm{T}} u_2$ respectively, for any $s \geq 1$. Lemma 8 (proved in Section 4.4.1) derives the variances. Recall $m_1 = \|u_1\|^2$ and $m_2 = \|u_2\|^2$.

Lemma 8

$$Var\left(\hat{d}\right) = \frac{1}{k}\left(2d^2 + (s-3)\sum_{i=1}^{D}\left(u_{1,i} - u_{2,i}\right)^4\right) \tag{4.16}$$

$$Var\left(\hat{a}\right) = \frac{1}{k}\left(m_1 m_2 + a^2 + (s-3)\sum_{i=1}^{D} u_{1,i}^2 u_{2,i}^2\right). \tag{4.17}$$

Compared with the corresponding variance formulas in *normal random projections* described in Chapter 3, the additional terms involve $(s-3)$. This suggests that if $s < 3$, one can obtain more accurate estimates in terms of the variances.

We are interested in $s \gg 3$, i.e., *very sparse random projections*. Note that the terms involving $(s-3)$ are negligibly small if the data have bounded fourth moment. For example, $d^2 = \left(\sum_{i=1}^D (u_{1,i}-u_{2,i})^2\right)^2 = \sum_{i=1}^D (u_{1,i}-u_{2,i})^4 + \sum_{i\neq i'} (u_{1,i}-u_{2,i})^2 (u_{1,i'}-u_{2,i'})^2$, with D diagonal terms and $\frac{D(D-1)}{2}$ cross-terms. By assuming bounded fourth moments, as $D \to \infty$, the diagonal terms are asymptotically negligible.

Lemma 9 studies the asymptotic distributing of $v_1 - v_2$, proved in Section 4.4.2.

Lemma 9 *Assume, for any fixed $\delta > 0$, the data satisfy*

$$\left(\frac{s}{D}\right)^{\frac{\delta}{2}} \frac{\sum_{i=1}^D |u_{1,i}-u_{2,i}|^{2+\delta}/D}{\left(\sum_{i=1}^D (u_{1,i}-u_{2,i})^2/D\right)^{(2+\delta)/2}} \to 0, \qquad as \quad D \to \infty. \tag{4.18}$$

Then

$$\frac{v_{1,j}-v_{2,j}}{\sqrt{d/k}} \overset{\mathcal{L}}{\Longrightarrow} N(0,1), \qquad \frac{\|v_1-v_2\|^2}{d/k} \overset{\mathcal{L}}{\Longrightarrow} \chi^2_k, \tag{4.19}$$

with the rate of convergence

$$|F_{v_{1,j}-v_{2,j}}(y) - \Phi(y)| \leq 0.7915\sqrt{s}\frac{\sum_{i=1}^D |u_{1,i}-u_{2,i}|^3}{d^{3/2}}$$

where $\overset{\mathcal{L}}{\Longrightarrow}$ denotes "convergence in distribution." $F_{v_{1,j}-v_{2,j}}(y)$ is the empirical cumulative density function (CDF) of $v_{1,j}-v_{2,j}$ and $\Phi(y)$ is $N(0,1)$ CDF.

The condition (4.18) is satisfied if we assume $s = o(D)$ and data have bounded $2+\delta$ moment, i.e., $\mathrm{E}|u_{1,i}-u_{2,i}|^{2+\delta} < \infty$, $\forall \delta > 0$, a very weak condition. When we make stronger assumptions, we can achieve faster convergence. For example,

assuming bounded third moment, $\mathrm{E}|u_{1,i} - u_{2,i}|^3 < \infty$, then the Berry-Esseen bound

$$0.7915\sqrt{s}\frac{\sum_{i=1}^{D}|u_{1,i} - u_{2,i}|^3}{d^{3/2}} \to 0.7915\sqrt{\frac{s}{D}}\frac{\mathrm{E}|u_{1,i} - u_{2,i}|^3}{(\mathrm{E}|u_{1,i} - u_{2,i}|^2)^{3/2}}. \tag{4.20}$$

In other words, if we assume bounded third moment on the data, the rate of convergence is in the order of $O\left(\sqrt{\frac{s}{D}}\right)$; and $s = \sqrt{D}$ appears to be a good trade-off.

4.2.2 Tail Bounds

In Section 4.1, we have studied the strict tail bounds for *sub-Gaussian random projections*. Due to the restriction $\tau^2 < 1 + \epsilon$ (i.e., s can only be slightly larger than 3), sub-Gaussian analysis seems hopeless in the worst case, if $s \gg 3$ is desired.

Lemma 9 suggests that assuming bounded third moment, as $D \to \infty$, one can approximately use the asymptotic tail bounds based on the chi-square distribution. Lemma 9, however, does not provide an answer when D is large but not too large.

This section conducts some *non-asymptotic* analysis while making reasonable data assumptions. We will assume that the data have bounded third moments, i.e., $\frac{\sum_{i=1}^{D}|u_{1,i}-u_{2,i}|^3}{D}$ convergences to a finite positive value. To fix ideas, we will further consider a Pareto distribution.

If z follows an η-Pareto distribution[1], then $\mathrm{E}(z^\gamma) < \infty$ if $\gamma < \eta$ and $= \infty$ is $\gamma \geq \eta$. We will consider a 3-Pareto distribution, although strictly speaking we should consider a $3 + \delta$-Pareto in order to have bounded third moments. This is probably the most "stringent" distribution assuming bounded third moments.

Let $z_i = u_{1,i} - u_{2,i}$. Similar to (4.13) and (3.9), we can still study the tail bounds

$$\mathbf{Pr}\left(\hat{d} - d \geq \epsilon d\right) \leq \frac{\mathrm{E}\left(\exp\left(k\hat{d}/dt\right)\right)}{\exp\left(k(1+\epsilon)t\right)}. \tag{4.21}$$

It is difficult to obtain the "optimal" t value, even if the data are constant (i.e., $z_i =$ constant for all i). However, we can "borrow" the optimal $t = t_{NR}$ for the normal case as in (3.9). The bound (4.21) will still hold because it is true for any $t > 0$ and

[1] If z is η-Pareto, the density $f(z;\eta) = \frac{\eta}{z^{\eta+1}}$, where $z \geq 1$.

it is going to be "asymptotically optimal" as long as the asymptotic normality holds. Therefore, plugging $t = t_{NR} = \frac{\epsilon}{2(1+\epsilon)}$ in (4.21) yields

$$\mathbf{Pr}\left(\hat{d} - d \geq \epsilon d\right) \leq \exp\left(-\frac{k}{2}\left(\epsilon - 2\log \mathrm{E}\left(\exp\left(\frac{\epsilon}{2(1+\epsilon)d}\left(\sum_{i=1}^{D} z_i r_{ij}\right)^2\right)\right)\right)\right). \quad (4.22)$$

Since evaluating the expectation in (4.22) is difficult even if all z_i are constants, we resort to simulations. If we "fix" the data z_i (i.e., conditioning on the data), a simulation is simply a widely acceptable approach for numerically evaluating the expectation.

Again, the "k inflation factor" would be

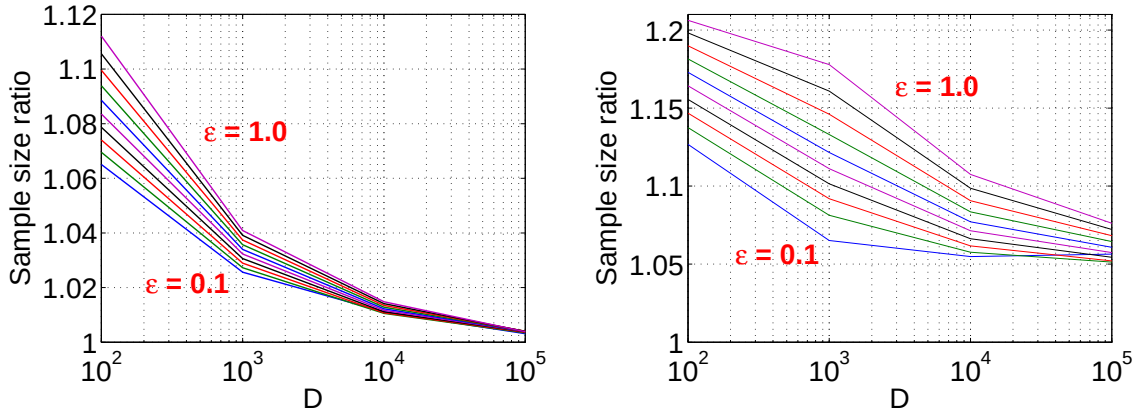
$$\frac{\epsilon - \log(1+\epsilon)}{\epsilon - 2\log \mathrm{E}\left(\exp\left(\frac{\epsilon}{2(1+\epsilon)d}\left(\sum_{i=1}^{D} z_i r_{ij}\right)^2\right)\right)}, \quad (4.23)$$

where the expectation is evaluated by simulations as follows:

- Simulate a data vector z_i, from a certain distribution (normal or 3-Pareto), of length $D = 100$, 10^3, 10^4, and 10^5.
- For every D, we simulate a vector of r_{ij} from (4.1) with length D and $s = \sqrt{D}$. We compute d and $\sum_{i=1}^{D} z_i r_{ij}$. We repeat this step 10^6 times for every D. This way, we can evaluate the expectation (4.23) conditioning on the data.
- We simulate the data z_i 15 times and average the conditional expectations.

We consider two types of data, normal and 3-Pareto. The normal data type is often what we hope the data should be like while the 3-Pareto data are probably the "worst case" among distributions with bounded third moments.

Figure 4.3 presents the simulations. The general trend is that the inflation factor is close to 1 even when D is not very large and it decreases monotonically as D increases. Also, as we would expect, the inflation factor is (slightly) higher for the 3-Pareto data than for the normal data. The simulations are promising that $s = \sqrt{D}$ is indeed a good trade-off between accuracy and cost, as long as the data are reasonable.



(a) Normal

(b) 3-Pareto

Figure 4.3: The "k inflation factor" as defined in (4.23). We simulate the expectation in (4.23) by first conditioning on the data (10^6 simulations for every D) and then generating the data 15 times for the unconditional expectation. We consider both normal and 3-Pareto. The inflation factor is close to 1. Each curve stands for a particular ϵ, from 0.1 to 1.0.

4.2.3 Very Sparse Random Projections for Classifying Microarray Data

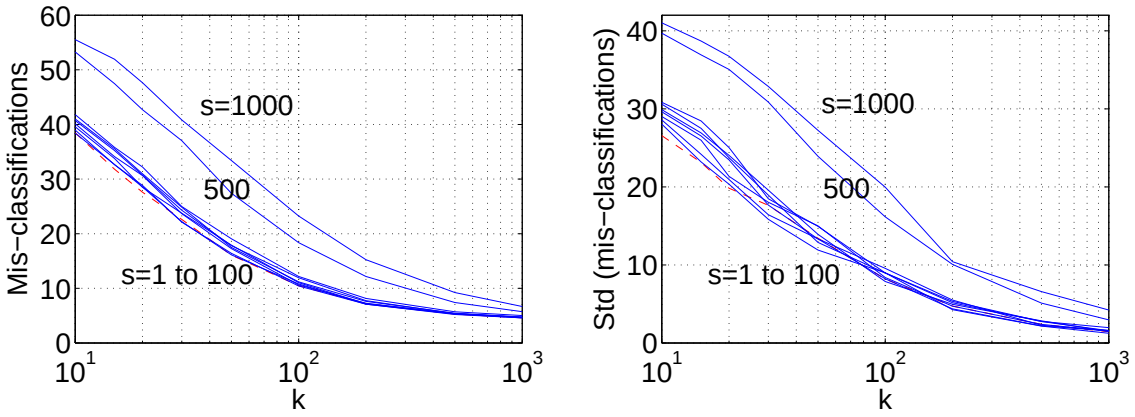
Usually the purpose of computing distances is for the subsequent tasks such as clustering, classification, information retrieval, etc. Here we consider the task of classifying diseases (cancer types) in the Harvard microarray dataset [24]. The original dataset contains 203 samples (specimens) in 12600 gene dimensions, including 139 lung adenocarcinomas (12 of which may be suspicious), 17 normal samples, 20 pulmonary carcinoids, 21 squamous cell lung carcinomas, and 6 SCLC cases. We select the first three classes (in total 139 + 17+20 = 176 samples).

A simple nearest neighbor classifier can classify the samples fairly well using the (l_2) correlation distance (i.e., 1− correlation coefficient), when m = 1, 3, 5, 7, 9, the m-nearest neighbor classifier mis-classifies 6, 4, 4, 4, 5, samples, respectively. In Chapter 5, we will show that using the l_1 distance can produce even better results.

We project the data from D = 12600 dimensions down to k dimensions by very

sparse random projections. We estimate the inner products (subsequently the correlation distances) from the projected data and use them to classify the cancers using a 5-nearest neighbor classifier.

Figure 4.4 presents the performance of *very sparse random projections*. Note that the original data dimension $D = 12600$ ($\sqrt{D} = 112$). We notice that the classification accuracies are similar when $s = 1$ to 100, especially after $k > 100$.



(a) Mean (b) Standard error (std)

Figure 4.4: Using *very sparse random projections* with s from 1 to 1000, after $k > 100$, we can achieve similar performance, in terms of classification errors with a 5-nearest neighbor classifier, as using the original correlation distances. In (a), the average classification errors (over 1000 runs) indicate that when $s = 1$ to 100, the classification accuracies are similar. In (b) we plot the standard errors, which have very similar trend as the mean. The dashed curve stands for $s = 1$.

Note that in this example, D is probably not large enough to show the full advantage of *very sparse random projections*. Also, the data are quite heavy-tailed. We will present a similar study using *Cauchy random projections* in Chapter 5.

4.3 Summary

For dimension reduction in the l_2 norm using random projections, it is beneficial to use a three-point sparse distribution $[-1, 0, 1]$ with probabilities $\left[\frac{1}{2s}, 1 - \frac{1}{s}, \frac{1}{2s}\right]$, as

opposed to normal. The *sub-Gaussian analysis* is a useful tool to prove that we can let s be slightly larger than 3 without hurting the accuracy, even in the worst case.

The worse-case scenario is not practically interesting. It is reasonable to assume that the data should have at least bounded second moments (otherwise one should not use the l_2 norm). For reasonable high-dimensional data, we propose *very sparse random projections* with $s \gg 3$, to achieve a significant s-fold speedup in computing $\mathbf{A} \times \mathbf{R}$, which can be prohibitive, for example, in massive data streams.

4.4 Proofs

4.4.1 Proof of Lemma 8

We only present the proof for $\mathrm{Var}\left(\hat{a}\right)$ since it is slightly more involved than $\mathrm{Var}\left(\hat{d}\right)$.

$$
\begin{aligned}
v_{1,j}v_{2,j} =& \frac{1}{k}\left(\sum_{i=1}^{D} r_{ij}u_{1,i}\right)\left(\sum_{i=1}^{D} r_{ij}u_{2,i}\right) = \frac{1}{k}\left(\sum_{i=1}^{D}\left(r_{ij}^2\right)u_{1,i}u_{2,i} + \sum_{i\neq i'}\left(r_{ij}\right)u_{1,i}\left(r_{i'j}\right)u_{2,i'}\right), \\
v_{1,j}^2v_{2,j}^2 =& \frac{1}{k^2}\left(\sum_{i=1}^{D}\left(r_{ij}^2\right)u_{1,i}u_{2,i} + \sum_{i\neq i'}\left(r_{ij}\right)u_{1,i}\left(r_{i'j}\right)u_{2,i'}\right)^2 \\
=& \frac{1}{k^2}\left(\begin{array}{l} \sum_{i=1}^{D}\left(r_{ij}^4\right)u_{1,i}^2u_{2,i}^2 + 2\sum_{i<i'}\left(r_{ij}^2\right)u_{1,i}u_{2,i}\left(r_{i'j}^2\right)u_{1,i'}u_{2,i'} + \\ \left(\sum_{i\neq i'}\left(r_{ij}\right)u_{1,i}\left(r_{i'j}\right)u_{2,i'}\right)^2 + \left(\sum_{i=1}^{D}\left(r_{ij}^2\right)u_{1,i}u_{2,i}\right)\left(\sum_{i\neq i'}\left(r_{ij}\right)u_{1,i}\left(r_{i'j}\right)u_{2,i'}\right) \end{array}\right).
\end{aligned}
$$

Thus,

$$\begin{aligned}
\mathrm{E}\left(v_{1,j}^2 v_{2,j}^2\right) =& \frac{1}{k^2}\left(s\sum_{j=1}^{D} u_{1,i}^2 u_{2,i}^2 + 4\sum_{i<i'} u_{1,i}u_{2,i}u_{1,i'}u_{2,i'} + \sum_{i\neq i'} u_{1,i}^2 u_{2,i'}^2\right)\\
=& \frac{1}{k^2}\left((s-2)\sum_{i=1}^{D} u_{1,i}^2 u_{2,i}^2 + \sum_{i\neq i'} u_{1,i}^2 u_{2,i'}^2 + 2a^2\right)\\
=& \frac{1}{k^2}\left(m_1 m_2 + (s-3)\sum_{i=1}^{D} u_{1,i}^2 u_{2,i}^2 + 2a^2\right),\\
\mathrm{Var}\left(v_{1,i}v_{2,i}\right) =& \frac{1}{k^2}\left(m_1 m_2 + a^2 + (s-3)\sum_{i=1}^{D} u_{1,i}^2 u_{2,i}^2\right),\\
\mathrm{Var}\left(v_1^{\mathrm{T}} v_2\right) =& \frac{1}{k}\left(m_1 m_2 + a^2 + (s-3)\sum_{i=1}^{D} u_{1,i}^2 u_{2,i}^2\right).
\end{aligned}$$

4.4.2 Proof of Lemma 9

The standard Lindeberg central limit theorem (CLT) and the Berry-Esseen theorem are needed for the proof [70, Theorems VIII.4.3 and XVI.5.2]. The best Berry-Esseen constant 0.7915 is from [155] (for the non-i.i.d. case), which seems not widely known.

Write $v_{1,j} - v_{2,j} = \sum_{i=1}^{D} z_i$, with $z_i = \frac{1}{\sqrt{k}}\left(r_{ij}\right)\left(u_{1,i} - u_{2,i}\right)$. Then

$$\mathrm{E}(z_i) = 0, \quad \mathrm{Var}(z_i) = \frac{\left(u_{1,i} - u_{2,i}\right)^2}{k}, \quad \mathrm{E}(|z_i|^{2+\delta}) = s^{\frac{\delta}{2}}\frac{|u_{1,i} - u_{2,i}|^{2+\delta}}{k^{(2+\delta)/2}}, \forall \delta > 0.$$

Let $s_D^2 = \sum_{i=1}^{D} \mathrm{Var}(z_i) = \frac{\sum_{i=1}^{D}(u_{1,i}-u_{2,i})^2}{k} = \frac{d}{k}$. Assume the Lindeberg condition

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

Then

$$\frac{\sum_{i=1}^{D} z_i}{s_D} = \frac{v_{1,j} - v_{2,j}}{\sqrt{d/k}} \xrightarrow{\mathcal{L}} N(0,1),$$

which immediately leads to

$$\frac{(v_{1,i}-v_{2,i})^2}{d/k} \xrightarrow{\mathcal{L}} \chi_1^2, \qquad \frac{\|v_1-v_2\|^2}{d/k} \xrightarrow{\mathcal{L}} \chi_k^2.$$

We need to go back and check the Lindeberg condition. By our assumption (4.18),

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

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

$$|F_{v_{1,j}-v_{2,j}}(y) - \Phi(y)| \leq 0.7915 \frac{\rho_D}{s_D^3} = 0.7915\sqrt{s}\frac{\sum_{i=1}^{D}|u_{1,i}-u_{2,i}|^3}{d^{3/2}}$$

This completes the proof.

Chapter 5

Cauchy Random Projections for l_1

While Chapter 3 and Chapter 4 are devoted to random projections for dimension reduction in the l_2 norm, this chapter concerns dimension reductions in the l_1 norm.

We will again consider a data matrix $\mathbf{A} \in \mathbf{R}^{n \times D}$. We generate a random projection matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$, sampled from i.i.d. standard Cauchy $C(0, 1)$. We will let the projected matrix be $\mathbf{B} = \mathbf{A} \times \mathbf{R} \in \mathbb{R}^{n \times k}$ without the normalization term $\frac{1}{\sqrt{k}}$ appearing in Chapter 3 and Chapter 4. The task also boils down to a statistical estimation task, for estimating the scale parameter from k i.i.d Cauchy random variables.

Because Cauchy does not have a finite mean, we can not use a linear estimator as in *normal random projections*. Furthermore, the impossibility result[32, 109, 33] has proved that, when using linear projections, there is no hope of using linear estimators without incurring large errors, i.e., no Johnson-Lindenstrauss (JL) Lemma for l_1.

In this chapter, we provide three nonlinear estimators and derive an analog of the JL Lemma for l_1. Since our estimators are not metrics, this analog of the JL Lemma is weaker than the classical JL Lemma for l_2; and we will leave it for future research to develop sub-linear nearest neighbor algorithms using these nonlinear estimators.

The main results in this chapter appeared in a conference proceeding[126] and will appear in a journal paper [127].

5.1 Summary of Main Results

We again consider the leading two rows, u_1 and u_2, in $\mathbf{A}$, and the leading two rows, v_1 and v_2, in $\mathbf{B}$. We denote by $d = \sum_{i=1}^{D} |u_{1,i} - u_{2,i}|$ the l_1 distance.

In *Cauchy random projections*, the key task is to recover the scale parameter of a Cauchy 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 $\mathrm{E}(x_j) = \infty$.

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

- **The sample median estimators**

  The sample median estimator $\hat{d}_{me}$ and the unbiased version $\hat{d}_{me,c}$ are

$$\hat{d}_{me} = \mathrm{median}(|x_j|, j = 1, 2, ..., k) \tag{5.1}$$

$$\hat{d}_{me,c} = \frac{\hat{d}_{me}}{b_{me}}, \tag{5.2}$$

$$b_{me} = \int_0^1 \frac{(2m+1)!}{(m!)^2} \tan\left(\frac{\pi}{2}t\right)\left(t - t^2\right)^m dt, \quad k = 2m+1 \tag{5.3}$$

  For convenience, we only consider $k = 2m + 1$, $m = 1, 2, ...$

  Among all quantile estimators, $\hat{d}_{me}$ (and $\hat{d}_{me,c}$) achieves the smallest asymptotic variance.

- **The geometric mean estimators**

  The geometric mean estimator $\hat{d}_{gm}$ and its unbiased version $\hat{d}_{gm,c}$ are

$$\hat{d}_{gm} = \prod_{j=1}^{k} |x_j|^{1/k}, \tag{5.4}$$

$$\hat{d}_{gm,c} = \cos^k\left(\frac{\pi}{2k}\right) \prod_{j=1}^{k} |x_j|^{1/k}. \tag{5.5}$$

  In terms of the asymptotic variances, the geometric mean estimators are asymptotically equivalent to the sample median estimators. In terms of the tail

bounds, however, the sample median estimators may require (up to) twice as many samples.

- **The maximum likelihood estimator**

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

$$\hat{d}_{MLE,c} = \hat{d}_{MLE}\left(1 - \frac{1}{k}\right), \tag{5.6}$$

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

$$-\frac{k}{\hat{d}_{MLE}} + \sum_{j=1}^{k} \frac{2\hat{d}_{MLE}}{x_j^2 + \hat{d}_{MLE}^2} = 0. \tag{5.7}$$

  The sample median estimators and the geometric mean estimators are about 80% as accurate the MLE, in terms of the asymptotic variances. While it is difficult to derive closed-form tail bounds, we show that the distribution of $\hat{d}_{MLE,c}$ can be accurately approximated by an inverse Gaussian.

5.2 The Sample Median Estimators

Recall that 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 [67, 68, 138]. Due to the symmetry, the quantile estimator becomes

$$\hat{d}_q = \frac{q\text{-Quantile}\{|x_j|, j = 1, 2, ..., k\}}{q\text{-Quantile}\{|C(0,1)|\}}. \tag{5.8}$$

As illustrated in Lemma 10, $\hat{d}_q$ achieves the smallest asymptotic variance when $q = 0.5$, i.e., sample median. Lemma 10 further summarizes some properties of the *sample median* estimator (denoted by $\hat{d}_{me}$).

Lemma 10 *The sample median estimator, $\hat{d}_{me}$,*

$$\hat{d}_{me} = median\{|x_j|, j = 1, 2, ..., k\}, \tag{5.9}$$

achieves the smallest asymptotic variance among all quantile estimators $\hat{d}_q$. Furthermore, $\hat{d}_{me}$ is asymptotically unbiased and normal

$$\sqrt{k}\left(\hat{d}_{me} - d\right) \overset{D}{\Longrightarrow} N\left(0, \frac{\pi^2}{4}d^2\right). \tag{5.10}$$

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

$$E\left(\hat{d}_{me}\right)^r = d^r\left(\int_0^1 \frac{(2m+1)!}{(m!)^2}\tan^r\left(\frac{\pi}{2}t\right)\left(t - t^2\right)^m dt\right), \quad m \geq r \tag{5.11}$$

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 $|C(0, d)|$:*

$$f(z; d) = \frac{2d}{\pi}\frac{1}{z^2 + d^2}, \quad F(z; d) = \frac{2}{\pi}\tan^{-1}\left(\frac{z}{d}\right), \quad z > 0. \tag{5.12}$$

The inverse of $F(z; d)$ is $F^{-1}(q; d) = d\tan\left(\frac{\pi}{2}q\right)$. The q-Quantile estimator defined in (5.8) can be written as

$$\hat{d}_q = \frac{q\text{-}Quantile\{|x_j|, j = 1, 2, ..., k\}}{\tan\left(\frac{\pi}{2}q\right)}. \tag{5.13}$$

By the asymptotic normality of sample quantiles [57, Theorem 9.2], we know

$$\sqrt{k}\left(\hat{d}_q - d\right) \overset{D}{\Longrightarrow} N\left(0, \frac{q - q^2}{\left(f\left(d\tan\left(\frac{\pi}{2}q\right); d\right) \times \tan\left(\frac{\pi}{2}q\right)\right)^2}\right), \tag{5.14}$$

i.e., $\hat{d}_q$ *is asymptotically unbiased and normal with the asymptotic variance*

$$Var\left(\hat{d}_q\right) = \frac{d^2}{k}\frac{q-q^2}{\left(\frac{2}{\pi}\frac{\tan\left(\frac{\pi}{2}q\right)}{\tan^2\left(\frac{\pi}{2}q\right)+1}\right)^2} + O\left(\frac{1}{k^2}\right) = \frac{d^2}{k}\pi^2\frac{q-q^2}{\sin^2(\pi q)} + O\left(\frac{1}{k^2}\right). \quad (5.15)$$

It is easy to check that $\frac{q-q^2}{\sin^2(\pi q)}$ *is a convex function of* $q \in [0,\ 1]$ *and reaches the minimum* $\frac{1}{4}$ *at* $q = 0.5$*. Thus, the sample median estimator* $\hat{d}_{me} = \hat{d}_{q=0.5}$ *has the asymptotic variance* $\frac{d^2}{k}\frac{\pi^2}{4}$*.*

For convenience, we assume $k = 2m + 1$*. By properties of sample quantile [57, Chapter 2.1], the probability density of* $\hat{d}_{me}$ *is*

$$f_{\hat{d}_{me}}(z) = \frac{(2m+1)!}{(m!)^2}\left(F(z;d)(1-F(z;d))\right)^m f(z;d), \quad (5.16)$$

from which we can write down the r^{th} *moment of* $\hat{d}_{me}$ *in (5.11).*

The next task is to analyze the exact tail behavior of $\hat{d}_{me}$, e.g., $\mathbf{Pr}\left(|\hat{d}_{me} - d| > \epsilon d\right)$. Proved in Section 5.7.1, Lemma 11 derives the precise tail bounds for $\hat{d}_{me}$.

Lemma 11

$$\mathbf{Pr}\left(\hat{d}_{me} \geq (1+\epsilon)d\right) \leq \exp\left(-k\frac{\epsilon^2}{G_{R,me}}\right), \quad \epsilon > 0 \quad (5.17)$$

$$\mathbf{Pr}\left(\hat{d}_{me} \leq (1-\epsilon)d\right) \leq \exp\left(-k\frac{\epsilon^2}{G_{L,me}}\right), \quad 0 < \epsilon < 1 \quad (5.18)$$

$$G_{R,me} = \frac{2\epsilon^2}{-\log\left(2 - \frac{4}{\pi}\tan^{-1}(1+\epsilon)\right) - \log\left(\frac{4}{\pi}\tan^{-1}(1+\epsilon)\right)}, \quad (5.19)$$

$$G_{L,me} = \frac{2\epsilon^2}{-\log\left(2 - \frac{4}{\pi}\tan^{-1}(1-\epsilon)\right) - \log\left(\frac{4}{\pi}\tan^{-1}(1-\epsilon)\right)}. \quad (5.20)$$

Moreover, $G_{R,me} \to \frac{\pi^2}{2}$*, and* $G_{L,me} \to \frac{\pi^2}{2}$*, as* $\epsilon \to 0+$*.*

Because the tail bounds in Lemma 11 are in exponential forms, they imply that an analog of the JL Lemma can be established using the *sample median* estimator.

$\hat{d}_{me}$ is biased, but we can easily remove its bias, as described in Lemma 12.

Lemma 12 *The estimator,*

$$\hat{d}_{me,c} = \frac{\hat{d}_{me}}{b_{me}}, \tag{5.21}$$

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

$$b_{me} = \frac{E\left(\hat{d}_{me}\right)}{d} = \int_0^1 \frac{(2m+1)!}{(m!)^2} \tan\left(\frac{\pi}{2}t\right) \left(t - t^2\right)^m dt, \qquad (k = 2m+1). \tag{5.22}$$

The variance of $\hat{d}_{me,c}$ *is* ($k = 2m+1 \geq 5$)

$$Var\left(\hat{d}_{me,c}\right) = d^2 \left(\frac{(m!)^2}{(2m+1)!} \frac{\int_0^1 \tan^2\left(\frac{\pi}{2}t\right) \left(t - t^2\right)^m dt}{\left(\int_0^1 \tan\left(\frac{\pi}{2}t\right) \left(t - t^2\right)^m dt\right)^2} - 1 \right), \tag{5.23}$$

$\hat{d}_{gm,c}$ *and* $\hat{d}_{gm}$ *are asymptotically equivalent, i.e.,*

$$\sqrt{k}\left(\hat{d}_{me,c} - d\right) \overset{D}{\Longrightarrow} N\left(0, \frac{\pi^2}{4} d^2\right). \tag{5.24}$$

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

$$b_{me} \geq 1, \qquad \lim_{m \to \infty} b_{me} = 1. \tag{5.25}$$

Proof: *Most of the results follow directly from Lemma 10. 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* r*th moment is* $E(z^r) = \frac{(2m+1)!(m+r)!}{(2m+1+r)!m!}$.

By Taylor expansions [85, 1.411.6],

$$\tan\left(\frac{\pi}{2}t\right) = \sum_{j=1}^{\infty} \frac{2^{2j}\left(2^{2j}-1\right)}{(2j)!}|B_{2j}|\left(\frac{\pi}{2}\right)^{2j-1} t^{2j-1}, \tag{5.26}$$

where B_{2j} is the Bernoulli number *[85, 9.61]. Therefore,*

$$b_{me} = \sum_{j=1}^{\infty} \frac{2^{2j}\left(2^{2j}-1\right)}{(2j)!}|B_{2j}|\left(\frac{\pi}{2}\right)^{2j-1} \frac{(2m+1)!(m+2j-1)!}{(2m+2j)!m!}. \tag{5.27}$$

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*

$$b_{me} \to \sum_{j=1}^{\infty} \frac{2^{2j}\left(2^{2j}-1\right)}{(2j)!}|B_{2j}|\left(\frac{\pi}{2}\right)^{2j-1}\left(\frac{1}{2}\right)^{2j-1} = \tan\left(\frac{\pi}{2}\frac{1}{2}\right) = 1. \tag{5.28}$$

It is well-known that bias-corrections are not always beneficial due to 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 but also reduces the variance.

The bias-correction factor b_{me} can be numerically evaluated and tabulated, at least for small k. Figure 5.1 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.

5.3 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 also allow us to derive tail bounds in explicit forms and (consequently) an analog of the JL Lemma.

In Chapter 6, we will study the geometric mean estimator for general $0 < \alpha \leq 2$. We still present the results on the geometric mean for l_1 because the derivations are more intuitive and much simpler, and may provide better insights. Moreover, we can provide sharper tail bounds in the l_1 case because the explicit density is available.

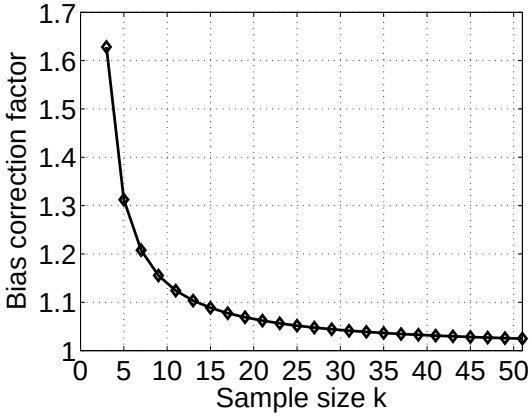


Figure 5.1: The bias correction factor, b_{me} in (5.22), as a function of $k = 2m + 1$. After $k > 50$, the bias is negligible.

Lemma 13 *Assume $x \sim C(0, d)$. Then*

$$E\left(|x|^\lambda\right) = \frac{d^\lambda}{\cos(\lambda\pi/2)}, \qquad |\lambda| < 1 \tag{5.29}$$

Proof: *Using the integral table [85, 3.221.1, page 337],*

$$E\left(|x|^\lambda\right) = \frac{2d}{\pi}\int_0^\infty \frac{y^\lambda}{y^2+d^2}dy = \frac{d^\lambda}{\pi}\int_0^\infty \frac{y^{\frac{\lambda-1}{2}}}{y+1}dy = \frac{d^\lambda}{\cos(\lambda\pi/2)}. \tag{5.30}$$

From Lemma 13, 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, proved in Section 5.7.2.

Lemma 14

$$\hat{d}_{gm,c} = \cos^k\left(\frac{\pi}{2k}\right)\prod_{j=1}^{k}|x_j|^{1/k}, \quad k > 1 \tag{5.31}$$

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

$$Var\left(\hat{d}_{gm,c}\right) = d^2\left(\frac{\cos^{2k}\left(\frac{\pi}{2k}\right)}{\cos^k\left(\frac{\pi}{k}\right)} - 1\right) = \frac{d^2}{k}\frac{\pi^2}{4} + \frac{\pi^4}{32}\frac{d^2}{k^2} + O\left(\frac{1}{k^3}\right). \tag{5.32}$$

The third and fourth central moments are

$$E\left(\hat{d}_{gm,c} - E\left(\hat{d}_{gm,c}\right)\right)^3 = \frac{3\pi^4}{16}\frac{d^3}{k^2} + O\left(\frac{1}{k^3}\right) \tag{5.33}$$

$$E\left(\hat{d}_{gm,c} - E\left(\hat{d}_{gm,c}\right)\right)^4 = \frac{3\pi^4}{16}\frac{d^4}{k^2} + O\left(\frac{1}{k^3}\right). \tag{5.34}$$

The higher (third or fourth) moments may be useful for approximating the distribution of $\hat{d}_{gm,c}$. In Section 5.5, 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 a similar technique to approximate $\hat{d}_{gm,c}$. Fortunately, we are able to derive the exact tail bounds for the geometric mean estimators.

Before presenting Lemma 16 for $\hat{d}_{gm,c}$, we prove tail bounds for $\hat{d}_{gm}$, the geometric mean estimator without bias-correction, in Lemma 15. This is for convenient in comparing the tail bounds of the *sample median* estimator $\hat{d}_{me}$ in Lemma 11. See the proof in Section 5.7.3.

Lemma 15 *Without the bias-correction, the geometric mean estimator*

$$\hat{d}_{gm} = \prod_{j=1}^{k} |x_j|^{1/k} \quad k > 1 \tag{5.35}$$

has tail bounds:

$$\mathbf{Pr}\left(\hat{d}_{gm} \geq (1+\epsilon)d\right) \leq \exp\left(-k\frac{\epsilon^2}{G_{R,gm}}\right), \quad \epsilon > 0 \tag{5.36}$$

$$\mathbf{Pr}\left(\hat{d}_{gm} \leq (1-\epsilon)d\right) \leq \exp\left(-k\frac{\epsilon^2}{G_{L,gm}}\right), \quad 0 < \epsilon < 1 \tag{5.37}$$

$$G_{R,gm} = \frac{\epsilon^2}{\left(-\frac{1}{2}\log\left(1 + \left(\frac{2}{\pi}\log(1+\epsilon)\right)^2\right) + \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1+\epsilon)\right)\log(1+\epsilon)\right)}, \tag{5.38}$$

$$G_{L,gm} = \frac{\epsilon^2}{\left(-\frac{1}{2}\log\left(1 + \left(\frac{2}{\pi}\log(1-\epsilon)\right)^2\right) + \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1-\epsilon)\right)\log(1-\epsilon)\right)}. \tag{5.39}$$

Moreover, $G_{R,gm} \to \frac{\pi^2}{2}$, *and* $G_{L,gm} \to \frac{\pi^2}{2}$, *as* $\epsilon \to 0+$.

Lemma 16

$$\mathbf{Pr}\left(\hat{d}_{gm,c} \geq (1+\epsilon)d\right) \leq \frac{\cos^{kt_1^*}\left(\frac{\pi}{2k}\right)}{\cos^k\left(\frac{\pi t_1^*}{2k}\right)(1+\epsilon)^{t_1^*}}, \qquad \epsilon \geq 0 \tag{5.40}$$

where

$$t_1^* = \frac{2k}{\pi}\tan^{-1}\left(\left(\log(1+\epsilon) - k\log\cos\left(\frac{\pi}{2k}\right)\right)\frac{2}{\pi}\right). \tag{5.41}$$

$$\mathbf{Pr}\left(\hat{d}_{gm,c} \leq (1-\epsilon)d\right) \leq \frac{(1-\epsilon)^{t_2^*}}{\cos^k\left(\frac{\pi t_2^*}{2k}\right)\cos^{kt_2^*}\left(\frac{\pi}{2k}\right)}, \qquad 0 \leq \epsilon \leq 1, \quad k \geq \frac{\pi^2}{8\epsilon} \tag{5.42}$$

where

$$t_2^* = \frac{2k}{\pi}\tan^{-1}\left(\left(-\log(1-\epsilon) + k\log\cos\left(\frac{\pi}{2k}\right)\right)\frac{2}{\pi}\right). \tag{5.43}$$

By restricting $0 \leq \epsilon \leq 1$, *the tail bounds can be written in exponential forms:*

$$\mathbf{Pr}\left(\hat{d}_{gm,c} \geq (1+\epsilon)d\right) \leq \exp\left(-k\frac{\epsilon^2}{8(1+\epsilon)}\right) \tag{5.44}$$

$$\mathbf{Pr}\left(\hat{d}_{gm,c} \leq (1-\epsilon)d\right) \leq \exp\left(-k\frac{\epsilon^2}{8(1+\epsilon)}\right), \quad k \geq \frac{\pi^2}{1.5\epsilon} \tag{5.45}$$

An analog of the JL bound for l_1 follows from the tail bounds (5.44) and (5.45).

Lemma 17 *Using* $\hat{d}_{gm,c}$ *with* $k \geq \frac{8(2\log n - \log\delta)}{\epsilon^2/(1+\epsilon)} \geq \frac{\pi^2}{1.5\epsilon}$, *then with probability at least* $1-\delta$, *the* l_1 *distance,* d, *between any pair of data points (among* n *data points), can be estimated with errors bounded by* $\pm\epsilon d$, *i.e.,* $|\hat{d}_{gm,c} - d| \leq \epsilon d$.

Remarks on Lemma 17: (1) We can replace the constant "8" in Lemma 17 with better (i.e., smaller) constants for specific values of ϵ. For example, If $\epsilon = 0.2$, we can replace "8" by "5". See the proof of Lemma 16. (2) This Lemma is weaker than the classical JL Lemma for l_2, because the geometric mean is not convex.

5.4 The Sample Median Estimators vs. the Geometric Mean Estimators

While the *sample median* estimators and the *geometric mean* estimators are asymptotically equivalent, in this section, we will compare them in terms of (a) (small sample) mean square errors, (b) tail bound constants, (c) simulated tail probabilities.

5.4.1 Mean Square Errors

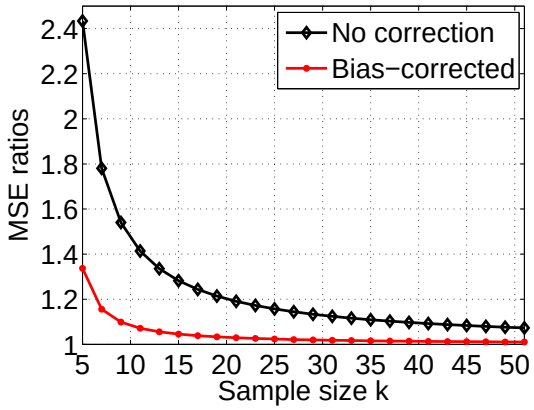
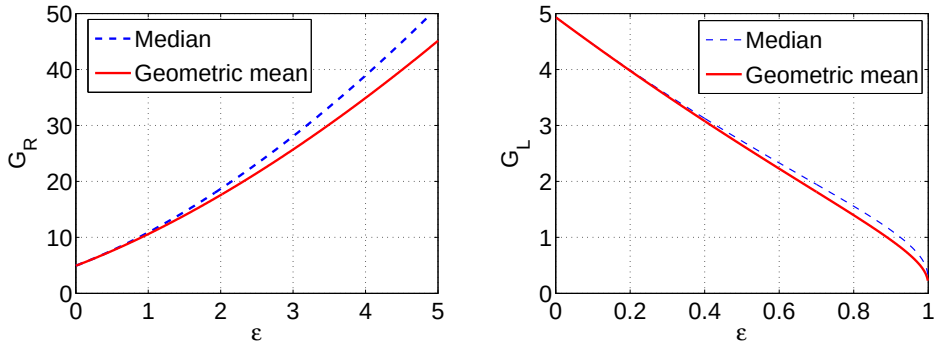


Figure 5.2: The ratios of the MSE, $\frac{\text{MSE}(\hat{d}_{me})}{\text{MSE}(\hat{d}_{gm,c})}$ and $\frac{\text{MSE}(\hat{d}_{me,c})}{\text{MSE}(\hat{d}_{gm,c})}$, demonstrate that the bias-corrected geometric mean estimator $\hat{d}_{gm,c}$ is considerably more accurate than the sample median estimator $\hat{d}_{me}$. The bias correction on $\hat{d}_{me}$ considerably reduces the MSE. When $k = 3$, the ratios are ∞.

Figure 5.2 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 (MSE). $\hat{d}_{gm,c}$ is considerably more accurate than $\hat{d}_{me}$ at small k; and the bias correction significantly reduces the mean square errors of $\hat{d}_{me}$. In terms of the MSE, the difference between $\hat{d}_{me,c}$ and $\hat{d}_{gm,c}$ is negligible after $k > 20$.

5.4.2 Tail Bounds

Figure 5.3 plots the tail bound constants in Lemma 11 and Lemma 15. For the left bound constants, $G_{L,me}$ and $G_{L,gm}$ in Figure 5.3(b), the difference between $G_{L,me}$ and $G_{L,gm}$ is very small, although $G_{L,me}$ is slightly larger than $G_{L,gm}$ when $\epsilon > 0.5$. For the right tail bound constants $G_{R,me}$ and $G_{L,gm}$ in Figure 5.3(a), we can see that the difference between $G_{R,me}$ and $G_{L,gm}$ becomes increasingly obvious as ϵ increases.



(a) $G_{R,me}$ v.s. $G_{R,gm}$ (b) $G_{L,me}$ v.s. $G_{L,gm}$

Figure 5.3: Tail bound constants for the sample median estimator and the geometric mean estimator, in Lemma 11 and Lemma 15 respectively. Smaller constants imply smaller required sample sizes.

The left bound $\mathbf{Pr}\left(\hat{d} - d \leq -\epsilon d\right)$ is often not as interesting because ϵ is limited to be $0 < \epsilon < 1$ and the bound is usually much smaller than the right bound $\mathbf{Pr}\left(\hat{d} - d \geq \epsilon d\right)$, which only requires $\epsilon > 0$. For example, in nearest neighbor problems, often the case only the right bound is considered and ϵ is sometimes taken to be much larger than one [100, 56, 15].

Therefore, we consider the difference between $G_{R,me}$ and $G_{L,gm}$ can be of practical interest. The larger a constant, the more samples are needed. While Figure 5.3(a) shows the difference between $G_{R,me}$ and $G_{R,gm}$ becomes larger as ϵ increases, Lemma 18 shows that $G_{R,me}$ will only be twice as large as $G_{R,gm}$ when $\epsilon \to \infty$.

Lemma 18

$$\lim_{\epsilon \to \infty} \frac{G_{R,me}}{G_{R,gm}} = 2 \tag{5.46}$$

Proof

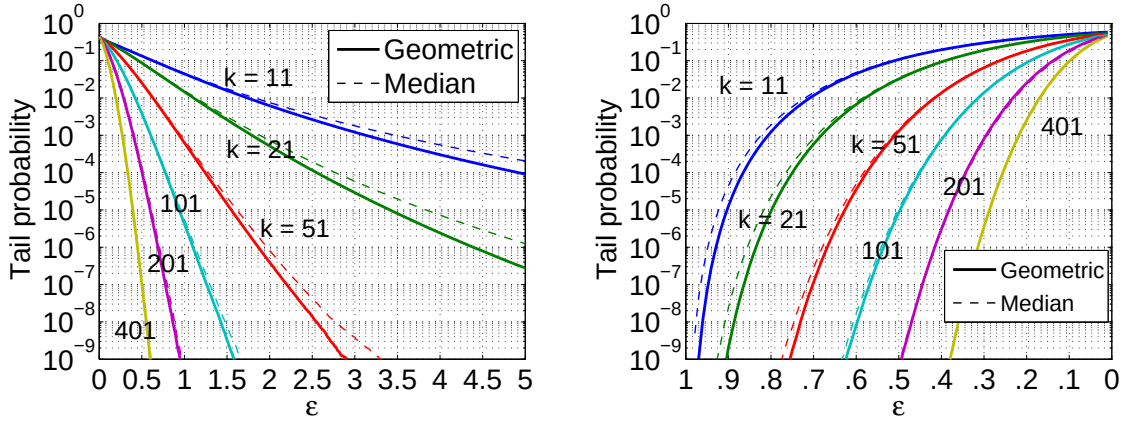
As $\epsilon \to \infty$, $\log(1+\epsilon) \sim \log(\epsilon)$ *and* $\tan^{-1}(\epsilon) \sim \frac{\pi}{2} - \frac{1}{\epsilon}$.

$$\begin{aligned}
\frac{G_{R,me}}{G_{R,gm}} =& \frac{\left(-\log\left(1+\left(\frac{2}{\pi}\log(1+\epsilon)\right)^2\right) + \frac{4}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1+\epsilon)\right)\log(1+\epsilon)\right)}{-\log\left(2-\frac{4}{\pi}\tan^{-1}(1+\epsilon)\right) - \log\left(\frac{4}{\pi}\tan^{-1}(1+\epsilon)\right)} \\
\sim& \frac{-2\log\left(\frac{2}{\pi}\right) - 2\log(\log(\epsilon)) + \frac{4}{\pi}\left(\frac{\pi}{2} - \frac{1}{\frac{\pi}{2}\log(\epsilon)}\right)\log(\epsilon)}{-\log\left(\frac{4}{\pi}\frac{1}{\epsilon}\right) - \log(2)} \\
\sim& \frac{2\log(\epsilon) - 2\log(\log(\epsilon)) - 2\log\left(\frac{2}{\pi}\right) - 2}{\log(\epsilon) - \log\left(\frac{4}{\pi}\right) - \log(2)} \\
\sim& 2\left(1 - \frac{\log\log(\epsilon)}{\log(\epsilon)}\right) \sim 2
\end{aligned}$$

Therefore, in the worst-case, using the *sample median* estimators may require twice as many samples as using the *geometric mean* estimators, according to the tail bound constants. Note that $\frac{G_{R,me}}{G_{R,gm}}$ grows to 2 like $2\left(1 - \frac{\log\log(\epsilon)}{\log(\epsilon)}\right)$ very slowly, unless ϵ is very large.

5.4.3 Simulated Tail Probabilities

Figure 5.4 compares the simulated tail probabilities for the geometric mean estimator $\hat{d}_{gm,c}$ and the sample median estimator $\hat{d}_{me,c}$, showing similar trend as indicated by comparing the tail bound constants. For the right tails, $\hat{d}_{gm,c}$ has smaller tail probabilities than $\hat{d}_{me,c}$; the difference becomes larger as ϵ increases. For the left tails, $\hat{d}_{gm,c}$ has slightly smaller tail probabilities, noticeable when $\epsilon > 0.5$.



(a) Right tail (b) Left tail

Figure 5.4: We simulate the tail probabilities with $k = 11, 21, 51, 101, 201$, and 401, for both $\hat{d}_{me,c}$ and $\hat{d}_{gm,c}$.

5.5 The Maximum Likelihood Estimators

This section is devoted to analyzing the maximum likelihood estimators (MLE), which are asymptotically optimum (in terms of the variances). 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 log joint likelihood of $\{x_j\}_{j=1}^k$ is

$$L(x_1, x_2, ...x_k; d) = k \log(d) - k \log(\pi) - \sum_{j=1}^{k} \log(x_j^2 + d^2), \tag{5.47}$$

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

$$L'(d) = \frac{k}{d} - \sum_{j=1}^{k} \frac{2d}{x_j^2 + d^2}, \tag{5.48}$$

$$L''(d) = -\frac{k}{d^2} - \sum_{j=1}^{k} \frac{2x_j^2 - 2d^2}{(x_j^2 + d^2)^2} = -\frac{L'(d)}{d} - 4 \sum_{j=1}^{k} \frac{x_j^2}{(x_j^2 + d^2)^2}. \tag{5.49}$$

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

$$-\frac{k}{\hat{d}_{MLE}} + \sum_{j=1}^{k} \frac{2\hat{d}_{MLE}}{x_j^2 + \hat{d}_{MLE}^2} = 0. \tag{5.50}$$

Because $L''(\hat{d}_{MLE}) \leq 0$, $\hat{d}_{MLE}$ indeed maximizes the joint likelihood and is the only solution to the MLE equation (5.50). For better accuracy, we recommend the following bias-corrected estimator:

$$\hat{d}_{MLE,c} = \hat{d}_{MLE}\left(1 - \frac{1}{k}\right). \tag{5.51}$$

Proved in Section 5.7.5, Lemma 19 concerns the asymptotic moments of the MLE.

Lemma 19 *Both $\hat{d}_{MLE}$ and $\hat{d}_{MLE,c}$ are asymptotically unbiased and normal. The first four moments of $\hat{d}_{MLE}$ are*

$$E\left(\hat{d}_{MLE} - d\right) = \frac{d}{k} + O\left(\frac{1}{k^2}\right) \tag{5.52}$$

$$Var\left(\hat{d}_{MLE}\right) = \frac{2d^2}{k} + \frac{7d^2}{k^2} + O\left(\frac{1}{k^3}\right) \tag{5.53}$$

$$E\left(\hat{d}_{MLE} - E(\hat{d}_{MLE})\right)^3 = \frac{12d^3}{k^2} + O\left(\frac{1}{k^3}\right) \tag{5.54}$$

$$E\left(\hat{d}_{MLE} - E(\hat{d}_{MLE})\right)^4 = \frac{12d^4}{k^2} + \frac{222d^4}{k^3} + O\left(\frac{1}{k^4}\right) \tag{5.55}$$

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

$$E\left(\hat{d}_{MLE,c} - d\right) = O\left(\frac{1}{k^2}\right) \tag{5.56}$$

$$Var\left(\hat{d}_{MLE,c}\right) = \frac{2d^2}{k} + \frac{3d^2}{k^2} + O\left(\frac{1}{k^3}\right) \tag{5.57}$$

$$E\left(\hat{d}_{MLE,c} - E(\hat{d}_{MLE,c})\right)^3 = \frac{12d^3}{k^2} + O\left(\frac{1}{k^3}\right) \tag{5.58}$$

$$E\left(\hat{d}_{MLE,c} - E(\hat{d}_{MLE,c})\right)^4 = \frac{12d^4}{k^2} + \frac{186d^4}{k^3} + O\left(\frac{1}{k^4}\right) \tag{5.59}$$

The order $O\left(\frac{1}{k}\right)$ term of the variance, i.e., $\frac{2d^2}{k}$, is known, e.g., [87]. We derive the bias-corrected estimator, $\hat{d}_{MLE,c}$, and the higher order moments using stochastic Taylor expansions [23, 154, 73, 54].

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).

5.5.1 A Numerical Example

The maximum likelihood estimators are tested on some MSN 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 5.5 illustrates that the bias correction is effective and these (asymptotic) formulas in Lemma 19 are accurate, especially when $k \geq 20$.

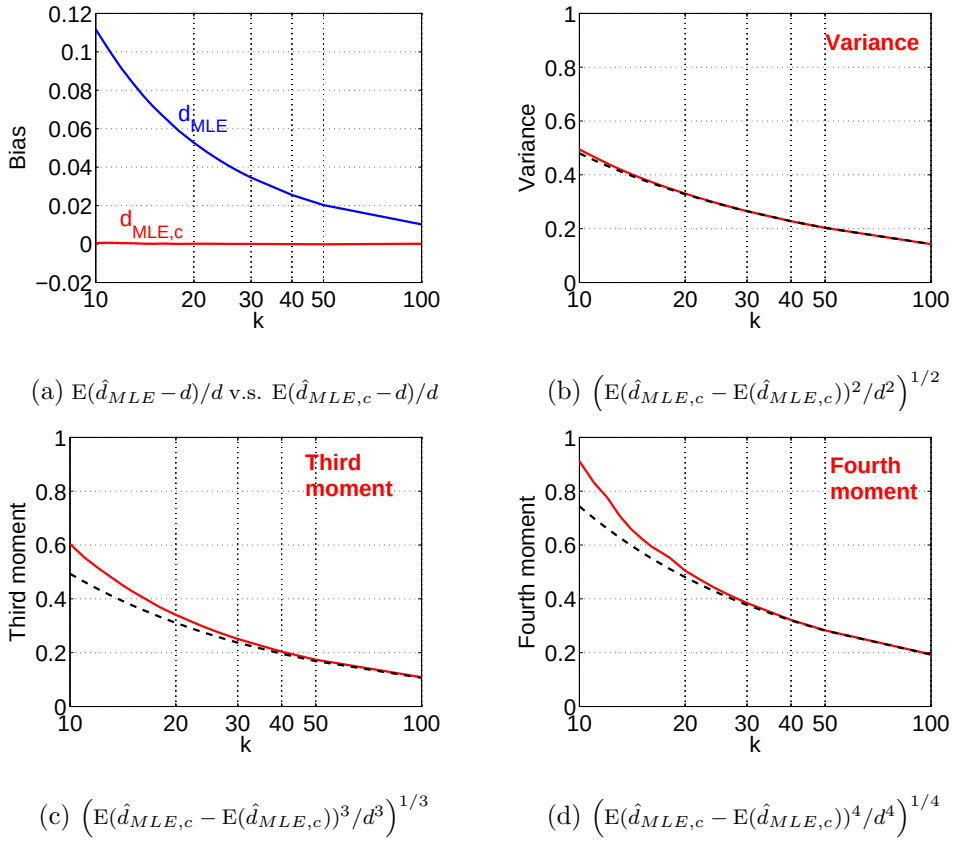
5.5.2 Approximate Distributions

Theoretical analysis on the exact distribution of a maximum likelihood estimator is difficult. It is a common practice to assume normality, which, however, is inaccurate[1]. The *Edgeworth expansion* improves the normal approximation by matching higher moments [70, 25, 153], 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*[102] in general improves Edgeworth expansions, often very considerably. Unfortunately, we can not apply the saddle-point approximation in our case (at least not directly), because the saddle-point approximation needs 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

[1]The simple normal approximation can also be improved by taking advantage of the conditional density on the ancillary configuration statistic, based on the observations x_1, x_2, ..., x_k.[74, 108, 93]



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

(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 5.5: We 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.

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 $\hat{d}_{MLE,c}$, i.e., $[0, \infty)$.

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, 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.

As the related work, [129] applied gamma and generalized gamma approximations to model the performance measure distribution and error probabilities in some wireless communication channels using random matrix theory.

The Gamma Approximation

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

$$\alpha\beta = d, \qquad \alpha\beta^2 = \frac{2d^2}{k} + \frac{3d^2}{k^2}, \quad \Longrightarrow \quad \alpha = \frac{1}{\frac{2}{k} + \frac{3}{k^2}}, \qquad \beta = \frac{2d}{k} + \frac{3d}{k^2}. \tag{5.60}$$

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

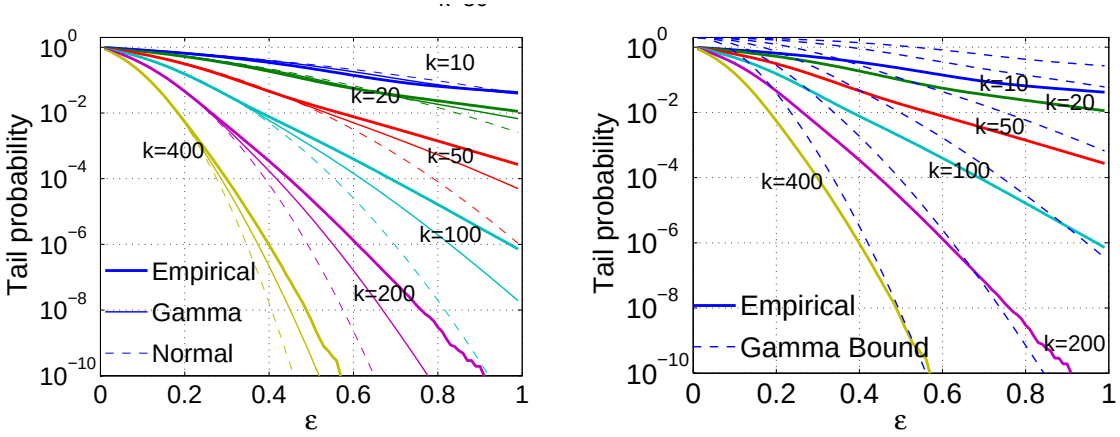
$$\mathbf{Pr}\left(\hat{d}_{MLE,c} \geq (1+\epsilon)d\right) \stackrel{\sim}{\leq} \exp\left(-\alpha\left(\epsilon - \log(1+\epsilon)\right)\right), \quad \epsilon \geq 0 \tag{5.61}$$

$$\mathbf{Pr}\left(\hat{d}_{MLE,c} \leq (1-\epsilon)d\right) \stackrel{\sim}{\leq} \exp\left(-\alpha\left(-\epsilon - \log(1-\epsilon)\right)\right), \quad 0 \leq \epsilon < 1, \tag{5.62}$$

where $\stackrel{\sim}{\leq}$ indicates 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 as shown in [16, 87]. Therefore, we can evaluate the accuracy of the gamma approximation by simulations with $d = 1$, as presented in Figure 5.6.

Figure 5.6(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 5.6(b) compares the empirical tail probabilities with the gamma Chernoff upper bound (5.61)+(5.62), indicating that these bounds are reliable, when the tail probability $\geq 10^{-5} \sim 10^{-6}$.



(a) (b)

Figure 5.6: We consider k from 10 to 400. For each k, we simulate standard Cauchy samples, from which we estimate the Cauchy parameter by the MLE $\hat{d}_{MLE,c}$. 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 (5.61)+(5.62).

The Inverse Gaussian (Generalized Gamma) Approximation

The distribution of $\hat{d}_{MLE,c}$ can be well approximated by an inverse Gaussian, which is a special case of the three-parameter generalized gamma distribution[95, 82, 129], denoted by $GG(\alpha, \beta, \eta)$. (The usual gamma distribution is a special case with $\eta = 1$.)

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

$$\mathrm{E}(z) = \alpha\beta, \quad \mathrm{Var}(z) = \alpha\beta^2, \quad \mathrm{E}\left(z - \mathrm{E}(z)\right)^3 = \alpha\beta^3(1+\eta). \tag{5.63}$$

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

$$\alpha\beta = d, \quad \alpha\beta^2 = \frac{2d^2}{k} + \frac{3d^2}{k^2}, \quad \alpha\beta^3(1+\eta) = \frac{12d^3}{k^2}, \tag{5.64}$$

from which we obtain

$$\alpha = \frac{1}{\frac{2}{k} + \frac{3}{k^2}}, \quad \beta = \frac{2d}{k} + \frac{3d}{k^2}, \quad \eta = 2 + O\left(\frac{1}{k}\right). \tag{5.65}$$

Taking $\eta = 2$, the generalized gamma approximation of $\hat{d}_{MLE,c}$ would be

$$GG\left(\frac{1}{\frac{2}{k} + \frac{3}{k^2}}, \frac{2d}{k} + \frac{3d}{k^2}, 2\right). \tag{5.66}$$

Although in general a generalized gamma distribution does not have a closed-form density, in our case, (5.66) is actually an inverse Gaussian (IG) distribution with a closed-form density function. Assuming $\hat{d}_{MLE,c} \sim IG(\alpha, \beta)$, with parameters α and β defined in (5.65), the moment generating function (MGF), the probability density function (PDF), and cumulative density function (CDF) would be [163, 164, 47, 152]

$$\mathrm{E}\left(\exp(\hat{d}_{MLE,c}t)\right) \cong \exp\left(\alpha\left(1 - (1 - 2\beta t)^{1/2}\right)\right), \tag{5.67}$$

$$f_{\hat{d}_{MLE,c}}(y) \cong \sqrt{\frac{\alpha d}{2\pi}} y^{-\frac{3}{2}} \exp\left(-\frac{(y-d)^2}{2y\beta}\right), \tag{5.68}$$

$$\mathbf{Pr}\left(\hat{d}_{MLE,c} \leq y\right) \cong \Phi\left(\sqrt{\frac{\alpha d}{y}}\left(\frac{y}{d} - 1\right)\right) + e^{2\alpha}\Phi\left(-\sqrt{\frac{\alpha d}{y}}\left(\frac{y}{d} + 1\right)\right), \tag{5.69}$$

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

$$\begin{aligned}
\mathrm{E}\left(\hat{d}_{MLE,c} - \mathrm{E}\left(\hat{d}_{MLE,c}\right)\right)^4 &\cong 15\alpha\beta^4 + 3\left(\alpha\beta^2\right)^2 \\
&= 15d\left(\frac{2d}{k} + \frac{3d}{k^2}\right)^3 + 3\left(\frac{2d^2}{k} + \frac{3d^2}{k^2}\right)^2 \\
&= \frac{12d^4}{k^2} + \frac{156d^4}{k^3} + O\left(\frac{1}{k^4}\right).
\end{aligned} \tag{5.70}$$

Lemma 19 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.

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

$$\mathbf{Pr}\left(\hat{d}_{MLE,c} \geq (1+\epsilon)d\right) \cong \Phi\left(-\epsilon\sqrt{\frac{\alpha}{1+\epsilon}}\right) - e^{2\alpha}\Phi\left(-(2+\epsilon)\sqrt{\frac{\alpha}{1+\epsilon}}\right),\ \epsilon \geq 0 \tag{5.71}$$

$$\mathbf{Pr}\left(\hat{d}_{MLE,c} \leq (1-\epsilon)d\right) \cong \Phi\left(-\epsilon\sqrt{\frac{\alpha}{1-\epsilon}}\right) + e^{2\alpha}\Phi\left(-(2-\epsilon)\sqrt{\frac{\alpha}{1-\epsilon}}\right),\ 0 \leq \epsilon < 1. \tag{5.72}$$

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

$$\mathbf{Pr}\left(\hat{d}_{MLE,c} \geq (1+\epsilon)d\right) \lesssim \exp\left(-\frac{\alpha\epsilon^2}{2(1+\epsilon)}\right), \quad \epsilon \geq 0 \tag{5.73}$$

$$\mathbf{Pr}\left(\hat{d}_{MLE,c} \leq (1-\epsilon)d\right) \lesssim \exp\left(-\frac{\alpha\epsilon^2}{2(1-\epsilon)}\right), \quad 0 \leq \epsilon < 1. \tag{5.74}$$

To see (5.73), assume $z \sim IG(\alpha, \beta)$. Then, using the Chernoff inequality:

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

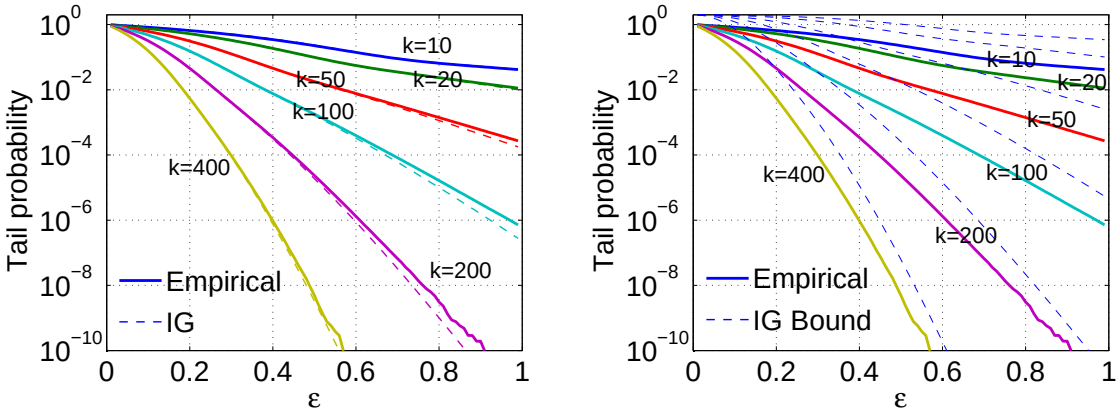
whose minimum is $\exp\left(-\frac{\alpha\epsilon^2}{2(1+\epsilon)}\right)$, attained at $t = \left(1 - \frac{1}{(1+\epsilon)^2}\right)\frac{1}{2\beta}$.

Combining (5.73) and (5.74) yields a symmetric bound

$$\mathbf{Pr}\left(\left|\hat{d}_{MLE,c} - d\right| \geq \epsilon d\right) \lesssim 2\exp\left(-\frac{\epsilon^2/(1+\epsilon)}{2\left(\frac{2}{k} + \frac{3}{k^2}\right)}\right), \quad 0 \leq \epsilon \leq 1 \tag{5.75}$$

Figure 5.7 compares the inverse Gaussian approximation with the same simulations as presented in Figure 5.6, 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 (tail probability $\geq 10^{-10}$).



(a) (b)

Figure 5.7: We compare the inverse Gaussian approximation with the same simulations in Figure 5.6. Panel (a) indicates that empirical tail probabilities can be highly accurately approximated by the inverse Gaussian tail probabilities. Panel (b) compares the empirical tail probabilities with the inverse Gaussian upper bound (5.73)+(5.74), indicating that they are reliable at least in our simulation range.

5.6 Summary

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*. We study 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.

The sample median and the geometric mean estimators are asymptotically equivalent but the latter is more accurate at small sample size. We have derived explicit tail bounds for the both the sample median and the geometric mean estimators, from

which 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.

Both the sample median and the 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 are provided, which, verified by simulations, hold at least in the $\geq 10^{-10}$ tail probability range.

5.7 Proofs

5.7.1 Proof of Lemma 11

$x_j \sim C(0,d)$, $j=1$ to k. Denote by $F(t;d)$ the cumulative density of $|C(0,d)|$, and by $F_k(t;d)$ the empirical cumulative density of $|x_j|$, $j=1$ to k. $kF_k(t;d)$ follows a binomial distribution, i.e., $kF_k(t;d) \sim Bin(k, F(t;d))$.

By the well-known binomial Chernoff bounds [46][2], we have

$$
\begin{aligned}
&\mathbf{Pr}\left(kF_k(t;d) \geq (1+\epsilon')kF(t;d)\right) \\
&\leq \left(\frac{k-kF(t;d)}{k-(1+\epsilon')kF(t;d)}\right)^{k-k(1+\epsilon')F(t;d)} \left(\frac{kF(t;d)}{(1+\epsilon')kF(t;d)}\right)^{(1+\epsilon')kF(t;d)} \\
&= \left[\left(\frac{1-F(t;d)}{1-(1+\epsilon')F(t;d)}\right)^{1-(1+\epsilon')F(t;d)} \left(\frac{1}{1+\epsilon'}\right)^{(1+\epsilon')F(t;d)}\right]^k,
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbf{Pr}\left(kF_k(t;d) \leq (1-\epsilon')kF(t;d)\right) \\
&\leq \left[\left(\frac{1-F(t;d)}{1-(1-\epsilon')F(t;d)}\right)^{1-(1-\epsilon')F(t;d)} \left(\frac{1}{1-\epsilon'}\right)^{(1-\epsilon')F(t;d)}\right]^k.
\end{aligned}
$$

[2]See the binomial Chernoff bounds in simpler forms in [140, Theorem 4.1, Theorem 4.2]. We have to use the more sophisticated formulations in order to achieve the optimal rate for the tail bounds in our problem.

For $\epsilon > 0$,

$$\mathbf{Pr}\left(\hat{d}_{me} \geq (1+\epsilon)d\right) = \mathbf{Pr}\left(kF_k\left((1+\epsilon)d; d\right) \leq \frac{k}{2}\right) = \mathbf{Pr}\left(kF_k(t;d) \leq (1-\epsilon')kF(t;d)\right),$$

where we let $t = (1+\epsilon)d$ and $\frac{1}{2} = (1-\epsilon')F(t;d) = (1-\epsilon')\frac{2}{\pi}\tan^{-1}(1+\epsilon)$. Thus, we can obtain

$$\begin{aligned}
&\mathbf{Pr}\left(\hat{d}_{me} \geq (1+\epsilon)d\right) \\
\leq& \left[(2 - 2F((1+\epsilon)d;d))^{1/2}(2F((1+\epsilon)d;d))^{1/2}\right]^k \\
=& \exp\left(-\frac{k}{2}\left(-\log\left(2 - 2F((1+\epsilon)d;d)\right) - \log(2F((1+\epsilon)d;d))\right)\right) \\
=& \exp\left(-\frac{k}{2}\left(-\log\left(2 - \frac{4}{\pi}\tan^{-1}(1+\epsilon)\right) - \log\left(\frac{4}{\pi}\tan^{-1}(1+\epsilon)\right)\right)\right).
\end{aligned}$$

Similarly, for $0 < \epsilon < 1$,

$$\mathbf{Pr}\left(\hat{d}_{me} \leq (1-\epsilon)d\right) = \mathbf{Pr}\left(kF_k\left((1-\epsilon)d; d\right) \geq \frac{k}{2}\right) = \mathbf{Pr}\left(kF_k(t;d) \geq (1+\epsilon')kF(t;d)\right),$$

where we let $t = (1-\epsilon)d$ and $\frac{1}{2} = (1+\epsilon')F(t;d) = (1+\epsilon')\frac{2}{\pi}\tan^{-1}(1-\epsilon)$. Thus, we can obtain

$$\mathbf{Pr}\left(\hat{d}_{me} \leq (1-\epsilon)d\right) \leq \exp\left(-\frac{k}{2}\left(-\log\left(2 - \frac{4}{\pi}\tan^{-1}(1-\epsilon)\right) - \log\left(\frac{4}{\pi}\tan^{-1}(1-\epsilon)\right)\right)\right).$$

$$\begin{aligned}
\lim_{\epsilon\to 0+} G_{R,me} =& \lim_{\epsilon\to 0+} \frac{2\epsilon^2}{-\log\left(2 - \frac{4}{\pi}\tan^{-1}(1+\epsilon)\right) - \log\left(\frac{4}{\pi}\tan^{-1}(1+\epsilon)\right)} \\
=& \lim_{\epsilon\to 0+} \frac{4\epsilon(1+(1+\epsilon)^2)}{\frac{1}{\frac{\pi}{2}-\tan^{-1}(1+\epsilon)} - \frac{1}{\tan^{-1}(1+\epsilon)}} = \lim_{\epsilon\to 0+} \frac{4(3\epsilon^2+2\epsilon+2)}{\frac{\frac{1}{1+(1+\epsilon)^2}}{\left(\frac{\pi}{2}+\tan^{-1}(1+\epsilon)\right)^2} - \frac{\frac{1}{1+(1+\epsilon)^2}}{\left(\tan^{-1}(1+\epsilon)\right)^2}} \\
=& 8\left(\frac{\pi}{4}\right)^2 = \frac{\pi^2}{2}.
\end{aligned}$$

5.7.2 Proof of Lemma 14

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 13,

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

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:

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

5.7.3 Proof of Lemma 15

We will use the Markov moment bound, because $\hat{d}_{gm}$ does not have a moment generating function ($\mathrm{E}\left(\hat{d}_{gm}\right)^t = \infty$ 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[145, 133].

By the Markov moment bound, for any $\epsilon > 0$,

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

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

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

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

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

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

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

It is easy to show $G_{R,gm} \to \frac{\pi^2}{2}$, and $G_{L,gm} \to \frac{\pi^2}{2}$, as $\epsilon \to 0+$.

5.7.4 Proof of Lemma 16

For any $\epsilon \geq 0$ and $0 \leq t < k$, Markov's inequality says

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

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

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

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

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

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

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

$$\begin{aligned}
\mathbf{Pr}\left(\hat{d}_{gm,c} \geq (1+\epsilon)d\right) &\leq \frac{\cos^{kt_1^{**}}\left(\frac{\pi}{2k}\right)}{\cos^k\left(\frac{\pi t_1^{**}}{2k}\right)(1+\epsilon)^{t_1^{**}}} = \left(\frac{\cos^{t_1^{**}}\left(\frac{\pi}{2k}\right)}{\cos\left(\frac{2\epsilon}{\pi}\right)(1+\epsilon)^{4\epsilon/\pi^2}}\right)^k \\
&\leq \left(\frac{1}{\cos\left(\frac{2\epsilon}{\pi}\right)(1+\epsilon)^{4\epsilon/\pi^2}}\right)^k \\
&= \exp\left(-k\left(\log\left(\cos\left(\frac{2\epsilon}{\pi}\right)\right) + \frac{4\epsilon}{\pi^2}\log(1+\epsilon)\right)\right) \\
&\leq \exp\left(-k\frac{\epsilon^2}{8(1+\epsilon)}\right), \quad 0 \leq \epsilon \leq 1 \qquad (5.76)
\end{aligned}$$

The last step in (5.76) needs some explanations. First, by the Taylor expansion,

$$\log\left(\cos\left(\frac{2\epsilon}{\pi}\right)\right)+\frac{4\epsilon}{\pi^2}\log(1+\epsilon)=\left(-\frac{2\epsilon^2}{\pi^2}-\frac{4}{3}\frac{\epsilon^4}{\pi^4}+...\right)+\frac{4\epsilon}{\pi^2}\left(\epsilon-\frac{1}{2}\epsilon^2+...\right)$$
$$=\frac{2\epsilon^2}{\pi^2}\left(1-\epsilon+...\right)$$

Therefore, we can seek the smallest constant γ_1 so that

$$\log\left(\cos\left(\frac{2\epsilon}{\pi}\right)\right)+\frac{4\epsilon}{\pi^2}\log(1+\epsilon)\geq\frac{\epsilon^2}{\gamma_1(1+\epsilon)}=\frac{\epsilon^2}{\gamma_1}(1-\epsilon+...)$$

It is easy to see that as $\epsilon\to 0$, $\gamma_1\to\frac{\pi^2}{2}$. Figure 5.8(a) illustrates that it suffices to let $\gamma_1=8$, which can be numerically verified. This is why the last step in (5.76) holds. Of course, we can get a better constant if (e.g.,) $\epsilon=0.5$.

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

$$\mathbf{Pr}\left(\hat{d}_{gm,c}\leq(1-\epsilon)d\right)=\mathbf{Pr}\left(\cos\left(\frac{\pi}{2k}\right)^k\prod_{j=1}^k|x_j|^{1/k}\leq(1-\epsilon)d\right)$$
$$=\mathbf{Pr}\left(\prod_{j=1}^k|x_j|^{-t/k}\geq\left(\frac{(1-\epsilon)d}{\cos^k\left(\frac{\pi}{2k}\right)}\right)^{-t}\right)\leq\left(\frac{(1-\epsilon)}{\cos^k\left(\frac{\pi}{2k}\right)}\right)^t\frac{1}{\cos^k\left(\frac{\pi t}{2k}\right)},$$

which is minimized at $t=t_2^*$

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

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

Again, t_2^* can be replaced by its approximation

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

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

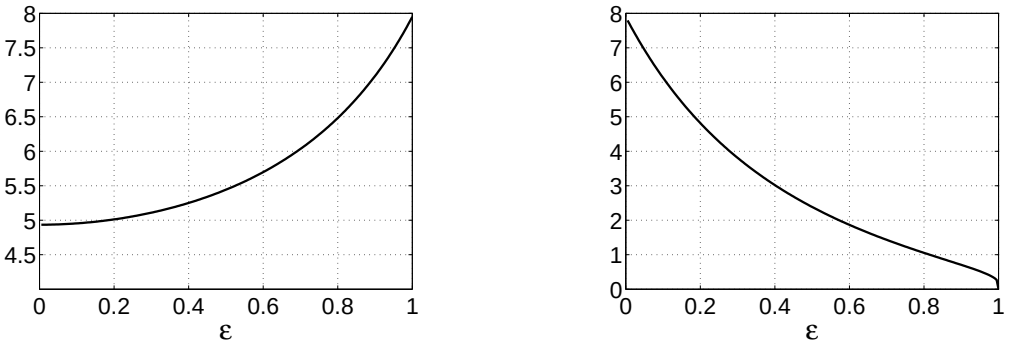
$$\begin{aligned}\mathbf{Pr}\left(\hat{d}_{gm,c} \leq (1-\epsilon)d\right) &\leq \left(\frac{(1-\epsilon)}{\cos^k\left(\frac{\pi}{2k}\right)}\right)^{t_2^{**}} \frac{1}{\cos^k\left(\frac{\pi t_2^{**}}{2k}\right)} \\ &= \exp\left(-k\left(\log\left(\cos\frac{2\epsilon}{\pi}\right) - \frac{4\epsilon}{\pi^2}\log(1-\epsilon) + \frac{4k\epsilon}{\pi^2}\log\left(\cos\frac{\pi}{2k}\right)\right)\right).\end{aligned}$$

We can bound $\frac{4k\epsilon}{\pi^2}\log\left(\cos\frac{\pi}{2k}\right)$ by restricting k.

In order to attain $\mathbf{Pr}\left(\hat{d}_{gm,c} \leq (1-\epsilon)d\right) \leq \exp\left(-k\left(\frac{\epsilon^2}{8(1+\epsilon)}\right)\right)$, we have to restrict k. We seek $k \geq \frac{\pi^2}{\gamma_2\epsilon}$, for some constant γ_2, and find $k \geq \frac{\pi^2}{1.5\epsilon}$ suffices, although readers can verify that a slightly better (smaller) restriction would be $k \geq \frac{1}{4/\pi^2 - 1/4}\frac{1}{\epsilon} = \frac{\pi^2}{1.5326\epsilon}$.

If $k \geq \frac{\pi^2}{1.5\epsilon}$, then $\frac{4k\epsilon}{\pi^2}\log\left(\cos\frac{\pi}{2k}\right) \geq \frac{8}{3}\log\left(\cos\frac{\epsilon}{3\pi}\right)$. Therefore,

$$\begin{aligned}\mathbf{Pr}\left(\hat{d}_{gm,c} \leq (1-\epsilon)d\right) &\leq \exp\left(-k\left(\log\left(\cos\frac{2\epsilon}{\pi}\right) - \frac{4\epsilon}{\pi^2}\log(1-\epsilon) + \frac{8}{3}\log\left(\cos\frac{\epsilon}{3\pi}\right)\right)\right) \\ &\leq \exp\left(-k\frac{\epsilon^2}{8(1+\epsilon)}\right), \quad k \geq \frac{\pi^2}{1.5\epsilon} \qquad (5.77)\end{aligned}$$



(a) (b)

Figure 5.8: (a): $\frac{\epsilon^2/(1+\epsilon)}{\log\left(\cos\left(\frac{2\epsilon}{\pi}\right)\right)+\frac{4\epsilon}{\pi^2}\log(1+\epsilon)}$. (b): $\frac{\epsilon^2/(1+\epsilon)}{\log\left(\cos\frac{2\epsilon}{\pi}\right)-\frac{4\epsilon}{\pi^2}\log(1-\epsilon)+\frac{8}{3}\log\left(\cos\frac{\epsilon}{3\pi}\right)}$. It suffices to use a constant 8 in (5.76) and (5.77). The optimal constant will be different for different ϵ. For example, if $\epsilon = 0.2$, we could replace the constant 8 by a constant 5.

5.7.5 Proof of Lemma 19

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[3] $\hat{d}_{MLE}$ 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 [23, 154]. We need to evaluate the expressions in [154, 16a-16d], involving tedious algebra:

[3] [64] proved the regularity condition for ensuring the asymptotic normality.

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

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:

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

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

$$
[12] = -\frac{1}{2}\frac{1}{d^3}, \quad [1^4] = \frac{3}{8}\frac{1}{d^4}, \quad [1^22] = \frac{1}{4}\frac{1}{d^4}, \quad [13] = \frac{3}{4}\frac{1}{d^4}, \quad [1^3] = 0.
$$

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, after some algebra, are

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

Chapter 6

Stable Random Projections for l_α

We have discussed random projections for the l_2 norm in Chapter 3 and Chapter 4, and for the l_1 norm in Chapter 5. This chapter will discuss dimension reductions in the l_α norm, for $0 < \alpha \leq 2$, including l_2 and l_1 norms as special cases.

The fundamental problem in *stable random projections* is the statistical estimation task, i.e., estimating the scale parameter of a symmetric stable distribution. Because the stable distribution does not have a closed-form density except for $\alpha = 1$ or 2, it is an interesting task to develop estimators, that are both statistically accurate and computationally efficient.

Estimators based on sample medians (or more generally, on sample quantiles) are well-known in statistics, but are not very accurate especially at small samples, and are not convenient for theoretical analysis such as tail bounds when $\alpha \neq 1$ or 2.

We will discuss various estimators based on the geometric mean, the harmonic mean, and the fractional power.

Part of the results in this chapter appeared in a technical report[116]. Section 6.7 appeared in a conference proceeding[117].

6.1 Main Results

Recall that, given two vectors u_1, $u_2 \in \mathbb{R}^D$ (e.g., u_1 and u_2 are the leading two rows in the data matrix $\mathbf{A}$), if $v_1 = \mathbf{R}^{\mathrm{T}} u_1$ and $v_2 = \mathbf{R}^{\mathrm{T}} u_2$, where $\mathbf{R} \in \mathbb{R}^{D \times k}$ consists of

i.i.d. samples in $S(\alpha, 1)$, then $x_j = v_{1,j} - v_{2,j}$, $j = 1, 2, ..., k$ are i.i.d. $S(\alpha, d_{(\alpha)})$, where $d_{(\alpha)} = \sum_{i=1}^{D} |u_{1,i} - u_{2,i}|^\alpha$. Thus, the key task is to estimate the scale parameter $d_{(\alpha)}$ from k i.i.d. samples $S(\alpha, d_{(\alpha)})$. In Chapter 2, we briefly reviewed stable distributions.

The widely used estimator in statistics is based on the sample *inter-quantiles*[67, 68, 138], which can be simplified to be the *sample median* estimator due to the symmetry of $S(\alpha, d_{(\alpha)})$,

$$\hat{d}_{(\alpha),me} = \frac{\text{median}\{|x_j|^\alpha, j = 1, 2, ..., k\}}{\text{median}\{S(\alpha, 1)\}^\alpha}. \tag{6.1}$$

Despite its simplicity, there are several problems with the *sample median* estimator $\hat{d}_{(\alpha),me}$. It is not accurate especially for small samples or small α. It is also difficult for precise theoretical analysis analysis such as tail bounds.

We provide several estimators based on the *geometric mean*, the *harmonic mean*, and the *fractional power*.

- The (unbiased) geometric mean estimator $\hat{d}_{(\alpha),gm}$,

$$\hat{d}_{(\alpha),gm} = \frac{\prod_{j=1}^{k} |x_j|^{\alpha/k}}{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1 - \frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^k}. \tag{6.2}$$

- The biased geometric mean estimator $\hat{d}_{(\alpha),gm,b}$

$$\hat{d}_{(\alpha),gm,b} = \frac{\prod_{j=1}^{k} |x_j|^{\alpha/k}}{\left[\frac{2}{\pi}\Gamma\left(\frac{1}{k}\right)\Gamma\left(1 - \frac{1}{k\alpha}\right)\sin\left(\frac{\pi}{2}\frac{1}{k}\right)\right]^{\alpha k}}, \tag{6.3}$$

  is asymptotically equivalent to $\hat{d}_{(\alpha),gm}$. However, when $0.25 \leq \alpha \leq 1$, $\hat{d}_{(\alpha),gm,b}$ exhibits smaller mean square errors than $\hat{d}_{(\alpha),gm}$

- The harmonic mean estimator $\hat{d}_{(\alpha),hm}$

$$\hat{d}_{(\alpha),hm} = \frac{-\frac{2}{\pi}\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)}{\sum_{j=1}^{k} |x_j|^{-\alpha}} \left(k - \left(\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2} - 1\right)\right), \tag{6.4}$$

is asymptotically optimal when $\alpha \to 0+$ and has smaller asymptotic variance than the *geometric mean* estimators when $\alpha \leq 0.344$.

- The arithmetic mean estimator

  When $\alpha = 2$, one of course should use the *arithmetic mean* estimator $\frac{1}{k}\sum_{j=1}^{k}|x_j|^2$, which can be further improved by the maximum likelihood estimator taking advantage of the marginal information as discussed in Chapter 3.

- The fractional power estimator $\hat{d}_{(\alpha),fp}$

$$\hat{d}_{(\alpha),fp} = \left(\frac{1}{k}\frac{\sum_{j=1}^{k}|x_j|^{\lambda^*\alpha}}{\frac{2}{\pi}\Gamma(1-\lambda^*)\Gamma(\lambda^*\alpha)\sin\left(\frac{\pi}{2}\lambda^*\alpha\right)}\right)^{1/\lambda^*} \times \left(1 - \frac{1}{k}\frac{1}{2\lambda^*}\left(\frac{1}{\lambda^*}-1\right)\left(\frac{\frac{2}{\pi}\Gamma(1-2\lambda^*)\Gamma(2\lambda^*\alpha)\sin\left(\pi\lambda^*\alpha\right)}{\left[\frac{2}{\pi}\Gamma(1-\lambda^*)\Gamma(\lambda^*\alpha)\sin\left(\frac{\pi}{2}\lambda^*\alpha\right)\right]^2}-1\right)\right), \quad (6.5)$$

  where

$$\lambda^* = \underset{-\frac{1}{2\alpha}<\lambda<\frac{1}{2}}{\text{argmin}} \frac{1}{\lambda^2}\left(\frac{\frac{2}{\pi}\Gamma(1-2\lambda)\Gamma(2\lambda\alpha)\sin\left(\pi\lambda\alpha\right)}{\left[\frac{2}{\pi}\Gamma(1-\lambda)\Gamma(\lambda\alpha)\sin\left(\frac{\pi}{2}\lambda\alpha\right)\right]^2}-1\right). \quad (6.6)$$

  When $\alpha = 2$, $\hat{d}_{(\alpha),fp}$ becomes the *arithmetic mean* estimator; and when $\alpha \to 0+$, $\hat{d}_{(\alpha),fp}$ becomes the *harmonic mean* estimator. Also, when $\alpha \to 1$, $\hat{d}_{(\alpha),fp}$ has the same asymptotic variance as the *geometric mean* estimator.

6.2 The Geometric Mean Estimators

The geometric mean estimators are convenient for theoretical analysis. For a certain range of α (e.g., $0.6 < \alpha < 1.1$), the geometric mean estimators can be close to be optimal ($\geq 80\%$ efficient).

We will first discuss the unbiased geometric mean estimator, including its variance and moments and its tail bounds in exponential forms.

The biased geometric mean estimator has smaller mean square errors in a certain range of α (e.g., $0.25 < \alpha \leq 1$), for finite k.

The geometric mean estimators are based the following result (Mellin Transform) about the stable distribution distribution $S(\alpha, d_{(\alpha)})$, in Lemma 20.

Lemma 20 *Suppose $x \sim S\left(\alpha, d_{(\alpha)}\right)$. Then for $-1 < \lambda \leq \alpha$,*

$$E\left(|x|^\lambda\right) = d_{(\alpha)}^{\lambda/\alpha} \frac{2}{\pi} \Gamma\left(1 - \frac{\lambda}{\alpha}\right) \Gamma(\lambda) \sin\left(\frac{\pi}{2}\lambda\right). \tag{6.7}$$

If $\alpha = 2$, i.e., $x \sim S(2, d_{(2)}) = N(0, 2d_{(2)})$, then for $\lambda > -1$,

$$E\left(|x|^\lambda\right) = d_{(2)}^{\lambda/2} \frac{2}{\pi} \Gamma\left(1 - \frac{\lambda}{2}\right) \Gamma(\lambda) \sin\left(\frac{\pi}{2}\lambda\right) = d_{(2)}^{\lambda/2} \frac{2\Gamma(\lambda)}{\Gamma\left(\frac{\lambda}{2}\right)}. \tag{6.8}$$

Proof: *For $0 < \alpha < 2$ and $-1 < \lambda < \alpha$, (6.7) follows from [171, Theorem 2.6.3], after some algebra.*

For the sake of verification, we can derive $E\left(|x|^\lambda\right)$ for $0 \leq \lambda < \alpha$, by completing the result in [151, Property 1.2.17], which says

$$E\left(|x|^\lambda\right) = d_{(\alpha)}^{\lambda/\alpha} \frac{2^{\lambda-1}\Gamma\left(1 - \frac{\lambda}{\alpha}\right)}{\lambda \int_0^\infty \frac{\sin^2 u}{u^{\lambda+1}} du}, \qquad 0 \leq \lambda < \alpha. \tag{6.9}$$

With the help of [85, 3.823], we obtain

$$\int_0^\infty \frac{\sin^2 u}{u^{\lambda+1}} du = -\frac{\Gamma(-\lambda)\cos\left(\frac{\pi}{2}\lambda\right)}{2^{-\lambda+1}} \tag{6.10}$$

By Euler's reflection formula, $\Gamma(1-z)\Gamma(z) = \frac{\pi}{\sin(\pi z)}$, we know $\Gamma(-\lambda) = -\frac{\pi}{\sin(\pi\lambda)} \frac{1}{\Gamma(\lambda+1)}$, and the desired expression follows after some algebra.

For $\alpha = 2$, the moment $E\left(|x|^\lambda\right)$ exists for any $\lambda > -1$. (6.8) can be found by directly integrating the Gaussian density (using the integral formula [85, 3.381.4]). The duplication formula $\Gamma(z)\Gamma\left(z + \frac{1}{2}\right) = 2^{1-2z}\sqrt{\pi}\Gamma(2z)$ will help the algebra.

Given k i.i.d. samples $x_j \sim S(\alpha, d_{(\alpha)})$, we can design an unbiased estimator of

$d_{(\alpha)}$ by plugging $\lambda = \alpha/k$ into (6.7), i.e.,

$$\hat{d}_{(\alpha),gm} = \frac{\prod_{j=1}^{k} |x_j|^{\alpha/k}}{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^k}, \quad k > 1. \tag{6.11}$$

6.2.1 The Moments and Tail bounds of $\hat{d}_{(\alpha),gm}$

Lemma 21 presents some theoretical properties of the moments for the unbiased geometric mean estimator $\hat{d}_{(\alpha),gm}$. See the proof in Section 6.9.1.

Lemma 21 *The estimator, $\hat{d}_{(\alpha),gm}$, as defined in (6.11):*

- *It is unbiased, i.e., $E\left(\hat{d}_{(\alpha),gm}\right) = d_{(\alpha)}$.*
- *As $k \to \infty$,*

$$\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^k \to \exp\left(-\gamma_e\left(\alpha-1\right)\right), \tag{6.12}$$

  where $\gamma_e = 0.577215665...$, is Euler's constant. It converges from above monotonically.

- *As $k \to \infty$, for fixed t,*

$$\begin{aligned} & E\left(\hat{d}_{(\alpha),gm}\right)^t \\ =& d_{(\alpha)}^t \frac{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha t}{k}\right)\Gamma\left(1-\frac{t}{k}\right)\sin\left(\frac{\pi\alpha t}{2k}\right)\right]^k}{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^{kt}} \qquad (-k/\alpha < t < k) \qquad (6.13) \\ =& d_{(\alpha)}^t \exp\left(\frac{1}{k}\frac{\pi^2}{6}\frac{(t^2-t)(2+\alpha^2)}{4} + \frac{1}{k^2}\zeta_3\frac{(t^3-t)(1-\alpha^3)}{3} + O\left(\frac{1}{k^3}\right)\right), \qquad (6.14) \end{aligned}$$

  where $\zeta_3 = \sum_{s=1}^{\infty}\frac{1}{s^3} = 1.2020569...$ is Apery's constant.

- *The variance, third and fourth central moments of* $\hat{d}_{(\alpha),gm}$ *are*

$$Var\left(\hat{d}_{(\alpha),gm}\right)$$

$$= d_{(\alpha)}^2 \left\{ \frac{\left[\frac{2}{\pi}\Gamma\left(\frac{2\alpha}{k}\right)\Gamma\left(1-\frac{2}{k}\right)\sin\left(\pi\frac{\alpha}{k}\right)\right]^k}{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^{2k}} - 1 \right\} \tag{6.15}$$

$$= d_{(\alpha)}^2 \left\{ \frac{1}{k}\frac{\pi^2}{12}\left(\alpha^2+2\right) + \frac{1}{k^2}\left(2(1-\alpha^3)\zeta_3 + \frac{\pi^4}{288}(\alpha^2+2)^2\right) + O\left(\frac{1}{k^3}\right) \right\}. \tag{6.16}$$

$$E\left(\hat{d}_{(\alpha),gm} - d_{(\alpha)}\right)^3 = \frac{d_{(\alpha)}^3}{k^2}\left\{\frac{\pi^4}{48}(2+\alpha^2)^2 + 2(1-\alpha^3)\zeta_3\right\} + O\left(\frac{1}{k^3}\right) \tag{6.17}$$

$$E\left(\hat{d}_{(\alpha),gm} - d_{(\alpha)}\right)^4 = \frac{d_{(\alpha)}^4}{k^2}\left\{\frac{\pi^4}{48}(2+\alpha^2)^2\right\} + O\left(\frac{1}{k^3}\right) \tag{6.18}$$

In (6.12), $\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^k$, converges to $\exp\left(-\gamma_e(\alpha-1)\right)$ fast. Figure 6.1 shows that as $k > 50 \sim 100$, this convenient asymptotic expression is reliable. The monotone convergence property will be useful in deriving various results later.

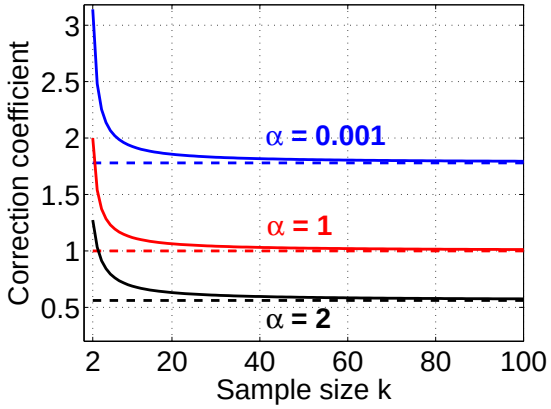


Figure 6.1: We plot the correction factor, $\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^k$, in (6.12) as a function of the sample size k, along with the asymptotic expression $\exp\left(-\gamma_e(\alpha-1)\right)$ (dashed horizontal lines). When $k > 50 \sim 100$, the asymptotic expression is accurate.

The asymptotic form (6.14) for the finite moment $\mathrm{E}\left(\hat{d}_{(\alpha),gm}\right)^t$ (fixed t, $k \to \infty$) will be useful in deriving asymptotic expressions for the variance, third and fourth central moments of $\hat{d}_{(\alpha),gm}$. It is also the tool which we will use to derive the tail bounds in the following Lemma 22 (proved in Section 6.9.2).

Lemma 22 *The right tail bound*

$$\mathbf{Pr}\left(\hat{d}_{(\alpha),gm} - d_{(\alpha)} > \epsilon d_{(\alpha)}\right) \leq \exp\left(-k\frac{\epsilon^2}{G_{R,\alpha,\epsilon}}\right), \quad 0 < \epsilon < e^{c_\alpha} - 1, \tag{6.19}$$

where $c_\alpha = \frac{\pi^2}{12}(\alpha^2 + 2)$ *and*

$$\frac{1}{G_{R,\alpha,\epsilon}} = \frac{\log^2(1+\epsilon)}{\epsilon^2 c_\alpha} - \frac{\log(1+\epsilon)}{\epsilon^2 c_\alpha}\gamma_e(\alpha - 1) - \frac{1}{\epsilon^2}\log\left(\frac{2}{\pi}\Gamma\left(\frac{\alpha\log(1+\epsilon)}{c_\alpha}\right)\Gamma\left(1 - \frac{\log(1+\epsilon)}{c_\alpha}\right)\sin\left(\frac{\pi\alpha\log(1+\epsilon)}{2c_\alpha}\right)\right). \tag{6.20}$$

The left tail bound

$$\mathbf{Pr}\left(\hat{d}_{(\alpha),gm} - d_{(\alpha)} < -\epsilon d_{(\alpha)}\right) \leq \exp\left(-k\frac{\epsilon^2}{G_{L,\alpha,\epsilon,k_0}}\right), \quad 0 < \epsilon < 1, k > k_0 > \frac{1}{\alpha}, \tag{6.21}$$

where

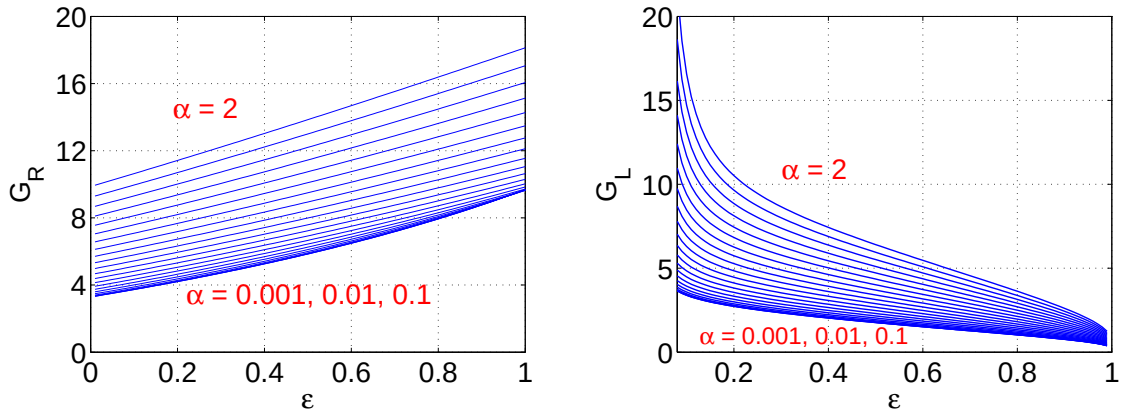
$$\frac{1}{G_{L,\alpha,\epsilon,k_0}} = -\frac{1}{c_\alpha\epsilon}\log(1-\epsilon) - \frac{1}{\epsilon^2}\log\left(-\frac{2}{\pi}\Gamma\left(1+\frac{\epsilon}{c_\alpha}\right)\Gamma\left(-\frac{\alpha\epsilon}{c_\alpha}\right)\sin\left(\frac{\pi}{2}\frac{\alpha\epsilon}{c_\alpha}\right)\right) - \frac{1}{c_\alpha\epsilon}k_0\log\left(\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k_0}\right)\Gamma\left(1-\frac{1}{k_0}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k_0}\right)\right]\right). \tag{6.22}$$

Figure 6.2 plots the constants to verify that they are reasonably small, at least for the range we are interested in.

Once we have derived the tail bounds in exponential forms, we can immediately establish an analog of the Johnson-Lindenstrauss (JL) Lemma in l_α, in Lemma 23.

Lemma 23 *Using the estimator* $\hat{d}_{(\alpha),gm}$*, for* $0 < \epsilon < 1$*, we only need* $k = O\left(\frac{\log(n)}{\epsilon^2}\right)$ *to guarantee that the* l_α *distance between any pair of points among* n *data points can be estimated within a factor of* $1 \pm \epsilon$*, with high probability. Moreover, the constant can be explicitly characterized from* $G_{R,\alpha,\epsilon}$ *and* $G_{L,\alpha,\epsilon,k_0}$ *defined in Lemma 22.*

This Lemma is weaker than classical JL Lemma for l_2, because the geometric mean is not a metric (e.g., it does not satisfy the triangle inequality).



(a) $G_{R,\alpha,\epsilon}$ (b) $G_{L,\alpha,\epsilon,k_0}$, $k_0 = 100$

Figure 6.2: We plot $G_{R,\alpha,\epsilon}$ and $G_{L,\alpha,\epsilon,k_0}$ in Lemma 22. The constants are reasonably small. We choose $\alpha = 0.001, 0.01$, and α from 0.1 to 2.0 spaced at 0.1.

6.2.2 The Biased Geometric Mean Estimator

We present another geometric mean estimator, which is biased and is asymptotically equivalent to the unbiased estimator $\hat{d}_{(\alpha),gm}$. Interestingly, this estimator exhibits smaller mean square errors than $\hat{d}_{(\alpha),gm}$ when $0.25 < \alpha < 1$, a good example of the bias-variance trade-off phenomenon.

Recall that, in this thesis, we define $d_{(\alpha)} = \sum_{i=1}^{D} |u_{1,i} - u_{2,i}|^\alpha$. Often we are also interested in $d_{(\alpha)}^{1/\alpha} = \left(\sum_{i=1}^{D} |u_{1,i} - u_{2,i}|^\alpha\right)^{1/\alpha}$. Suppose we have i.i.d. samples $x_j = S(\alpha, d_{(\alpha)})$, $j = 1, 2, ..., k$. It is easy to show that

$$\frac{\prod_{j=1}^{k} |x_j|^{1/k}}{\left[\frac{2}{\pi}\Gamma\left(\frac{1}{k}\right)\Gamma\left(1-\frac{1}{k\alpha}\right)\sin\left(\frac{\pi}{2}\frac{1}{k}\right)\right]^k}, \quad (k > 1/\alpha) \tag{6.23}$$

is an unbiased estimator of $d_{(\alpha)}^{1/\alpha}$. Therefore, we can define a biased estimator as

$$\hat{d}_{(\alpha),gm,b} = \frac{\prod_{j=1}^{k} |x_j|^{\alpha/k}}{\left[\frac{2}{\pi}\Gamma\left(\frac{1}{k}\right)\Gamma\left(1-\frac{1}{k\alpha}\right)\sin\left(\frac{\pi}{2}\frac{1}{k}\right)\right]^{\alpha k}}, \quad (k > \max(1/\alpha, 1)) \tag{6.24}$$

As $k \to \infty$, $\hat{d}_{(\alpha),gm,b}$ and $\hat{d}_{(\alpha),gm}$ are asymptotically equivalent. For finite k, $\hat{d}_{(\alpha),gm,b}$

may exhibit smaller mean square error than $\hat{d}_{(\alpha),gm}$, depending on α.

The mean square error (MSE) of $\hat{d}_{(\alpha),gm,b}$ is (for $k > \max(1/\alpha, 2)$)

$$\begin{aligned}&\mathrm{MSE}\left(\hat{d}_{(\alpha),gm,b}\right)\\ =&d_{(\alpha)}^2\left(\frac{\left[\frac{2}{\pi}\Gamma\left(\frac{2\alpha}{k}\right)\Gamma\left(1-\frac{2}{k}\right)\sin\left(\frac{\pi\alpha}{k}\right)\right]^k}{\left[\frac{2}{\pi}\Gamma\left(\frac{1}{k}\right)\Gamma\left(1-\frac{1}{k\alpha}\right)\sin\left(\frac{\pi}{2}\frac{1}{k}\right)\right]^{2k\alpha}}-\frac{2\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi\alpha}{2k}\right)\right]^k}{\left[\frac{2}{\pi}\Gamma\left(\frac{1}{k}\right)\Gamma\left(1-\frac{1}{k\alpha}\right)\sin\left(\frac{\pi}{2}\frac{1}{k}\right)\right]^{k\alpha}}+1\right)\end{aligned} \tag{6.25}$$

Using the monotonicity of $\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi\alpha}{2k}\right)\right]^k$, it is easy to show that

$$\mathrm{MSE}\left(\hat{d}_{(\alpha),gm,b}\right) > \mathrm{MSE}\left(\hat{d}_{(\alpha),gm}\right) = \mathrm{Var}\left(\hat{d}_{(\alpha),gm}\right), \quad \text{if} \quad \alpha > 1. \tag{6.26}$$

$$\mathrm{MSE}\left(\hat{d}_{(\alpha),gm,b}\right) < \mathrm{MSE}\left(\hat{d}_{(\alpha),gm}\right), \quad \text{if} \quad 0.5 < \alpha < 1. \tag{6.27}$$

That is, when $0.5 < \alpha < 1$, the biased estimator $\hat{d}_{(\alpha),gm,b}$ should be preferred over the unbiased estimator $\hat{d}_{(\alpha),gm}$.

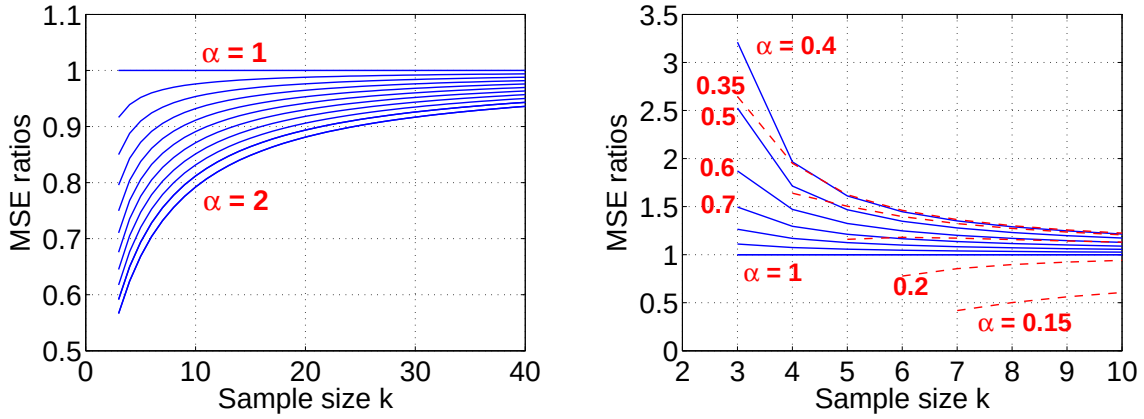
What happens when $0 < \alpha < 0.5$? We can numerically show that $\mathrm{MSE}\left(\hat{d}_{(\alpha),gm,b}\right) < \mathrm{MSE}\left(\hat{d}_{(\alpha),gm}\right)$, if $0.25 < \alpha < 0.5$. Figure 6.3 plots the ratios of the mean square errors, $\frac{\mathrm{MSE}\left(\hat{d}_{(\alpha),gm}\right)}{\mathrm{MSE}\left(\hat{d}_{(\alpha),gm,b}\right)}$, to help visualize the numerical results.

In a summary, we can benefit from the biased geometric mean estimator $\hat{d}_{(\alpha),gm,b}$ when $0.25 < \alpha < 1$. In Section 6.4, we will provide a *harmonic mean* estimator, which asymptotically outperforms the geometric mean estimators when $0 < \alpha \leq 0.344$. Therefore, we only recommend $\hat{d}_{(\alpha),gm,b}$ for $0.344 < \alpha < 1$.

6.3 Estimators for $\alpha \to 0+$

The case $\alpha \to 0+$ is interesting and practically useful. For example, [49] and [51] applied stable random projections for approximating the Hamming norm and the max-dominance norm, respectively, using very small α.

We will compare three estimators for $\alpha \to 0+$, including the geometric mean estimator, the sample median estimator, and the *harmonic mean* estimator. In terms



(a) $\alpha \geq 1$ (b) $\alpha \leq 1$

Figure 6.3: The MSE ratios $\frac{\mathrm{MSE}\left(\hat{d}_{(\alpha),gm}\right)}{\mathrm{MSE}\left(\hat{d}_{(\alpha),gm,b}\right)}$. (a): when $\alpha \geq 1$, the ratios are always below 1. α from 1.0 to 2.0, spaced at 0.1. (b): When $0.4 \leq \alpha \leq 1$, the ratios (solid curves) are all above 1.0 and follow the obvious pattern. The ratios for $\alpha = 0.35, 0.30, 0.25, 0.20$, and 0.15 (dashed curves) follow a reverse pattern. Overall, when $0.25 < \alpha < 1$, the MSE ratios are larger than 1, i.e., $\hat{d}_{(\alpha),gm,b}$ should be preferred.

of the variances, the harmonic mean estimator is asymptotically optimal and is substantially more accurate than the geometric mean estimator, which is considerably more accurate than the sample median estimator.

As shown by [52] (or see [113]), when $x \sim S(\alpha, 1)$, as $\alpha \to 0+$, $|x|^\alpha$ converges to $1/E_1$, where E_1 stands for an exponential distribution with mean 1. This property will become very intuitive when we discuss about sampling from an α-stable distribution in Section 6.7. Practically, we will have to use a small α to approximate $\alpha = 0+$. Smaller α values lead to better approximations but may be more sensitive to numerical instability. To simplify the discussion, we will assume that we have k i.i.d. samples

$$z_j \sim h/E_1, \quad j = 1, 2, ..., k, \tag{6.28}$$

where h is the Hamming distance:

$$h = \lim_{\alpha \to 0+} d_{(\alpha)} = \lim_{\alpha \to 0+} \sum_{i=1}^{D} |u_{1,i} - u_{2,i}|^\alpha = \#\{|u_{1,i} - u_{2,i}| \neq 0, i = 1, 2, ..., D\}. \quad (6.29)$$

Next, we will analyze and compare three estimators.

6.3.1 The (Unbiased) Geometric Mean Estimator

When $\alpha \to 0+$, the (unbiased) geometric mean estimator, denoted by $\hat{h}_{gm}$, becomes

$$\hat{h}_{gm} = \hat{d}_{(0+),gm} = \frac{\prod_{j=1}^{k} z_j^{1/k}}{\Gamma\left(1 - \frac{1}{k}\right)^k}, \quad (6.30)$$

with the variance

$$\mathrm{Var}\left(\hat{h}_{gm}\right) = h^2 \left(\frac{\Gamma\left(1 - \frac{2}{k}\right)^k}{\Gamma\left(1 - \frac{1}{k}\right)^{2k}} - 1 \right) = h^2 \left(\frac{\pi^2}{6} \frac{1}{k} \right) + O\left(\frac{1}{k^2}\right). \quad (6.31)$$

6.3.2 The Sample Median Estimator

Since the median of $1/E_1$ is $1/\log(2)$, the sample median estimator would be

$$\hat{h}_{me} = \frac{\mathrm{median}\left\{z_j, j = 1, 2, ..., k\right\}}{1/\log(2)}. \quad (6.32)$$

We know the distribution of $z_j \sim h/E_1$ exactly:

$$\mathbf{Pr}(z_j \leq t) = \exp(-h/t), \qquad f_{z_j}(t) = \exp(-h/t)\frac{h}{t^2}, \qquad t > 0. \quad (6.33)$$

The asymptotic variance of the sample median estimator $\hat{h}_{me}$ can be shown using known statistical results on sample quantiles [57, Theorem 9.2].

$$\mathrm{Var}\left(\hat{h}_{me}\right) = \frac{1}{4\left(f_{z_j}(h/\log(2))\right)^2 (1/\log(2))^2} \frac{1}{k} + O\left(\frac{1}{k^2}\right) = \frac{h^2}{\log^2(2)} \frac{1}{k} + O\left(\frac{1}{k^2}\right). \quad (6.34)$$

Therefore, asymptotically the ratio of the variances $\frac{\mathrm{Var}\left(\hat{h}_{me}\right)}{\mathrm{Var}\left(\hat{h}_{gm}\right)} = \frac{6}{\pi^2 \log^2(2)} \approx 1.27$. In other words, asymptotically, the geometric mean estimator is about 27% more accurate than the sample median estimator when $\alpha \to 0+$.

Non-asymptotically, however, the geometric mean estimator can be much more accurate when k is not too large. The moments of the sample median estimator $\hat{h}_{me}$ can be written as an integral:

$$\begin{aligned}\mathrm{E}\left(\hat{h}_{me}^s\right) =& h^s \int_0^\infty \frac{t^s}{1/\log^s(2)} \left(\exp\left(-\frac{1}{t}\right)\right)^m \left(1-\exp\left(-\frac{1}{t}\right)\right)^m \exp\left(-\frac{1}{t}\right) \frac{1}{t^2} \frac{(2m+1)!}{(m!)^2} dt \\ =& h^s \int_0^1 \frac{\log^s(2)}{(-\log(t))^s} (t-t^2)^m \frac{(2m+1)!}{(m!)^2} dt, \end{aligned} \tag{6.35}$$

by properties of sample quantiles; see [57, Chapter 2.1]. Here, for convenience, we only consider $k = 2m+1$. When $s = 1$ and 2, this integral can be expressed as finite (binomial-type) summations [85, 4.267.41, 4.268.5], which, however, are numerically unstable when $m > 12$.

We compare the estimators in terms of their mean square errors (MSE). The MSE of $\hat{h}_{me}$ is ∞ if $k < 5$ and it is about 3.26 times as large as that of $\hat{h}_{gm}$ if $k = 5$. The ratio of their MSEs, converges to about 1.27 as $k \to \infty$, as illustrated in Figure 6.4.

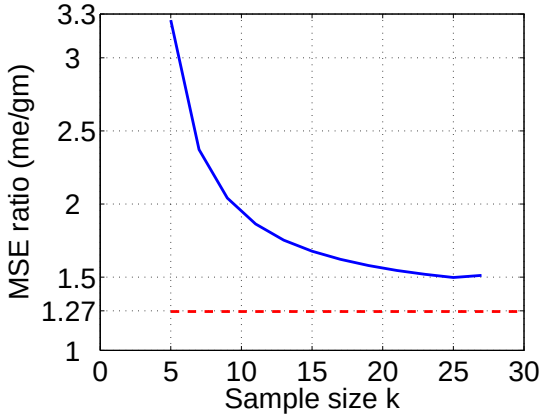


Figure 6.4: Ratios of the MSE $\frac{\mathrm{MSE}\left(\hat{h}_{me}\right)}{\mathrm{MSE}\left(\hat{h}_{gm}\right)}$. The horizontal (dashed) line is the theoretical asymptotic value (i.e., 1.27). The solid curve is obtained by analytically integrating (6.35). Note that (6.35) is numerically unstable when $k > 25$.

6.3.3 The Harmonic Mean Estimator

The *harmonic mean* estimator coincides with the maximum likelihood estimator (MLE) when $\alpha \to 0+$; and hence it is asymptotically optimal in term of the variance.

For $z_j \sim h/E_1$, it is easy to show $\mathrm{E}(\frac{1}{z_j}) = \frac{1}{h}$, suggesting a straightforward estimator

$$\hat{h}_{mle} = \frac{1}{\frac{1}{k}\sum_{j=1}^{k}\frac{1}{z_j}}. \tag{6.36}$$

$\hat{h}_{mle}$ is indeed the MLE. We recommend the bias-corrected version

$$\hat{h}_{mle,c} = \frac{1}{\frac{1}{k}\sum_{j=1}^{k}\frac{1}{z_j}}\left(1-\frac{1}{k}\right) = \frac{k-1}{\sum_{j=1}^{k}\frac{1}{z_j}}. \tag{6.37}$$

The moments of $\hat{h}_{mle}$ and $\hat{h}_{mle,c}$ are analyzed in the following lemma.

Lemma 24 *The first two moments of the maximum likelihood estimator (6.36) are*

$$E\left(\hat{h}_{mle}\right) = h\left(1+\frac{1}{k}\right) + O\left(\frac{1}{k^2}\right) \tag{6.38}$$

$$Var\left(\hat{h}_{mle}\right) = h^2\left(\frac{1}{k}+\frac{4}{k^2}\right) + O\left(\frac{1}{k^3}\right). \tag{6.39}$$

The first two moments of the bias-corrected maximum likelihood estimator (6.37) are

$$E\left(\hat{h}_{mle,c}\right) = h + O\left(\frac{1}{k^2}\right) \tag{6.40}$$

$$Var\left(\hat{h}_{mle,c}\right) = h^2\left(\frac{1}{k}+\frac{2}{k^2}\right) + O\left(\frac{1}{k^3}\right). \tag{6.41}$$

The tail bounds are

$$\begin{aligned}
&\mathbf{Pr}\left(\hat{h}_{mle,c} > (1+\epsilon)h\right) \\
\leq &\exp\left(-k\left(\log(1+\epsilon) - \frac{\epsilon}{1+\epsilon} + \log\left(\frac{k}{k-1}\right) - \frac{1}{k}\frac{1}{1+\epsilon}\right)\right), \epsilon > 0 \qquad (6.42)\\
&\mathbf{Pr}\left(\hat{h}_{mle,c} < (1-\epsilon)h\right) \\
\leq &\exp\left(-k\left(\log(1-\epsilon) + \frac{\epsilon}{1-\epsilon} + \log\left(\frac{k}{k-1}\right) - \frac{1}{k}\frac{1}{1-\epsilon}\right)\right), 0 < \epsilon < 1 \qquad (6.43)
\end{aligned}$$

See the proof in Section 6.9.3. These tail bounds are sharper than those based on the *geometric mean* estimator. Also, asymptotically, $\lim_{k\to\infty}\frac{\mathrm{Var}\left(\hat{h}_{gm}\right)}{\mathrm{Var}\left(\hat{h}_{mle,c}\right)} = \frac{\pi^2}{6} = 1.645$, $\lim_{k\to\infty}\frac{\mathrm{Var}\left(\hat{h}_{me}\right)}{\mathrm{Var}\left(\hat{h}_{mle,c}\right)} = \frac{1}{\log^2(2)} = 2.081$. When k is not too large, the harmonic mean estimator improves the other two estimators even much more considerably, as can be shown in Figure 6.5.

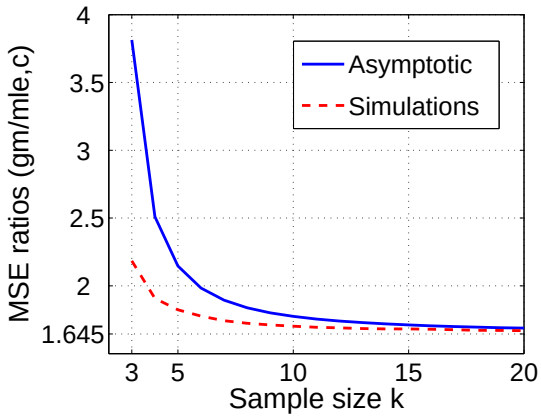


Figure 6.5: The solid curve represents the theoretical MSE ratios $\frac{\mathrm{MSE}(h_{gm})}{\mathrm{MSE}(h_{mle,c})}$, where we take $\mathrm{MSE}(h_{mle,c}) = h^2\left(\frac{1}{k} + \frac{2}{k^2}\right)$ to generate the curve. The dashed curve replaces $\mathrm{MSE}(h_{mle,c})$ with simulations. The harmonic mean estimator $h_{mle,c}$ considerably outperforms the geometric mean estimator h_{gm}. The plot also indicates that the asymptotic variance formula in Lemma 24 is reliable when $k > 15$.

6.3.4 Comparisons with Normal Random Projections in Boolean (0/1) Data

High-dimensional boolean (0/1) data are common in practice. To approximate the Hamming distances (h) in boolean data, we can apply stable random projections using very small α. Alternatively, we can also apply *normal random projections* directly. Suppose we are only interested in the estimation accuracy, one might ask which approach is better.

We have shown that, when $\alpha \to 0+$, using stable random projections with the harmonic mean estimator $h_{mle,c}$, the variance is $\frac{h^2}{k} + O\left(\frac{1}{k^2}\right)$.

If instead we apply normal random projections, the variance will be $\frac{2h^2}{k}$ (see Chapter 3), which is about twice as large as $\text{Var}\left(\hat{h}_{mle,c}\right)$.

Therefore, stable random projections with very small α not only provide a solution to approximating the Hamming norms in dynamic settings (i.e., data are subject to frequent additions/subtractions), but also are preferable even in static boolean data.

6.4 The Harmonic Mean Estimator for Small α

The harmonic mean estimator for $\alpha \to 0+$ can be extended to small α. This estimator outperforms the geometric mean estimator when $\alpha \leq 0.344$.

We present the harmonic mean estimator, denoted by $\hat{d}_{(\alpha),hm}$, in the following Lemma, which is proved in Section 6.9.4.

Lemma 25 *Suppose $x_j \sim S(\alpha, d_{(\alpha)})$, $j = 1, 2, ..., k$. We define the harmonic mean estimator, $\hat{d}_{(\alpha),hm}$, as*

$$\hat{d}_{(\alpha),hm} = \frac{-\frac{2}{\pi}\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)}{\sum_{j=1}^{k}|x_j|^{-\alpha}}\left(k - \left(\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2} - 1\right)\right), \qquad 0 < \alpha < 1.$$

This estimator is nearly unbiased,

$$E\left(\hat{d}_{(\alpha),hm}\right) = d_{(\alpha)} + O\left(\frac{1}{k^2}\right). \tag{6.44}$$

The variance of $\hat{d}_{(\alpha),hm}$ *is*

$$Var\left(\hat{d}_{(\alpha),hm}\right) = d_{(\alpha)}^2 \frac{1}{k}\left(\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2} - 1\right) + O\left(\frac{1}{k^2}\right), \qquad 0 < \alpha < 1/2.$$

Moreover,

$$\lim_{\alpha\to 0+} -\frac{2}{\pi}\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right) = 1, \tag{6.45}$$

$$\lim_{\alpha\to 0+} \frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2} - 1 = 1, \tag{6.46}$$

and, there exists a unique α_0 *(*$0 < \alpha_0 < 1/2$*), such that, for any* $0 < \alpha < \alpha_0$*, we have*

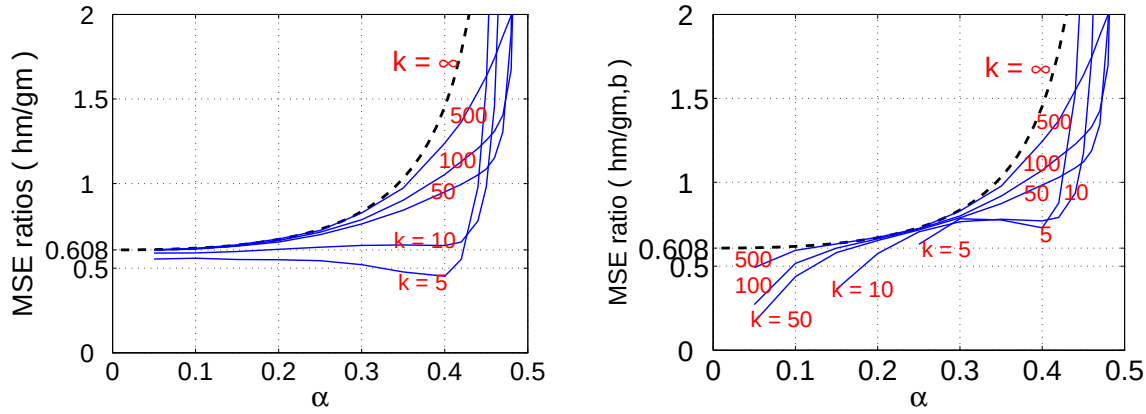
$$\left(\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2} - 1\right) < \frac{\pi^2}{12}(\alpha^2 + 2). \tag{6.47}$$

Thus, as long as $0 < \alpha < \alpha_0$*, the asymptotic variance of the harmonic mean estimator is always smaller than the asymptotic variance of the geometric mean estimator.*

We can numerically find the break-even point to be $\alpha_0 \approx 0.3441$. Therefore, for any $0 < \alpha \leq 0.344$, the harmonic mean estimator $\hat{d}_{(\alpha),hm}$ should be preferred over the geometric mean estimators. This can be visualized from Figure 6.6. The figure also shows that for finite k, the break-even point for α is actually (slightly) larger than 0.344. Therefore, using $\hat{d}_{(\alpha),hm}$ for $0 < \alpha \leq 0.344$ is a "safe" recommendation.

6.5 The Fractional Power Estimator

The *fractional power* estimator is defined in Lemma 26. See the proof in Section 6.9.5.



(a) $\hat{d}_{(\alpha),hm}$ v.s. $\hat{d}_{(\alpha),gm}$ (b) $\hat{d}_{(\alpha),hm}$ v.s. $\hat{d}_{(\alpha),gm,b}$

Figure 6.6: We plot the MSE ratios to verify that, when $0 < \alpha \leq 0.344$, the harmonic mean estimator $\hat{d}_{(\alpha),hm}$ should be preferred over the geometric mean estimators. In both panels, the dashed curves are the asymptotic MSE ratios, i.e., $\left(\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2} - 1\right) / \left(\frac{\pi^2}{12}(\alpha^2+2)\right)$. To show the non-asymptotic behavior, we also plot the MSE ratios at finite k ($k = 5, 10, 50, 100$, and 500). Panel (a) plots $\frac{\mathrm{MSE}\left(\hat{d}_{(\alpha),hm}\right)}{\mathrm{MSE}\left(\hat{d}_{(\alpha),gm}\right)}$, and Panel (b) plots $\frac{\mathrm{MSE}\left(\hat{d}_{(\alpha),hm}\right)}{\mathrm{MSE}\left(\hat{d}_{(\alpha),gm,b}\right)}$, where we simulate $\mathrm{MSE}\left(\hat{d}_{(\alpha),hm}\right)$. We can visualize from these curves that for finite k, the break-even points of α (i.e., the MSE ratio = 1) are actually slightly larger than 0.344.

Lemma 26 *Denoted by $\hat{d}_{(\alpha),fp}$, the fractional power estimator is defined as*

$$\hat{d}_{(\alpha),fp} = \left(\frac{1}{k}\frac{\sum_{j=1}^k |x_j|^{\lambda^*\alpha}}{\frac{2}{\pi}\Gamma(1-\lambda^*)\Gamma(\lambda^*\alpha)\sin\left(\frac{\pi}{2}\lambda^*\alpha\right)}\right)^{1/\lambda^*} \times \left(1 - \frac{1}{k}\frac{1}{2\lambda^*}\left(\frac{1}{\lambda^*}-1\right)\left(\frac{\frac{2}{\pi}\Gamma(1-2\lambda^*)\Gamma(2\lambda^*\alpha)\sin\left(\pi\lambda^*\alpha\right)}{\left[\frac{2}{\pi}\Gamma(1-\lambda^*)\Gamma(\lambda^*\alpha)\sin\left(\frac{\pi}{2}\lambda^*\alpha\right)\right]^2} - 1\right)\right), \quad (6.48)$$

where

$$\lambda^* = \underset{-\frac{1}{2\alpha}<\lambda<\frac{1}{2}}{argmin}\ g\left(\lambda;\alpha\right), \qquad g\left(\lambda;\alpha\right) = \frac{1}{\lambda^2}\left(\frac{\frac{2}{\pi}\Gamma(1-2\lambda)\Gamma(2\lambda\alpha)\sin\left(\pi\lambda\alpha\right)}{\left[\frac{2}{\pi}\Gamma(1-\lambda)\Gamma(\lambda\alpha)\sin\left(\frac{\pi}{2}\lambda\alpha\right)\right]^2} - 1\right). \quad (6.49)$$

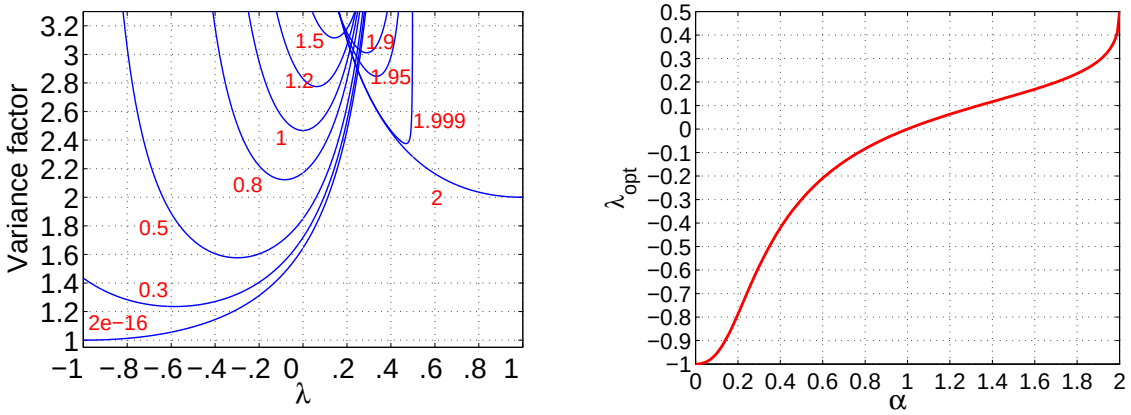
Asymptotically (i.e., as $k \to \infty$), the bias and variance of $\hat{d}_{(\alpha),fp}$ are

$$E\left(\hat{d}_{(\alpha),fp}\right) - d_{(\alpha)} = O\left(\frac{1}{k^2}\right), \tag{6.50}$$

$$Var\left(\hat{d}_{(\alpha),fp}\right) = d_{(\alpha)}^2 \frac{1}{k}\frac{1}{\lambda^{*2}}\left(\frac{\frac{2}{\pi}\Gamma(1-2\lambda^*)\Gamma(2\lambda^*\alpha)\sin\left(\pi\lambda^*\alpha\right)}{\left[\frac{2}{\pi}\Gamma(1-\lambda^*)\Gamma(\lambda^*\alpha)\sin\left(\frac{\pi}{2}\lambda^*\alpha\right)\right]^2} - 1\right) + O\left(\frac{1}{k^2}\right). \tag{6.51}$$

Note that in calculating $\hat{d}_{(\alpha),fp}$, the real computation only involves $\left(\sum_{j=1}^k |x_j|^{\lambda^*\alpha}\right)^{1/\lambda^*}$, because all other terms are basically constants and can be pre-computed.

Figure 6.7(a) plots $g\left(\lambda;\alpha\right)$ as a function of λ for different values of α. Figure 6.7(b) plots the optimal λ^* as a function of α. We can see that $g\left(\lambda;\alpha\right)$ is a convex function of λ and $-1 < \lambda^* < \frac{1}{2}$ (when $\alpha < 2$), which will be proved in Lemma 27.



(a) (b)

Figure 6.7: Panel (a) plots the variance factor $g\left(\lambda;\alpha\right)$ as functions of λ for different values of α. We can see that $g\left(\lambda;\alpha\right)$ is a convex function of λ and the optimal solution (lowest points on the curves) are between -1 and 0.5 ($\alpha < 2$). Note that there is a discontinuity between $\alpha \to 2-$ and $\alpha = 2$. Panel (b) plots the optimal λ^* as a function of α. Since $\alpha = 2$ is not included, we only see $\lambda^* < 0.5$ in the figure.

In Table 6.1, we tabulate the values for the optimal λ^*, the variance factor, $\frac{1}{\lambda^{*2}}\left(\frac{\frac{2}{\pi}\Gamma(1-2\lambda^*)\Gamma(2\lambda^*\alpha)\sin(\pi\lambda^*\alpha)}{\left[\frac{2}{\pi}\Gamma(1-\lambda^*)\Gamma(\lambda^*\alpha)\sin\left(\frac{\pi}{2}\lambda^*\alpha\right)\right]^2} - 1\right)$, as well as the maximum orders (γ_m) of bounded moments, for selected α. For example, if $\alpha = 1.90$, then $\mathrm{E}\left(\hat{d}_{(1.90),fp}\right)^\gamma < \infty$ if $\gamma < \gamma_m = 3.5$. The magnitude of γ_m is a good indicator of the tail behavior of the

estimator $\hat{d}_{(\alpha),fp}$, while, depending on the requirements of specific applications, the theoretical (asymptotic) variance is not always the best evaluation criterion. From the table, we can expect good tail behavior when $\alpha \to 0+$ or 1.

Table 6.1: The optimal λ^*, the variance factor $\frac{1}{\lambda^{*2}}\left(\frac{\frac{2}{\pi}\Gamma(1-2\lambda^*)\Gamma(2\lambda^*\alpha)\sin(\pi\lambda^*\alpha)}{\left[\frac{2}{\pi}\Gamma(1-\lambda^*)\Gamma(\lambda^*\alpha)\sin\left(\frac{\pi}{2}\lambda^*\alpha\right)\right]^2}-1\right)$, and the maximum order (γ_m) of bounded moments are tabulated. For example, when $\alpha = 1.90$, $\mathrm{E}\left(\hat{d}_{(1.90),fp}\right)^\gamma < \infty$ if $\gamma < \gamma_m = 3.5$.

α	λ^*	Var factor	Moments
0.001	-0.9999972	1.000002	< 1000
0.01	-0.9996995	1.000169	< 100
0.05	-0.9907232	1.004739	< 20.2
0.10	-0.9543453	1.021523	< 10.5
0.15	-0.8839427	1.053513	< 7.5
0.20	-0.7877194	1.101431	< 6.3
0.30	-0.5844494	1.235287	< 5.7
0.344	-0.5065260	1.304899	< 5.7
0.40	-0.4207041	1.399130	< 5.9
0.50	-0.2995625	1.576568	< 6.7
0.60	-0.2090153	1.759186	< 8.0
0.70	-0.1392770	1.942324	< 10.3
0.80	-0.0838209	2.122944	< 14.9
0.90	-0.0383596	2.298683	< 29.0
0.95	-0.0184348	2.384053	< 57.1
0.99	-0.0035803	2.450906	< 282

α	λ^*	Var factor	Moments
1.05	0.01717657	2.548452	< 58.2
1.10	0.03329429	2.626915	< 30.0
1.20	0.06303470	2.774800	< 15.9
1.30	0.09042653	2.908174	< 11.1
1.40	0.11653101	3.023410	< 8.9
1.50	0.14242018	3.115667	< 7.0
1.60	0.16938764	3.178010	< 5.9
1.70	0.19936117	3.199507	< 5.0
1.80	0.23603107	3.160157	< 4.2
1.85	0.25941186	3.105168	< 3.9
1.90	0.28958180	3.011529	< 3.5
1.95	0.33479644	2.847017	< 3.0
1.99	0.41211762	2.561692	< 2.4
1.999	0.46871987	2.375615	< 2.13
1.9999	0.48969067	2.312855	< 2.04

6.5.1 Special cases

The discontinuity, $\lambda^*(2-) = 0.5$ and $\lambda^*(2) = 1$, reflects the fact that, for $x \sim S(\alpha, d)$, $\mathrm{E}\left(|x|^\lambda\right)$ exists for $-1 < \lambda < \alpha$ when $\alpha < 2$ and exists for any $\lambda > -1$ when $\alpha = 2$.

When $\alpha = 2$, since $\lambda^*(2) = 1$, the *fractional power* estimator becomes $\frac{1}{k}\sum_{j=1}^k |x_j|^2$, i.e., the *arithmetic mean* estimator. We will from now on only consider $0 < \alpha < 2$.

when $\alpha \to 0+$, since $\lambda^*(0+) = -1$, the *fractional power* estimator approaches the *harmonic mean* estimator, which is asymptotically optimal when $\alpha = 0+$[116].

When $\alpha \to 1$, since $\lambda^*(1) = 0$ in the limit, the *fractional power* estimator has the same asymptotic variance as the *geometric mean estimator*.

6.5.2 Theoretical Properties

We study more theoretical properties of the *fractional power* estimator $\hat{d}_{(\alpha),fp}$. In particular, we will show that to find the optimal λ^* only involves searching for the minimum on a convex curve in the narrow range $\lambda^* \in \left(\max\left\{-1, -\frac{1}{2\alpha}\right\}, 0.5\right)$. These properties theoretically ensure that this estimator is well-defined and is numerically easy to compute. The proof of Lemma 27 is in Section 6.9.6.

Lemma 27 *Part 1:*

$$g\left(\lambda;\alpha\right) = \frac{1}{\lambda^2}\left(\frac{\frac{2}{\pi}\Gamma(1-2\lambda)\Gamma(2\lambda\alpha)\sin\left(\pi\lambda\alpha\right)}{\left[\frac{2}{\pi}\Gamma(1-\lambda)\Gamma(\lambda\alpha)\sin\left(\frac{\pi}{2}\lambda\alpha\right)\right]^2} - 1\right), \tag{6.52}$$

is a convex function of λ.

Part 2: For $0 < \alpha < 2$, *the optimal* $\lambda^* = \underset{-\frac{1}{2\alpha}\lambda<\frac{1}{2}}{argming}\left(\lambda;\alpha\right)$, *satisfies* $-1 < \lambda^* < 0.5$.

6.5.3 Comparing Variances at Finite Samples

Figure 6.8 plots the empirical mean square errors (MSE) from simulations for the *fractional power* estimator, the *harmonic mean* estimator, and the *sample median* estimator. The MSE for the *geometric mean* estimators can be computed exactly without simulations.

Figure 6.8 indicates that the *fractional power* estimator $\hat{d}_{(\alpha),fp}$ has good small sample performance unless α is close to 2. After $k \geq 50$, the advantage of $\hat{d}_{(\alpha),fp}$ becomes noticeable even when α is very close to 2. It is also clear that the *sample median* estimator has poor small sample performance; but even at very large k, its performance is not that good except when α is about 1.

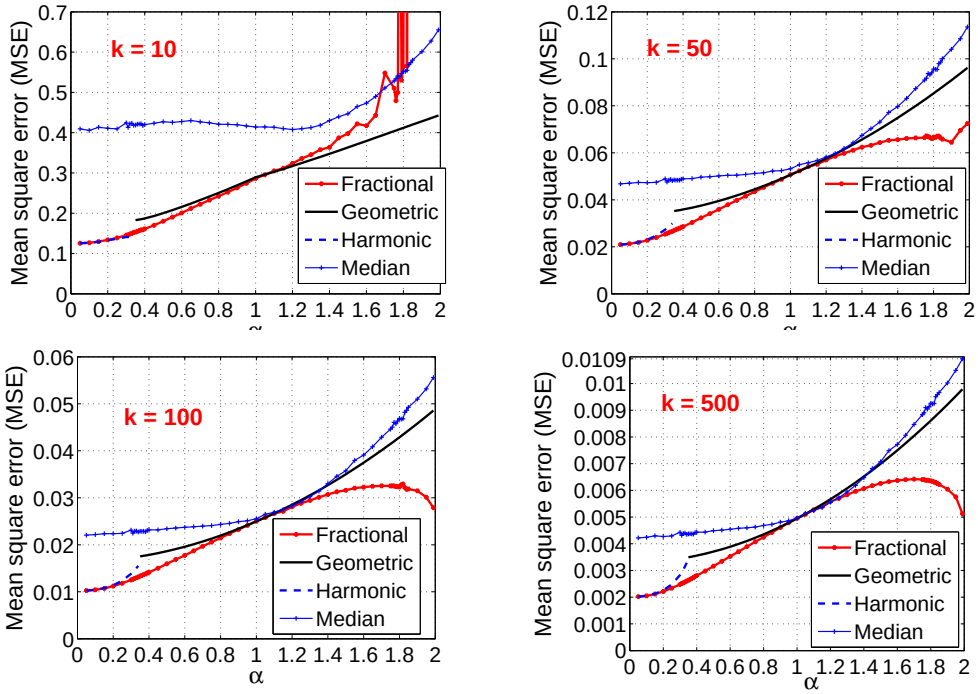


Figure 6.8: We simulated the MSEs for various estimators, 10^6 times for every α and k. The MSEs for the *geometric mean* estimators were computed exactly. The *harmonic mean* estimator was used for $\alpha \leq 0.344$, the *biased geometric mean* for $0.344 < \alpha \leq 1$, and the *unbiased geometric mean* for $1 < \alpha < 2$. The *fractional power* has good accuracy except when k is small and α is close to 2.

6.6 Asymptotic (Cramér-Rao) Efficiencies

For an estimator $\hat{d}_{(\alpha)}$, its variance, under certain regularity condition, is lower-bounded by the Information inequality (also known as the Cramér-Rao bound)[110, Chapter 2], i.e., $\text{Var}\left(\hat{d}_{(\alpha)}\right) \geq \frac{1}{k\text{I}(\alpha)}$. The Fisher Information $\text{I}(\alpha)$ can be approximated by computationally intensive procedures[137].

When $\alpha = 2$, it is well-known that the *arithmetic mean* estimator attains the Cramér-Rao bound. When $\alpha = 0+$, we have shown that both the *fractional power* and the *harmonic mean* estimators are asymptotically optimal.

The asymptotic (Cramér-Rao) efficiency is defined as the ratio of $\frac{1}{k\text{I}(\alpha)}$ to the asymptotic variance of $\hat{d}_{(\alpha)}$ ($d_{(\alpha)} = 1$ for simplicity). Figure 6.9 plots the efficiencies for all estimators we have mentioned.

In terms of the asymptotic variance, it appears that the *fractional power* estimator would be the "best" choice. However, we often choose estimators not only based on the asymptotic variances but also based on the tail behaviors. When α is close to 2, the *fractional* power estimator does not have moments much higher than the second order as shown in Table 6.1 and hence it is expected that the *fractional power* estimator would not have good tail behavior when α is close to 2.

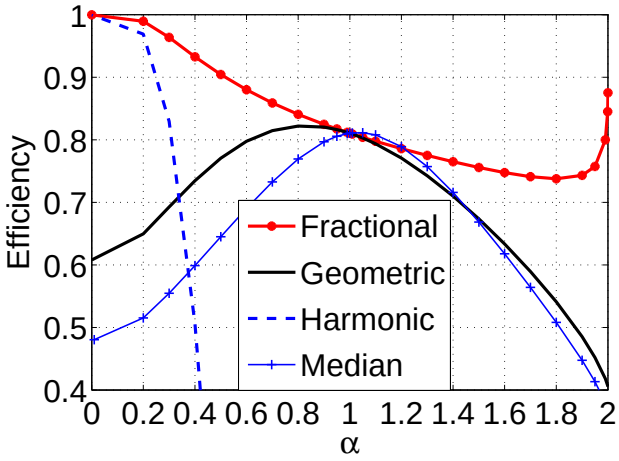


Figure 6.9: The asymptotic Cramér-Rao efficiencies of various estimators for $0 < \alpha < 2$. The asymptotic variance of the *sample median* estimator $\hat{d}_{(\alpha),me}$ is computed from known statistical theory for sample quantiles. We can see that the *fractional power* estimator $\hat{d}_{(\alpha),fp}$ is close to be optimal in a wide range of α; and it always outperforms both the *geometric mean* and the *harmonic mean* estimators. Note that since we only consider $\alpha < 2$, the efficiency of $\hat{d}_{(\alpha),fp}$ does not achieve 100% when $\alpha \to 2-$.

6.7 Very Sparse Stable Random Projections

There are practical issues with *stable random projections* using stable distributions.

First, the processing time for the matrix multiplication $\mathbf{AR}$ is expensive in modern massive data streams, one of the open questions raised in a recent workshop on data streams, `http://www.cse.iitk.ac.in/users/sganguly/data-stream-probs.pdf`.

Another problem is that we have to store the projection matrix $\mathbf{R}$ at the memory

cost[1] $O(Dk)$, which is also expensive for large D. We need to store $\mathbf{R}$ because (1) new data points may arrive, (2) data may be subject to frequent updating, (3) data do not necessarily arrive in order.

In addition, sampling from the stable distribution is in general expensive. As described in [151, Proposition 1.71.1], we sample W_1 uniform on $(-\frac{\pi}{2}, \frac{\pi}{2})$ and E_1 from an exponential distribution with mean 1. If W_1 and E_1 are independent, then

$$\frac{\sin(\alpha W_1)}{\cos(W_1)^{1/\alpha}} \left(\frac{\cos\left((1-\alpha)W_1\right)}{E_1} \right)^{(1-\alpha)/\alpha} \tag{6.53}$$

is distributed as $S(\alpha, 1)$. Clearly, this procedure is not inexpensive.

In contrast, a Pareto distribution has the same tail behaviors as $S(\alpha, 1)$, but it is much easier to sample from. Thus, we suggest *very sparse stable random projections*, using a Pareto to replace stable distributions.

6.7.1 Our Solution: Very Sparse Stable Random Projections

We propose *very sparse stable random projections* by replacing the α-stable random variable $S(\alpha, 1)$ in the projection matrix $\mathbf{R}$ with a mixture of a symmetric α-Pareto distribution (with probability $0 < \beta \leq 1$) and a point mass at the origin (with probability $1 - \beta$), i.e., $\mathbf{R}$ consists of i.i.d. entries

$$r_{ij} = \begin{cases} P_\alpha & \text{with prob. } \frac{\beta}{2} \\ 0 & \text{with prob. } 1-\beta \\ -P_\alpha & \text{with prob. } \frac{\beta}{2} \end{cases}, \tag{6.54}$$

where P_α denotes an α-Pareto variable, $\mathbf{Pr}\left(P_\alpha > t\right) = \frac{1}{t^\alpha}$ if $t \geq 1$; and 0 otherwise. This procedure is beneficial because

- It is much easier to sample from an α-Pareto distribution P_α than from an α-stable distribution $S(\alpha, 1)$.

[1] Another possibility is to store only the seed for generating the random projection matrix $\mathbf{R}$ and then re-create the specific entries of $\mathbf{R}$ on the fly. This approach, of course, would be computationally prohibitive and may be not suitable for real-time applications.

- Computing $\mathbf{A} \times \mathbf{R}$ costs only $O(\beta nDk)$ as opposed to $O(nDk)$, i.e., a $\frac{1}{\beta}$-fold speedup.
- The storage (of $\mathbf{R}$) cost is reduced from $O(Dk)$ to $O(\beta Dk)$.

With this scheme, the projected data, under reasonable regularity conditions, are asymptotically (instead of exactly) α-stable. When the convergence is "fast enough," we might treat the projected data as if they were exactly α-stable so that we could use the theoretical estimators developed for stable random projections.

Lemma 28 *Suppose z_i, $i = 1, 2, ..., D$, are i.i.d. random variables defined in (6.54). g_1, g_2, ..., g_D are any constants. Then, as $D \to \infty$,*

$$\frac{\sum_{i=1}^D z_i g_i}{\left(\beta \sum_{s=1}^D |g_s|^\alpha\right)^{\frac{1}{\alpha}}} \stackrel{D}{\Longrightarrow} S\left(\alpha, \Gamma(1-\alpha)\cos\left(\frac{\pi}{2}\alpha\right)\right), \tag{6.55}$$

provided

$$\frac{\max\limits_{1\le i\le D}(|g_i|)}{\left(\sum_{s=1}^D |g_s|^\alpha\right)^{\frac{1}{\alpha}}} \to 0. \tag{6.56}$$

At the risk of not being rigorous, we sometimes write (6.55) in Lemma 28 conveniently as

$$\sum_{i=1}^D z_i g_i \stackrel{D}{\Longrightarrow} S\left(\alpha, \Gamma(1-\alpha)\cos\left(\frac{\pi}{2}\alpha\right)\beta \sum_{i=1}^D |g_i|^\alpha\right). \tag{6.57}$$

When the convergence is fast, we might assume the projected data after *very sparse stable random projections* to be "exactly stable" and we might still use the various estimators proposed. Of course, we need to adjust the estimates for the additional constant factor $\Gamma(1-\alpha)\cos\left(\frac{\pi}{2}\alpha\right)\beta$.

6.7.2 Classifying Cancers in Microarray Data

We consider the same task as in Chapter 4 for classifying cancer types from the Harvard microarray expression data[24]. We will use a nearest neighbor classifier based on the l_1 norm, as opposed to the l_2 norm in Chapter 4

A simple nearest neighbor method can classify the samples nearly perfectly using the l_1 distance. When $m = 1, 3, 5, 7, 9$, the m-nearest neighbor classifier mis-classifies 3, 2, 2, 2, 2, samples, respectively.

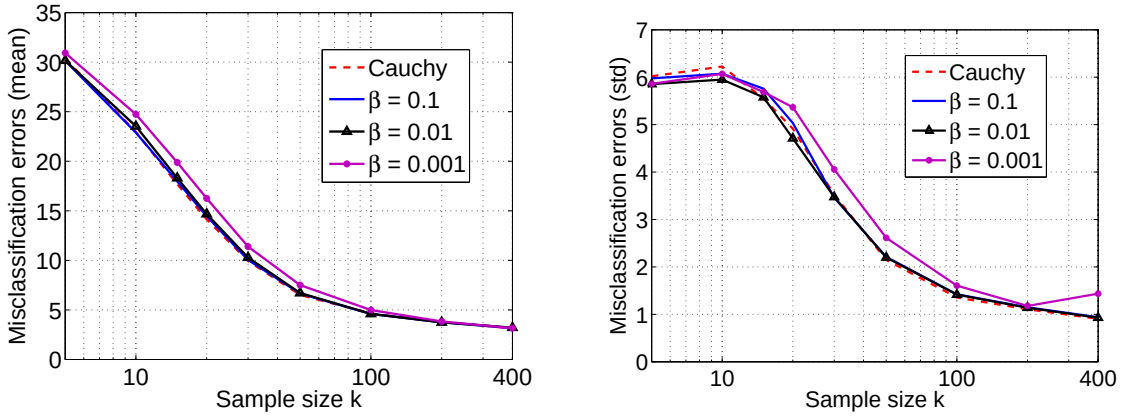
We conduct both *Cauchy random projections* and *very sparse stable random projections* ($\beta = 0.1$, 0.01, and 0.001) and classify the specimens using a 5-nearest neighbor classifier based on the estimated l_1 distances (using the geometric mean estimator). Figure 6.10 indicates that (A): stable random projections can achieve similar classification accuracy using about 100 projections; (B): very sparse stable random projections work well when $\beta = 0.1$ and 0.01. Even with $\beta = 0.001$, the classification results are only slightly worse.

6.8 Summary

In *stable random projections*, recovering the original l_α properties from the projected data boils down to estimating the scale parameter of a symmetric α-stable distribution. The popular statistical estimator based on the sample inter-quantiles is often not accurate (especially at small samples) and it is difficult for theoretical analysis. We provide various estimators based on the geometric mean, the harmonic mean. and the fractional power. The geometric mean estimator is convenient for theoretical analysis and the fractional power estimator is near-optimal in the Cramér-Rao sense.

Our theoretical analysis on the (unbiased) geometric mean estimator provides the tail bounds in exponential forms with constants explicitly given. These tail bounds consequently establish an analog of the Johnson-Lindenstrauss (JL) Lemma that $k = O\left(\frac{\log n}{\epsilon^2}\right)$ suffices, for dimension reduction in l_α. This Lemma is weaker than the classical JL Lemma in l_2 as the geometric mean is not a metric.

Assuming reasonable regularity conditions on the data, *very sparse stable random*



(a) Mean

(b) Standard errors

Figure 6.10: We apply *Cauchy random projections* and *very sparse stable random projections* ($\beta = 0.1,\ 0.01,\ 0.001$) and classify the microarray specimens using a 5-nearest neighbor classifier and l_1 distances. Panel (a) show that *Cauchy random projections* can achieve almost the same classification accuracy with 100 projections. Panel (a) also shows that *very sparse stable random projections* with $\beta = 0.1$ and 0.01 perform almost indistinguishably from *Cauchy random projections*. Each curve is averaged over 1000 runs. Panel (b) plots the standard errors, indicating that the classification accuracy does not fluctuate much.

projections can considerably speedup the processing (matrix multiplication) time, which may be critical in certain (e.g.,) massive data stream applications.

6.9 Proofs

6.9.1 Proof of Lemma 21

The estimator defined in (6.11)

$$\hat{d}_{(\alpha),gm} = \frac{\prod_{j=1}^{k} |x_j|^{\alpha/k}}{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^k}, \quad k > 1$$

is obviously unbiased, i.e., $\mathrm{E}\left(\hat{d}_{(\alpha),gm}\right) = d_{(\alpha)}$, because x_j's are i.i.d. in $S(\alpha, d_{(\alpha)})$ with

$$\mathrm{E}\left(|x_j|^{\alpha/k}\right) = d_{(\alpha)}^{\alpha/k}\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right).$$

The variance of $\hat{d}_{(\alpha),gm}$ is then

$$\mathrm{Var}\left(\hat{d}_{(\alpha),gm}\right) = d_{(\alpha)}^2\left\{\frac{\left[\frac{2}{\pi}\Gamma\left(\frac{2\alpha}{k}\right)\Gamma\left(1-\frac{2}{k}\right)\sin\left(\pi\frac{\alpha}{k}\right)\right]^k}{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^{2k}} - 1\right\}.$$

We now show

$$\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^k \to \exp\left(-\gamma_e\left(\alpha-1\right)\right), \text{ as } k\to\infty,$$

where $\gamma_e = 0.577215665...$, is Euler's constant.

By Euler's reflection formula, $\Gamma(1-z)\Gamma(z) = \frac{\pi}{\sin(\pi z)}$,

$$\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^k = \left[\frac{2\Gamma\left(\frac{\alpha}{k}\right)\sin\left(\frac{\pi\alpha}{2k}\right)}{\Gamma\left(\frac{1}{k}\right)\sin\left(\frac{\pi}{k}\right)}\right]^k$$

We use the infinite-product representation of the Gamma function [85, 8.322],

$$\Gamma(z) = \frac{\exp\left(-\gamma_e z\right)}{z}\prod_{s=1}^{\infty}\left(1+\frac{z}{s}\right)^{-1}\exp\left(\frac{z}{s}\right),$$

and the infinite-product representation of the sine function [85, 1.431.1],

$$\sin(z) = z\prod_{s=1}^{\infty}\left(1-\frac{z^2}{s^2\pi^2}\right),$$

to obtain

$$\left[\frac{2\Gamma\left(\frac{\alpha}{k}\right)\sin\left(\frac{\pi\alpha}{2k}\right)}{\Gamma\left(\frac{1}{k}\right)\sin\left(\frac{\pi}{k}\right)}\right]^k = \exp\left(-\gamma_e(\alpha-1)\right)\times$$
$$\left(\prod_{s=1}^{\infty}\exp\left(\frac{1}{sk}(\alpha-1)\right)\left(1+\frac{\alpha}{ks}\right)^{-1}\left(1+\frac{1}{ks}\right)\left(1-\frac{\alpha^2}{4k^2s^2}\right)\left(1-\frac{1}{s^2k^2}\right)^{-1}\right)^k$$
$$= \exp\left(-\gamma_e(\alpha-1)\right)\times\left(\prod_{s=1}^{\infty}\exp\left(\frac{1}{sk}(\alpha-1)\right)\left(1+\frac{1}{sk}(1-\alpha)\right)+O\left(\frac{1}{k^2}\right)\right)^k$$
$$= \exp\left(-\gamma_e(\alpha-1)\right)\times\left(\prod_{s=1}^{\infty}\left(1+\frac{1}{sk}(\alpha-1)+O\left(\frac{1}{s^2k^2}\right)\right)\left(1+\frac{1}{sk}(1-\alpha)+O\left(\frac{1}{s^2k^2}\right)\right)\right)^k$$
$$= \exp\left(-\gamma_e(\alpha-1)\right)\times\left(\prod_{s=1}^{\infty}\left(1+O\left(\frac{1}{s^2k^2}\right)\right)\right)^k$$
$$= \exp\left(-\gamma_e(\alpha-1)\right)\times\exp\left(k\sum_{s=1}^{\infty}\log\left(1+O\left(\frac{1}{s^2k^2}\right)\right)\right)$$
$$= \exp\left(-\gamma_e(\alpha-1)\right)\times\exp\left(k\sum_{s=1}^{\infty}O\left(\frac{1}{s^2k^2}\right)\right)$$
$$= \exp\left(-\gamma_e(\alpha-1)\right)\times\exp\left(O\left(\frac{1}{k}\right)\right) \qquad \left(\text{because } \sum_{s=1}^{\infty}\frac{1}{s^2}=\frac{\pi^2}{6}\right)$$
$$\to \exp\left(-\gamma_e(\alpha-1)\right) \qquad (\text{as } k\to\infty)$$

To show that

$$\left[\frac{2\Gamma\left(\frac{\alpha}{k}\right)\sin\left(\frac{\pi\alpha}{2k}\right)}{\Gamma\left(\frac{1}{k}\right)\sin\left(\frac{\pi}{k}\right)}\right]^k \to \exp\left(-\gamma_e(\alpha-1)\right)$$

converges from above monotonically, it is equivalent to show that

$$\left(\prod_{s=1}^{\infty}\left(1+\frac{\alpha}{ks}\right)^{-1}\left(1+\frac{1}{ks}\right)\left(1-\frac{\alpha^2}{4k^2s^2}\right)\left(1-\frac{1}{s^2k^2}\right)^{-1}\right)^k$$

decreases monotonically with increasing k. It suffices to show that for any $s\geq 1$

$$\left(\left(1+\frac{\alpha}{ks}\right)^{-1}\left(1+\frac{1}{ks}\right)\left(1-\frac{\alpha^2}{4k^2s^2}\right)\left(1-\frac{1}{s^2k^2}\right)^{-1}\right)^k$$

decreases monotonically, which is equivalent to show the monotonicity of

$$k \log \left(\frac{4k^3s^3 + 4k^2s^2 - \alpha^2 ks - \alpha^2}{4\left(k^3s^3 + \alpha k^2 s^2 - ks - \alpha\right)} \right),$$

or, equivalently, the monotonicity of

$$g(t) = t \log \left(\frac{4t^3 + 4t^2 - \alpha^2 t - \alpha^2}{4\left(t^3 + \alpha t^2 - t - \alpha\right)} \right), \qquad (t > 1).$$

We can check $g'(t) \leq 0$.

$$\begin{aligned} g'(t) =& \log \left(\frac{4t^2 - \alpha^2}{4(t-1)(t+\alpha)} \right) + \frac{\alpha}{t+\alpha} - \frac{1}{t-1} + \frac{2\alpha^2}{4t^2 - \alpha^2} \\ \leq& -1 + \left(\frac{4t^2 - \alpha^2}{4(t-1)(t+\alpha)} \right) + \frac{\alpha}{t+\alpha} - \frac{1}{t-1} + \frac{2\alpha^2}{4t^2 - \alpha^2} \\ =& \frac{\alpha\left(t^2(-16+4\alpha) + 8t(\alpha^2 - \alpha) + \alpha^2(\alpha - 4)\right)}{4(t-1)(t+\alpha)(4t^2 - \alpha^2)}. \end{aligned}$$

Note that $\log(x) \leq x - 1$ if $0 < x < 2$. It is easy to verify that $0 < \frac{4t^2 - \alpha^2}{4(t-1)(t+\alpha)} < 2$.

Therefore, to show $g'(t) \leq 0$, it suffices to show $t^2(-16+4\alpha) + 8t(\alpha^2 - \alpha) + \alpha^2(\alpha - 4) \leq 0$, which is true, because $-16 + 4\alpha < 0$ and

$$\left(8(\alpha^2 - \alpha)\right)^2 - 4(-16 + 4\alpha)\alpha^2(\alpha - 4) = 48\alpha^2(\alpha^2 - 4) \leq 0.$$

Thus, we have completed the proof for the monotonicity of $\left[\frac{2\Gamma\left(\frac{\alpha}{k}\right)\sin\left(\frac{\pi\alpha}{2k}\right)}{\Gamma\left(\frac{1}{k}\right)\sin\left(\frac{\pi}{k}\right)} \right]^k$.

Next, we will show that, as $k \to \infty$, for any fixed t,

$$\begin{aligned} \mathrm{E}\left(\hat{d}_{(\alpha),gm}\right)^t =& d_{(\alpha)}^t \frac{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha t}{k}\right)\Gamma\left(1 - \frac{t}{k}\right)\sin\left(\frac{\pi\alpha t}{2k}\right)\right]^k}{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1 - \frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^{kt}} \qquad (-k/\alpha < t < k) \\ =& d_{(\alpha)}^t \exp\left(\frac{1}{k}\frac{\pi^2}{6}\frac{(t^2 - t)(2 + \alpha^2)}{4} + \frac{1}{k^2}\zeta_3 \frac{(t^3 - t)(1 - \alpha^3)}{3} + O\left(\frac{1}{k^3}\right) \right), \end{aligned}$$

where $\zeta_3 = \sum_{s=1}^{\infty} \frac{1}{s^3} = 1.2020569...$ is Apery's constant.

We again resort to Euler's reflection formula, the infinity-product representations

of the Gamma function, and the sine function, to obtain

$$\frac{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha t}{k}\right)\Gamma\left(1-\frac{t}{k}\right)\sin\left(\frac{\pi\alpha t}{2k}\right)\right]^k}{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^{kt}} = \left(\prod_{s=1}^{\infty}\left(1+\frac{\alpha t}{sk}\right)^{-1}\left(1+\frac{1}{sk}\right)^{-t}\left(1+\frac{t}{sk}\right)\left(1+\frac{\alpha}{sk}\right)^{t}\right)^k \times \left(\prod_{s=1}^{\infty}\left(1-\frac{1}{s^2k^2}\right)^{t}\left(1-\frac{\alpha^2t^2}{4s^2k^2}\right)\left(1-\frac{t^2}{s^2k^2}\right)^{-1}\left(1-\frac{\alpha^2}{4s^2k^2}\right)^{-t}\right)^k$$

$$= \left(\prod_{s=1}^{\infty}\left(1+\frac{1}{s^2k^2}\frac{t^2-t}{4}(\alpha^2+2)+\frac{1}{s^3k^3}\frac{t^3-t}{3}(1-\alpha^3)+O\left(\frac{1}{k^4}\right)\right)\right)^k$$

$$= \exp\left(k\sum_{s=1}^{\infty}\log\left(1+\frac{1}{s^2k^2}\frac{t^2-t}{4}(\alpha^2+2)+\frac{1}{s^3k^3}\frac{t^3-t}{3}(1-\alpha^3)+O\left(\frac{1}{k^4}\right)\right)\right)$$

$$= \exp\left(\frac{1}{k}\frac{t^2-t}{4}(\alpha^2+2)\sum_{s=1}^{\infty}\frac{1}{s^2}+\frac{1}{k^2}\frac{t^3-t}{3}(1-\alpha^3)\sum_{s=1}^{\infty}\frac{1}{s^3}+O\left(\frac{1}{k^3}\right)\right)$$

$$= \exp\left(\frac{1}{k}\frac{t^2-t}{4}(\alpha^2+2)\frac{\pi^2}{6}+\frac{1}{k^2}\frac{t^3-t}{3}(1-\alpha^3)\zeta_3+O\left(\frac{1}{k^3}\right)\right)$$

$$= 1+\frac{1}{k}\frac{t^2-t}{4}(\alpha^2+2)\frac{\pi^2}{6}+\frac{1}{k^2}\frac{t^3-t}{3}(1-\alpha^3)\zeta_3+\frac{1}{2k^2}\left(\frac{t^2-t}{4}(\alpha^2+2)\frac{\pi^2}{6}\right)^2+O\left(\frac{1}{k^3}\right)$$

$$= 1+\frac{1}{k}\frac{\pi^2}{24}(t^2-t)(\alpha^2+2)+\frac{1}{k^2}\left(\frac{\zeta_3}{3}(t^3-t)(1-\alpha^3)+\frac{1}{2}\left(\frac{\pi^2}{24}(t^2-t)(\alpha^2+2)\right)^2\right)+O\left(\frac{1}{k^3}\right)$$

Now it becomes straightforward to derive the asymptotic expressions for the variance, third, and fourth central moments.

$$\begin{aligned}\mathrm{Var}\left(\hat{d}_{(\alpha),gm}\right) =& \mathrm{E}\left(\hat{d}^2_{(\alpha),gm}\right)-d^2_{(\alpha)}\\ =& d^2_{(\alpha)}\left(\frac{1}{k}\frac{\pi^2}{12}(\alpha^2+2)+\frac{1}{k^2}\left(2\zeta_3(1-\alpha^3)+\frac{\pi^4}{288}(\alpha^2+2)^2\right)+O\left(\frac{1}{k^3}\right)\right)\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left(\hat{d}_{(\alpha),gm} - d_{(\alpha),gm}\right)^3 &= \mathrm{E}\hat{d}^3_{(\alpha),gm} - 3d_{(\alpha)}\mathrm{E}\hat{d}^2_{(\alpha),gm} + 2d^3_{(\alpha)} \\
&= \frac{d^3_{(\alpha)}}{k^2}\left(2(1-\alpha^3)\zeta_3 + \frac{\pi^4}{48}(2+\alpha^2)^2\right) + O\left(\frac{1}{k^3}\right) \\
\mathrm{E}\left(\hat{d}_{(\alpha),gm} - d_{(\alpha),gm}\right)^4 &= \mathrm{E}\hat{d}^4_{(\alpha),gm} - 4d_{(\alpha)}\mathrm{E}\hat{d}^3_{(\alpha),gm} + 6d^2_{(\alpha)}\mathrm{E}\hat{d}^2_{(\alpha),gm} - 3d^4_{(\alpha)} \\
&= \frac{d^4_{(\alpha)}}{k^2}\left(\frac{\pi^4}{48}(2+\alpha^2)^2\right) + O\left(\frac{1}{k^3}\right).
\end{aligned}$$

This completes of the proof of Lemma 21.

6.9.2 Proof of Lemma 22

We first find the constant $G_{R,\alpha,\epsilon}$ in the right tail bound

$$\mathbf{Pr}\left(\hat{d}_{(\alpha),gm} - d_{(\alpha)} > \epsilon d_{(\alpha)}\right) \leq \exp\left(-k\frac{\epsilon^2}{G_{R,\alpha,\epsilon}}\right), \quad 0 < \epsilon < \exp\left(\frac{\pi^2}{12}(2+\alpha^2)\right) - 1.$$

Since $\hat{d}_{(\alpha),gm}$ only has moments less than k, we have to use the Markov moment bound instead of the Chernoff bound. The Markov moment bound yields

$$\begin{aligned}
\mathbf{Pr}\left(\hat{d}_{(\alpha),gm} - d_{(\alpha)} > \epsilon d_{(\alpha)}\right) \leq & \frac{\mathrm{E}\left(\hat{d}_{(\alpha),gm}\right)^t}{(1+\epsilon)^t d^t_{(\alpha)}}, \quad (0 < t < k). \\
= & (1+\epsilon)^{-t}\frac{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha t}{k}\right)\Gamma\left(1-\frac{t}{k}\right)\sin\left(\frac{\pi\alpha t}{2k}\right)\right]^k}{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^{kt}}. \qquad (6.58)
\end{aligned}$$

Ideally, we would like to find the "optimal" t to minimize the moment bound. However, the expression (6.58) is too complicated to be exactly optimized analytically. We resort to a "sub-optimal" (but asymptotically optimal) solution using asymptotic expression of $\mathrm{E}\left(\hat{d}_{(\alpha),gm}\right)^t$ we have derived in Lemma 21, i.e.,

$$\frac{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha t}{k}\right)\Gamma\left(1-\frac{t}{k}\right)\sin\left(\frac{\pi\alpha t}{2k}\right)\right]^k}{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^{kt}} = \exp\left(\frac{1}{k}\frac{\pi^2}{6}\frac{(t^2-t)(2+\alpha^2)}{4} + O\left(\frac{1}{k^2}\right)\right).$$

Therefore, for a sub-optimal moment bound, we can find the t which minimizes

$$(1+\epsilon)^{-t}\exp\left(\frac{1}{k}\frac{\pi^2}{6}\frac{(t^2-t)\,(2+\alpha^2)}{4}\right),$$

whose minimum is attained at

$$t=\frac{\log(1+\epsilon)}{\frac{\pi^2}{12}(2+\alpha^2)}k+\frac{1}{2}\approx\frac{\log(1+\epsilon)}{c_\alpha}k,\qquad c_\alpha=\frac{\pi^2}{12}(2+\alpha^2).$$

In order to use the approximate optimum $t^*=\frac{\log(1+\epsilon)}{c_\alpha}k$, we need to check

$$\frac{\log(1+\epsilon)}{c_\alpha}k<k,\ \Longrightarrow\epsilon<\exp\left(\frac{\pi^2}{12}(2+\alpha^2)\right)-1=\begin{cases}4.18 & \text{if } \alpha=0\\ 10.79 & \text{if } \alpha=1\\ 138.04 & \text{if } \alpha=2\end{cases}.$$

Plugging $t^*=\frac{\log(1+\epsilon)}{c_\alpha}k$ into the moment bound (6.58), we obtain

$$\begin{aligned}&\mathbf{Pr}\left(\hat{d}_{(\alpha),gm}-d_{(\alpha)}>\epsilon d_{(\alpha)}\right)\\ \leq&(1+\epsilon)^{-t^*}\frac{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha t^*}{k}\right)\Gamma\left(1-\frac{t^*}{k}\right)\sin\left(\frac{\pi\alpha t^*}{2k}\right)\right]^k}{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^{kt^*}}\\ \leq&(1+\epsilon)^{-t^*}\frac{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha t^*}{k}\right)\Gamma\left(1-\frac{t^*}{k}\right)\sin\left(\frac{\pi\alpha t^*}{2k}\right)\right]^k}{\exp\left(-t^*\gamma_e(\alpha-1)\right)}\\ =&\exp\left(-k\frac{\epsilon^2}{G_{R,\alpha,\epsilon}}\right),\end{aligned}$$

where

$$\begin{aligned}\frac{1}{G_{R,\alpha,\epsilon}}=&\frac{\log^2(1+\epsilon)}{\epsilon^2 c_\alpha}-\frac{\log(1+\epsilon)}{\epsilon^2 c_\alpha}\gamma_e(\alpha-1)\\ &-\frac{1}{\epsilon^2}\log\left(\frac{2}{\pi}\Gamma\left(\frac{\alpha\log(1+\epsilon)}{c_\alpha}\right)\Gamma\left(1-\frac{\log(1+\epsilon)}{c_\alpha}\right)\sin\left(\frac{\pi\alpha\log(1+\epsilon)}{2c_\alpha}\right)\right).\end{aligned}$$

Next, we will find the constant $G_{L,\alpha,\epsilon,k_0}$ in the left tail bound

$$\mathbf{Pr}\left(\hat{d}_{(\alpha),gm} - d_{(\alpha)} < -\epsilon d_{(\alpha)}\right) \leq \exp\left(-k\frac{\epsilon^2}{G_{L,\alpha,\epsilon,k_0}}\right), \quad 0 < \epsilon < 1, k > k_0 > \frac{1}{\alpha}.$$

From Lemma 21, we know that, for any $0 < t < k/\alpha$,

$$\begin{aligned}
&\mathbf{Pr}\left(\hat{d}_{(\alpha),gm} \leq (1-\epsilon)d_{(\alpha)}\right) = \mathbf{Pr}\left(\hat{d}^{-t}_{(\alpha),gm} \geq (1-\epsilon)^{-t}d^{-t}_{(\alpha)}\right) \\
\leq &\frac{\mathrm{E}\left(\hat{d}^{-t}_{(\alpha),gm}\right)}{(1-\epsilon)^{-t}d^{-t}_{(\alpha)}} = (1-\epsilon)^t \frac{\left[-\frac{2}{\pi}\Gamma\left(-\frac{\alpha t}{k}\right)\Gamma\left(1+\frac{t}{k}\right)\sin\left(\frac{\pi\alpha t}{2k}\right)\right]^k}{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^{-kt}}
\end{aligned} \tag{6.59}$$

Again, we resort to a sub-optimal solution by finding the t that minimizes

$$(1-\epsilon)^t \exp\left(\frac{1}{k}\frac{\pi^2}{6}\frac{(t^2+t)(2+\alpha^2)}{4}\right),$$

whose minimum is attained at

$$t = -\frac{\log(1-\epsilon)}{\frac{\pi^2}{12}(2+\alpha^2)}k - \frac{1}{2} \approx \frac{-\log(1-\epsilon)}{c_\alpha}k.$$

Suppose we consider $t = \frac{-\log(1-\epsilon)}{c_\alpha}k$. In order to satisfy $t < k/\alpha$, we need to make sure that $\epsilon < 1 - \exp\left(-\frac{c_\alpha}{\alpha}\right)$, which may be somewhat restrictive. We will instead consider the Taylor approximation $t = t^{**} = \frac{\epsilon}{c_\alpha}k$, which imposes no additional restrictions on $0 < \epsilon < 1$.

Plugging $t^{**} = \frac{\epsilon}{c_\alpha} k$ into the moment bound (6.59), we obtain

$$\begin{aligned}
&\mathbf{Pr}\left(\hat{d}_{(\alpha),gm} - d_{(\alpha)} < -\epsilon d_{(\alpha)}\right) \\
\leq &(1-\epsilon)^{t^{**}} \frac{\left[-\frac{2}{\pi}\Gamma\left(-\frac{\alpha t^{**}}{k}\right)\Gamma\left(1+\frac{t^{**}}{k}\right)\sin\left(\frac{\pi\alpha t^{**}}{2k}\right)\right]^k}{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k}\right)\Gamma\left(1-\frac{1}{k}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k}\right)\right]^{-kt^{**}}} \\
\leq &(1-\epsilon)^{t^{**}} \frac{\left[-\frac{2}{\pi}\Gamma\left(-\frac{\alpha t^{**}}{k}\right)\Gamma\left(1+\frac{t^{**}}{k}\right)\sin\left(\frac{\pi\alpha t^{**}}{2k}\right)\right]^k}{\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k_0}\right)\Gamma\left(1-\frac{1}{k_0}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k_0}\right)\right]^{-k_0 t^{**}}} \\
= &\exp\left(-k\frac{\epsilon^2}{G_{L,\alpha,\epsilon,k_0}}\right) \qquad\qquad (k > k_0),
\end{aligned}$$

where

$$\begin{aligned}
\frac{1}{G_{L,\alpha,\epsilon,k_0}} = &-\frac{1}{c_\alpha \epsilon}\log(1-\epsilon) - \frac{1}{\epsilon^2}\log\left(-\frac{2}{\pi}\Gamma\left(1+\frac{\epsilon}{c_\alpha}\right)\Gamma\left(-\frac{\alpha\epsilon}{c_\alpha}\right)\sin\left(\frac{\pi}{2}\frac{\alpha\epsilon}{c_\alpha}\right)\right) \\
&- \frac{1}{c_\alpha\epsilon}k_0\log\left(\left[\frac{2}{\pi}\Gamma\left(\frac{\alpha}{k_0}\right)\Gamma\left(1-\frac{1}{k_0}\right)\sin\left(\frac{\pi}{2}\frac{\alpha}{k_0}\right)\right]\right).
\end{aligned}$$

We have completed the proof.

6.9.3 Proof of Lemma 24

Assume $z \sim h/E_1$, where E_1 stands for the exponential distribution with mean 1. The log likelihood, $l(z;h)$, and first three derivatives (w.r.t. h) are

$$l(z;h) = -\frac{h}{z} + \log(h) - 2\log(z), \quad l'(h) = \frac{1}{h} - \frac{1}{z}, \quad l''(h) = -\frac{1}{h^2}, \quad l'''(d) = \frac{2}{h^3}.$$

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

$$\mathrm{I} = \mathrm{I}(h) = \mathrm{E}\left(-l''(h)\right) = \frac{1}{h^2}.$$

Using the formulas for the bias and higher moments of the MLE in [23, 154]:

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

where, after re-formatting,

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

Note that, for any integer $m > 0$,

$$\mathrm{E}\left(\frac{h}{z}\right)^m = \int_0^\infty \left(\frac{h}{z}\right)^m \exp\left(-\frac{h}{z}\right)\frac{h}{z^2}dz$$
$$= \int_0^\infty s^m \exp(-s)ds = m\int_0^\infty s^{m-1}\exp(-s)ds = m!,$$

from which it follows that

$$\mathrm{E}\,(l')^3 = -\frac{2}{h^3}, \quad \mathrm{E}\,(l'l'') = 0, \quad \mathrm{E}\,(l')^4 = \frac{9}{h^4}, \quad \mathrm{E}(l''(l')^2) = -\frac{1}{h^4}, \quad \mathrm{E}\,(l'l''') = 0.$$

Hence

$$[12] = -\frac{2}{h^3}, \quad [1^4] = \frac{9}{h^4}, \quad [1^2 2] = \frac{8}{h^4}, \quad [13] = \frac{6}{h^4}, \quad [1^3] = -\frac{2}{h^3}.$$

Thus, we obtain

$$\mathrm{E}\left(\hat{h}_{mle}\right) = h + \frac{h}{k} + O\left(\frac{1}{k^2}\right),$$
$$\mathrm{Var}\left(\hat{h}_{mle}\right) = \frac{h^2}{k} + \frac{4h^2}{k^2} + O\left(\frac{1}{k^3}\right).$$

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

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

whose first two moments are

$$\mathrm{E}\left(\hat{h}_{mle,c}\right) = h + O\left(\frac{1}{k^2}\right),$$
$$\mathrm{Var}\left(\hat{h}_{mle,c}\right) = \left(1 - \frac{1}{k}\right)^2\left(\frac{h^2}{k} + \frac{4h^2}{k^2}\right) + O\left(\frac{1}{k^3}\right) = \frac{h^2}{k} + \frac{2h^2}{k^2} + O\left(\frac{1}{k^3}\right).$$

Now we prove the tail bounds for $\hat{h}_{mle,c}$. For $t > 0$,

$$\mathbf{Pr}\left(\hat{h}_{mle,c} - h > \epsilon h\right) = \mathbf{Pr}\left(\frac{k-1}{\sum_{j=1}^k \frac{1}{z_j}} > (1+\epsilon)h\right) = \mathbf{Pr}\left(-\sum_{j=1}^k \frac{1}{z_j}t > -t\frac{k-1}{(1+\epsilon)h}\right)$$
$$\leq \left(\prod_{j=1}^k \mathrm{E}\left(\exp\left(\frac{-t}{z_j}\right)\right)\right)\exp\left(t\frac{k-1}{(1+\epsilon)h}\right) = \left(\frac{h}{h+t}\right)^k \exp\left(t\frac{k-1}{(1+\epsilon)h}\right)$$
$$= \exp\left(k\log\left(\frac{h}{h+t}\right) + t\frac{k-1}{(1+\epsilon)h}\right)$$

whose minimum is attained at $t = h\left(\epsilon + \frac{1+\epsilon}{k-1}\right)$ and therefore,

$$\mathbf{Pr}\left(\hat{h}_{mle,c} - h > \epsilon h\right) \leq \exp\left(-k\left(\log(1+\epsilon) - \frac{\epsilon}{1+\epsilon} + \log\left(\frac{k}{k-1}\right) - \frac{1}{k}\frac{1}{1+\epsilon}\right)\right)$$

$$\mathbf{Pr}\left(\hat{h}_{mle,c} - h < -\epsilon h\right) = \mathbf{Pr}\left(\frac{k-1}{\sum_{j=1}^k \frac{1}{z_j}} < (1-\epsilon)h\right) = \mathbf{Pr}\left(\sum_{j=1}^k \frac{1}{z_j}t > t\frac{k-1}{(1-\epsilon)h}\right)$$
$$\leq \left(\prod_{j=1}^k \mathrm{E}\left(\exp\left(\frac{t}{z_j}\right)\right)\right)\exp\left(-t\frac{k-1}{(1-\epsilon)h}\right) = \left(\frac{h}{h-t}\right)^k \exp\left(-t\frac{k-1}{(1-\epsilon)h}\right)$$
$$= \exp\left(k\log\left(\frac{h}{h-t}\right) - t\frac{k-1}{(1-\epsilon)h}\right)$$

whose minimum is attained at $t = h\left(\epsilon - \frac{1-\epsilon}{k-1}\right)$ and therefore,

$$\mathbf{Pr}\left(\hat{h}_{mle,c} - h < -\epsilon h\right) \leq \exp\left(-k\left(\log(1-\epsilon) + \frac{\epsilon}{1-\epsilon} + \log\left(\frac{k}{k-1}\right) - \frac{1}{k}\frac{1}{1-\epsilon}\right)\right)$$

This completes the proof of Lemma 24.

6.9.4 Proof of Lemma 25

Suppose $x_j \sim S(\alpha, d_{(\alpha)})$, $j = 1, 2, ..., k$. Then if $0 < \alpha < 1$, we know from Lemma 20

$$\mathrm{E}\left(|x_j|^{-\alpha}\right) = \frac{1}{d_{(\alpha)}}\left(-\frac{2}{\pi}\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right),$$

which suggests an unbiased estimator for $\frac{1}{d_{(\alpha)}}$, denoted by $\hat{R}_{(\alpha)}$,

$$\hat{R}_{(\alpha)} = \frac{1}{k}\sum_{j=1}^{k}\frac{|x_j|^{-\alpha}}{-\frac{2}{\pi}\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)}, \quad (0 < \alpha < 1)$$

$$\mathrm{Var}\left(\hat{R}_{(\alpha)}\right) = \frac{1}{d_{(\alpha)}^2}\frac{1}{k}\left(\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2} - 1\right), \quad (0 < \alpha < 1/2)$$

Once we have an unbiased estimator for $\frac{1}{d_{(\alpha)}}$, we can estimate $d_{(\alpha)}$ by $\frac{1}{\hat{R}_{(\alpha)}}$, which will be asymptotically unbiased. We can remove the $O\left(\frac{1}{k}\right)$ bias of $\frac{1}{\hat{R}_{(\alpha)}}$ by Taylor expansions[2]. See [110, Theorem 6.1.1].

Ideally, we would like to estimate $\hat{d}_{(\alpha)}$ from $\frac{1}{\hat{R}_{(\alpha)}}$ with bias corrections:

$$\frac{1}{\hat{R}_{(\alpha)}} - \frac{\mathrm{Var}\left(\hat{R}_{(\alpha)}\right)}{2}\left(\frac{2}{d_{(\alpha)}^{-3}}\right) = \frac{1}{\hat{R}_{(\alpha)}} - \frac{1}{k}\left(\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2} - 1\right)d_{(\alpha)}$$

We have to replace $d_{(\alpha)}$ by the estimated value, i.e., $\frac{1}{\hat{R}_{(\alpha)}}$. This way, we obtain a

[2]We will need the first and second derivatives of $\frac{1}{x}$, i.e., $\left(\frac{1}{x}\right)' = \frac{-1}{x^2}$, $\left(\frac{1}{x}\right)'' = \frac{2}{x^3}$.

harmonic mean type of estimator with only $O\left(\frac{1}{k^2}\right)$ bias.

$$\hat{d}_{(\alpha),hm} = \frac{1}{\frac{1}{k}\sum_{j=1}^{k}\frac{|x_j|^{-\alpha}}{-\frac{2}{\pi}\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)}}\left(1-\frac{1}{k}\left(\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2}-1\right)\right)$$
$$= \frac{-\frac{2}{\pi}\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)}{\sum_{j=1}^{k}|x_j|^{-\alpha}}\left(k-\left(\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2}-1\right)\right)$$

The asymptotic variance would be

$$\mathrm{Var}\left(\hat{d}_{(\alpha),hm}\right) = \mathrm{Var}\left(\hat{R}_{(\alpha)}\right)\left(-\frac{1}{d_{(\alpha)}^{-2}}\right)^2 + O\left(\frac{1}{k^2}\right)$$
$$= d_{(\alpha)}^2\frac{1}{k}\left(\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2}-1\right)+O\left(\frac{1}{k^2}\right)$$

As $\alpha \to 0+$, it is easy to show that

$$\lim_{\alpha\to 0+} -\frac{2}{\pi}\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right) = 1,$$
$$\lim_{\alpha\to 0+}\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2}-1 = 1.$$

Finally, we prove that there exists a unique α_0 $(0 < \alpha_0 < 1/2)$, such that

$$\left(\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2}-1\right) < \frac{\pi^2}{12}(\alpha^2+2), \quad \forall 0<\alpha<\alpha_0.$$

Because

$$\left(\frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2}-1\right)-\frac{\pi^2}{12}(\alpha^2+2) = \begin{cases} 1-\frac{\pi^2}{6}<0 & \text{if } \alpha\to 0+ \\ +\infty & \text{if } \alpha\to 0.5- \end{cases},$$

we only need to show that

$$g(\alpha) = \frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2}-\frac{\pi^2}{12}\alpha^2,$$

increases monotonically with increasing α, i.e., $g'(\alpha) > 0$ for $0 < \alpha < 1/2$.

To overcome the difficulty in finding $g'(\alpha)$, we again resort to the infinite-product representations of the Gamma function and the sine function to obtain

$$p(\alpha) = \frac{-\pi\Gamma(-2\alpha)\sin(\pi\alpha)}{\left[\Gamma(-\alpha)\sin\left(\frac{\pi}{2}\alpha\right)\right]^2} = 2\prod_{s=1}^{\infty} f_s(\alpha)$$
$$f_s(\alpha) = \left(1-\frac{2\alpha}{s}\right)^{-1}\left(1-\frac{\alpha}{s}\right)^2\left(1-\frac{\alpha^2}{s^2}\right)\left(1-\frac{\alpha^2}{4s^2}\right)^{-2}.$$

Because

$$\frac{\partial \log f_s(\alpha)}{\partial \alpha} = \frac{2\alpha\left(\alpha^3 - 2\alpha^2 s + 8\alpha s^2 + 2s^3\right)}{(s^2-\alpha^2)(s-2\alpha)(4s^2-\alpha^2)} > 0,$$

we know $f_s(\alpha)$ is monotonically increasing with increasing α, which implies that $p(\alpha)$ is monotonically increasing, i.e., $p(\alpha) > p(0+) = 2$.

To show the monotonicity of $g(\alpha)$, it suffices to show $g'(\alpha) > 0$, i.e.,

$$p(\alpha)\sum_{s=1}^{\infty}\frac{\partial \log f_s(\alpha)}{\partial \alpha} - \frac{\pi^2}{6}\alpha > 0,$$

for which it suffices to show that

$$2\sum_{s=1}^{\infty}\frac{2\alpha\left(\alpha^3 - 2\alpha^2 s + 8\alpha s^2 + 2s^3\right)}{(s^2-\alpha^2)(s-2\alpha)(4s^2-\alpha^2)} - \frac{\pi^2}{6}\alpha > 0,$$

which is true, because

$$\sum_{s=1}^{\infty}\frac{2\alpha\left(\alpha^3 - 2\alpha^2 s + 8\alpha s^2 + 2s^3\right)}{(s^2-\alpha^2)(s-2\alpha)(4s^2-\alpha^2)} > \alpha\sum_{s=1}^{\infty}\frac{4s^3}{4s^5} = \alpha\sum_{s=1}^{\infty}\frac{1}{s^2} = \alpha\frac{\pi^2}{6}.$$

Therefore, we have proved $g'(\alpha) > 0$, i.e., the monotonicity of $g(\alpha)$. This completes the proof of Lemma 25.

6.9.5 Proof of Lemma 26

By Lemma 20, we first seek an unbiased estimator of $d_{(\alpha)}^\lambda$, denoted by $\hat{R}_{(\alpha),\lambda}$,

$$\hat{R}_{(\alpha),\lambda} = \frac{1}{k} \frac{\sum_{j=1}^k |x_j|^{\lambda\alpha}}{\frac{2}{\pi}\Gamma(1-\lambda)\Gamma(\lambda\alpha)\sin\left(\frac{\pi}{2}\lambda\alpha\right)}, \qquad -1/\alpha < \lambda < 1$$

whose variance is

$$\mathrm{Var}\left(\hat{R}_{(\alpha),\lambda}\right) = \frac{d_{(\alpha)}^{2\lambda}}{k}\left(\frac{\frac{2}{\pi}\Gamma(1-2\lambda)\Gamma(2\lambda\alpha)\sin\left(\pi\lambda\alpha\right)}{\left[\frac{2}{\pi}\Gamma(1-\lambda)\Gamma(\lambda\alpha)\sin\left(\frac{\pi}{2}\lambda\alpha\right)\right]^2} - 1\right), \qquad -\frac{1}{2\alpha} < \lambda < \frac{1}{2}$$

A biased estimator of $d_{(\alpha)}$ would be simply $\left(\hat{R}_{(\alpha),\lambda}\right)^{1/\lambda}$, which has $O\left(\frac{1}{k}\right)$ bias. This bias can be removed to an extent by Taylor expansions [110, Theorem 6.1.1]. While it is well-known that bias-corrections are not always beneficial because of the bias-variance trade-off phenomenon, in our case, it is a good idea to conduct the bias-correction because the function $f(x) = x^{1/\lambda}$ is convex for $x > 0$. Note that $f'(x) = \frac{1}{\lambda}x^{1/\lambda-1}$ and $f''(x) = \frac{1}{\lambda}\left(\frac{1}{\lambda}-1\right)x^{1/\lambda-2} > 0$, assuming $-\frac{1}{2\alpha} < \lambda < \frac{1}{2}$. Because $f(x)$ is convex, removing the $O\left(\frac{1}{k}\right)$ bias will also lead to a smaller variance.

We call this new estimator the "fractional power" estimator:

$$\hat{d}_{(\alpha),fp,\lambda} = \left(\hat{R}_{(\alpha),\lambda}\right)^{1/\lambda} - \frac{\mathrm{Var}\left(\hat{R}_{(\alpha),\lambda}\right)}{2}\frac{1}{\lambda}\left(\frac{1}{\lambda}-1\right)\left(d_{(\alpha)}^\lambda\right)^{1/\lambda-2}$$

$$= \left(\frac{1}{k}\frac{\sum_{j=1}^k |x_j|^{\lambda\alpha}}{\frac{2}{\pi}\Gamma(1-\lambda)\Gamma(\lambda\alpha)\sin\left(\frac{\pi}{2}\lambda\alpha\right)}\right)^{1/\lambda}\left(1 - \frac{1}{k}\frac{1}{2\lambda}\left(\frac{1}{\lambda}-1\right)\left(\frac{\frac{2}{\pi}\Gamma(1-2\lambda)\Gamma(2\lambda\alpha)\sin\left(\pi\lambda\alpha\right)}{\left[\frac{2}{\pi}\Gamma(1-\lambda)\Gamma(\lambda\alpha)\sin\left(\frac{\pi}{2}\lambda\alpha\right)\right]^2} - 1\right)\right),$$

where we plug in the estimated $d_{(\alpha)}^\lambda$. The asymptotic variance would be

$$\mathrm{Var}\left(\hat{d}_{(\alpha),fp,\lambda}\right) = \mathrm{Var}\left(\hat{R}_{(\alpha),\lambda}\right)\left(\frac{1}{\lambda}\left(d_{(\alpha)}^\lambda\right)^{1/\lambda-1}\right)^2 + O\left(\frac{1}{k^2}\right)$$

$$= d_{(\alpha)}^2 \frac{1}{\lambda^2 k}\left(\frac{\frac{2}{\pi}\Gamma(1-2\lambda)\Gamma(2\lambda\alpha)\sin\left(\pi\lambda\alpha\right)}{\left[\frac{2}{\pi}\Gamma(1-\lambda)\Gamma(\lambda\alpha)\sin\left(\frac{\pi}{2}\lambda\alpha\right)\right]^2} - 1\right) + O\left(\frac{1}{k^2}\right).$$

The optimal λ, denoted by λ^*, is then

$$\lambda^* = \underset{-\frac{1}{2\alpha}\lambda<\frac{1}{2}}{\text{argmin}} \left\{ \frac{1}{\lambda^2} \left(\frac{\frac{2}{\pi}\Gamma(1-2\lambda)\Gamma(2\lambda\alpha)\sin(\pi\lambda\alpha)}{\left[\frac{2}{\pi}\Gamma(1-\lambda)\Gamma(\lambda\alpha)\sin\left(\frac{\pi}{2}\lambda\alpha\right)\right]^2} - 1 \right) \right\}.$$

6.9.6 Proof of Lemma 27

Part 1

In this section, we prove that $g(\lambda;\alpha) = \frac{1}{\lambda^2}\left(\frac{\frac{2}{\pi}\Gamma(1-2\lambda)\Gamma(2\lambda\alpha)\sin(\pi\lambda\alpha)}{\left[\frac{2}{\pi}\Gamma(1-\lambda)\Gamma(\lambda\alpha)\sin\left(\frac{\pi}{2}\lambda\alpha\right)\right]^2} - 1\right)$ is a convex function of λ, for $0 < \alpha \leq 2$. We need to show that $\frac{\partial^2 g(\lambda;\alpha)}{\partial\lambda^2} > 0$.

We use the infinite-product representations of the Gamma and the sine functions,

$$\Gamma(z) = \frac{\exp(-\gamma_e z)}{z}\prod_{s=1}^{\infty}\left(1+\frac{z}{s}\right)^{-1}\exp\left(\frac{z}{s}\right), \qquad \sin(z) = z\prod_{s=1}^{\infty}\left(1-\frac{z^2}{s^2\pi^2}\right),$$

to obtain

$$g(\lambda;\alpha) = \frac{1}{\lambda^2}\left(M(\lambda;\alpha)-1\right) = \frac{1}{\lambda^2}\left(\prod_{s=1}^{\infty} f_s(\lambda;\alpha) - 1\right),$$

where

$$f_s(\lambda;\alpha) = \left(1-\frac{\lambda}{s}\right)^2\left(1+\frac{2\lambda\alpha}{s}\right)^{-1}\left(1-\frac{\lambda\alpha}{s}\right)\left(1+\frac{\lambda\alpha}{s}\right)^3\left(1-\frac{\lambda^2\alpha^2}{4s^2}\right)^{-2}\left(1-\frac{2\lambda}{s}\right)^{-1}.$$

With respect to λ, the first two derivatives of $g(\lambda;\alpha)$ are

$$\frac{\partial g}{\partial\lambda} = \frac{1}{\lambda^2}\left(-\frac{2}{\lambda}(M-1) + \sum_{s=1}^{\infty}\frac{\partial\log f_s}{\partial\lambda}M\right)$$

$$\frac{\partial^2 g}{\partial\lambda^2} = \frac{M}{\lambda^2}\left(\frac{6}{\lambda^2} + \sum_{s=1}^{\infty}\frac{\partial^2\log f_s}{\partial\lambda^2} + \left(\sum_{s=1}^{\infty}\frac{\partial\log f_s}{\partial\lambda}\right)^2 - \frac{4}{\lambda}\sum_{s=1}^{\infty}\frac{\partial\log f_s}{\partial\lambda}\right) - \frac{6}{\lambda^4},$$

and

$$\frac{\partial M}{\partial \lambda} = M \sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda}, \qquad \frac{\partial^2 M}{\partial \lambda} = M \left(\sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2} + \left(\sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda} \right)^2 \right)$$

$$\sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda} = 2\lambda \sum_{s=1}^{\infty} \frac{1}{s^2 - 3s\lambda + 2\lambda^2} + \alpha^2 \left(\frac{2}{4s^2 - \lambda^2\alpha^2} + \frac{1}{s^2 + 3s\lambda\alpha + 2\lambda^2\alpha^2} - \frac{1}{s^2 - \lambda^2\alpha^2} \right),$$

$$\sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2} = \sum_{s=1}^{\infty} -\frac{2}{(s-\lambda)^2} + \frac{4}{(s-2\lambda)^2} + \alpha^2 \left(\frac{2}{(2s-\lambda\alpha)^2} - \frac{1}{(s-\lambda\alpha)^2} - \frac{3}{(s+\lambda\alpha)^2} + \frac{4}{(s+2\lambda\alpha)^2} + \frac{2}{(2s+\lambda\alpha)^2} \right),$$

$$\sum_{s=1}^{\infty} \frac{\partial^3 \log f_s}{\partial \lambda^3} = \sum_{s=1}^{\infty} \frac{-4}{(s-\lambda)^3} + \frac{16}{(s-2\lambda)^3} + 2\alpha^3 \left(\frac{2}{(2s-\lambda\alpha)^3} - \frac{1}{(s-\lambda\alpha)^3} + \frac{3}{(s+\lambda\alpha)^3} - \frac{8}{(s+2\lambda\alpha)^3} - \frac{2}{(2s+\lambda\alpha)^3} \right),$$

$$\sum_{s=1}^{\infty} \frac{\partial^4 \log f_s}{\partial \lambda^4} = \sum_{s=1}^{\infty} -\frac{-12}{(s-\lambda)^4} + \frac{96}{(s-2\lambda)^4} + 6\alpha^4 \left(\frac{2}{(2s-\lambda\alpha)^4} - \frac{1}{(s-\lambda\alpha)^4} - \frac{3}{(s+\lambda\alpha)^4} + \frac{16}{(s+2\lambda\alpha)^4} + \frac{2}{(2s+\lambda\alpha)^4} \right).$$

We will soon show that $\lambda \sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda} > 0$, and $\sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2} > 0$, assuming which suggests $\lambda \frac{\partial M}{\partial \lambda} > 0$ and $\frac{\partial^2 M}{\partial \lambda^2} > 0$. Because $M(0;\alpha) = 1$, we know $M(\lambda;\alpha) > 1$ provided $\lambda \neq 0$.

To show $\frac{\partial^2 g}{\partial \lambda^2} > 0$, it suffices to show that

$$F_1(\lambda;\alpha) = 6(M-1) + \lambda^2 \sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2} + \lambda^2 \left(\sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda} \right)^2 - 4\lambda \sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda} > 0,$$

for which it suffices to show that for $\lambda \neq 0$, $\lambda \frac{\partial F_1}{\partial \lambda} > 0$ (because $F_1(0;\alpha) = 0$), where

$$\begin{aligned} \frac{\partial F_1}{\partial \lambda} =& 6M \sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda} - 4 \sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda} - 2\lambda \sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2} + \lambda^2 \sum_{s=1}^{\infty} \frac{\partial^3 \log f_s}{\partial \lambda^3} \\ & + 2\lambda \left(\sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda} \right)^2 + 2\lambda^2 \sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda} \sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2}. \end{aligned}$$

It suffices to show the second derivative $\frac{\partial^2 F_1}{\partial \lambda^2} > 0$. Because $M > 1$, it actually suffices if we replace $\frac{\partial F_1}{\partial \lambda}$ by F_2, by substituting M with 1. Now it suffices to show

that

$$\frac{\partial F_2}{\partial \lambda} = \lambda^4 \sum_{s=1}^{\infty} \frac{\partial^4 \log f_s}{\partial \lambda^4} + 2\left(\sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2}\right)^2 + 8\lambda \sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda} \sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2} + 2\lambda^2 \left(\sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2}\right)^2 + 2\lambda^2 \sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda} \sum_{s=1}^{\infty} \frac{\partial^3 \log f_s}{\partial \lambda^3} > 0,$$

which is true, because we will soon show that $\lambda \sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda} > 0$, and $\sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2} > 0$, and using similar algebra (skipped) we can analogously show that

$$\sum_{s=1}^{\infty} \frac{\partial^4 \log f_s}{\partial \lambda^4} > 0. \qquad 4\sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2} + \lambda \sum_{s=1}^{\infty} \frac{\partial^3 \log f_s}{\partial \lambda^3} > 0.$$

Therefore, to complete the proof of the convexity of $g(\lambda; \alpha)$, it only remains to show $\sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2} > 0$, which implies $\lambda \sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda} > 0$ for $\lambda \neq 0$. We will proceed with direct algebra.

First, we consider $-\frac{1}{2\alpha} < \lambda < 0$, i.e., $-\frac{1}{2} < \lambda\alpha < 0$. In this case, it is obvious that

$$\frac{2}{(2s-\lambda\alpha)^2} + \frac{2}{(2s+\lambda\alpha)^2} - \frac{1}{(s-\lambda\alpha)^2} > 0, \qquad -\frac{3}{(s+\lambda\alpha)^2} + \frac{3}{(s+2\lambda\alpha)^2} > 0$$

Thus,

$$\begin{aligned}
&\sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2} > \sum_{s=1}^{\infty} -\frac{2}{(s-\lambda)^2} + \frac{4}{(s-2\lambda)^2} \\
&= -\left\{\frac{2}{(1-\lambda)^2} + \frac{2}{(2-\lambda)^2} + \frac{2}{(3-\lambda)^2} + ...\right\} + \left\{\frac{4}{(1-2\lambda)^2} + \frac{4}{(2-2\lambda)^2} + \frac{4}{(3-2\lambda)^2} + ...\right\} \\
&= -\left\{\frac{1}{(1-\lambda)^2} + \frac{1}{(2-\lambda)^2} + \frac{1}{(3-\lambda)^2} + ...\right\} + \left\{\frac{4}{(1-2\lambda)^2} + \frac{4}{(3-2\lambda)^2} + \frac{4}{(5-2\lambda)^2} ...\right\} \\
&= \left\{\frac{4}{(1-2\lambda)^2} - \frac{1}{(1-\lambda)^2}\right\} + \left\{\frac{4}{(3-2\lambda)^2} - \frac{1}{(2-\lambda)^2}\right\} + \left\{\frac{4}{(5-2\lambda)^2} - \frac{1}{(3-\lambda)^2}\right\} + ... \\
&= \sum_{s=1}^{\infty} \frac{4}{(2s-1-2\lambda)^2} - \frac{1}{(s-\lambda)^2} = \sum_{s=1}^{\infty} \frac{4s-1-4\lambda}{(2s-1-2\lambda)^2(s-\lambda)^2} > 0.
\end{aligned}$$

Next, we consider $0 < \lambda < \frac{1}{2}$. Note that $\lambda\alpha < \frac{\alpha}{2} < 1$. Using the similar trick,

$$
\begin{aligned}
&\sum_{s=1}^{\infty} -\frac{3}{(s+\lambda\alpha)^2} + \frac{4}{(s+2\lambda\alpha)^2} + \frac{2}{(2s+\lambda\alpha)^2} \\
=&-\left(\frac{3}{(1+\lambda\alpha)^2} + \frac{3}{(2+\lambda\alpha)^2} + \frac{3}{(3+\lambda\alpha)^2} + \frac{3}{(4+\lambda\alpha)^2} + \frac{3}{(5+\lambda\alpha)^2} + \frac{3}{(6+\lambda\alpha)^2} + ...\right) \\
&+\left(\frac{4}{(1+2\lambda\alpha)^2} + \frac{4}{(2+2\lambda\alpha)^2} + \frac{4}{(3+2\lambda\alpha)^2} + \frac{4}{(4+2\lambda\alpha)^2} + \frac{4}{(5+2\lambda\alpha)^2} + \frac{4}{(6+2\lambda\alpha)^2} + ...\right) \\
&+\left(\frac{2}{(2+\lambda\alpha)^2} + \frac{2}{(4+\lambda\alpha)^2} + \frac{2}{(6+\lambda\alpha)^2} + \frac{2}{(8+\lambda\alpha)^2} + \frac{2}{(10+\lambda\alpha)^2} + \frac{2}{(12+\lambda\alpha)^2} + ...\right) \\
=&-\left(\frac{2}{(1+\lambda\alpha)^2} + \frac{2}{(3+\lambda\alpha)^2} + \frac{2}{(5+\lambda\alpha)^2} + \frac{2}{(7+\lambda\alpha)^2} + \frac{2}{(9+\lambda\alpha)^2} + ...\right) \\
&+\left(\frac{4}{(1+2\lambda\alpha)^2} + \frac{4}{(3+2\lambda\alpha)^2} + \frac{4}{(5+2\lambda\alpha)^2} + \frac{4}{(7+2\lambda\alpha)^2} + \frac{4}{(9+2\lambda\alpha)^2} + ...\right) \\
=&\sum_{s=1}^{\infty} \frac{2(s^2-\lambda^2\alpha^2)}{(s+2\lambda\alpha)^2(s+\lambda\alpha)^2} > 0
\end{aligned}
$$

Thus, to show $\sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2} > 0$ for $0 < \lambda < \frac{1}{2}$, it suffices to show that

$$
\sum_{s=1}^{\infty} -\frac{2}{(s-\lambda)^2} + \frac{4}{(s-2\lambda)^2} + \alpha^2\left(\frac{2}{(2s-\lambda\alpha)^2} - \frac{1}{(s-\lambda\alpha)^2}\right) > 0,
$$

which is true, because

$$
\begin{aligned}
&\sum_{s=1}^{\infty} -\frac{2}{(s-\lambda)^2} + \frac{4}{(s-2\lambda)^2} + \alpha^2\left(\frac{2}{(2s-\lambda\alpha)^2} - \frac{1}{(s-\lambda\alpha)^2}\right) \\
=&\sum_{s=1}^{\infty} \frac{4}{(2s-1-2\lambda)^2} - \frac{1}{(s-\lambda)^2} - \frac{\alpha^2}{4}\left(\frac{4}{(2s-1-2\lambda\alpha)^2} - \frac{1}{(s-\lambda\alpha)^2}\right) \\
=&\sum_{s=1}^{\infty} \left\{\frac{1}{\left(s-\frac{1}{2}-\lambda\right)^2} - \frac{1}{(s-\lambda)^2}\right\} - \frac{\alpha^2}{4}\sum_{s=1}^{\infty}\left\{\frac{1}{\left(s-\frac{1}{2}-\lambda\alpha/2\right)^2} - \frac{1}{(s-\lambda\alpha/2)^2}\right\} \\
>&\sum_{s=1}^{\infty} \left\{\frac{s-\lambda-\frac{1}{4}-\frac{\alpha^2}{4}\left(s-\frac{\lambda\alpha}{2}-\frac{1}{4}\right)}{\left(s-\frac{1}{2}-\lambda\right)^2(s-\lambda)^2}\right\} > 0,
\end{aligned}
$$

because $s-\lambda-\frac{1}{4}-\frac{\alpha^2}{4}\left(s-\frac{\lambda\alpha}{2}-\frac{1}{4}\right) > 0$ for $s \geq 1$, $0 < \alpha < 2$, and $0 < \lambda < \frac{1}{2}$.

In the above steps, we have (implicitly) used the infinite countability, and also the

fact that, for $b > 1$, $\sum_{s=1}^{\infty} \frac{1}{(s-a)^b}$ is the Riemann's Zeta function which is well-defined (and hence we can exchange summation with differentiation).

Therefore, we have proved that $\sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2} > 0$; and hence we have also completed the proof of the convexity of $g(\lambda; \alpha)$.

Part 2

We need to show that the optimal $-1 < \lambda^* < \frac{1}{2}$. It suffices to consider $\alpha < 0.5$.

Setting $\frac{\partial g(\lambda;\alpha)}{\partial \lambda} = 0$ yields

$$h(\lambda^*) = M(\lambda^*) \left(1 - \frac{\lambda^*}{2} \sum_{s=1}^{\infty} \left. \frac{\partial \log f_s}{\partial \lambda} \right|_{\lambda^*} \right) = 1,$$

If we can show that $h(\lambda) < h(-1) < 1$ whenever $\lambda < -1$, we are done.

First, we notice that $\sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda}\big|_{\lambda=-1} < -1$, as $\frac{1}{1+2} - \frac{1}{1+1} + \frac{1}{2+2} - \frac{1}{2+1} + ... = -\frac{1}{2}$.

Assuming $\lambda < -1$, then

$$h'(\lambda) = \frac{M}{2} \left(\sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda} - \lambda \left(\sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda} \right)^2 - \lambda \sum_{s=1}^{\infty} \frac{\partial^2 \log f_s}{\partial \lambda^2} \right) > 0$$

which implies $h(\lambda) < h(-1)$. Thus, it remains to show

$$h(-1) = M(-1; \alpha) \left(1 + \frac{1}{2} \sum_{s=1}^{\infty} \left. \frac{\partial \log f_s}{\partial \lambda} \right|_{-1} \right) < -1.$$

Because α has to be < 0.5 if $\lambda < -1$, and

$$\frac{\partial \sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda}\big|_{-1}}{\partial \alpha} = \sum_{s=1}^{\infty} s \left(-\frac{1}{(s+\alpha)^2} + \frac{4}{(2s+\alpha)^2} + \frac{3}{(s-\alpha)^2} - \frac{4}{(2s-\alpha)^2} - \frac{2}{(s-2\alpha)^2} \right)$$

$$\frac{\partial M(-1;\alpha)}{\partial \alpha} = M(-1;\alpha) \sum_{s=1}^{\infty} \frac{1}{s+\alpha} - \frac{2}{2s+\alpha} - \frac{3}{s-\alpha} + \frac{2}{2s-\alpha} + \frac{2}{s-2\alpha},$$

it is straightforward to show that

$$\frac{\partial h(-1)}{\partial \alpha} = \frac{\partial \left\{ M(-1;\alpha) \left(1 + \frac{1}{2} \sum_{s=1}^{\infty} \frac{\partial \log f_s}{\partial \lambda}\big|_{-1} \right) \right\}}{\partial \alpha} < 0, \qquad 0 < \alpha < 0.5$$

Thus $h(-1;\alpha) < h(-1;0+) = 1$. This completes the proof.

6.9.7 Proof of Lemma 28

To simplify the notation, we let $c_i = \frac{g_i}{\left(\sum_{s=1}^{D} |g_s|^\alpha\right)^{\frac{1}{\alpha}}}$. The fundamental probability theory says that, in order to show the convergence in distribution, it suffices to show the convergence of the characteristic function (Fourier transform of the density function). That is, it suffices to show, as $D \to \infty$,

$$\log\left(\mathrm{E}\left(\exp\left(\sqrt{-1} t \beta^{-1/\alpha} \sum_{i=1}^{D} c_i z_i \right) \right) \right) \to -\Gamma(1-\alpha)\cos\left(\frac{\pi}{2}\alpha\right)|t|^\alpha.$$

By our definition of z_i in (6.54), we have

$$\begin{aligned} &\mathrm{E}\left(\exp(\sqrt{-1} z_i t)\right) \\ =&1 - \beta + \beta \int_1^\infty \frac{\alpha \cos(xt)}{x^{1+\alpha}} dx = 1 - \beta + \alpha\beta \times \\ &\left(\int_0^\infty \frac{\cos(xt) - 1}{x^{1+\alpha}} dx - \int_0^1 \frac{\cos(xt) - 1}{x^{1+\alpha}} dx + \int_1^\infty \frac{1}{x^{1+\alpha}} dx \right) \end{aligned}$$

Using the integral formula [85, 3.823, page 484], we obtain

$$\int_0^\infty \frac{\cos(xt) - 1}{x^{1+\alpha}} dx = -\frac{1}{\alpha}|t|^\alpha \Gamma(1-\alpha)\cos\left(\frac{\pi}{2}\alpha\right).$$

Also, by the Taylor expansion,

$$\int_0^1 \frac{\cos(xt) - 1}{x^{1+\alpha}} dx = -\frac{1}{2}\frac{|t|^2}{2-\alpha} + \frac{1}{4!}\frac{|t|^4}{4-\alpha} ...$$

Combining the results, we obtain

$$\begin{aligned}
&\mathrm{E}\left(\exp(\sqrt{-1}z_i t)\right)\\
=&1-\beta+\beta\left(1-|t|^\alpha\Gamma(1-\alpha)\cos\left(\frac{\pi}{2}\alpha\right)+\frac{\alpha}{2}\frac{|t|^2}{2-\alpha}-\frac{\alpha}{4!}\frac{|t|^4}{4-\alpha}...\right)\\
=&1-\beta|t|^\alpha\Gamma(1-\alpha)\cos\left(\frac{\pi}{2}\alpha\right)+\frac{\alpha\beta}{2}\frac{|t|^2}{2-\alpha}...
\end{aligned}$$

Once we know $\mathrm{E}\left(\exp(\sqrt{-1}z_i t)\right)$, by Taylor expansions we obtain

$$\begin{aligned}
&\log\left(\mathrm{E}\left(\exp\left(\sqrt{-1}t\beta^{-1/\alpha}\sum_{i=1}^D c_i z_i\right)\right)\right)\\
=&\log\prod_{i=1}^D \mathrm{E}\left(\exp\left(\sqrt{-1}t\beta^{-1/\alpha}c_i z_i\right)\right)\\
=&\sum_{i=1}^D \log\left(1-|c_i|^\alpha|t|^\alpha\Gamma(1-\alpha)\cos\left(\frac{\pi}{2}\alpha\right)+\frac{\alpha}{2}\frac{|c_i|^2|t|^2}{\beta^{2/\alpha-1}(2-\alpha)}...\right)\\
=&\sum_{i=1}^D -|c_i|^\alpha|t|^\alpha\Gamma(1-\alpha)\cos\left(\frac{\pi}{2}\alpha\right)+\frac{\alpha}{2}\frac{|c_i|^2|t|^2}{\beta^{2/\alpha-1}(2-\alpha)}-\\
&\qquad\frac{1}{2}|c_i|^{2\alpha}|t|^{2\alpha}\left(\Gamma(1-\alpha)\cos\left(\frac{\pi}{2}\alpha\right)\right)^2+....
\end{aligned}$$

If $\max_{1\le i\le D}(|c_i|)\to 0$, then

$$\begin{aligned}
&\log\left(\mathrm{E}\left(\exp\left(\sqrt{-1}t\beta^{-1/\alpha}\sum_{i=1}^D c_i z_i\right)\right)\right)\\
=&-|t|^\alpha\Gamma(1-\alpha)\cos\left(\frac{\pi}{2}\alpha\right)\sum_{i=1}^D|c_i|^\alpha+...\\
\to&-|t|^\alpha\Gamma(1-\alpha)\cos\left(\frac{\pi}{2}\alpha\right).
\end{aligned}$$

Chapter 7

Conditional Random Sampling

The method of *stable random projections* is attractive in part because one does not need to know anything about the data and still has the worst-case performance guarantees. This also implies that other methods can sometimes outperform *stable random projections* if additional information about the data is available, for example, when the data are highly sparse.

Another disadvantage of *stable random projections* is that one has to choose a fixed α ($0 < \alpha \leq 2$) in advance. For example, we have to maintain two sets of projected data if we would like to estimate both the l_2 and the l_1 distances; and, if later we also hope to study the $l_{0.5}$ or l_0 distances, we have to start everything over.

In this chapter, we introduce *Conditional Random Sampling (CRS)*, which is specifically designed for sampling massive sparse data. CRS is simpler than *stable random projections* and often outperforms *stable random projections* when the data are reasonably sparse. In some cases, for example, boolean (0/1) data, it is almost guaranteed that CRS will outperform *stable random projections*. Another advantage of CRS is that it can approximate multi-way distances while *stable random projections* are limited to the pair-wise situation.

This chapter will present the main methodology of CRS and compare CRS with *stable random projections*. Chapter 8 will focus on the boolean (0/1) data. Boolean data are common in practice and are convenient for rigorous theoretical analysis; thus we allocate a separate chapter for this case.

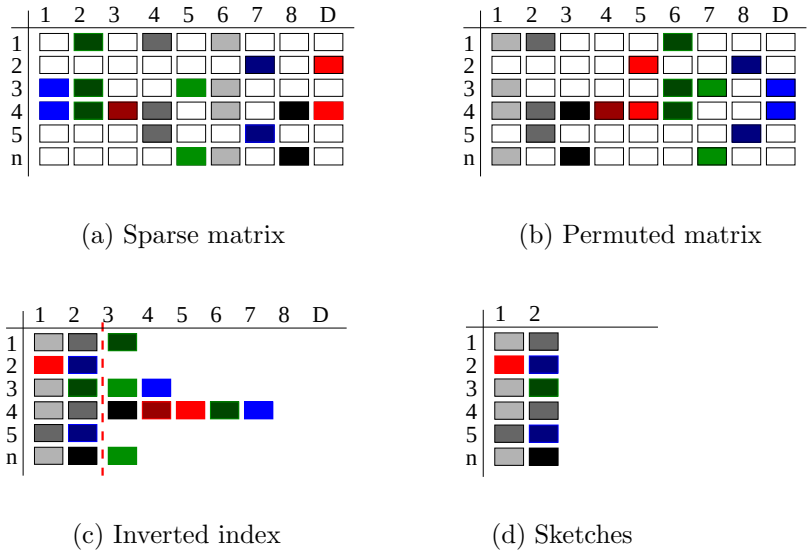
Part of the results presented in this chapter and Chapter 8 appeared in two conference proceedings[119, 123] and in two technical reports[120, 122], and will appear in a journal paper[121].

7.1 The Procedures of CRS

Conditional Random Sampling (CRS) is applicable for approximating many kinds of summary statistics, including multi-way associations, pair-wise l_α distances for any $\alpha > 0$, particularly suitable when the datasets are sparse. This technique can be applied in both static data and dynamic data (e.g., data streams).

CRS is a two-stage procedure, including a *sketching* stage and an *estimation* stage. In the *sketching* stage, we scan the data matrix once and store a fraction of the non-zero elements in each data point, as "sketches." In the *estimation* stage, we generate *conditional random samples* online pair-wise (for two-way) or group-wise (for multi-way); hence we name our algorithm *Conditional Random Sampling (CRS)*.

7.1.1 The Sampling/Sketching Procedure



(a) Sparse matrix (b) Permuted matrix

(c) Inverted index (d) Sketches

Figure 7.1: The *sketching* procedure of *Conditional Random Sampling (CRS)*.

Figure 7.1 provides a global view of the *sketching* stage. The columns of a sparse data matrix (a) are randomly permuted (b). The inverted index (c) only considers the non-zero entries. *Sketches* are simply the front of inverted index (d). In the actual implementation, of course, we only need to maintain a permutation mapping on the column IDs for those non-zero entries.

More specifically, suppose we work with a data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$, our sketching procedure works as follows:

- Generate a sample of random permutation mapping (e.g., [105, Algorithm 3.4.2.P]), $\pi : \Omega \to \Omega$, where $\Omega = \{1, 2, 3, ..., D\}$.
- For each data point (row) $u_i \in \mathbb{R}^D$ in $\mathbf{A}$, we apply the mapping π on the column IDs of the non-zero entries and store the k_i entries with the smallest permuted IDs. That is, if the permuted ID of a non-zero entry is $\leq k_i$, we store this entry (including its ID and value). The stored non-zero entries form a sketch K_i.

Apparently, it is reasonable to assume that f_i's are known, because the sketching procedure scans the data matrix once. At least as an option, the marginal information can be taken advantage of for sharper estimates.

CRS in Dynamic Data

The above sketching procedure also works with the *dynamic data*, e.g., data streams.

A data stream, say u_1, contains pairs $(i,\ u_{1,i})$, $i \in \Omega = \{1, 2, ..., D\}$, often arriving in random order. When a pair $(i,\ u_{1,i})$ arrives, if $\pi(i) \leq k_1$ (k_1 is the sketch size of u_1), we update the "old" $u_{1,i}$ in the sketch. If $\pi(i) > k_1$, we simply do nothing.

From Sketches to Random Samples

Sketches are not conventional random samples, which may make the estimation task difficult. We show, in Figure 7.2, that sketches are almost random samples pair-wise (or group-wise). Figure 7.2(a) constructs *conventional random samples* from a data matrix; and we show one can generate (retrospectively) the same random samples from sketches in Figure 7.2(b)(c).

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

(a) Data matrix and random samples

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

(b) Postings

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

(c) Sketches

Figure 7.2: (a): A data matrix with two rows and $D = 15$. If the column IDs are random, the first $D_s = 10$ columns constitute a random sample. u_i denotes the ith row. (b): Inverted index consists of tuples "ID (Value)." (c): Sketches are the first k_i entries of the inverted index sorted ascending by IDs. In this example, $k_1 = 5$, $k_2 = 6$, $D_s = \min(10, 11) = 10$. Excluding 11(3) in K_2, we obtain the same samples as if we directly sampled the first $D_s = 10$ columns in the data matrix.

In Figure 7.2(a), after the columns have been randomly permuted, we can construct random samples by simply taking the first D_s columns from the data matrix (of D columns).

For sparse data, we only store the non-zero elements in the form of tuples "ID (Value)," a structure called *inverted index*, denoted by P_i for each row u_i. Figure 7.2(b) shows the inverted index for the same data matrix in Figure 7.2(a). The tuples are sorted ascending by their IDs. A *sketch*, K_i, of inverted index P_i, is the first k_i entries (i.e., the smallest k_i IDs) of P_i, as shown in Figure 7.2(c).

The central observation is that if we exclude all elements in sketches whose IDs are larger than

$$D_s = \min\left(\max(\mathrm{ID}(K_1)), \max(\mathrm{ID}(K_2))\right), \tag{7.1}$$

we obtain exactly the same samples as if we directly sampled the first D_s columns from the data matrix in Figure 7.2(a). This way, we convert sketches into random samples by *conditioning* on D_s, which differs pair-wise and we do not know beforehand.

In general, suppose some summary statistic involves m data points, say u_1, u_2, ...,

u_m, then the (conditional) sample size D_s would be

$$D_s = \min\left(\max(\mathrm{ID}(\mathrm{K}_1)), \max(\mathrm{ID}(\mathrm{K}_2)), ..., \max(\mathrm{ID}(\mathrm{K}_m))\right). \tag{7.2}$$

7.1.2 The Estimation Procedure

The estimation task for CRS can be extremely simple. After we construct the conditional random samples from sketches K_1 and K_2 with the effective sample size D_s, we can compute any distances (l_2, l_1, or inner products) from the samples and multiply them by $\frac{D}{D_s}$ to estimate the original space. (Later, we will show how to improve the estimates by taking advantage of the marginal information.)

We use $\tilde{u}_{1,j}$ and $\tilde{u}_{2,j}$ ($j = 1$ to D_s) to denote the conditional random samples (of size D_s) obtained by CRS. For example, in Figure 7.2, we have $D_s = 10$, and the non-zero $\tilde{u}_{1,j}$ and $\tilde{u}_{2,j}$ are

$$\tilde{u}_{1,2} = 3,\ \tilde{u}_{1,4} = 2,\ \tilde{u}_{1,6} = 1,\ \tilde{u}_{1,9} = 1,\ \tilde{u}_{1,10} = 2$$
$$\tilde{u}_{2,1} = 1,\ \tilde{u}_{2,2} = 3,\ \tilde{u}_{2,5} = 1,\ \tilde{u}_{2,6} = 2,\ \tilde{u}_{2,8} = 1.$$

Denote the inner product and l_α distance, by a and $d_{(\alpha)}$ (for any α, not limited to $0 < \alpha \leq 2$), respectively,

$$a = \sum_{i=1}^{D} u_{1,i} u_{2,i}, \quad d_{(\alpha)} = \sum_{i=1}^{D} |u_{1,i} - u_{2,i}|^\alpha, \tag{7.3}$$

Once we have the random samples, we can then use simple linear estimators:

$$\hat{a}_{MF} = \frac{D}{D_s} \sum_{j=1}^{D_s} \tilde{u}_{1,j} \tilde{u}_{2,j}, \quad \hat{d}_{(\alpha),MF} = \frac{D}{D_s} \sum_{j=1}^{D_s} |\tilde{u}_{1,j} - \tilde{u}_{2,j}|^\alpha, \tag{7.4}$$

7.2 Theoretical Properties of D_s

The *conditional sampling size* D_s is random. It is important to study its theoretical properties (e.g., expectation) because we hope, at least on average, D_s should be as

large as possible for more coverage of the data. We have an important approximation:

$$\mathrm{E}\left(\frac{D_s}{D}\right) = \min\left(\frac{k_1}{f_1+1}, \frac{k_2}{f_2+1}, ..., \frac{k_m}{f_m+1}\right) + O\left(\sum_{i=1}^{m} \frac{k_i}{f_i+1}\sqrt{\frac{1}{k_i}\frac{f_i-k_i}{f_i}}\right) + O\left(\frac{m}{D}\right), \tag{7.5}$$

which can be inferred from the following lemma.

Lemma 29 *Denote $Z_i = \max(ID(K_i))$, $D_s = \min(Z_1, Z_2, ..., Z_m)$, the numbers of non-zeros f_i ($i = 1$ to m), and the sketch sizes k_i ($i = 1$ to m). Then*

$$E(Z_i) = \frac{k_i(D+1)}{f_i+1}, \tag{7.6}$$

$$Var(Z_i) = \frac{(D+1)(D-f_i)k_i(f_i+1-k_i)}{(f_i+1)^2(f_i+2)} \leq \frac{1}{k_i}\frac{f_i-k_i}{f_i}\left(E(Z_i)\right)^2. \tag{7.7}$$

$$0 \leq \min\left(E(Z_1), E(Z_2), ..., E(Z_m)\right) - E\left(D_s\right) \leq \sum_{i=1}^{m}\sqrt{\frac{1}{k_i}\frac{f_i-k_i}{f_i}}E(Z_i), \tag{7.8}$$

Proof: *The column IDs are uniform in $\Omega = \{1, 2, ..., D\}$ at random. $Z_i = \max(ID(K_i))$ is the (k_i)th order statistics of the set $ID(K_i) \in \Omega = \{1, 2, ..., D\}$, with the probability mass function (PMF) and moments[57, Exercise 2.1.4]*[1]

$$P(Z_i = t) = \frac{\binom{t-1}{k_i-1}\binom{D-t}{f_i-k_i}}{\binom{D}{f_i}}, \quad E(Z_i) = \frac{k_i(D+1)}{f_i+1},$$
$$Var(Z_i) = \frac{(D+1)(D-f_i)k_i(f_i+1-k_i)}{(f_i+1)^2(f_i+2)} \leq \frac{1}{k_i}\frac{f_i-k_i}{f_i}\left(E(Z_i)\right)^2.$$

By Jensen's inequality, we know that

$$E(D_s) = E\left(\min\left(Z_1, Z_2, ..., Z_m\right)\right) \leq \min\left(E(Z_1), E(Z_2), ..., E(Z_m)\right)$$

[1] Also see `www.ds.unifi.it/VL/VL_EN/urn/index.html`

Assuming $E(Z_1)$ is the smallest among all $E(Z_i)$ yields

$$\begin{aligned}
&\min\left(E(Z_1), E(Z_2), ..., E(Z_m)\right) - E\left(D_s\right) \\
=&E\left(\max\left(E(Z_1) - Z_1, E(Z_1) - Z_2, ..., E(Z_1) - Z_m\right)\right) \\
\leq&E\left(\max\left(E(Z_1) - Z_1, E(Z_2) - Z_2, ..., E(Z_m) - Z_m\right)\right) \\
\leq&\sum_{i=1}^{m} E\left(E(Z_i) - Z_i\right) \leq \sum_{i=1}^{m} \sqrt{Var\left(Z_i\right)} \leq \sum_{i=1}^{m} \sqrt{\frac{1}{k_i}\frac{f_i - k_i}{f_i}} E(Z_i).
\end{aligned}$$

This completes the proof. (Note that we actually only need to sum $m-1$ terms because one term will cancel.)

From (7.5), we can learn some important properties of our algorithm:

- The effective (conditional) sample size D_s is directly related to the data sparsity $\frac{f_i}{D}$. Suppose all $f_i = f$ and $k_i = k$, then $\mathrm{E}(D_s) \approx k\frac{D}{f}$. In other words, if the data only have about 1% non-zeros, then the effective sample size is roughly $100k$. This is why our algorithm can dramatically reduce the estimation variances compared to conventional *random coordinate sampling*.

- The performance of our algorithm will degrade with increasing m. In real applications, $m = 2$ (i.e., pair-wise) is probably the most important; and we do not expect many applications will be interested in $m > 4$. It is unreliable to estimate very high-order interactions (associations) because we will soon run out of data as m increases.

Later, we will also need to evaluate $\mathrm{E}\left(\frac{D}{D_s}\right)$, which can be well approximated by $\frac{1}{\mathrm{E}\left(\frac{D}{D_s}\right)} + O\left(\sum_{i=1}^{m} \frac{1}{k_i}\frac{f_i - k_i}{f_i}\right)$ by the *delta method*. By Jensen's inequality, we know

$$\mathrm{E}\left(\frac{D}{D_s}\right) \geq \frac{1}{\mathrm{E}\left(\frac{D_s}{D}\right)} \geq \max\left(\frac{f_1 + 1}{k_1}, \frac{f_2 + 1}{k_2}, ..., \frac{f_m + 1}{k_m}\right), \tag{7.9}$$

and the inequality becomes equality asymptotically.

For convenience, we recommend the approximation

$$\mathrm{E}\left(\frac{D}{D_s}\right) \approx \max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}, ..., \frac{f_m}{k_m}\right). \tag{7.10}$$

We have conducted extensive simulations and learned that the errors caused by using (7.10) are usually within 5% when $k_i \geq 20 \sim 50$ and $m \leq 4$. Real applications will probably use $k_i \geq 100$ or even more. Therefore, it is reasonable to consider (7.10) is exact, even though it is only asymptotically exact. Figure 7.3 presents some simulations for the pair-wise ($m = 2$) case.

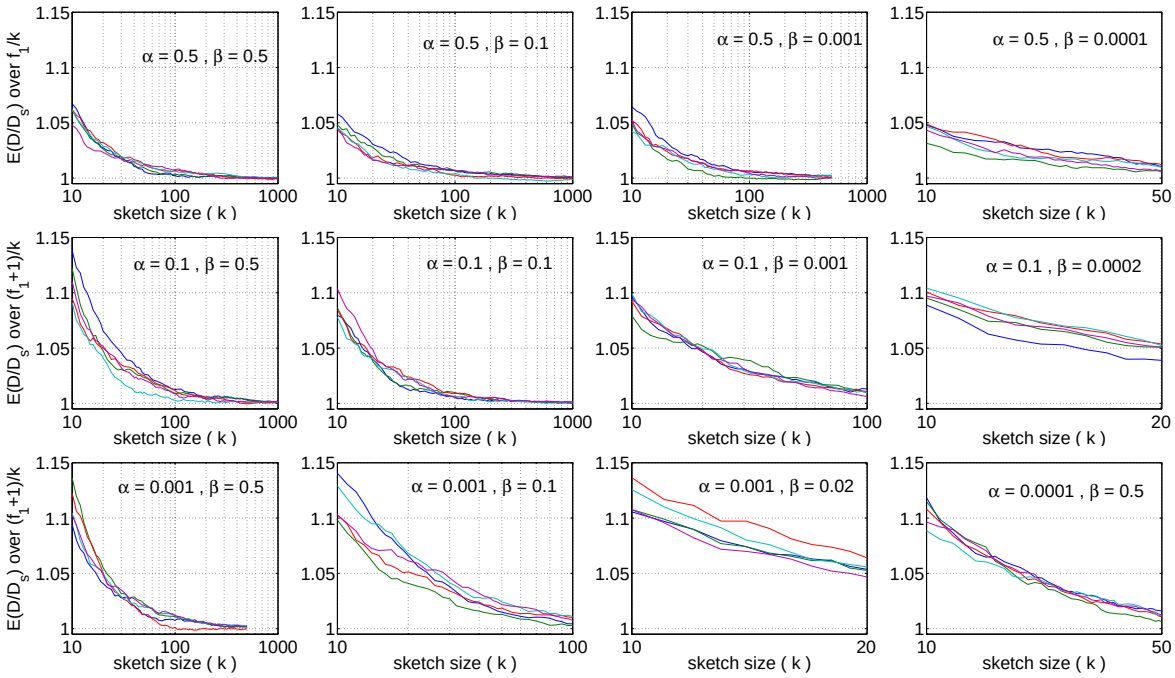


Figure 7.3: The ratios $\mathrm{E}\left(\frac{D}{D_s}\right) / \frac{\max(f_1+1, f_2+1)}{k}$ show that the errors from using (7.10) are usually within 5% when $k \geq 20$. $D = 10^6$, $f_1 = \alpha D$, and $f_2 = \beta f_1$. In each panel, the five curves correspond to different values of $\#\{j : u_{1,j} > 0 \ \& \ u_{2,j} > 0, 1 \leq j \leq D\}$, which measure the correlations and are set to be γf_2, with $\gamma = 0.01, 0.1, 0.5, 0.8, 1.0$, respectively. It is clear that γ does not affect D_s strongly. In each panel, we only simulated 10^3 permutations for each k; and hence the curves are not very smooth.

It is possible to improve (7.10). Figure 7.3 implies that $\mathrm{E}(D_s)$ is not affected strongly by the correlations (i.e., the γ values in Figure 7.3). Therefore, we can

assume that ID(K_1) and ID(K_2) are independent for the purpose of estimating E(D_s). With this assumption, we can in principle express E(D_s) in some summation form, which can be evaluated numerically. However, we still prefer the approximation (7.10) because it is intuitive, simple, and accurate enough. Note that, we only need E $\left(\frac{D}{D_s}\right)$ for theoretically evaluating the estimation variances; it is not needed in our sketching procedure nor in the estimation stage.

7.2.1 Sample-Without-Replacement

Our scheme is "sample-without-replacement." When a data vector u_i is extremely sparse (e.g., $f_i < 100$), it is probably affordable to take the whole inverted index as a sketch. In general, we expect that $D_s \ll D$ holds in most cases. For simplicity, we will assume "sample-with-replacement" for the analysis in this chapter. In Chapter 8, we will consider *sample-without-replacement* for the boolean data case. Note that when we do not use the marginal information, the estimators will be the same whether or not we consider "sample-without-replacement."

7.3 Theoretical Variance Analysis of CRS

We first consider $\hat{a}_{MF} = \frac{D}{D_s}\sum_{j=1}^{D_s}\tilde{u}_{1,j}\tilde{u}_{2,j}$. By assuming "sample-with-replacement," the samples, $(\tilde{u}_{1,j}\tilde{u}_{2,j})$, $j = 1$ to D_s, are i.i.d, conditional on D_s. Thus,

$$\mathrm{Var}(\hat{a}_{MF}|D_s) = \left(\frac{D}{D_s}\right)^2 D_s \mathrm{Var}\left(\tilde{u}_{1,1}\tilde{u}_{2,1}\right) = \frac{D}{D_s} D\left(\mathrm{E}\left(\tilde{u}_{1,1}\tilde{u}_{2,1}\right)^2 - \mathrm{E}^2\left(\tilde{u}_{1,1}\tilde{u}_{2,1}\right)\right), \tag{7.11}$$

$$\mathrm{E}\left(\tilde{u}_{1,1}\tilde{u}_{2,1}\right) = \frac{1}{D}\sum_{i=1}^{D}\left(u_{1,i}u_{2,i}\right) = \frac{a}{D}, \qquad \mathrm{E}\left(\tilde{u}_{1,1}\tilde{u}_{2,1}\right)^2 = \frac{1}{D}\sum_{i=1}^{D}\left(u_{1,i}u_{2,i}\right)^2, \tag{7.12}$$

$$\mathrm{Var}(\hat{a}_{MF}|D_s) = \frac{D}{D_s} D\left(\frac{1}{D}\sum_{i=1}^{D}\left(u_{1,i}u_{2,i}\right)^2 - \left(\frac{a}{D}\right)^2\right) = \frac{D}{D_s}\left(\sum_{i=1}^{D} u_{1,i}^2 u_{2,i}^2 - \frac{a^2}{D}\right). \tag{7.13}$$

The unconditional variance would be simply

$$\mathrm{Var}(\hat{a}_{MF}) = \mathrm{E}\left(\mathrm{Var}(\hat{a}_{MF}|D_s)\right) = \mathrm{E}\left(\frac{D}{D_s}\right)\left(\sum_{i=1}^{D} u_{1,i}^2 u_{2,i}^2 - \frac{a^2}{D}\right).$$

We can similarly derive the variances for $\hat{d}_{(\alpha),MF}$. In a summary, we obtain (when $k_1 = k_2 = k$)

$$\mathrm{Var}\left(\hat{a}_{MF}\right) = \mathrm{E}\left(\frac{D}{D_s}\right)\left(\sum_{i=1}^{D} u_{1,i}^2 u_{2,i}^2 - \frac{a^2}{D}\right) \approx \frac{\max(f_1, f_2)}{D}\frac{1}{k}\left(D\sum_{i=1}^{D} u_{1,i}^2 u_{2,i}^2 - a^2\right), \tag{7.14}$$

$$\mathrm{Var}\left(\hat{d}_{(\alpha),MF}\right) = \mathrm{E}\left(\frac{D}{D_s}\right)\left(d_{(2\alpha)} - \frac{[d_{(\alpha)}]^2}{D}\right) \approx \frac{\max(f_1, f_2)}{D}\frac{1}{k}\left(Dd_{(2\alpha)} - [d_{(\alpha)}]^2\right), \tag{7.15}$$

using the approximation (7.10).

The *sparsity* term $\frac{\max(f_1,f_2)}{D}$ reduces the variances significantly. If $\frac{\max(f_1,f_2)}{D} = 0.01$, the variances can be reduced by a factor of 100, compared to *conventional random sampling*. On the other hand, the variances of CRS estimators are still severely affected by higher moments. Therefore, CRS would not provide worst case performance guarantees.

7.4 Improving CRS Using Marginal Information

As mentioned previously, it is reasonable to assume that we know the marginal information such as marginal norms, numbers of non-zero elements, or even marginal histograms. We utilize the marginal information by maximizing the joint likelihood under marginal constraints.

In some cases, we know the explicit likelihood. For example, when the data are integer valued (which is also the case in histograms), the joint likelihood is equivalent to that of a two-way or multi-way contingency table. In the 0/1 case, we can even express the MLE solution explicitly and derive the closed-form (asymptotic) variance.

In general real-valued data, the joint likelihood is not available. One option is to use non-parametric maximum likelihood such as the "Empirical Likelihood" [143]. In this study, we make some assumptions on the data and propose an approximate MLE, which is conceptually reasonable and works well in real data.

The use of margins for estimating contingency tables was suggested in the 1940s [59, 157] for a census application. They developed a straightforward iterative estimation method called *iterative proportional scaling*, which was an approximation to the maximum likelihood estimator.

We first describe a procedure for estimating two-way histograms (i.e., integer-valued data), which amounts to estimating contingency tables under margin constraints. Once we have estimated the contingency table, we can compute the inner product (and other summary statistics) easily.

7.4.1 Integer-valued Data (Histograms)

Histograms are useful for answering queries like $\mathbf{Pr}(1 < u_1 < 2 \,\&\, u_2 > 2)$. Histograms contain more information than the mere inner product $a = u_1^{\mathrm{T}} u_2$. While univariate histograms are easy to compute and store, joint histograms are much more difficult especially for high-order joins. We will focus only on two-way histograms, as the notation gets messy in the multi-way case.

Without loss of generality, we number each histogram bin $\{0, 1, 2, ...\}$ as shown in Figure 7.4(a). Apparently, we can also think these are the original data, which happen to be integer-valued. Histograms can be conveniently represented by contingency tables, e.g., Figure 7.4(b)

We denote the original contingency table by $\mathbf{X} = \{x_{i,j}\}_{i=0}^{I}{}_{j=0}^{J}$. Similarly, we denote the sample contingency table by $\mathbf{S} = \{s_{i,j}\}_{i=0}^{I}{}_{j=0}^{J}$. An example of sample contingency table is shown in Figure 7.4(c) by taking the first $D_s = 10$ columns from the binned data matrix. Of course, we generate the equivalent sample table online by *Conditional Random Sampling*.

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

(a) Data matrix

u_1 \ u_2	0	1	2	3	
0	3	3	1	0	7
1	2	0	1	2	5
2	2	1	0	0	3
	7	4	2	2	

(b) Original table

u_1 \ u_2	0	1	2	3	
0	2	3	0	0	5
1	1	0	1	1	3
2	2	0	0	0	2
	5	3	1	1	

(c) Sample table

Figure 7.4: (a): A data matrix with binned (integer) data, $D = 15$. The entries of $u_1 \in \{0, 1, 2\}$ and $u_2 \in \{0, 1, 2, 3\}$. We construct a 3×4 contingency table for u_1 and u_2 in (b). For example, in three columns ($j = 3$, $j = 7$, and $j = 12$), we have $u_{1,j} = u_{2,j} = 0$, hence the (0,0) entry in the table is 3. Suppose the column IDs of the data matrix are random, we can construct a random sample by taking the first $D_s = 10$ columns, represented by a sample contingency table in (c).

Estimators

Conditioning on D_s, we can estimate the original table in a straightforward fashion:

$$\hat{x}_{i,j,MF} = \frac{D}{D_s} s_{i,j}, \tag{7.16}$$

$$\mathrm{Var}\left(\hat{x}_{i,j,MF}\right) = \mathrm{E}\left(\frac{D}{D_s}\right) \frac{1}{\frac{1}{x_{i,j}} + \frac{1}{D - x_{i,j}}}. \tag{7.17}$$

Next, we would like to take advantage of marginal histograms, i.e., the row and column sums of the contingency table. There are $I + 1$ row sums and $J + 1$ column sums. The total number of degrees of freedom is $(I+1)\times(J+1)-(I+1)-(J+1)+1 = I \times J$.[2] Denote the row sums by $\{x_{i+}\}_{i=0}^{I}$ and the column sums by $\{x_{+j}\}_{j=0}^{J}$.

When margins are known, we can expect more accurate estimates, especially when the number of degrees of freedom IJ is not too large. The sample contingency table $\mathbf{S} = \{s_{i,j}\}_{i=0}^{I}{}_{j=0}^{J}$ follows a multinomial. Therefore, the maximum likelihood estimator

[2]The sum of the row sums has to be equal to the sum of the column sums, which is equal to D (sum of all cells). Therefore, the effective number of constraints is $I + J + 1$, instead of $I + J + 2$.

(MLE) of $x_{i,j}$ under marginal constraints amounts to a convex program:

$$\begin{aligned} &\text{Minimize } -\sum_{i=0}^{I}\sum_{j=0}^{J} s_{i,j}\log x_{i,j} \\ &\text{such that } \sum_{j=0}^{J} x_{i,j} = x_{i+},\ \sum_{i=0}^{I} x_{i,j} = x_{+j},\ x_{i,j} \geq s_{i,j}, \end{aligned} \tag{7.18}$$

which can be solved easily using any standard convex optimization algorithm such as Newton's method.[3] We denote the estimated table cells as $\hat{x}_{i,j,MLE}$.

We can estimate the inner product $a = u_1^{\mathrm{T}} u_2$ from the estimated contingency table

$$a = u_1^{\mathrm{T}} u_2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(ij)x_{i,j}. \tag{7.19}$$

Therefore, we can estimate a by one of the following:

$$\hat{a}_{MF,c} = \sum_{i=1}^{I}\sum_{j=1}^{J}(ij)\hat{x}_{i,j,MF}, \tag{7.20}$$

$$\hat{a}_{MLE,c} = \sum_{i=1}^{I}\sum_{j=1}^{J}(ij)\hat{x}_{i,j,MLE}, \tag{7.21}$$

where the subscript "c" indicates that we compute a from contingency tables. We can also compute many other summary statistics such as the chi-square statistic from the estimated contingency table.

Section 7.8 presents the derivation for the variances of $\hat{x}_{i,j,MLE}$ and $\hat{a}_{MLE,c}$.

A Numerical Example

Two words "THIS" and "HAVE" are taken from a chunk of MSN Web crawl data ($D = 2^{16}$). The data are quantized into a few histogram bins. Two experiments are conducted, with 5 bins and 3 bins, respectively, as shown in Table 7.1.

[3] In the implementation, the linear dependency in the constraints can be removed by discarding the two constraints on x_{+0} and x_{0+}, and adding a new constraint: $\sum_{i=0}^{I}\sum_{j=0}^{J} x_{i,j} = D$.

Table 7.1: The two word vectors "THIS" and "HAVE" are quantized. (a) Exp #1: 5 bins numbered from 0 to 4. (b) Exp #2: 3 bins.

Bin ID	Data
0	0
1	1
2	$2 \sim 3$
3	$4 \sim 9$
4	≥ 10

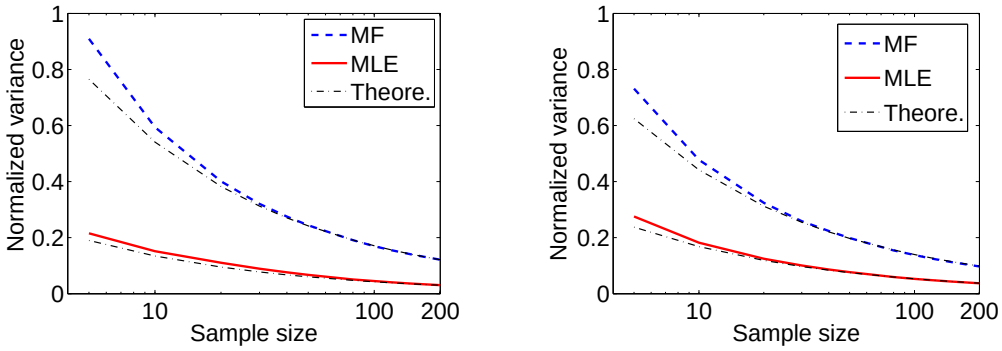
(a) Exp.#1

Bin ID	Data
0	0
1	1
2	≥ 2

(b) Exp.#2

The two (quantized) vectors are sampled by *Conditional Random Sampling* with sketch sizes ranging from 5 to 200. Sample contingency tables are then constructed (online) from sketches and the original contingency tables are estimated using both margin-free (MF) and MLE estimators. Then we estimate the inner product a from the estimated contingency table.

Figure 7.5 compares the empirical variances with the theoretical predicts for $\hat{a}_{MF,c}$ and $\hat{a}_{MLE,c}$, verifying that our theoretical variances are accurate after sample sizes $\geq 10 \sim 20$. The errors are mostly due to the approximation $\mathrm{E}\left(\frac{D}{D_s}\right) = \max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right)$ (we let $k_1 = k_2 = k$). In this case, the marginal histograms help considerably.



(a) Exp. #1 (b) Exp. #2

Figure 7.5: The inner product a (after quantization) between "THIS" and "HAVE" is estimated by both $\hat{a}_{MF,c}$ and $\hat{a}_{MLE,c}$. Results are reported in $\frac{\sqrt{\mathrm{Var}(\hat{a})}}{a}$. The two thin dashed lines both labeled "theore." are theoretical variances, which match the empirical values well especially when sketch sizes $\geq 10 \sim 20$. In this case, marginal histograms help considerably.

The Important Special case: Boolean (0/1) Data

Chapter 8 is devoted entirely to the boolean data case. In 0/1 data, estimating the inner product becomes estimating a two-way contingency table, which has four cells. Because of the margin constraints, there is only one degree of freedom. Therefore, the MLE of a is the solution, denoted by $\hat{a}_{0/1,MLE}$, to a cubic equation

$$\frac{s_{11}}{a} - \frac{s_{10}}{f_1 - a} - \frac{s_{01}}{f_2 - a} + \frac{s_{00}}{D - f_1 - f_2 + a} = 0, \tag{7.22}$$

where $s_{11} = \#\{j : \tilde{u}_{1,j} = \tilde{u}_{2,j} = 1\}$, $s_{10} = \#\{j : \tilde{u}_{1,j} = 1, \tilde{u}_{2,j} = 0\}$, $s_{01} = \#\{j : \tilde{u}_{1,j} = 0, \tilde{u}_{2,j} = 1\}$, $s_{00} = \#\{j : \tilde{u}_{1,j} = 0, \tilde{u}_{2,j} = 0\}$, $j = 1, 2, ..., D_s$.

The (asymptotic) variance of $\hat{a}_{0/1,MLE}$ is proved in Chapter 8 to be

$$\mathrm{Var}(\hat{a}_{0/1,MLE}) = \mathrm{E}\left(\frac{D}{D_s}\right) \frac{1}{\frac{1}{a} + \frac{1}{f_1 - a} + \frac{1}{f_2 - a} + \frac{1}{D - f_1 - f_2 + a}}. \tag{7.23}$$

7.4.2 Real-valued Data

A practical solution is to assume some parametric form of the (bivariate) data distribution based on prior knowledge, and then solve an MLE considering various constraints. Suppose the samples $(\tilde{u}_{1,j},\ \tilde{u}_{2,j})$ are i.i.d. bivariate normal with moments determined by the population moments, i.e.,

$$\begin{bmatrix} \tilde{v}_{1,j} \\ \tilde{v}_{2,j} \end{bmatrix} = \begin{bmatrix} \tilde{u}_{1,j} - \bar{u}_1 \\ \tilde{u}_{2,j} - \bar{u}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tilde{\Sigma} \right), \tag{7.24}$$

$$\tilde{\Sigma} = \frac{1}{D_s}\frac{D_s}{D} \begin{bmatrix} \|u_1\|^2 - D\bar{u}_1^2 & u_1^{\mathrm{T}}u_2 - D\bar{u}_1\bar{u}_2 \\ u_1^{\mathrm{T}}u_2 - D\bar{u}_1\bar{u}_2 & \|u_2\|^2 - D\bar{u}_2^2 \end{bmatrix} = \frac{1}{D_s} \begin{bmatrix} \ddot{m}_1 & \ddot{a} \\ \ddot{a} & \ddot{m}_2 \end{bmatrix}, \tag{7.25}$$

where $\bar{u}_1 = \sum_{i=1}^{D} u_{1,i}/D$, $\bar{u}_2 = \sum_{i=1}^{D} u_{2,i}/D$. $\ddot{m}_1 = \frac{D_s}{D}\left(\|u_1\|^2 - D\bar{u}_1^2\right)$, $\ddot{m}_2 = \frac{D_s}{D}\left(\|u_2\|^2 - D\bar{u}2^2\right)$, $\ddot{a} = \frac{D_s}{D}\left(u_1^{\mathrm{T}}u_2 - D\bar{u}_1\bar{u}_2\right)$. Suppose that $\bar{u}_1$, $\bar{u}_2$, $m_1 = \|u_1\|^2$ and $m_2 = \|u_2\|^2$ are known, an MLE for $a = u_1^{\mathrm{T}}u_2$, denoted by $\hat{a}_{MLE,N}$, is

$$\hat{a}_{MLE,N} = \frac{D}{D_s}\hat{\ddot{a}} + D\bar{u}_1\bar{u}_2, \tag{7.26}$$

where, similar to Lemma 4 in Chapter 3, $\hat{\ddot{a}}$ is the solution to a cubic equation:

$$\ddot{a}^3 - \ddot{a}^2\left(\tilde{v}_1^{\mathrm{T}}\tilde{v}_2\right) + \ddot{a}\left(-\ddot{m}_1\ddot{m}_2 + \ddot{m}_1\|\tilde{v}_2\|^2 + \ddot{m}_2\|\tilde{v}_1\|^2\right) - \ddot{m}_1\ddot{m}_2\tilde{v}_1^{\mathrm{T}}\tilde{v}_2 = 0. \qquad (7.27)$$

$\hat{a}_{MLE,N}$ is fairly robust, although sometimes we observe the biases are quite noticeable. In general, this is a good bias-variance trade-off (especially when k is not too large). Intuitively, the reason why this (seemingly crude) assumption of bivariate normality works well is because, once we have fixed the margins, we have removed to a large extent the non-normal component of the data.

7.5 Theoretical Comparisons of CRS With Random Projections

As reflected by their variances, for general data types, whether CRS is better than random projections depends on two competing factors: data sparsity and data heavy-tailedness. In the following two important scenarios, however, CRS outperforms random projections.

7.5.1 Boolean (0/1) data

In this case, the marginal norms are the same as the numbers of non-zeros, i.e., $m_i = \|u_i\|^2 = f_i$. Figure 7.6 plots the ratio, $\frac{\mathrm{Var}(\hat{a}_{MF})}{\mathrm{Var}\left(\hat{a}_{NRP,MF}\right)}$, verifying that CRS is (considerably) more accurate:

$$\frac{\mathrm{Var}\left(\hat{a}_{MF}\right)}{\mathrm{Var}\left(\hat{a}_{NRP,MF}\right)} = \frac{\max(f_1, f_2)}{f_1 f_2 + a^2}\frac{1}{\frac{1}{a} + \frac{1}{D-a}} \leq \frac{\max(f_1, f_2)a}{f_1 f_2 + a^2} \leq 1,$$

where $\mathrm{Var}\left(\hat{a}_{NRP,MF}\right)$ is the variance of the inner product estimator in *normal random projections*, i.e., (3.6) in Chapter 3.

Figure 7.7 plots $\frac{\mathrm{Var}\left(\hat{a}_{0/1,MLE}\right)}{\mathrm{Var}\left(\hat{a}_{NRP,MLE}\right)}$, where $\mathrm{Var}\left(\hat{a}_{NRP,MLE}\right)$ is the asymptotic variance of the MLE in *normal random projections* using margin constraints, i.e., (3.15) in Chapter 3. In most possible range of the data, this ratio is less than 1. When u_1

and u_2 are very close (e.g., $a \approx f_2 \approx f_1$), random projections appear more accurate. However, when this does occur, the absolute variances are so small (even zero) that their ratio does not matter.

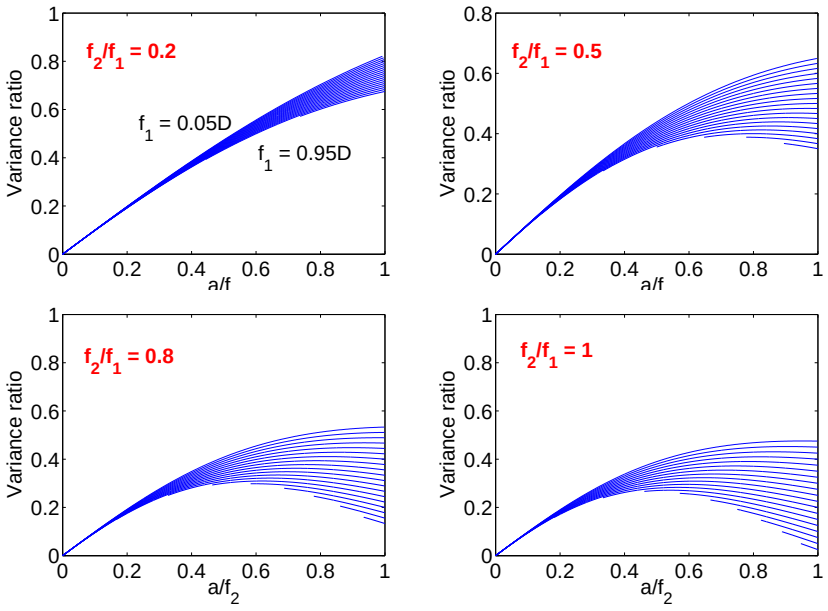


Figure 7.6: The variance ratios, $\frac{\mathrm{Var}(\hat{a}_{MF})}{\mathrm{Var}\left(\hat{a}_{NRP,MF}\right)}$, show that CRS has smaller variances than random projections, when no marginal information is used. We let $f_1 \geq f_2$ and $f_2 = \alpha f_1$ with $\alpha = 0.2, 0.5, 0.8, 1.0$. For each α, we plot from $f_1 = 0.05D$ to $f_1 = 0.95D$ spaced at $0.05D$.

7.5.2 Nearly Independent Data

Suppose two data points u_1 and u_2 are independent (or less strictly, uncorrelated to the second order), it is easy to show that the variance of CRS is always smaller:

$$\mathrm{Var}\left(\hat{a}_{MF}\right) \leq \frac{\max(f_1, f_2)}{D}\frac{m_1 m_2}{k} \leq \mathrm{Var}\left(\hat{a}_{NRP,MF}\right) = \frac{m_1 m_2 + a^2}{k}, \tag{7.28}$$

even if we ignore the data sparsity. Therefore, CRS will be much better for estimating inner products in nearly independent data. Note that, in high dimensions, it is often the case that most of the data points are only very weakly correlated.

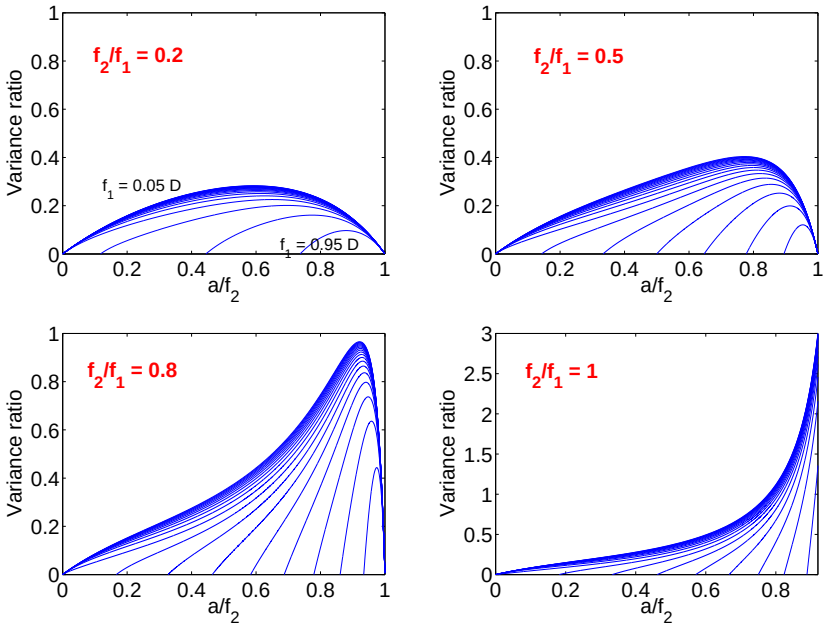


Figure 7.7: The ratios, $\frac{\mathrm{Var}\left(\hat{a}_{0/1,MLE}\right)}{\mathrm{Var}\left(\hat{a}_{NRP,MLE}\right)}$, show that CRS usually has smaller variances than random projections, except when $f_1 \approx f_2 \approx a$.

7.5.3 Comparing the Computational Efficiency

As previously mentioned, the cost of constructing sketches for $\mathbf{A} \in \mathbb{R}^{n \times D}$ would be $O(nD)$ (or more precisely, $O(\sum_{i=1}^{n} f_i)$). The cost of (normal) random projections would be $O(nDk)$. Therefore, it is possible that CRS is considerably more efficient than random projections in the sampling stage.

In the estimation stage, CRS costs $O(2k)$ to compute the sample distance for each pair. This cost is only $O(k)$ in random projections. Since k is very small, the difference should not be a concern.

7.6 Empirical Evaluations

We evaluate *Conditional Random Sampling* and compare it with random projections using various datasets listed in Table 7.2.

The NSF dataset[4] [60] contains 13297 documents (in 5298 dimensions). We randomly sampled 100 documents (i.e., 4950 pairs).

The NEWSGROUP dataset [26] were used for experimenting with random projections. It contains 2262 documents (in 5000 dimensions). Again, we randomly sampled 100 documents.

The COREL dataset has been used in quite a few SVM image classification papers (e.g., [42]). We selected one class (out of 14), which contains 80 images in 4096 dimensions.

The DEXTER dataset is taken from NIPS 2003 workshop on feature extraction.[5] We take the first 100 data points (out of 300) from the Dexter training dataset.

For the above four datasets, we use the original values without any preprocessing.

Finally, we randomly selected 100 words, each in 65536 (document) dimensions, provided by MSN. The original data are extremely heavy-tailed and highly skewed as can be seen from Table 7.2. We process data by both *square root* weighting (i.e., replacing any entries with its square root), and *logarithmic* weighting (i.e., replacing any non-zero entry with 1 + log(original value)). We include results for the original data as well as the weighted data.

	n	D	Sparsity	Kurtosis	Skewness
NSF	100	5298	1.09%	349.8	16.3
NEWSGROUP	100	5000	1.01%	352.9	16.5
COREL	80	4096	4.82%	765.9	24.7
DEXTER	100	20000	0.45%	1729.0	35.8
MSN (original)	100	65536	3.65%	4161.5	49.6
MSN (square root)	100	65536	3.65%	175.3	10.7
MSN (logarithmic)	100	65536	3.65%	111.8	9.5

Table 7.2: For each dataset, we compute the overall sparsity, median kurtosis, and median skewness. Our data are highly heavy-tailed and severely skewed. Recall normal has zero kurtosis and zero skewness.

We estimate the inner product, l_1 distance, and l_2 distance for every pair, using *Conditional Random Sampling (CRS)* as well as random projections. For l_2 distances, we use both the margin-free estimator and the MLE using margins, in *normal random*

[4] `ftp://ftp/cs.utexas.edu/pub/inderjit/Data/Text/Matrices/nsf_tfn.dat`

[5] `http://www.nipsfsc.ecs.soton.ac.uk/datasets/`

projections. For l_1 distances, we use the MLE since it is the most accurate.

For each pair, we conduct 50 runs and average the absolute errors. Since in each dataset, we have $\frac{n(n-1)}{2}$ pairs, we compare the median errors and the percentage of pairs in which CRS outperforms random projections.

The results are presented in Figures 7.8, 7.9, 7.10, 7.11, 7.12, 7.13, and 7.14. In each figure, the upper four panels are the ratio (CRS over random projections) of the median (among $n(n-1)/2$ pairs) average (over 50 runs) absolute errors. The bottom four panels are the percentage of pairs for which our algorithm has smaller absolute errors than random projections.

In each panel, the dashed curve indicates that we sample each data point with equal sample size (k). For CRS, we can adjust the sample size according to the sparsity, reflected by the solid curves. We adjust sample sizes only crudely (not optimizing anything). The data points are divided into 3 (5 for the MSN data) groups according to sparsity. Data in different groups are assigned different sketch sizes. For random projections, we use the "average" sample size. For example, in Figure 7.8, when the "sample size k" is 15, we assign sketch sizes 10, 15, and 20 to the data points whose sparsities fall in [0 1/3), [1/3, 2/3), and [2/3, 1] quantiles, respectively; and we let the sample size of random projections to be 15.

The results agree well with what we would expect:

- CRS works particularly well in estimating inner products, even in extremely heavy-tailed datasets (e.g., Figure 7.12).
- CRS is comparable to (Cauchy) random projections in estimating l_1 distances.
- Without using marginal information, CRS often performs poorly in estimating l_2 distances in extremely heavy-tailed data (e.g., Figures 7.10 and 7.12).
- Using marginal information, CRS can be improved dramatically and, in our experiments, always outperforms random projections even in extremely heavy-tailed data, in estimating l_2 distances.
- Adjusting the sketch size according to data sparsity in general improves the overall performance of CRS.

7.7 Summary

Large sparse datasets are often highly sparse, however, *stable random projections* can not take advantage of this important information. We propose *Conditional Random Sampling (CRS)*, which is provably better than *normal random projections*, at least for the important special cases of boolean data and nearly independent data. More generally, CRS compares favorably, both theoretically and empirically, especially when we take advantage of the margins.

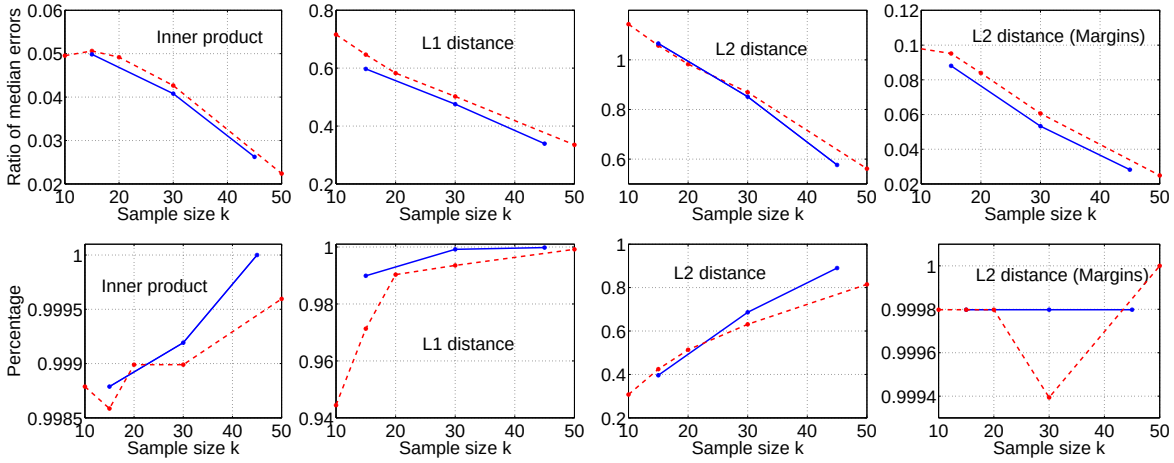


Figure 7.8: NSF data. CRS is overwhelmingly better than random projections. Upper four panels: ratios (CRS over random projections) of the median absolute errors; values < 1 indicate that CRS does better. Bottom four panels: percentage of pairs for which CRS has smaller errors than random projections; values > 0.5 indicate that CRS does better. Dashed curves correspond to fixed sample sizes while solid curves indicate that we (crudely) adjust sketch sizes according to data sparsity.

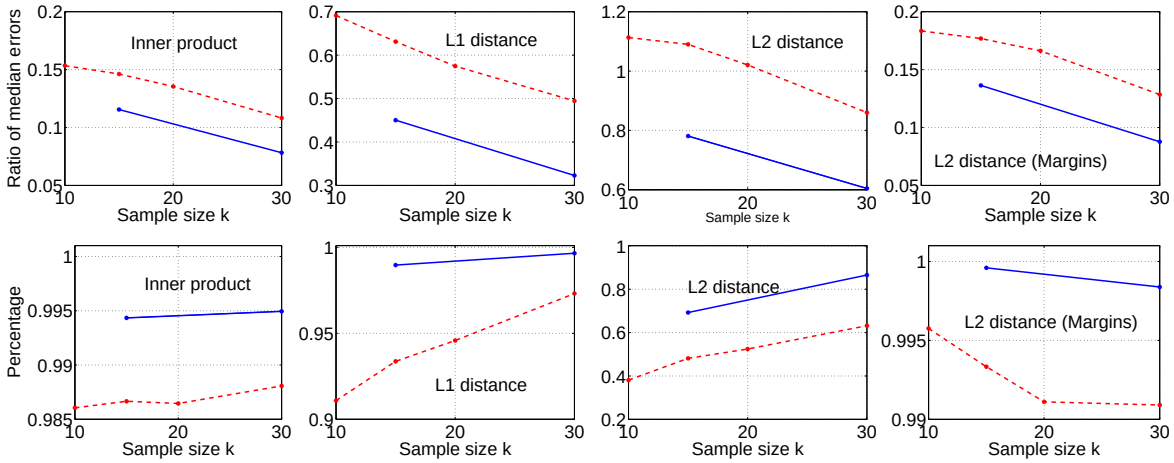


Figure 7.9: NEWSGROUP data. The results are similar to Figure 7.8 for the NSF data. CRS is overwhelmingly better than random projections in approximating inner products and l_2 distances (using margins).CRS is also significantly better in approximating l_1 distances. In this case, it is more obvious that adjusting sketch sizes helps.

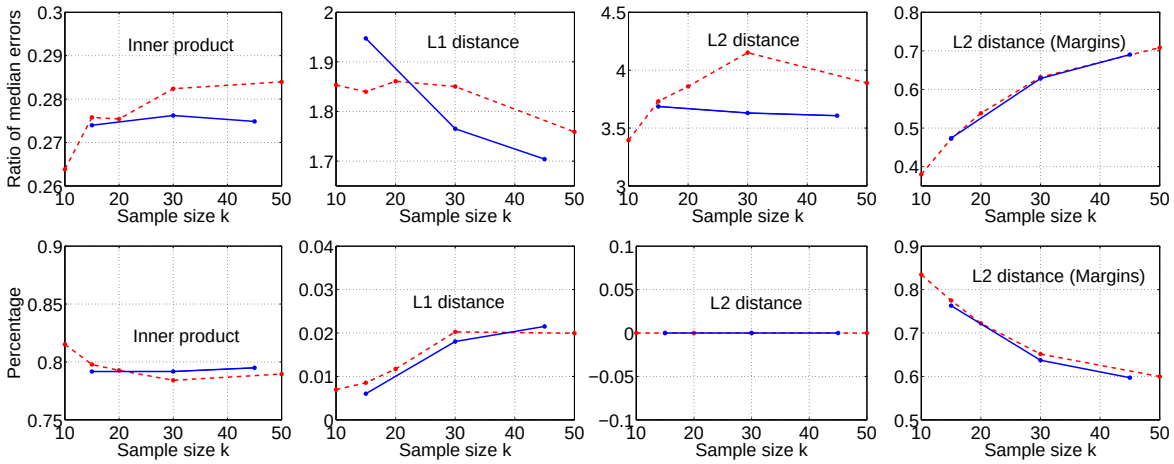


Figure 7.10: COREL image data. This dataset is more heavy-tailed than NSF and NEWSGROUP datasets but not as sparse. CRS is still much better than random projections in approximating inner products as well as l_2 distances (using margins). We observe that at large sample sizes, using margins (i.e., assuming normality) may cause quite noticeable biases. This, however, is a good bias-variance trade-off.

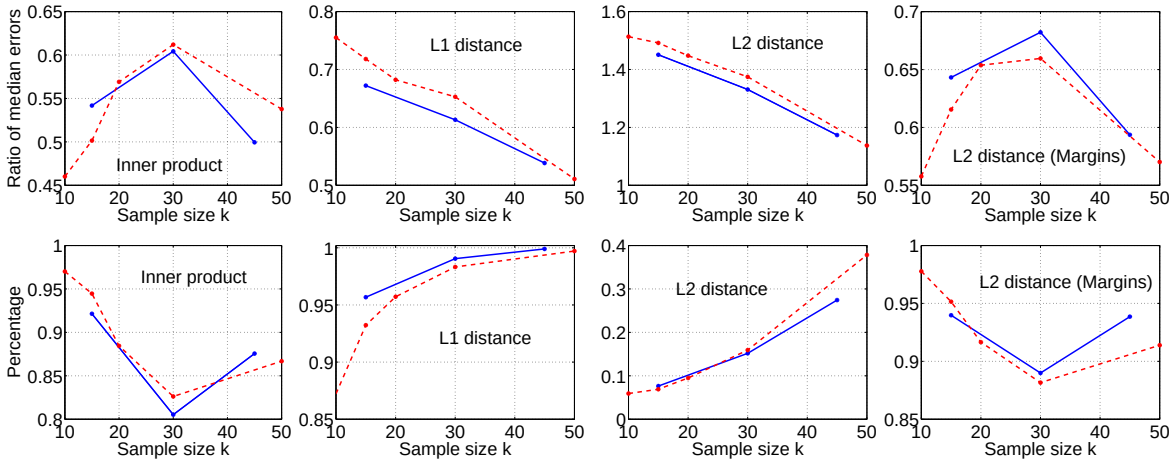


Figure 7.11: DEXTER data. The data are more heavy-tailed than NSF, NEWSGROUP, and COREL data. CRS still outperforms random projections in approximating inner products, l_1, and l_2 distances (using margins).

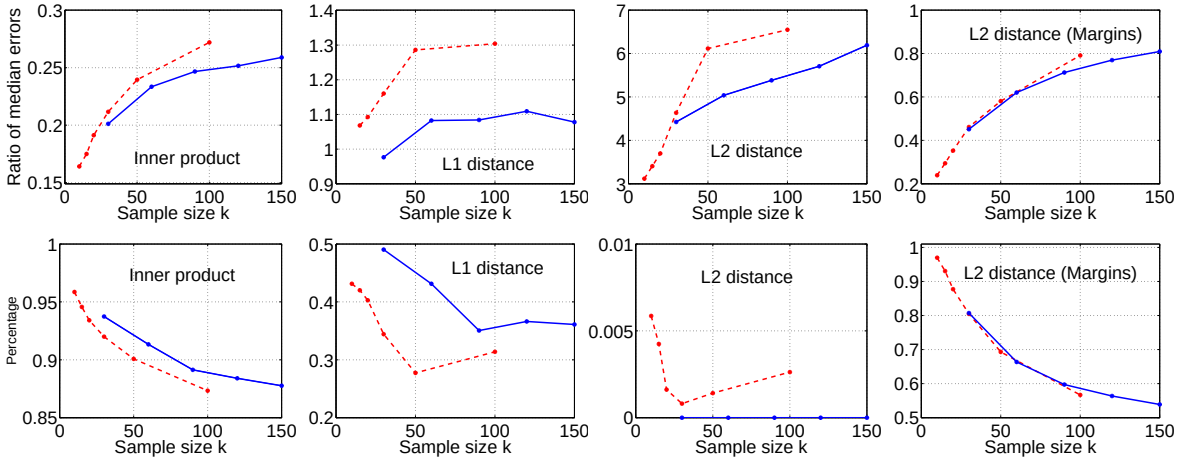


Figure 7.12: MSN data (original). The data are extremely heavy-tailed. Even so, *Conditional Random Sampling* is still significantly better than random projections in approximating inner products, and are also better in approximating l_2 distances using margins.

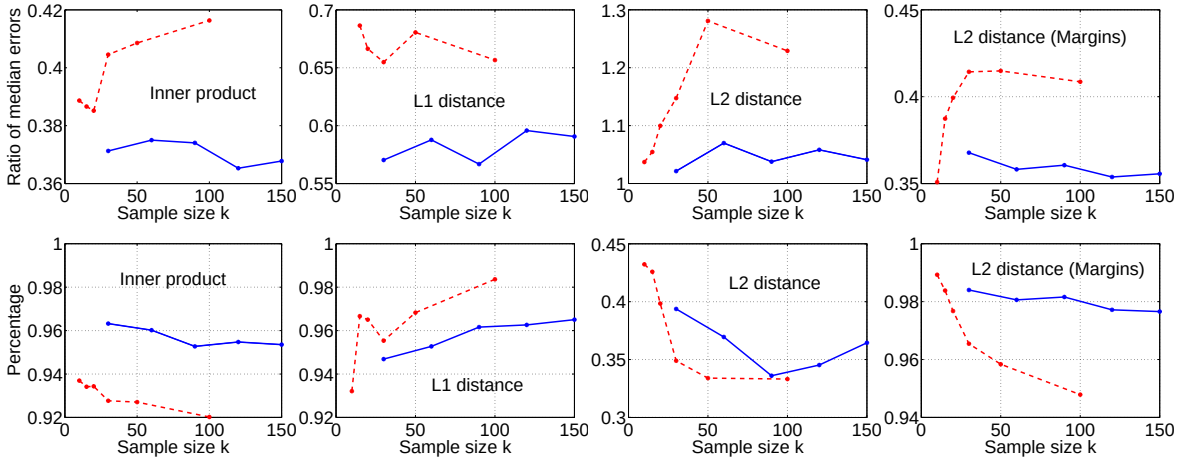


Figure 7.13: MSN data (square root weighting). CRS is significantly better than random projections in estimating inner products, l_1 distances, and l_2 distances (using margins). Without using margins, CRS is about the same as random projections in approximating l_2 distances, especially when the sketch sizes are adjusted.

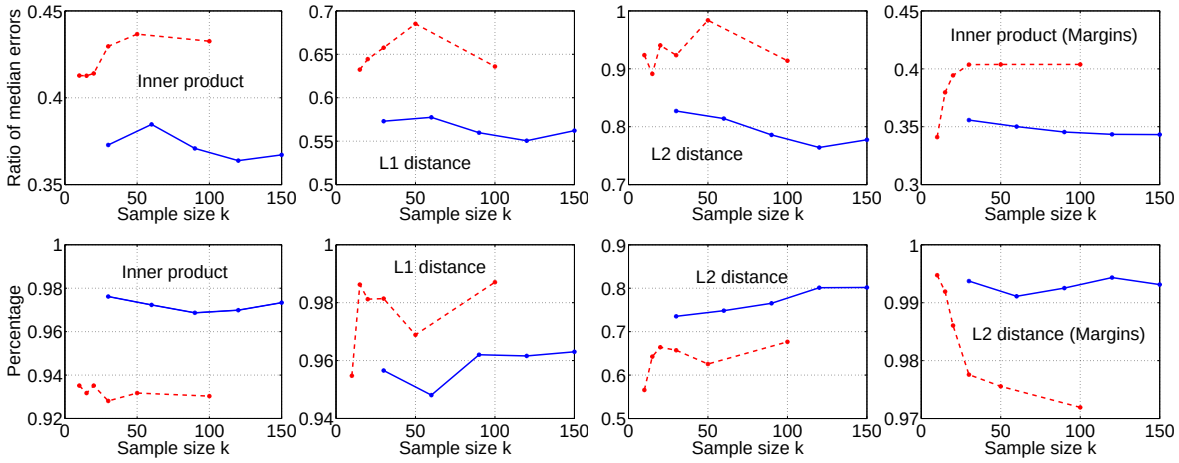


Figure 7.14: MSN data (logarithmic weighting). The results are even better, compared to Figure 7.13. In particular, when estimating l_2 distances without using margins, CRS is strictly better than random projections.

7.8 Covariance Matrix in Estimating Contingency Tables with Marginal Constraints

Recall that $\mathrm{S} = \{s_{i,j}\}_{i=0}^{I}{}_{j=0}^{J}$ denotes the sample contingency table and $\mathbf{X} = \{x_{i,j}\}_{i=0}^{I}{}_{j=0}^{J}$ denotes the original contingency table to be estimated. We vectorize the tables row-wise, i.e., $\mathbf{Z} = \mathrm{vec}\{\mathbf{X}\} = \{z_m\}_{m=1}^{(I+1)(J+1)}$ for the original contingency table and $\mathrm{H} = \mathrm{vec}\{\mathbf{S}\} = \{h_m\}_{m=1}^{(I+1)(J+1)}$ for the observed sample contingency table. We will give simple examples for $I = J = 2$ and $I = J = 1$ to help visualize the procedure.

There are in total $I + J + 1$ constraints, i.e., row sums $\{x_{i+}\}_{i=1}^{I}$, column sums $\{x_{+j}\}_{j=1}^{J}$, and the total sum $\sum_{m=1}^{(I+1)(J+1)} z_m = D$. Since the effective number of degrees of freedom is $I \times J$, we will partition the table into two parts: $\mathbf{Z}_1$ and $\mathbf{Z}_2$. $\mathbf{Z}_1$ corresponds to $\mathbf{X}_1 = \{x_{i,j}\}_{i=1}^{I}{}_{j=1}^{J}$ and $\mathbf{Z}_2$ corresponds to the rest of the table. The trick is to represent $\mathbf{Z}_2$ in terms of $\mathbf{Z}_1$ so that we can apply the multivariate large sample theory for the asymptotic covariance matrix of $\mathbf{Z}_1$. It is not hard to show that

$$\mathbf{Z}_2 = \mathbf{C}_1 - \mathbf{C}_2\mathbf{Z}_1, \tag{7.29}$$

where

$$\mathbf{C}_1 = \begin{bmatrix} x_{0+} + x_{+0} - D \\ \{x_{+j}\}_{j=1}^{J} \\ \{x_{i+}\}_{i=1}^{I} \end{bmatrix}, \qquad \mathbf{C}_2 = \begin{bmatrix} & -1_{IJ}^{\mathrm{T}} & & \\ \mathrm{I}_J & \mathrm{I}_J & ... & \mathrm{I}_J \\ 1_J^{\mathrm{T}} & 0_J^{\mathrm{T}} & ... & 0_J^{\mathrm{T}} \\ 0_J^{\mathrm{T}} & 1_J^{\mathrm{T}} & ... & 0_J^{\mathrm{T}} \\ & ... & & \\ 0_J^{\mathrm{T}} & 0_J^{\mathrm{T}} & ... & 1_J^{\mathrm{T}} \end{bmatrix}, \tag{7.30}$$

where I_J denotes the identity matrix of size $J \times J$, 1_J denotes a vector of ones of length J and 0_J denotes a vector of zeros of length J.

Assuming "sample-with-replacement," $\mathbf{Z}$ follows a multinomial distribution, with a log likelihood function (let $N = (I+1)(J+1)$):

$$Q(\mathbf{Z}) \propto \sum_{m=1}^{N} h_m \log z_m \tag{7.31}$$

In order to apply the large sample theory, we need to compute the Hessian ($\bigtriangledown^2 Q$), which is a matrix whose $(i,j)^{th}$ entry is the partial derivative $\frac{\partial^2 Q}{\partial z_i z_j}$, i.e.,

$$\bigtriangledown^2 Q = -\begin{bmatrix} \frac{h_1}{z_1^2} & 0 & \cdots & 0 \\ 0 & \frac{h_2}{z_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{h_N}{z_N^2} \end{bmatrix} = -\mathrm{diag}\left[\frac{h_1}{z_1^2}, \frac{h_2}{z_2^2}, ..., \frac{h_N}{z_N^2}\right]. \tag{7.32}$$

The log likelihood function Q, which is separable, can then be expressed as

$$Q(\mathbf{Z}) = Q_1(\mathbf{Z}_1) + Q_2(\mathbf{Z}_2). \tag{7.33}$$

By the matrix derivative chain rule, the Hessian of Q with respect to $\mathbf{Z}_1$ is

$$\bigtriangledown_1^2 Q = \bigtriangledown_1^2 Q_1 + \bigtriangledown_1^2 Q_2 = \bigtriangledown_1^2 Q_1 + \mathbf{C}_2^{\mathrm{T}} \bigtriangledown_2^2 Q_2 \mathbf{C}_2, \tag{7.34}$$

where $\bigtriangledown_1^2$ and $\bigtriangledown_2^2$ indicate the Hessians are with respect to $\mathbf{Z}_1$ and $\mathbf{Z}_2$, respectively.

The expected Fisher Information of $\mathbf{Z}_1$ is

$$\mathrm{I}(\mathbf{Z}_1) = \mathrm{E}\left(-\bigtriangledown_1^2 Q\right) = -\mathrm{E}(\bigtriangledown_1^2 Q_1) - \mathbf{C}_2^{\mathrm{T}}\mathrm{E}(\bigtriangledown_2^2 Q_2)\mathbf{C}_2. \tag{7.35}$$

Because $\mathrm{E}(h_m) = \frac{D_s}{D} z_m$, we can evaluate the above expectations, i.e.,

$$\mathrm{E}(-\bigtriangledown_1^2 Q_1) = \mathrm{diag}\left[\mathrm{E}\left(\frac{h_m}{z_m^2}\right), z_m \in \mathbf{Z}_1\right] = \frac{D_s}{D}\mathrm{diag}\left[\frac{1}{z_m}, z_m \in \mathbf{Z}_1\right], \tag{7.36}$$

$$\mathrm{E}(-\bigtriangledown_2^2 Q_2) = \frac{D_s}{D}\mathrm{diag}\left[\frac{1}{z_m}, z_m \in \mathbf{Z}_2\right]. \tag{7.37}$$

By the large sample theory, the asymptotic covariance matrix of $\mathbf{X}_1$ would be

$$\mathrm{Cov}(\mathbf{Z}_1) = \mathrm{I}(\mathbf{Z}_1)^{-1} = \frac{D}{D_s}\left(\mathrm{diag}\left[\frac{1}{z_m}, z_m \in \mathbf{Z}_1\right] + \mathbf{C}_2^{\mathrm{T}}\mathrm{diag}\left[\frac{1}{z_m}, z_m \in \mathbf{Z}_2\right]\mathbf{C}_2\right)^{-1}. \tag{7.38}$$

Of course, to obtain the unconditional covariance, we need to replace $\frac{D}{D_s}$ by $\mathrm{E}\left(\frac{D}{D_s}\right)$.

7.8.1 An Example with $I = 2, J = 2$

$$\begin{aligned}
&\mathrm{Z} = [z_1\ z_2\ z_3 \quad z_4\ z_5\ z_6 \quad z_7\ z_8\ z_9]^{\mathrm{T}} = [x_{0,0}\ x_{0,1}\ x_{0,2} \quad x_{1,0}\ x_{1,1}\ x_{1,2} \quad x_{2,0}\ x_{2,1}\ x_{2,2}]^{\mathrm{T}} \\
&\mathrm{Z}_1 = [z_5\ z_6\ z_8\ z_9]^{\mathrm{T}} = [x_{1,1}\ x_{1,2}\ x_{2,1}\ x_{2,2}]^{\mathrm{T}}, \\
&\mathrm{Z}_2 = [z_1\ z_2\ z_3\ z_4\ z_7]^{\mathrm{T}} = [x_{0,0}\ x_{0,1}\ x_{0,2}\ x_{1,0}\ x_{2,0}]^{\mathrm{T}}.
\end{aligned}$$

$$\mathbf{C}_1 = \begin{bmatrix} x_{0+} + x_{+0} - D \\ x_{+1} \\ x_{+2} \\ x_{1+} \\ x_{2+} \end{bmatrix}, \qquad \mathbf{C}_2 = \begin{bmatrix} -1 & -1 & -1 & -1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix},$$

$$\mathrm{Cov}(\mathbf{Z}_1) = \frac{D}{D_s} \times$$
$$\left(\mathrm{diag}\left[\frac{1}{x_{1,1}},\ \frac{1}{x_{1,2}},\ \frac{1}{x_{2,1}},\ \frac{1}{x_{2,2}} \right] + \mathbf{C}_2^{\mathrm{T}} \mathrm{diag}\left[\frac{1}{x_{0,0}},\ \frac{1}{x_{0,1}},\ \frac{1}{x_{0,2}},\ \frac{1}{x_{1,0}},\ \frac{1}{x_{2,0}} \right] \mathbf{C}_2 \right)^{-1}.$$

7.8.2 An Example with $I = 1, J = 1$ (0/1 Data)

$$\begin{aligned} \mathrm{Z} &= [z_1\ z_2\ z_3\ z_4]^{\mathrm{T}} = [x_{0,0}\ x_{0,1}\ x_{1,0}\ x_{1,1}]^{\mathrm{T}} \\ \mathrm{Z}_1 &= [z_4]^{\mathrm{T}} = [x_{1,1}]^{\mathrm{T}}, \\ \mathrm{Z}_2 &= [z_1\ z_2\ z_3]^{\mathrm{T}} = [x_{0,0}\ x_{0,1}\ x_{1,0}]^{\mathrm{T}}. \end{aligned}$$

$$\mathbf{C}_2 = [-1\ \ 1\ \ 1]^{\mathrm{T}},$$

$$\begin{aligned} \mathrm{Cov}(\mathbf{Z}_1) &= \frac{D}{D_s} \left(\mathrm{diag}\left[\frac{1}{x_{1,1}} \right] + \mathbf{C}_2^{\mathrm{T}} \mathrm{diag}\left[\frac{1}{x_{0,0}},\ \frac{1}{x_{0,1}},\ \frac{1}{x_{1,0}} \right] \mathbf{C}_2 \right)^{-1} \\ &= \frac{D}{D_s} \frac{1}{\frac{1}{x_{1,1}} + \frac{1}{x_{0,0}} + \frac{1}{x_{0,1}} + \frac{1}{x_{1,0}}} \end{aligned}$$

7.8.3 The Variance of $\hat{a}_{MLE,c}$

Since estimate the inner product from the contingency table as

$$\hat{a}_{MLE,c} = \sum_{i=1}^{I} \sum_{j=1}^{J} (ij) \hat{x}_{i,j,MLE}.$$

the variance would be

$$\mathrm{Var}\left(\hat{a}_{MLE,c}\right) = \mathbf{e}^{\mathrm{T}} \mathrm{Cov}\left(\hat{\mathbf{Z}}_1\right) \mathbf{e}, \tag{7.39}$$

$$\mathbf{e}^{\mathrm{T}} = [(ij)]_{i=1\ \ j=1}^{I\ \ J} = [1, 2, ..., J,\ \ 2, 4, ..., 2J,\ \ ...,\ \ I, ..., IJ]^{\mathrm{T}}. \tag{7.40}$$

Chapter 8

Conditional Random Sampling in Boolean Data

Boolean (0/1) data are important in practice and quite easy for theoretical analysis. Applying *Conditional Random Sampling (CRS)* to boolean data amounts to estimating (two-way) and (multi-way) contingency tables. Also, in boolean data, the benefits of using the margins (row and column sums of the table) are considerable and can be analyzed accurately. In addition, towards the end of this chapter, we will compare CRS with "Broder's sketch," which was originally developed for boolean data.

In this chapter, we will use *word associations* as the running example to explain the methodology and evaluate the estimators.

8.1 An Example of CRS in Boolean Data

Figure 8.1 shows an example of *Conditional Random Sampling (CRS)* with more than two words. There are D=15 documents in the collection. We generate a random permutation π in Figure 8.1(b). For every document ID in the inverted index P_i in Figure 8.1(a), we apply the random permutation π, but we only store the k_i smallest IDs as a sketch K_i, denoted by $\mathrm{K}_i = \mathrm{MIN}_{k_i}(\pi(\mathrm{P}_i))$. In this example, we choose $k_1 = 4$, $k_2 = 4$, $k_3 = 4$, $k_4 = 3$, $k_5 = 6$. The sketches are stored in Figure 8.1(c). In addition, since $\pi(\mathrm{P}_i)$ operates on every ID in P_i, we know the total number of

P_1	1 4 5 7 11 13 15
P_2	2 4 7 8 10 11 13
P_3	1 3 4 6 9 10 14
P_4	2 5 8
P_5	1 2 5 6 8 9 11 12 13 14 15

(a) Inverted index

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

(b) π

K_1	1 6 8 11
K_2	3 6 10 12
K_3	2 5 7 9
K_4	3 6 12
K_5	1 2 3 4 5 6

(c) Sketches

Figure 8.1: The original inverted index is given in (a). $D = 15$ documents in the collection. We generate a random permutation π in (b). We apply π to P_i and store the sketch $K_i = MIN_{k_i}(\pi(P_i))$. For example, $\pi(P_1) = \{11, 13, 1, 12, 15, 6, 8\}$ and we choose $k_1 = 4$; hence $K_1 = \{1, 6, 8, 11\}$. We choose $k_2 = 4$, $k_3 = 4$, $k_4 = 3$, and $k_5 = 6$.

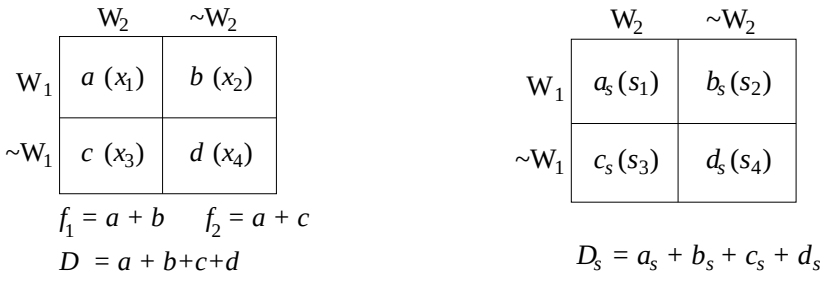
non-zeros in P_i, denoted by $f_i = |P_i|$. f_i is the margin, also known as the *document frequency* in information retrieval (IR).

The estimation is straightforward if we ignore the margins. For example, suppose we need to estimate the number of documents containing the first two words, i.e., the inner product between P_1 and P_2, denoted by $a_{(1,2)}$. (We have to use the subscript $_{(1,2)}$ because we have more than just two words in the vocabulary.) We calculate, from sketches K_1 and K_2, the sample inner product $a_{s,(1,2)} = |\{6\}| = 1$, and the corresponding sample size, denoted by $D_{s,(1,2)} = \min(\max(K_1), \max(K_2)) = \min(11, 12) = 11$. Therefore, the "margin-free" estimate of $a_{s,(1,2)}$ is simply $a_{s,(1,2)} \frac{D}{D_{s,(1,2)}} = 1\frac{15}{11} = 1.4$.

The procedure can be easily extended to more than two rows. Suppose we would like to estimate the three-way inner product among P_1, P_4, and P_5, denoted by $a_{(1,4,5)}$. We calculate the three-way sample inner product from K_1, K_4, and K_5, $a_{s,(1,4,5)} = |\{6\}| = 1$, and the corpus sample size $D_{s,(1,4,5)} = \min(\max(K_1), \max(K_4), \max(K_5)) = \min(11, 12, 6) = 6$. Then the "margin-free" estimate of $a_{(1,4,5)}$ is $1\frac{15}{6} = 2.5$.

8.2 Estimators for Two-Way Associations

In the pair-wise case, conditioning on D_s, the task is to estimate the contingency table (a, b, c, d) from the sample contingency table (a_s, b_s, c_s, d_s), shown in Figure 8.2.



(a) Contingency table (b) Sample table

Figure 8.2: (a): A contingency table for word W_1 and word W_2. Cell a is the number of documents that contain both W_1 and W_2, b is the number that contain W_1 but not W_2, c is the number that contain W_2 but not W_1, and d is the number that contain neither. The margins, $f_1 = a + b$ and $f_2 = a + c$ are known as document frequencies. For consistency with the notation for multi-way associations, (a, b, c, d) are also denoted by (x_1, x_2, x_3, x_4). (b): A sample contingency table (a_s, b_s, c_s, d_s), also denoted by (s_1, s_2, s_3, s_4).

We factor the likelihood (probability mass function, PMF) $\mathbf{Pr}(a_s, b_s, c_s, d_s; a)$ into

$$\mathbf{Pr}(a_s, b_s, c_s, d_s; a) = \mathbf{Pr}(a_s, b_s, c_s, d_s | D_s; a) \times \mathbf{Pr}(D_s; a) \tag{8.1}$$

We seek the a that maximizes the *partial likelihood* $\mathbf{Pr}(a_s, b_s, c_s, d_s | D_s; a)$, i.e.,

$$\hat{a}_{MLE} = \underset{a}{\text{argmax}} \mathbf{Pr}\left(a_s, b_s, c_s, d_s | D_s; a\right) = \underset{a}{\text{argmax}} \log \mathbf{Pr}\left(a_s, b_s, c_s, d_s | D_s; a\right). \tag{8.2}$$

$\mathbf{Pr}(a_s, b_s, c_s, d_s | D_s; a)$ is the PMF of a two-way sample contingency table. That is relatively straightforward, but $\mathbf{Pr}(D_s; a)$ is difficult. As illustrated in Figure 7.3, there is no strong dependency of D_s on a, and therefore, we focus on the easy part.

Before we delve into maximizing $\mathbf{Pr}(a_s, b_s, c_s, d_s | D_s; a)$ under margin constraints, we will first consider two simplifications, which lead to two baseline estimators.

8.2.1 The Independence Baseline

Independence assumptions are often made in databases [81, Chapter 16.4] and NLP [135, Chapter 13.3]. When two words W_1 and W_2 are independent, the size of intersections, a, follows a hypergeometric distribution,

$$\mathbf{Pr}(a) = \binom{f_1}{a}\binom{D-f_1}{f_2-a}\Big/\binom{D}{f_2}, \tag{8.3}$$

where $\binom{n}{m} = \frac{n!}{m!(n-m)!}$. This distribution suggests an estimator

$$\hat{a}_{IND} = \mathrm{E}(a) = \frac{f_1 f_2}{D}. \tag{8.4}$$

(8.3) is also a common null-hypothesis distribution in testing independence of a two-way contingency table, i.e., the so-called "Fisher's exact test" [11, Section 3.5.1].

8.2.2 The Margin-free Baseline

Conditional on D_s, the sample contingency table (a_s, b_s, c_s, d_s) follows the *multivariate hypergeometric* distribution with moments[1]

$$\begin{aligned}&\mathrm{E}(a_s|D_s) = \frac{D_s}{D}a, \quad \mathrm{E}(b_s|D_s) = \frac{D_s}{D}b, \quad \mathrm{E}(c_s|D_s) = \frac{D_s}{D}c, \quad \mathrm{E}(d_s|D_s) = \frac{D_s}{D}d,\\ &\mathrm{Var}(a_s|D_s) = D_s\frac{a}{D}\left(1-\frac{a}{D}\right)\frac{D-D_s}{D-1},\end{aligned} \tag{8.5}$$

where the term $\frac{D-D_s}{D-1} \approx 1 - \frac{D_s}{D}$, is known as the "finite population correction factor."

An unbiased estimator and its variance would be

$$\hat{a}_{MF} = \frac{D}{D_s}a_s, \quad \mathrm{Var}(\hat{a}_{MF}|D_s) = \frac{D^2}{D_s^2}\mathrm{Var}(a_s|D_s) = \frac{D}{D_s}\frac{1}{\frac{1}{a}+\frac{1}{D-a}}\frac{D-D_s}{D-1}, \tag{8.6}$$

which we refer to as "margin-free" because it does not take advantage of the margins.

The multivariate hypergeometric distribution can be simplified to a *multinomial* assuming "sample-with-replacement," often a good approximation when $\frac{D_s}{D}$ is small.

[1] `http://www.ds.unifi.it/VL/VL_EN/urn/urn4.html`

According to the multinomial model, an estimator and its variance would be

$$\hat{a}_{MF,r} = \frac{D}{D_s} a_s, \quad \mathrm{Var}(\hat{a}_{MF,r}|D_s) = \frac{D}{D_s} \frac{1}{\frac{1}{a} + \frac{1}{D-a}}. \tag{8.7}$$

Note that these expectations in (8.5) hold both when the margins are known, as well as when they are not known, because the samples (a_s, b_s, c_s, d_s) are obtained without consulting the margins.

8.2.3 The Exact MLE with Margin Constraints

Considering the margin constraints, the partial likelihood $\mathbf{Pr}\left(a_s, b_s, c_s, d_s | D_s; a\right)$ can be expressed as a function of a single unknown parameter, a:

$$\begin{aligned}
\mathbf{Pr}\left(a_s, b_s, c_s, d_s | D_s; a\right) &= \frac{\binom{a}{a_s}\binom{b}{b_s}\binom{c}{c_s}\binom{d}{d_s}}{\binom{a+b+c+d}{a_s+b_s+c_s+d_s}} = \frac{\binom{a}{a_s}\binom{f_1-a}{b_s}\binom{f_2-a}{c_s}\binom{D-f_1-f_2+a}{d_s}}{\binom{D}{D_s}} \\
&\propto \frac{a!}{(a-a_s)!} \times \frac{(f_1-a)!}{(f_1-a-b_s)!} \times \frac{(f_2-a)!}{(f_2-a-c_s)!} \times \frac{(D-f_1-f_2+a)!}{(D-f_1-f_2+a-d_s)!} \\
&= \prod_{i=0}^{a_s-1}(a-i) \times \prod_{i=0}^{b_s-1}(f_1-a-i) \times \prod_{i=0}^{c_s-1}(f_2-a-i) \times \prod_{i=0}^{d_s-1}(D-f_1-f_2+a-i).
\end{aligned} \tag{8.8}$$

Let $\hat{a}_{MLE}$ be the value of a that maximizes the partial likelihood (8.8), or equivalently, maximizes the log likelihood, $\log \mathbf{Pr}\left(a_s, b_s, c_s, d_s | D_s; a\right)$:

$$\sum_{i=0}^{a_s-1} \log(a-i) + \sum_{i=0}^{b_s-1} \log\left(f_1-a-i\right) + \sum_{i=0}^{c_s-1} \log\left(f_2-a-i\right) + \sum_{i=0}^{d_s-1} \log\left(D-f_1-f_2+a-i\right),$$

whose first derivative, $\frac{\partial \log \mathbf{Pr}(a_s,b_s,c_s,d_s|D_s;a)}{\partial a}$, is

$$\sum_{i=0}^{a_s-1} \frac{1}{a-i} - \sum_{i=0}^{b_s-1} \frac{1}{f_1-a-i} - \sum_{i=0}^{c_s-1} \frac{1}{f_2-a-i} + \sum_{i=0}^{d_s-1} \frac{1}{D-f_1-f_2+a-i}. \tag{8.9}$$

Since the second derivative, $\frac{\partial^2 \log \mathbf{Pr}(a_s,b_s,c_s,d_s|D_s;a)}{\partial a^2}$,

$$-\sum_{i=0}^{a_s-1}\frac{1}{(a-i)^2}-\sum_{i=0}^{b_s-1}\frac{1}{(f_1-a-i)^2}-\sum_{i=0}^{c_s-1}\frac{1}{(f_2-a-i)^2}-\sum_{i=0}^{d_s-1}\frac{1}{(D-f_1-f_2+a-i)^2},$$

is negative, the log likelihood function is concave, and therefore, there is a unique maximum. One could solve (8.9) for $\frac{\partial \log \mathbf{Pr}(a_s,b_s,c_s,d_s|D_s;a)}{\partial a}=0$ numerically, but it turns out there is a more direct solution using the updating formula from (8.8):

$$\mathbf{Pr}\left(a_s,b_s,c_s,d_s|D_s;a\right)=\mathbf{Pr}\left(a_s,b_s,c_s,d_s|D_s;a-1\right)\times g(a). \tag{8.10}$$

Since the MLE exists and is unique, it suffices to find the a such that $g(a)=1$,

$$g(a)=\frac{a}{a-a_s}\frac{f_1-a+1-b_s}{f_1-a+1}\frac{f_2-a+1-c_s}{f_2-a+1}\frac{D-f_1-f_2+a}{D-f_1-f_2+a-d_s}=1, \tag{8.11}$$

which is cubic in a, because the a^4 term in the numerator cancels the a^4 term in the denominator after expanding (8.11).

8.2.4 The "Sample-with-replacement" Simplification

Assuming "sample-with-replacement," the likelihood function is slightly simpler:

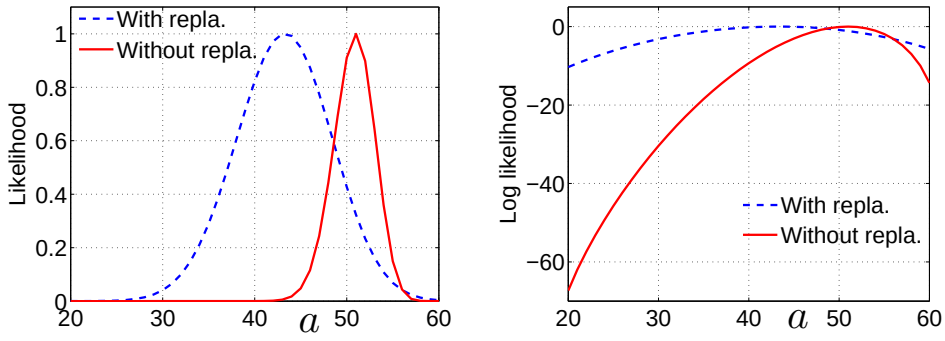
$$\begin{aligned}\mathbf{Pr}(a_s,b_s,c_s,d_s|D_s;a,r)&=\binom{D_s}{a_s,b_s,c_s,d_s}\left(\frac{a}{D}\right)^{a_s}\left(\frac{b}{D}\right)^{b_s}\left(\frac{c}{D}\right)^{c_s}\left(\frac{d}{D}\right)^{d_s}\\&\propto a^{a_s}(f_1-a)^{b_s}(f_2-a)^{c_s}(D-f_1-f_2+a)^{d_s}.\end{aligned} \tag{8.12}$$

Setting the first derivative of the log likelihood to be zero yields a cubic equation:

$$\frac{a_s}{a}-\frac{b_s}{f_1-a}-\frac{c_s}{f_2-a}+\frac{d_s}{D-f_1-f_2+a}=0. \tag{8.13}$$

As shown in Section 8.2.2, using the margin-free model, the "sample-with-replacement" assumption amplifies the variance but does not change the estimation. With the proposed MLE, the "sample-with-replacement" assumption will change the estimation,

although in general we do not expect the differences to be large. Figure 8.3 gives an (exaggerated) example, to show the concavity of the log likelihood and the difference caused by assuming "sample-with-replacement."



(a) Likelihood (b) Log likelihood

Figure 8.3: An example: $a_s = 20$, $b_s = 40$, $c_s = 40$, $d_s = 800$, $f_1 = f_2 = 100$, $D = 1000$. The estimated $\hat{a} = 43$ for "sample-with-replacement," and $\hat{a} = 51$ for "sample-without-replacement." (a): The likelihood profile, normalized to have a maximum = 1. (b): The log likelihood profile, normalized to have a maximum = 0.

8.2.5 A Convenient Practical Quadratic Approximation

Solving a cubic equation for the exact MLE may be so inconvenient that one may prefer the less accurate margin-free baseline because of its simplicity. This section derives a convenient closed-form quadratic approximation to the exact MLE.

The idea is to assume "sample-with-replacement" and that one can identify a_s from K_1 without knowledge of K_2. In other words, we assume $a_s^{(1)} \sim \text{Binomial}(a_s + b_s, \frac{a}{f_1})$, $a_s^{(2)} \sim \text{Binomial}(a_s + c_s, \frac{a}{f_2})$, $a_s^{(1)}$ and $a_s^{(2)}$ are independent with $a_s^{(1)} = a_s^{(2)} = a_s$. The PMF of $(a_s^{(1)}, a_s^{(2)})$ is a product of two binomials:

$$\left[\binom{f_1}{a_s + b_s}\left(\frac{a}{f_1}\right)^{a_s}\left(\frac{f_1 - a}{f_1}\right)^{b_s}\right] \times \left[\binom{f_2}{a_s + c_s}\left(\frac{a}{f_2}\right)^{a_s}\left(\frac{f_2 - a}{f_2}\right)^{c_s}\right]$$
$$\propto a^{2a_s}\left(f_1 - a\right)^{b_s}\left(f_2 - a\right)^{c_s}. \tag{8.14}$$

Setting the first derivative of the logarithm of (8.14) to be zero, we obtain

$$\frac{2a_s}{a} - \frac{b_s}{f_1 - a} - \frac{c_s}{f_2 - a} = 0, \tag{8.15}$$

which is quadratic in a and has a convenient closed-form solution:

$$\hat{a}_{MLE,a} = \frac{f_1 (2a_s + c_s) + f_2 (2a_s + b_s) - \sqrt{(f_1 (2a_s + c_s) - f_2 (2a_s + b_s))^2 + 4 f_1 f_2 b_s c_s}}{2 (2a_s + b_s + c_s)}. \tag{8.16}$$

The second root can be ignored because it is always out of range:

$$\begin{aligned}
&\frac{f_1 (2a_s + c_s) + f_2 (2a_s + b_s) + \sqrt{(f_1 (2a_s + c_s) - f_2 (2a_s + b_s))^2 + 4 f_1 f_2 b_s c_s}}{2 (2a_s + b_s + c_s)} \\
\geq &\frac{f_1 (2a_s + c_s) + f_2 (2a_s + b_s) + |f_1 (2a_s + c_s) - f_2 (2a_s + b_s)|}{2 (2a_s + b_s + c_s)} \\
\geq &\begin{cases} f_1 & \text{if } f_1 (2a_s + c_s) \geq f_2 (2a_s + b_s) \\ f_2 & \text{if } f_1 (2a_s + c_s) < f_2 (2a_s + b_s) \end{cases} \geq \min(f_1, f_2).
\end{aligned}$$

The evaluation in Section 8.3 showed that $\hat{a}_{MLE,a}$ is close to $\hat{a}_{MLE}$.

8.2.6 The Variance and Bias

Usually, a maximum likelihood estimator is nearly unbiased. Furthermore, assuming "sample-with-replacement," we can apply the large sample theory[2], which says that $\hat{a}_{MLE}$ is asymptotically unbiased and converges in distribution to a normal with

[2]See [148, 149] for regularity conditions to ensure convergence in "sample-without-replacement."

variance $\frac{1}{\mathrm{I}(a)}$, where $\mathrm{I}(a)$, the expected Fisher Information, is

$$\begin{aligned}
\mathrm{I}(a) &= -\mathrm{E}\left(\frac{\partial^2}{\partial a^2}\log \mathbf{Pr}\left(a_s, b_s, c_s, d_s | D_s; a, r\right)\right) \\
&= \mathrm{E}\left(\frac{a_s}{a^2} + \frac{b_s}{(f_1-a)^2} + \frac{c_s}{(f_2-a)^2} + \frac{d_s}{(D-f_1-f_2+a)^2}\middle| D_s\right). \\
&= \frac{\mathrm{E}(a_s|D_s)}{a^2} + \frac{\mathrm{E}(b_s|D_s)}{(f_1-a)^2} + \frac{\mathrm{E}(c_s|D_s)}{(f_2-a)^2} + \frac{\mathrm{E}(d_s|D_s)}{(D-f_1-f_2+a)^2} \\
&= \frac{D_s}{D}\left(\frac{1}{a} + \frac{1}{f_1-a} + \frac{1}{f_2-a} + \frac{1}{D-f_1-f_2+a}\right), \qquad (8.17)
\end{aligned}$$

where we evaluate $\mathrm{E}(a_s|D_s)$, $\mathrm{E}(b_s|D_s)$, $\mathrm{E}(c_s|D_s)$, $\mathrm{E}(d_s|D_s)$ by (8.5).

For "sample-without-replacement," we correct the asymptotic variance $\frac{1}{\mathrm{I}(a)}$ by multiplying by the finite population correction factor $1 - \frac{D_s}{D}$:

$$\mathrm{Var}\left(\hat{a}_{MLE}|D_s\right) \approx \frac{1}{\mathrm{I}(a)}\left(1 - \frac{D_s}{D}\right) = \frac{\frac{D}{D_s} - 1}{\frac{1}{a} + \frac{1}{f_1-a} + \frac{1}{f_2-a} + \frac{1}{D-f_1-f_2+a}}. \qquad (8.18)$$

Comparing (8.6) with (8.18), we know that $\mathrm{Var}\left(\hat{a}_{MLE}|D_s\right) < \mathrm{Var}\left(\hat{a}_{MF}|D_s\right)$, and the difference could be substantial.

The unconditional variance is (asymptotically)

$$\mathrm{Var}\left(\hat{a}_{MLE}\right) \approx \frac{\mathrm{E}\left(\frac{D}{D_s}\right) - 1}{\frac{1}{a} + \frac{1}{f_1-a} + \frac{1}{f_2-a} + \frac{1}{D-f_1-f_2+a}}, \qquad (8.19)$$

$$\approx \frac{\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right) - 1}{\frac{1}{a} + \frac{1}{f_1-a} + \frac{1}{f_2-a} + \frac{1}{D-f_1-f_2+a}}, \qquad (8.20)$$

using the approximation in (7.10) proposed in Chapter 7.

8.3 Evaluation of Two-Way Associations

We consider a chunk of MSN Web crawls ($D = 2^{16}$). We collected two sets of English words which we refer to as the "small dataset" and the "large dataset". The small

dataset contains just four high frequency words: "THIS," "HAVE," "HELP" and "PROGRAM" (see Table 8.1), whereas the large dataset contains 968 words (i.e., 468,028 pairs). The large dataset was constructed by taking a random sample of English words that appeared in at least 20 documents in the collection. The histograms of the margins and co-occurrences have long tails, as expected (see Figure 8.4).

Table 8.1: Small dataset: co-occurrences and margins for the population. The task is to estimate these values, which are referred to as the gold standard.

Case #	Words	Co-occurrence (a)	Margin (f_1)	Margin (f_2)
Case 2-1	THIS & HAVE	13,517	27633	17369
Case 2-2	THIS & HELP	7221	27633	10791
Case 2-3	THIS & PROGRAM	3682	27633	5327
Case 2-4	HAVE & HELP	5781	17369	10791
Case 2-5	HAVE & PROGRAM	3029	17369	5327
Case 2-6	HELP & PROGRAM	1949	17369	5327

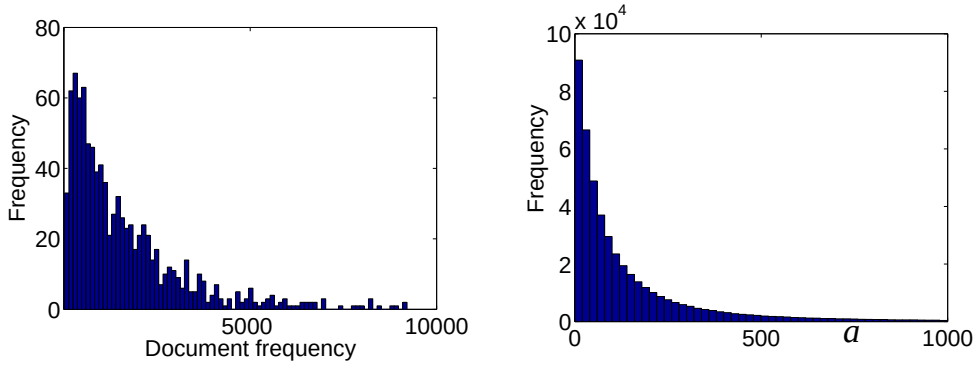


Figure 8.4: Large dataset: histograms of document frequencies, df (left panel), and co-occurrences, a (right panel). Left panel: max document frequency df = 42,564, median = 1135, mean = 2135, standard deviation = 3628. Right panel: max co-occurrence a = 33,045, mean = 188, median = 74, standard deviation = 459.

For the small dataset, we applied 10^5 independent random permutations to the $D = 2^{16}$ document IDs, $\Omega = \{1, 2, ..., D\}$. High frequency words were selected so we could study a large range of sampling rates ($\frac{k}{f}$), from 0.002 to 0.95. A pair of sketches was constructed for each of the 6 pairs of words in Table 8.1, each of the 10^5 permutations and each sampling rate. Mean square errors (MSE) and other statistics

were computed by aggregating over the 10^5 Monte Carlo trials.

The large dataset experiment contains many words with a large range of frequencies; and hence the experiment was repeated just six times (i.e., six different permutations). With such a large range of frequencies and sampling rates, there is a danger that some samples would be too small. A floor was imposed to make sure that every sample contains at least 20 documents.

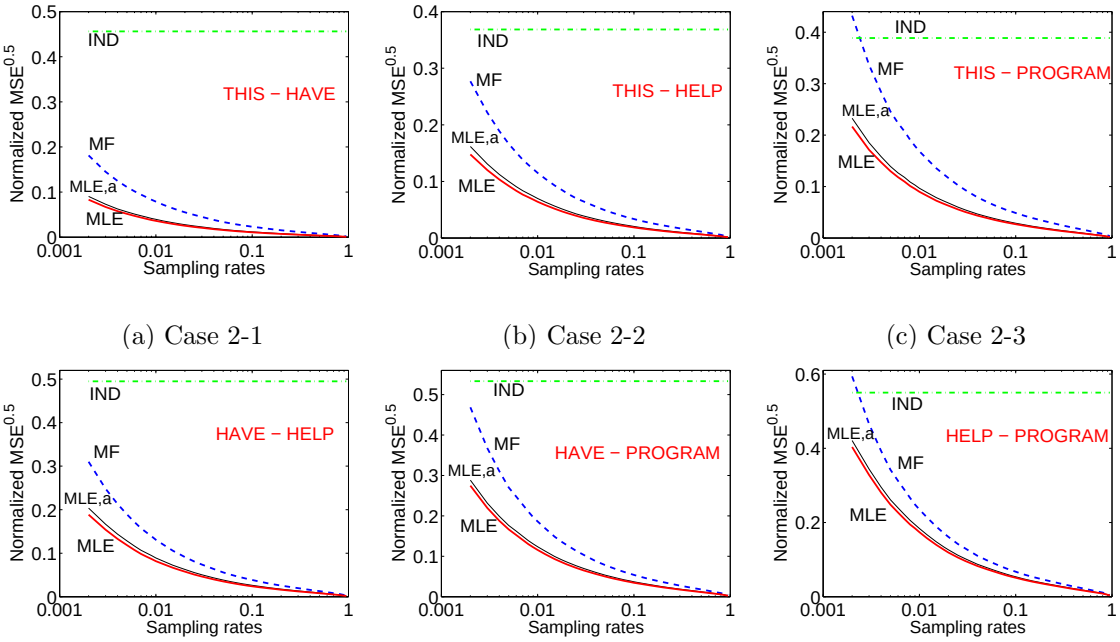
8.3.1 Results from Small Dataset Experiment

Figure 8.5 shows that the proposed methods (solid lines) are better than the baselines (dashed lines), in terms of MSE, estimated by the large Monte Carlo experiment over the small dataset, as described above. The independence baseline ($\hat{a}_{IND}$), which does not take advantage of the sample, has very large errors. The sample is a very useful source of information; even a small sample is much better than no sample.

The recommended quadratic approximation, $\hat{a}_{MLE,a}$, is remarkably close to the exact MLE solution. Both of the proposed methods, $\hat{a}_{MLE,a}$ and $\hat{a}_{MLE}$ (solid lines), have much smaller MSE than the margin-free baseline $\hat{a}_{MF}$ (dashed lines), especially at low sampling rates. When we know the margins, we ought to use them.

Margin Constraints Improve Smoothing

Though not a major emphasis, Figure 8.6 shows that smoothing is effective at low sampling rates, but only for those methods that take advantage of the margin constraints (solid lines as opposed to dashed lines). Figure 8.6 compares smoothed estimates ($\hat{a}_{MLE}$, $\hat{a}_{MLE,a}$, and $\hat{a}_{MF}$) with their un-smoothed counterparts. The y-axis reports percentage improvement of the MSE due to smoothing. Smoothing helps the proposed methods (solid lines) for all six word pairs, and hurts the baseline methods (dashed lines), for most of the six word pairs. We believe margin constraints keep the smoother from wandering too far astray; without margin constraints, smoothing can easily do more harm than good, especially when the smoother isn't very good. In this experiment, we used the simple "add-one" smoother that replaces a_s, b_s, c_s and d_s with $a_s + 1$, $b_s + 1$, $c_s + 1$ and $d_s + 1$, respectively.



(a) Case 2-1 (b) Case 2-2 (c) Case 2-3

(d) Case 2-4 (e) Case 2-5 (f) Case 2-6

Figure 8.5: The proposed estimator, $\hat{a}_{MLE}$, outperforms the margin-free baseline, $\hat{a}_{MF}$, in terms of $\frac{\sqrt{\text{MSE}}}{a}$. The quadratic approximation, $\hat{a}_{MLE,a}$, is close to $\hat{a}_{MLE}$. All methods are better than assuming independence (IND).

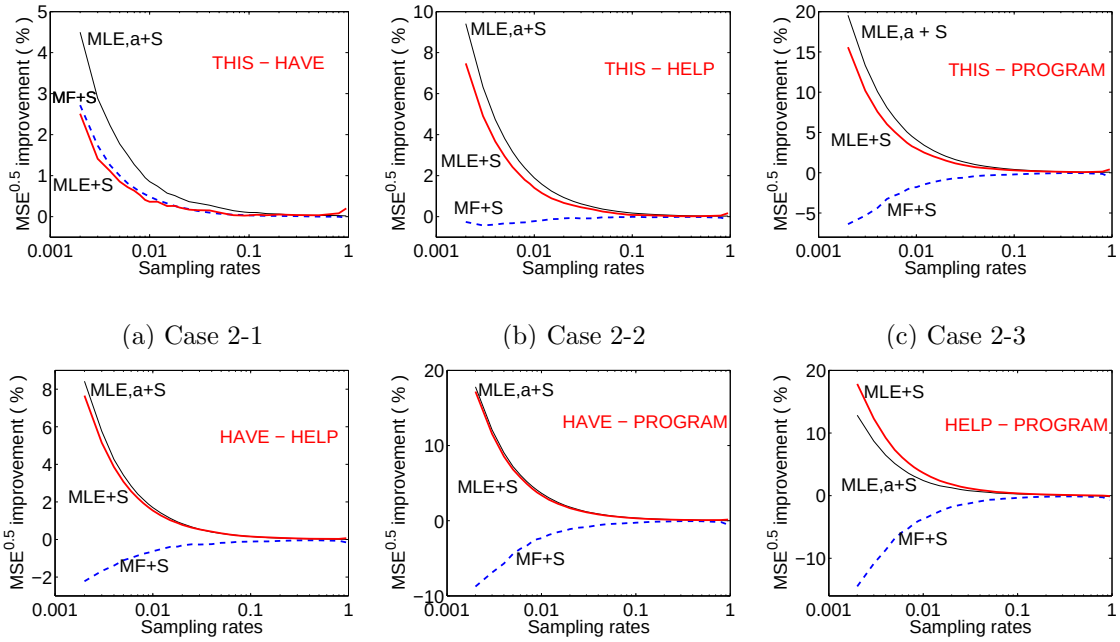
8.3.2 Results from Large Dataset Experiment

In Figure 8.7, the large dataset experiment confirms the findings of the large Monte Carlo Experiment: the proposed MLE method is better than the margin-free and independence baselines. The recommended quadratic approximation, $\hat{a}_{MLE,a}$, is close to the exact solution, $\hat{a}_{MLE}$.

8.3.3 Rank Retrieval by Cosine

We are often interested in finding top ranking pairs according to some similarity measure such as cosine. Figure 8.8 shows that we can find many of the top ranking pairs, even at low sampling rates.

Note that the estimate of cosine, $\frac{a}{\sqrt{f_1 f_2}}$, depends solely on the estimate of a, since we know the margins, f_1 and f_2. If we sort word pairs by their cosines, using estimates



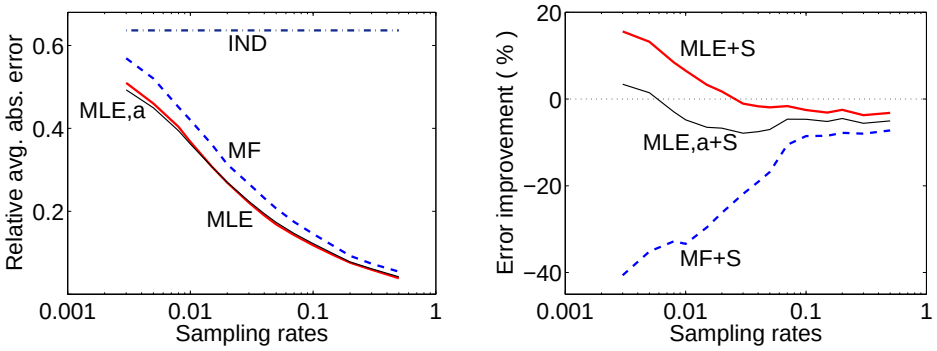
(a) Case 2-1 (b) Case 2-2 (c) Case 2-3

(d) Case 2-4 (e) Case 2-5 (f) Case 2-6

Figure 8.6: Smoothing improves the proposed MLE estimators but hurts the margin-free estimator in most cases. The vertical axis is the percentage of relative improvement in $\sqrt{\text{MSE}}$ of each smoothed estimator with respect to its un-smoothed version.

of a based on a small sample, the rankings will hopefully be close to what we would obtain if we used the entire dataset. This section will compare the rankings based on a small sample to a gold standard, the rankings based on the entire dataset.

How should we evaluate rankings? We follow the suggestion in [146] of reporting the percentage of agreements in the top-S. That is, we compare the top-S pairs based on a sample with the top-S pairs based on the entire dataset. We report the intersection of the two lists, normalized by S. Figure 8.8(a) emphasizes high precision region ($3 \leq S \leq 200$), whereas Figure 8.8(b) emphasizes higher recall, extending S to cover all 468,028 pairs in the large dataset experiment. Of course, agreement rates are high at high sampling rates. For example, we have nearly $\approx 100\%$ agreement at a sampling rate of 0.5. It is reassuring that agreement rates remain fairly high ($\approx 70\%$), even when we push the sampling rate way down (0.003). In other words, we can find many of the most obvious associations with very little work.



(a) (b)

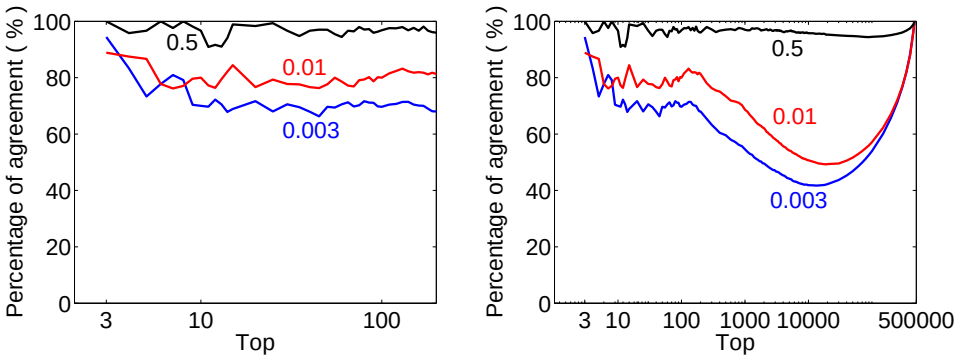
Figure 8.7: (a): The proposed MLE methods (solid lines) have smaller errors than the baselines (dashed lines). We report the mean absolute errors (normalized by the mean co-occurrences, 188), averaged over six permutations. The proposed MLE and the recommended quadratic approximation are very close. Both are well below the margin-free (MF) baseline and the independence (IND) baseline. (b): Percentage of improvement due to smoothing. Smoothing helps MLE, but hurts MF.

The same comparisons can be evaluated in terms of precision and recall, by fixing the top-L_G gold standard list but varying the length of the sample list L_S. More precisely, recall = relevant/L_G, and precision = relevant/L_S, where "relevant" means the retrieved pairs in the gold standard list. Figure 8.9 gives a graphical representation of this evaluation scheme, using notation in [135, Chapter 8.1].

Figure 8.10 presents the precision-recall curves for $L_G = 1\%L$ and $10\%L$, where $L = 468,028$. For each L_G, there is one precision-recall curve corresponding to each sampling rate. All curves indicate the precision-recall trade-off and that the only way to improve both precision and recall simultaneously is to increase the sampling rate.

8.4 Multi-way Associations

Many applications involve multi-way associations, e.g., association rules, databases, and Web search. Fortunately, our sketch construction and estimation algorithm can be naturally extended to multi-way associations. We have already presented such an example in Section 8.1. When we do not consider the margins, the estimation task is



(a) Top 3 - 200 (b) Top 3 - All

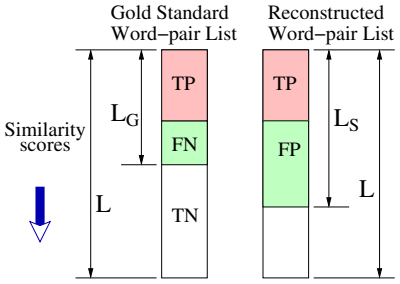
Figure 8.8: We can find many of the most obvious associations with very little work. Two sets of cosine scores were computed for the 468,028 pairs. The gold standard scores were computed over the entire dataset, whereas sample scores were computed over a sample. The plots shows the percentage of agreement between these two lists, as a function of S. As expected, agreement rates are high ($\approx 100\%$) at high sampling rates (0.5). But it is reassuring that agreement rates remain pretty high ($\approx 70\%$) even when we crank the sampling rate way down (0.003).

as simple as in the pair-wise case. When we take advantage of margins, estimating multi-way associations amounts to a convex program.

8.4.1 Multi-way Sketches

Suppose we are interested in the associations among m words, denoted by W_1, W_2, ..., W_m. The document frequencies are f_1, f_2, ..., and f_m. $f_i = |\mathrm{P}_i|$. There are $N = 2^m$ combinations of associations, denoted by $x_1, x_2, ..., x_N$. For example,

$$
\begin{aligned}
a = x_1 &= |\mathrm{P}_1 \cap \mathrm{P}_2 \cap ... \cap \mathrm{P}_{m-1} \cap \mathrm{P}_m|, \\
x_2 &= |\mathrm{P}_1 \cap \mathrm{P}_2 \cap ... \cap \mathrm{P}_{m-1} \cap \neg \mathrm{P}_m|, \\
x_3 &= |\mathrm{P}_1 \cap \mathrm{P}_2 \cap ... \cap \neg \mathrm{P}_{m-1} \cap \mathrm{P}_m|, \\
&... \\
x_{N-1} &= |\neg \mathrm{P}_1 \cap \neg \mathrm{P}_2 \cap ... \cap \neg \mathrm{P}_{m-1} \cap \mathrm{P}_m|, \\
x_N &= |\neg \mathrm{P}_1 \cap \neg \mathrm{P}_2 \cap ... \cap \neg \mathrm{P}_{m-1} \cap \neg \mathrm{P}_m|,
\end{aligned} \tag{8.21}
$$

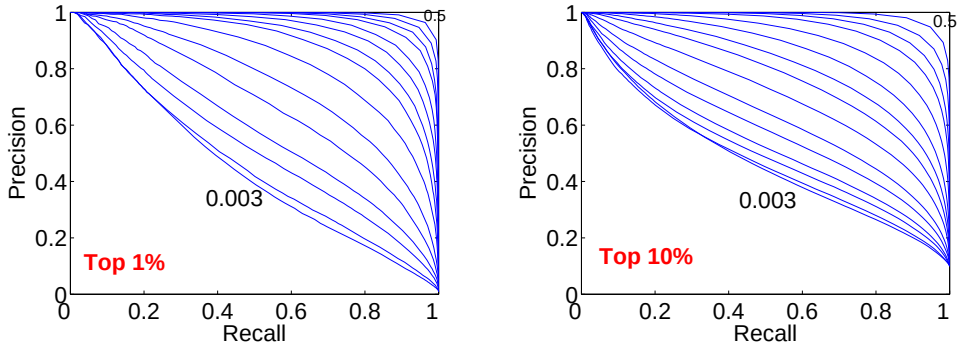


	Relevant	Irrelevant
Retrieved	TP (a)	FP (b)
Not Retrieved	FN (c)	TN (d)

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{a}{a + b}$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{a}{a + c}$$

Figure 8.9: Definitions of recall and precision. L = total number of pairs. L_G = number of pairs from the top of the gold standard similarity list. L_S = number of pairs from the top of the reconstructed similarity list.



(a) Top 1% (b) Top 10%

Figure 8.10: Precision-recall curves in retrieving the top-1% and top-10% pairs, at different sampling rates from 0.003 to 0.5. Note that the precision is $\geq \frac{L_G}{L}$.

which can be directly corresponded to the binary representation of integers.

Using the vector and matrix notation, $\mathbf{X} = [x_1, x_2, ..., x_N]^{\mathrm{T}}$, $\mathbf{F} = [f_1, f_2, ..., f_m, D]^{\mathrm{T}}$, we can write down the margin constraints in terms of a linear matrix equation as

$$\mathbf{AX} = \mathbf{F}, \tag{8.22}$$

where $\mathbf{A}$ is the constraint matrix. If necessary, we can use $\mathbf{A}^{(m)}$ to identify $\mathbf{A}$ for

different m values. For example, when $m = 2$ or $m = 3$,

$$\mathbf{A}^{(2)} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \qquad \mathbf{A}^{(3)} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}. \tag{8.23}$$

For each word W_i, we sample the k_i smallest elements from its permuted inverted index, $\pi(\mathrm{P}_i)$, to form a sketch, K_i. We compute the (conditional) sample size D_s

$$D_s = \min\{\max(\mathrm{K}_1), \max(\mathrm{K}_2), ..., \max(\mathrm{K}_m)\}. \tag{8.24}$$

After removing the elements in all m K_i's that are larger than D_s, we generate the sample table counts. The samples are denoted as $\mathbf{S} = [s_1, s_2, ..., s_N]^{\mathrm{T}}$.

Conditional on D_s, the conditional PMF and log PMF would be

$$\mathbf{Pr}(\mathbf{S}|D_s; \mathbf{X}) = \frac{\binom{x_1}{s_1}\binom{x_2}{s_2}...\binom{x_N}{s_N}}{\binom{D}{D_s}} \propto \prod_{i=1}^{N}\prod_{j=0}^{s_i-1}(x_i - j), \tag{8.25}$$

$$\log \mathbf{Pr}(\mathbf{S}|D_s; \mathbf{X}) \propto Q = \sum_{i=1}^{N}\sum_{j=0}^{s_i-1}\log(x_i - j). \tag{8.26}$$

The log PMF is concave, as in two-associations. A partial likelihood MLE solution, i.e., the $\hat{\mathbf{X}}$ that maximizes $\log \mathbf{Pr}(\mathbf{S}|D_s; \hat{\mathbf{X}})$, will again be adopted, which leads to a convex optimization problem. But first, we shall discuss two baseline estimators.

8.4.2 Baseline Independence Estimator

Assuming independence, an estimator of x_1 would be

$$\hat{x}_{1,IND} = D\prod_{i=1}^{m}\frac{f_i}{D}, \tag{8.27}$$

which can be easily proved using a conditional expectation argument.

By property of the hypergeometric distribution, $\mathrm{E}(|\mathrm{P}_i \cap \mathrm{P}_j|) = \frac{f_i f_j}{D}$. Therefore,

$$\begin{aligned}\mathrm{E}(x_1) &= \mathrm{E}(|\mathrm{P}_1 \cap \mathrm{P}_2| \cap ... \cap \mathrm{P}_m|) = \mathrm{E}(|\cap_{i=1}^m \mathrm{P}_i|) \\ &= \mathrm{E}(\mathrm{E}(|\mathrm{P}_1 \cap (\cap_{i=2}^m \mathrm{P}_i) \,||\, (\cap_{i=2}^m \mathrm{P}_i))) = \frac{f_1}{D}\mathrm{E}(|\cap_{i=2}^m \mathrm{P}_i|) \\ &= \frac{f_1 f_2 ... f_{m-2}}{D^{m-2}}\mathrm{E}(|\mathrm{P}_{m-1} \cap \mathrm{P}_m|) = D\prod_{i=1}^m \frac{f_i}{D}. \end{aligned} \tag{8.28}$$

8.4.3 Baseline Margin-free Estimator

The conditional PMF $\mathbf{Pr}(\mathbf{S}|D_s;\mathbf{X})$ is a multivariate hypergeometric distribution, based on which we can derive the margin-free estimator:

$$\mathrm{E}(s_i|D_s) = \frac{D_s}{D}x_i, \quad \hat{x}_{i,MF} = \frac{D}{D_s}s_i, \quad \mathrm{Var}(\hat{x}_{i,MF}|D_s) = \frac{D}{D_s}\frac{1}{\frac{1}{x_i}+\frac{1}{D-x_i}}\frac{D-D_s}{D-1}, \tag{8.29}$$

i.e., the margin-free estimator remains its simplicity in the multi-way case.

8.4.4 The MLE

The exact MLE can be formulated as a standard convex program,

$$\begin{aligned} \text{minimize} \quad & -Q = -\sum_{i=1}^N \sum_{j=0}^{s_i-1} \log(x_i - j), \\ \text{subject to} \quad & \mathbf{AX} = \mathbf{F}, \ \text{ and } \ \mathbf{X} \succeq \mathbf{S}, \end{aligned} \tag{8.30}$$

where $\mathbf{X} \succeq \mathbf{S}$ is a compact representation for $x_i \geq s_i, 1 \leq i \leq N$.

This optimization problem can be solved by a variety of standard methods such as Newton's method[27, Chapter 10.2]. Note that we can ignore the implicit inequality constraints, $\mathbf{X} \succeq \mathbf{S}$, if we start with a feasible initial guess.

It turns out that the formulation in (8.30) will encounter numerical difficulty due to the inner summation in the objective function Q. Smoothing will bring in more numerical issues. Recall in estimating two-way associations we do not have this problem, because we have eliminated the inner summation in the objective function,

using an (integer) updating formula. In multi-way associations, it seems not easy to reformulate the objective function Q in a similar form.

To avoid the numerical problems, a simple solution is to assume "sample-with-replacement," under which the conditional likelihood and log likelihood become

$$\mathbf{Pr}(\mathbf{S}|D_s; \mathbf{X}, r) \propto \prod_{i=1}^{N} \left(\frac{x_i}{D}\right)^{s_i} \propto \prod_{i=1}^{N} x_i^{s_i}, \tag{8.31}$$

$$\log \mathbf{Pr}(\mathbf{S}|D_s; \mathbf{X}, r) \propto Q_r = \sum_{i=1}^{N} s_i \log x_i. \tag{8.32}$$

Our MLE problem can then be reformulated as a convex program

$$\begin{aligned} \text{minimize} \quad & -Q = -\sum_{i=1}^{N} s_i \log x_i, \\ \text{subject to} \quad & \mathbf{AX} = \mathbf{F}, \quad \text{and} \quad \mathbf{X} \succeq \mathbf{S}, \end{aligned} \tag{8.33}$$

where, to simplify the notation, we neglect the subscript "r."

We can compute the gradient ($\bigtriangledown Q$) and Hessian ($\bigtriangledown^2 Q$). The gradient is a vector of the first derivatives of Q with respect to x_i, for $1 \leq i \leq N$,

$$\bigtriangledown Q = \left[\frac{\partial Q}{\partial x_i}, 1 \leq i \leq N\right] = \left[\frac{s_1}{x_1}, \frac{s_2}{x_2}, ..., \frac{s_N}{x_N}\right]^{\mathrm{T}}. \tag{8.34}$$

The Hessian is a matrix whose $(i,j)^{th}$ entry is the partial derivative $\frac{\partial^2 Q}{\partial x_i x_j}$, i.e.,

$$\bigtriangledown^2 Q = -\mathrm{diag}\left[\frac{s_1}{x_1^2}, \frac{s_2}{x_2^2}, ..., \frac{s_N}{x_N^2}\right]. \tag{8.35}$$

8.4.5 The Covariance Matrix

We apply the large sample theory to estimate the covariance matrix of the MLE. Recall that we have $N = 2^m$ variables and $m + 1$ constraints. The effective number of variables would be $2^m - (m + 1)$, which is the dimension of the covariance matrix.

We seek a partition of $\mathbf{A} = [\mathbf{A_1}, \mathbf{A_2}]$, such that $\mathbf{A_2}$ is invertible. We may have to

switch some columns of $\mathbf{A}$ in order to find an invertible $\mathbf{A_2}$. In our construction, the jth column of $\mathbf{A}_2$ is the column of $\mathbf{A}$ such that last entry of the jth row of $\mathbf{A}$ is 1. An example for $m = 3$ would be

$$\mathbf{A_1^{(3)}} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad \mathbf{A_2^{(3)}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \tag{8.36}$$

where $\mathbf{A_1^{(3)}}$ is the [1 2 3 5] columns of $\mathbf{A^{(3)}}$ and $\mathbf{A_2^{(3)}}$ is the [4 6 7 8] columns of $\mathbf{A^{(3)}}$. $\mathbf{A}_2$ constructed this way is always invertible because its determinant is always one.

Corresponding to the partition of $\mathbf{A}$, we partition $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]^{\mathrm{T}}$. For example, when $m = 3$, $\mathbf{X}_1 = [x_1, x_2, x_3, x_5]^{\mathrm{T}}$, $\mathbf{X}_2 = [x_4, x_6, x_7, x_8]^{\mathrm{T}}$. $\mathbf{X}_2$ can be expressed as

$$\mathbf{X}_2 = \mathbf{A}_2^{-1} \left(\mathbf{F} - \mathbf{A}_1 \mathbf{X_1} \right) = \mathbf{A}_2^{-1} \mathbf{F} - \mathbf{A}_2^{-1} \mathbf{A}_1 \mathbf{X_1}. \tag{8.37}$$

The log likelihood function Q, which is separable, can then be expressed as

$$Q(\mathbf{X}) = Q_1(\mathbf{X}_1) + Q_2(\mathbf{X}_2). \tag{8.38}$$

By the matrix derivative chain rule, the Hessian of Q with respect to $\mathbf{X}_1$ would be

$$\bigtriangledown_1^2 Q = \bigtriangledown_1^2 Q_1 + \bigtriangledown_1^2 Q_2 = \bigtriangledown_1^2 Q_1 + \left(\mathbf{A}_2^{-1} \mathbf{A}_1 \right)^{\mathrm{T}} \bigtriangledown_2^2 Q_2 \left(\mathbf{A}_2^{-1} \mathbf{A}_1 \right), \tag{8.39}$$

where $\bigtriangledown_1^2$ and $\bigtriangledown_2^2$ indicate the Hessians are with respect to $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively.

Conditional on D_s, the Expected Fisher Information of $\mathbf{X}_1$ is

$$\mathrm{I}(\mathbf{X}_1) = \mathrm{E}\left(- \bigtriangledown_1^2 Q | D_s \right) = -\mathrm{E}(\bigtriangledown_1^2 Q_1 | D_s) - \left(\mathbf{A}_2^{-1} \mathbf{A}_1 \right)^{\mathrm{T}} \mathrm{E}(\bigtriangledown_2^2 Q_2 | D_s) \left(\mathbf{A}_2^{-1} \mathbf{A}_1 \right), \tag{8.40}$$

where

$$\mathrm{E}(-\bigtriangledown_1^2 Q_1|D_s) = \mathrm{diag}\left[\mathrm{E}\left(\frac{s_i}{x_i^2}\right), x_i \in \mathbf{X}_1\right] = \frac{D_s}{D}\mathrm{diag}\left[\frac{1}{x_i}, x_i \in \mathbf{X}_1\right], \tag{8.41}$$

$$\mathrm{E}(-\bigtriangledown_2^2 Q_2|D_s) = \frac{D_s}{D}\mathrm{diag}\left[\frac{1}{x_i}, x_i \in \mathbf{X}_2\right]. \tag{8.42}$$

By the large sample theory, and also considering the finite population correction factor, we can approximate the (conditional) covariance matrix of $\mathbf{X}_1$ to be

$$\mathrm{Cov}(\mathbf{X}_1|D_s) \approx \mathrm{I}(\mathbf{X}_1)^{-1}\left(1 - \frac{D_s}{D}\right) = \left(\frac{D}{D_s} - 1\right) \times \left(\mathrm{diag}\left[\frac{1}{x_i}, x_i \in \mathbf{X}_1\right] + \left(\mathbf{A}_2^{-1}\mathbf{A}_1\right)^{\mathrm{T}} \mathrm{diag}\left[\frac{1}{x_i}, x_i \in \mathbf{X}_2\right]\left(\mathbf{A}_2^{-1}\mathbf{A}_1\right)\right)^{-1}. \tag{8.43}$$

For a sanity check, we verify that this approach recovers the same variance formula in the two-way association case. Recall that, when $m = 2$, we have

$$\bigtriangledown^2 Q = -\begin{bmatrix} \frac{s_1}{x_1^2} & 0 & 0 & 0 \\ 0 & \frac{s_2}{x_2^2} & 0 & 0 \\ 0 & 0 & \frac{s_3}{x_3^2} & 0 \\ 0 & 0 & 0 & \frac{s_4}{x_4^2} \end{bmatrix}, \ \bigtriangledown_1^2 Q_1 = -\frac{s_1}{x_1^2}, \ \bigtriangledown_2^2 Q_2 = -\begin{bmatrix} \frac{s_2}{x_2^2} & 0 & 0 \\ 0 & \frac{s_3}{x_3^2} & 0 \\ 0 & 0 & \frac{s_4}{x_4^2} \end{bmatrix}$$

$$\mathbf{A^{(2)}} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \ \mathbf{A_1^{(2)}} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \ \mathbf{A_2^{(2)}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

$$\left(\mathbf{A}_2^{-1}\mathbf{A}_1\right)^{\mathrm{T}} \bigtriangledown_2^2 Q_2 \mathbf{A}_2^{-1}\mathbf{A}_1 = -\begin{bmatrix} 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{s_2}{x_2^2} & 0 & 0 \\ 0 & \frac{s_3}{x_3^2} & 0 \\ 0 & 0 & \frac{s_4}{x_4^2} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} = -\frac{s_2}{x_2^2} - \frac{s_3}{x_3^2} - \frac{s_4}{x_4^2}.$$

Hence,

$$-\bigtriangledown_1^2 Q = \frac{s_1}{x_1^2} + \frac{s_2}{x_2^2} + \frac{s_3}{x_3^2} + \frac{s_4}{x_4^2} = \frac{a_s}{a^2} + \frac{b_s}{(f_1 - a)^2} + \frac{c_s}{(f_2 - a)^2} + \frac{d_s}{(D - f_1 - f_2 + a)^2},$$

which leads to the same Fisher Information for the two-way case as we have derived.

Similar to two-way associations, the unconditional variance of the proposed MLE can

be estimated by replacing $\frac{D}{D_s}$ in (8.43) with $\mathrm{E}\left(\frac{D}{D_s}\right)$, i.e.,

$$\mathrm{Cov}(\mathbf{X}_1) \approx \left(\mathrm{E}\left(\frac{D}{D_s}\right) - 1\right) \times \left(\mathrm{diag}\left[\frac{1}{x_i}, x_i \in \mathbf{X}_1\right] + \left(\mathbf{A}_2^{-1}\mathbf{A}_1\right)^{\mathrm{T}} \mathrm{diag}\left[\frac{1}{x_i}, x_i \in \mathbf{X}_2\right] \left(\mathbf{A}_2^{-1}\mathbf{A}_1\right)\right)^{-1}. \quad (8.44)$$

8.4.6 Empirical Evaluation

We use the same four words as in Table 8.1 to evaluate the multi-way association algorithm, as merely a sanity check. There are four different combinations of three-way associations and one four-way associations, as numbered in Table 8.2.

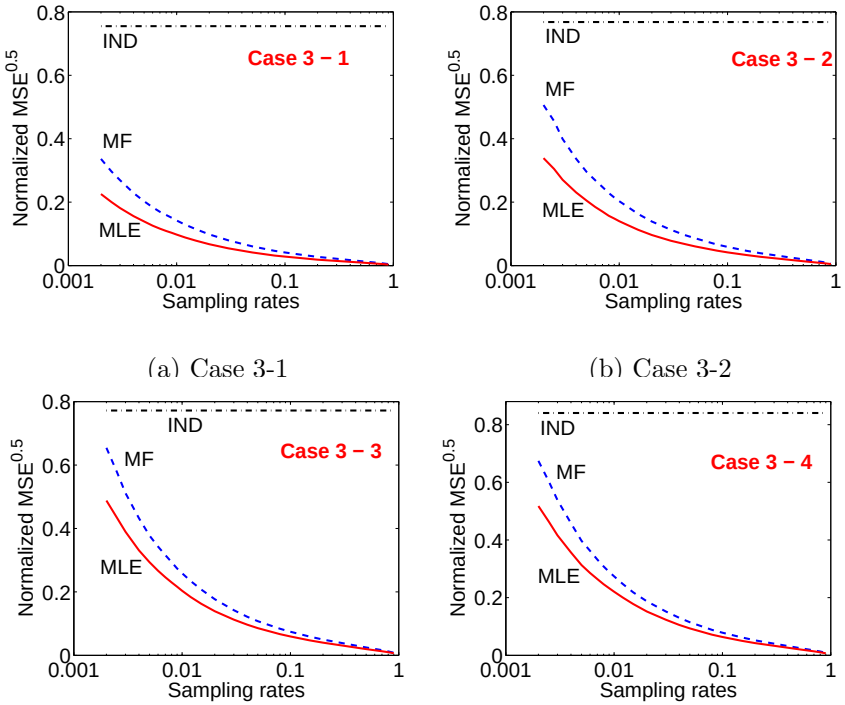
Table 8.2: The same four words as in Table 8.1 are used for evaluating multi-way associations. There are four three-way combinations and one four-way combination.

	Case #	Words	Co-occurrences
Three-way	Case 3-1	THIS & HAVE & HELP	4940
	Case 3-2	THIS & HAVE & PROGRAM	2575
	Case 3-3	THIS & HELP & PROGRAM	1626
	Case 3-4	HAVE & HELP & PROGRAM	1460
Four-way	Case 4	THIS & HAVE & HELP & PROGRAM	1316

We present results for x_1 (i.e., a in two-way associations) for all cases. The evaluations for four three-way cases are presented in Figures 8.11, and 8.12. From these figures, we see that the proposed MLE has lower MSE than the margin-free baseline (MF). As in the two-way case, smoothing helps MLE but still hurts MF in most cases.

Figure 8.13 presents the evaluation results for the four-way association case. The results are similar to the three-way case.

Combining the results of two-way associations for the same four words, we can study the trend how the proposed MLE improve the MF baseline. Figure 8.14(a) suggests that the proposed MLE is a big improvement over the MF baseline for two-way associations, but the improvement becomes less noticeable with higher order associations. This observation is not surprising, because the number of degrees of freedom, $2^m - (m+1)$, increases exponentially with m.



(a) Case 3-1 (b) Case 3-2

(c) Case 3-3 (d) Case 3-4

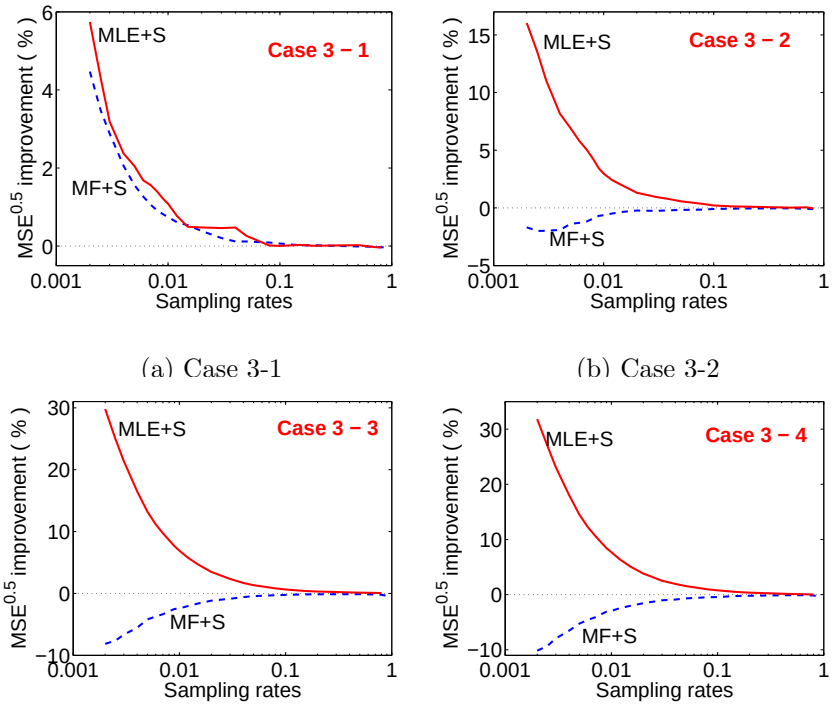
Figure 8.11: In terms of $\frac{\sqrt{\mathrm{MSE}(x_1)}}{x_1}$, the proposed MLE is consistently better than the margin-free baseline (MF), which is better than the independence baseline (IND), for four three-way association cases.

On the other hand, smoothing becomes more important as m increases, as shown in Figure 8.14(b), partly because of the data sparsity in high-order associations.

8.5 Comparison with Broder's Sketches

Broder's Sketches[34], originally introduced for removing duplicates in the AltaVista index, have been applied to a variety of applications[38, 89, 90]. [36, 37] presented some theoretical aspects of the sketch algorithm. There has been considerable exciting work following up on this line of research including[98, 43, 101].

Broder and his colleagues introduced two algorithms, which we refer to as the "original sketch" and the "minwise sketch" for estimating resemblance, $R = \frac{|\mathrm{P}_1 \cap \mathrm{P}_2|}{|\mathrm{P}_1 \cup \mathrm{P}_2|}$. The original



(a) Case 3-1

(b) Case 3-2

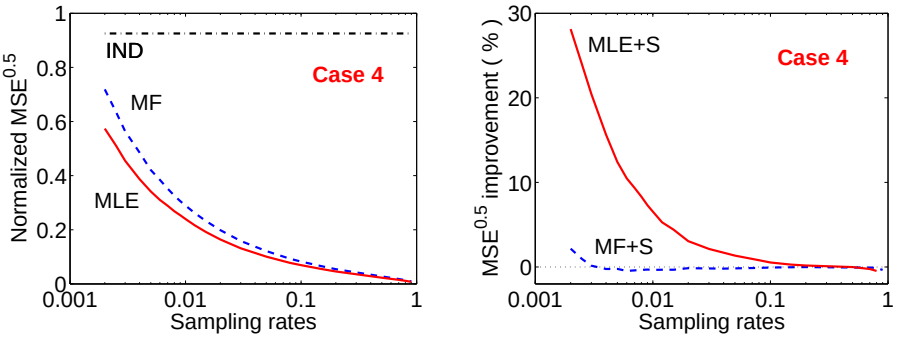
(c) Case 3-3

(d) Case 3-4

Figure 8.12: The simple "add-one" smoothing improves the accuracies for the MLE. Smoothing, however, in all cases except Case 3-1 hurts the margin-free estimator.

sketch uses a single random permutation, while the minwise sketch uses k permutations. Both algorithms have similar estimation accuracies, as will see.

Our proposed sketch algorithm is closer to Broder's original sketch, with a few important differences. One difference is that Broder's original sketch throws out about half of the sample, whereas we throw out less. In addition, the sketch sizes are fixed over all words for Broder, whereas we allow different sizes for different words. Broder's method was designed for pair-wise associations, whereas our method generalizes to multi-way. Finally, Broder's method was designed for boolean data, whereas our method generalizes to reals.



(a) MSE (b) Smoothing

Figure 8.13: Four-way associations (Case 4). (a): The MLE has smaller MSE than the margin-free (MF) baseline, which has smaller MSE than the independence baseline. (b): Smoothing considerably improve the accuracy for MLE.

8.5.1 Broder's Minwise Sketch

Suppose a random permutation π_1 is performed on the document IDs. We denote the smallest IDs in P_1 and P_2, by $\min(\pi_1(\mathrm{P}_1))$ and $\min(\pi_1(\mathrm{P}_2))$, respectively. Obviously,

$$\mathbf{Pr}\left(\min(\pi_1(\mathrm{P}_1)) = \min(\pi_1(\mathrm{P}_2))\right) = \frac{|\mathrm{P}_1 \cap \mathrm{P}_2|}{|\mathrm{P}_1 \cup \mathrm{P}_2|} = R. \tag{8.45}$$

After k minwise independent permutations, denoted as π_1, π_2, ..., π_k, we can estimate R without bias, as a binomial probability, i.e.,

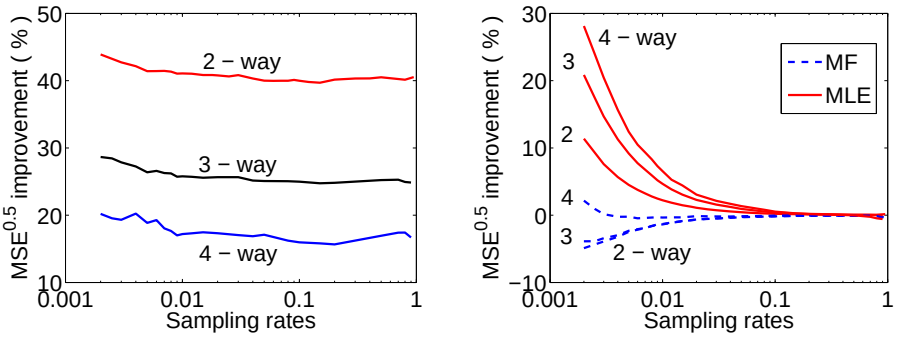
$$\hat{R}_{B,r} = \frac{1}{k}\sum_{i=1}^{k}\{\min(\pi_i(\mathrm{P}_1)) = \min(\pi_i(\mathrm{P}_2))\}, \qquad \mathrm{Var}\left(\hat{R}_{B,r}\right) = \frac{1}{k}R(1-R). \tag{8.46}$$

8.5.2 Broder's Original Sketch

A single random permutation π is applied to the document IDs. Two sketches are constructed: $\mathrm{K}_1 = \mathrm{MIN}_{k_1}(\pi(\mathrm{P}_1))$, $\mathrm{K}_2 = \mathrm{MIN}_{k_2}(\pi(\mathrm{P}_2))$.[3] [34] proposed an unbiased estimator for the resemblance:

$$\hat{R}_B = \frac{|\mathrm{MIN}_k(\mathrm{K}_1 \cup \mathrm{K}_2) \cap \mathrm{K}_1 \cap \mathrm{K}_2|}{|\mathrm{MIN}_k(\mathrm{K}_1 \cup \mathrm{K}_2)|}. \tag{8.47}$$

[3]Broder required fix sketch sizes: $k_1 = k_2 = k$, a restriction that we find convenient to relax.



(a) (b)

Figure 8.14: (a): Combining the three-way, four-way, and the two-way association results for the four words in the evaluations, the average relative improvements of $\sqrt{\text{MSE}}$ suggests that the proposed MLE is consistently better the MF baseline but the improvement decreases monotonically as the order of associations increases. (b): Average $\sqrt{\text{MSE}}$ improvements due to smoothing imply that smoothing becomes more and more important as the order of association increases.

Note that intersecting by $\text{MIN}_k(\text{K}_1 \cup \text{K}_2)$ throws out about half the samples, which is undesirable (and unnecessary).

The following explanation for (8.47) is slightly different from [34]. We can divide the set $\text{P}_1 \cup \text{P}_2$ (of size $a+b+c = f_1+f_2-a$) into two disjoint sets: $\text{P}_1 \cap \text{P}_2$ and $\text{P}_1 \cup \text{P}_2 - \text{P}_1 \cap \text{P}_2$. Within the set $\text{MIN}_k(\text{K}_1 \cup \text{K}_2)$ (of size k), the document IDs that belong to $\text{P}_1 \cap \text{P}_2$ would be $\text{MIN}_k(\text{K}_1 \cup \text{K}_2) \cap \text{K}_1 \cap \text{K}_2$, whose size is denoted by a_s^{B}. This way, we have a hypergeometric sample, i.e., we sample k document IDs from $\text{P}_1 \cup \text{P}_2$ randomly without replacement and obtain a_s^{B} IDs that belong to $\text{P}_1 \cap \text{P}_2$. By property of the hypergeometric distribution, the expectation of a_s^{B} would be

$$\text{E}\left(a_s^{\text{B}}\right) = \frac{ak}{f_1+f_2-a} \Longrightarrow \text{E}\left(\frac{a_s^{\text{B}}}{k}\right) = \frac{a}{f_1+f_2-a} = \frac{|\text{P}_1 \cap \text{P}_2|}{|\text{P}_1 \cup \text{P}_2|} \Longrightarrow \text{E}(\hat{R}_B) = R.$$

The variance of $\hat{R}_B$, according to the hypergeometric distribution, is:

$$\text{Var}\left(\hat{R}_B\right) = \frac{1}{k} R(1-R) \frac{f_1+f_2-a-k}{f_1+f_2-a-1}, \tag{8.48}$$

where the term $\frac{f_1+f_2-a-k}{f_1+f_2-a-1}$ is the "finite population correction factor."

The minwise sketch can be considered as a "sample-with-replacement" variate of the original sketch. The analysis of minwise sketch is slightly simpler mathematically while the original sketch is more efficient. The original sketch requires only one random permutation and has slightly smaller variance than the minwise sketch, i.e., $\text{Var}\left(\hat{R}_{B,r}\right) \geq \text{Var}\left(\hat{R}_B\right)$. When k is reasonably small, as is common in practice, two sketch algorithms have similar errors.

8.5.3 Why Our Algorithm Improves Broders's Sketch

Our proposed sketch algorithm starts with Broder's original (one permutation) sketch; but our estimation method differs in two important aspects.

Firstly, Broder's estimator (8.47) uses k out of $2 \times k$ samples. In particular, it uses only $a_s^{\text{B}} = |\text{MIN}_k(\text{K}_1 \cup \text{K}_2) \cap \text{K}_1 \cap \text{K}_2|$ intersections, which is always smaller than $a_s = |\text{K}_1 \cap \text{K}_2|$ available in the samples. In contrast, our algorithm takes advantage of all useful samples up to $D_s = \min(\max(\text{K}_1), \max(\text{K}_2))$, particularly all a_s intersections. If $\frac{k_1}{f_1} = \frac{k_2}{f_2}$, i.e., if we sample proportionally to the margins:

$$k_1 = 2k\frac{f_1}{f_1 + f_2}, \qquad k_2 = 2k\frac{f_2}{f_1 + f_2}, \tag{8.49}$$

it is expected that almost all samples will be utilized.

Secondly, Broder's estimator (8.47) considers a two-cell hypergeometric model (a, $b+c$) while the two-way association is a four-cell model (a, b, c, d). Simpler data model often results in simpler estimation method but with larger errors.

8.5.4 Comparison of Variances

Broder's method was designed to estimate resemblance. Thus, this section will compare the proposed method with Broder's sketches in terms of resemblance, R.

We can compute R from our estimated association $\hat{a}_{MLE}$:

$$\hat{R}_{MLE} = \frac{\hat{a}_{MLE}}{f_1 + f_2 - \hat{a}_{MLE}}. \tag{8.50}$$

$\hat{R}_{MLE}$ is slightly biased. However, since the second derivative $R''(a)$

$$R''(a) = \frac{2(f_1+f_2)}{(f_1+f_2-a)^3} \leq \frac{2(f_1+f_2)}{\max(f_1,f_2)^3} \leq \frac{4}{\max(f_1,f_2)^2}, \tag{8.51}$$

is small (i.e., the nonlinearity is weak), it is unlikely that the bias will be noticeable. By the delta method, the variance of $\hat{R}_{MLE}$ is approximately:

$$\mathrm{Var}\left(\hat{R}_{MLE}\right) \approx \frac{\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right)}{\frac{1}{a}+\frac{1}{f_1-a}+\frac{1}{f_2-a}+\frac{1}{D-f_1-f_2+a}} \frac{(f_1+f_2)^2}{(f_1+f_2-a)^4}, \tag{8.52}$$

conservatively ignoring the "finite population correction factor," for convenience.

Define the ratio of the variances to be $V_B = \frac{\mathrm{Var}\left(\hat{R}_{MLE}\right)}{\mathrm{Var}\left(\hat{R}_B\right)}$, then

$$V_B = \frac{\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right)}{\frac{1}{a}+\frac{1}{f_1-a}+\frac{1}{f_2-a}+\frac{1}{D-f_1-f_2+a}} \frac{(f_1+f_2)^2}{(f_1+f_2-a)^2} \frac{k}{a(f_1+f_2-2a)}. \tag{8.53}$$

To help our intuitions, let us consider some reasonable simplifications to V_B. Assuming $a << \min(f_1,f_2) < \max(f_1,f_2) << D$, then approximately

$$V_B \approx \frac{k\max(\frac{f_1}{k_1}, \frac{f_2}{k_2})}{f_1+f_2} = \begin{cases} \frac{\max(f_1,f_2)}{f_1+f_2} & \text{if } \; k_1=k_2=k \\ \\ \frac{1}{2} & \text{if } k_1 = 2k\frac{f_1}{f_1+f_2}, \;\; k_2 = 2k\frac{f_2}{f_1+f_2} \end{cases}, \tag{8.54}$$

which indicates that the proposed method is a considerable improvement over Broder's sketch. To achieve the same accuracy, our method requires only half as many samples.

Figure 8.15 plots the V_B in (8.53) for the whole range of f_1, f_2, and a, assuming equal samples: $k_1 = k_2 = k$. We can see that $V_B \leq 1$ always holds and $V_B = 1$ only when $f_1 = f_2 = a$. There is also the possibility that V_B is close to zero.

Proportional samples further reduce V_B, as shown in Figure 8.16.

We show algebraically that in (8.53) $V_B < 1$, unless $f_1 = f_2 = a$. For convenience, we use the notion a, b, c, d in (8.53). Assuming $k_1 = k_2 = k$ and $f_1 > f_2$, we obtain

$$V_B = \frac{a+b}{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \frac{(2a+b+c)^2}{(a+b+c)^2} \frac{1}{a(b+c)}. \tag{8.55}$$

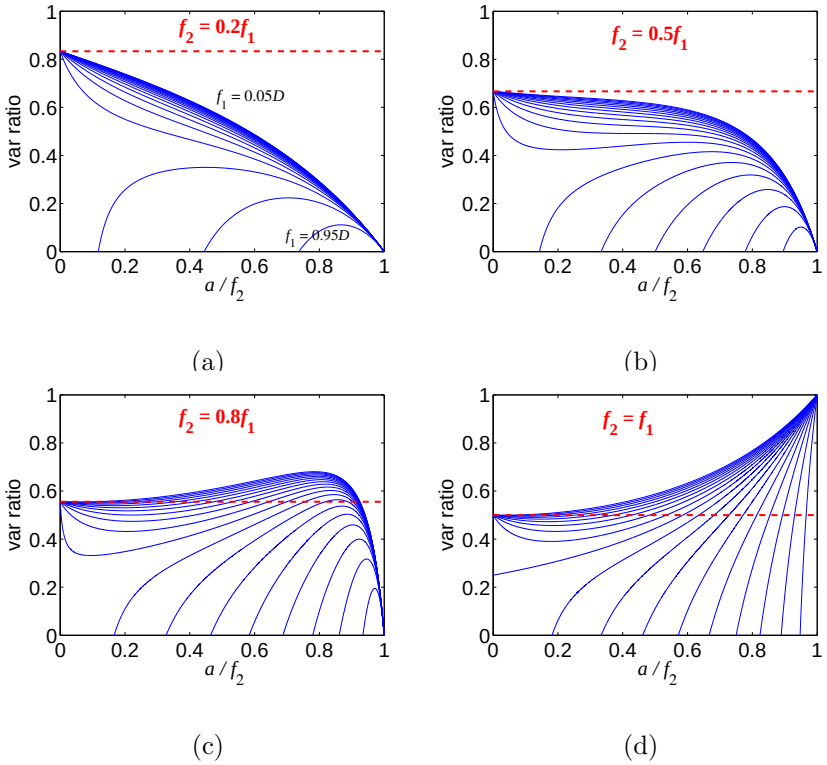


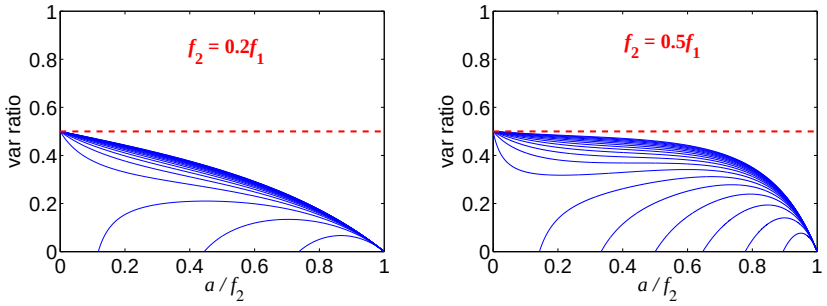
Figure 8.15: We plot V_B in (8.53) for the whole range of f_1, f_2, and a, assuming equal samples: $k_1 = k_2 = k$. (a), (b), (c) and (d) correspond to $f_2 = 0.2f_1$, $f_2 = 0.5f_1$, $f_2 = 0.8f_1$ and $f_2 = f_1$, respectively. Different curves are for different f_1's, ranging from $0.05D$ to $0.95D$ spaced at $0.05D$. The horizontal lines are $\frac{\max(f_1,f_2)}{f_1+f_2}$. We can see that for all cases, $V_B \leq 1$ holds. $V_B = 1$ when $f_1 = f_2 = a$, a trivial case. When a/f_2 is small, $V_B \approx \frac{\max(f_1,f_2)}{f_1+f_2}$ holds well. It is also possible that V_B is very close to zero.

To show $V_B \leq 1$, it suffices to show

$$(a+b)(2a+b+c)^2bcd \leq (bcd+acd+abd+abc)(a+b+c)^2(b+c), \tag{8.56}$$

which is equivalent to the following true statement (note that a, b, c, and d are positive):

$$\begin{aligned}&\left(a^3(b-c)^2+bc^2(b+c)^2+a^2(2b+c)(b^2-bc+2c^2)+a(b+c)(b^3+4bc^2+c^2)\right)d\\&+abc(b+c)(a+b+c)^2 \geq 0.\end{aligned} \tag{8.57}$$



(a) (b)

Figure 8.16: Compared with Figure 8.15, proportional samples further reduce V_B. Note that (a) and (b) are "squashed" versions of (a) and (b), respectively, in Figure 8.15.

8.5.5 Empirical Evaluations

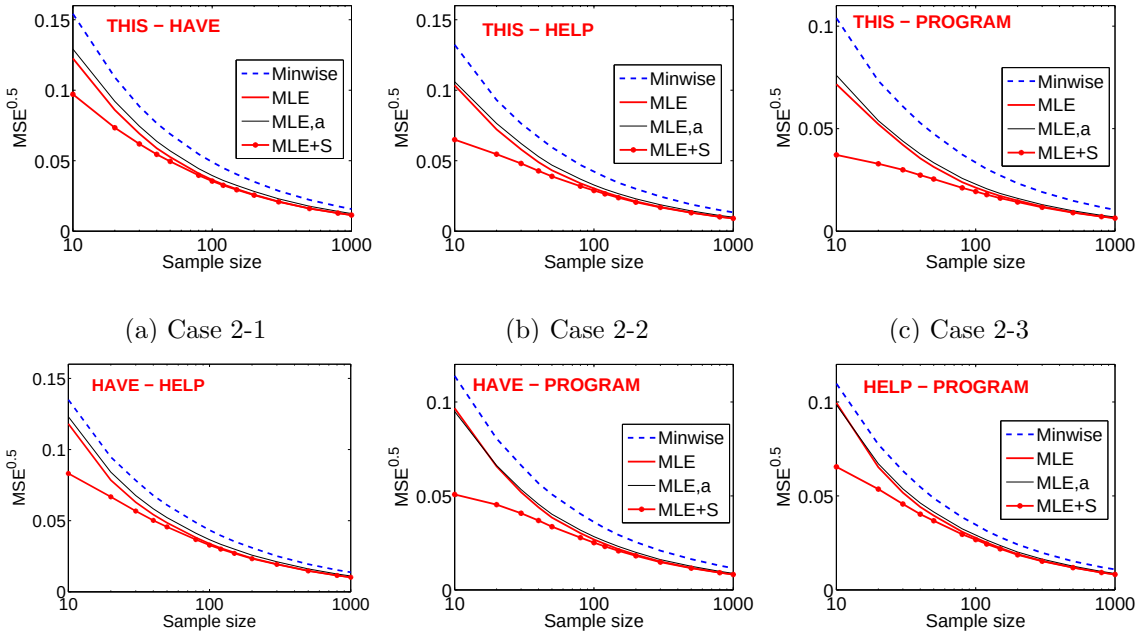
We have theoretically shown that our proposed method is a considerable improvement over Broder's sketch. Next, we would like to evaluate these theoretical results using the same experiment data as in evaluating two-way associations (i.e., Table 8.1).

Figure 8.17 compares the MSE. Here we assume equal samples and later we will show that proportional samples further improves the results. The figure shows that the MLE is consistently better than Broder's sketch. In addition, the approximate MLE $\hat{a}_{MLE,a}$ still gives very close answers to the exact MLE, and the simple "add-one" smoothing improves the estimations at low sampling rates, quite substantially.

Finally, in Figure 8.18, we show that with proportional samples, our algorithm further improves the estimates. In terms of MSE. With equal samples, our estimators improve Broder's sketch by $30\% - 50\%$. With proportional samples, improvements become $40\% - 80\%$. Note that the maximum possible improvement is 100%.

8.6 Summary

The proposed method generates random sample contingency tables directly from the sketch, the front of the inverted index. Because the term-by-document matrix is extremely sparse, it is possible for a relatively small sketch, k, to characterize a large sample of D_s documents. The front of the inverted index not only tells us about the presence of the word in the



(a) Case 2-1 (b) Case 2-2 (c) Case 2-3

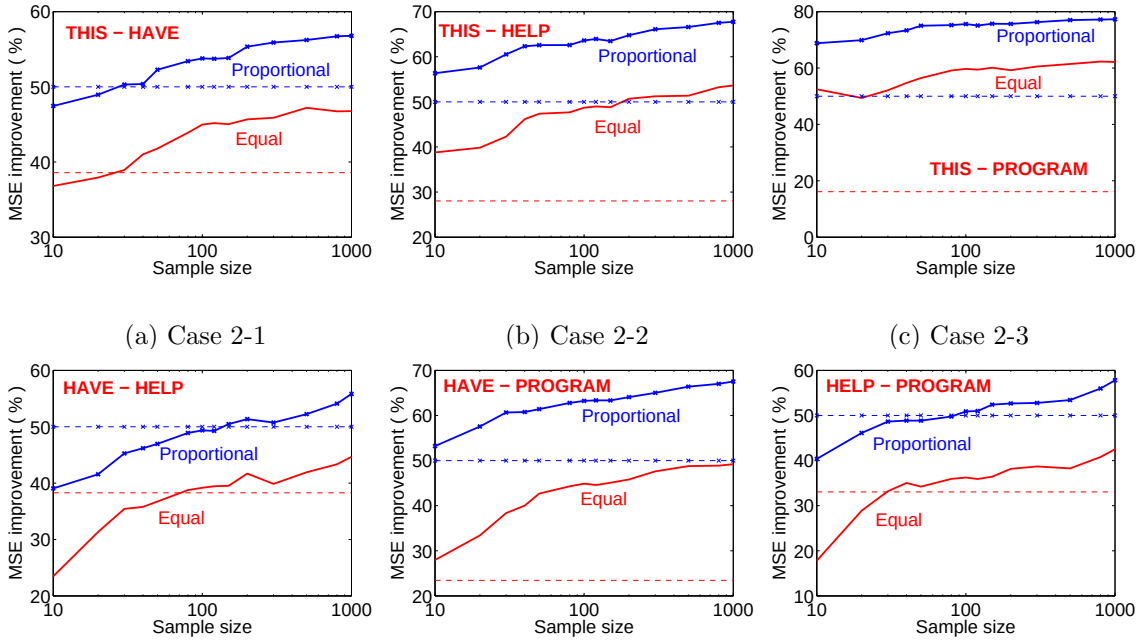
(d) Case 2-4 (e) Case 2-5 (f) Case 2-6

Figure 8.17: When estimating the resemblance, our algorithm gives consistently more accurate answers than Broder's sketch. In our experiments, Broder's "minwise" construction gives almost the same answers as the "original" sketch, thus only the "minwise" results are presented here. The approximate MLE again gives very close answers to the exact MLE. Also, smoothing improves at low sampling rates.

first k documents, but also in the remaining $D_s - k$ documents. This observation becomes increasingly important with larger web collections (with ever increasing sparsity). Typically, $D_s \gg k$.

To estimate the contingency table for the entire population, one can use the "margin-free" baseline, which simply multiplies the sample contingency table by the appropriate scaling factor. However, we recommend taking advantage of the margins. The maximum likelihood solution under margin constraints is a cubic equation, which has a remarkably accurate quadratic approximation. The proposed MLE methods were compared empirically and theoretically to the margin-free (MF) baseline, finding large improvements. When we know the margins, we ought to use them.

Our proposed method differs from Broder's sketches in important aspects. (1) Our sketch construction allows more flexibility in that the sketch size can be different from



(a) Case 2-1 (b) Case 2-2 (c) Case 2-3

(d) Case 2-4 (e) Case 2-5 (f) Case 2-6

Figure 8.18: Compared with Broder's sketch, the relative MSE improvement should be, approximately, $\frac{\min(f_1,f_2)}{f_1+f_2}$ with equal samples, and $\frac{1}{2}$ with proportional samples, i.e., the two horizontal lines. The actual improvements could be lower or higher. The figure verifies that proportional samples can considerably improve the accuracies.

one word to the next. (2) Our estimation is more accurate. The estimator in Broder's sketches uses one half of the samples while our method always uses more. More samples lead to smaller errors. (3) Broder's method considers a two-cell model while our method works with a more refined (hence more accurate) four-cell contingency table model. (4) Our method extends naturally to estimating multi-way associations. (5) Our method extends naturally to real-valued data, as elaborated in Chapter 7.

Chapter 9

Conclusions

The ubiquitous phenomenon of massive datasets in modern applications has brought many interesting and challenging problems to scientists and engineers. Many sampling techniques have been developed for (1) storing massive data in compact representations using small (memory) space, (2) approximating original summary statistics (such as inner products and distances) from the compact representations.

The conventional "simple random sampling" often does not perform well in massive data mainly for two reasons. Firstly, simple random sampling is often not accurate, unless we use a very large sample, because large-scale datasets are often heavy-tailed. Secondly, large-scale datasets are often sparse (e.g., text data), but simple random sampling will miss most of the informative (non-zero) elements. To overcome these challenges, two sampling techniques are elaborated in this thesis, *stable random projections* and *Conditional Random Sampling (CRS)*.

The method of *stable random projections* multiplies the original data matrix $\mathbf{A} \in \mathbb{R}^{n \times D}$ by a random matrix $\mathbf{R} \in \mathbb{R}^{D \times k}$ ($k \ll D$), resulting in $\mathbf{B} = \mathbf{A} \times \mathbf{R} \in \mathbb{R}^{n \times k}$. The projection matrix $\mathbf{R}$ is typically sampled i.i.d. from a symmetric α-stable distribution ($0 < \alpha \leq 2$). The projected matrix $\mathbf{B}$ contains enough information to estimate the original l_α distances in $\mathbf{A}$. Because many statistical and learning applications only require the distances, we can often "throw away" the original data, $\mathbf{A}$.

In the l_2 case, the advantage of *stable random projections* is highlighted by the celebrated Johnson-Lindenstrauss (JL) Lemma, which says it suffices to let $k = O\left(\frac{\log n}{\epsilon^2}\right)$ so that any pair-wise l_2 distance in $\mathbf{A}$ can be estimated within a $1 \pm \epsilon$ factor of the truth with high

probability.

In this thesis, we prove an analog of the JL Lemma for general $0 < \alpha \leq 2$. The method of *stable random projections* boils down to a statistical estimation problem, for estimating the scale parameter of a symmetric α-stable distribution. This problem is interesting because we seek estimators that are both statistically accurate and computationally efficient. We study and compare various estimators including the arithmetic mean, the geometric mean, the harmonic mean, the fractional power, as well as the maximum likelihood estimator. The analog of the JL Lemma for general $0 < \alpha \leq 2$ is proved base on the geometric mean estimator. For l_1 and l_{0+}, we also prove similar lemmas using the sample median estimator and the harmonic mean estimator, respectively.

This thesis also addresses several special cases of *stable random projections*. For the l_2 norm case (i.e., *normal random projections*), we propose improving the estimates by taking advantage of the marginal information. Also for the l_2 case, one can sample the projection matrix $\mathbf{R}$ from a much simpler *sub-Gaussian* distribution instead of normal. Under reasonable regularity conditions, a special *sub-Gaussian* distribution in $[-1,\ 0,\ 1]$ with probabilities $\left\{\frac{1}{s}, 1-\frac{1}{2s}, \frac{1}{s}\right\}$ and a large s (i.e., *very sparse random projections*) can work just as well as using normal random projections. For the l_1 case, i.e., *Cauchy random projections*, the estimation task is also particularly interesting. For example, the maximum likelihood estimator (MLE) is computationally feasible in the l_1 case and we propose using an *inverse Gaussian* distribution to accurately model the distribution of the MLE.

The method of *stable random projections* is appealing partly because one does not need to know anything about the data and still has the worst-case performance guarantees for estimating the pair-wise distances (i.e., the JL Lemma or the analog of the JL Lemma), although not for the inner products. This also implies that other techniques may outperform *stable random projections* when additional information about the data is available, for example, when the data are sparse.

The method of *Conditional Random Sampling (CRS)* is proposed particularly for sampling *massive sparse* data. Large-scale datasets are often highly sparse, e.g., text data. *Conditional Random Sampling (CRS)* takes advantage of the data sparsity by sampling only from the informative (i.e., non-zero) elements. In the estimation stage, CRS re-construct equivalent (conditional) *random samples* on the fly, on a pair-wise or group-wise basis, with the (equivalent) sample size retrospectively determined. Theoretical analysis is conditional on the (retrospective) sample size, which is a random variable.

Compared with *stable random projections*, CRS can be advantageous in various important scenarios.

- When the data are highly sparse, CRS can be (considerably) more accurate.
- In the l_2 case, when the data points are nearly independent (common in high-dimensional data), CRS can be significantly more accurate even when the data are not sparse.
- In the boolean data case, it is almost guaranteed that CRS outperforms *stable random projections*.
- One can re-use the same samples (sketches) from CRS to estimate many summary statistics including l_α distances (for any α). In contrast, *stable random projections* are limited to a particular l_α ($0 < \alpha \leq 2$) norm; and one has to start over for different α values.
- CRS can be used for estimating multi-way distances and histograms, a significant advantage over *stable random projections*.

These two sampling techniques can be complementary to each other, for solving large-scale problems, in search engines & information retrieval, databases, modern data streaming systems, numerical linear algebra, and many machine learning and data mining tasks involving massive computations of distances.

Bibliography

[1] James Abello, Panos M. Pardalos, and Mauricio G. C. Resende, editors. *HANDBOOK OF MASSIVE DATA SETS*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.

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

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

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

[5] Charu C. Aggarwal, editor. *Data Streams: Models and Algorithms*. Springer, New York, NY, 2007.

[6] Charu C. Aggarwal, Cecilia Magdalena Procopiuc, Joel L. Wolf, Philip S. Yu, and Jong Soo Park. Fast algorithms for projected clustering. In *SIGMOD*, pages 61–72, Philadelphia, PA, 1999.

[7] 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.

[8] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. In *SIGMOD*, pages 207–216, Washington, DC, 1993.

[9] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press, 1996.

[10] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499, Santiago de Chile, Chile, 1994.

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

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

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

[14] Shun-Ichi Amari. Differential geometry of curved exponential families-curvatures and information loss. *Annals of Statistics*, 10(2):357–385, 1982.

[15] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *FOCS*, pages 459–468, Berkeley, CA, 2006.

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

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

[18] Rosa Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63(2):161–182, 2006.

[19] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and issues in data stream systems. In *PODS*, pages 1–16, Madison, WI, 2002.

[20] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. On kernels, margins, and low-dimensional mappings. In *ALT*, pages 194 – 205, Padova, Italy, 2004.

[21] O.E. Barndorff-Nielsen and D. R. Cox. *Inference and Asymptotics*. Chapman & Hall, London, UK, 1994.

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

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

[24] Arindam Bhattacharjee, William G. Richards, Jane Staunton, Cheng Li, Stefano Monti, Priya Vasa, Christine Ladd, Javad Beheshti, Raphael Bueno, Michael Gillette, Massimo Loda, Griffin Weber, Eugene J. Mark, Eric S. Lander, Wing Wong, Bruce E. Johnson, Todd R. Golub, David J. Sugarbaker, and Matthew Meyerson. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *PNAS*, 98(24):13790–13795, 2001.

[25] R. N. Bhattacharya and J. K. Ghosh. On the validity of the formal edgeworth expansion. *The Annals of Statistics*, 6(2):434–451, 1978.

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

[27] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK. Also online `www.stanford.edu/~boyd/bv_cvxbook.pdf`, 2004.

[28] Sergey Brin, James Davis, and Hector Garcia-Molina. Copy detection mechanisms for digital documents. In *SIGMOD*, pages 398–409, San Jose, CA, 1995.

[29] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *WWW*, pages 107–117, Brisbane, Australia, 1998.

[30] Sergy Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *SIGMOD*, pages 265–276, Tucson, AZ, 1997.

[31] Sergy Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *SIGMOD*, pages 265–276, Tucson, AZ, 1997.

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

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

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

[35] Andrei Z. Broder. Filtering near-duplicate documents. In *FUN*, Isola d'Elba, Italy, 1998.

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

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

[38] Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. Syntactic clustering of the web. In *WWW*, pages 1157 – 1166, Santa Clara, CA, 1997.

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

[40] V. V. Buldygin and Yu. V. Kozachenko. *Metric Characterization of Random Variables and Random Processes*. American Mathematical Society, Providence, RI, 2000.

[41] Emmanuel J. Candes, Justin Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[42] 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.

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

[44] Surajit Chaudhuri, Rajeev Motwani, and Vivek R. Narasayya. On random sampling over joins. In *SIGMOD*, pages 263–274, Philadelphia, PA, 1999.

[45] Bin Chen, Peter Haas, and Peter Scheuermann. New two-phase sampling based algorithm for discovering association rules. In *KDD*, pages 462–468, Edmonton, Alberta, Canada, 2002.

[46] 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.

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

[48] Kenneth Church and Patrick Hanks. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29, 1991.

[49] 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.

[50] 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.

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

[52] N. Cressie. A note on the behaviour of the stable distributions for small index. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 31(1):61–64, 1975.

[53] Mark E. Crovella and Azer Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. *IEEE/ACM Trans. Networking*, 5(6):835–846, 1997.

[54] 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.

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

[56] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokn. Locality-sensitive hashing scheme based on p-stable distributions. In *SCG*, pages 253 – 262, Brooklyn, NY, 2004.

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

[58] Scott Deerwester, Susan T. Dumais, George W. Furnas, and Thomas K. Landauer. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1999.

[59] W. Edwards Deming and Frederick F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

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

[61] David L. Donoho. Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and Computational Harmonic Analysis*, 1(1):100–115, 1993.

[62] David L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

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

[64] William H. DuMouchel. On the asymptotic normality of the maximum-likelihood estimate when sampling from a stable distribution. *Annals of Statistics*, 1(5):948–957, 1973.

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

[66] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the Internet topology. In *SIGMOD*, pages 251–262, Cambridge,MA, 1999.

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

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

[69] 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.

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

[71] Marin Ferecatu, Michel Crucianu, and Nozha Boujemaa. Retrieval of difficult image classes using SVD-based relevance feedback. In *ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 23–30, New York, NY, 2004.

[72] Xiaoli Zhang Fern and Carla E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *ICML*, pages 186–193, Washington, DC, 2003.

[73] 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.

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

[75] Jessica H. Fong and Martin Strauss. An approximate lp difference algorithm for massive data streams. *Discrete Mathematics & Theoretical Computer Science*, 4(2):301–322, 2001.

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

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

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

[79] Jerome H. Friedman, F. Baskett, and L. Shustek. An algorithm for finding nearest neighbors. *IEEE Transactions on Computers*, 24:1000–1006, 1975.

[80] Jerome H. Friedman, J. Bentley, and R. Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3:209–226, 1977.

[81] Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom. *Database Systems: the Complete Book*. Prentice Hall, New York, NY, 2002.

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

[83] Anna C. Gilbert, Yannis Kotidis, S. Muthukrishnan, and Martin J. Strauss. One-pass wavelet decompositions of data streams. *IEEE Transactions on Knowledge and Data Engineering*, 15(3):541–554, 2003.

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

[85] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, New York, fifth edition, 1994.

[86] Warren R. Greiff. A theory of term weighting based on exploratory data analysis. In *SIGIR*, pages 11–19, Melbourne, Australia, 2003.

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

[88] Trevor J. Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning:Data Mining, Inference, and Prediction*. Springer, New York, NY, 2001.

[89] Taher H. Haveliwala, Aristides Gionis, and Piotr Indyk. Scalable techniques for clustering the web. In *WebDB*, pages 129–134, 2000.

[90] Taher H. Haveliwala, Aristides Gionis, Dan Klein, and Piotr Indyk. Evaluating strategies for similarity search on the web. In *WWW*, pages 432–442, Honolulu, HI, 2002.

[91] Monika R. Henzinger, Prabhakar Raghavan, and Sridhar Rajagopalan. *Computing on data streams*. American Mathematical Society, Boston, MA, USA, 1999.

[92] Christian Hidber. Online association rule mining. In *SIGMOD*, pages 145–156, Philadelphia PA, 1999.

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

[94] Albert Sydney Hornby, editor. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, Oxford, UK, fourth edition, 1989.

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

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

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

[98] Piotr Indyk. A small approximately min-wise independent family of hash functions. *Journal of Algorithm*, 38(1):84–90, 2001.

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

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

[101] Toshiya Itoh, Yoshinori Takei, and Jun Tarui. On the sample size of k-restricted min-wise independent permutations and other k-wise distributions. In *STOC*, pages 710–718, San Diego, CA, 2003.

[102] Jens Ledet Jensen. *Saddlepoint approximations*. Oxford University Press, New York, 1995.

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

[104] W. B. Johnson and G. Schechtman. Embedding l_p into l_1. *Acta. Math.*, 149:71–85, 1982.

[105] Donald E. Knuth. *The Art of Computer Programming (V. 2): Seminumerical Algorithms*. Addison-Wesley, New York, NY, third edition, 1997.

[106] C. Kraft and L. LeCam. A remark on the roots of the maximum likelihood equation. *The Annals of Mathematical Statistics*, 27(3):1174–1177, 1956.

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

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

[109] 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.

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

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

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

[113] Raoul LePage, Michael Woodroofe, and Joel Zinn. Convergence to a stable distribution via order statistics. *The Annals of Probability*, 9(4):624–632, 1981.

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

[115] Ping Li. Very sparse stable random projections, estimators and tail bounds for stable random projections. Technical report, `http://arxiv.org/PS_cache/cs/pdf/0611/0611114v2.pdf`, 2006.

[116] Ping Li. Estimators and tail bounds for dimension reduction in l_α $(0 < \alpha \leq 2)$ using stable random projections. Technical Report 2007-01, Department of Statistics, Stanford University, `http://www.stanford.edu/~pingli98/publications/stable.pdf`, 2007.

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

[118] Ping Li, Christopher J.C. Burges, and Qiang Wu. Learning to rank using classifications and gradient boosting. Technical Report MSR-TR-2007-74, Microsoft Research, 2007.

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

[120] Ping Li and Kenneth W. Church. Using sketches to estimate two-way and multi-way associations. Technical Report TR-2005-115, Microsoft Research, Redmond, WA, September 2005.

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

[122] Ping Li, Kenneth W. Church, and Trevor J. Hastie. Conditional random sampling: A sketched-based sampling technique for sparse data. Technical Report 2006-08, Department of Statistics, Stanford University, 2006.

[123] 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, 2007.

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

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

[126] 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*, 2007.

[127] Ping Li, Trevor J. Hastie, and Kenneth W. Church. Nonlinear estimators and tail bounds for dimensional reduction in l_1 using cauchy random projections. *Accepted to Journal of Machine Learning Research*, 2007.

[128] Ping Li, Sandy Napel, Burak Acar, David S. Paik, R. Brooke Jeffrey Jr., and Christopher F. Beaulieu. Registration of central paths and colonic polyps between supine and prone scans in computed tomography colonography: pilot study. *Medical Physics*, 31(10):2912–2923, 2004.

[129] 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, 2006.

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

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

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

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

[134] Gurmeet Singh Manku, Sridhar Rajagopalan, and Bruce G. Lindsay. Random sampling techniques for space efficient online computation of order statistics of large datasets. In *SIGCOMM*, pages 251–262, Philadelphia, PA, 1999.

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

[136] Yossi Matias, Jeffrey Scott Vitter, and Min Wang. Wavelet-based histograms for selectivity estimation. In *SIGMOD*, pages 448–459, Seattle, WA, 1998.

[137] M. Matsui and A. Takemura. Some improvements in numerical evaluation of symmetric stable density and its derivatives. *Communications on Statistics-Theory and Methods*, 35(1):149–172, 2006.

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

[139] Sally A. McKee. Reflections on the memory wall. In *CF*, pages 162–167, Ischia, Italy, 2004.

[140] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, New York, NY, 1995.

[141] S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1, 2 2005.

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

[143] Art B. Owen. *Empirical Likelihood*. Chapman & Hall/CRC, New York, NY, 2001.

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

[145] 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.

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

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

[148] Bengt Rosen. Asymptotic theory for successive sampling with varying probabilities without replacement, I. *The Annals of Mathematical Statistics*, 43(2):373–397, 1972.

[149] Bengt Rosen. Asymptotic theory for successive sampling with varying probabilities without replacement, II. *The Annals of Mathematical Statistics*, 43(3):748–776, 1972.

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

[151] Gennady Samorodnitsky and Murad S. Taqqu. *Stable Non-Gaussian Random Processes*. Chapman & Hall, New York, 1994.

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

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

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

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

[156] Christopher G. Small, Jinfang Wang, and Zejiang Yang. Eliminating multiple root problems in estimation. *Statistical Science*, 15(4):313–341, 2000.

[157] Frederick F. Stephan. An iterative method of adjusting sample frequency tables when expected marginal totals are known. *The Annals of Mathematical Statistics*, 13(2):166–178, 1942.

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

[159] Kyuseok Shim Sudipto Guha, Rajeev Rastogi. Cure: An efficient clustering algorithm for large databases. In *SIGMOD*, pages 73–84, Seattle, WA, 1998.

[160] Vivek R. Narasayya Surajit Chaudhuri, Rajeev Motwani. Random sampling for histogram construction: How much is enough? In *SIGMOD*, pages 436–447, Seattle, WA, 1998.

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

[162] Hannu Toivonen. Sampling large databases for association rules. In *VLDB*, pages 134–145, Bombay, India, 1996.

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

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

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

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

[167] Abraham Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949.

[168] Wm. A. Wulf and Sally A. McKee. Hitting the memory wall: implications of the obvious. *ACM SIGARCH Computer Architecture News*, 23(1):20–24, 1995.

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

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

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