DEEP COMPRESSION FOR RANDOM FEATURES: THEORY AND APPLICATIONS TO MACHINE LEARNING

By

XIAOYUN LI

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Statistics

Written under the direction of

Ping Li

And approved by

__________________________________

__________________________________

__________________________________

__________________________________

New Brunswick, New Jersey

May 2021
ABSTRACT OF THE DISSERTATION

Deep Compression for Random Features: Theory and Applications to Machine Learning

by XIAOYUN LI

Dissertation Director: Prof. Ping Li

With the rapid development of modern technologies, “big data” problems involving massive datasets have become more and more common in machine learning applications. The challenges of such large datasets call for the need to design efficient strategies for large-scale machine learning and data storage. In this thesis, we will study the data compression methods with randomized sketching techniques, from linear to non-linear methods.

In the linear case, we will consider the method of random projection (RP) and its compression method achieved by quantization. Firstly, we provide thorough statistical analysis on the cosine estimators given by quantized random projections (QRPs). By allowing two data vectors to be quantized by different precision, our analysis generalizes to various practical scenarios. Under Lloyd-Max (LM) quantization scheme, simple inner product estimator, normalized estimator and Maximum Likelihood Estimator (MLE) are studied and compared, and the correctness of the estimators is justified. Our theoretical and empirical results reveal the connection between the variance of debiased cosine estimators with similarity search accuracy.

We then consider the privacy of compressed linear sketches by noise addition. Based on an information-theoretic approach, we demonstrate how to find the optimal noise added to the QRPs that minimizes the mutual information. We then propose differentially private
(DP) methods for RP and QRP release with smaller noise magnitude compared with classical Gaussian mechanism, and show the improvement of QRP over RP in terms of differential privacy. Utility of Euclidean estimation using RPs and QRPs is also provided.

Next, we consider a new application of QRPs where the quantized linear sketches are used to train machine learning models. We study three models: nearest neighbor classification, linear classification and linear regression, where generalization error bounds are analyzed. Our analysis provides guidance on the proper choice of quantizers for different models in practice. We show that for linear classification and regression, the Lloyd-Max (LM) quantization is recommended.

In Part II of this dissertation, we consider approximate non-linear learning. Firstly, we demonstrate an example on how the method of random Fourier features (RFF) can be used in approximate kernel learning, by applying it to the Kernel Multi-view Discriminant Analysis (KMvDA) problem. Theoretical analysis on the perturbation bounds and numerical experiments are provided, showing the efficacy of RFFs in practice.

Then, we design two distortion optimal quantizers for RFF under Lloyd-Max (LM) scheme, based on our analysis on the probability distribution of RFF. The method is called LM-RFF. Detailed statistical analysis is provided, and experiments illustrate superior learning performance of LM-RFF over stochastic rounding technique. The proposed LM-RFF can reach the accuracy of full-precision RFF with high compression ratio, significantly reducing the memory cost in large-scale systems. Lastly, we propose a strategy to address the potential limitations of LM-RFF, by directly extracting non-linear sketches from QRPs (compressed linear measurements). The problem is highly efficient since we only need to store one set of compressed QRPs in the database, for both linear and non-linear learning. Moreover, it also provides a solution to very practical scenarios. Kernel estimation and uniform approximation error bound are provided. Empirical results confirm that it can be a convenient alternative for RFF compression in practice.
ACKNOWLEDGMENTS

It has been an exciting, challenging and life-changing experience for me to pursue a PhD degree during the past few years. Please allow me to express my deepest gratitude to people who gave me great support and guidance throughout this wonderful journey.

Firstly, I am more than grateful to my advisor, Professor Ping Li, for his warm help, continuous guidance and generous support. I always recall those days and nights when we worked together to get breakthroughs in our exciting research. His patience, passion, encouragement and sense of humor kept motivating me to devote myself in what I was doing and to be a person like him. Prof. Li gave me great flexibility in choosing the topics that I am interested in, and offered me the best opportunities to collaborate with other brilliant researchers which has been extremely helpful to my research. Prof. Li is more than an academic advisor for me. Those experience and suggestions he shared with me will also influence my career, my life, and help me to become a better person. I am so lucky to spend these years with Prof. Li and have him as my advisor.

Secondly, I sincerely appreciate Prof. Cun-Hui Zhang, Prof. Min Xu and Prof. William Steiger for their constructive advice and kind help on my research. I would like to thank the Department of Statistics, Rutgers University for the generous support on my research and finance. I appreciate all the support and help from the faculty, stuff and colleague. Thanks for creating a motivating, collaborating and comfortable environment, which allowed me to work on diverse topics in statistics and machine learning. I am grateful to the department chairs, Prof. Regina Liu and Prof. Rong Chen, and also the graduate directors, Prof. John Kolassa, Prof. Harry Crane and Prof. Tirthankar Dasgupta for giving me the chance to join the PhD program, and their kind support and guidance. Moreover, I would like to thank Prof. Min-ge Xie for his generous and constructive suggestions. I would like to say thank you to Prof. William Strawderman. His encouraging words gave me confidence when I started my PhD study.
Last but not the least, words cannot express my gratitude to my parents for their endless love and support. I would also like to thank my lovely girlfriend for her accompany and encouragement. I am so fortunate to have you by my side.
DEDICATION

To my parents, for their care, love, support and encouragement.

In memory of my beloved grandma. I know you will be proud of me.
TABLE OF CONTENTS

Abstract ................................................................. ii

Acknowledgments ......................................................... iv

Dedication ................................................................. vi

List of Tables ............................................................ xiii

List of Figures ........................................................... xiv

Chapter 1 Introduction ................................................... 1

1.1 Background and Motivation: Outline ............................... 1

1.1.1 Linear Sketching: Random Projections (RP) ................. 1

1.1.2 Non-linear Sketching: Random Fourier Features (RFF) .... 2

1.1.3 Quantized Random Projections (QRP) ....................... 4

1.1.4 Quantization Schemes for RFF .............................. 6

1.1.5 From QRP to Non-linear Sketching ......................... 7

1.2 Introduction to Quantization Methods ........................... 8

1.2.1 Fixed Lloyd-Max (LM) Quantization ....................... 9

1.2.2 Fixed Uniform Quantization ............................... 10

1.2.3 Stochastic Quantization (StocQ) ......................... 11
1.3 Outline of the Dissertation

I Linear Learning with Compressed Random Features

Chapter 2 Quantization for Random Projections

2.1 Preliminaries

2.1.1 Random Projection (RP) and Cosine Estimation

2.1.2 Lloyd-Max (LM) Quantizer Used for QRP

2.1.3 Existing Work on Cosine Estimation by Symmetric Quantized RP

2.1.4 Asymmetric Quantization: Two Motivating Scenarios

2.2 Scenario 1: Quantization vs. Full-precision

2.2.1 Three Cosine Estimators

2.2.2 Benefits of Normalized Estimators and Debiased Estimators

2.2.3 Analysis of Mis-ordering Probabilities in Similarity Search

2.2.4 Variance-Optimal Quantizer Design

2.3 Scenario 2: Quantization with Different Bits

2.3.1 The Cosine Estimators

2.3.2 Monotonicity of the Inner Product Estimates

2.4 Empirical Study: Similarity Search

Chapter 3 Privacy of Compressed Linear Sketches

3.1 Enhancing the Privacy of QRP: An Information-Theoretic Approach

3.1.1 Optimal Noise by Mutual Information Minimization

3.1.2 Example

3.2 Differential Privacy of QRP
Chapter 4  Generalization Error Analysis of Quantized Compressive Learning 49

4.1 Preliminaries ................................. 50

4.2 Quantized Compressive Nearest Neighbor Classification ............... 51
  4.2.1 Finite Sample Analysis ....................... 52
  4.2.2 Asymptotic Analysis ......................... 56
  4.2.3 Implication to Practice ....................... 58

4.3 Quantized Compressive Linear Classification with (0,1)-loss .......... 59
  4.3.1 Generalization Error Analysis ................ 60
  4.3.2 Choice of Quantizer $Q$ ...................... 62

4.4 Quantized Compressive Least Squares Regression .................... 63
  4.4.1 Generalization Error Analysis ................ 63
  4.4.2 Choice of Quantizer $Q$ ...................... 65

4.5 Numerical Study ................................ 65
  4.5.1 Quantized Compressive Classification .............. 66
  4.5.2 Quantized Compressive Regression .............. 68

Conclusion of Part I .................................. 70

II  Non-linear Learning with Compressed Random Features 72

Introduction to Random Fourier Features (RFF) ......................... 73
6.2.1 Stochastic Quantization (StocQ) for RFF ............................................. 101
6.2.2 Lloyd-Max (LM) Quantization for RFF ............................................. 102
6.2.3 Optimization for LM Quantizers .................................................. 103
6.2.4 Quantized Kernel Estimators ..................................................... 106

6.3 Theoretical Analysis ................................................................. 106
6.3.1 StocQ Estimator ................................................................. 106
6.3.2 LM Estimators ................................................................. 107
6.3.3 Variance Comparisons ......................................................... 110
6.3.4 Monotonicity of Mean Kernel Estimation ................................. 112

6.4 Numerical Experiments ............................................................. 114
6.4.1 Kernel SVM (KSVM) ............................................................ 115
6.4.2 Kernel Ridge Regression (KRR) ............................................. 117
6.4.3 Scale-invariant Kernel Approximation Error ............................. 117

6.5 Conclusion .............................................................................. 119

Chapter 7 One-Sketch-for-All: Non-linear Random Features from Compressed Linear Measurements ............................................... 121

7.1 Motivation of This Chapter ....................................................... 121
7.1.1 Practical Significance ............................................................ 122
7.1.2 Notations ........................................................................... 123

7.2 QRP-RFF Scheme ................................................................ 124

7.3 Analysis ................................................................................. 126
7.3.1 Equivalence in kernel learning when $b = 1$ ............................ 126
7.3.2 Information loss of QRP-RFF ............................................... 127

xi
7.3.3 Mean and variance ........................................... 128
7.3.4 Uniform Approximation Error ................................. 131
7.4 Experiments ....................................................... 133

Conclusion of Part II ................................................ 136

Bibliography ............................................................. 137

Appendix A Proofs of Chapter 2 ................................. 151
Appendix B Proofs of Chapter 3 ................................. 176
Appendix C Proofs of Chapter 4 ................................. 184
Appendix D Proofs of Chapter 5 ................................. 199
Appendix E Proofs of Chapter 6 ................................. 209
Appendix F Proofs of Chapter 7 ................................. 229
## LIST OF TABLES

2.1 Datasets used in the empirical study. Mean $\rho$ is the average pair-wise cosine similarity for sample pairs. Mean 1-NN $\rho$ is the average cosine similarity of each point to its nearest neighbor. ........................................ 31

4.1 Summary statistics of datasets used for experiments in Chapter 2, all standardized to unit norm. .............................................................. 66

5.1 Randomized KMvDA: Rank-1 recognition rate (%) of linear kernel, RBF kernel and RFF approximation. ................................................. 91

5.2 Multi-PIE: Linear, rank-1 recognition rate (%). ................................. 94

5.3 Multi-PIE: RBF, rank-1 recognition rate (%). .................................... 94

5.4 Multi-PIE: RFF $\dot{f}$, rank-1 recognition rate (%). ............................ 94

6.1 Constructed borders and reconstruction levels of LM-RFF quantizers, $b = 1, 2, 3, 4$, keeping three decimal places. Since the quantizers are symmetric about 0, we only present the positive part for conciseness. ............... 105

6.2 Constructed borders and reconstruction levels of LM$^2$-RFF quantizers, $b = 1, 2, 3, 4$, keeping three decimal places. Since the quantizers are symmetric about 0, we only present the positive part for conciseness. ............... 105
LIST OF FIGURES

1.1 Diagram of the dissertation structure. .................................................. 13

2.1 Scenario 1: variance of (debiased) estimators \( \hat{\rho}_{b,f}^{db}, \hat{\rho}_{b,f,n}^{db} \) and MLE. **Left panel:** the variance factor \( V_{b,f}^{db} \) for \( b = 1, 3, 5 \) and MLE. **Middle panel:** the variance factor \( V_{b,f,n}^{db} \) and MLE. **Right panel:** the variance ratio: \( \frac{V_{b,f,n}^{db}}{V_{b,f,ML}} \) (solid), \( \frac{V_{b,f,n}^{db}}{V_{b,f,ML}} \) (dash). ................................................................. 22

2.2 Variance factor \( V_{b,f}^{db} \) when using refined \( Q_{\rho, opt} \) for \( b = 2, 3, 4 \), at different \( \rho \). ........................................... 26

2.3 **Left panel:** the absolute bias (solid curves, in \( \log_{10} \) scale) of \( \hat{\rho}_{b_1,b_2} \) by simulations, along with the upper bound (red dashed curves) and lower bound (blue dashed curves) in Eq. (2.21). **Middle panel:** the absolute bias of \( \hat{\rho}_{b_1,b_2} \) with \( b_1 = b_2 \) (the symmetric case) along with the upper and lower bounds in Eq. (2.22). **Right panel:** The variance \( V_{b_1,b_2,n} \) of the normalized estimator in Theorem 2.3.3. ................................................................. 29

2.4 Scenario 1: Nearest neighbor search recovery results using cosine similarity and quantized estimators, from random projections. Columns 1: \( \hat{\rho}_{b,f} \). Column 2: \( \hat{\rho}_{b,f,n} \). ................................................................. 33

2.5 Scenario 2: Nearest neighbor search recovery results using cosine similarity and quantized estimators, from random projections. Columns 1: \( \hat{\rho}_{b_1,b_2} \). Column 2: \( \hat{\rho}_{b_1,b_2,n} \). ................................................................. 34

3.1 The probability mass function of QRP \( (Z = Q(X), X \sim N(0,1)) \), optimal noise \( R \) and the perturbed output \( Y = Z + R \), at different variance constraint \( \eta \). The first column is 3-bit uniform quantizer, and the second column is 4-bit uniform quantizer, both with maximal border 4. In the \( P(Y) \) plots, the red lines are the p.m.f. of the original QRP, \( Z \), and the blue lines represent the p.m.f. of \( Y \). ................................................................. 40
3.2 The required standard deviation of Gaussian noise. Red: Our algorithm (Algorithm 2). Blue: analysis based on standard Gaussian mechanism (Theorem 3.2.4). 43

3.3 The $\sqrt{E[\hat{d}]}$ and $Var[\hat{d}]$ of DP-RP and DP-QRP distance estimators, $\epsilon = 1$, $\delta = 0.1$. Dash curves are the theoretical values of Theorem 3.2.7 and Theorem 3.2.8, which overlap with the empirical curves. 48

4.1 Illustration of sign flipping in quantized space. Points above the black decision boundary are classified as 1, and 0 otherwise. Green dashed lines are borders of the quantizer. **Left:** data space classifier predicts 1. **Right:** quantized space prediction (using $Q(h^TR)$) changes to 0. 61

4.2 Empirical debiased variance of tested quantizers. To be divided by $k$. 66

4.3 Test accuracy of quantized compressive nearest neighbor classification. 68

4.4 Test accuracy of quantized compressive linear SVM. 68

4.5 Test MSE of QCLS. 69

5.1 Examples from CUFSF multi-view dataset. **View 1 (first row):** actual face photo. **View 2 (second row):** sketch drawn by artist. 77

5.2 An illustration of the MvDA framework, which jointly learns a projection for each view and conducts classification in the common subspace. 79

5.3 RFFs: rank-1 recognition rate vs. number of random Fourier features. The upper panel is for HFB dataset and the lower panel is for CUFSF dataset. 92

5.4 Linear kernel, RBF kernel and RFFs: rank-1 recognition rate vs. projection dimensionality. Upper panel: HFB. Lower panel: CUFSF. 93

5.5 Multi-PIE dataset. **Left panel:** Accuracy vs. number of RFFs, $-45^\circ \rightarrow 0^\circ$ cross-view. **Middle panel:** Accuracy vs. projection dimensionality, $-45^\circ \rightarrow 0^\circ$ cross-view. **Right panel:** Average accuracy over all cross-views vs. dimensionality. 95

6.1 Probability density functions $f_Z$ and $f_{Z_2}$ derived in Theorem 6.1.2. 98

6.2 The joint density of RFF (Theorem 6.1.4) with example $\rho = -0.9, 0.9$, and $\gamma = 1, 3, 5$. 100
6.3 Left: LM-RFF and LM\textsuperscript{2}-RFF quantizers, \( b = 2 \). Black curve is the RFF marginal density. Right two: Distortion \( D_1 \) and \( D_2 \) of LM-RFF, LM\textsuperscript{2}-RFF and StocQ (with uniform borders).

6.4 Observation 6.3.1 and 6.3.2 (black dash curves) vs. empirical bias (blue curves) of LM-RFF. Red dots are the biases given in the theorems at specific \( \rho \) values.

6.5 Variance (scaled by \( m \)) of StocQ, LM, and normalized LM kernel estimators, with \( \gamma = 0.5, 1, 2 \). The dashed curves are the variances of corresponding full-precision estimators, to which the variance of quantized estimators converges as \( b \) increases.

6.6 Debiased variance ratio of LM\textsuperscript{2}-RFF over LM-RFF, at various \( \gamma, b = 2 \). Left panel: vanilla estimator \( \frac{\text{Var}_{db}[\hat{k}_{Q,(2)}]}{\text{Var}_{db}[\hat{k}_Q]} \). Right panel: normalized estimator \( \frac{\text{Var}_{db}[\hat{k}_{n,Q,(2)}]}{\text{Var}_{db}[\hat{k}_{n,Q}]} \).

6.7 Debiased variance ratio of normalized LM estimators against simple LM estimators, \( b = 2, 3 \). Left panel: LM-RFF, \( \frac{\text{Var}_{db}[\hat{k}_{n,Q}]}{\text{Var}_{db}[\hat{k}_Q]} \). Right panel: LM\textsuperscript{2}-RFF, \( \frac{\text{Var}_{db}[\hat{k}_{n,Q,(2)}]}{\text{Var}_{db}[\hat{k}_{n,Q}]} \).

6.8 The derivatives obtained by Lemma 6.3.5 and Theorem 6.3.6 with different \( g_1 \) and \( g_2 \) functions, at multiple \( \gamma \). Black curves are the function value, and red lines are the theoretical derivatives at multiple points. Left panel: \( g_1(x) = x^3, g_2(x) = e^{5x} \), increasing functions. Middle panel: \( g_1(x) = g_2(x) = x^2 \), even functions. The region \( \rho \geq 0 \) is magnified for clarity. Right panel: \( g_1(x) = g_2(x) = Q \), where \( Q \) is the 1-bit LM-RFF quantizer. Here the derivative is approximated by the continuous approximation \( \tilde{g}(x) = \mu_2 \text{sign}(x)(1 - e^{-50|x|}) \), where \( \mu_2 \) is the positive reconstruction level of \( Q \).

6.9 Left two columns: Test accuracy of kernel SVM using different compression schemes of RFFs vs. number of random features \( m \). Right two columns: Test accuracy vs. memory per sample \((m \times b)\).

6.10 Kernel ridge regression: test MSE of KRR, StocQ vs. LM-RFF. Upper panels: Test MSE vs. number of random features \( m \). Bottom panels: Test MSE vs. memory per sample \((m \times b \text{ bits})\).

6.11 Scale-invariant kernel approximation error (Definition 6.4.1) of LM-RFF vs. StocQ, \( m = 2^{10} \). Black curves are full-precision RFF, blue curves are StocQ, and curves in red represent our proposed LM-RFF. For all metrics, the smaller the better.
7.1 Recap: applications of QRP in large-scale systems. Here, the full-precision RPs are never stored in memory. QRP can be used for linear learning and compressed sensing. The box in red is the new application studied herein—constructing non-linear random features for non-linear learning. 123

7.2 Illustration of QRP-RFF framework. QRP-RFF is constructed directly from QRP. Full-precision RPs can be discarded after QRP is generated. 125

7.3 Information loss of QRP-RFF: Mean of 1-bit QRP-RFF estimate from LM quantized QRP, at \( \rho = -1 \). The red curve is the true kernel, and the blue curve is the 1-bit QRP-RFF mean. 128

7.4 Solid curves: the mean of QRP-RFF estimate (Theorem 7.3.1). Dash curves: the true RBF kernel. We see some large deviations when \( b = 1, 2 \). 129

7.5 Variance of a random feature of FP-RFF, StocQ and QRP-RFF (Theorem 7.3.1, to be scaled by \( k \)). The variance of StocQ follows from Theorem 6.3.1. 131

7.6 Test accuracy of (linearized) kernel SVM using RFF with different quantization strategies. The dash line is for standard RBF kernel SVM. 134

C.1 An illustration of bounding the distance in projected space \( S_R \) with two covers. For simplicity we omit the scaling term \( \frac{1}{\sqrt{k}} \). Boxes resulted from red dash lines are the \( \epsilon \)-cover constructed by hand, and boxes surrounded by blue solid lines are induced cover by \( Q \). 186
CHAPTER 1
INTRODUCTION

1.1 Background and Motivation: Outline

In the era of big data, methods for efficient storage and computation of massive datasets become increasingly important. This dissertation presents a thorough study on the theory and practice of machine learning with memory-efficient data compression. Lying at the core of this thesis is the random sketching techniques for data processing. We study quantization methods for those random sketches, to further reduce the storage and computational expenses when dealing with large-scale data. The roadmap of this thesis starts from compressing the linear sketches. Afterwards, we extend it to non-linear sketches, which provide stronger learning capacity. Along the way, we develop several new applications of quantized random sketches in machine learning, with theoretical insights and empirical justification. Practically, the introduced methods can be used in various large-scale applications with good learning performance and significantly improved efficiency.

1.1.1 Linear Sketching: Random Projections (RP)

With the rapid development of modern technologies, massive datasets emerge from many sources (e.g., social networks, bioinformatics, webpages) have posed great challenges on the “big data” problem, where storing such large datasets and subsequently using them in machine learning models can be extremely costly. Therefore, it has been an active research area to design efficient methods for large-scale learning problems, in terms of both storage and computation. The main “big data” challenge is composed of two parts: (i) huge number of data points; (ii) high dimensionality. Sketching methods are popular to tackle the second challenge by reducing the data dimensionality. For example, the method
of Random Projection (RP) (Vempala, 2004), also known as linear sketching, projects the originally high-dimensional data onto a random lower-dimensional subspace. By the celebrated Johnson-Lindenstrauss (JL) Lemma (Johnson and Lindenstrauss, 1984), the pairwise distances between the data points can be preserved with high probability in the projected space. Concretely, consider two normalized data vectors (with unit \( l_2 \) norm) \( u, v \in \mathbb{R}^d \) with \( \rho = u^T v \). We generate a random Gaussian vector \( w \in \mathbb{R}^d \) with i.i.d. entries from \( N(0, 1) \). It then holds that

\[ \mathbb{E}[\langle w^T u, w^T v \rangle] = \langle u, v \rangle = \rho, \]

which is the basic idea of using random projection (RP) to approximate inner product. This nice unbiased estimation property, along with the approximation guarantee of JL Lemma, facilitates numerous applications of RP in theory, dimensionality reduction, nearest neighbor search, clustering, classification, regression, compressed sensing, etc. (Bingham and Mannila, 2001; Buhler, 2001; Buhler and Tompa, 2002; Fern and Brodley, 2003; Candès et al., 2006; Donoho, 2006a; Li et al., 2006b; Freund et al., 2007; Li, 2007; Indyk and Motwani, 1998; Dasgupta, 2000; Charikar, 2002; Fagin et al., 2003; Donoho, 2006b; Freund et al., 2007; Dasgupta and Freund, 2008; Dahl et al., 2013).

1.1.2 Non-linear Sketching: Random Fourier Features (RFF)

While linear sketching has been proven to be simple and effective in many applications, the learning performance can usually be improved since the non-linearity among the data features is not considered. For example, it has been shown in rich literature that learning with non-linear kernels can elevate the performance of various learning models, e.g., Kernel Ridge Regression (KRR), Kernel Support Vector Machine (KSVM), Kernel Discriminant Analysis (KDA), Kernel Principle Component Analysis (KPCA), etc. In this dissertation, we mainly consider the well-known Gaussian kernel, which is defined between sample
$u, v \in \mathcal{X}$ as

$$K_\gamma(u, v) = \langle \xi(u), \xi(v) \rangle = e^{-\frac{\gamma^2\|u-v\|^2}{2}},$$

where $\xi(\cdot) : \mathcal{X} \mapsto \mathbb{H}$ is the implicit feature map and $\gamma$ is a hyper-parameter. $\mathbb{H}$ denotes the associated Reproducing Kernel Hilbert Space (RKHS). In this thesis, in several chapters we will assume that $\mathcal{X}$ is the unit sphere, i.e., all the data points are normalized to have unit $l_2$ norm. This will save us from book-keeping the sample norms in our calculations, plus normalizing data vectors before classification is a fairly standard (or recommended) procedure in practice. Denoting $\rho = \cos(u, v) = u^Tv$, we can formulate

$$K_\gamma(u, v) = e^{-\frac{\gamma^2(2-2\rho)}{2}} = e^{-\gamma^2(1-\rho)}. \quad (1.1)$$

We will drop the subscript “$\gamma$” and use $K(\cdot)$ to denote the kernel when everything is clear from the context. Two major general issues with large-scale non-linear kernel learning arise. Firstly, storing/materializing a kernel matrix for a dataset of $n$ samples would need $n^2$ entries, which may not be realistic even just for medium datasets (e.g., $n = 10^6$). To avoid this problem, the entries of the kernel matrix can be computed on the fly from the original dataset. This however will increase the computation time, plus storing the original high-dimensional dataset for on-demand distance computations can also be costly. Secondly, the training procedure for nonlinear kernel algorithms is also well-known to be expensive (Platt, 1998; Bottou et al., 2007). Therefore, it has been an active area of research to speed up kernel machines, and using various types of random projections has become popular, e.g., (Li et al., 2006c; Achlioptas, 2001; Rahimi and Recht, 2007; Hsieh et al., 2014).

One popular way to solve the aforementioned limitation of non-linear kernel learning is called kernel linearization. Similar to the idea of random projection, we can construct random sketches with the goal to approximate the non-linear kernel among data samples. This scheme is generally named as non-linear sketching. By casting trigonometric functions on random projections, the method of Random Fourier Features (RFF) (Rahimi and Recht,
2007) is a popular data-independent approach for kernel linearization. For two $d$-dimensional data samples $u, v \in \mathcal{X}$, the inner product of the RFFs of $u$ and $v$ is an unbiased estimator of the kernel $K(u, v)$. Thus, taking the inner product of the RFFs between two data points would approximate the Gaussian kernel. This implies that, training linear machines on RFFs would approximate training a non-linear kernel machine on the original data, which may significantly accelerate training and alleviate memory burden for storing the large kernel matrix. This has become very popular in literature with numerous applications, e.g., (Raginsky and Lazebnik, 2009; Yang et al., 2012; Dai et al., 2014; Yen et al., 2014; Lopez-Paz et al., 2014; Shah and Ghahramani, 2015; Sutherland and Schneider, 2015; Avron et al., 2017; Sun et al., 2018).

In Chapter 5, as an example on how RFF can be applied in machine learning, we apply RFF to a new application: Multi-view Discriminant Analysis (MvDA). In this model, we have data from different views of the objects, and the goal is to identify the object from one view, given the data from another view. The solution of MvDA is to learn a common subspace, and conduct classification based on the pair-wise distances in the projected space. By kernelizing the problem, we develop Kernel MvDA (KMvDA) and combine RFF to randomize the training procedure. Perturbation of the subspace resulting from RFF approximation is analyzed, and empirically we show that using RFF can match the accuracy of using exact Gaussian kernel in KMvDA.

1.1.3 Quantized Random Projections (QRP)

For linear learning, despite that the data dimensionality can be effectively reduced by random projection, the issue of huge expense for storing these projections, when we have massive data samples, still remains. Consequently, in recent years, “random projection + quantization” has been an active research topic to further compress the linear sketches. That is, the projected data, which are in general real-valued (i.e., infinite precision), are quantized into integers with a small number of bits. Applying quantization on top of random
projections has at least two major advantages: (i) the storage cost is further (substantially) reduced; and (ii) some important applications such as hashing-table-based near neighbor search, require using quantized data for indexing purposes.

The pioneering example of quantized random projections is the so-called “1-bit” (sign) random projections, initially used for analyzing the MaxCut problem (Goemans and Williamson, 1995). Afterwards, existing literature has mainly adopted QRP in the following application:

**Application 1.1.1. Similarity search and compressed sensing:** 1-bit QRP has been adopted for near neighbor search (Charikar, 2002), and compressed sensing (Boufounos and Baraniuk, 2008; Jacques et al., 2013; Plan and Vershynin, 2013; Li, 2016). The more general multi-bit QRP is also applied to these two problems in (Datar et al., 2004; Zymnis et al., 2010; Li, 2017a; Li and Slawski, 2017). That is, similarity search and compressed sensing are conducted on the randomly projected data, after being quantized.

In this dissertation, we will start in Chapter 2 by studying the distortion optimal quantization under Lloyd-Max (LM) scheme of random projections, considering multiple general and practical scenarios where two data sources are coded through different strategy (i.e., asymmetric quantization case). Detailed statistical analysis is provided on the simple cosine estimator, normalized cosine estimator and maximum likelihood estimator (MLE), which then motivates some new quantizer design. Furthermore, we show that the simplest 1-bit QRP may achieve better performance in search problems compared with full-precision RPs (FP-RP). In addition, in Chapter 3, we consider the data privacy of QRPs, in terms of information-theoretic and differential privacy (DP). Optimal noise distribution is discussed, and the DP guarantee of QRP is analyzed.

Random projections (linear sketches) can also be used in regression and classification tasks. This line of research, called compressive learning (Garg et al., 2002; Calderbank et al.,
studies the learning problems in the projected space instead of the original data space. In Chapter 4, we consider a new application of QRP:

**Application 1.1.2. Quantized compressive learning:** The QRP s can be effectively used in regression and classification models for large-scale learning. In particular, we study three learning models: nearest neighbor classification, linear classification and linear regression, when trained with the QRP s.

The benefit of quantized compressive learning is straightforward. If one can achieve same learning performance with e.g. 4-bit encoded RPs, it is then a substantial saving the memory compared with full-precision (32 or 64 bits) representation. Theoretical analysis on the generalization error bounds are developed for each model, when learned in the quantized projected space. Our results identify the factor of quantizers that influences each learning model, leading to practical guidance on the choice of proper quantizers for different statistical models. In general, the distortion optimal quantization is favorable for all above models, and the trick of normalizing the random projections may improve nearest neighbor classification accuracy.

### 1.1.4 Quantization Schemes for RFF

Similar to quantizing the random projections, we can also condense the full-precision random Fourier features (FP-RFFs). In Chapter 6, we will provide thorough statistical analysis on this problem. After identifying the probability distribution of RFFs, we propose the corresponding Lloyd-Max (LM) distortion optimal quantizers. Interestingly, the marginal distribution of RFF is free of parameter $\gamma$, which largely simplifies the LM quantizer design. We compare the so-called LM-RFF quantization with stochastic rounding method. Moreover, robust kernel approximation error metrics are defined and investigated. Extensive experiments on large-scale learning problems are provided to demonstrate the efficacy of
the proposed compression schemes. Particularly, the proposed LM-RFF method can achieve the same accuracy as FP-RFF, with substantially reduced memory cost. Thus, it can serve as an effective and highly efficient compression strategy for large-scale non-linear learning in practice.

1.1.5 From QRP to Non-linear Sketching

In LM-RFF, the practitioner may have to store various sets of compressed RFFs for different $\gamma$ in the RBF kernel. This might not be feasible in terms of memory in many cases. To address this potential issue, in Chapter 7 we introduce a highly efficient scheme that accomplishes both approximate linear and non-linear learning while only requiring one set of compressed linear measurements (i.e., QRP). This, at the same time, is more practical since in large-scale databases one typically would like to discard the full-precision random projections after the quantized RPs have been stored.

Consider the following practical scenario. Suppose a server has collected random projections of massive data samples and stored quantized version of RPs in the database to save memory cost. In this procedure, we have lost access to the full-precision RPs (otherwise, quantization becomes meaningless). Now, in order to achieve better learning performance, a data scientist wants to apply non-linear kernel method for large-scale learning. Typically, this can be done in a standard way by learning with RFF generated by RP. Yet, FP-RPs are discarded after quantization, and re-collecting the past data might be inconvenient or even impossible (e.g., due to data loss or privacy reason). This practical scenario prompts us to explore a new application of QRP:

**Application 1.1.3. Non-linear learning:** QRP can be used to construct non-linear random features for large-scale non-linear learning. We call it QRP-RFF scheme. By extracting RFF directly from quantized RP on hand, QRP-RFF provides a convenient compression scheme that only needs one sketch (the QRP) for both linear and non-
linear learning, sparing the (unnecessary) space for the original FP-RP. Indeed, in the above practical scenario where we have erased full-precision RP (which is a natural choice), the QRP-RFF approach becomes a requisite approach.

In other words, Chapter 7 explores a new application of QRPs, apart from \( A_1 \) and \( A_2 \). Remarkably, our studied problem is “one-sketch-for-all” because of the fact that we only need one set of highly compressed linear measurements (i.e., QRPs) for both linear and non-linear learning. As stated before, this is a more practical setting in real-world applications. Naturally, one may think that very careful quantizer design is needed to make this scheme work. Perhaps surprisingly, we will show that the common quantizers used for QRP suffices to provide good performance. As one would expect, this QRP-RFF approach is less accurate than LM-RFF which is directly optimized on RFFs. Yet, it still outperforms stochastic rounding strategy especially when the number of bits is small. Therefore, it is a feasible alternative in practice for memory-efficient non-linear learning with random features.

1.2 Introduction to Quantization Methods

As we stated preceding, the primary focus of this dissertation is to study compression algorithms for large-scale machine learning. Our goal is achieved by tools from Quantization Theory. Usually, a full-precision number stored in computer systems takes 32 or 64 bits memory. By only storing a few bits per number, the memory efficiency of quantization is significant and straightforward. For example, in the extreme case, 1-bit quantization means that we only need to store binary 0/1 values in the database. Such substantial storage saving would be extremely helpful and welcome in industrial applications.

In the sequel, we introduce three quantization strategies that will appear in this dissertation. The Lloyd-Max (LM) quantization scheme will be the main focus throughout this thesis because of its nice theoretical property. The other two methods, uniform quantizer and stochastic rounding, will be compared with LM quantizer on some tasks. We will present the general formulation of each type of quantizer, and in later chapters we will describe the
specific quantizers used for different problems.

1.2.1 Fixed Lloyd-Max (LM) Quantization

In signal processing and quantization theory, *Lloyd-Max (LM) quantization* (Max, 1960; Lloyd, 1982) is an important scheme which leads to many well-known algorithms such as the $k$-means clustering. In this dissertation, we consider fixed LM quantizers. An $M$-level fixed scalar quantizer $Q_M(\cdot)$ is specified by $M + 1$ decision borders $t_0 < t_1 < \cdots < t_M$ and $M$ reconstruction levels (or codes) $\mu_i$, $i = 1, \ldots, M$, with the quantizing operator defined as

$$Q_M(x) = \mu_{i^*}, \quad i^* = \{i : t_{i-1} < x \leq t_i, \quad 1 \leq i \leq M \}. \quad (1.2)$$

Following the convention, we assume that $M = 2^b$ with $b$ a positive integer representing the number of bits of the quantizer. Thus, we will write $Q_b$ instead of $Q_M$ whenever needed. Under the LM quantization framework, each quantizer is associated with a signal distribution, denoted as $X \sim f$ herein. The distortion is an important quantity that measures how much information is lost from the original signal due to quantization.

**Definition 1.2.1.** The distortion of a $b$-bit quantizer $Q_b(\cdot)$ with respect to signal distribution $f$ is defined as

$$D_Q = \mathbb{E} \left[ (X - Q_b(X))^2 \right] = \int (x - Q_b(x))^2 f(x) dx. \quad (1.3)$$

Then, the Lloyd-Max (LM) quantizer is constructed such that the distortion is minimized:

$$\textbf{LM quantizer:} \quad Q_b = \arg \min_Q D_Q, \quad (1.4)$$

That is, the LM quantizers are distortion optimal. In practice, Lloyd’s algorithm (Lloyd, 1982) is used to find the solution, which alternates between updating borders and reconstruction points until convergence. The updating rule is derived by setting the derivative
with respect to each border and reconstruction level to zero to find a stationary point. The concrete procedure is summarized in Algorithm 1. Here, we assume that the support of the signal distribution is known, such that the two ends \( t_0 \) and \( t_2^b \) can be fixed (possibly infinite). According to (Wu, 1992a), the convergence of Algorithm 1 is guaranteed since the squared loss is convex and symmetric.

**Algorithm 1** General construction of Lloyd-Max (LM) quantizer

**Input:** Density \( f(x) \); Number of bits \( b \); \( t_0, t_2^b \)

**Output:** LM quantizer \( t = [t_0, ..., t_2^b] \), \( \mu = [\mu_1, ..., \mu_2^b] \)

While true

For \( i = 1 \) to \( 2^b \)

Update \( \mu_i \) by \( \mu_i = \frac{\int_{t_i-1}^{t_i} xf(x)dx}{\int_{t_i-1}^{t_i} f(x)dx} \)

End For

For \( i = 1 \) to \( 2^b - 1 \)

Update \( t_i \) by \( t_i = \frac{\mu_i + \mu_{i+1}}{2} \)

End For

Until Convergence

### 1.2.2 Fixed Uniform Quantization

Compared with distortion optimal LM quantization, a rather simple strategy is the *uniform quantization* which does not involve any optimization. The construction of \( b \)-bit uniform quantizer is as follows:

- If the signal distribution is supported on a finite interval \([a, b]\), then the uniform quantizer partitions \([a, b]\) into \( 2^b \) bins with equal size \( \Delta = \frac{b-a}{2^b} \), and the reconstruction levels are the mid points of each bin.

- If \( b = \infty \), a threshold \( c > 0 \) is chosen, and \([a, c]\) is divided into \((2^b - 1)\) length-\( \Delta \) intervals, with \( \Delta = \frac{b-a}{2^{b-1}} \). If the value to be quantized, \( x \), is greater than \( c \), it is quantized to \( c + \Delta/2 \). Similar cutting strategy holds reversely when \( a = -\infty \).

Since uniform quantization does not reach any optimality theoretically, the cutting point \( c \) are usually chosen based on experience. Additionally, for the same reason, in many cases
uniform quantization tends to perform worse than LM quantization.

1.2.3 Stochastic Quantization (StocQ)

The previously introduced LM and uniform quantization both rely on fixed quantizers. Alternatively, another popular coding strategy is the so-called *stochastic quantization* (StocQ), also known as *stochastic rounding* in signal processing literature. More specifically, a $b$-bit StocQ quantizer has $2^b$ borders, denoted as $-\infty < t_0 < \cdots < t_{2^b-1} < \infty$. Let $[t_i, t_{i+1}]$ be the interval containing $x$, the value being quantized. Unlike fixed quantization, StocQ pushes $x$ to either $t_i$ or $t_{i+1}$, depending on its distance to the borders. Concretely, denoting $\Delta_i = t_{i+1} - t_i$,

$$
P(Q(x) = t_i) = \frac{t_{i+1} - x}{\Delta_i}, \quad P(Q(x) = t_{i+1}) = \frac{x - t_i}{\Delta_i}. \quad (1.5)
$$

By the sampling procedure, $Q(x)$ is conditionally unbiased of $x$. Yet, it pays a price of extra variance especially with few bits (e.g., $b = 1, 2$). Note that, in most applications, the borders in StocQ are set to be uniform for simplicity. In this dissertation, we consider the more general approach where the borders are not necessarily uniform.

1.3 Outline of the Dissertation

Following previous introduction, the dissertation will be divided into two parts. In the first part, we study the compression schemes for random projections (RP) in linear learning. In the second part, we consider various compression methods for random Fourier features (RFF) with applications in non-linear learning. A graphical demonstration is presented in Figure 1.1.

In Chapter 2, we study the general procedure of adopting Lloyd-Max (LM) quantization to random projections, where two data sources may be quantized by different number of bits (asymmetric quantization). Detailed statistical analysis is provided on the cosine
estimation, monotonicity and implications to similarity search problems. Various cosine estimators are analyzed, and a new quantizer is developed. In Chapter 3, we consider the privacy of QRP, from information-theoretic and differential privacy (DP) perspectives. Optimal noise addition scheme is proposed, and we also analyze a differentially private QRP publishing strategy based on Gaussian noise addition. In Chapter 4, QRP is applied to three learning models: nearest neighbor classification, linear classification and linear regression. Generalization error bounds to investigated, providing insights on the choice of proper quantizer for different tasks.

In Part II, we focus on the random Fourier features (RFFs) for approximate non-linear learning. In Chapter 5, we apply RFF to Kernel Multi-view Discriminant Analysis (KMvDA) problem. In Chapter 6, we design distortion optimal LM quantizers for RFF (LM-RFF), and compare it with stochastic quantization method. Besides thorough theoretical analysis, extensive numerical experiments are conducted, which illustrate that with the compressed random sketches, machine learning can be highly memory-efficient while achieving same accuracy. In Chapter 7, we address a potential limitation of LM-RFF in practice, and explore a new application of QRP where it is conveniently used in approximate non-linear kernel learning. Specifically, non-linear random features are directly extracted from compressed linear sketches (QRPs). Our study also provides a solution to a very practical scenario.
Figure 1.1: Diagram of the dissertation structure.
Part I

Linear Learning with Compressed
Random Features
CHAPTER 2
QUANTIZATION FOR RANDOM PROJECTIONS

As introduced in Section 1.1.1, the linear sketching method of random projection (RP) is a popular dimensionality reduction technique for high-dimensional data. In this chapter, we consider compression schemes for random projections for further storage and computation efficiency. We provide thorough analysis on the distance (cosine) estimators using QRP, as well as the practical implications to similarity search problems in Application 1.1.1. To begin with, we first introduce the backgrounds, notations and problem setting.

2.1 Preliminaries

2.1.1 Random Projection (RP) and Cosine Estimation

Let \( U = [u_1, \ldots, u_n]^T \in \mathbb{R}^{n \times d} \) be the original data matrix (with \( d \) possibly being large). Random projections are realized by

\[
Z = [z_1, \ldots, z_n]^T = U \times R,
\]

where \( R \in \mathbb{R}^{d \times k} \), \( k \ll d \) is a random matrix with i.i.d. standard Gaussian entries. In words, for each data point \( u_i \), we generate \( k \) linear sketches by \( k \) independent random projections.

Let \( \| \cdot \|_2 \) denote the \( l_2 \) Euclidean norm. Recall that throughout this dissertation, we assume that every data point is normalized to unit norm\(^1\), i.e., \( \| u_i \|_2 = 1 \), \( \forall 1 \leq i \leq n \). We will hence use the terms “inner product” and “cosine similarity” interchangeably.

\(^1\)Normalizing each data vector to the unit norm is a standard (or recommended) preprocessing procedure in many applications. We adopt this assumption merely for convenience. When data is unnormalized, our results still hold, although we will need to keep track of the norms.
For the convenience of presentation, our results (estimators and properties) will be given for two pairs of data vectors, \( u \) and \( v \) (and correspondingly \( z_u \) and \( z_v \)). Let \( \rho = \langle u, v \rangle \) be the inner product between \( u \) and \( u \). We also denote \( x = z_u \) and \( y = z_v \) for simplicity of notation. It is then easy to verify that \( (x, y) \) is bi-variate normal:

\[
\begin{pmatrix} x \\ y \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right). \tag{2.1}
\]

By RP, we derive \( k \) independent sketches \( (x_i, y_i), 1 \leq i \leq k \), each following (2.1). We are interested in using these random projections to estimate cosine \( \rho \). One can use the usual simple estimator

\[
\hat{\rho}_f = \frac{1}{k} \sum_{i=1}^{k} x_i y_i, \quad \text{with } \mathbb{E} (\hat{\rho}_f) = \rho, \quad \text{Var} (\hat{\rho}_f) = \frac{1 + \rho^2}{k}. \tag{2.2}
\]

where \( \mathbb{E} (\hat{\rho}) \) is the expectation and \( \text{Var} (\hat{\rho}) \) is the variance. Note that the variance grows as \( |\rho| \) increases. Moreover, we can take advantage of the following so-called “normalized estimator”:

\[
\hat{\rho}_{f,n} = \frac{\sum_{i=1}^{k} x_i y_i}{\sqrt{\sum_{i=1}^{k} x_i^2} \sqrt{\sum_{i=1}^{k} y_i^2}}, \quad \mathbb{E} (\hat{\rho}_{f,n}) = \rho + O \left( \frac{1}{k} \right), \quad \text{Var} (\hat{\rho}_{f,n}) = \frac{(1 - \rho^2)^2}{k^2} + O \left( \frac{1}{k^2} \right). \tag{2.3}
\]

\( \hat{\rho}_{f,n} \) is nearly unbiased, and it substantially reduces the estimation variance, especially near two extreme points \( \rho = \pm 1 \) where the variance tends to zero. We refer readers to Anderson (2003); Li et al. (2006b) for more details on the calculation. In this chapter, we will use \( \phi(x) \) and \( \Phi(x) \) to denote the probability density function (p.d.f.) and cumulative density function (c.d.f.) of standard normal distribution, respectively.
2.1.2 Lloyd-Max (LM) Quantizer Used for QRP

We will study using Lloyd-Max (LM) quantizer to compress the random projections (see Section 1.2.1 for background). One key step is to determine the signal distribution. As we see from (2.1), the marginal distribution of the projected data, \( x \), is standard normal. Therefore, we set the signal distribution as standard normal \( \phi(x) \) in Algorithm 1 to build the LM quantizer for distortion minimization (1.4). That is, the LM quantizer minimizes the distortion

\[
D_Q = \mathbb{E}_{X \sim N(0,1)} [(X - Q_b(X))^2] = \int (x - Q_b(x))^2 \phi(x) dx.
\]

2.1.3 Existing Work on Cosine Estimation by Symmetric Quantized RP

Using quantized random projections (QRP) in cosine estimation has been considered in prior literature, for the sake of its significant storage saving. In general, existing approaches can be divided into two categories: collision-based methods and scalar-based methods. Note that, all existing work assumes that both sources of data (i.e., \( x \) and \( y \)) are quantized under same scheme, which we call symmetric quantization.

In the collision-based approach, each QRP is treated as the index of the bin in which the RP is contained. The cosine can be estimated by the collision counts and a proper inverse map, i.e., \( \hat{\rho}_{col} = g^{-1} \left( \frac{\sum_{i=1}^{k} 1\{Q(x_i) = Q(y_i)\}}{k} \right) \), where \( g : [-1, 1] \mapsto [0, 1] \) is defined by \( \rho \mapsto g(\rho) = P[Q(x) = Q(y)] \). This approach is studied in detail in (Li et al., 2014, 2016b). In practice, the estimation can be efficiently computed once we tabulate down the map \( g \) a priori on a fine grid and resort to table lookup. The statistical efficiency of such estimator, when measured by the Mean Squared Error (MSE), decreases with more bits and increases with \( \rho \to 1 \), i.e., in high similarity regime.

In this dissertation, we will mainly focus on the scalar-based methods. That is, the quantized projections are simply regarded as scalars, based on which similar estimators as (2.2)
and (2.3) can be constructed. Li and Slawski (2017) studied such inner product estimators under LM quantization scheme, by analyzing the biases and variances of estimators in the symmetric case (the observations $x_i$ and $y_i$ are quantized by the same LM scheme with same $b$). In this dissertation, we study the asymmetric setting by using $b_1$ number of bits for quantizing $x_i$ and $b_2$ number of bits for $y_i$. The motivation will be provided in next section. Apparently, symmetric quantization is a special case of our results (i.e., $b_1 = b_2$). Interestingly, our analysis also leads to a more refined bound on the estimation bias in the symmetric case compared to the corresponding bound in (Li and Slawski, 2017). See Section 2.3 for the detailed results.

2.1.4 Asymmetric Quantization: Two Motivating Scenarios

Everyday, massive data is collected from every possible place that one can think of, but it is often impractical to cast a universal encoding strategy on data storage methods for every place/device. As a consequence, it becomes a meaningful task to look into the estimation problems with data encoded by different algorithms, or namely, the asymmetric case. In this thesis, we provide insights on this type of problems, and particularly, we consider recovering inner products from asymmetrically quantized random projections, arising from the following two practical scenarios.

- **Scenario 1: quantization vs. full-precision.**  Consider, for example, a retrieval system which uses random projections to process every data vector. To save storage, the projected data stored in the repository are quantized into a small number of bits. When a query data vector arrives, it is first processed by random projections. We then have the option of quantizing the projected query data vector before conducting the similarity search (with vectors in the repository); but we do not have to do the quantization step since we still have the projected query data vector in full-precision (why waste?). This natural and practical situation hence creates the “quantization vs. full-precision” estimation problem.
• **Scenario 2: quantization with different bits.** In applications such as large ad hoc networks (Younis and Fahmy, 2004; Li, 2008), data are collected and processed by different nodes (e.g., sensors or mobile devices) at different locations before sent to the central unit or cloud server. However, distinct nodes may use different quantization methods (or different bits) due to many possible reasons, e.g., memory capacity or purpose of data usage. In this situation, information retrieval among data sources using different quantization schemes could be on the cards. As a tightly related topic, asymmetric distributed source coding (with different bits from different sources) has also been considered in (Berger, 1961; Pradhan and Ramchandran, 2000) among others for sensor networks.

We may notice that, Scenario 1 is in fact an important special case of Scenario 2, where one source of data is quantized with infinite bits. In this chapter, we provide thorough statistical analysis on the above two scenarios. All the proofs in this chapter are deferred to Appendix A.

### 2.2 Scenario 1: Quantization vs. Full-precision

Recall that, we have i.i.d. observations \( \{x_i, y_i\}, i = 1, 2, \ldots, k \), from the standard bi-variate normal distribution with \( x_i \sim N(0, 1) \), \( y_i \sim N(0, 1) \), and \( \mathbb{E}(x_i y_i) = \rho \). In this section, we study Scenario 1: quantization vs. full-precision. That is, we quantize \( x_i \) with \( b \) bits and we leave \( y_i \) intact. The task is to estimate \( \rho \) from \( \{Q_b(x_i), y_i\}, i = 1, 2, \ldots, k \).

#### 2.2.1 Three Cosine Estimators

We now study three cosine estimators from QRP under LM quantization scheme, namely, the simple estimator, the normalized estimator and Maximum Likelihood Estimator (MLE).
**Simple Estimator.** Firstly, the simple estimator, similar to (2.2), is defined via the (asymmetric) inner product as

\[
\hat{\rho}_{b,f} = \frac{1}{k} \sum_{i=1}^{k} Q_b(x_i)y_i.
\]  

(2.4)

As one would expect, this estimator \(\hat{\rho}_{b,f}\) is no longer unbiased. In this chapter, we define

\[
\gamma_{\alpha,\beta} = \mathbb{E} \left( Q_b(x)^\alpha y^\beta \right), \quad \xi_{\alpha,\beta} = \mathbb{E} \left( Q_b(x)^\alpha x^\beta \right).
\]  

(2.5)

Note that \(\xi_{\alpha,\beta}\) can be represented by \(\gamma_{\alpha,\beta}\), but we use both for convenience. We can show that \(\mathbb{E}(\hat{\rho}_{b,f}) = \xi_{1,1}\rho\). Hence, we can attempt to remove the bias explicitly by using the following “debiased estimator”:

\[
\hat{\rho}_{db}^{b,f} = \frac{\hat{\rho}_{b,f}}{\xi_{1,1}} = \frac{1}{k} \frac{1}{\xi_{1,1}} \sum_{i=1}^{k} Q_b(x_i)y_i.
\]  

(2.6)

Also note that \(\xi_{1,1} = \xi_{2,0} = 1 - D_b\) from definitions. For \(b = 1, 2, 3, 4, \infty\), we can compute \(\xi_{1,1} = 0.6366, 0.8825, 0.9655, 0.9905, 1\), respectively (keeping four decimal points). In fact, it is also known that \(D_b = \frac{3^{3/2} \pi}{12} 2^{-2b}\), i.e., the bias of QRP cosine estimator decays at the rate of \(O(2^{-2b})\). In the following, Theorem 2.2.1 summarizes the expectations and variances of the two estimators \(\hat{\rho}_{b,f}\) and \(\hat{\rho}_{db}^{b,f}\).

**Theorem 2.2.1.** Let \(\xi_{i,j}\) be defined as (2.5). For simple estimators,

\[
\mathbb{E}(\hat{\rho}_{b,f}) = \xi_{1,1}\rho, \quad \mathbb{E}(\hat{\rho}_{db}^{b,f}) = \rho,
\]  

(2.7)

\[
\text{Var}(\hat{\rho}_{b,f}) = \frac{V_{b,f}}{k}, \text{ with } V_{b,f} = (\xi_{2,2} - \xi_{2,0} - \xi_{1,1}^2)\rho^2 + \xi_{2,0}
\]  

(2.8)

\[
\text{Var}(\hat{\rho}_{db}^{b,f}) = \frac{V_{db}^{b,f}}{k}, \text{ with } V_{db}^{b,f} = \frac{(\xi_{2,2} - \xi_{2,0} - \xi_{1,1}^2)\rho^2 + \xi_{2,0}}{\xi_{1,1}^2}.
\]  

(2.9)
Normalized Estimator. Similar to the full-precision case, we can also take advantage of
the (beneficial) effect of normalization by defining the normalized estimator:

$$\hat{\rho}_{b,f,n} = \frac{\sum_{i=1}^{k} Q_b(x_i)y_i}{\sqrt{\sum_{i=1}^{k} Q_b^2(x_i)}}.$$

(2.10)

The asymptotic moments are summarized in Theorem 2.2.2.

**Theorem 2.2.2.** Let $\xi_{i,j}$ and $\gamma_{i,j}$ be defined as (2.5). For normalized estimators, as $k \to \infty$, we have

$$\hat{\rho}_{b,f,n} = \frac{\sum_{i=1}^{k} Q_b(x_i)y_i}{\sqrt{\sum_{i=1}^{k} Q_b^2(x_i)}}$$

$$\mathbb{E}(\hat{\rho}_{b,f,n}) = \sqrt{\xi_{1,1}} + O\left(\frac{1}{k}\right),$$

(2.11)

$$\hat{\rho}_{db}^{b,f,n} = \sqrt{\xi_{1,1}},$$

$$\mathbb{E}(\hat{\rho}_{db}^{b,f,n}) = \rho + O\left(\frac{1}{k}\right),$$

(2.12)

$$\text{Var}(\hat{\rho}_{b,f,n}) = \frac{V_{b,f,n}}{k} + O\left(\frac{1}{k^2}\right),$$

$$\text{Var}(\hat{\rho}_{db}^{b,f,n}) = \frac{V_{db}^{b,f,n}}{k} + O\left(\frac{1}{k^2}\right),$$

(2.13)

$$V_{b,f,n} = \left(\gamma_{4,0} + 3 \gamma_{2,0} + \frac{1}{2} \gamma_{2,2}\right) \rho^2 - \left(\gamma_{3,1} \gamma_{2,0} + \gamma_{1,3}\right) \rho + \frac{\gamma_{2,2}}{\gamma_{2,0}},$$

$$V_{db}^{b,f,n} = \frac{V_{b,f,n}}{\xi_{1,1}}.$$

(2.14)

Maximum likelihood estimator (MLE). In statistical theory, maximum likelihood estimator (MLE) is consistent and achieves the minimal variance among all unbiased estimators.

**Theorem 2.2.3.** Let $\{Q(x_i), y_i\}, i = 1, \ldots, k$ be i.i.d. from (2.1), and $Q$ be a $b$-bit quantizer with borders $-\infty = t_0 < \cdots < t_M = \infty$, $r_i = \{j : t_{j-1} \leq x_i < t_j\}$. The MLE $\hat{\rho}_{b,f,ML}$ solves equation

$$\sum_{i=1}^{k} \mathbb{I}_{\{r_i \notin (1,M)\}} \frac{\alpha_i(t_{r_i}) - \alpha_i(t_{r_{i-1}})}{\beta_i(t_{r_i}) - \beta_i(t_{r_{i-1}})} + \mathbb{I}_{\{r_{i=1}\}} \frac{\alpha_i(t_{r_i})}{\beta_i(t_{r_i})} - \mathbb{I}_{\{r_{i=M}\}} \frac{\alpha_i(t_{r_{i-1}})}{1 - \rho^2} \Phi\left(\frac{t_{r_{i-1}} - \rho y_i}{\sqrt{1 - \rho^2}}\right) = 0,$$

where $\alpha_i(t) = (\rho t - y_i) \phi\left(\frac{t_{r_{i-1}} - \rho y_i}{\sqrt{1 - \rho^2}}\right)$, $\beta_i(t) = (1 - \rho^2)^{3/2} \Phi\left(\frac{t_{r_{i-1}} - \rho y_i}{\sqrt{1 - \rho^2}}\right)$. We provide the variance of $\hat{\rho}_{b,ML} = \frac{V_{b,ML}}{k}$ in the Appendix because the expression is rather complicated.

The variance factor $V_{b,ML}$ can be solved by simulation or numerical methods. Note that,
MLE is mainly for the theoretical interest, which could help us understand the minimal variance that can be achieved in the limit.

2.2.2 Benefits of Normalized Estimators and Debiased Estimators

Figure 2.1 plots the variances for two debiased estimators $\hat{\rho}_{b,f}^{db}, \hat{\rho}_{b,f,n}^{db}$ and the MLE. This comparison is reasonable since MLE is asymptotically unbiased. To illustrate the benefits of normalization, we see that the variance of the normalized estimator, when both compared under the unbiased estimation regime, is always smaller than the simple estimator (see the right panel for variance ratio), and substantially so as $\rho$ away from zero. Moreover, we observe that the variance of normalized estimator is very close to that of MLE, implying that normalized estimators are almost optimal in achieving the minimal possible variance.

To elaborate on the benefit of debiased estimators, we evaluate the Mean Square Errors (MSE): bias$^2 +$ variance. Given the benefit of normalization, we consider the two normalized estimators:

$$MSE(\hat{\rho}_{b,f,n}) = \left(1 - \sqrt{\xi_{1,1}}\right)^2 \rho^2 + \frac{V_{b,f,n}}{k} + O\left(\frac{1}{k^2}\right), \quad MSE(\hat{\rho}_{b,f,n}^{db}) = \frac{V_{b,f,n}}{\xi_{1,1}} + O\left(\frac{1}{k^2}\right).$$

Thus, to compare the MSE, it suffices to examine the ratio: $\xi_{1,1} + k\rho^2 \frac{\xi_{1,1}(1-\sqrt{\xi_{1,1}})^2}{V_{b,f,n}}$, which
will be larger than 1 quickly as \( k \) increases. Note that \( \xi_{1,1} \leq 1 \) but it is very close to 1 when \( b \geq 3 \). In summary, the MSE of the debiased estimator quickly becomes smaller as \( k \) increases, leading to higher statistical efficiency.

### 2.2.3 Analysis of Mis-ordering Probabilities in Similarity Search

In similarity search problem, the estimates of inner products are subsequently used for ordering data vectors to identify the nearest neighbor for a given query data point. Intuitively, a more accurate estimator should provide a more accurate ordering. To quantify this argument, we provide a precise analysis of the “mis-ordering” probabilities defined as below.

**Definition 2.2.1. (Mis-ordering probability)** Suppose \( u_1, u_2, u_3 \in \mathbb{R}^d \) are three data points (with \( u_1 \) being a query point) with unit norm and pair-wise cosine similarity \( \rho_{12}, \rho_{13} \) and \( \rho_{23} \), respectively. For an estimator \( \hat{\rho} \), the probability of mis-ordering is defined as

\[
P_M(u_1; u_2, u_3) = Pr(\hat{\rho}_{12} > \hat{\rho}_{13} | \rho_{12} < \rho_{13}).
\]

This probability essentially controls the accuracy of similarity search. Consider a case where \( u_3 \) is the nearest point of \( u_1 \) in the data space (which implies \( \rho_{12} < \rho_{13} \)). If the cosine estimation gives \( \hat{\rho}_{12} > \hat{\rho}_{13} \), we then make a wrong decision that \( u_3 \) is not the nearest neighbor of \( u_1 \). Theorem 2.2.4 builds the connection of mis-ordering probability with debiased cosine estimator variance.

**Theorem 2.2.4. (Asymptotic mis-ordering)** Suppose \( u_1, u_2, u_3 \in \mathbb{R}^d \) are three data points (with \( u_1 \) being a query point) on a unit sphere with pair-wise inner products \( \rho_{12}, \rho_{13} \) and \( \rho_{23} \), respectively. Denote two estimators \( \hat{\rho} \) and \( \hat{\rho}' \) based on \( k \) random projections such that as \( k \to \infty \), the normality \( \hat{\rho} \sim N(\alpha \rho, \hat{\sigma}_\rho^2) \) and \( \hat{\rho}' \sim N(\alpha' \rho, \hat{\sigma}'_\rho^2) \) hold, with constants \( \alpha, \alpha' > 0 \). Denote \( \delta_\rho^2 = \frac{\hat{\sigma}_\rho^2}{\alpha^2} \), \( \delta'_\rho^2 = \frac{\hat{\sigma}'_\rho^2}{\alpha'^2} \) and the correlations \( C = \text{corr}(\hat{\rho}_{12}, \hat{\rho}_{13}), C' = \text{corr}(\hat{\rho}'_{12}, \hat{\rho}'_{13}) \).
respectively. If

\[ \delta'_{\rho_{12}} = a \delta_{\rho_{12}}, \quad \delta'_{\rho_{13}} = a' \delta_{\rho_{13}}, \quad C - aa'C' < \frac{(1 - a^2)\delta_{\rho_{12}}^2 + (1 - a'^2)\delta_{\rho_{13}}^2}{2\delta_{\rho_{12}} \delta_{\rho_{13}}}, \]  

(2.15)

with some \( 0 < a < 1, \quad 0 < a' < 1 \), then as \( k \to \infty \) we have \( \hat{P}_M(u_1; u_2, u_3) > \hat{P}'_M(u_1; u_2, u_3) \), where \( \hat{P}_M(u_1; u_2, u_3) \) and \( \hat{P}'_M(u_1; u_2, u_3) \) are the mis-ordering probability of \( \hat{\rho} \) and \( \hat{\rho}' \), respectively.

**Remark 2.2.1.** Condition (2.15) basically assumes that the variance of the debiased \( \hat{\rho}' \) is smaller than that of the debiased \( \hat{\rho} \) at \( \rho_{12} \) and \( \rho_{13} \), respectively. In a special case where \( a = a' \) and \( C = C' \), the last constraint in (2.15) reduces to \( C < \frac{\delta_{\rho_{12}}^2 + \delta_{\rho_{13}}^2}{2\delta_{\rho_{12}} \delta_{\rho_{13}}} \), which always holds since the RHS is greater than 1.

The important message given by Theorem 2.2.4 is that estimators with lower “debiased variance” (\( \delta \)) tend to have lower mis-ordering probability, which leads to a more accurate estimation of nearest neighbors in the original data space. Since the result holds for general cosine estimators, it also applies to our QRP approach. For example, in many search applications, the “target similarity” \( \rho_{13} \) is usually high. From Figure 2.1, we see that the normalized estimator typically has smaller debiased variance with \( \rho \) close to 1 than simple estimators. Thus, when using QRP for similarity search tasks, we expect that normalizing the QRP would improve the accuracy. This will be justified in our numerical experiments.

### 2.2.4 Variance-Optimal Quantizer Design

Back to Figure 2.1, while the variance of \( \hat{\rho}^{db}_{b,f,n} \) is close to that of MLE, there is a substantial room for the simple estimator to improve. Given the connection between debiased estimation variance and similarity search performance (see Section 2.2.3), we design a new quantizer that achieves point-wise minimal \( \hat{\rho}^{db}_{b,f} \), specifically for the simple cosine estimator. Given a
fixed \( \rho \), consider the optimization system

\[
Q_{\rho, \text{opt}} = \arg\min_{Q} D_{\rho, \text{opt}}, \quad \text{with} \quad D_{\rho, \text{opt}} \triangleq V_{b, f}^{\text{db}} + \rho^2 \frac{\rho^2 \mathbb{E}[Q(x)^2 x^2] + (1 - \rho^2) \mathbb{E}[Q(x)^2]}{\mathbb{E}[Q(x)]},
\]

which directly minimizes the (debiased) variance (2.9). The solution to (2.16) is provided in the following. Note that, \( Q_{\rho, \text{opt}} \) is different at different \( \rho \). Since we are interested in high similarity region, we focus on \( \rho \geq 0 \) for conciseness.

**Proposition 2.2.5.** Let \( D_{\rho, \text{opt}} \) and \( Q_{\rho, \text{opt}} \) be defined by (2.16). For each code \( \mu_i \) and border \( t_i \), a stationary point is given by

\[
\mu_i = \frac{\lambda \int_{t_{i-1}}^{t_i} x \phi(x) dx}{2(\rho^2 \int_{t_{i-1}}^{t_i} x^2 \phi(x) dx + (1 - \rho^2) \int_{t_{i-1}}^{t_i} \phi(x) dx)},
\]

\[
t_i = \frac{\lambda - \sqrt{\lambda^2 - 4\rho^2(1 - \rho^2)(\mu_i + \mu_{i+1})^2}}{2\rho^2(\mu_i + \mu_{i+1})}.
\]

In addition, \( t_i = \frac{\mu_i + \mu_{i+1}}{\lambda} \) when \( \rho = 0 \), and \( t_i = \frac{\lambda}{\mu_i + \mu_{i+1}} \) when \( \rho = 1 \). Here \( \lambda > 0 \) controls the overall scale of quantizer \( Q_{\rho, \text{opt}} \) and is independent of \( D_{\rho, \text{opt}} \).

We still use the iterative approach in Algorithm 1 to update the quantizer till convergence. For conciseness, the detailed implementation is omitted. Empirically, a moderate \( \lambda \in \{1, 2, 3\} \) is suggested to produce a nice-looking quantizer. An interesting observation is that, by setting \( \rho = 0 \) and \( \lambda = 2 \), \( Q_{0, \text{opt}} \) becomes the classical LM quantizer (1.4), which indicates that LM quantizer actually achieves minimal \( \text{Var}(\hat{\rho}_{b, f}^{\text{db}}) \) at \( \rho = 0 \) (minimizes \( D_{0, \text{opt}} \)).

We justify the point-wise optimality of \( Q_{\rho, \text{opt}} \) in Figure 2.2. The \( x \)-axis is the target \( \rho^* \) used for constructing \( Q_{\rho^*, \text{opt}} \). Each curve corresponds to \( \text{Var}(\hat{\rho}_{b, f}^{\text{db}}) \) when true cosine is \( \rho \) and \( Q_{\rho^*, \text{opt}} \) is used to construct the estimator. We observe that the red stars (minimal \( \text{Var}(\hat{\rho}_{b, f}^{\text{db}}) \) for all \( \rho \) are achieved at \( \rho^* = \rho \). The variance reduction of using \( Q_{\rho, \text{opt}} \) in \( \hat{\rho}_{b, f}^{\text{db}} \) gets more substantial as \( \rho \to \pm 1 \) (see Figure 2.1).

The quantizer \( Q_{\rho, \text{opt}} \) can be applied to similarity search problems in practice, when we have some prior information on the range of cosine between the nearest neighbor points.
Figure 2.2: Variance factor $V_{b, f}$ when using refined $Q_{\rho, opt}$ for $b = 2, 3, 4$, at different $\rho$.

This may be acquired based on expertise experience, or some information about the data. For instance, we we know that the cosine between most data points and their nearest points are in $[0.7, 0.9]$, then by Theorem 2.2.4, adopting $Q_{0.8, opt}$ would be a good choice to improved the search accuracy using simple QRP estimates.

### 2.3 Scenario 2: Quantization with Different Bits

We now consider the more general case (Scenario 2) where the data vectors are LM quantized with different numbers of bits, with the symmetric quantization a special case. That is, given observations $\{x_i, y_i\}$, $1 \leq i \leq n$, we quantize $x_i$ using $b_1$ bits and $y_i$ using $b_2$ bits. Without loss of generality, we assume $b_1 < b_2$. Furthermore, we denote two Lloyd-Max (LM) quantizers as $Q_{b_1}$ and $Q_{b_2}$ and distortion $D_{b_1}$ and $D_{b_2}$, respectively.

#### 2.3.1 The Cosine Estimators

Similar to Scenario 1, we define the simple estimator and the corresponding normalized estimator as

$$
\hat{\rho}_{b_1, b_2} = \frac{1}{k} \sum_{i=1}^{k} Q_{b_1}(x_i)Q_{b_2}(y_i), \quad \hat{\rho}_{b_1, b_2, n} = \frac{\sum_{i=1}^{k} Q_{b_1}(x_i)Q_{b_2}(y_i)}{\sqrt{\sum_{i=1}^{k} Q_{b_1}^2(x_i)} \sqrt{\sum_{i=1}^{k} Q_{b_2}^2(y_i)}}. \quad (2.17)
$$

As one might expect, the analysis will become somewhat more difficult. Similar to the
analysis for Scenario 1, in this section we will use the following notations:

\[ \xi_{\alpha,\beta} = \mathbb{E} \left( Q_{b_1}(x)^{\alpha}x^{\beta} \right), \quad \gamma_{\alpha,\beta} = \mathbb{E} \left( Q_{b_2}(x)^{\alpha}x^{\beta} \right), \quad \zeta_{\alpha,\beta} = \mathbb{E} \left( Q_{b_1}(x)^{\alpha}Q_{b_2}(y)^{\beta} \right), \]

(2.18)

which allow us to express the expectation and variance of \( \hat{\rho}_{b_1,b_2} \) as follows.

\[ \mathbb{E} (\hat{\rho}_{b_1,b_2}) = \zeta_{1,1}, \quad \text{Var} (\hat{\rho}_{b_1,b_2}) = \frac{V_{b_1,b_2}}{k}, \quad V_{b_1,b_2} = \zeta_{2,2} - \zeta_{1,1}^2 \]

(2.19)

The expectation \( \zeta_{1,1} \) can be expressed as an infinite summation, but it appears difficult to be further simplified. Nevertheless, it can be bounded as in Theorem 2.3.1.

**Theorem 2.3.1.** *The following two bounds hold for \( \rho \in [-1, 1] *:

\[
\left| \mathbb{E} (\hat{\rho}_{b_1,b_2}) - (1 - D_{b_1})(1 - D_{b_2})\rho \right| \leq \Delta_1, \quad \text{and}
\]

\[
\Delta_2 - \Delta_1 \leq |\mathbb{E} (\hat{\rho}_{b_1,b_2}) - \rho| \leq \Delta_1 + \Delta_2, \quad \text{where}
\]

\[
\Delta_1 = \sqrt{D_{b_1}D_{b_2}(1 - D_{b_1})(1 - D_{b_2})|\rho|^3}, \quad \Delta_2 = (D_{b_1} + D_{b_2} - D_{b_1}D_{b_2})|\rho|. \]

(2.20)

(2.21)

**Remark 2.3.1.** *When \( b_2 \to \infty \) (i.e., Scenario 1), we have \( D_{b_2} \to 0 \) and the bound reduces to an equality \( \mathbb{E} (\hat{\rho}_{b_1,\infty}) = (1 - D_{b_1})\rho \), which matches the result in Section 2.2.*

Equation (2.21) provides upper and lower bounds for the absolute bias of \( \hat{\rho}_{b_1,b_2} \). When \( b_1 = b_2 \) (i.e., the symmetric quantization case), Theorem 2.3.2 presents more refined bounds of the bias of \( \hat{\rho}_{b_1,b_2} \).

**Theorem 2.3.2. (Symmetric quantization)** *Suppose \( b_1 = b_2 = b \). For \( \rho \in [-1, 1] \), we have

\[
(2D_b - D_b^2)|\rho| - D_b(1 - D_b)|\rho|^3 \leq |\mathbb{E} (\hat{\rho}_{b,b}) - \rho| \leq (2D_b - D_b^2)|\rho|. \]

(2.22)

**Remark 2.3.2.** *Compared to Li and Slawski (2017), which derived \( |\mathbb{E} (\hat{\rho}_{b,b}) - \rho| \leq 2D_b|\rho| \), our bounds are more tight.*
What about the debiased estimator of $\hat{\rho}_{b_1,b_2}$? It is slightly tricky because $\mathbb{E}(\hat{\rho}_{b_1,b_2}) = \zeta_{1,1}$ cannot be explicitly expressed as $c\rho$ for some constant $c$ (otherwise, the debiased estimator would be simply $\hat{\rho}_{b_1,b_2}/c$). In Theorem 2.3.1, (2.20) implies that the expectation of $\hat{\rho}_{b_1,b_2}$ is close to $(1 - D_{b_1})(1 - D_{b_2})\rho$. Thus, we recommend $\frac{\hat{\rho}_{b_1,b_2}}{(1-D_{b_1})(1-D_{b_2})}$ as the surrogate for the debiased estimator.

Next, we provide the expectation and variance of the normalized estimator in Theorem 2.3.3.

**Theorem 2.3.3. (Normalized estimator)** Let $\xi_{i,j}$, $\gamma_{i,j}$ and $\zeta_{i,j}$ be defined as (2.18). As $k \to \infty$, we have

$$\mathbb{E}(\hat{\rho}_{b_1,b_2,n}) = \frac{\zeta_{1,1}}{\sqrt{\xi_{2,0}^2 \gamma_{2,0}^2}} + O\left(\frac{1}{k}\right), \quad \text{Var}(\hat{\rho}_{b_1,b_2,n}) = \frac{V_{b_1,b_2,n}}{k} + O\left(\frac{1}{k^2}\right), \quad (2.23)$$

$$V_{b_1,b_2,n} = \frac{\zeta_{2,2} - \zeta_{1,1}^2 - \zeta_{1,1} \xi_{3,1} - \zeta_{1,1}^2 \xi_{2,0}^2 - \xi_{2,0}^2 \gamma_{2,0}^2}{\xi_{2,0}^2 \gamma_{2,0}^2} - \frac{\zeta_{1,1} \xi_{1,3} - \zeta_{1,1}^2 \gamma_{2,0}^2}{\xi_{2,0}^2 \gamma_{2,0}^2} + \frac{\zeta_{1,1} \xi_{2,2} - \zeta_{1,1}^2 \xi_{2,0} \gamma_{2,0}}{2 \xi_{2,0}^2 \gamma_{2,0}^2} + \frac{\zeta_{1,1} \xi_{4,0} - \zeta_{1,1}^2 \xi_{2,0}^2}{4 \xi_{2,0}^2 \gamma_{2,0}^2} + \frac{\zeta_{1,1} \gamma_{4,0} - \zeta_{1,1}^2 \gamma_{2,0}^2}{4 \xi_{2,0}^2 \gamma_{2,0}^2}. \quad (2.24)$$

**Remark 2.3.3.** When $b_2 = \infty$, the expected value of $\hat{\rho}_{b_1,b_2,n}$ reduces to that of $\hat{\rho}_{b_1,f,n}$ in Scenario 1. Additionally, we have $\zeta_{1,1} = \zeta_{2,0} \rho$, $\gamma_{2,0} = 1$, and $\gamma_{4,0} = 3$. It is easy to check that the expression of the variance will reduce to the corresponding formula in Theorem 2.2.2.

Also, note that $\xi_{2,0} = 1 - D_{b_1}$, $\gamma_{2,0} = 1 - D_{b_2}$, and $\zeta_{1,1} \approx (1 - D_{b_1})(1 - D_{b_2})\rho$. This means that we can practically use $\frac{\hat{\rho}_{b_1,b_2,n}}{\sqrt{(1-D_{b_1})(1-D_{b_2})}}$ as surrogate for the debiased estimator of $\hat{\rho}_{b_1,b_2,n}$.

We plot the related results in Figure 2.3, which verifies the theories in Theorems 2.3.1, Theorem 2.3.2 and Theorem 2.3.3.

### 2.3.2 Monotonicity of the Inner Product Estimates

In applications such as nearest neighbor retrieval, the order of distances tends to matter more than the exact values. In this context, given an estimator $\hat{\rho}$, one would hope that $\mathbb{E}(\hat{\rho})$ is
Figure 2.3: **Left panel**: the absolute bias (solid curves, in $\log_{10}$ scale) of $\hat{\rho}_{b_1,b_2}$ by simulations, along with the upper bound (red dashed curves) and lower bound (blue dashed curves) in Eq. (2.21). **Middle panel**: the absolute bias of $\hat{\rho}_{b_1,b_2}$ with $b_1 = b_2$ (the symmetric case) along with the upper and lower bounds in Eq. (2.22). **Right panel**: The variance $V_{b_1,b_2,n}$ of the normalized estimator in Theorem 2.3.3.

monotone in $\rho$. This guarantees that asymptotically as $k \to \infty$, the cosine comparison is always correct, i.e. $\hat{\rho}(u, v_1) > \hat{\rho}(u, v_2)$ if $\rho(u, v_1) > \rho(u, v_2)$ ($u$ is a query point). Otherwise (say, $\mathbb{E}[\hat{\rho}]$ decreasing in $\rho$ on $[s, t]$), the comparison of estimated cosine would be wrong for $\rho \in [s, t]$ even with infinite much data. Recall that, in Section 2.1 and Section 2.2, the simple full-precision estimator and the Scenario 1 simple estimator is monotone in $\rho$ in the expectation because $\mathbb{E}(\hat{\rho}_f) = \rho$ and $\mathbb{E}(\hat{\rho}_{b,f}) = \xi_{1,1}\rho$. Naturally, one will ask if the expectation of Scenario 2 simple estimator $\hat{\rho}_{b_1,b_2} = \frac{1}{k} \sum_{i=1}^{k} Q_{b_1}(x_i)Q_{b_2}(y_i)$, is also monotone in $\rho$. This turns out to be non-trivial question.

We solve this important problem rigorously through several steps. Our analysis is general and not restricted to LM quantizers. To do so, we will need the following definition of “increasing quantizer”.

**Definition 2.3.1. (Increasing quantizer)** Let $Q$ be an $M$-level quantizer with boarders $t_0 < \cdots < t_M$ and reconstruction levels $\mu_1, \ldots, \mu_M$. We say that $Q$ is an increasing quantizer if $\mu_1 < \cdots < \mu_M$.

To proceed, we will exploit the decomposition of quantizers. Specifically, we prove the following three Lemmas for increasing quantizers.

**Lemma 2.3.4. (1-bit vs. others)** Suppose $Q_{b_1}, Q_{b_2}$ are increasing quantizers symmetric
about 0, with \( b_1 \geq 1 \), and \( b_2 = 1 \). Then \( \mathbb{E}(Q_{b_1}(x)Q_{b_2}(y)) \) is strictly increasing in \( \rho \) on \([-1, 1]\).

**Lemma 2.3.5. (2-bit vs. 2-bit)** Suppose \( Q_{b_1}, Q_{b_2} \) are any two increasing quantizers symmetric about 0, with \( b_1 = b_2 = 2 \). Then \( \mathbb{E}(Q_{b_1}(x)Q_{b_2}(y)) \) is strictly increasing in \( \rho \) on \([-1, 1]\).

**Lemma 2.3.6. (Universal decomposition)** For any increasing discrete quantizer \( Q_b, b \geq 3 \) which is symmetric about 0, there exist a 2-bit symmetric increasing quantizer \( Q_2 \) and a \((b-1)\)-bit symmetric increasing quantizer \( Q_{b-1} \) such that \( Q_b = Q_{b-1} + Q_2 \).

With the above lemmas, we are ready to prove the monotonicity of \( \mathbb{E}(Q_{b_1}(x)Q_{b_2}(y)) \).

**Theorem 2.3.7. (Monotonicity)** For any increasing quantizers \( Q_{b_1} \) and \( Q_{b_2} \) symmetric about 0 with bits \( b_1 \geq 1 \) and \( b_2 \geq 1 \), the function \( \mathbb{E}(Q_{b_1}(x)Q_{b_2}(y)) \) is increasing in \( \rho \).

Recall that, in Section 2.2.3, we have proved the result for the mis-ordering probability, i.e., Theorem 2.2.4, which actually assumes estimators have expectations monotone in \( \rho \). Therefore, Theorem 2.3.7 provides the necessary proof to support the assumption in Theorem 2.2.4 for Scenario 2 estimators.

### 2.4 Empirical Study: Similarity Search

In this section, we test the proposed estimators on similarity search tasks on 3 datasets from the UCI repository (Dua and Graff, 2017). The summary statistics is given in Table 2.1. Note that, the “Mean 1-NN \( \rho \)” metric is the average cosine similarity of each data point to its nearest neighbor, which could be set as the target \( \rho \) level when constructing the variance optimal quantizer \( Q_{\rho, opt} \). The experiments clearly confirm the benefit of using \( Q_{\rho, opt} \) (in Scenario 1) and QRP normalization on similarity search accuracy. The results also, to an extent, illustrate the influence of mis-ordering probability studied in Theorem 2.2.4 on search problems.
### Table 2.1: Datasets used in the empirical study.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># samples</th>
<th># features</th>
<th># classes</th>
<th>Mean $\rho$</th>
<th>Mean 1-NN $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arcene</td>
<td>200</td>
<td>10000</td>
<td>2</td>
<td>0.63</td>
<td>0.86</td>
</tr>
<tr>
<td>Isolet</td>
<td>1560</td>
<td>617</td>
<td>26</td>
<td>0.53</td>
<td>0.77</td>
</tr>
<tr>
<td>COIL20</td>
<td>1440</td>
<td>1024</td>
<td>20</td>
<td>0.61</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Mean $\rho$ is the average pair-wise cosine similarity for sample pairs. Mean 1-NN $\rho$ is the average cosine similarity of each point to its nearest neighbor.

For each dataset, all the examples are preprocessed to have unit norm. Each data sample $u_i, i = 1, \ldots, n$ is treated as a query point, and other samples act as the candidates in the database from which the nearest neighbor of $u_i$ is retrieved. For every data sample, we first apply random projection (RP), and then quantize the resulting linear sketches by Lloyd-Max (LM) scheme. The simple estimators are realized by taking the inner product of the QRPs. To use normalized estimators, we normalized the corresponding QRPs per sample.

The evaluation metric we adopt is the **1-NN precision**, which is the proportion of nearest neighbors (NN) we can recover from the nearest neighbor, averaged over all the examples.

**Results.** We summarize the search accuracy in Figure 2.4 (Scenario 1) and Figure 2.5 (Scenario 2). First of all, we can see that, in both scenarios, as the number of bits increases, the performance of LM quantized estimators converges to those of the full-precision estimators, which is as expected. Importantly, the normalization step of the estimators substantially improves the performances, by comparing Column 2 with Column 1 in both figures. In addition, we validate the advantage of $Q_{\rho, opt}$ for Scenario 1 QRP estimators. This, to some extent, also justifies the assertions in Theorem 2.2.4 that smaller variance of debiased estimators could improve NN recovery precision. More precisely,

- **Scenario 1 simple estimator:** In Figure 2.1 (left panel), we see that the variance of debiased estimate $\hat{\rho}_{b,f}^{db}$ with $b = 1$ is much smaller than using $b \geq 2$ in high similarity region (e.g. $|\rho| > 0.85$), while roughly the same at $\rho = 0.75$. Since *Arcene* and *COIL20* have high mean 1-NN $\rho$ (0.86 and 0.93 respectively), Theorem 2.2.4 implies that the cosine estimation of $\hat{\rho}_{1,f}^{db}$ should (in general) have smaller mis-ordering
probability than that with $b \geq 2$, leading to higher 1-NN precision. On the other hand, the average 1-NN $\rho$ of *Isolet* is 0.77, so $\hat{\rho}_{b,f}$ with all $b = 1, 2, ..., \infty$ would likely give similar performance. These claims are consistent with Column 1 of Figure 2.4.

- **Scenario 1 normalized estimator:** The variance of the debiased normalized estimator $\hat{\rho}_{b,f,n}^d$ (Figure 2.1, middle panel) decreases as $b$ increases, uniformly for any $\rho$. Hence by Theorem 2.2.4 we expect the 1-NN precision should increase with larger $b$. This is confirmed by Column 2 of Figure 2.4, on all three datasets.

- **Scenario 1 refined quantizer $Q_{\rho,opt}$:** In Column 1 of Figure 2.4, we see that the refined quantizer $Q_{\rho,opt}$ significantly improves the search accuracy when simple estimates are used. For datasets with high mean 1-NN $\rho$ (*Arcene* and *COIL20*), $Q_{0.9,opt}$ outperforms $Q_{0.5,opt}$ because the former is targeted at the region near $\rho = 0.9$. On *Isolet*, the result is opposite since the mean 1-NN $\rho$ becomes much smaller. Therefore, as we mentioned previously, in practice we should choose the target $\rho$ for $Q_{\rho,opt}$ depending on different applications.

For Scenario 2, the conclusions are very similar. In particular, $Q_{\rho,opt}$ can also empirically improves Scenario 2 simple estimator. To conclude this chapter, our empirical study corroborates our theoretical findings, and further shows that

1. With $b = 4$, using QRP in similarity search yields same accuracy as using full-precision random projections.

2. Normalizing the QRPs substantially improves the search accuracy.

3. Adopting the proposed variance optimal quantizer $Q_{\rho,opt}$ in Scenario 1 (quantization vs. full-precision) can also elevate the performance of similarity search with QRPs.

Therefore, using quantized random projections (QRP) is a memory efficient and effective strategy in similarity search applications. In some cases, the simplest 1-bit QRP can outperform full-precision RP in terms of search accuracy.
Figure 2.4: Scenario 1: Nearest neighbor search recovery results using cosine similarity and quantized estimators, from random projections. Columns 1: $\hat{\rho}_{b,f}$. Column 2: $\hat{\rho}_{b,f,n}$. 
Figure 2.5: Scenario 2: Nearest neighbor search recovery results using cosine similarity and quantized estimators, from random projections. Columns 1: $\hat{\rho}_{b_1,b_2}$. Column 2: $\hat{\rho}_{b_1,b_2,n}$. 
In recent years, with the rapid growth of capable electronic devices, protecting data privacy in the transmission process has been an important topic in statistics and machine learning. For example, people may speculate from public datasets about the private information of individual users, companies or government agencies. Even for non-public data, an adversary may still break into the database to get confidential information. This calls for the need of private machine learning, or private data release. Basically, the goal is to transfer or release “manipulated” data to confuse potential adversaries from extracting useful private information, while still providing good utility for statistical inference or learning tasks.

In this context, adding noise to the data is a natural solution to this problem. Typically, as more noise is introduced, the original information of the data would be harder to speculate. However, this would also cast difficulty on the learning models to retrieve truly useful information. Consequently, one of the most critical topics in data privacy literature is to find the “best” noise adding strategy (in terms of various possible metrics), to achieve high utility under certain privacy constraints.

There are multiple ways to define and evaluate the data privacy. Besides some attack-specific methods (e.g., $k$-anonymity (Samarati and Sweeney, 1998)), a classical approach is based on information-theoretic analysis. Suppose the original data follows a probability distribution $f$, and the perturbed data has distribution $\tilde{f}$. One can then measure the privacy preserving power by metrics such as Mutual Information (MI), Kullback-Leibler(K-L) divergence and Fisher Information (FI) (Wang et al., 2014; Farokhi and Sandberg, 2017; Salamatian et al., 2015; Murguia et al., 2018) between $f$ and $\tilde{f}$. In particular, one would like to make the original data ($f$) and the perturbed data ($\tilde{f}$) as dissimilar as possible, under some utility constraints.
Another popular scheme in privacy analysis is called Differential Privacy (DP) (Dwork et al., 2014), which has gained boosting attention in machine learning community recently. The intuitive description of DP is as follows: we want to find a randomized data output procedure, such that a small change in the database can hardly be detected by an adversary. In differential privacy, adding noise is also a common practice, where Laplacian mechanism and Gaussian mechanism are two most popular methods. The design of DP noise adding methods (e.g., the noise level) largely relies on the sensitivity of the algorithm, which quantifies the largest possible change in the output when the database is modified slightly. In most cases, differential privacy provides more stringent privacy guarantee than information-theoretic approaches, which protects the data from any type of attacks, regardless of how the adversary utilizes the data subsequently. Another attractive property of DP is that, multiple private mechanisms and ingredients can be easily composed together, still with nice DP guarantee.

In this section, we consider the privacy of quantized random projection (QRP). We will consider two distinct aspects: (i) adding noise to the QRP to enhance its privacy; and (ii) invoking differential privacy techniques to propose a DP-QRP release strategy. More specifically, we will first use an information-theoretic approach to identify the optimal noise adding strategy to QRPs based on Mutual Information (MI) minimization. Then, we will discuss the differentially private data release through RPs and QRPs, where the utility is evaluated by Euclidean distance estimation.

3.1 Enhancing the Privacy of QRP: An Information-Theoretic Approach

We have random projections $X = UW$. Every single projection $x = w^T u$ follows i.i.d. $N(0, 1)$. Given a quantizer $Q$, the QRP $z = Q(x)$ follows a discrete distribution with probability mass function (p.m.f.)

$$P_{Z,i} \triangleq P[z_i = \mu_i] = \int_{t_{i-1}}^{t_i} \phi(x)dx, \quad i = 1, ..., 2^b,$$
where $\mu_i$’s are the reconstruction levels of a quantizer. In this subsection, we consider improving the privacy of QRP by noise addition. We will borrow the ideas from information theory literature to find the optimal noise adding method (with some distributional assumptions), under the constraint on the noise magnitude.

3.1.1 Optimal Noise by Mutual Information Minimization

Consider two discrete random variables, $S$ and $T$, supported on $\mathcal{S}$ and $\mathcal{T}$ with p.m.f. as $p_S(s)$ and $P_T(t)$, respectively. We denote $p(s, t)$ as the joint mass function. We define the entropy and mutual information as follows.

**Definition 3.1.1.** The entropy of a discrete random variable $S$ is:

$$H[S] = -\sum_{s \in \mathcal{S}} p_S(s) \log p_S(s).$$

The joint entropy of $S$ and $T$ is:

$$H[S, T] = -\sum_{s \in \mathcal{S}} \sum_{t \in \mathcal{T}} p(s, t) \log p(s, t).$$

The conditional entropy is defined as:

$$H[S|T] = -\sum_{s \in \mathcal{S}} \sum_{t \in \mathcal{T}} p(s, t) \log p(s|t).$$

The Mutual Information (MI) between $S$ and $T$ is defined as

$$I(S; T) = \sum_{s \in \mathcal{S}} \sum_{t \in \mathcal{T}} p(s, t) \log \frac{p(s, t)}{p_S(s)p_T(t)}.$$

Conceptually, $I(S; T)$ measures the dependence between $S$ and $T$, and is always non-negative. In particular, $I(S; T) = 0$ if and only if $S$ and $T$ are independent. The method of our approach is simple: we add i.i.d. discrete mean-zero noise $R$ to each QRP vector,
denoted as $Z = Q(x_i^T W)$. To get tractable analytical solution, we need to constrain the support of $R$. Here, we consider $b$-bit uniform quantizer, and assume that $R$ has the same support as $Z$, such that the perturbed QRP s can still be coded by integers resulting from a uniform quantizer.

Our goal is to minimize the mutual information between the QRP and the perturbed QRP, $I(Z + R; Z)$, under the constraint that $\mathbb{E}[\|R\|^2] \leq \eta$ (i.e., bounded noise variance). In this way, the perturbed data is not very different from the truth, while the dependence between them is minimized (thus it is harder to speculate about the true data). A known result (Cover, 1999) of mutual information is that, for independent $Z = (Z_1, ..., Z_k)$ and $R = (R_1, ..., R_k)$, it holds that

$$I(Z + R; Z) = \sum_{i=1}^{k} H[Z_i + R_i] - H[Z_i].$$

Hence, it suffices to consider a single QRP. We finally arrive at our objective function to find the distribution of $R$ such that

$$\min_{p_R(r)} \quad J_R \triangleq H[Z + R] - H[Z],$$

$$\text{s.t.} \quad \mathbb{E}[R^2] \leq \eta.$$ 

Denote the probability $P_{R,i} = P[R = \mu_i], i = 1, ..., 2^b$. It can be seen that the support of $Y \triangleq Z + R$ is expanded to $\tilde{\mu} = \{-2\mu_1, -2\mu_1 + \Delta, ..., 2\mu_{2^b} - \Delta, 2\mu_{2^b}\}$, with $|\tilde{\mu}| = 2^{b+1} - 1$. Thus, the perturbed QRP can still be regarded as quantized and coded by memory-efficient techniques. By standard computation, we have

$$P_{Y,j} \triangleq P[Y = \tilde{\mu}_j] = \begin{cases} \sum_{i=1}^{j} P_{Z,j+1-i} P_{R,i}, & j = 1, ..., 2^b \\
\sum_{i=j+2^b}^{2^b} P_{Z,j+1-i} P_{R,i}, & j = 2^b + 1, ..., 2^{b+1} - 1. \end{cases}$$
We can now write the objective function as

\[
\min_{PR(r)} H[Z + R] - H[R] = - \sum_{j=1}^{2^{b+1}-1} P_{Y,j} \log P_{Y,j} + \sum_{j=1}^{2^b} P_{R,j} \log P_{R,j}.
\]

The following result states that the optimization problem is convex. Thus, it can be solved efficiently by convex optimization tools.

**Theorem 3.1.1.** \( J_R \) is convex in \( P_{R,i} \), \( \forall i = 1, \ldots, 2^b \).

The proof of Theorem 3.1.1 and all results in this chapter can be found in Appendix B.

### 3.1.2 Example

In Figure 3.1, we show examples on the distribution of obtained noise \( R \) and the perturbed output \( Y = Z + R \), under different constraint \( \eta \) on the noise variance. 3-bit and 4-bit uniform quantizer with largest border as 4 is tested. As we can see, the noise \( R \), when its variance is constrained to be small (i.e., small \( \eta \)), concentrates around 0. As we allow more variance, the noise spreads out. Consequently, when \( \eta \) is small, the distribution of perturbed output \( Y \) is close to that of original \( Z \) (though the support is changed due to the addition of noise). With larger \( \eta \), the distribution of \( Y \) becomes “flat”, providing less information about the true distribution of \( Z \).
Figure 3.1: The probability mass function of QRP ($Z = Q(X), X \sim N(0, 1)$), optimal noise $R$ and the perturbed output $Y = Z + R$, at different variance constraint $\eta$. The first column is 3-bit uniform quantizer, and the second column is 4-bit uniform quantizer, both with maximal border $4$. In the $P(Y)$ plots, the red lines are the p.m.f. of the original QRP, $Z$, and the blue lines represent the p.m.f. of $Y$. 
3.2 Differential Privacy of QRP

We now consider the differential privacy property of RP and QRP. Our goal is to publish the data matrix through RP or QRP, with strict differential privacy guarantee. This published data matrix can then be used for subsequent computations (e.g., distance estimation) with privacy. Following the convention, we will consider unnormalized data, and focus on the goal of Euclidean distance estimation. We have a database containing \( n \) samples, denoted as \( U = [u_1, ..., u_n]^T \), where \( u_i \in \mathcal{U} \). We assume that all the features are bounded between \([-1, 1]\), such that \( \mathcal{U} = [-1, 1]^p \). Denote \( m_i = \|u_i\|^2 \). To study the distance estimator, we look at a pair of data samples, \( u_1 \) and \( u_2 \). Let \( d = \|u_1 - u_2\|^2 = m_1 + m_2 - 2u_1^T u_2 \).

Differential privacy (DP) provides a rigorous mathematical concept to measure the amount of information leakage of an algorithm.

**Definition 3.2.1** (Differential Privacy (Dwork et al., 2014)). For a randomized algorithm \( \mathcal{A} : \mathcal{U} \mapsto \mathbb{R}^k \), if for any two adjacent data sets \( U \) and \( U' \) with symmetric difference one, \( \Pr[\mathcal{A}(U) \in O] \leq e^\epsilon \Pr[\mathcal{A}(U') \in O] + \delta \) holds for any \( O \subset \text{Range}(\mathcal{A}) \) and \( \epsilon, \delta \geq 0 \), then algorithm \( \mathcal{A} \) is called \((\epsilon, \delta)\)-differentially private.

There are various ways to define the “adjacent databases”, which, do not lead to essentially different privacy analysis. In this paper, we will use one of the standard definitions that two data vectors are said to be adjacent if \( u \) and \( u' \) only differ in one dimension, and \( \|u - u'\| = 1 \). Note that this implies \( \|u\|^2 - \|u'\|^2 \leq 1 \). One of the nice facts about DP is that, multiple DP algorithms can be composed together, where the final output is still DP.

**Theorem 3.2.1.** (Advanced Composition Theorem (Dwork et al., 2010)) For any \( \epsilon', \delta', \tilde{\delta} \geq 0 \), the adaptive composition of \( k \) \((\epsilon', \delta')\)-differentially private mechanisms is \((\epsilon, \delta)\)-differentially private with

\[
\epsilon = k\epsilon'(e^{\epsilon'} - 1) + \epsilon' \sqrt{2k \log(1/\tilde{\delta})},
\]

\[
\delta = k\delta' + \tilde{\delta}.
\]
3.2.1 Differential Privacy of Random Projection

We now propose our differentially private random projection (DP-RP) method in Algorithm 2, where Gaussian noise is added to the projected data.

**Algorithm 2** DP Random Projection (DP-RP)

**Input:** $n \times p$ matrix $U$; Privacy parameters $0 < \epsilon, \delta < 1$; Number of projections $k$.

**Output:** Differentially private random projections

1. Build a random $p \times k$ projection $W$, where $W_{ij} \sim N(0, 1)$.
2. Generate random noise matrix $\Sigma \in \mathbb{R}^{n \times k}$, with i.i.d. $N(0, \sigma^2)$ entry where $\sigma = \sqrt{\frac{\log \frac{2k}{\delta}}{2k \log 2}}$.
3. Publish perturbed projected data $X = UW + \Sigma$

**Theorem 3.2.2.** Algorithm 2 achieves $(\epsilon, \delta)$-differential privacy.

Note that, in Algorithm 2, the Gaussian noise adding procedure is exactly what classical “Gaussian mechanism” in differential privacy do. How does our analysis compare with the result obtained from the standard Gaussian mechanism? To start with, we first introduce a key definition.

**Definition 3.2.2** ($l_2$ Sensitivity). The $l_2$ sensitivity of a function $h : U \mapsto \mathbb{R}^k$ is defined by

$$\Delta = \max_{u, u' \in U, u' \in \text{Adj}(u)} \| h(u) - h(u') \|_2.$$

The $l_2$ sensitivity bounds the maximal change in $l_2$ norm of the output, when the databases is modified slightly. Equipped with the sensitivity, we restate the standard Gaussian mechanism as below.

**Theorem 3.2.3.** ([Dwork et al. (2014)](Gaussian Mechanism)) Suppose an algorithm $A : U \mapsto \mathbb{R}^k$ outputs $A(U)$ for a data $U$. Let $\epsilon, \delta \in (0, 1)$. Denote $\Delta$ as the $l_2$ sensitivity of $A$. Then, if we add i.i.d. Gaussian noise from $N(0, \sigma^2)$ with $\sigma = \frac{\Delta \sqrt{2 \log \frac{1}{\delta}}}{\epsilon}$ to each output element, then the perturbation mechanism is $(\epsilon, \delta)$-DP.
Applying Theorem 3.2.3 to our problem setting, we obtain Theorem 3.2.4 as the consequence of classical DP Gaussian mechanism.

**Theorem 3.2.4** (RP-Gaussian Mechanism). Let $\epsilon, \delta \in (0, 1)$. In Algorithm 2, by adding $N(0, \sigma^2)$ i.i.d. Gaussian noise matrix $\Sigma \in \mathbb{R}^{n \times k}$ to each projected sketch, publishing the perturbed projected data $\tilde{X} = UW + \Sigma$ is $(\epsilon, \delta)$-DP, provided that

$$\sigma \geq \sqrt{k + 2 \sqrt{k \log \frac{\epsilon^{2.25}}{\delta} + 2 \log \frac{\epsilon^{2.25}}{\delta}} \sqrt{2 \log \frac{\epsilon^{2.25}}{\delta}}}.$$ 

In Figure 3.2, we compare the noise level required by our method (Algorithm 2), with the noise level induced by classical analysis of Gaussian mechanism (Theorem 3.2.3). As we can see, the $\sigma$ of our approach is significantly smaller than that required by Gaussian mechanism, especially in the high privacy region (i.e., $\epsilon \to 0$). This is consistent with the recent findings of (Balle and Wang, 2018) that the noise level obtained by classical Gaussian mechanism analysis is sub-optimal in the high privacy region, due to its proof strategy based on a “loose” sufficient condition. Moreover, as $\epsilon \to \infty$ increases, the $\sigma$ in Algorithm 2 scales as $\Theta(\frac{1}{\sqrt{\epsilon}})$, which aligns with the order of the Gaussian noise in low privacy region established theoretically in Balle and Wang (2018).

![Figure 3.2](image)
3.2.2 Differential Privacy of Quantized Random Projection

Algorithm 3 DP Quantized Random Projection (DP-QRP)

**Input:** $n \times p$ matrix $U$; Privacy parameters $0 < \epsilon, \delta < 1$; Number of projections $k$.

**Output:** Differentially private random projections

Build a random $p \times k$ projection $W$, where $W_{ij} \sim N(0,1)$.

Generate random noise matrix $\Sigma \in \mathbb{R}^{n \times k}$, with i.i.d. $N(0,\sigma^2)$ entry where $\sigma = \sqrt{\log \frac{2}{\delta} \sqrt{2k \log \frac{2}{\delta}}}$.

Perturb the projected data $X = UW + \Sigma$.

Publish the quantized RP $Z = Q(X)$ \hspace{1cm} //quantization step

Now, we consider discretizing the projected data by quantization, as depicted by Algorithm 3. Since $\mathcal{U} = [-1,1]^p$, the norm of each data point is upper bounded by $\sqrt{p}$. This implies that, each projected sketch, after adding perturbation, has variance at most $(p + \sigma^2)$.

In this subsection, we in particular consider a uniform quantizer with the largest borders of $Q$ as $\pm \sqrt{2 \log \frac{2k}{\delta} \sqrt{p + a^2}}$. If a projected value falls outside of this range, we set it as the nearest reconstruction level depending on its sign. Theorem 3.2.5 states that by quantizing the perturbed RP, DP-QRP release (Algorithm 3) enhances the privacy compared with DP-RP.

**Theorem 3.2.5.** Algorithm 3 is $(\epsilon_Q, \delta)$-DP, where $\epsilon_Q < \epsilon$ as in Theorem 3.2.2.

This result has several implications:

- The smaller privacy parameter of DP-QRP indicates that, it is more difficult for an adversary to distinguish a database from its adjacent neighbor.

- Although the $\epsilon_Q$ is non-trivial to be analytically quantified and depends on the specific quantizer, $\epsilon_Q < \epsilon$ means that we can actually reduce $\sigma$ (add less noise) to achieve same $(\epsilon, \delta)$-DP compared with Algorithm 2. Fewer bits allows us to add less noise. In an extreme case, we have

**Theorem 3.2.6.** 1-bit (sign) random projection achieves perfect differential privacy, i.e., it is $(0,0)$-DP.
The proof is trivial by noticing that \( \int_0^\infty f_x(z)dz = \frac{1}{2} = \int_0^\infty f_x'(x)dz \), and same argument holds for the negative part. Note that, 1-bit QRP can only be used for cosine estimation, since it drops all the norm information of the data. The utility of 1-bit cosine estimation has been extensively studied in Chapter 2. Next, we will consider using QRP for distance estimation.

### 3.2.3 Utility of Euclidean Distance Estimation

Recall the notation \( u_i^T W \) as the randomly projected vector. Denote \( \|u_1 - u_2\|^2 = d \). With (non-DP) random projections, define the distance estimator of \( d \) as

\[
\hat{d}_{RP} = \frac{1}{k}\|u_1^T W - u_2^T W\|^2. \tag{3.1}
\]

Classical results reveal that

\[
\mathbb{E}[\hat{d}_{RP}] = d, \quad \text{Var}[\hat{d}_{RP}] = \frac{2}{k}d^2. \tag{3.2}
\]

In the following, we consider the utility of the distance estimators from Algorithm 2 and Algorithm 3. With a little abuse of notation, we will from now on use \( x_i = u_i^W + \Sigma_i \) to denote the perturbed projected vector, where \( \Sigma_i \sim N(0, \sigma^2 I_k) \) is the independent Gaussian noise. \( z_i = Q(x_i) \) is noted as the quantized projection vector. Correspondingly, define the estimators of DP-RP and DP-QRP as

\[
\hat{d}_{DP-RP} = \frac{1}{k}\|x_1 - x_2\|^2 - 2\sigma^2 - 2\sigma^2; \tag{3.3}
\]

\[
\hat{d}_{DP-QRP} = \frac{1}{k}\|z_1 - z_2\|^2 - 2\sigma^2 - 2\sigma^2. \tag{3.4}
\]

The utility of distance estimation based on differentially private random projection is given in Theorem 3.2.7.
**Theorem 3.2.7.** For DP-RP estimator (3.3), we have

\[
\mathbb{E}[\hat{d}_{DP-RP}] = d,
\]

\[
\text{Var}[\hat{d}_{DP-RP}] = \frac{1}{k}(2d^2 + 8\sigma^2d + 8\sigma^4).
\]

Recall that, we use a uniform quantizer with largest borders \( \pm \sqrt{2 \log \frac{2k}{\delta} \sqrt{p + a^2}} \). We define the following “re-scaled” quantizer as follows.

**Definition 3.2.3.** For a quantizer \( Q \) with borders \( t = [t_0, ..., t_2b] \) and reconstruction levels \( \mu = [\mu_1, ..., \mu_{2b}] \), define \( Q/\alpha \) as the scaled quantizer with borders \( t/\alpha \) and reconstruction levels \( \mu/\alpha \).

**Theorem 3.2.8.** Let \( d = \|u_1 - u_2\|^2 \), and \( \|u_1\|^2 = m_1, \|u_2\|^2 = m_2 \). Let \( m_1 = m_1 + \sigma^2, m_2 = m_2 + \sigma^2 \) and \( \bar{\rho} = \frac{u_1^Tu_2}{\sqrt{m_1m_2}} \). For \( (x, y) \sim N(0, \begin{pmatrix} 1 & \bar{\rho} \\ \bar{\rho} & 1 \end{pmatrix}) \), denote \( \zeta_{i,j} = \mathbb{E}[Q/\sqrt{m_1}(x)^iQ/\sqrt{m_2}(y)^j] \). Then, for DP-QRP estimator (3.4),

\[
\mathbb{E}[\hat{d}_{DP-QRP}] = \zeta_{2,0}m_1 + \zeta_{0,2}m_2 - 2\zeta_{1,1}\sqrt{m_1m_2},
\]

\[
\text{Var}[\hat{d}_{DP-QRP}] = \zeta_{4,0}m_1^2 + \zeta_{0,4}m_2^2 + 6\zeta_{2,2}m_1m_2 - 4(\zeta_{3,1}m_1^{3/2}\sqrt{m_2} + \zeta_{1,3}\sqrt{m_1m_2^{3/2}} - \mathbb{E}[\hat{d}_{DP-QRP}]^2).
\]

As one would expect, using more bits for quantization would increase the utility. How many bits are recommended? In general, we can roughly quantify the goodness of quantization by the number of reconstruction levels used for the items (perturbed projections \( Z \)) with smallest magnitude. Since the variance of \( Z \) falls in the range of \([\sigma^2, p + \sigma^2]\), we can choose the number of quantization intervals, \( N \), that covers \([0, \sigma]\), to control the approximation precision of quantization. In general, \( N \geq 2 \) would be a fairly good choice in practice. As the window size of \( Q \) is \( \sqrt{2 \log \frac{2k}{\delta} \sqrt{p + \sigma^2/2b}} \), the condition implies...
\[ N \cdot \sqrt{2 \log \frac{2k}{\delta}} \sqrt{p + \sigma^2/2^b} \geq \sigma, \text{ leading to} \]

\[ b \geq \log_2 \frac{N \sqrt{2 \log \frac{2k}{\delta}} \sqrt{p + \sigma^2}}{\sigma} + 1 \tag{3.5} \]

which could be used in practice for good utility.

We simulate a pair of data vectors, \((u_1, u_2)\) in \([-1, 1]^{2500}\), with fixed norm \(\|u_1\| = 5\) and varying \(l_2\) distance. We set \(\epsilon = 1\) and \(\delta = 0.1\), and validate the theoretical mean and variance of Theorem 3.2.7 and Theorem 3.2.8 in Figure 3.3, where the theory aligns with the empirical observation. The trend is the same for other \(\|u_1\|\) values. We observe that, the relative variation of DP-RP and DP-QRP for estimating small distances is larger, since the quantity to be estimated becomes smaller compared with the magnitude of Gaussian noise. According to (3.5), when \(k = 1000\) and for \(N = 2\) and \(N = 3\), \(b\) is required to be at least 4.8 and 5.4, respectively. Indeed, we see that DP-QRP with \(b = 6\) suffices to match the utility of DP-RP. In this case, with same estimation utility, the storage cost is reduced, and the privacy of the released data is also improved.
Figure 3.3: The $\sqrt{\mathbb{E}[d]}$ and $\text{Var}[\hat{d}]$ of DP-RP and DP-QRP distance estimators, $\epsilon = 1$, $\delta = 0.1$. Dash curves are the theoretical values of Theorem 3.2.7 and Theorem 3.2.8, which overlap with the empirical curves.
CHAPTER 4

GENERALIZATION ERROR ANALYSIS OF QUANTIZED COMPRESSION LEARNING

As we discussed in Chapter 2, the product of random projections (RP) between two data samples is an unbiased estimator of their cosine, when the data space is normalized to the unit sphere. A more general statement on the approximation guarantee of RP is given by the famous Johnson-Lindenstrauss (JL) Lemma (Johnson and Lindenstrauss, 1984). In short, under some conditions, we can always project a set of $n$ points $X \in \mathbb{R}^{n \times d}$ in a high-dimensional space onto a lower $k$-dimensional space such that pair-wise distances are approximately preserved (within $1 \pm \epsilon$ error), with high probability. Recall that $k \ll d$ is the number of random projections (linear sketches). This nice theoretical result has originated the study of learning in the reduced dimensional space instead of the original space. This line of work is called Compressive Learning (Garg et al., 2002; Calderbank et al., 2009; Maillard and Munos, 2009; Durrant and Kabán, 2013; Kabán, 2014; Slawski, 2017; Thanei et al., 2017; Kabán, 2015). Intuitively, as the pairwise distances among the data points are approximately preserved, learning algorithms based on the data geometry would still work in the lower-dimensional projected space (and approximate learning in the original data space). The benefit is the significantly reduced data dimensionality, from $d$ to $k$, which may usually translate into considerable savings in storage and computation.

In Chapter 2, we have shown theoretically and practically that quantized random projection (QRP) can be very conveniently and effectively applied to cosine estimation and similarity search problems. However, as stated in Section 1.1.3, those are the most considered application of QRP in machine learning community so far. In this chapter, we exploit the new (and natural) Application 1.1.2 of QRP, extending the idea of compressive learning: training classification or regression models using QRPs (i.e., in the quantized projected
space). We call the studied problem as Quantized Compressive Learning. The merit of our approach is again straightforward: using a few bits to represent and store the linear sketches brings favorable storage saving in large-scale databases. In particular, we study in detail the generalization error bounds of applying QRP in three learning models: nearest neighbor (NN) classifier, linear classifier and least squares regression, all of which are popular and important models in a wide range of applications. Empirical experiments are also provided to justify the theory.

An important implication of our analysis is to answer the following question—Which factors of a quantizer determine the generalization performance when the model is trained using QRPs? Our theoretical analysis illustrates that, for nearest neighbor and linear classification, the learning performance is determined by the variance of debiased inner product estimator when data samples are allocated on the unit sphere. For regression problems, the distortion of a quantizer becomes crucial. Our theoretical findings are validated by empirical study. Practically, our results also suggest appropriate quantizing strategies for different learning models, which would be helpful in various applications. Proofs of theoretical result in this chapter are placed in Appendix C.

4.1 Preliminaries

Problem setting. The problem setup is basically the same as in Chapter 2. We will use $X$ to denote the data matrix, instead of $U$. Assume dataset $(X, Y) \sim D^n$ with $X = [x_1, ..., x_n]^T \in \mathbb{R}^{n \times d}$ and $Y = [y_1, ..., y_n]$, where $x_i, i = 1, ..., n$ are i.i.d. drawn from some marginal data distribution $D_m$ and $y_i$ is the label of the $i$-th sample. As before, we assume that every sample in $X$ is standardized to have unit Euclidean norm\(^1\). Therefore, the domain of $X$ is a subset of the unit Euclidean sphere $S^{d-1}$, which allows us to call “inner product” and ”cosine” interchangeably. For classification problems, $Y \in [0, 1]^n$, while in regression model $Y \in \mathcal{R}^n$. We will focus on the Gaussian random projection matrix

---

\(^1\)Again, our analysis in this chapter can be adapted to cases without data normalization.
$R = [r_1, ..., r_k] \in \mathbb{R}^{d \times k}$ with i.i.d. standard normal entries. Random projection is realized by $X_R = \frac{1}{\sqrt{k}} X R$, where the scaling factor $\frac{1}{\sqrt{k}}$ is for the ease of presentation.

The Lloyd-Max (LM) quantization for Gaussian RP is derived in the same way as in Section 2.1.2, i.e., with respect to the standard normal distribution. Yet, in this chapter our analysis applies to more general quantizers not restricted to the Lloyd-Max (LM) scheme. Let $Q(\cdot)$ be a general quantizing function that operates element-wise on matrix, with borders $t_0 < \cdots < t_M$. The quantized random projection (QRP) of data $X$ is thus defined as $X_Q = \frac{1}{\sqrt{k}} Q(X R)$. We are interested in learning with $X_Q$ instead of $X$.

The following definition will be used in our analysis.

**Definition 4.1.1. (Maximal gap)** For an $M$-level quantizer $Q$ and an interval $[a, b]$, denote $\alpha = \{i : t_{i-1} < a \leq t_i\}$ and $\beta = \{i : t_i < b \leq t_{i+1}\}$. The maximal gap of $Q$ on $[a, b]$ is defined as the largest distance between any two nearby borders on $[a, b]$, $g_Q(a, b) = \max\{\max_{i, \alpha \leq i \leq \beta - 1} |t_{i+1} - t_i|, |t_{\alpha} - a|, |b - t_{\beta}|\}$ if $t_{\alpha} \in [a, b]$, and $g_Q(a, b) = |b - a|$ otherwise.

This chapter has deep connection with Chapter 2, in particular with the cosine estimators studied therein. Now, we formally specify the definition of debiased variance.

**Definition 4.1.2. (Debiased variance)** Denote the space of expectation of a cosine estimator $\hat{\rho}_Q$ as $\mathcal{E}$. If there exists a map $g : [-1, 1] \to \mathcal{E}$ given by $\rho \mapsto \mathbb{E}[\hat{\rho}_Q | \rho]$, the debiased estimator is defined by applying the inverse map $\hat{\rho}_Q^{db} = g^{-1}(\hat{\rho}_Q)$ to correct for the bias. The variance of $\hat{\rho}_Q^{db}$ is called the debiased variance.

Note that, debiasing the cosine estimators might not be realistic in practice. Definition 4.1.2 is mainly for analytical purposes.

### 4.2 Quantized Compressive Nearest Neighbor Classification

We first look at the generalization error incurred by learning using $X_Q$ instead of $X$ on nearest neighbor (NN) classification problem, which is a simple but powerful non-parametric algorithm that is popular in practice. Given a dataset $S = (X, Y) \sim \mathcal{D}^n$ and a test sample
Let \((x, y) \sim \mathcal{D}\) where \(y\) is unknown, the algorithm finds the nearest neighbors of \(x\) in \(X\), denoted by \((x^{(1)}, y^{(1)})\), and classifies \(u\) as \(\hat{y} = y^{(1)}\). We denote the classifier of NN as \(h_S(x) = y^{(1)}\), in the original sample space. Denote the conditional distribution of \(y\) given \(x \sim \mathcal{X}\) as \(\eta(x) = P(y = 1 | x)\). A Bayes classifier, \(h^*(x) = \mathbb{1}\{\eta(x) > 1/2\}\), is well known as the optimal solution in minimizing the risk \(L(h(x)) = \mathbb{E}_u[\mathbb{1}\{h(x) \neq y\}]\) over all hypothesis. (Cover and Hart, 1967) showed that the risk of NN classifier converges to \(2L(h^*(x))\) as sample size \(n \to \infty\). See additional asymptotic analysis in (Fritz, 1975; Wagner, 1971; Györfi and Györfi, 1978). In finite \(n\) case, (Shalev-Shwartz and Ben-David, 2014; Gottlieb et al., 2014; Chaudhuri and Dasgupta, 2014) studied the error bounds and convergence rate of NN classifier, all of which require the sample size \(n\) increases exponentially in dimensionality \(d\), under some Lipschitz-type assumptions on the conditional probability function \(\eta(x)\). As discussed in (Shalev-Shwartz and Ben-David, 2014; Kabán, 2015), by the celebrated No-Free-Lunch Theorem (Wolpert and Macready, 1997), this exponential sample complexity comes from the nature of this problem and cannot be reduced in general.

### 4.2.1 Finite Sample Analysis

In the finite sample setting, Kabán (2015) demonstrates that when data has small “metric size” measured by metric entropy integral \(\gamma\) (which will be defined later), it is possible to reduce the sample complexity from \(\tilde{O}(e^d)\) to \(\tilde{O}(e^{\gamma^2})\) by working in the randomly projected space using \(X_R\), i.e., training the NN classifier with RPs. This is called compressive NN classification. The following definitions are necessary for our analysis.

**Definition 4.2.1.** Let \((\mathcal{T}, \| \cdot \|)\) be a totally bounded metric space, and \(\alpha > 0\). \(\mathcal{T}\) is \(\alpha\)-separated if \(\forall a - b \in \mathcal{T}, a \neq b, \|a - b\| \geq \alpha\) holds. The \(\alpha\)-packing number of \(\mathcal{T}\) is \(N_{\| \cdot \|}(\alpha, \mathcal{T}) = \max\{|T'| : T' \text{ is } \alpha\text{-separable}, T' \subset \mathcal{T}\}\).

**Definition 4.2.2.** The \(\alpha\)-entropy of \(\mathcal{T}\) is defined as \(Z(\alpha, \mathcal{T}) = \log N(\alpha, \mathcal{T})\), and the function \(Z(\cdot, \mathcal{T})\) is called the metric entropy of \(\mathcal{T}\).
Theorem 4.2.1. (Klartag and Mendelson, 2005). Let $\mathcal{X} \subset \mathbb{R}^d$, and $R \in \mathbb{R}^{d \times k}$ a random matrix with i.i.d. Gaussian or Rademacher entries with mean 0 and variance $\sigma^2$. $T = \{ \frac{a - b}{\|a - b\|} : a, b \in \mathcal{X} \}$ be the set of all pair-wise normalized chords. Define metric entropy integral as $\gamma(T) = \int_0^1 \sqrt{Z(\alpha, T)} d\alpha$, then there exists an absolute constant $c$, such that for all $\omega, \delta \in (0, 1)$, if $k \geq c\omega^{-2}(\gamma(T)^2 + \log(2/\delta))$, then with probability at least $1 - \delta$, we have $R$ is $\omega$-isometry on $\mathcal{X}$, namely,

\[ (1 - \omega)k\sigma^2\|s - t\|^2 \leq \|RTs - RTt\|^2 \leq (1 + \omega)k\sigma^2\|s - t\|^2, \forall s, t \in \mathcal{X}. \]

Theorem 4.2.1 is a generalization of Johnson-Lindenstrauss (JL) Lemma, which characterizes the probability of getting a “good” projection matrix with nice isometry property. By a careful analysis under a slightly different assumption on the domain $\mathcal{X}$ (unit sphere), we present the generalization bound on compressive NN classifier (learning with $X_R$) in Kabán (2015) as follows.

Theorem 4.2.2. $X \sim \mathcal{X}^n$, $Y \sim \{0, 1\}^n$ with $X = [x_1, ..., x_n]^T \in \mathbb{R}^{n \times d}$ and $\mathcal{X}$ is the unit sphere. Assume that $\eta(x) = Pr(y = 1|x)$ is $L$-Lipschitz. Let $R \in \mathbb{R}^{d \times k}$, $k < d$ a random matrix with i.i.d. Gaussian entries following $N(0, 1)$. $(x, y)$ is a test sample with unknown $y$. Denote $(u_R^{(1)}, y_R^{(1)}) \in (X, Y)$ the training sample such that $\frac{1}{\sqrt{k}}R^Tu_R^{(1)}$ is the nearest neighbor of $\frac{1}{\sqrt{k}}RTu$ in the projected space, and the compressive NN classifier $h_R(x) = y_R^{(1)}$. Denote $\mathcal{L}(h^*)$ the risk of Bayes classifier. Let $T$ be defined as Theorem 4.2.1. Then for all $\omega, \delta \in (0, 1)$, if $k = O(\omega^{-2}(\gamma(T)^2 + \log(2/\delta)))$, with probability $1 - \delta$ over random draws of $R$, we have the risk of compressive NN classifier

\[ \mathbb{E}_{X,Y}[\mathcal{L}(h_R(x))] \leq 2\mathcal{L}(h^*(x)) + 2\sqrt{2}(L\sqrt{\frac{1 + \omega}{1 - \omega}})\frac{1}{\sqrt{k}}(ne)^{-\frac{1}{4\sqrt{r}}} \sqrt{k}. \] (4.1)

Equipped with above tools, we are now ready to state our first result on the risk of uniformly quantized compressive nearest neighbor classifier, with finite $n$ and $k$. 
Theorem 4.2.3. (Finite sample analysis) Let $X, Y, R$ and $\eta(x)$ be the same as in Theorem 4.2.2. $Q$ is a $b$-bit uniform quantizer with bin width $\triangle$. Suppose $(x, y)$ is a test sample with unknown $y$. Denote $(x_Q^{(1)}, y_Q^{(1)}) \in (X, Y)$ the training sample such that $\frac{1}{\sqrt{k}} Q(R^T x_Q^{(1)})$ is the nearest neighbor of $\frac{1}{\sqrt{k}} Q(R^T x)$ in the quantized projected space, and the quantized compressive NN classifier $h_Q(x) = y_Q^{(1)}$. Let $\mathcal{T}$ be defined as Theorem 4.2.1. Then $\forall \omega, \delta \in (0, 1)$, if $k = O(\omega^{-2}(\gamma(\mathcal{T})^2 + \log(2/\delta)))$, $[-\sqrt{1 + \omega}, \sqrt{1 + \omega}] \subset [t_0, t_2]$ and the maximal gap $g_Q = g_Q(-\sqrt{1 + \omega}, \sqrt{1 + \omega}) < 2\sqrt{1 + \omega}$, then with probability $1 - \delta$ over random draws of $R$, the risk of quantized compressive NN classifier is bounded by

$$\mathbb{E}_{X,Y}[L(h_Q(x))] \leq 2L(h^*(x)) + 2\sqrt{2(\frac{L\triangle}{g_Q} \sqrt{\frac{1 + \omega}{1 - \omega}})(ne)^{-\frac{1}{k+1}} \sqrt{k} + \frac{2L\triangle \sqrt{k}}{\sqrt{1 - \omega}}}, \quad (4.2)$$

where $L(h^*)$ is the risk of Bayes rule.

Remark 4.2.1. The assumption that $Q$ is uniform quantizer is mainly for the ease of presentation. For an arbitrary quantizer, the bound also holds with $\triangle$ replaced by a more complicated and quantizer-specific term. When $\triangle \rightarrow 0$ which means no quantization applied, we have $g_Q = \triangle$ and the bound reduces to (4.1) in Theorem 4.2.2.

The proof involves two interleaving covers of the projected space, which, by Theorem 4.2.1, has bounded diameter with high probability. Now we compare Theorem 4.2.3 with Theorem 4.2.2. We call the second term in (4.2) as the random projection error and the last term as quantization error. We first observe the following sample complexity bound.

Corollary 4.2.1. Under the setting as Theorem 4.2.3, suppose $\triangle \leq \frac{c\sqrt{1 - \omega}}{2L\sqrt{k}} \epsilon$ for some constant $0 < c < 1$. Then we have $\mathbb{E}_{X,Y}[L(h_Q(x))] \leq 2L(h^*(x)) + \epsilon$, if $n \geq \tilde{O}(e^{\gamma(T)^2})$.

Proof. We would like to solve $n$ for

$$2\sqrt{2(\frac{L\triangle}{g_Q} \sqrt{\frac{1 + \omega}{1 - \omega}})(ne)^{-\frac{1}{k+1}} \sqrt{k} + \frac{2L\triangle \sqrt{k}}{\sqrt{1 - \omega}}} \leq \epsilon.$$
By assumption, \( \frac{2L\Delta\sqrt{k}}{\sqrt{1-\omega}} \leq c\epsilon \). Thus we need
\[
2\sqrt{2}\left(\frac{L\Delta}{gQ}\right)\sqrt{\frac{1+\omega}{1-\omega}}^{-\frac{k}{k+1}}(ne)^{-\frac{1}{k+1}}\sqrt{k} \leq (1-c)\epsilon.
\]
Solving for \( n \) and plugging in the required order of \( k \) gives the desired result.

Therefore, when the quantizer is precise enough, by conducting NN classification in the quantized projected space, the sample complexity is reduced from \( \tilde{O}(e^d) \) to \( \tilde{O}(e^{\gamma(T)^2}) \), which essentially relies on the metric entropy of the sample space \( \mathcal{X} \). This quantity is in general smaller than \( d \). Two special cases are:

- When \( \mathcal{X} \) is a finite set of size \( N \), the squared metric entropy \( \gamma(T)^2 = O(\log N) \), implying that \( k \) is required to be \( O(\log N) \), recovering the JL Lemma.

- When the data samples are \( p \)-sparse, we have \( \gamma(T)^2 \leq O(p\log \frac{d}{p}) \) (Boucheron et al., 2013), which makes the sample complexity in Corollary 4.2.1 be \( \tilde{O}(\frac{d}{p}e^p) \). This is typically smaller than the exponential dependence \( O(e^d) \) when \( p \) is small.

Regarding the last term in (4.2), the extra quantization error decreases with smaller bin length \( \Delta \), which is reasonable since small \( \Delta \) implies better approximation to the full-precision RP’s in general. Nevertheless, the factor \( \sqrt{k} \) implies that the error incurred by “random projection + quantization” becomes larger as \( k \) increases. Intuitively, however, large \( k \) provides more accurate estimation of the pair-wise cosines (i.e. pair-wise distances since \( \mathcal{X} \subset S^{d-1} \)), which should actually reduce the extra loss as the nearest neighbors would be more precisely estimated. This unsatisfactory pattern of quantization error in Theorem 4.2.3 comes from the finite sample setting and proof methodology, since the bound is a worst case bound with \( n \) and \( k \) both finite. Thus, this bound is less meaningful for practical purposes.
4.2.2 Asymptotic Analysis

We should notice that, the key difference between NN classifier, compressive NN and quantized compressive NN is simply the space in which we search for the neighbors. More importantly, this procedure essentially depends on the distance estimation. Given that $X$ is defined on the unit sphere, finding NN in projected or quantized space is identical to finding $x_i \in X$ that has largest estimated cosine between test example $x$. It is essentially the same task for compressive and quantized compressive NN classification, where in this case, we do not need to care about the specific space from which we derive the estimator, while the statistical property becomes the major concern.

Conceptually, nearest neighbor classification is very similar to nearest neighbor search, while the former also considers the label information. Hence, the next result will be deeply connected with the mis-ordering probability (Definition 2.2.1) introduced in Chapter 2 when analyzing the similarity search accuracy.

**Theorem 4.2.4. (Asymptotic analysis)** Let data $X, Y$ and projection matrix $R$ be same as Theorem 4.2.3. Let $(x, y)$ be a test sample with unknown $y$. $Q$ is any arbitrary quantizer with increasing reconstruction levels. We estimate the cosine between any two points $s, t \in X$ with $\langle s, t \rangle = \rho_{s,t}$ in the quantized space by estimator $\hat{\rho}_Q(s, t)$ with asymptotic normality. Assume that $\forall s, t \sim X$, $E[\hat{\rho}_Q(s, t)] = \alpha \rho_{s,t}$ for some $\alpha > 0$. Denote $(x_Q^{(1)}, y_Q^{(1)}) \in (X, Y)$ the training sample such that $\frac{1}{\sqrt{k}} Q(R^T u_Q^{(1)})$ is the nearest neighbor of $\frac{1}{\sqrt{k}} Q(R^T u)$, and the quantized compressive NN classifier $h_Q(x) = y_Q^{(1)}$. Then we have as $k \to \infty$,

$$E_{X, Y, R}[\mathcal{L}(h_Q(x))] \leq E_{X, Y}[\mathcal{L}(h_S(x))] + r_k,$$

where $r_k = E_{X,x}[\sum_{i: x_i \in G} \Phi\left(\frac{\sqrt{k(\cos(x,x_i) - \cos(x,x_Q^{(1)})}}{\sqrt{\xi_{x,x_i}^2 + \xi_{x,x_Q^{(1)}}^2 + 2\text{Corr}(\hat{\rho}_Q(x,x_i), \hat{\rho}_Q(x,x_Q^{(1)})\xi_{x,x_i}\xi_{x,x_Q^{(1)}})}}\right)]$, with $\xi_{s,t}^2/k$ the debiased variance of $\hat{\rho}_Q(s, t)$ and $G = x/x_Q^{(1)}$. $\mathcal{L}(h_S(x))$ is the risk of data space NN classifier, i.e. $h_S(x) = y^{(1)}$ with $(x^{(1)}, y^{(1)})$ the nearest neighbor of $x$.

**Remark 4.2.2.** In Theorem 4.2.4, we express the bound in terms of $E_{X,Y}[\mathcal{L}(h_S(x))]$ to
highlight the extra quantization error. We can further bound the risk $\mathbb{E}_{X,Y,R}[\mathcal{L}(h_Q(x))]$ by directly adopting any bound on $\mathbb{E}_{X,Y}[\mathcal{L}(h_S(x))]$. The assumption that $\hat{\rho}_Q$ has expectation linear in $\rho$ is mainly for the ease of analytical consideration. Similar result also holds in general situations, under additional minor assumptions.

The bound is intuitive, in the sense that the quantization error term $r_k$ represents the probability of picking a different nearest neighbor in data space and in the quantized projected space, aggregating the mis-ordering probabilities as introduced in Definition 2.2.1. The benefit of Theorem 4.2.4 is that, we factor out $\mathcal{L}(h_S(x))$, instead of $\mathcal{L}(h_R(x))$ as in Theorem 4.2.3. Conceptually, we get rid of the error incurred by searching in the projected space as an intermediate step. Regarding the quantization error term $r_k$ is interesting, note that for $\forall i \in \mathcal{G} = X/x^{(1)}$, $\cos(x, x_i) - \cos(x, x^{(1)}) < 0$ holds. Consequently, when $k \to \infty$, all the $\Phi(\cdot)$ terms in $r_k$ would decrease towards 0 (since $\Phi(t) \to 0$ as $t \to -\infty$). Therefore, we derive a well-behaving quantization error term in the asymptotic case: the extra quantization error indeed decreases with $k$, and the risk of quantized compressive NN converges to that of the data space nearest neighbor classifier, asymptotically.

In the following, we derive a corollary regarding the asymptotic risk of compressive NN classifier, $h_R(x)$, by noticing that the full-precision RP corresponds to applying quantization with infinite bits.

**Lemma 4.2.5.** Suppose $u, s, t \in \mathcal{X}$ are three data points on a unit sphere with inner products $\rho_{us}, \rho_{ut}$ and $\rho_{st}$ respectively. Let full-precision linear estimator $\hat{\rho}_R$ be defined as $\hat{\rho}_R(s, u) = \frac{s^T RR^T u}{k}$ for any $s, u \in \mathcal{X}$. Then we have

$$\text{Cov}(\hat{\rho}_R(u, s), \hat{\rho}_R(u, t)) = \frac{1}{k}(\rho_{su} + \rho_{us}\rho_{ut}).$$

**Corollary 4.2.2.** Let the data $(X, Y), (x, y)$ and projection matrix $R$ be same as Theorem 4.2.3. We estimate the cosine between any two points $s, t \in X$ with $\langle s, t \rangle = \rho_{st}$ in the projected space by $\hat{\rho}_R(s, t) = \frac{s^T RR^T t}{k}$. Denote $(x_R^{(1)}, y_R^{(1)}) \in (X, Y)$ the training sample
such that $R^T x_Q^{(1)}$ is the nearest neighbor of $R^T x$ in the projected space, and the NN classifier $h_R(x) = y_R^{(1)}$. Then, as $k \to \infty$,

$$
\mathbb{E}_{X,Y,R}[\mathcal{L}(h_R(x))] \leq \mathbb{E}_{X,Y}[\mathcal{L}(h_S(x))] + r_k,
$$

where $r_k = \mathbb{E}_{X,x} \left[ \sum_{i: u_i \in G} \Phi \left( \frac{\sqrt{k} \cdot (\cos(x,x_i) - \cos(x,x_i^{(1)}))}{\sqrt{\cos(x,x_i) - \cos(x,x_i^{(1)})^2 + 2(1 - \cos(x,x_i^{(1)}))}} \right) \right]$, with $G = X/x^{(1)}$.

### 4.2.3 Implication to Practice

As mentioned before, the excessive risk $r_k$ in Theorem 4.2.3 involves the debiased variance of the cosine estimator, which is closely related to the mis-ordering probability. Thus, by a similar argument as in Theorem 2.2.4, under mild conditions, for a general estimation approach, smaller debiased variance lowers the bound on the quantization error. Regarding the most influential region, in Theorem 4.2.4, given a fixed $k$ and a query $x$, points near $x^{(1)}$ (i.e. $x_i$ with small $|\cos(x,x_i) - \cos(x,x_i^{(1)})|$) tend to affect the quantization error more substantially due to the sub-exponential tail of Gaussian. Hence, for 1-NN classification, we should choose quantizers with low debiased variance around $\rho^* = \cos(x,x_i^{(1)})$, provided that it can be known (or estimated) a priori. In particular, if a quantized estimator has lower debiased variance than using full-precision RPs (especially around $\rho^*$), then learning with $X_Q$ would outperforms learning with $X_R$ in nearest neighbor classification.

Is there a way to reduce the debiased variance of inner product estimates, for better generalization in NN classification? In Chapter 2, we have shown that normalizing the random projected vectors (i.e. $R^T x_i, i = 1, ..., n$) yields smaller debiased variance, especially in high similarity region (i.e. large $|\rho|$ region). This is exactly the situation for most of the NN classifications where $\rho^* = \cos(x,x_i^{(1)})$ is relatively high. More specifically, for $s, t \in \mathcal{X}$, we can use the normalized estimator

$$
\hat{\rho}_{Q,n}(s,t) = \frac{Q(R^T s)^T Q(R^T t)}{\|Q(R^T s)\|\|Q(R^T t)\|}, \quad (4.3)
$$
for the general estimator \( \hat{\rho}_Q(s, t) \) in Theorem 4.2.4.

### 4.3 Quantized Compressive Linear Classification with (0,1)-loss

In this section, we consider the generalization error for binary linear classifiers, which include some of the most popular learning models such as logistic regression and linear support vector machine (SVM).

We first clarify the problem setting and notations. Let \( \mathcal{H} \) be a hypothesis class of functions on \( \mathcal{X} \to \{0, 1\} \). For the original data, we assume that a function \( H \in \mathcal{H} \) separates the dataset \( S \) by a hyperplane, and classify each side as a distinct class 0/1. Mathematically, for a test data point \( u \), the label returned by \( H \) is

\[
H(u) = \mathbb{1}\{h^T x > 0\},
\]

where \( h \) is a vector in \( \mathbb{R}^d \) and orthogonal to the separating plane. Since all \( u_i \)'s are normalized to unit norm, we may without loss of generality assume that \( h \) also lies on the unit sphere passing though the origin. The optimal classifier, \( \hat{H} \), is the minimizer of (0,1)-loss, defined as

\[
\hat{L}_{(0,1)}(S, h) = \frac{1}{n} \sum_{i=1}^{n} L_{(0,1)}(H(x_i), y_i), \quad L_{(0,1)}(H(x_i), y_i) = \begin{cases} 
0, & \text{if } H(x_i) = y_i, \\
1, & \text{otherwise.} 
\end{cases}
\]

(4.4)

We denote \( \hat{h} \in \mathbb{R}^d \) the learned vector associated with \( \hat{H} \). The model \((\hat{H}, \hat{h})\) is called the empirical risk minimization (ERM) classifier. In the projected space and the quantized projected space, the ERM classifiers are denoted by similar notations with corresponding subscripts as \( \hat{H}_R, \hat{h}_R \in \mathbb{R}^k, \hat{H}_Q \) and \( \hat{h}_Q \in \mathbb{R}^k \), as follows,

\[
\hat{H}(x) = \mathbb{1}\{\hat{h}^T u > 0\}, \quad \hat{H}_R(x) = \mathbb{1}\{\hat{h}_R^T R^T u > 0\}, \quad \hat{H}_Q(x) = \mathbb{1}\{\hat{h}_Q^T Q(R^T u) > 0\}.
\]

(4.5)
4.3.1 Generalization Error Analysis

Now suppose \( x \) is a test sample with unknown class \( y \), we are interested in the probability of making a wrong prediction by training the linear classifier with QRPs,

\[
Pr[\hat{H}_Q(x) \neq y] = \mathbb{E}[\mathcal{L}(\hat{H}_Q(x), y)].
\]

Existing results on such compressive linear classifier have studied bounds on the same type of objective in the projected space, with exact expression in finite \( k \) case (Durrant and Kabán, 2013). Here we look at this problem in the asymptotic domain. When studying the error incurred by learning in the projected space, an important tool is the following definition.

**Definition 4.3.1.** Let \( \hat{h}, x \in \mathbb{R}^d \) be defined above, \( ||\hat{h}|| = ||x|| = 1 \). Let \( \langle \hat{h}, X \rangle = \cos(\hat{h}, x) = \rho > 0 \), and \( \hat{\rho}_R = \frac{\hat{h}^T RR^T x}{k} \). \( R \in \mathbb{R}^{d \times k} \) a i.i.d. standard Gaussian random matrix. The **flipping probability** is defined as

\[
f_k(\rho) = Pr[\rho_R < 0 | \rho > 0].
\]

(4.6)

Intuitively, this quantity measures the probability of changed prediction when we project the data space by \( R \), and use \( R^T \hat{h} \) as the classifier. Durrant and Kabán (2013) gives the exact formula of this quantity,

\[
f_k(\rho) = \frac{\Gamma(k)}{\Gamma(k/2)^2} \int_0^{1/\rho^2} \frac{z^{(k-2)/2}}{(1+z)^k} dz = F_{k,k}(\frac{1 - \rho}{1 + \rho}),
\]

(4.7)

where \( F_{k,k} \) is the cumulative distribution function (CDF) of F-distribution with \((k, k)\) degrees of freedom. This formula also holds for \( \rho < 0 \) by simply plugging in \( \rho = -\rho \). By symmetry, it suffices to consider \( \rho > 0 \). As it is well-known that \( \mathbb{E}[\hat{\rho}_R] = \rho \) and \( \text{Var}[\hat{\rho}_R] = \frac{1 + \rho^2}{k} \), the full-precision simple linear estimator \( \hat{\rho}_R \) should asymptotically follow \( N(\rho, \frac{1 + \rho^2}{k}) \) as
Thus, the asymptotic flipping probability should be

$$\tilde{f}_k(\rho) = \Phi\left(-\frac{\sqrt{k}\rho}{\sqrt{1 + \rho^2}}\right).$$

(4.8)

The following calculation confirms this asymptotic convergence.

**Proposition 4.3.1.** As $k \to \infty$, we have $f_k(\rho) \to \tilde{f}_k(\rho)$ for $\rho > 0$.

![Figure 4.1](image_url)

Figure 4.1: Illustration of sign flipping in quantized space. Points above the black decision boundary are classified as 1, and 0 otherwise. Green dashed lines are boarders of the quantizer. **Left:** data space classifier predicts 1. **Right:** quantized space prediction (using $Q(\hat{h}^T R)$) changes to 0.

For quantized compressive classifier, sign flipping may also happen (an illustrative example is given in Figure 4.1) and can be regarded as the extra loss additional to classifying in the original data space. By analyzing this event, in the following we state the asymptotic generalization error bound for quantized compressive linear classifiers.

**Theorem 4.3.2.** Let the (0,1)-loss and ERM classifier be defined as (4.4) and (4.5). $R \in \mathbb{R}^{d \times k}$ is i.i.d standard normal random matrix. Let $\hat{L}_{(0,1)}(S, \hat{h}) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{(0,1)}(H(x_i), y_i)$ be the empirical loss in the data space. Assume $Q$ is a quantizer and the quantized estimator $\hat{\rho}_Q(s, t)$ is asymptotically normal, has mean $\alpha \rho$ for some $\alpha > 0$ and debiased variance $\xi^2_{\rho}/k$ at $\rho = \cos(s, t)$, $\forall s, t \sim \mathcal{X}$. Given $(u, y)$ a test sample with $y$ unknown, when $k \to \infty$, with
probability at least $1 - 2\delta$ we have

$$
Pr[\hat{H}_Q(x) \neq y] \leq \hat{L}_{(0,1)}(S, \hat{h}) + 2\sqrt{\frac{(k + 1) \log \frac{2en}{k+1} + \log \frac{1}{\delta}}{n}}
+ \frac{1}{n} \sum_{i=1}^{n} f_{k,Q}(\rho_i) + \min \left\{ \sqrt{3 \log \frac{1}{\delta}}, \frac{1}{n} \sum_{i=1}^{n} f_{k,Q}(\rho_i), \frac{1-\delta}{\delta n} \sum_{i=1}^{n} f_{k,Q}(\rho_i) \right\},
$$

where the flipping probability $f_{k,Q}(\rho_i) = \Phi(-\frac{\sqrt{k}\rho_i}{\xi_{\rho_i}})$, with $\rho_i$ the cosine between training sample $u_i$ and ERM classifier $\hat{h}$ in the data space.

In Theorem 4.3.2, the first term is the empirical training loss in the data space, and the second term is the generic sample complexity in VC learning theory. When $b \to \infty$ (full-precision RPs), the bound reduces to that derived in (Durrant and Kabán, 2013) for compressive linear classifier, according to Proposition 4.3.1. One important observation is that the quantization error again depends on the debiased variance of the quantized inner product estimator, averaging over all the data samples at $\rho_i, i = 1, ..., n$.

### 4.3.2 Choice of Quantizer $Q$

Unlike nearest neighbor classifier, the extra generalization error depends more on the region near 0 for linear classifier. To see this, we notice that the flipping probability $f_{k,Q}(\rho_i) = \Phi(-\frac{\sqrt{k}\rho_i}{\xi_{\rho_i}})$ in Theorem 4.3.2 decreases with larger $\rho$. Intuitively, label flipping is much more likely to occur for the points near the decision boundary (i.e. with small $\hat{h}^T u$). As a result, one needs to choose a quantizer with small debiased variance around $\rho = 0$ for linear classification tasks, e.g., logistic regression, linear SVM. Recall that in Chapter 2, Proposition 2.2.5 illustrates that the Lloyd-Max (LM) quantizer achieves minimal variance\(^2\) at $\rho = 0$ in Scenario 1—“Quantization vs. Full-precision” estimation problem.

By similar arguments, we can show that in the symmetric case where both linear sketching vectors are quantized with same $Q$, among all non-trivial quantizers such that $\mathbb{E}[Q(x)x] > 0$

\(^2\)This is also the debiased variance since $\mathbb{E}[\hat{\rho}] = 0$ when $\rho = 0$.\]


\[ x \sim N(0, 1), \]  
LM quantizer also minimizes the (debiased) variance as \( \rho = 0. \) Hence, for linear classification problem, we recommend using Lloyd-Max (LM) quantization.

### 4.4 Quantized Compressive Least Squares Regression

Besides classification problem, linear sketches can also be applied to regression problems. Compressive Least Squares (CLS) regression has been studied in several papers, e.g., Maillard and Munos (2009); Kabán (2014). Slawski (2017) shows that in many cases, CLS can match the performance of Principle Component Regression (PCR), but runs faster by avoiding large scale Singular Value Decomposition (SVD) or optimization, especially on high-dimensional data. In CLS, the projected design matrix \( X_R \), instead of the original \( X \), is used for ordinary least squares (OLS) regression. We are interested in the extra error brought by further quantizing the projections, where \( X_Q \) is used as the new design matrix. We call this approach Quantized CLS (QCLS). In particular, we consider a fix design problem where data \( X \in \mathbb{R}^{n \times d} \) is determinant and \( Y \in \mathbb{R}^n \) are treated as random. Standard OLS regression with Gaussian noise is modeled by

\[ y_i = x_i^T \beta + \epsilon_i, \quad (4.9) \]

with \( \beta \in \mathbb{R}^d \) the parameter to be estimated, and \( \epsilon_i \)'s are i.i.d. Gaussian with mean 0 and variance \( \gamma^2 \).

#### 4.4.1 Generalization Error Analysis

We now formulate the problem in compressed spaces. For projected data and quantized projected data, \( y_i \) is the same while the predictors becomes \( \frac{1}{\sqrt{k}} R^T x_i \) and \( \frac{1}{\sqrt{k}} Q(R^T x_i) \), respectively. The corresponding models are then given by

\[ y_i = \frac{1}{\sqrt{k}} x_i^T R \beta_R + \epsilon_i, \quad y_i = \frac{1}{\sqrt{k}} Q(x_i^T R) \beta_Q + \epsilon_i, \quad (4.10) \]
where $\beta_R$ and $\beta_Q$ are length-$k$ parameter vectors. The squared loss in the data space, the projected space and the quantized projected space are defined respectively as

$$ L(\beta) = \frac{1}{n} \mathbb{E}_Y \| Y - X \beta \|^2, \quad L_R(\beta_R) = \frac{1}{n} \mathbb{E}_{Y|R} \| Y - \frac{1}{\sqrt{k}} X R \beta_R \|^2, \quad L_Q(\beta_Q) = \frac{1}{n} \mathbb{E}_{Y|R} \| Y - \frac{1}{\sqrt{k}} Q(X R) \beta_Q \|^2. \quad (4.11) $$

Note that here the expectation is taken w.r.t. $Y$, and $R$ is given. Denote the true minimizers to above losses as $\beta^* \in \mathbb{R}^d$, $\beta^*_R \in \mathbb{R}^k$ and $\beta^*_Q \in \mathbb{R}^k$, respectively. The risk of an estimator $\hat{\beta}$ in the data space is defined as $r(\beta) = L(\beta) - L(\beta^*)$, and analogously $r_R(\beta_R)$ and $r_Q(\beta_Q)$ can be defined in projected and quantized projected spaces. On the other hand, we have the empirical losses

$$ \hat{L}(\beta) = \frac{1}{n} \| Y - X \beta \|^2, \quad \hat{L}_R(\beta_R) = \frac{1}{n} \| Y - \frac{1}{\sqrt{k}} X R \beta_R \|^2, \quad \hat{L}_Q(\beta_Q) = \frac{1}{n} \| Y - \frac{1}{\sqrt{k}} Q(X R) \beta_Q \|^2, \quad (4.12) $$

which are computed over the training data. The OLS estimates minimize the empirical losses in the associated space, namely,

$$ \hat{\beta}^* = \arg\min_{\beta \in \mathbb{R}^d} \hat{L}(\beta), \quad \hat{\beta}^*_R = \arg\min_{\beta \in \mathbb{R}^k} \hat{L}_R(\beta), \quad \hat{\beta}^*_Q = \arg\min_{\beta \in \mathbb{R}^k} \hat{L}_Q(\beta). \quad (4.13) $$

**Theorem 4.4.1.** Consider the OLS problem defined in (4.9), (4.10) with losses given by (4.11) and (4.12). Suppose all samples in $X$ have unit norm, $\Sigma = X^T X / n$, and $R \in \mathbb{R}^{d \times k}$ are i.i.d. standard normal, with $k < n$. $Q$ is a LM quantizer with distortion $D_Q$ defined as (1.3) w.r.t. standard Gaussian distribution. Further define $\xi_{2,2} = \mathbb{E}[Q(x)^2 x^2]$ with $x \sim N(0, 1)$. The expected QCLS excess risk of the ERM estimator $\hat{\beta}_Q^*$ (4.13) over the minimal loss in the data space is bounded by

$$ \mathbb{E}_{Y,R}[L_Q(\hat{\beta}_Q^*)] - L(\beta^*) \leq \gamma \frac{k}{n} + \frac{1}{k} \| \beta^* \|^2_{\Omega}, \quad (4.14) $$
where $\Omega = \left[ \xi_2^2 \frac{1 + D_Q}{(1 - D_Q)^2} - 1 \right] \Sigma + \frac{1}{1 - D_Q} I_d$, with $\|w\|_\Omega = \sqrt{w^T \Omega w}$ the Mahalanobis norm and $I_d$ the identity matrix with rank $d$.

**Remark 4.4.1.** Lloyd-Max (LM) quantizer is considered for the ease of presentation. Similar result can be derived for general quantizers under extra technical assumptions. When $D_Q = 0$ (no quantization is applied), $\xi_2^2 = 3$ holds and the bound reduces to the classical bound (Kabán, 2014) for CLS.

### 4.4.2 Choice of Quantizer $Q$

In Theorem 4.4.1, if we treat the term $\xi_2$ as a minor constant, the distortion $D_Q$ in general controls the excess risk. In particular, smaller $D_Q$ reduces the expected loss. Intuitively, classification usually allows more noise in the data representation (brought by quantization) than regression. Smaller distortion in general implies less deviation of QRP from the compressed linear sketches, thus improving the generalisation performance of QCLS. Although the bound in Theorem 4.4.1 assumes LM quantizer, we expect similar results for other quantizers, under extra technical conditions. Hence, with a fix number of bits, Lloyd-Max (LM) quantizer, which is built naturally for the purpose of distortion minimization, should be the first choice for QCLS.

### 4.5 Numerical Study

In this section, we validate the theoretical findings through experiments on real-world datasets from UCI repository (Dua and Graff, 2017). Table 4.1 provides summary statistics. All the data samples are normalized to have unit $l_2$ norm. Same as Table 2.1 in Chapter 2, “Mean $\rho$” is the average cosine among all pairs of data points, and “Mean 1-NN $\rho$” is the average cosine of each point to its nearest neighbor.
Table 4.1: Summary statistics of datasets used for experiments in Chapter 2, all standardized to unit norm.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># samples</th>
<th># features</th>
<th># classes</th>
<th>Mean $\rho$</th>
<th>Mean 1-NN $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>arcene</td>
<td>200</td>
<td>10000</td>
<td>2</td>
<td>0.63</td>
<td>0.86</td>
</tr>
<tr>
<td>BASEHOCK</td>
<td>1993</td>
<td>4862</td>
<td>2</td>
<td>0.33</td>
<td>0.59</td>
</tr>
<tr>
<td>orlraws10P</td>
<td>100</td>
<td>10304</td>
<td>10</td>
<td>0.80</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Figure 4.2: Empirical debiased variance of tested quantizers. To be divided by $k$.

4.5.1 Quantized Compressive Classification

The experiment setup for classification problems is as follows. Given a dataset, we first project the data by random Gaussian projection, and then quantize the derived sketches to get the QRPs. We test three quantizers: 1-bit Lloyd-Max (LM) quantizer, 3-bit Lloyd-Max (LM) quantizer and 3-bit uniform quantizer. LM quantizers are optimized w.r.t. standard normal distribution, and the uniform quantizer is symmetric about 0 with $\Delta = 1$, and cut-off points $x = -3.5$ if $x < -3$; $x = 3.5$ if $x > 3$. As discussed in Chapter 2, the debiased variance of $\hat{\rho}_Q = \frac{Q(XR)^T Q(XR)}{k}$ cannot be analytically exactly computed. Here we plot the debiased variance of different quantizers via numerical simulation, which is presented in Figure 4.2. For 1-NN classification, we take each data point as test sample and
the rest as training data over all the examples, and report the mean test accuracy. For linear classifier, we feed the inner product estimation matrix $X_QX_Q^T$ as the kernel matrix into a linear SVM solver (Chang and Lin, 2011). We randomly split the data to 60% training and 40% testing, and the best test accuracy among all hyper-parameter $C$ is reported, averaged over 5 repetitions.

**Nearest Neighbor Classification**

The test accuracy of nearest neighbor classification is provided in Figure 4.3. Theorem 4.2.4 suggests that, small debiased variance around the “mean 1-NN $\rho$” should be beneficial for 1-NN classification. BASEHOCK dataset has mean 1-NN $\rho = 0.59$, the point at which the debiased variance is compared as 1-bit LM>3-bit uniform>full-precision≈3-bit LM. Hence, we see in Figure 4.3 that the NN classification error is consistent with this order on BASEHOCK. On the other hand, the mean 1-NN $\rho$ of arcene and orlraws10P is high (around 0.9). At this point, 1-bit LM quantizer has much smaller debiased variance than others. Therefore, we expect 1-bit LM to provide highest test accuracy on these two datasets, which again agrees with the empirical results. In conclusion, our experiments validate the theoretical results and analysis in Theorem 4.2.4 on the influence of debiased estimator variance on nearest neighbor classification.

**Linear SVM**

In Figure 4.4, we report the test accuracy of linear SVM trained using QRPs. At $\rho = 0$, the debiased variance of estimators using different quantizers are in the order of 1-bit LM>3-bit uniform>3-bit LM>full-precision. Therefore, following the discussion in Theorem 4.3.2, the test errors are expected to be sorted in the same order. This is confirmed by 4.4. In particular, under same 3-bit coding, Lloyd-Max (LM) quantizer outperforms uniform quantizer on all three datasets, validating the superiority of LM quantizers in linear classification tasks.
Figure 4.3: Test accuracy of quantized compressive nearest neighbor classification.

Figure 4.4: Test accuracy of quantized compressive linear SVM.

4.5.2 Quantized Compressive Regression

For this task, we simulate data $X \in \mathbb{R}^{3000 \times 1200}$ and the true parameter $\beta \in \mathbb{R}^{1200}$ from i.i.d $N(0, 1)$, and noise $\epsilon \sim N(0, 0.2)$. We compare uniform quantizers with equal-bit LM quantizers, with $b = 3, 4, 5$. For each $b$, the uniform quantizer is constructed to have same outer reconstruction levels ($\mu_1$ and $\mu_2b$) as the corresponding LM quantizer. The distortions are $(0.035, 0.009, 0.002)$ for LM quantizers and $(0.043, 0.026, 0.019)$ for uniform quantizers, respectively. In Figure 4.5, we see that the order of test MSE perfectly agrees with the comparison of distortion from high to low, and LM quantizers always outperform uniform quantizers with same number of bits. As the distortion gets smaller, the performance of QCLS approaches that of CLS (with full-precision RPs). Again, these observations verify the conclusion in Theorem 4.4.1 that quantizers with smaller distortion generalize better in QCLS problem.
Figure 4.5: Test MSE of QCLS.
CONCLUSION OF PART I

In Part I of the dissertation, we study in detail the compression of the linear sketching method of random projection (RP) by applying scalar quantization to the projected data. By discretizing the full-precision RPs (which take 64 bits per sketch) into only a few bits per sketch, the cost would be significantly reduced to store these random features. In large-scale databases which become more and more important in the era of big data, such memory efficiency could be extremely beneficial. Besides theoretical analysis, our results also bring practical insights on machine learning with these highly compressed random features.

In Chapter 2, we mainly study the statistical properties of the cosine estimators built upon the quantized random projections (QRPs). We consider a very general case allowing two RP vectors to be quantized by different precision (asymmetric quantization), which is practical in real-world applications. We study two scenarios: (i) quantization vs. full-precision, and (ii) quantization with different bits. Simple inner product estimators, normalized estimators and Maximum Likelihood Estimators (MLE) are analyzed and compared. Based on our theoretical results which connect the variance of debiased cosine estimators with similarity search, we also propose a new “variance optimal” quantizer that benefits search problems. Empirically, we conduct similarity search experiments to justify our theoretical findings and show the effectiveness of searching with QRPs in practice.

In Chapter 3, we consider the privacy aspect of QRP. By an information-theoretic approach, we establish an optimal discrete noise adding strategy on top on QRP, with numerical examples. For differential privacy (DP), we study DP-RP and DP-QRP algorithms with strict DP guarantees, by introducing Gaussian noise. The required noise variance is much smaller than that derived from the classical Gaussian mechanism in DP literature. Utility is analyzed in Euclidean distance estimation problem.

In Chapter 4, we study a new application of compressed linear sketches, by applying
them to classification and regression tasks. That is, models are learned in the quantized projected spaces (i.e., with QRPs) instead of the original data space. This is called Quantized Compressive Learning framework. We develop generalization error bounds for three models: nearest neighbor classification, linear classification and linear regression. Our results reveals the connection between generalisation performance of the two classification models with the debiased variance of QRP cosine estimators. Moreover, for regression problems, the generalization performance is closely related to the distortion of the used quantizer. Our analysis provides implications on the proper choice of quantization method for each learning model. Finally, numerical experiments are implemented on above learning tasks that validate our theoretical claims.
Part II

Non-linear Learning with Compressed Random Features
INTRODUCTION TO RANDOM FOURIER FEATURES (RFF)

In Part I, the analysis of theory and applications of quantized random projections (QRPs) is restricted to linear learning. That said, the estimation procedure and learning models considered are all based on the cosine (inner product) distance, i.e. learning with linear kernel. For most real-world datasets, the data features and the response usually exhibit (possibly strong) non-linear relationship. Therefore, it is well-known that by exploiting the non-linearity of the data, learning with non-linear kernels in many cases leads to considerably better performance than linear learning. In Part II of the thesis, we move to the non-linear learning domain. Firstly, in this chapter, we show how randomized non-linear learning is applied in practice, on multi-view discriminant analysis applications.

It has been shown in rich literature that learning with non-linear kernels can elevate the performance of various learning models, e.g., Kernel Ridge Regression (KRR), Kernel Support Vector Machine (KSVM), Kernel Discriminant Analysis (KDA), Kernel Principle Component Analysis (KPCA), etc. In this dissertation, we mainly consider the well-known Gaussian kernel, which is defined between sample $u, v \in \mathcal{X}$ as

$$k_{\gamma}(u, v) = \langle \xi(u), \xi(v) \rangle = e^{-\frac{\gamma^2 \|u-v\|^2}{2}},$$

where $\xi(\cdot) : \mathcal{X} \mapsto \mathbb{H}$ is the implicit feature map and $\gamma$ is a hyper-parameter. $\mathbb{H}$ denotes the associated Reproducing Kernel Hilbert Space (RKHS). Note that from now on, we will use $k$ to denote the kernel function, instead of the number of RPs in previous chapters. We will drop the subscript “$\gamma$” and use $k(\cdot)$ to denote the kernel when everything is clear from the context. Moreover, we will use $K \in [0, 1]^{n \times n}$ to denote the kernel matrix, where $K_{ij} = k(u_i, u_j)$ is the RBF kernel between the $i$-th and $j$-th data sample.

Two major general issues with large-scale non-linear kernel learning arise. Firstly,
storing/materializing a kernel matrix for a dataset of \( n \) samples would need \( n^2 \) entries, which may not be realistic even just for medium datasets (e.g., \( n = 10^6 \)). To avoid this problem, the entries of the kernel matrix can be computed on the fly from the original dataset. This however will increase the computation time, plus storing the original high-dimensional dataset for on-demand distance computations can also be costly. Secondly, the training procedure for nonlinear kernel algorithms is also well-known to be expensive (Platt, 1998; Bottou et al., 2007). Therefore, it has been an active area of research to speed up kernel machines, and using various types of random projections has become popular, e.g., (Li et al., 2006c; Achlioptas, 2001; Rahimi and Recht, 2007; Hsieh et al., 2014).

One popular way to solve the aforementioned limitation of non-linear kernel learning is called \textit{kernel linearization}. Similar to the idea of random projection, we can construct random sketches with the goal to approximate the non-linear kernel among data samples. This scheme is generally named as \textit{non-linear sketching}. By casting trigonometric functions on random projections, the method of \textit{Random Fourier Features} (RFF) (Rahimi and Recht, 2007) is a popular data-independent approach for kernel linearization. It is an application of Bochner’s Theorem (Rudin, 1990), which says that a shift-invariant kernel is positive definite (which is true for RBF kernel) if and only if it is the inverse Fourier transform of a non-negative measure \( \Psi \). For data points \( u, v \in \mathbb{R}^d \), it then holds that

\[
K(u, v) = (\mathcal{F}^{-1}\Psi)(u - v) = \int e^{iw^T(u-v)}d\Psi(w) \\
= \mathbb{E}_{w \sim \Psi}[\cos(w^T(u-v))],
\]

where \( \mathcal{F} \) denotes the Fourier transform operator. Based on trigonometric identities, there are two general formulations of RFF:

\textbf{RFF-I:} \quad F(u) = \sqrt{2}\cos(w^T u + \tau), \quad (4.15)

\textbf{RFF-II:} \quad F(u) = [\sin(w^T u) \; \cos(w^T u)]^T, \quad (4.16)
where $w \sim N(0, \gamma^2 I_d)$, $\tau \sim \text{Uniform}(0, 2\pi)$ for the RBF kernel. Empirically, these two formulations usually perform similarly. The inner product of RFFs admits: $\mathbb{E}[F(u)^T F(v)] = k(u, v)$, i.e., unbiasedness. Hence, by using $m$ independent $w_i$ and $\tau_i$ to generate i.i.d. random features $F_i$, $i = 1, ..., m$, we obtain an unbiased kernel estimator as

$$\hat{k}(u, v) = \frac{1}{m} \sum_{i=1}^{m} F_i(u)^T F_i(v) \approx k(u, v),$$

(4.17)

where we treat each second type of RFF as a 2-dimensional vector. Imposing linear kernel on the RFFs would be equivalent to learning with the RBF kernel on the original data. That means, we can simply feed the RFFs of the dataset into linear learning models, to approximate the learning capacity of non-linear Gaussian kernels. This builds the foundation of approximate non-linear learning with RFF, with numerous applications in clustering, classification, regression, computer vision, etc. (Raginsky and Lazebnik, 2009; Yang et al., 2012; Affandi et al., 2013; Hernández-Lobato et al., 2014; Dai et al., 2014; Yen et al., 2014; Hsieh et al., 2014; Shah and Ghahramani, 2015; Chwialkowski et al., 2015; Richard et al., 2015; Sutherland and Schneider, 2015; Li, 2017b; Avron et al., 2017; Sun et al., 2018; Tompkins and Ramos, 2018).

In Part II, we will address learning and compression with above introduced Random Fourier Features (RFF), of both formulations. In Chapter 5, we first show how RFFs can be used in practice by studying its application on Multi-view Discriminant Analysis (MvDA) problem. Then, in Chapter 2, we study in detail the distortion optimal Lloyd-Max (LM) quantization scheme for RFF. Last, in Chapter 7, we consider a convenient variant of RFF compression, where non-linear features are directly extracted from linear QRPs, addressing an important practical scenario.
CHAPTER 5
RANDOMIZED KERNEL MULTI-VIEW DISCRIMINANT ANALYSIS

5.1 Introduction

Multi-view learning (Kan et al., 2016; Trapp et al., 2017; Xin et al., 2017; Gui and Li, 2018) or learning with multiple different feature sets is a rapidly growing research area with practical success in many important applications. For example, a person can be described by visual light face image, sketch, near infrared face image, iris, fingerprint, palmprint or signature with information secured from many different sources (e.g., distinct angles). Our task is to classify an object from one view, given the information from other views. For instance, Figure 5.1 shows some samples from two different views in the CUFSF multi-view dataset (more detailed dataset description is provided in the experiments section). Here, each person represents a object class, and we would like to classify a sketch given all the label information of the photos, or vise versa. In many cases, the views can be quite different as well. An example is the content-based web-image retrieval, where an object can be identified by the text depicting the image or visual features from the image itself. Here the text and image can be regarded as two distinct views, of the same object.

In this chapter, we focus on multi-view subspace learning, which aims to learn a common subspace shared by all different views. The research on multi-view learning started with “two-view” learning. The canonical correlation analysis (CCA) (Hotelling, 1936; Hastie et al., 2009; Xu and Li, 2019) is perhaps the most well-known two-view unsupervised algorithm. CCA finds the linear projections for two views respectively which have maximum correlation with each other. The discriminative variants of CCA were studied in (Ma et al., 2007; Sun et al., 2008). The paper (Lin and Tang, 2006) provided common discriminant feature extraction to maximize the inter-class separability and meanwhile minimize the intra-
Figure 5.1: Examples from CUFSF multi-view dataset. **View 1 (first row):** actual face photo. **View 2 (second row):** sketch drawn by artist.

class scatter. Multi-view CCA (MCCA) (Nielsen, 2002; Rupnik and Shawe-Taylor, 2010) was proposed to secure one common subspace for all views, under unsupervised setting. Generalized multi-view analysis (GMA) framework (Sharma et al., 2012) took advantage of class information, resulting in a discriminant common space. The authors of (Sim et al., 2009) presented the multi-model discriminant analysis (MMDA) to decompose variations in a dataset into independent modes (factors). The multi-view discriminant analysis (MvDA) was proposed by (Kan et al., 2016), which learned projections for different views jointly via Fisher discriminant analysis. In Cao et al. (2018), the authors proposed a variant called MvMDA, which differs from standard MvDA in the definition of inter-class and intra-class covariance matrices.

In many cases, non-linear kernel has stronger learning capacity than linear kernel. Hence, to enhance performance of standard MvDA algorithm, we seek to kernelize multi-view discriminant analysis, and derive so-called kernel MvDA (KMvDA). However, it is known that a direct implementation of nonlinear kernels is difficult for large-scale datasets, since even for a medium-sized dataset with only 100,000 instances, the $100,000 \times 100,000$ kernel matrix has $10^{10}$ entries, which is essentially not feasible for most machines that people use daily. Therefore, in practical applications, being capable of linearizing nonlinear kernels is highly welcome (Bottou et al., 2007; Li et al., 2011, 2013). Random Fourier features
(RFFs) (Rahimi and Recht, 2007; Li, 2017b) is a celebrated algorithm to linearly approximate Gaussian (RBF) kernel, which has been widely used and studied in literature (Lopez-Paz et al., 2014; Sutherland and Schneider, 2015) on clustering, CCA, PCA, classification, etc. For two data points, the inner product of linearized Fourier features is unbiased estimator of the true RBF kernel. By using faster linearized algorithms, we are able to exploit the learning power of non-linear kernel (e.g., RBF kernel) in linear time when dealing with large-scale datasets.

5.2 Preliminaries

5.2.1 Problem Setting and Notations

In this chapter, for the convenience of notations, we denote a multi-view dataset by \( X = \{x_{ijk} | i = 1, \cdots, c; j = 1, \cdots, v; k = 1, \cdots, n_{ij} \} \) with the instances where \( x_{ijk} \in \mathbb{R}^{d_j} \) is the \( k \)-th instance from the \( i \)-th class of the \( j \)-th view of \( d_j \) dimension, \( c \) denotes the number of classes, \( v \) is the number of views. \( X_j \) represents the instances from the \( j \)-th view. \( n_{ij} \) denotes the number of instances from the \( i \)-th class of the \( j \)-th view, and \( n_i \) is the number of observations from the \( i \)-th class of all views. Let \( n \) denote the total number of examples from all views. Let \( C(x) \) denote the class label of \( x \). Let \( w_1, \ldots, w_v \) denote the view-specific linear projections that we aim to learn. Throughout the chapter, \( \| \cdot \|_F \) denotes matrix Frobenius norm and \( \| \cdot \| \) is the operator norm for matrices and Euclidean norm for vectors.

5.2.2 Multi-view Discriminant Analysis (MvDA)

Multi-view discriminant analysis (MvDA) (Kan et al., 2016) aims to find \( v \) view-specific linear projections \( w_1, w_2, \cdots, w_v \) which can respectively transform the instances from \( v \) views to one discriminant common space, by minimizing the within-class variation and maximizing the between-class variation. Instances from \( v \) views are then projected onto the same common space by \( w_1, w_2, \cdots, w_v \). Note that here the definition of projections is different from the random projections in previous chapters. Figure 5.2 depicts the idea and
Figure 5.2: An illustration of the MvDA framework, which jointly learns a projection for each view and conducts classification in the common subspace.

framework of MvDA. To achieve cross-view discrimination, the within-class variation from all views should be minimized while the between-class variation from all views should be maximized in the common space.

More specifically, MvDA is a generalization of linear discriminant analysis (LDA) (Hastie et al., 2009) for multi-view learning. They share the same type of objective function (i.e., the Rayleigh quotient),

$$ (w_1^*, w_2^*, \ldots, w_v^*) = \arg \max_{w_1, \ldots, w_v} \text{tr} \left( \frac{W^T D W}{W^T S W} \right), $$

where

$$ S = \begin{bmatrix} S_{11} & \cdots & S_{1v} \\ \vdots & \ddots & \vdots \\ S_{v1} & \cdots & S_{vv} \end{bmatrix}, \quad D = \begin{bmatrix} D_{11} & \cdots & D_{1v} \\ \vdots & \ddots & \vdots \\ D_{v1} & \cdots & D_{vv} \end{bmatrix}. $$

Particularly, $S$ and $D$ can be seen as the within-class scatter matrix and between-class scatter matrix for multi-view learning, respectively. The $r$-th column and the $j$-th row block matrix...
of $S$, which is denoted by $S_{jr}$, is defined as:

$$
S_{jr} = \begin{cases} 
\sum_{i=1}^{c} \left( \sum_{k=1}^{n_{ij}} x_{ijk} x_{ijk}^T - \frac{n_{ij} n_{ir}}{n_i} u_{ij}^{(x)} u_{ij}^{(x)^T} \right), & j = r, \\
- \sum_{i=1}^{c} \frac{n_{ij} n_{ir}}{n_i} u_{ij}^{(x)} u_{ir}^{(x)^T}, & \text{otherwise.}
\end{cases}
$$

The term $D_{jr}$ is the $r$-th column and the $j$-th row block matrix of $D$ and defined as

$$
D_{jr} = \left( \sum_{i=1}^{c} \frac{n_{ij} n_{ir}}{n_i} u_{ij}^{(x)} u_{ir}^{(x)^T} \right) - \frac{1}{n} \left( \sum_{i=1}^{c} n_{ij} u_{ij}^{(x)} \left( \sum_{i=1}^{c} n_{ir} u_{ir}^{(x)} \right)^T \right),
$$

with $u_{ij}^{(x)} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} x_{ijk}$.

Basically, MvDA extends LDA to multi-view setting with carefully designed block covariance matrices that aim to achieve accurate multi-view classification. The standard approach for solving the optimization problem (B.1) is by transforming it into a generalized eigenvalue problem, which will be introduced in the next sub-section.

### 5.2.3 Eigenspace Comparison

A standard eigenvalue problem (SEP), given a square matrix $A$, is to solve $Aw = \lambda w$ for vector $w$ and scalar $\lambda$. Feasible $\lambda$’s are called the eigenvalues (or spectrums), and $w$’s are called eigenvectors. For SEP, there exist many well-known results on the eigenvalues (e.g., Weyl’s theorem) when a small perturbation is added to $A$. The Davis-Kahan theorem (e.g., the $\sin \Theta$ theorem) provides bounds on the change of angles between eigenvectors, which could be regarded as a measure of the change of eigenspace. These theorems cast additional restrictions on the eigenvalues by assuming the existence of eigengaps.

A generalized eigenvalue problem (GEP), given $A, B \in \mathbb{R}^{n \times n}$, is to find the solution to the system

$$
\beta Ax = \alpha Bx,
$$

and each pair of $(\alpha, \beta)$ and $x$ that satisfies this equation is called a pair of generalized
eigenvalue and generalized eigenvector. One natural idea to study the perturbation of generalized eigen system (5.2) is to left-multiply the inverse of $B$ on both sides and yields an ordinary eigenvalue problem $B^{-1}Ax = \alpha \beta x$, provided that $B$ is invertible. However, when $B$ is singular, this approach would fail but feasible solution to GEP may still exists (Stewart, 1979; Sun, 1983; Crawford, 1976). More specifically, when matrices $A$ and $B$ have common null space, the set of eigenvalues may become the whole complex plane. In this case, the problem is said to be *ill-disposed* since the spectrum is extremely unstable. The *Crawford number*, defined as

$$C(A, B) = \min_{\|x\|=1} \{|x^H (A + iB)x|\},$$

is very important in this context. Here $x^H$ means the conjugate transpose. For real matrices, we could also write it as $C(A, B) = \min_{\|x\|=1} \{(x^T Ax)^2 + (x^T Bx)^2\}^{1/2}$. Matrix pair $(A, B)$ is said to be definite if $C(A, B) > 0$ holds. In this case, the problem is called a definite problem. This technical condition ensures that $A$ and $B$ not having interlacing null space.

It is shown in Stewart (1979) that without special information, the eigenvectors may be very sensitive to small perturbations, but the subspace spanned by them may be stable. In the following, we summarize some related concepts and results on subspace perturbation. Notations with tildes denote the counterparts in perturbed problem.

**Definition 5.2.1.** Suppose $(A, B)$ is a definite matrix pair. A subspace $\mathcal{X}$ is an eigenspace of $(A, B)$ if $\dim(A\mathcal{X} + B\mathcal{X}) \leq \dim \mathcal{X}$.

For a definite pair, there always exists $Z = (Z_1, Z_2)$ with $Z_1 \in C_{n \times l}$, $Z_2 \in C_{n \times (n-l)}$, such that

$$Z^HAZ = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}, \quad Z^HBZ = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix}$$

where $A_1, B_1 \in C^{l \times l}$, $Z_1^HZ_1 = I_l$ and $Z_2^HZ_2 = I_{n-l}$, and a similar decomposition holds for perturbed matrices. Clearly, $Z_1$ is an eigenspace for $(A, B)$. Let $\mathcal{R}(A)$ be the column space
of $A$. Analyzing a rotation between $\mathcal{R}(Z_1)$ and $\mathcal{R}(\tilde{Z}_1)$ shows that

$$\Theta = \cos^{-1}(Z_1^H \tilde{Z}_1 \tilde{Z}_1^H Z_1)^{1/2}$$

(5.5)

represents the canonical angles between some sets of suitably chosen base vectors of $\mathcal{R}(Z_1)$ and $\mathcal{R}(\tilde{Z}_1)$. Hence, $\sin \Theta$ becomes a good measure of the difference between these two subspaces. The following theorem depicts the relationship between $\sin \Theta$, the gap between subspaces and corresponding projection operators.

**Theorem 5.2.1.** *(Stewart, 1979)* Let $P_R$ and $P_{\tilde{R}}$ be the orthogonal projections onto $\mathcal{R}(Z_1)$ and $\mathcal{R}(\tilde{Z}_1)$. Let $\Theta$ be defined by (5.5). Furthermore, define the gap between subspaces $\mathcal{R} \triangleq \mathcal{R}(Z_1)$ and $\tilde{\mathcal{R}} \triangleq \mathcal{R}(\tilde{Z}_1)$ as $G(\mathcal{R}, \tilde{\mathcal{R}}) = \max \left\{ \sup_{\|x\|=1} \inf_{y \in \tilde{\mathcal{R}}} \|x - y\|, \sup_{\|y\|=1} \inf_{x \in \mathcal{R}} \|x - y\| \right\}$.

Then

$$G(\mathcal{R}, \tilde{\mathcal{R}}) = \|P_R - P_{\tilde{R}}\| = \|\sin \Theta\|,$$

$$\sqrt{2}G(\mathcal{R}, \tilde{\mathcal{R}}) \leq \|P_R - P_{\tilde{R}}\|_F = \sqrt{2}\|\sin \Theta\|_F.$$

This equivalence makes $\sin \Theta$ a commonly used measure for the difference between two subspaces. We also define the *chordal distance* between points $p_1 = (a_1, b_1)$, $p_2 = (a_2, b_2)$ as

$$\rho(p_1, p_2) = \frac{|a_1 b_2 - a_2 b_1|}{\sqrt{|a_1|^2 + |b_1|^2} \sqrt{|a_2|^2 + |b_2|^2}}.$$ (5.6)

which is crucial for comparing eigenvalues in generalized eigen problems. It will also characterize the eigengap in GEP. It is invariant under rotation about the origin and can handle large, or infinite eigenvalues by measuring the distances on the Riemann sphere.
5.3 Kernel Multi-view Discriminant Analysis (KMvDA)

Kernel trick enables us to access a much higher, possibly infinite dimensional feature space (as given by the Gaussian kernel) by operating in an inner product space associated with a proper Reproducing Kernel Hilbert Space (RKHS) (Aronszajn, 1950). In this section, we combine kernel trick with MvDA and derive kernel MvDA (KMvDA).

5.3.1 Formulation

Formulation. Without loss of generality, we look at one projection direction (e.g., the top eigenvector). Based on previous definitions, we rewrite $S_{jr}$ and $D_{jr}$ in matrix form:

\[
S_{jr} \triangleq X_jH^S_{jr}X_r = \begin{cases} 
X_j \left( I - \frac{1}{n} \sum_{i=1}^{c} e^i_j (e^i_j)^T \right) X_r, & j = r, \\
X_j \left( - \frac{1}{n} \sum_{i=1}^{c} e^i_j (e^i_r)^T \right) X_r, & \text{otherwise},
\end{cases}
\]

(5.7)

\[
D_{jr} = \sum_{i=1}^{c} \frac{1}{n_i} \left( n_{ij} \mu^{(x)}_{ij} \right) \left( n_{ir} \mu^{(x)}_{ir} \right)^T - \frac{1}{n} \left( \sum_{i=1}^{c} \sum_{k=1}^{n_{ir}} x_{irk} \right) \left( \sum_{i=1}^{c} \sum_{k=1}^{n_{ir}} x_{irk} \right)^T
\]

\[
= \sum_{i=1}^{c} \frac{1}{n_i} X_j e^i_j (X_r e^i_r)^T - \frac{1}{n} X_j e^i_j (X_r e^i_r)^T
\]

\[
= X_j \left( - \frac{1}{n} \sum_{i=1}^{c} e^i_j (e^i_r)^T \right) X_r
\]

\[
\triangleq X_jH^D_{jr}X_r,
\]

(5.8)

where $e_r$ is a vector with all elements equal to one and the dimensionality of $e_r$ is the same as the number of the examples of the $r$-th view; $e^i_r$ is a vector whose dimensionality is the same as that of $e_r$ and with the $i$-th class equal to one and zero otherwise. In the rest of this section, the same computation is described in another inner product space $\mathcal{F}$, which is associated with the input space by map $\phi: \mathbb{R}^d \rightarrow \mathcal{F}$, $x \mapsto \phi(x)$ and the RBF kernel
function \( k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \) in a reproducing kernel Hilbert space (RKHS) such that for all \( x, y \in \mathcal{X} \),

\[
k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{F}}.
\]

Note that the feature space \( \mathcal{F} \) could have an arbitrarily large, possibly infinite dimensionality. However, explicit representation of the function \( \phi(\cdot) \) is unnecessary as long as \( \mathcal{F} \) is a proper inner product space. By this mapping, the objective function of KMvDA becomes

\[
J = \sum_{j=1}^{v} \sum_{r=1}^{v} w_j^T \phi(X_j) \left( \sum_{i=1}^{c} \frac{1}{n_i} e_i^T e_i - \frac{1}{n} e e^T \right) K_r z_r + \sum_{j=1}^{v} \sum_{r=1,r \neq j}^{v} w_j^T \phi(X_j) \left( \sum_{i=1}^{c} \frac{1}{n_i} e_i^T e_i \right) K_r z_r,
\]

where \((w_1, ..., w_v)\) are projection directions of distinct views. By the well-known Representer Theorem in RKHS, there exists \( z_j \) such that \( w_j = \phi(X_j) z_j, \forall j = 1, ..., v \). Therefore, we can re-write the objective function using the inner products in the feature space \( \mathcal{F} \),

\[
J = \sum_{j=1}^{v} \sum_{r=1}^{v} z_j^T K_j \left( \sum_{i=1}^{c} \frac{1}{n_i} e_i^T e_i - \frac{1}{n} e e^T \right) K_r z_r + \sum_{j=1}^{v} \sum_{r=1,r \neq j}^{v} z_j^T K_j \left( \sum_{i=1}^{c} \frac{1}{n_i} e_i^T e_i \right) K_r z_r.
\]

(5.9)

where \( H^D \) and \( H^S \) are block matrices with entries \( H^D_{jr}, H^S_{jr} \) respectively, and \( K = \text{diag}(K_1, \ldots, K_v) \) is a block diagonal Gaussian kernel matrix. After some standard derivation, (5.9) eventually turns into solving the GEP

\[
Dz = \lambda Sz.
\]

As the eigenvalues are invariant of scale, we denote that all eigenvalues for MvDA (and KMvDA) are of the form \((\lambda_i, 1)\), along with the paired eigenvectors \( z_i, i = 1, ..., n \). Note that, every \( z_i \) is an \( n \)-dimensional vector (recall that \( n \) is the total number of samples). The \( i \)-th projection direction of the \( j \)-th view, \( z^j_i \), is set to be the slice at corresponding positions.
of the $j$-th view. For our task, we choose projection directions as the eigenvectors associated with the largest eigenvalues. More precisely, we sort $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and project $X_j$ onto an $l$-dimensional space with respect to $Z_j = (z_j^1, \ldots, z_j^l)$.

**Testing phase.** Given a new test set $Y = (Y_1, \ldots, Y_v)$, for a test example $y = (y_1, \ldots, y_v)$, we compute projections of the $j$-th view in the kernel space by

$$Proj(y_j) = (W_j)^T \phi(y_j) = (Z_j)^T \phi(X_j)^T \phi(y_j) = (Z_j)^T k(X_j, y_j),$$

where $y_j$ is the $j$-th view of $y$ and $k(\cdot, y)$ represents the element-wise kernel function. If our goal is to classify $y_j$ based on view $Y_m$, we assign $y_j$ with the label of nearest neighbor of $Y_m$ in the projected space, $\hat{C}(y_j) = C(\arg\min_{y' \in Y_m} \|Proj(y') - Proj(y_j)\|)$.

### 5.3.2 KMvDA with Randomized Kernels

As discussed preceding, in many practical tasks, computing kernels is very expensive when the data size is large. Therefore, we adopt Random Fourier Features (RFFs) as (4.15) for KMvDA problem. Therefore, we do not need to explicitly train the KMvDA, whose complexity is at least $O(n^2)$ for the eigen decomposition. Instead, we train a linear MvDA model with generated $m$ RFFs per sample of the dataset. The complexity would be around $O(m^2)$. In cases where $n$ is much larger than $m$ (i.e., large-scale MvDA problems), the computational cost would be significantly reduced.

### 5.4 Analysis of Randomized KMvDA

Recall that, the key ingredient of KMvDA is the learned subspaces. In this section, we investigate the subspace perturbation of using linearized RBF kernels (approximated by RFF), which directly determines the approximation effectiveness of randomized KMvDA to the exact KMvDA method. For each view with data $X$, denote the approximated kernel
matrix as $\hat{K}$, where $\hat{K}_{ij} = \hat{k}(x_i, x_j)$ estimated by (4.17). In this sequel, notations with hats are defined for objects associated with the approximated kernel. Without loss of generality, we assume that the number of examples in each view is the same, i.e., $\bar{n} = n/v$. Moreover, in each view, the number of observations per class is also the same (all equal to $\bar{n}/c$). Besides, the classes are ordered in the same way in all views. The proofs are given in Appendix D.

**Lemma 5.4.1.** Let $H^S, H^D, H^S_{(,)}$ and $H^D_{(,)}$ be defined in (5.7), (5.8) and (5.9). For $\forall j, r \leq v, \|H^D_{jr}\| = \frac{1}{v}$. For $\forall j \neq r, \|H^S_{jj}\| = 1, \|H^S_{jr}\| = \frac{1}{v}$. Moreover, $\|H^D\| = \|H^S\| = 1$, and $D, S, \hat{D}$ and $\hat{S}$ are positive semi-definite matrices.

Lemma 5.4.1 summarizes the spectral property of covariance structure sub-matrices. In particular, it illustrates that the generalized eigen problem arise from KMvDA is definite, and thus the following analysis would be valid.

### 5.4.1 A General Perturbation Bound

As discussed in preliminaries, a feasible solution to (5.10) exists as long as the GEP is definite, which does not require $S$ to be invertible. We first consider this general situation. The next lemma is a modified version of Theorem 3 in Lopez-Paz et al. (2014), which characterizes the kernel approximation error.

**Lemma 5.4.2.** Suppose $X \subset X^n$. Define linear approximation $\hat{K}_{n \times n}$ using $m$ random samples as (11). Then with probability $1 - \eta$,

$$
\|\hat{K} - K\| \leq \frac{2n \log \frac{2n}{\eta}}{3m} + \sqrt{\frac{4n^2(\log \frac{2n}{\eta})^2 + 18mn\|K\| \log \frac{2n}{\eta}}{3m}}.
$$

It is worth mentioning that because of the correlated entries of $\hat{K}$, in general this bound cannot be reduced in the absence of more structural assumptions. Now we are ready to study the eigenspace perturbation caused by kernel approximation.
Theorem 5.4.3. For the GEP associated with KMvDA (i.e., (5.10)), assume that $D$, $S$, $\hat{D}$ and $\hat{S}$ admit decompositions (5.4) in the form of $M = \text{diag}(M_1, M_2)$ correspondingly. Let $\lambda(D, S)$ denote the set of eigenvalues of (5.10)). Assume the Crawford number $C(D, S) > 0$, $C(\hat{D}, \hat{S}) > 0$, and there are $\alpha \geq 0$, $\delta > 0$ satisfying $\alpha + \delta \leq 1$, and a real number $\gamma$, such that

$$\frac{|\gamma - \lambda_i|}{\sqrt{\gamma^2 + 1/\lambda_i^2 + 1}} \leq \alpha, \quad \forall \lambda_i \in \lambda(D_1, S_1),$$

$$\frac{|\gamma - \hat{\lambda}_i|}{\sqrt{\hat{\gamma}^2 + 1/\hat{\lambda}_i^2 + 1}} \geq \alpha + \delta, \quad \forall \hat{\lambda}_i \in \lambda(\hat{D}_2, \hat{S}_2).$$

(5.11)

Denote $\|K^*\| = \max_{i=1,\ldots,v} \|K_i\|$, $\|\hat{K}^*\| = \max_{i=1,\ldots,v} \|\hat{K}_i\|$. Then the following inequality holds with probability $1 - \eta$,

$$\| \sin \Theta \| \leq \frac{p(\alpha, \delta, \gamma) \|K^*\| \|\hat{K}^*\| \xi}{C(D, S)C(\hat{D}, \hat{S})},$$

where

$$p(\alpha, \delta, \gamma) = \frac{q(\gamma)[(\alpha + \delta)\sqrt{1 - \alpha^2} + \alpha\sqrt{1 - (\alpha + \delta)^2}]}{2\alpha + \delta}$$

with $q(\gamma) = 2\sqrt{2}$ for $\gamma \neq 0$ and $q(0) = 2$. Also, we have

$$\xi = \frac{2n \log \frac{2n/v}{1 - (1 - \eta)^{1/v}}}{3vm} + \frac{\sqrt{4(n/v)^2(\log \frac{2n}{1 - (1 - \eta)^{1/v}})^2 + \frac{18}{v}mn\|K^*\| \|\hat{K}^*\| \log \frac{2n/v}{1 - (1 - \eta)^{1/v}}}}{3m}.$$

where $m$ is the number of random features.

Condition (5.11) characterizes the separation of the generalized eigenvalues, where the eigengap can be interpreted in terms of chordal distance, defined by (5.6). Since the generalized eigenvalues are invariant of scale, we may force them on a unit semicircle in the upper plane. Note that in our problem, the generalized eigenvalues are in the form $(\lambda_i, 1)$. Hence, we can scale each eigenvalue to $(s_i, t_i) \triangleq \left( \frac{\lambda_i}{\sqrt{\lambda_i^2 + 1}}, \frac{1}{\sqrt{\lambda_i^2 + 1}} \right)$. For any two pairs, we have
\[
\sin((s_i, t_i), (\tilde{s}_i, \tilde{t}_i)) = \frac{|\lambda_i - \tilde{\lambda}_i|}{\sqrt{\lambda_i^2 + 1} \sqrt{\tilde{\lambda}_i^2 + 1}} = \rho((s_i, t_i), (\tilde{s}_i, \tilde{t}_i)).
\]

That is, the chordal distance between two eigenvalue pairs is the sine between the two rays with slopes \(\frac{1}{\lambda_i}\) and \(\frac{1}{\tilde{\lambda}_i}\) extended from the origin. Now we can translate (5.11) into angles (defined anti-clockwise): there exist a real number \(\gamma, \alpha \geq 0, \delta > 0\) and \(\alpha + \delta \leq 1\), such that

\[
\max_{\lambda_i \in \lambda(D_1, S_1)} \sin((\lambda_i, 1), (\gamma, 1)) \leq \alpha \triangleq \sin \theta,
\]

\[
\min_{\tilde{\lambda}_i \in \lambda(\tilde{D}_1, \tilde{S}_1)} \sin((\tilde{\lambda}_i, 1), (\gamma, 1)) \geq \alpha + \delta \triangleq \sin \tilde{\theta}.
\]

Define \(\theta_g = \min_{\lambda_i \in \lambda(D_1, S_1), \tilde{\lambda}_i \in \lambda(\tilde{D}_1, \tilde{S}_1)} \sin((\lambda_i, 1), (\tilde{\lambda}_i, 1))\) as the gap between eigenvalue sets \(\lambda(D_1, S_1)\) and \(\lambda(\tilde{D}_1, \tilde{S}_1)\). It is easy to check that \(\theta_g = \theta - \tilde{\theta}\), and

\[
\sin(\theta_g) = \sin(\tilde{\theta}) \cos(\theta) - \cos(\tilde{\theta}) \sin(\theta) \geq (\alpha + \delta) \sqrt{1 - \alpha^2} - \alpha \sqrt{1 - (\alpha + \delta)^2} > 0,
\]

which implies that two sets of eigenvalues are well separated.

### 5.4.2 Perturbation of Regularized Problem

In practice, a regularization term is often added to GEP to handle singularity and make the system more stable and theoretically justifiable. Consider the following regularized GEP,

\[
Dw = (S + \epsilon I)w,
\] (5.12)

with \(\epsilon > 0\) a small constant. The problem is guaranteed to be definite, since \((S + \epsilon I)\), by Lemma 5.4.1, now becomes positive definite. More importantly, the invertibility of \((S + \epsilon I)\) allows us to transform (5.12) into an SEP.

**Theorem 5.4.4.** Let \(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n\) denote eigenvalues of \((S + \epsilon I)^{-1}D\), and
\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n \text{ be the eigenvalues of } (\hat{S} + \epsilon I)^{-1} \hat{D}. \text{ Assume } \lambda_l - \hat{\lambda}_{l+1} = \delta > 0, \text{ then with probability } 1 - \eta,

\| \sin \Theta \| \leq \frac{\xi_\eta}{\delta} \cdot \left\{ C \frac{\| K^* \|^2 (\| K^* \| + \| \hat{K}^* \|)}{\epsilon^2} + \left( \frac{\| K^* \| + \| \hat{K}^* \|}{\epsilon} \right) \right\},

where \( C = \frac{1+\sqrt{5}}{2} \). \| K^* \|, \| \hat{K}^* \| \text{ and } \xi_\eta \text{ are defined in Theorem 5.4.3.}

5.4.3 \( \sin \Theta \) Theorem Based on Population Eigengap Condition

In both Theorem 5.4.3 and Theorem 5.4.4, we see that the eigengap condition involves both \( \lambda \)'s of the population eigenvalues, and \( \hat{\lambda} \)'s of the approximated (perturbed) eigenvalues. In the following, we establish a general \( \sin \Theta \) theorem with more natural assumptions—the eigengap is only required for the population eigenvalues. Our result holds for general matrices and Singular Value Decomposition (SVD). For any real matrix \( A \in \mathbb{R}^{p \times q} \), denote the SVD as \( A = U \Sigma V^T \) where \( U \) and \( V \) are orthonormal matrices and \( \Sigma \) is diagonal.

**Theorem 5.4.5.** Let \( A, \tilde{A} \in \mathbb{R}^{p \times q} \) be two matrices with with singular values \( \sigma_1 \geq \cdots \geq \sigma_{\min\{p,q\}} \) and \( \tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_{\min\{p,q\}} \) respectively. Suppose \( l \leq \text{rank}(A) < \min\{p, q\} \) and \( \sigma_l - \sigma_{l+1} > 0 \). Let the right singular matrices \( V = (v_1, \ldots, v_l) \in \mathbb{R}^{p \times l} \) and \( \tilde{V} = (\tilde{v}_1, \ldots, \tilde{v}_l) \in \mathbb{R}^{p \times l} \) have orthonormal columns satisfying \( Av_j = \sigma_j u_j \) and \( \tilde{A} \tilde{v}_j = \tilde{\sigma}_j \tilde{u}_j \) for \( j = 1, \ldots, l \), where \( u_j \) and \( \tilde{u}_j \) are corresponding left singular vectors. Then we have

\[ \| \sin \Theta(V, \tilde{V}) \| \leq \frac{2(\sigma_1 + \tilde{\sigma}_1)\| A - \tilde{A} \|}{\sigma_l^2 - \tilde{\sigma}_{l-1}^2}. \]

Again, note that in Theorem 5.4.5, it is in some cases more natural and favorable as the eigengap condition is required only for the population singular values \( \sigma_1, \ldots, \sigma_p \), independent of the perturbed counterparts.

**Theorem 5.4.6.** Consider eigenvalue problem \( (S + \epsilon I)^{-1} Dw = \lambda w \) where \( S, D \in \mathbb{R}^{n \times n} \) are defined in (5.9), and \( (\hat{S} + \epsilon I)^{-1} \hat{D} w = \hat{\lambda} w \) its randomized approximation. Let the
the eigenvalues be \( \lambda_1 \geq \cdots \geq \lambda_n \) and \( \hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n \) respectively, with corresponding eigenvectors \((w_1, \ldots, w_n)\) and \((\hat{w}_1, \ldots, \hat{w}_n)\). Assume for \( 1 \leq l < n \), \( \lambda_l - \lambda_{l+1} > 0 \). Denote \( W = (w_1, \ldots, w_l) \) and \( \hat{W} = (\hat{w}_1, \ldots, \hat{w}_l) \). Then we have

\[
\| \sin \Theta(W, \hat{W}) \| \leq \frac{(1 + \sqrt{5})(\lambda_1 + \hat{\lambda}_1)}{\lambda_l^2 - \hat{\lambda}_l^2} \left( \frac{\| K^* \| \| \hat{K}^* \|}{\varepsilon^2} + \frac{(\| K^* \| + \| \hat{K}^* \|)}{\varepsilon} \right).
\]

The proof immediately follows from Theorem 5.4.4 and Theorem 5.4.5.

5.5 Experiments

In this section, we present empirical results that illustrate the performance of KMvDA and linearized KMvDA using random Fourier features. The major goal is to show 1) KMvDA improves linear MvDA, and 2) randomized KMvDA is able to well approximate the performance of KMvDA with sufficient number of RFFs.

5.5.1 Datasets

We test our algorithms on 3 popular datasets for multi-view learning research and applications. All datasets are publicly available.

**Heterogeneous Face Biometrics (HFB)** database has 100 persons in total, with 4 composed of visual (VIS) and 4 near infrared (NIR) face images for each person. This gives us a 2-view classification problem. For each view, we have 400 examples in total from 100 different people. We use the first 65 persons for training and the remaining 35 persons for testing. Each example is a \( 32 \times 32 \) image, which is transformed into 1024 features.

**CUHK Face Sketch FERET (CUFSF)** database is designed for research on face sketch synthesis and face sketch recognition. It includes 1194 persons (i.e., categories) from the FERET database. An example is given in Figure 5.1. This dataset contains two views: 1) face photo with lighting variation, and 2) sketch with shape exaggeration drawn by an artist when viewing this photo, both with dimensionality 5120. We use the first 650 examples as
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Views</th>
<th>Linear f</th>
<th>RBF f</th>
<th>RFF f</th>
</tr>
</thead>
<tbody>
<tr>
<td>HFB</td>
<td>NIR-VIS VIS-NIR</td>
<td>56.4</td>
<td>59.3</td>
<td>60.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>47.2</td>
<td>51.4</td>
<td>51.4</td>
</tr>
<tr>
<td>CUFSF</td>
<td>Photo-Sketch</td>
<td>45.2</td>
<td>52.4</td>
<td>51.0</td>
</tr>
<tr>
<td></td>
<td>Sketch-Photo</td>
<td>52.6</td>
<td>52.2</td>
<td>52.4</td>
</tr>
<tr>
<td>Multi-PIE</td>
<td>Avg. Accuracy</td>
<td>93.6</td>
<td>94.8</td>
<td>94.8</td>
</tr>
</tbody>
</table>

Table 5.1: Randomized KMvDA: Rank-1 recognition rate (%) of linear kernel, RBF kernel and RFF approximation.

training set and the rest 544 examples for testing.

We use Multi-PIE dataset to test the performance of KMvDA on dealing with multiple views and larger sample size. The whole dataset contains more than 750,000 face images of 337 people, under different poses and from distinct views. In our experiment, we choose 7 different views (left 45°, 30°, 15°, frontal, right 15°, 30°, 45°), three facial expressions (smile, neutral, disgust), and no flush illumination as the evaluation data. Each example is a 5,120 dimensional vector. This subset is divided into two parts: the images from the first 248 subjects with 4 randomly selected images under each pose of each person are utilized as training data and the images from the rest are utilized as test data.

5.5.2 Parameters and Performance Evaluation

**Kernels.** There is no tuning parameter for linear kernel. For RBF kernel, we fine-tune the parameter $\sigma$ over a fine grid in the range of $\{0.001, 100\}$. The number of random Fourier features are chosen to be $m = \{2^6, 2^7, \ldots, 2^{15}\}$ for each view. We set $\sigma$ for RFFs the same as fine tuned parameter value for RBF kernel to compare the approximation effectiveness of linearized methods. RFF vectors are normalized to have unit norm.

**Evaluation.** We mainly use the classification test accuracy to evaluate the model performance. We denote “$v_2$–$v_1$” when using training examples from view $v_1$ to classify test examples from view $v_2$. The metric we use is the rank-1 recognition rate, which is the highest test accuracy among all parameter $\sigma$ and projection dimensionality $d$. 
5.5.3 Experiment Results

**Overall performance.** Table 5.1 summarizes the rank-1 recognition rate of different approaches on HFB and CUFSF datasets, and the average rank-1 recognition rate among all 7 views for Multi-PIE dataset. As we can see, RBF kernel significantly outperforms linear kernel in almost all cases. In addition, the accuracy of using linearized approximation is very close to that of using RBF kernel directly, sometimes even slightly better.

![Graph](image1.png)

![Graph](image2.png)

**Figure 5.3:** RFFs: rank-1 recognition rate vs. number of random Fourier features. The upper panel is for HFB dataset and the lower panel is for CUFSF dataset.

**Number of features.** In Figure 5.3, we plot the highest test accuracy against different $m$, the number of random features. For HFB dataset, the recognition rate becomes stable at around $m = 2^{11}$. For CUFSF and Multi-PIE (Figure 5.5) dataset, this number is between $2^{12}$ to $2^{13}$. This is consistent with the observation in Băzăvan et al. (2012) that a few thousands
of RFFs are often required in order to provide good approximation.

Figure 5.4: Linear kernel, RBF kernel and RFFs: rank-1 recognition rate vs. projection dimensionality. Upper panel: HFB. Lower panel: CUFSF.

Number of projections. Figure 5.4 shows the rank-1 accuracy against the subspace dimensionality $l$. We observe for all cross-views, the performance of KMvDA stabilizes after the dimensionality reaches 50, which appears to be a good recommendation in practice. Also, adding more projection directions may deteriorate the test accuracy of linear kernel, since we observe significant decrease in recognition rate in all figures after $l = 50$. In this sense, RBF kernel (as well as RFFs) is much more robust.

Multi-PIE dataset. Tables 5.2, 5.3, and 5.4 demonstrate the best recognition rate among all views of Multi-PIE dataset. Here gallery means training view, and probe refers to test
We see that RBF kernel improves the accuracy on almost all cross-views. The pair 
(0, $-45^\circ$) and (0, $45^\circ$) are most challenging tasks since the front face is most different from 
the face seen from $\pm 45^\circ$ angle. For these cross-views, RBF can increase the accuracy by 
around 5%. Figure 5.5 shows the results on this cross-view. The right panel of Figure 5.5 also 
plots the average accuracy among all pairs of views, which again confirms the convergence 
since the curves of RBF and RFFs almost overlap.
Figure 5.5: Multi-PIE dataset. **Left panel:** Accuracy vs. number of RFFs, $-45^\circ \rightarrow 0^\circ$ cross-view. **Middle panel:** Accuracy vs. projection dimensionality, $-45^\circ \rightarrow 0^\circ$ cross-view. **Right panel:** Average accuracy over all cross-views vs. dimensionality.
CHAPTER 6
QUANTIZATION ALGORITHMS FOR RANDOM FOURIER FEATURES

In Chapter 5, we have shown an application of Random Fourier Features (RFF) in training approximate non-linear MvDA models. In this chapter, we focus on the general compression scheme of RFF, for more substantial memory saving. From now on, we will assume that \( \mathcal{X} \) is the unit sphere, i.e., all the data points are normalized to have unit \( l_2 \) norm. This will save us from book-keeping the sample norms in our calculations, plus normalizing data vectors before classification is a fairly standard (or recommended) procedure in practice. Denoting \( \rho = \cos(u, v) = u^T v \), we can formulate

\[
k_{\gamma}(u, v) = e^{-\gamma^2(2-2\rho)} = e^{-\gamma^2(1-\rho)}. \tag{6.1}
\]

Again, the subscript “\( \gamma \)” may be dropped if there is no risk of confusion.

As introduced in Section 1.2.1, since each Lloyd-Max (LM) quantizer is associated with a specific random signal distribution, at the first glance, designing LM quantizers for the random Fourier features and the Gaussian kernel might appear challenging, due to the tuning parameter \( \gamma \), which is a crucial component of the Gaussian kernel. Initially, one might expect that a different LM quantizer would be needed for a different \( \gamma \) value. In this chapter, our contribution begins with an interesting finding that the marginal distribution of the RFF is actually free of the parameter \( \gamma \). This result greatly simplifies the design of LM quantization schemes for the RFF, because only one quantizer would be needed for all \( \gamma \) values. Once we have derived the marginal distribution of the RFF, we incorporate the idea of distortion optimal quantization theory to nonlinear random feature compression by providing a thorough study on the theoretical properties and practical performance. Extensive simulations and machine learning experiments validate the effectiveness of the
proposed LM quantization schemes for the RFF. We provide the proofs of the Theorems in Appendix E.

6.1 The Probability Distributions of RFF

We start the introduction to our proposed method by providing analysis on the probability distribution of RFF given by (4.15), which is key to the design of quantization schemes in Section 6.2. First, we introduce some notations.

Let $u, v \in \mathcal{X} \subseteq \mathbb{R}^d$ be two normalized data points, and $w \in \mathbb{R}^d$ be a random vector with i.i.d. $N(0, 1)$ entries. The randomly projected data $w^T u$ and $w^T v$ follow a bivariate normal distribution:

$$
\begin{pmatrix}
    w^T u \\
    w^T v
\end{pmatrix}
\sim
N
\left(
0, 
\begin{pmatrix}
    1 & \rho \\
    \rho & 1
\end{pmatrix}
\right),
\text{ where } \rho = u^T v, \|u\| = \|v\| = 1.
$$

Following previous notation, we will use the following two definitions for $\phi_\sigma(t)$ and $\Phi(t)$:

$$
\phi_\sigma(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2\sigma^2}},
$$

$$
\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \int_{-\infty}^{t} \phi_1(z) dz.
$$

That is, $\Phi(t)$ is the cumulative distribution function (CDF) of the standard normal $N(0, 1)$ and $\phi_\sigma(t)$ is the probability density function (pdf) of $N(0, \sigma^2)$.

6.1.1 Marginal Distributions

We first consider the marginal distributions of the RFF, which serve as the foundation of our proposed quantization schemes. The following Lemma is a result of the convolution of normal and uniform distributions.

Lemma 6.1.1. Suppose $X \sim N(0, 1)$ and $\tau \sim \text{uniform}(0, 2\pi)$ are independent, $\gamma > 0$. 

Then

\[ \gamma X + \tau \sim \frac{1}{2\pi} \left[ \Phi\left( \frac{2\pi - y}{\gamma} \right) - \Phi\left( -\frac{y}{\gamma} \right) \right]. \]

In the following, we formally give the distribution of the RFF.

**Theorem 6.1.2.** Let $X \sim N(0, 1)$, $\tau \sim \text{uniform}(0, 2\pi)$ be independent. Denote $Z = \cos(\gamma X + \tau)$, and $Z_2 = \cos^2(\gamma X + \tau)$. We have the probability density functions

- $f_{Z}(z) = \frac{1}{\pi \sqrt{1 - z^2}}$, $z \in [-1, 1]$, \hspace{1cm} (6.2)
- $f_{Z_2}(z) = \frac{1}{\pi \sqrt{z - z^2}}$, $z \in [0, 1]$. \hspace{1cm} (6.3)

for any $\gamma > 0$. In particular, $Z \overset{d}{\sim} \cos(\tau)$ in distribution.

The density plots can be found in Figure 6.1. Theorem 6.1.2 says that for any kernel parameter $\gamma$, the (unscaled) RFF follows the same distribution as the cosine of the uniform noise itself. Intuitively, this is because cosine is a $2\pi$-periodic function and normal distribution is symmetric. As will be introduced in Section 6.2, each Lloyd-Max (LM) quantizer is associated with a signal distribution. We will characterize two LM-type quantizers w.r.t. density (6.2) and (6.3), respectively. This interesting result is favorable for our purpose as it...
implies we only need to construct one LM quantizer, which covers all the Gaussian kernels with different $\gamma$ value. Thus, the design of LM quantizer for RFF is convenient.

**Remark 6.1.1.** In Theorem 6.1.2 we consider $X \sim N(0, 1)$ because we assume data samples are normalized for conciseness. It is easy to see that this result also holds without data normalization (i.e., $X$ is Gaussian with arbitrary variance) since we can offset the variance of $X$ by altering $\gamma$. Therefore, Theorem 6.1.2 is a universal result implying that the LM quantizer also works without data normalization.

### 6.1.2 Joint Distribution

In the sequel, we analyze the joint distribution of RFFs of two data samples with correlation $\rho$’s. The joint distribution will play an important role in later theoretical analysis. The following Lemma 6.1.3 leads to the desired result presented in Theorem 6.1.4.

**Lemma 6.1.3.** Denote $z_x = \gamma X + \tau$, $z_y = \gamma Y + \tau$ with $(X, Y) \sim N(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$, $\tau \sim \text{uniform}(0, 2\pi)$. We have the joint distribution

$$f(z_x, z_y) = \frac{1}{2\pi} \phi \sqrt{2(1-\rho)\gamma} (z_x - z_y) \left[ \Phi\left( \frac{4\pi - (z_x + z_y)}{\gamma \sqrt{2(1+\rho)}} \right) - \Phi\left( -\frac{z_x + z_y}{\gamma \sqrt{2(1+\rho)}} \right) \right].$$

**Theorem 6.1.4.** Denote $z_x = \cos(\gamma X + \tau)$, $z_y = \cos(\gamma Y + \tau)$ where $X, Y, \tau$ are the same as Lemma 6.1.3. Then we have the joint density function for $(z_x, z_y) \in [-1, 1]^2$,

$$f(z_x, z_y) = \frac{1}{\pi \sqrt{1 - z_x^2} \sqrt{1 - z_y^2}} \sum_{k=-\infty}^{\infty} \phi \phi \left[ \frac{a_x^* - a_y^* + 2k\pi}{\sqrt{2(1-\rho)\gamma}} a_x^* + a_y^* + 2k\pi \right],$$

where $a_x^* = \cos^{-1}(z_x)$, $a_y^* = \cos^{-1}(z_y)$. In addition, $(\sin(\gamma X + \tau), \sin(\gamma Y + \tau))$ follows the same distribution.

In Figure 6.2, we plot the joint density at several $\gamma$ values. We conclude several properties of the joint distribution. Firstly, it is obvious that $z_x$ and $z_y$ are exchangeable, i.e.,
Secondly, it is symmetric which means \( f(Z_x, Z_y) = f(-Z_x, -Z_y) \). Moreover, we have the following important result, which is helpful for our analysis on the monotonicity and variance of quantized kernel estimators in Section 6.3.

**Proposition 6.1.5.** Let the density function \( f \) be defined as Theorem 6.1.4. If \( \sqrt{2(1 - \rho)\gamma} \leq \pi \), then \( f(z_x, z_y) > f(z_x, -z_y) \) for \( \forall (z_x, z_y) \in (0, 1]^2 \) or \( (z_x, z_y) \in [-1, 0)^2 \).

In Proposition 6.1.5, the quantity \( \sqrt{2(1 - \rho)\gamma} \) will be reduced if we either increase \( \rho \) or decrease \( \gamma \). In Figure 6.2, we see how this term characterizes the joint density of RFF. In particular, smaller \( \sqrt{2(1 - \rho)\gamma} \) reinforces the dependency between \( z_x \) and \( z_y \). The density around \((1, 1)\) and \((-1, -1)\) reaches the highest when \( \rho = 0.9 \) and \( \gamma = 1 \). As \( \rho \) decreases or \( \gamma \) increases, the density is “flattened”.

![Figure 6.2: The joint density of RFF (Theorem 6.1.4) with example \( \rho = -0.9, 0.9 \), and \( \gamma = 1, 3, 5 \).](image)

### 6.2 Quantization Schemes for RFF

Quantization is, to a good extent, an ancient topic in information theory and signal processing (Widrow and Kollár, 2008). On the other hand, many interesting research works appear in the literature even very recently, for achieving better efficiency in data storage, data transmission, computation, and energy consumption. Quantization for random projections has been heavily studied. In this chapter, we focus on developing quantization schemes for random Fourier features, in particular, based on the Lloyd-Max (LM) framework.
This article https://www.eetimes.com/an-introduction-to-different-rounding-algorithms/, published in EETimes 2006, provides a good summary of common quantization (rounding) schemes: “rounding in decimal”, “round-toward-nearest”, “round-half-up”, “round-half-down”, “round-half-even”, “Round-half-odd”, “round-alternate”, “round-random”, “round-ceiling”, “round-floor”, “round-toward-zero”, “round-away-from-zero”, “round-up”, “round-down”, “truncation”. For random projections, quantization methods such as Datar et al. (2004); Li et al. (2014); Li (2017a) belong to those categories. One should expect that, as long as we use a sufficient number (such as 32 or 64) of bits, any reasonable quantization scheme should achieve a good accuracy.

In this thesis, we mainly focus on quantization for RFF with a small number (such as 1, 2, 3, or 4) bits. We consider general multi-bit quantizers, with 1-bit quantization as a special case, for the cosine feature in RFF bounded in \([-1, 1]\). Here, “\(b\)-bit” means the quantizer has \(2^b\) levels. For simplicity, we will denote \(z = \cos(\gamma w^T u + \tau)\) as the item to be quantized. In particular, we focus on two algorithms: “round-random” (which we refer to as the “stochastic quantization (StocQ)”) and “Lloyd-Max (LM) quantization”. For clarity, we will repeat the general procedures as follows, and specify the concrete signal domain and distribution used for quantizer construction.

### 6.2.1 Stochastic Quantization (StocQ) for RFF

A \(b\)-bit StocQ quantizer, also known as stochastic rounding, splits \([-1, 1]\) (since \(z \in [-1, 1]\)) into \((2^b - 1)\) intervals, with consecutive borders \(-1 = t_0 < ... < t_{2^b-1} = 1\). Let \([t_i, t_{i+1}]\) be the interval containing \(z\). StocQ pushes \(z\) to either \(t_i\) or \(t_{i+1}\), depending on its distance to the borders. Concretely, denoting \(\Delta_i = t_{i+1} - t_i\),

\[
P(Q(z) = t_i) = \frac{t_{i+1} - z}{\Delta_i}, \quad P(Q(z) = t_{i+1}) = \frac{z - t_i}{\Delta_i}.
\]  

(6.4)
It is not difficult to see that by the sampling procedure, conditional on the full-precision RFF \( z \), the quantized value \( Q(z) \) by StocQ is unbiased of \( z \). On the other hand, also due to the Bernoulli sampling approach, StocQ has the extra variance especially when the number of bits is small (e.g., 1-bit or 2-bit quantization). Note that in Zhang et al. (2019), the authors applied StocQ with uniform borders in large-scale machine learning tasks with RFF. Here we consider a more general approach where the borders are not necessarily uniform.

6.2.2 Lloyd-Max (LM) Quantization for RFF

In information theory and signal processing, the Lloyd-Max (LM) (Max, 1960; Lloyd, 1982) quantization scheme has a long history that also leads to many well-known methods (e.g., the \( k \)-means clustering). Interestingly, it has not been adopted to RFF in the prior literature. In contrast to StocQ, the proposed LM-RFF constructs a fixed quantizer \( Q_b \). We call \([\mu_1, ..., \mu_{2b}] \in \mathbb{C}\) the reconstruction levels, with \( \mathbb{C} \) the “codebook” of \( Q_b \). Also, \(-1 = t_0 < ... < t_{2b} = 1\) represent the borders of the quantizer. Then, LM-RFF quantizer \( Q_b \) defines a mapping: \([-1, 1] \mapsto \mathbb{C}\), where \( Q_b(x) = \mu_i \) if \( t_{i-1} < x \leq t_i \). By choosing the error function as the squared difference, given an underlying signal distribution \( f(z) \) with support \( S \), the LM quantizer minimizes the distortion defined as

\[
D_Q = \int_S (z - Q(z))^2 f(z) dz,
\]

aiming to keep most amount of information of the original signal. For the signal distribution, we consider two variants. First, it is natural to set the target distribution as the distribution of RFF itself (6.2). Consequently, the first LM quantizer is subject to the distortion:

**LM-RFF:** \( D_1 \triangleq \int_{[-1,1]} (z - Q(z))^2 \frac{1}{\pi \sqrt{1 - z^2}} dz \). \((6.5)\)

Conceptually, optimizing (6.5) minimizes the average difference between RFF \( z \) and \( Q(z) \). Alternatively, we may also choose to address more on the high similarity region (which is
more important sometimes). Denote \( z_u \) and \( z_v \) as the RFFs of \( u \) and \( v \). As \( \rho \to 1 \), we have \( z_x \cdot z_y \to z_x^2 \), with density given by (6.3). Thus, our second variant, \( \text{LM}^2\)-RFF, is designed to approximate \( z^2 \) by \( Q(z)^2 \) with distortion

\[
\text{LM}^2\text{-RFF}: \quad D_2 \triangleq \int_{[-1,1]} \left( z^2 - Q(z)^2 \right)^2 \frac{1}{\pi \sqrt{1 - z^2}} dz. \tag{6.6}
\]

Minimizing (6.5) and (6.6) leads to our proposed two LM-type quantizers for RFF compression in this chapter. \(^1\)

### 6.2.3 Optimization for LM Quantizers

To solve the introduced optimization problems, we exploit classical Lloyd’s algorithm, which alternatively updates two parameters until convergence. By e.g., Wu (1992b), the algorithm converges to the globally optimal solution since the squared loss is convex and symmetric. The algorithm terminates when the total absolute change in borders and reconstruction levels in two consecutive iterations is smaller than a given threshold (e.g., \( 10^{-5} \)). We provide the concrete steps for LM-RFF and \( \text{LM}^2\)-RFF in Algorithm 4 and Algorithm 5, respectively.

For LM-RFF, we see that the procedure is standard (exactly the same as above derivation).

Denote \( z \) as the unscaled RFF, \( z = \cos(\gamma X + \tau) \), \( X \sim N(0, 1) \) and \( \tau \sim \text{uniform}(0, 2\pi) \).

For \( \text{LM}^2\)-RFF, recall that our objective is to minimize (by change of random variable)

\[
D_Q = \int_{-1}^{1} (Q(z)^2 - z^2)^2 f_Z(z) dz = \int_{0}^{1} (\tilde{Q}(s) - s)^2 f_{Z^2}(s) ds,
\]

where \( f_Z(z) \) is the density in Theorem 6.1.2 (6.2) of RFF (supported on \([-1, 1]\)), and \( f_{Z^2}(s) \) is the density in Theorem 6.1.2 (6.3) of squared RFF \( z^2 \) (on \([0, 1]\)). Here we let \( \tilde{Q}(s) = Q(z)^2 \) on \([0, 1]\). This provides us a way to first optimize the positive part according to the density of \( z^2 \), then get the complete \( \text{LM}^2\)-RFF quantizer by taking the square root and

\(^1\)Note that, in Eq. (6.6), \( \int_{0}^{1} \frac{1}{\pi \sqrt{1-z^2}} dz = \int_{0}^{1} \frac{1}{\pi \sqrt{1-z}} dz^{1/2} = \frac{1}{2} \int_{0}^{1} \frac{1}{\pi \sqrt{1-z^2}} dz = \frac{1}{2} \int_{0}^{1} f_{Z^2}(z) dz \), where \( f_{Z^2} \) is the density (6.3).
symmetry. In particular, the SymmetricExpand function in Algorithm 5 applies reverted expansion with symmetry. For example, the output of \( \{ t = (0, 0.5, 1), \mu = (0.2, 0.7) \} \) would be \( \{ t = (-1, -0.5, 0, 0.5, 1), \mu = (-0.7, -0.2, 0.2, 0.7) \} \). For practitioners to use our quantizers forthrightly, in Table 6.1 and 6.2 we summarize precisely the derived LM-RFF and LM^2-RFF quantizers for \( b = 1, 2, 3, 4 \).

**Algorithm 4** Construction of LM-RFF quantizer

**Input:** Density \( f_Z(z) \) (Theorem 6.1.2, (6.2)), Number of bits \( b \)

**Output:** LM-RFF quantizer \( t = [t_0, ..., t_{2^b}] \), \( \mu = [\mu_1, ..., \mu_{2^b}] \)

Fix \( t_0 = -1, t_{2^b} = 1 \)

While true

For \( i = 1 \) to \( 2^b \)

Update \( \mu_i \) by \( \mu_i = \frac{\int_{t_{i-1}}^{t_i} zf_Z(z)dz}{\int_{t_{i-1}}^{t_i} f_Z(z)dz} \)

End For

For \( i = 1 \) to \( 2^b - 1 \)

Update \( t_i \) by \( t_i = \frac{\mu_i + \mu_{i+1}}{2} \)

End For

Until Convergence

**Algorithm 5** Construction of LM^2-RFF quantizer

**Input:** Density \( f_{Z_2}(z) \) (Theorem 6.1.2, (6.3)), Number of bits \( b \)

**Output:** LM-RFF quantizer \( t = [t_0, ..., t_{2^b}] \), \( \mu = [\mu_1, ..., \mu_{2^b}] \)

Fix \( t_{2^b-1} = 0, t_{2^b} = 1 \)

While true

For \( i = 2^{b-1} + 1 \) to \( 2^b \)

Update \( \mu_i \) by \( \mu_i = \frac{\int_{t_{i-1}}^{t_i} zf_{Z_2}(z)dz}{\int_{t_{i-1}}^{t_i} f_{Z_2}(z)dz} \)

End For

For \( i = 2^{b-1} + 1 \) to \( 2^b - 1 \)

Update \( t_i \) by \( t_i = \frac{\mu_i + \mu_{i+1}}{2} \)

End For

Until Convergence

\( \{ t, \mu \} = \text{SymmetricExpand}(\sqrt{t}, \sqrt{\mu}) \)

In Figure 6.3, we plot the 2-bit LM-RFF and LM^2-RFF quantizer as an example, along with the distortions of LM quantizers and uniform stochastic quantization (StocQ), with various number of bits. We see that both LM methods give non-uniform quantization borders and codes. LM^2-RFF “expands” more towards two ends since it tries to approximate \( z^2 \).
Table 6.1: Constructed borders and reconstruction levels of LM-RFF quantizers, \( b = 1, 2, 3, 4 \), keeping three decimal places. Since the quantizers are symmetric about 0, we only present the positive part for conciseness.

<table>
<thead>
<tr>
<th>( b )</th>
<th>Borders</th>
<th>Reconstruction Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>([0, 1])</td>
<td>({0.637})</td>
</tr>
<tr>
<td>2</td>
<td>([0, 0.576, 1])</td>
<td>({0.297, 0.854})</td>
</tr>
<tr>
<td>3</td>
<td>([0, 0.286, 0.563, 0.819, 1])</td>
<td>({0.144, 0.428, 0.699, 0.939})</td>
</tr>
<tr>
<td>4</td>
<td>([0, 0.142, 0.283, 0.421, 0.557, 0.687, 0.811, 0.922, 1])</td>
<td>({0.071, 0.213, 0.353, 0.49, 0.624, 0.751, 0.87, 0.974})</td>
</tr>
</tbody>
</table>

Table 6.2: Constructed borders and reconstruction levels of LM\(^2\)-RFF quantizers, \( b = 1, 2, 3, 4 \), keeping three decimal places. Since the quantizers are symmetric about 0, we only present the positive part for conciseness.

<table>
<thead>
<tr>
<th>( b )</th>
<th>Borders</th>
<th>Reconstruction Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>([0, 1])</td>
<td>({0.707})</td>
</tr>
<tr>
<td>2</td>
<td>([0, 0.707, 1])</td>
<td>({0.426, 0.905})</td>
</tr>
<tr>
<td>3</td>
<td>([0, 0.461, 0.707, 0.888, 1])</td>
<td>({0.27, 0.593, 0.805, 0.963})</td>
</tr>
<tr>
<td>4</td>
<td>([0, 0.301, 0.467, 0.596, 0.707, 0.802, 0.884, 0.954, 1])</td>
<td>({0.175, 0.39, 0.535, 0.654, 0.756, 0.845, 0.92, 0.985})</td>
</tr>
</tbody>
</table>

From the distortion plots, we validate that LM-RFF provides smallest \(D_1\) and LM\(^2\)-RFF gives smallest \(D_2\).

As a final remark before ending this subsection, we note that LM quantization is more convenient and faster compared with StocQ in practical implementation, for quantizing the full-precision RFFs. While LM is a fixed quantization approach, StocQ requires generating an extra random number for each sketch and each data point. For large datasets, producing these additional random numbers might be rather slow in practice.

Figure 6.3: **Left**: LM-RFF and LM\(^2\)-RFF quantizers, \( b = 2 \). Black curve is the RFF marginal density. **Right two**: Distortion \(D_1\) and \(D_2\) of LM-RFF, LM\(^2\)-RFF and StocQ (with uniform borders).
6.2.4 Quantized Kernel Estimators

In the following, we define the kernel estimators built upon quantized RFFs, which lie at the core of understanding the behavior of different compression schemes. For a fixed $\gamma$ (Gaussian kernel parameter), with a general RFF quantizer $\tilde{Q}$ (i.e., proper kind introduced above), a simple quantized kernel estimator using $m$ random features can be constructed as

$$\hat{k}_{Q}(u, v) = \frac{2}{m} \sum_{i=1}^{m} \tilde{Q}(z_{u,i}) \tilde{Q}(z_{v,i}), \quad (6.7)$$

with $z_{u,i} = \cos(w^Tu + \tau_i)$ and $z_{v,i} = \cos(w^Tv + \tau_i)$ the $i$-th unscaled RFF of $u$ and $v$, respectively. Moreover, for the proposed LM quantizers, we consider normalized estimator,

$$\hat{k}_{n,Q}(u, v) = \frac{\sum_{i=1}^{m} Q(z_{u,i})Q(z_{v,i})}{\sqrt{\sum_{i=1}^{m} Q(z_{u,i})^2} \sqrt{\sum_{i=1}^{m} Q(z_{v,i})^2}}. \quad (6.8)$$

This estimator can also be conveniently used, as we only need to normalize the quantized RFFs (per data point) before learning. We will use $\hat{k}_{Q,(2)}$ and $\hat{k}_{n,Q,(2)}$ to denote the corresponding estimators using LM$^2$-RFF quantization. We will analyze and compare the estimators using different quantization methods, theoretically and practically, in the remaining sections of the chapter.

6.3 Theoretical Analysis

In this section, we first analyze the mean, variance, and monotonicity property of the quantized kernel estimators, then discuss kernel matrix approximation property based on some new evaluation metrics that can well align with the generalization performance.

6.3.1 StocQ Estimator

We start this section by analyzing the stochastic rounding method for RFF. In Zhang et al. (2019), the exact variance of the kernel estimator is not provided. In the following, we
establish the precise variance calculation in Theorem 6.1.4, which is in fact a more general result on any symmetric stochastic quantizer.

**Theorem 6.3.1 (StocQ).** Suppose a b-bit StocQ quantizer (6.4) applies stochastic rounding corresponding to arbitrary bin split \(-1 = t_0 < ... < t_{2^b-1} = 1\) that is symmetric about 0. Denote \(S_i = t_{i-1} + t_i\) and \(P_i = t_{i-1}t_i, i = 1, ..., 2^b - 1\). Let \(f(z_u, z_v)\) be the RFF joint distribution in Theorem 6.1.4. Denote \(\kappa_{i,j} = \int_{t_{i-1}}^{t_i} \int_{t_{j-1}}^{t_j} z_u z_v f(z_u, z_v) dz_u dz_v,\) and \(p_{i,j} = \int_{t_{i-1}}^{t_i} \int_{t_{j-1}}^{t_j} f(z_u, z_v) dz_u dz_v.\) Then we have \(\mathbb{E}[\hat{k}_{\text{StocQ}}] = k(u, v)\) and \(\text{Var}[\hat{k}_{\text{StocQ}}] = \frac{V_{\text{StocQ}}}{m}\) with

\[
V_{\text{StocQ}} = 4 \sum_{i=1}^{2^b-1} \sum_{j=1}^{2^b-1} \left[ S_i S_j \kappa_{i,j} + P_i P_j p_{i,j} \right] - k(u, v)^2,
\]

which is always greater than \(\text{Var}[\hat{k}(u, v)]\) defined in (4.17).

The important take-away messages are: 1) the StocQ kernel estimator is unbiased of the Gaussian kernel; 2) the variance is always larger than full-precision RFF estimate. Further, we have the following result for 1-bit StocQ, which is a straightforward consequence of Theorem 6.3.1

**Corollary 6.3.1.** With 1-bit, \(V_{\text{StocQ}} = 4 - k(u, v)^2.\)

### 6.3.2 LM Estimators

In this subsection, we study the moments of the proposed LM kernel estimators. Since our following results generalize to both LM-RFF and LM²-RFF, we will unify the notation as \(Q\) to denote a LM-type quantizer. First, we have the following formulation of the mean estimate of LM quantized estimator (6.7) based on Chebyshev functional approximation.

**Theorem 6.3.2 (LM).** Let \(u, v\) be two normalized data samples with correlation \(\rho\), and \(\hat{k}_Q(u, v)\) be as (6.7) with LM quantizer \(Q\) (of either kind). Let \(z_x = \cos(\gamma X + \tau)\), \(z_y = \)
\[
\cos(\gamma Y + \tau) \text{ where } (X,Y) \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \ \tau \sim \text{uniform}(0,2\pi), \text{ and denote } \theta_{s,t} = \mathbb{E}[z_{s}^{*}Q(z_{s})^{t}], \ \zeta_{s,t} = \mathbb{E}[Q(z_{s})^{s}Q(z_{y})^{t}]. \]

Further define \( \alpha_{i} = \frac{2}{\pi} \int_{-1}^{1} Q(x)T_{i}(x) \frac{dx}{\sqrt{1-x^{2}}} \), \( \psi_{i,j} = \mathbb{E}[T_{i}(z_{x})T_{j}(z_{y})] \), where \( T_{i}(x) \) is the \( i \)-th Chebyshev polynomial of the first kind. Then we have

\[
\mathbb{E}[\hat{k}_{Q}(u,v)] = 4\theta_{1,1}^{2}k(u,v) + \sum_{i=1,\text{odd}}^{\infty} \sum_{j=3,\text{odd}}^{\infty} \alpha_{i}\alpha_{j}\psi_{i,j},
\]

\[
\text{Var}[\hat{k}_{Q}(u,v)] = \frac{4}{m}(\zeta_{2,2} - \zeta_{1,1}^{2}).
\]

In particular, \( \mathbb{E}[\hat{k}_{Q}(u,v)] = 4\theta_{1,1}^{2}k(u,v) \) when \( \rho = 0 \), and \( \mathbb{E}[\hat{k}_{Q}(u,v)] = 2\theta_{1,1} \) when \( \rho = 1 \).

Note that, for LM-RFF quantizer, we have \( 4\theta_{1,1}^{2} = (1 - 2D_{1})^{2} \) with \( D_{1} \) defined in (6.5).

Since Chebyshev polynomials form an orthogonal basis of the function space on \([-1, 1]\) with finite number of discontinuities, we can show that \( \alpha_{i} = \sqrt{2\theta_{1,1}}c_{i} \) where \( c_{i} \) is the cosine between \( Q(x) \) and \( T_{i}(x) \), and \( \sum_{i=3,\text{odd}}^{\infty} \alpha_{i}^{2} = 2(\theta_{1,1} - 2\theta_{1,1}^{2}) \) which is typically very small and decreases as the quantizer has more bits. Also, we have \( |\psi_{i,j}| \leq \mathbb{E}[T_{i}(z_{x})^{2}] = 1/2 \). Consequently, in Theorem 6.3.2 the last term approximates zero in most cases. This translates into the following observation.

**Observation 6.3.1.** \( \mathbb{E}[\hat{k}_{Q}(u,v)] \approx 4\theta_{1,1}^{2}k(u,v) \).

Next, we provide an asymptotic analysis on the normalized quantized kernel estimate (6.8) under LM scheme.

**Theorem 6.3.3 (Normalized estimator).** Under same setting as Theorem 6.3.2, as \( m \to \infty \),

\[
\mathbb{E}[\hat{k}_{n,Q}] = \frac{\zeta_{1,1}}{\zeta_{2,0}} + \mathcal{O}\left(\frac{1}{m}\right), \ \text{Var}[\hat{k}_{n,Q}] = \frac{V_{n}}{m} + \mathcal{O}\left(\frac{1}{m^{2}}\right),
\]

with \( V_{n} = \frac{\zeta_{2,2}}{\zeta_{2,0}^{2}} - \frac{2\zeta_{1,1}\zeta_{3,1}}{\zeta_{2,0}^{3}} + \frac{\zeta_{4,0}^{2} + \zeta_{2,2}}{2\zeta_{2,0}^{4}} \).
In particular, $E[\hat{k}_{n,Q}(u,v)] = 1$ when $\rho = 1$.

By the property of LM quantization that $E[Q(x)x] = E[Q(x)^2]$, we have

**Observation 6.3.2.** $E[\hat{k}_{n,Q}] \approx 2\theta_1 k(u,v)$ as $m \to \infty$.

Observation 6.3.1 and 6.3.2 says that, $E[\hat{k}_Q]$ and $E[\hat{k}_{n,Q}]$ approximately equal to some scaled version of true kernel, which will motivate our discussion in Section 6.4.3 on the robust kernel approximation error metrics.

![Figure 6.4: Observation 6.3.1 and 6.3.2 (black dash curves) vs. empirical bias (blue curves) of LM-RFF. Red dots are the biases given in the theorems at specific $\rho$ values.](image)

**Validation.** We plot the empirical bias of LM-RFF against Observations 6.3.1 and 6.3.2 in Figure 6.4. As we see, the proposed surrogates for bias align with true biases very well when $\rho$ is not very close to 1. The biases shrink to 0 as $b$ increases (e.g., negligibly $O(10^{-3})$ with $b = 4$). As $\rho \to 1$, at some "disjoint point" the absolute biases have sharp drops and quickly converge to the theoretical values (red dots) given in Theorem 6.3.2 and 6.3.3. As $b$ or $\gamma$ increases, the “disjoint point” gets closer to $\rho = 1$. 
Figure 6.5: Variance (scaled by \( m \)) of StocQ, LM, and normalized LM kernel estimators, with \( \gamma = 0.5, 1, 2 \). The dashed curves are the variances of corresponding full-precision estimators, to which the variance of quantized estimators converges as \( b \) increases.

6.3.3 Variance Comparisons

Figure 6.5 provides variance comparisons, where full-precision estimator variances are plotted for reference. As \( b \) gets larger, the variances of LM-type quantized estimators converges to those of full-precision estimators. The variance of StocQ is significantly larger than RFF and LM quantization, especially when \( b = 1, 2 \). This to a good extent explains why StocQ performs poorly in approximate kernel learning with low bits (Section 7.4).

**Variance of debiased kernel estimates.** As shown previously, LM estimators are slightly biased which brings theoretical challenges on finding a method to “properly” compare their variances. Similar to Chapter 2 and Chapter 4, we investigate the concept of “debiased variance” of the Gaussian kernel estimator, which refers to the estimator variance after bias corrections.

**Definition 6.3.1** (DB-variance). For data points \( u \) and \( v \) with \( \rho = u^T v \), and a kernel estimator \( \hat{k}(u, v; \rho) \) with \( E[\hat{k}] = E(\rho) > 0 \) and \( Var[\hat{k}] = V(\rho) \), the debiased variance of
\[ \hat{k}(u, v; \rho) \text{ at } \rho \text{ is defined as } \text{Var}^{db}[\hat{k}] = V(\rho) \cdot k(\rho)^2 / E(\rho)^2. \]

Note that, the debiasing step is only for analytical purpose. Intuitively, Definition 6.3.1 is reasonable in that it compares the variation of different estimation procedures given that they have same mean. It is worth mentioning that, DB-variance is invariant of linear scaling, i.e., \( c\hat{k} \) and \( \hat{k} \) have same DB-variance for \( c > 0 \). Classical metrics for estimation quality, such as the Mean Squared Error (MSE), might be largely affected by such simple scaling. Note that, the DB-variance of all 1-bit estimators (both simple and normalized) from fixed quantizers are essentially identical. This can be easily verified by writing every 1-bit quantizer as \( Q(z) = \text{sign}(z) \cdot C_Q \) for some \( C_Q > 0 \) and substituting it into (6.7) and (6.8). Thus, we will focus on multi-bit quantizers (i.e., \( b \geq 2 \)).

**LM-RFF v.s. LM\(^2\)-RFF.** In Figure 6.6, we provide the DB-variance ratio of LM\(^2\)-RFF estimator against that of LM-RFF estimator in the 2-bit case. (The observed pattern is the same for more bits.) For the simple kernel estimator, we see that in general LM-RFF has smaller DB-variance. Yet, the DB-variance of LM\(^2\)-RFF sharply drops towards 0 and beats LM-RFF as \( \rho \rightarrow 1 \), i.e., in high similarity region, which meets the goal of LM\(^2\)-RFF quantizer design (to favor high similarity region). However, for normalized estimators, \( \hat{k}_{n,Q} \) has consistently smaller DB-variance than \( \hat{k}_{n,Q,(2)} \).

**Benefit of normalization.** Next we prove the theoretical merit of normalizing RFFs, in terms of DB-variance.

**Theorem 6.3.4.** Suppose \( u, v \) are two samples with correlation \( \rho \). Let the simple and normalized kernel estimator, \( \hat{k}_Q \) and \( \hat{k}_{n,Q} \), be defined as (6.7) and (6.8), respectively, where \( Q \) is any LM-type quantizer. Assume \( \gamma \leq \pi / \sqrt{2} \). Then, \( \text{Var}^{db}[\hat{k}_{n,Q}] \leq \text{Var}^{db}[\hat{k}_Q] \) on \( \rho \in [0, 1] \) as \( m \rightarrow \infty \).

Theorem 6.3.4 says that when \( \gamma \leq \pi / \sqrt{2} \approx 2.2 \), normalization is guaranteed to reduce the DB-variance at any \( \rho \in [0, 1] \). In Figure 6.7, we plot the DB-variance ratio of \( \frac{\text{Var}^{db}[\hat{k}_{n,Q}(x,y)]}{\text{Var}^{db}[\hat{k}_Q(x,y) \]} \) at multiple \( \gamma \) and \( b \), for LM-RFF and LM\(^2\)-RFF respectively. We corroborate
Figure 6.6: Debiased variance ratio of LM$^2$-RFF over LM-RFF, at various $\gamma$, $b = 2$. **Left panel:** vanilla estimator $\frac{\text{Var}_{\text{db}}[\hat{k}_{Q,(2)}]}{\text{Var}_{\text{db}}[k_{Q}]}$. **Right panel:** normalized estimator $\frac{\text{Var}_{\text{db}}[\hat{k}_{n,Q,(2)}]}{\text{Var}_{\text{db}}[k_{n,Q}]}$.

the advantage of normalized estimates over simple estimators in terms of DB-variance (ratio always < 1), especially with large $\rho$.

6.3.4 Monotonicity of Mean Kernel Estimation

Similar to QRP (Chapter 2), monotonicity of the mean estimation against $\rho$ is also important for non-linear kernel estimation. For a kernel estimator $\hat{k}(\rho)$ (written as a function of $\rho$), the monotonicity of its mean estimation $\mathbb{E}[\hat{k}(\rho)]$ against $\rho$ is important to ensure its “correctness”. It guarantees that asymptotically ($m \to \infty$), the comparison of estimated kernel distances is always correct, i.e., $\hat{k}(u, v_1) > \hat{k}(u, v_2)$ if $k(u, v_1) > k(u, v_2)$ for data points $u, v_1, v_2$. Otherwise (say, $\mathbb{E}[\hat{k}]$ decreasing in $\rho$ on $[s, t]$), the comparison of estimated kernel would be wrong for $\rho \in [s, t]$ even with infinite much data. By Theorem 6.3.1, StocQ estimator is unbiased with $\mathbb{E}[\hat{k}_{\text{StocQ}}] = e^{-\gamma^2(1-\rho)}$ strictly increasing in $\rho$. Hence, we will focus on the fixed LM quantization.

**Lemma 6.3.5.** Suppose $X, Y, \tau$ are same as Theorem 6.3.2, and denote $s_x = \gamma X + \tau$, $s_y = \gamma Y + \tau$, such that $z_x = \cos(s_x)$ and $z_y = \cos(s_y)$ are RFFs. Assume $g_1, g_2 : [-1, 1] \mapsto \mathbb{R}$
Figure 6.7: Debiased variance ratio of normalized LM estimators against simple LM estimators, \( b = 2, 3 \). **Left panel:** LM-RFF, \( \frac{\text{Var}^{db}[\hat{k}_n,Q]}{\text{Var}^{db}[k_Q]} \). **Right panel:** LM\(^2\)-RFF, \( \frac{\text{Var}^{db}[\hat{k}_n,Q,(2)]}{\text{Var}^{db}[k_Q,(2)]} \).

are twice differentiable and bounded functions. Then,

\[
\frac{\partial \mathbb{E}[g_1(z_x)g_2(z_y)]}{\partial \rho} = \gamma^2 \mathbb{E}[g'_1(z_x) \sin(s_x)g'_2(z_y) \sin(s_y)].
\]

Furthermore, when \( \sqrt{2(1-\rho)} \gamma \leq \pi \), if \( g_1 \) and \( g_2 \) are both increasing odd functions or non-constant even functions, then the mean is increasing in \( \rho \), i.e., \( \frac{\partial \mathbb{E}[g_1(z_x)g_2(z_y)]}{\partial \rho} > 0 \).

Figure 6.8: The derivatives obtained by Lemma 6.3.5 and Theorem 6.3.6 with different \( g_1 \) and \( g_2 \) functions, at multiple \( \gamma \). Black curves are the function value, and red lines are the theoretical derivatives at multiple points. **Left panel:** \( g_1(x) = x^3 \), \( g_2(x) = e^{5x} \), increasing functions. **Middle panel:** \( g_1(x) = g_2(x) = x^2 \), even functions. The region \( \rho \geq 0 \) is magnified for clarity. **Right panel:** \( g_1(x) = g_2(x) = Q \), where \( Q \) is the 1-bit LM-RFF quantizer. Here the derivative is approximated by the continuous approximation \( \tilde{g}(x) = \mu_2 \text{sign}(x)(1 - e^{-50|x|}) \), where \( \mu_2 \) is the positive reconstruction level of \( Q \).
Lemma 6.3.5 is a general result for the monotonicity when RFFs are processed by continuous functions. In the left and mid panels of Figure 6.8, we plot two examples of $E[g_1(s_x)g_2(s_y)]$ against $\rho$, with continuously increasing functions $g_1(x) = x^3$ and $g_2(x) = e^{5x}$ and even functions $g_1(x) = g_2(x) = x^2$, respectively. As we can see, the expectation is increasing in $\rho$ with true derivatives given by Lemma 6.3.5.

Next, in the following theorem, we extend the above result to discrete functions, which include our proposed LM quantizers as special cases.

**Theorem 6.3.6.** Suppose $Q_1$ and $Q_2$ are bounded, discrete, and non-decreasing odd functions or non-constant even functions, with finite many discontinuities. Let $z_x$ and $z_y$ be defined as Lemma 6.3.5. If $\sqrt{2(1 - \rho)}\gamma \leq \pi$, then $E[Q_1(z_x)Q_2(z_y)]$ is increasing in $\rho$.

**Remark 6.3.1.** The condition $\sqrt{2(1 - \rho)}\gamma \leq \pi$ in Theorem 6.3.6 implies that $E[Q(z_x)Q(z_y)]$ for any LM quantizer increases in $\rho \in \left[\max(-1, 1 - \frac{\pi^2}{4\gamma^2}), 1\right]$. Thus, larger $\gamma$ typically requires higher $\rho$ for this condition to hold. For example, when $\gamma = 1$ and $\gamma = 5$, monotonicity is ensured for $\rho \in [-1, 1]$ and $\rho \geq 0.8$, respectively, which is consistent with numerical observations, e.g., see Figure 6.8.

In the right panel of Figure 6.8, we plot the mean kernel estimation $E[Q(z_x)Q(z_y)]$ (with $z_x, z_y$ being the RFFs) of 1-bit LM-RFF quantization, along with the derivatives predicted by Lemma 6.3.5. Here we use the continuous approximation $\tilde{g}(x) = \mu_2\text{sign}(x)(1 - e^{-50|x|})$, where $\mu_2$ is the reconstruction level of $Q$, as a surrogate to compute $Q'(x)$. We see that the mean estimation increases in $\rho$, and our theory on the derivative well aligns with the numerical result.

### 6.4 Numerical Experiments

We conduct experiments with compressed RFFs on both approximate kernel SVM (KSVM) classification and kernel ridge regression (KRR) tasks. Our results illustrate the effectiveness of large-scale kernel learning with highly compressed RFFs, especially highlighting the
superior advantage of the proposed LM-RFF quantization. Moreover, we also propose and evaluate robust kernel approximation error metrics to consolidate our claims.

![Graphs showing test accuracy vs memory for different datasets and compression schemes](image)

Figure 6.9: **Left two columns:** Test accuracy of kernel SVM using different compression schemes of RFFs vs. number of random features $m$. **Right two columns:** Test accuracy vs. memory per sample ($m \times b$).

### 6.4.1 Kernel SVM (KSVM)

For this task, we use four popular public datasets from UCI repository\(^2\) (Dua and Graff, 2017) and ASU database\(^3\) (Li et al., 2016a). All the data samples are pre-processed by instance normalization, and we randomly split each dataset into 60% for training and 40% for testing. For each task and each quantization method, the best test tuned accuracy is

\(^3\)https://jundongl.github.io/scikit-feature/datasets.html
reported, averaged over 10 independent runs.

To compare the learning power of different compression schemes, we provide the test accuracy vs. number of RFFs in the left two columns of Figure 6.9, with $b = 1, 2$. We observe: 1) LM-RFF substantially outperforms StocQ on all datasets with low bits. In particular, 1-bit StocQ performs very poorly, while 1-bit LM-RFF achieves similar accuracy as full-precision RFF; 2) On all datasets, LM-RFF with $b = 2$ already approaches the accuracy of full-precision RFF with moderate $m \approx 4000$, indicating the superior learning capacity of LM-RFF under deep feature compression. As one would expect, with larger $b$, the performance of StocQ approaches that of LM-RFF. In particular, when $b = 4$, LM-RFF and StocQ perform similarly on those datasets.

To characterize the memory efficiency\(^4\), note that under $b$-bit compression, each data sample requires $m \times b$ bits in total for storage. If we assume that each full-precision RFF is represented by 32 bits, then the storage cost per sample for full-precision RFF is $32m$. This allows us to plot the test accuracy against the total memory cost per sample, as shown in the right two columns of Figure 6.9. A curve near upper-left corner is more desirable, which means that the method requires less memory to achieve some certain test accuracy.

- We observe significant advantage of LM-RFF over full-precision RFF in terms of memory efficiency. For example, to achieve 95% accuracy on Isolet, LM-RFF (both 1-bit and 2-bit) requires $\approx 2000$ bits per sample, while RFF needs $\approx 18000$ bits, leading to a 9x compression ratio. Similar comparison holds for all datasets, and in general the compression ratio of LM-RFF is around 10x.

- When compared with StocQ, we see consistently advantage of LM-RFF in memory cost. In general, LM-RFF can further improve the compression ratio of StocQ by 2x~4x. Additionally, LM-RFF typically requires fewer-bit quantizers (smaller $b$) than StocQ to achieve satisfactory accuracy.

---

\(^4\)For simplicity, we mainly consider the memory cost for (quantized) RFF storage, which dominates in large-scale learning.
6.4.2 Kernel Ridge Regression (KRR)

For kernel ridge regression (KRR), we use a synthetic dataset admitting high non-linearity. Precisely, each data sample \( u \in \mathbb{R}^{10} \) is drawn from i.i.d. \( N(0, 1) \). We generate the response by
\[
y_i = \sum_{p=1}^{3} \beta_p x^p + \epsilon,
\]
where \( \beta_1 = [1, 2, \ldots, 10] \), \( \beta_2 = [1, 1, \ldots, 1] \), \( \beta_3 \) and \( \epsilon \) also follow i.i.d. \( N(0, 1) \). We simulate 40,000 independent samples for training and 10,000 for testing.

We summarize KRR results in Figure 6.10. Again, with same \( b \) and number of RFFs, LM-RFF consistently beats StocQ especially with low bits. We see that 1-bit LM-RFF even outperforms 2-bit StocQ, and when \( b = 4 \), we still observe considerable advantage of LM-RFF over StocQ. In the second row, we present the memory efficiency comparison. Note that, due to high-order terms in the true model, the test MSE of linear kernel is 20.8, while learning with full-precision RFF significantly reduces it to 3.5. With largest memory budget that is tested, 1-bit and 2-bit LM-RFF yield 5.9 and 4.1 test MSE respectively, which are already quite close to 3.5, while for 1-bit and 2-bit StocQ, the test losses are 14.5 and 5.0 respectively, much worse than those of LM-RFF. We again see significant storage saving of LM-RFF. For instance, to reach the same test MSE (e.g., 10), the compression ratio is about 5x for \( b = 4 \) compared with full-precision RFF. Moreover, the advantage of LM-RFF over StocQ is also significant for this regression problem.

6.4.3 Scale-invariant Kernel Approximation Error

Recall the notation \( U = [u_1, \ldots, u_n]^T \) as the data matrix. Let \( \mathcal{K} \) be the \( n \times n \) Gaussian kernel matrix, with \( \mathcal{K}_{ij} = k(u_i, u_j) \). Denote \( \hat{\mathcal{K}} \) as the estimated kernel matrix by an approximation algorithm. Kernel Approximation Error (KAE) has been shown to play an important role in the generalization of learning with random features, including the norms (Cortes et al., 2010; Gittens and Mahoney, 2013; Sutherland and Schneider, 2015) of \( \hat{\mathcal{K}} - \mathcal{K} \) and spectral approximations (Bach, 2013; Alaoui and Mahoney, 2015; Avron et al., 2017; Zhang et al., 2019). We investigate the KAEs to better justify the impressive generalization ability of LM-RFF from a theoretical aspect.
Figure 6.10: Kernel ridge regression: test MSE of KRR, StocQ vs. LM-RFF. **Upper panels:** Test MSE vs. number of random features $m$. **Bottom panels** Test MSE vs. memory per sample ($m \times b$ bits).

However, existing KAE metrics are not robust to bias. Consider $E[\hat{K}] = \beta K$ with some $\beta > 0$. Obviously, learning with $\beta K$ is equivalent to learning with $K$ for kernel-distance based models like KSVM and KRR, since with proper scaling of model parameters, the objective functions/predictions are invariant of multiplying the input kernel matrix with a scalar. However, traditional KAEs do not generalize to this case. For example, when $\beta = 0.1$, the 2-norm error $\|0.1\hat{K} - K\|_2$ would be very large. To make the KAE metrics more robust, we define the scale-invariant KAE metrics as follows.

**Definition 6.4.1 (Scale-Invariant KAE).** Let $K$ be a kernel matrix and $\hat{K}$ be its randomized approximation. We define

\[
\|\hat{K} - K\|_2 = \min_{\beta > 0} \|\beta \hat{K} - K\|_2, \quad \|\hat{K} - K\|_F = \min_{\beta > 0} \|\beta \hat{K} - K\|_F.
\]
Denote the minimizers as $\beta_{2}^*$ and $\beta_F^*$, respectively. Define

$$
(\delta_1^*, \delta_2^*) = \inf_{\beta \in \{\beta_{2}^*, \beta_F^*\}} \{ \delta_1, \delta_2 : (1 - \delta_1) K \preceq \beta \hat{K} \preceq (1 + \delta_2) K \}.
$$

Our new KAE metrics are more general, adapted to the best scaling factor $\beta_{2}^*$ or $\beta_F^*$ of the estimated kernel. Since LM-RFF estimators are slightly biased (recall Observations 6.3.1 and 6.3.2), Definition 6.4.1 is important for appropriately evaluating our proposed LM-RFF kernel estimation approach. In Figure 6.11, we provide scale-invariant $\| \cdot \|_2^*$, $\| \cdot \|_F^*$ and $\delta_2^*$ metrics\(^5\) on Isolet and BASEHOCK dataset as representatives. As we can see, LM-RFF always has smaller KAEs than StocQ with equal bits. In particular, with extreme 1-bit compression, StocQ has exceedingly large loss due to its large variance, while in many cases the KAEs of 1-bit LM-RFF are already quite small. The KAE comparison well aligns with, and to an extent explains, our experimental results in Section 6.4.1 and Section 6.4.2 that 1) LM-RFF consistently outperforms StocQ, and 2) 1-bit StocQ generalizes very poorly. Thus, it provides a general justification of the superior effectiveness of LM-RFF in machine learning.

### 6.5 Conclusion

The technique of random Fourier features (RFF) is a popular method to solve the computational bottleneck in large-scale (Gaussian) kernel learning tasks. In this chapter, we study quantization methods to compress RFFs for substantial memory savings and efficient computations. In particular, we focus on developing quantization algorithms based on the Lloyd-Max (LM) framework and propose two methods named LM-RFF and LM\(^2\)-RFF. In addition, we also analyze a method based on stochastic rounding (StocQ). Both theoretically and empirically, LM-RFF significantly outperforms StocQ on many tasks, especially when

---

\(^5\)Zhang et al. (2019) found that for kernel approximation methods, $\delta_2$ is fairly predictive of the generalisation performance.
Figure 6.11: Scale-invariant kernel approximation error (Definition 6.4.1) of LM-RFF vs. StocQ, $m = 2^{10}$. Black curves are full-precision RFF, blue curves are StocQ, and curves in red represent our proposed LM-RFF. For all metrics, the smaller the better.

the number of bits is not large. Compared to full-precision (e.g., 32- or 64-bit) RFFs, the experiments imply that often a 2-bit LM-RFF quantizer achieve comparable performance with full-precision, at a substantial (e.g., 10x) saving in memory cost, which would be highly beneficial in practical applications.
CHAPTER 7
ONE-SKETCH-FOR-ALL: NON-LINEAR RANDOM FEATURES FROM COMPRESSED LINEAR MEASUREMENTS

In Chapter 6, we studied distortion optimal Lloyd-Max (LM) quantizer for RFF, and compared it with stochastic rounding technique. Practically, the performance of LM quantization is consistently better, which can even match the full-precision RFF under very high compression ratio (e.g., 1 bit or 2 bits per RFF). Thus, significant storage cost reduction can be achieved by LM-RFF. Yet, this method may have some practical limitations, see Section 7.1.

In this chapter, our goal is to resolve these limitations, by proposing an alternative highly convenient RFF compression approach. More specifically, we directly construct the non-linear random features from quantized RPs (QRP), which has been thoroughly studied in Chapter 2 and Chapter 4. This strategy only requires storing one set of QRPs in memory, which can be used for both large-scale linear and non-linear learning.

7.1 Motivation of This Chapter

In this chapter, we consider the second formulation of RFF, namely (4.16). Nevertheless, the idea and analysis also apply to the first type of RFF. In (4.16), constructing RFF can actually be considered as a two-stage procedure: (i) random projection (RP): $w^T x$; (ii) applying non-linearity ($\sin$ and $\cos$ function). All the existing works on RFF compression, including Chapter 6, seek to quantize the RFFs after the non-linearity, thus requiring access to the original full-precision random projections. In other words, in such procedure, one would have to store a set of quantized RFFs for each different tuning parameter $\gamma$ in the Gaussian kernel. When the best tuning parameter is already known (e.g., from prior experience), then the methods can be adopted feasibly. Yet, the best tuning parameter might be unknown, especially in the early stage of the exploration. This means that practitioners might have to
store multiple sets of quantized RFFs. Furthermore, in the databases where linear sketches (i.e., the random projections (RPs)) are also stored at the same time for some learning tasks (e.g., similarity search), requiring a copy of full-precision RPs may also take a lot of memory for large-scale datasets. The motivation of this chapter is to remove this burden, by developing an even more memory-efficient and convenient compression approach for RFF.

In this chapter, we consider an alternative approach, where the quantization is applied **before the non-linearity**. That is, we first derive the quantized RP as $Q(w^T x)$ in step (i), which is then used to construct non-linear RFF in step (ii). We call it Quantized Random Fourier Feature (QRP-RFF) scheme. The title of this chapter, inspired by Gilbert et al. (2007); Li et al. (2006a), characterizes the key advantage of the proposed QRP-RFF approach—it achieves “one-sketch-for-all” because we only need one set of highly compressed linear measurements (i.e., quantized RPs), for both linear and non-linear learning. That is, by sparing the (unnecessary) space for the full-precision RPs, QRP-RFF fulfills memory-efficient linear and non-linear learning at the same time. Compared with LM-RFF scheme in Chapter 6, the scheme studied herein would lose some accuracy as one would expect. Nevertheless, the proposed method still perform noticeably better than the quantization scheme based on random rounding, especially with low precision compression. We will provide theoretical analysis on the QRP-RFF kernel estimator and the approximation error, and experiments are conducted to demonstrate the efficacy of the proposed approach.

### 7.1.1 Practical Significance

Our QRP-RFF framework allows us to only store one set of quantized random projections (QRPs) in the database, without requiring access to the full-precision RPs (FP-RPs). As we have discussed throughout the dissertation, the “intermediate product”, namely the QRP, is itself a useful tool in machine learning, with a wide range of applications in theory, linear learning, similarity search, compressed sensing, etc. See cited references in Chapter 1. We also recap the applications of QRP in Figure 7.1. Our proposed strategy also has significant
Figure 7.1: Recap: applications of QRP in large-scale systems. Here, the full-precision RPs are never stored in memory. QRP can be used for linear learning and compressed sensing. The box in red is the new application studied herein—constructing non-linear random features for non-linear learning.

practical value from the perspective of QRP.

Consider the following practical scenario, where a server has collected RPs of massive data samples and stored the quantized RPs (QRPs) in the database to save storage. In this procedure, we have lost access to the full-precision RPs (otherwise, quantization becomes meaningless). In order to achieve better learning performance, a data scientist wants to apply non-linear kernel learning. Typically, this can be done in a standard way by learning with RFFs generated by RPs. Yet, full-precision RPs (FP-RPs) are discarded after quantization, and re-collecting the data might be inconvenient or even impossible (e.g., due to data loss or privacy). Our QRP-RFF method exactly provides a solution in this case, by directly extracting non-linear sketches from QRPs. Therefore, QRP-RFF can be viewed as the first extended application of QRP to non-linear kernel learning, arising from very practical settings.

7.1.2 Notations

In this chapter, for the ease of notations, we switch to denote the data matrix as \( X \in \mathbb{R}^{n \times d} \). Same as before, we assume all the samples are normalized. \( \rho \) is used to denote the correlation
between sample $x$ and $y$. Since we have shown in previous chapters that fixed quantization typically performs better than stochastic quantization, we will mainly consider the fixed Lloyd-Max (LM) and uniform quantization for RPs, same as in Chapter 2 and Chapter 4. Also, we will mainly compare our method with StocQ quantization method. Detailed introduction on these quantizers can be found in Chapter 1.2.

7.2 QRP-RFF Scheme

As shown in (4.16), RFF is built upon RP with one extra step of casting non-linearity. Given the popularity and wide application of QRP, one natural question arises: can we extract RFF from QRP for fast non-linear kernel learning? Recalling Figure 7.1, since one typically would like to discard the full-precision RPs to spare unnecessary storage once they have been quantized and stored in the database, this is a more practical setting worth studying. We call the proposed method “QRP-RFF”, as depicted in Figure 7.2. It consists of three steps:

1. Apply random projection $X_W = X W_\gamma$.

2. Quantize the projected data $X_Q = Q(X_W)$ which will be stored in database. We can discard $X_W$ afterwards and use compressed $X_Q$ for various subsequent linear learning tasks.

3. Extract QRP-RFFs from the quantized $X_Q$, which can then be fed into linear machines for fast approximate non-linear kernel learning.

In step 2, we propose to use “linear” quantizers that satisfy the following property.

**Definition 7.2.1.** In the context of Gaussian QRP, let $Q$ be the quantizer w.r.t. $N(0, 1)$. The quantization scheme is called linear if for any $\gamma > 0$, $\gamma Q$ is the corresponding quantizer for $N(0, \gamma^2)$.

In particular, it is easy to check that fixed uniform and LM quantizer both fall into this category. The linearity of quantizer allows us to derive QRP for any $\gamma$, from the QRP with
Figure 7.2: Illustration of QRP-RFF framework. QRP-RFF is constructed directly from QRP. Full-precision RPs can be discarded after QRPs are generated.

\(\gamma = 1\). This has a crucial impact on the parameter tuning of QRP-RFF. We can simply store one set of compressed linear sketch (e.g., QRPs with \(\gamma = 1\)) in memory to tune QRP-RFF with any \(\gamma\) by scaling—This is the essential reason that QRP-RFF does not require FP-RP and achieves “one-sketch-for-all”. On the contrary, if linearity does not hold, we will have to use the original FP-RP to re-construct quantized sketches for each distinct \(\gamma\), violating our problem setting.

Following step 3, we formally define the QRP-RFF as

\[
F_Q(x) = [\sin(Q(w^T x)) \cos(Q(w^T x))]^T,
\]

(7.1)

where \(w \sim N(0, \gamma^2)\) and \(Q\) is a linear quantizer (Definition 7.2.1). Here we consider the second type of RFF (4.16) because it is more convenient for building look-up tables (the QRP-RFF is also discrete). Moreover, this type of RFF has been shown to have smaller kernel estimation variance than Type-I. Analogously, by \(m\) i.i.d. projections, we defined
the QRP-RFF kernel estimator by

$$\hat{k}_Q(x, y) = \frac{1}{m} \sum_{i=1}^{m} F_{Q,i}(x)^T F_{Q,i}(y),$$  (7.2)

where $F_{Q,i}(x)$ is the $i$-th QRP-RFF of $x$ associated with projection $w_i$. We re-emphasize the significance of our QRP-RFF scheme: by directly retrieving non-linear random features from QRP, QRP-RFF does not need FP-RP or many sets of quantized RFFs (for different $\gamma$ values). Instead, only one set of compressed QRP is needed for both linear and non-linear learning.

7.3 Analysis

In this section, we discuss properties of QRP-RFF kernel estimators and provide the theoretical analysis.

7.3.1 Equivalence in kernel learning when $b = 1$

In practice, 1-bit compression, e.g. 1-bit random projection, is an important special case of quantization because it achieves highest compression ratio. For linear RP, one can easily point out that all 1-bit quantization methods are equivalent, when the task is to estimate the cosine $\rho$. Every 1-bit quantizer can be written as $Q(w^T x) = \gamma c_Q \cdot \text{sign}(w^T x)$ with some quantizer-specific constant $c_Q$. That is, different 1-bit quantizers only differ by a constant scaling factor, which can be easily fixed by re-scaling when estimating $\rho$ by the inner product $Q(w^T x)Q(w^T y)$. However, it is obvious that for QRP-RFF, linear scaling of the kernel estimate no longer holds due to the high non-linearity of sine and cosine functions. Nevertheless, we have a weaker statement of equivalence. For QRP-RFF, all 1-bit fixed linear quantizers are equivalent in non-linear kernel learning models, provided that $\gamma$ is tuned properly. When $b = 1$, Eq. (7.1) can be written in the general form $F_Q(x) = [\sin(\gamma c_Q \cdot \text{sign}(w^T x)) \cos(\gamma c_Q \cdot \text{sign}(w^T x))]^T$ with some quantizer-specific $c_Q$. Then, the QRP-RFFs generated by $Q_1$ with $\gamma_1$ can be produced by $Q_2$ with $\gamma_2 = \frac{c_1}{c_2} \gamma_1$. In
words, the difference in $c_Q$ can be eliminated in practice by tuning $\gamma$ adequately. As a result, in principle, the learning performance of all 1-bit quantizers for QRP-RFF are expected to be the same with fine tuning.

7.3.2 Information loss of QRP-RFF

From now on, we will denote $k_Q(x, y) = \mathbb{E}[\hat{k}_Q(x, y)]$, or $k_Q$ in short. This is sometimes referred as “expected kernel” in kernel approximation literature. One inevitable issue of extracting RFF directly from fixed quantized random projections, is the information loss in the transaction from linear projections to highly non-linear sine and cosine functions, especially for large $\gamma$ and small bits $b$. The reason is that, when $\gamma$ is large, the difference $|Q(z) - z|$ might be so large that $\sin(Q(z))$ and the FP-RFF $\sin(z)$ (and cosine) are very different, where $z = w^T x \sim N(0, \gamma^2)$ is the RP. When $b$ is small, the deviation is even larger.

We will use the estimation at a single point as an example. When $b = 1$, we can compute the LM quantizer as $Q(z) = \text{sign}(z) \times 0.7979\gamma$. Thus we can explicitly compute $k_Q$ at $\rho = -1$ as $-\sin(0.7979\gamma)^2 + \cos(0.7979\gamma)^2$. This is a periodic function in $\gamma$ that deviates significantly from the true kernel value at $\rho = -1$, as depicted in Figure 7.3. We see that the mean (at $\rho = -1$) is only reasonable with $\gamma \leq 1$. With larger $\gamma$, the estimation becomes wild. Similarly instability holds for other $\rho$. Unfortunately, this unstable behavior is caused by the nature of the problem, i.e., the information loss of coding with discrete $Q$ in the “linear” QRP to “non-linear” QRP-RFF transaction. Nevertheless, as will be shown in Section 7.3.3, when $b \geq 3$, the information loss becomes acceptable as the mean estimation of QRP-RFF approaches the true RBF kernel.
Figure 7.3: Information loss of QRP-RFF: Mean of 1-bit QRP-RFF estimate from LM quantized QRP, at \( \rho = -1 \). The red curve is the true kernel, and the blue curve is the 1-bit QRP-RFF mean.

7.3.3 Mean and variance

**Theorem 7.3.1.** Let \( Q \) be a \( b \)-bit fixed quantizer with borders \( -\infty = t_0 < t_1 < \ldots < t_{2^b} = +\infty \) and reconstruction levels \( \mu_1 < \ldots < \mu_{2^b} \). Suppose \( u, v \sim N \left( 0, \gamma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \), and let \( p_{ij} = P(u \in [t_{i-1}, t_i], v \in [t_{j-1}, t_j]) \) for \( 1 \leq i, j \leq 2^b \). Denote \( s_i = \sin(\mu_i) \) and \( c_i = \cos(\mu_i) \). For normalized data vectors \( x \) and \( y \),

\[
 k_Q \triangleq \mathbb{E}[\hat{k}_Q(x, y)] = \sum_{i=1}^{2^b} \sum_{j=1}^{2^b} (s_is_j + c_ic_j) p_{ij},
\]

\[
 Var[\hat{k}_Q(x, y)] = \frac{1}{k} \left\{ \sum_{i=1}^{2^b} \sum_{j=1}^{2^b} (s_is_j + c_ic_j)^2 p_{ij} - k_Q^2 \right\}.
\]

With the potentially severe instability of low-bit QRP-RFF estimate in mind, in Figure 7.4 we plot \( k_Q \) with different \( b \) and \( \gamma \), when the QRP is quantized by LM quantizers. We observe the mild estimation when \( b = 1, 2 \) at some \( \gamma \) value. As expected, as \( b \) increases, \( k_Q \) converges to the true kernel \( K \).

**Remark 7.3.1.** It is important to understand that, \( k_Q \) deviating from the exact RBF kernel...
does not imply bad generalization performance of QRP-RFF. On the one hand, the performance of randomized algorithms also largely relies on the variance (e.g., the large variance of low-bit LP-RFF results in poor learning capacity, though it is unbiased). On the other hand, in some sense we can regard the QRP-RFF estimators as converging to some other kernel, and comparing the learning capacity of two non-linear kernels is non-trivial and in general data-dependent. In Section 7.4, we will provide empirical results demonstrating the empirical effectiveness of our proposed QRP-RFF method.

Due to the information loss, the intrinsic instability of QRP-RFF estimator makes it difficult to obtain elegant theoretical results on the expected kernel $k_Q$ (e.g., recalling Figure 7.3). Nonetheless, we still provide analytical bounds on $k_Q$ showing its concentration around the RBF kernel. The following is a general result holding for any quantizer $Q$. 

Figure 7.4: Solid curves: the mean of QRP-RFF estimate (Theorem 7.3.1). Dash curves: the true RBF kernel. We see some large deviations when $b = 1, 2$. 

\[
\begin{align*}
\text{Kernel} & \quad \rho \\
\text{b = 1} & \\
\gamma = 0.1 & \\
\gamma = 0.5 & \\
\gamma = 1 & \\
\gamma = 2 & \\
\gamma = 5 & \\
\text{b = 2} & \\
\gamma = 0.1 & \\
\gamma = 0.5 & \\
\gamma = 1 & \\
\gamma = 2 & \\
\gamma = 5 & \\
\text{b = 3} & \\
\gamma = 0.1 & \\
\gamma = 0.5 & \\
\gamma = 1 & \\
\gamma = 2 & \\
\gamma = 5 & \\
\text{b = 4} & \\
\gamma = 0.1 & \\
\gamma = 0.5 & \\
\gamma = 1 & \\
\gamma = 2 & \\
\gamma = 5 & 
\end{align*}
\]
Theorem 7.3.2. For any fixed $\gamma$, let $z \sim N(0, \gamma^2)$, define $D_s = \mathbb{E}[(\sin(Q(z)) - \sin(z))^2]$, $\zeta_s = \text{Cov}(\sin(Q(z)) - \sin(z), \sin(z))$, and $D_c$ and $\zeta_c$ analogously for cosine function. Further denote $\Delta_c = \mathbb{E}[\cos(Q(z))] - e^{-\gamma^2/2}$ and $\tilde{D}_c = D_c - \Delta_c^2$. Denote $V_s^* = \frac{1}{2} \left[ 1 - e^{-2\rho^2\gamma^2} \right]$ and $V_c^* = \frac{1}{2} \left[ 1 + e^{-2\rho^2\gamma^2} \right] - e^{-\gamma^2/2}$. Assume $x, y$ are two normalized samples with correlation $\rho$. Then at $\gamma$, $k_Q(x, y)$ is lower and upper bounded respectively by

$$k(x, y) - D_s - D_c + 2e^{-\gamma^2(1-\rho^2)/2} (C_{1-} + C_{2-}),$$

$$k(x, y) + D_s + D_c + 2e^{-\gamma^2(1-\rho^2)/2} (C_{1+} + C_{2+}),$$

where

$$C_{1\pm} = (C_1 C_2 \pm \sqrt{(1-C_2^2)(1-C_3^2)}) \sqrt{D_s V_s^*},$$

$$C_{2\pm} = [(C_3 C_4 \pm \sqrt{(1-C_3^2)(1-C_4^2)}) \sqrt{D_c V_c^*} + e^{-\gamma^2/2} \Delta_c],$$

with

$$C_1 = \sqrt{\frac{2\zeta_s}{D_s(1 - e^{-2\gamma^2})}}, \quad C_2 = \frac{e^{-\gamma^2(1-\rho^2)/2} - e^{-\gamma^2(1+\rho^2)/2}}{\sqrt{2(1 - e^{-2\gamma^2})V_s^*}},$$

$$C_3 = \sqrt{\frac{\zeta_c}{D_c(\frac{1}{2} [1 + e^{-2\gamma^2}] - e^{-\gamma^2})}}, \quad C_4 = \frac{1}{2} \left[ e^{-\gamma^2(1-\rho^2)/2} + e^{-\gamma^2(1+\rho^2)/2} \right] - e^{-\gamma^2(1+\rho^2)/2} \sqrt{(\frac{1}{2} [1 + e^{-2\gamma^2}] - e^{-\gamma^2})V_c^*}. $$

The proof is given in Appendix F. Theorem 7.3.2 gives a universal bound on the QRP-RFF mean for any $Q$ at any $\gamma$ and $\rho$, which states that $k_Q$ “concentrates” around $K$ with error no more than $O(D_s + D_c + \sqrt{D_s} + \sqrt{D_c})$. Thus, smaller non-linear distortions lead to stronger concentration, as smaller $D_s$ and $D_c$ imply better approximation of QRP-RFF to RFF. As $b \to \infty$, the distortions go to 0 and $k_Q$ converges to $K$.

**Variance.** Another important factor that affects the learning performance with randomized algorithms is the variance of the estimation. In Figure 7.5, we plot variances of full-precision
Figure 7.5: Variance of a random feature of FP-RFF, StocQ and QRP-RFF (Theorem 7.3.1, to be scaled by $k$). The variance of StocQ follows from Theorem 6.3.1.

RFF, QRP-RFF with LM quantization and StocQ estimators at representative $\gamma$ levels. We see that at low bit constraint $b = 1, 2$, the stochastic StocQ has much larger variance than QRP-RFF estimators. Consequently, StocQ may perform worse than QRP-RFF in low-bit training. We omit the figures for more bits, since the variance of QRF-RFF converges to that of FP-RFF as expected.

7.3.4 Uniform Approximation Error

In practice, it is favorable to produce and store as few RFFs as possible to achieve small approximation error to the true RBF kernel. Similarly, we are also interested in the sufficient number of QRP-RFFs to approximate $k_Q$ within some pre-defined error. In this context, it is important to understand the sample complexity of QRP-RFF, measured by the uniform approximation error $\sup_{x,y\in \mathcal{X}} |\hat{k}_Q(x, y) - k_Q(x, y)|$. That is, we would like to bound the deviation over the data domain $\mathcal{X}$. For full-precision RFF (Rahimi and Recht, 2007; Sutherland and Schneider, 2015), $m$ is required to be at least $O(\frac{d}{\epsilon^2} \log \frac{1}{\epsilon})$ to guarantee $\epsilon$-approximation with high probability. To proceed, we first introduce the following definition.
**Definition 7.3.1. (Mean Smooth Quantizer)** We say a quantizer $Q(\cdot)$ is mean Lipschitz smooth w.r.t. distribution $\Gamma$ and function $f$ with constant $L^f_Q$, if for $\forall \delta > 0$, the following holds,

$$
\mathbb{E}_{t \sim \Gamma} \left[ \sup_{|r| \leq \delta} |f(Q(t + r)) - f(Q(t))| \right] \leq L^f_Q \delta. \tag{7.3}
$$

Basically, quantizer $Q$ is mean smooth if the average maximal deviation of a function $f$ applied to the quantized random measurements from $\Gamma$ is bounded in a Lipschitz way. This is a an “averaged” version of Lipschitz continuity, which also works for discrete functions. Definition 7.3.1 is a generalization of (Schellekens and Jacques, 2020) which was restricted to periodic functions. In our problem where $f$ is sine or cosine, $f \circ Q$, when composited as one function, is no longer periodic. By extending the characterization to a more general setting, the uniform approximation error of QRP-RFF is given as below.

**Theorem 7.3.3. (Uniform Approximation Error)** Assume the sample space $S$ is the unit sphere (normalized data). Let QRP-RFF estimators be defined as (7.2). Let $\Gamma \sim N(0, \gamma^2)$ in Definition 7.3.1. Suppose a quantizer $Q$ is mean smooth w.r.t. sin and cos functions with Lipschitz constant $L^s_Q$ and $L^c_Q$, respectively. Then for $\forall \epsilon > 0$, with probability at least $1 - 4e^{-ke^2/256}$,

$$
|\hat{k}_Q(x, y) - k_Q(x, y)| \leq \epsilon, \quad \text{for } \forall x, y \in S,
$$

when $m \geq \frac{512d}{\epsilon^2} \log \left( \frac{64 \max\{L^s_Q, L^c_Q\} \gamma}{\epsilon} + 1 \right)$.

The proof is given in Appendix F. Theorem 7.3.3 says that to achieve $\epsilon$-error, the sample complexity of QRP-RFF is the same as that of full-precision RFF, within constant factor. We now show that for our problem where $f$ is the sin or cos function and $\Gamma \sim N(0, \gamma^2)$, every bounded quantizer with finite bits is mean smooth. Hence, the error bound in Theorem 7.3.3 is a general theoretical result holding for all finite fixed quantization, including the LM and uniform quantizers.
Proposition 7.3.4. When $f$ is sin or cos function and $\Gamma \sim N(0, \gamma^2)$ in Definition 7.3.1, every finite-bit bounded quantizer is mean Lipschitz smooth.

7.4 Experiments

In this section, we test the learning performance of QRP-RFF scheme in kernel classification problems. The main purpose is to show that (i) QRP-RFF performs better than StocQ with low bits; and (ii) when $b$ is as large as 4, the performance of QRP-RFF is similar to the full-precision RFF.

Setting. We compare three randomized approximations: 1) the full-precision RFF; 2) QRP-RFF with underlying LM quantization; and 3) stochastic quantization (StocQ). For approaches involving quantization, after the FP-RFFs are generated, we process them with corresponding quantization strategy, then feed them into a linear SVM solver. We tune the parameters $C$ for SVM and $\gamma$ for RBF kernel over a wide range of values. We use public datasets from UCI machine learning repository (Dua and Graff, 2017). For all datasets, the samples are normalized to have unit norm. On each dataset, we randomly split the samples into 60% for training and 40% for testing. For each method, the best test accuracy among $C$ and $\gamma$ is reported, averaged over 10 independent repetitions.

Results. In Figure 7.6, we report the classification test accuracy against the number of RFFs used, with $b = 1, 2, 4$. We observe the following:

- Low-bit training. For $b = 1$, on all datasets, we observe significant higher accuracy of QRP-RFF over StocQ. On ISOLET, QRP-RFF with $b = 1$ almost achieves the same accuracy as full-precision RFF. For $b = 2$, QRP-RFF also compares favorably with StocQ, when the number of random sketches is moderate (i.e., more than $2^{10}$). The poor performance of StocQ can be partially explained by its large variance with low bits (see Figure 7.5 and Theorem 6.3.1).

\footnote{We implemented StocQ to both aforementioned formulations of RFFs, and found very similar performance.}
Figure 7.6: Test accuracy of (linearized) kernel SVM using RFF with different quantization strategies. The dash line is for standard RBF kernel SVM.

- **More bits.** As we use more bits, the test accuracy of both quantization methods gets improved. For QRP-RFF, on all datasets, \( b = 4 \) is sufficient to approach the performance of FP-RFF, with moderate number of random features.

- **Memory saving.** The benefit of QRP-RFF in terms of storage saving becomes obvious
given Figure 7.6. Since 4-bit QRP-RFF almost has same test accuracy as using FP-RFF, the storage can typically be reduced by at least \( \frac{32}{4} = 8x \) or \( 16x \), when FP-RFFs are represented by 32 bits or 64 bits, respectively.
CONCLUSION OF PART II

In Part II of the dissertation, we extend Part I by considering non-linear learning with random features, and studying the corresponding compression strategies.

In Chapter 5, we first show how Random Fourier Features (RFF) can be applied to multi-view discriminant analysis problems to elevate the classification accuracy by kernel approximation. Perturbation bounds are provided on the approximated subspaces, and a new variant of Davis-Kahan’s $\sin \Theta$ Theorem is derived.

In Chapter 6, we apply distortion optimal quantization under the Lloyd-Max (LM) scheme to RFF, which is named LM-RFF method. Since the marginal distribution of RFF is independent of the Gaussian kernel parameter $\gamma$, which leads to convenient quantizer design. We also study a variant called LM$^2$-RFF which is theoretically better in high similarity region. The moments of LM-RFF kernel estimators, debiased variance comparison and monotonicity of the mean estimation is analyzed. Empirical experiments show that the proposed LM-RFF is superior on kernel classification and regression tasks compared with stochastic rounding, especially under very low-bit compression.

In Chapter 7, we address a potential limitation of LM-RFF approach in practice, that one has to store many sets of quantized non-linear sketches for different $\gamma$. We proposed to directly construct RFFs based on QRPs, which enables us to only store one set of highly compressed QRPs in the database, for both linear and non-linear learning. This is typically a more practical setting, since full-precision RPs are usually discarded after they are quantized and stored in the data server. Of course, this approach is less accurate than LM-RFF. Yet, it still outperforms stochastic rounding strategy especially when the number of bits is small. Therefore, it is a feasible alternative in practice for memory-efficient non-linear learning with Random Fourier Features.


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


**APPENDIX A**

**PROOFS OF CHAPTER 2**

**Proofs of Scenario 1**

For Scenario 1, recall the notations,

\[
\gamma_{\alpha,\beta} = \mathbb{E} \left( Q_b(x)^\alpha y^\beta \right), \quad \xi_{\alpha,\beta} = \mathbb{E} \left( Q_b(x)^\alpha x^\beta \right),
\]

(A.1)

where

\[
\begin{pmatrix}
    x \\
    y
\end{pmatrix}
\sim
\mathcal{N}
\left(
    \begin{pmatrix}
        0 \\
        0
    \end{pmatrix},
    \begin{pmatrix}
        1 & \rho \\
        \rho & 1
    \end{pmatrix}
\right).
\]

(A.2)

Also denote \( x = (x_1, \ldots, x_k), \ y = (y_1, \ldots, y_k) \).

The following Lemma is a known result of Lloyd-Max (LM) quantizer. We provide a proof here which will be helpful for our analysis and will appear at several places.

**Lemma A.1.** Let \( Q_b \) be a \( b \)-bit Lloyd-Max quantizer optimized with respect to an arbitrary probability distribution \( f \), with distortion \( D_b \) Suppose random variable \( x \sim f \), then

\[
\xi_{1,1} = \xi_{2,0} = 1 - D_b.
\]

Furthermore, if \( f \) is standard normal distribution, then \( \xi_{1,1} = \xi_{2,0} \leq 1 \).

**Proof.** Recall that each reconstruction levels of LM quantizer are the conditional expectations on the corresponding separated regions. Let \( t_0 < t_1 < \cdots < t_M \) be the boarders. We
have
\[
\mathbb{E}(Q_b(x)x) = \sum_{i=1}^{M} \int_{t_{i-1}}^{t_i} \frac{xf(x)dx}{\int_{t_{i-1}}^{t_i} f(x)dx} \int_{t_{i-1}}^{t_i} xf(x)dx
\]
\[
= \sum_{i=1}^{M} \left( \frac{\int_{t_{i-1}}^{t_i} xf(x)dx}{\int_{t_{i-1}}^{t_i} f(x)dx} \right)^2 \int_{t_{i-1}}^{t_i} f(x)dx = \mathbb{E}(Q_b(x)^2).
\]

If \( f(x) = \phi(x) \) which is standard Gaussian density, we have
\[
1 - D_b = 1 - \mathbb{E}((x - Q_b(x)^2)) = 2\mathbb{E}(Q_b(x)x) - \mathbb{E}(Q_b(x)^2) = \xi_{1,1}.
\]

The proof is complete.

Proof of Theorem 2.2.1

Proof. We have \( y_i = \rho x_i + \sqrt{1 - \rho^2}Z \) in distribution, with \( Z \sim N(0, 1) \) independent of \( x \). Hence,
\[
\mathbb{E}(\hat{\rho}_{b,f}) = \gamma_{1,1} = \mathbb{E}(Q_b(x)(\rho x_i + \sqrt{1 - \rho^2}Z))
\]
\[
= \rho \mathbb{E}(Q_b(x_i)x_i)
\]
\[
= \xi_{1,1}\rho.
\]

Moreover, we have
\[
\gamma_{2,2} = \mathbb{E}(Q_b(x)^2(\rho x_i + \sqrt{1 - \rho^2}Z)^2)
\]
\[
= (\xi_{2,2} - \xi_{2,0})\rho^2 + \xi_{2,0}.
\]
Therefore, the variance can be expressed as

\[ \text{Var}(\hat{\rho}_{b,f}) = \frac{1}{k} \text{Var}(Q_b(x_i)y_i) = \frac{1}{k} (\mathbb{E}(Q_b(x_i)^2y_i^2) - \mathbb{E}(Q_b(x_i)y_i)^2) \]

\[ = \frac{1}{k} (\gamma_{2,2} - \xi_{1,1}^2 \rho^2) \]

\[ = \frac{(\xi_{2,2} - \xi_{2,0} - \xi_{1,1}^2 \rho^2 + \xi_{2,0})}{k} . \]

The variance of debiased estimator follows easily. The proof is complete. \( \square \)

**Proof of Theorem 2.2.2**

**Proof.** Using first order Taylor expansion of \( \frac{x}{y} \) at \( x_0, y_0 \) we get

\[ \frac{x}{y} = \frac{x_0}{y_0} + \frac{x-x_0}{y_0} - \frac{(y-y_0)x_0}{y_0^2} + O\left(\frac{(y-y_0)^2}{y_0^3}\right). \]  

(A.3)

Therefore,

\[ \mathbb{E}(\hat{\rho}_{b,f,n}) = \mathbb{E}\left( \frac{\frac{1}{k} \langle Q_b(x), y \rangle}{\sqrt{\frac{1}{k^2} \|Q_b(x)\|_2^2 \|y\|_2^2}} \right) = \frac{\mathbb{E}(\hat{\rho}_{b,f})}{\mathbb{E}\left( \sqrt{\frac{1}{k^2} \|Q_b(x)\|_2^2 \|y\|_2^2} \right)} + O\left(\frac{1}{k}\right) . \]

Let \( T = \frac{1}{k^2} \|Q_b(x)\|_2^2 \|y\|_2^2 \) and \( \mathbb{E}(T) = E_0 \). Using another Taylor expansion we have:

\[ \mathbb{E}(\sqrt{T}) = \mathbb{E}\left[ \sqrt{E_0} + \frac{T - E_0}{2\sqrt{E_0}} + O((T - E_0)^2) \right] \]

\[ = \sqrt{E_0} + O\left(\frac{1}{k}\right), \text{ as } k \to \infty , \]
and

\[ E_0 = \mathbb{E}(T) = \frac{1}{k^2} \mathbb{E}[\left( \sum_{i=1}^{k} Q_b(x_i)^2 \right) \left( \sum_{i=1}^{k} y_i^2 \right)] \]

\[ = \frac{1}{k^2} \left( \mathbb{E}[\sum_{i \neq j} Q_b(x_i)^2 y_j^2] + \mathbb{E}[\sum_{i=1}^{k} Q(x_i)^2 y_i^2] \right) \]

\[ = \frac{k(k-1)}{k^2} \mathbb{E}(Q(x_1)^2) + \frac{1}{k} \mathbb{E}(Q(x_1)^2 y_1^2) \]

\[ = \frac{k-1}{k} \xi_{2,0} + \frac{\gamma_{2,2}}{k} + O\left(\frac{1}{k}\right), \text{ as } k \to \infty. \]

Put above parts together, we obtain the expected value as \( k \to \infty \),

\[ \mathbb{E}(\hat{\rho}_{b,f,n}) = \frac{\xi_{1,1}\rho}{\sqrt{\xi_{2,0}}} + O\left(\frac{1}{k}\right). \]

To derive the asymptotic variance, let \( a = \frac{\langle Q_b(x), y \rangle}{k}, b = \frac{\|Q_b(x)\|^2}{k}, c = \frac{\|y\|^2}{k}, \) and hence

\[ \hat{\rho}_{b,f,n} = \frac{a}{\sqrt{b \cdot c}}. \]

We have

\[ \mathbb{E}(a) = \xi_{1,1}\rho = \xi_{2,0}\rho = \gamma_{2,0}\rho, \quad Var(a) = \frac{\gamma_{2,2} - \gamma_{2,0}^2\rho^2}{k}, \]

\[ \mathbb{E}(b) = \xi_{2,0} = \gamma_{2,0}, \quad Var(b) = \frac{\gamma_{4,0} - \gamma_{2,0}^2}{k}, \]

\[ \mathbb{E}(c) = 1, \quad Var(c) = \frac{2}{k}, \]

\[ Cov(a,b) = \mathbb{E} \left[ \frac{1}{k^2} \left( \sum_{i=1}^{k} Q_b(x_i) y_i \right) \left( \sum_{i=1}^{k} Q_b(x_i)^2 \right) \right] - \mathbb{E}(a)\mathbb{E}(b), \]

\[ = \frac{1}{k^2} \left[ k(k - 1) \gamma_{2,0} \cdot \gamma_{2,0}\rho + k \gamma_{3,1} \right] - \gamma_{2,0}^2\rho, \]

\[ = \frac{\gamma_{3,1} - \gamma_{2,0}^2\rho}{k}. \]

Similarly, we can get

\[ Cov(a,c) = \frac{\gamma_{1,3} - \gamma_{2,0}^2\rho}{k}, \quad Cov(b,c) = \frac{\gamma_{2,2} - \gamma_{2,0}^2}{k}. \]
Hence the covariance matrix is formulated as

\[
Cov(a, b, c) = \frac{1}{k} \begin{pmatrix}
\gamma_{2,2} - \gamma_{2,0}^2 \rho^2 & \gamma_{3,1} - \gamma_{2,0}^2 \rho & \gamma_{1,3} - \gamma_{2,0} \rho \\
\gamma_{3,1} - \gamma_{2,0} \rho & \gamma_{4,0} - \gamma_{2,0}^2 & \gamma_{2,2} - \gamma_{2,0} \\
\gamma_{1,3} - \gamma_{2,0} \rho & \gamma_{2,2} - \gamma_{2,0} & 2
\end{pmatrix},
\]

and the gradients

\[
\nabla (a, b, c) = \left( \frac{1}{\sqrt{bc}}, -\frac{a}{2b^2 \sqrt{c}}, -\frac{a}{2c^2 \sqrt{b}} \right).
\]

Second order Taylor expansion gives

\[
Var(\hat{\rho}_{b,f,n}) = \nabla (E(a), E(b), E(c))^T Cov(a, b, c) \nabla (E(a), E(b), E(c)) + O\left( \frac{1}{k^2} \right),
\]

and the final result is derived by plugging in the expressions and collecting terms:

\[
Var(\hat{\rho}_{b,f,n}) = \frac{1}{k} \left[ \left( \frac{\gamma_{4,0}}{4\gamma_{2,0}} + \frac{3}{4} \gamma_{2,0} + \frac{1}{2} \gamma_{2,2} \right) \rho^2 - \left( \frac{\gamma_{3,1}}{\gamma_{2,0}} + \gamma_{1,3} \right) \rho + \frac{\gamma_{2,2}}{\gamma_{2,0}} \right] + O\left( \frac{1}{k^2} \right).
\]

This concludes the proof.

\[\square\]

**Proof of Theorem 2.2.3**

**Proof.** We have \( \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \) and \( \Phi = \int_{-\infty}^{x} \phi(x) dx. \) Note that \( \phi'(x) = -x \phi(x). \)

\[
P(X \in [s, t], Y = y) = \int_{s}^{t} \frac{1}{2\pi} e^{-\frac{x^2 - 2\rho xy + y^2}{2(1 - \rho^2)}} dx
\]

\[
= \phi(y) \int_{s}^{t} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1 - \rho^2}} e^{-\frac{(x - \rho y)^2}{2(1 - \rho^2)}} dx
\]

\[
= \phi(y) \int_{s - \rho y}^{\frac{t - \rho y}{\sqrt{1 - \rho^2}}} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1 - \rho^2}} e^{-\frac{u^2}{2}} du
\]

\[
= \phi(y) \left[ \Phi\left( \frac{t - \rho y}{\sqrt{1 - \rho^2}} \right) - \Phi\left( \frac{s - \rho y}{\sqrt{1 - \rho^2}} \right) \right].
\]
Define $S(j) = [t_{j-1}, t_j], j = 1, 2, ..., m$ the $j$-th region separated by $Q$. Here $t_0 = \infty$ and $t_m = \infty$. Let $r_i = \{j : x_i \in S_j\}$. The joint likelihood function is given by:

$$f = \prod_{i=1}^k P(X_i \in S(r_i), Y_i = y_i),$$

and the log-likelihood function is thus

$$L(\rho) = \log(f) = \sum_{i=1}^k \log P(x_i \in S(r_i), y_i = y_i)$$

$$= \sum_{i=1}^k \log(\phi(y_i)) + \sum_{i=1}^k \log[\Phi(\frac{t_{r_i} - \rho y_i}{\sqrt{1 - \rho^2}}) - \Phi(\frac{t_{r_i-1} - \rho y_i}{\sqrt{1 - \rho^2}})]$$

$$= C + \sum_{i=1}^k \delta_i(\rho).$$

We can compute the first derivative as follows. If $r_i \notin \{1, m\}$, we have

$$\frac{\partial \delta_i(\rho)}{\partial \rho} = \frac{(\rho t_{r_i} - y_i)\phi(\frac{t_{r_i} - \rho y_i}{\sqrt{1 - \rho^2}}) - (\rho t_{r_i-1} - y_i)\phi(\frac{t_{r_i-1} - \rho y_i}{\sqrt{1 - \rho^2}})}{(1 - \rho^2)^{3/2}\Phi(\frac{t_{r_i} - \rho y_i}{\sqrt{1 - \rho^2}}) - \Phi(\frac{t_{r_i-1} - \rho y_i}{\sqrt{1 - \rho^2}})} \triangleq \Theta(\rho).$$

If $r_i = 1$,

$$\frac{\partial \delta_i(\rho)}{\partial \rho} = \frac{(\rho t_{r_i} - y_i)\phi(\frac{t_{r_i} - \rho y_i}{\sqrt{1 - \rho^2}})}{(1 - \rho^2)^{3/2} \Phi(\frac{t_{r_i} - \rho y_i}{\sqrt{1 - \rho^2}})} \triangleq \Theta_1(\rho).$$

If $r_i = m$, we have

$$\frac{\partial \delta_i(\rho)}{\partial \rho} = \frac{-(\rho t_{r_i-1} - y_i)\phi(\frac{t_{r_i-1} - \rho y_i}{\sqrt{1 - \rho^2}})}{(1 - \rho^2)^{3/2}[1 - \Phi(\frac{t_{r_i-1} - \rho y_i}{\sqrt{1 - \rho^2}})]} \triangleq \Theta_m(\rho).$$

Hence we have

$$\frac{\partial L(\rho)}{\partial \rho} = \sum_{i=1}^k \mathbb{1}\{r_i \notin \{1, m\}\} \Theta(\rho) + \mathbb{1}\{r_i = 1\} \Theta_1(\rho) + \mathbb{1}\{r_i = m\} \Theta_m(\rho).$$

(A.4)
Therefore, we have

\[ \Theta = \cdots \]

Similarly, we have for \( \Theta \), denote \( \Theta = \frac{A}{B} \).

\[ A' = t_r \left[ \phi \left( \frac{t_r - \rho y_i}{\sqrt{1 - \rho^2}} \right) - (t_r - y_i) \phi \left( \frac{t_r - \rho y_i}{1 - \rho^2} \right) \right] \]

\[ = \left( t_r - \frac{(t_r - y_i)^2(t_r - \rho y_i)}{(1 - \rho^2)^2} \right) \phi \left( \frac{t_r - \rho y_i}{\sqrt{1 - \rho^2}} \right), \]

\[ B' = -3 \rho \sqrt{1 - \rho^2} \phi \left( \frac{t_r - \rho y_i}{\sqrt{1 - \rho^2}} \right) + (t_r - y_i) \phi \left( \frac{t_r - \rho y_i}{\sqrt{1 - \rho^2}} \right). \]

Therefore, we have

\[ \frac{\partial \Theta}{\rho} = \frac{A' B_1 - B_1 A_1}{B_1^2} \]  \hspace{1cm} (A.5)

For \( \Theta_m = \frac{A_m}{B_m} \), we have

\[ A'_m = - \left( t_{r-1} - \frac{(t_{r-1} - y_i)^2(t_{r-1} - \rho y_i)}{(1 - \rho^2)^2} \right) \phi \left( \frac{t_{r-1} - \rho y_i}{\sqrt{1 - \rho^2}} \right), \]

\[ B'_m = -3 \rho \sqrt{1 - \rho^2} \left[ 1 - \Phi \left( \frac{t_{r-1} - \rho y_i}{\sqrt{1 - \rho^2}} \right) \right] - (t_{r-1} - y_i) \phi \left( \frac{t_{r-1} - \rho y_i}{\sqrt{1 - \rho^2}} \right), \]

and

\[ \frac{\partial \Theta_m}{\rho} = \frac{A'_m B_m - B'_m A_m}{B_m^2} \]  \hspace{1cm} (A.6)

Similarly, we have for \( \Theta = \frac{A}{B} \),

\[ A' = \left( t_r - \frac{(t_r - y_i)^2(t_r - \rho y_i)}{(1 - \rho^2)^2} \right) \phi \left( \frac{t_r - \rho y_i}{\sqrt{1 - \rho^2}} \right) - \left( t_{r-1} - \frac{(t_{r-1} - y_i)^2(t_{r-1} - \rho y_i)}{(1 - \rho^2)^2} \right) \phi \left( \frac{t_{r-1} - \rho y_i}{\sqrt{1 - \rho^2}} \right), \]

\[ B' = -3 \rho \sqrt{1 - \rho^2} \phi \left( \frac{t_r - \rho y_i}{\sqrt{1 - \rho^2}} \right) + (t_r - y_i) \phi \left( \frac{t_r - \rho y_i}{\sqrt{1 - \rho^2}} \right) + 3 \rho \sqrt{1 - \rho^2} \phi \left( \frac{t_{r-1} - \rho y_i}{\sqrt{1 - \rho^2}} \right) - (t_{r-1} - y_i) \phi \left( \frac{t_{r-1} - \rho y_i}{\sqrt{1 - \rho^2}} \right), \]
and
\[
\frac{\partial \Theta(\rho)}{\rho} = \frac{A'B - B'A}{B^2}.
\]

Combining (A.5), (A.6) and (A.7), we derive the expression of the Fisher Information
\[
I(\rho) = -\mathbb{E}\left(\frac{\partial^2 L}{\partial \rho^2}\right)
\]
as
\[
\frac{\partial^2 L(\rho)}{\partial \rho^2} = \sum_{i=1}^{k} \mathbb{1}\{r_i \notin \{1, m\}\} \frac{\partial \Theta(\rho)}{\rho} + \mathbb{1}\{r_i = 1\} \frac{\partial \Theta_1(\rho)}{\rho} + \mathbb{1}\{r_i = m\} \frac{\partial \Theta_m(\rho)}{\rho}.
\]

By the consistency and asymptotic normality of MLE, by Cramer-Rao bound gives
\[
\hat{\rho}_{MLE} \sim N(\rho, I^{-1}(\rho)).
\]
The variance of MLE can be computed numerically.

Proof of Theorem 2.2.4

Proof. By normality assumption, we can compute,
\[
P_M(u_1, u_2, u_3) = 1 - \Phi\left(\frac{\alpha_1 - \alpha_3}{\sqrt{\alpha_1^2 + \alpha_3^2 - 2C_1 \alpha_1 \alpha_3}}\right),
\]
\[
P'_M(u_1, u_2, u_3) = 1 - \Phi\left(\frac{\alpha'_1 - \alpha'_3}{\sqrt{\alpha'_1^2 + \alpha'_3^2 - 2C' \alpha'_1 \alpha'_3}}\right).
\]

We can rewrite in terms of debiased variances by
\[
\sigma^2_\rho = \delta^2 \alpha^2 \quad \text{and} \quad \sigma^2'_\rho = \delta'^2 \alpha'^2 \quad \text{for} \quad \forall \rho:
\]
\[
P_M(u_1, u_2, u_3) = 1 - \Phi\left(\frac{\rho_12 - \rho_13}{\sqrt{\delta_{\rho_12}^2 + \delta_{\rho_13}^2 - 2C \delta_{\rho_12} \delta_{\rho_13}}}\right),
\]
\[
P'_M(u_1, u_2, u_3) = 1 - \Phi\left(\frac{\rho_12 - \rho_13}{\sqrt{\delta'_{\rho_12}^2 + \delta'_{\rho_13}^2 - 2C' \delta'_{\rho_12} \delta'_{\rho_13}}}\right)
\]
\[
= 1 - \Phi\left(\frac{\rho_12 - \rho_13}{\sqrt{\alpha^2 \delta_{\rho_12}^2 + \alpha'^2 \delta'_{\rho_13}^2 - 2\alpha \alpha' \delta_{\rho_12} \delta_{\rho_13}}}\right).
\]
with $0 < a < 1$, $0 < a' < 1$ by assumption. To compare the probabilities it suffices to consider the denominators. To make $\hat{P}_A(u_1, u_2, u_3) < \hat{P}_M(u_1, u_2, u_3)$, we need

$$\delta_{p_{12}}^2 + \delta_{p_{13}}^2 - 2C\delta_{p_{12}}\delta_{p_{13}} > a^2\delta_{p_{12}}^2 + a'^2\delta_{p_{13}}^2 - 2aa'C\delta_{p_{12}}\delta_{p_{13}},$$

which after some simplification gives the condition

$$C - aa'C < \frac{(1 - a^2)\delta_{p_{12}}^2 + (1 - a'^2)\delta_{p_{13}}^2}{2\delta_{p_{12}}\delta_{p_{13}}}.$$ 

The proof is complete.\qed

**Proof of Proposition 2.2.5**

**Proof.** First, we notice that minimizing $\frac{\rho^2 E[Q(x)^2 x^2] + (1 - \rho^2) E[Q(x)^2]}{E[Q(x)x^2]}$ is equivalent to

$$\min_Q \rho^2 E[Q(x)^2 x^2] + (1 - \rho^2) E[Q(x)^2], \text{ subject to } E[Q(x)x] = C.$$ 

Here we restrict $C$ to be a positive constant. We can tell easily that $C$ will not affect our solution because the objective is in a quotient form. Now we may express the expectations as integrals and use Lagrangian multipliers safely. We have

$$L = \rho^2 \sum_{i=1}^{M} \int_{t_{i-1}}^{t_i} \mu_i^2 x^2 \phi(x) dx + (1 - \rho^2) \sum_{i=1}^{M} \int_{t_{i-1}}^{t_i} \mu_i^2 \phi(x) dx - \lambda \left( \sum_{i=1}^{M} \int_{t_{i-1}}^{t_i} \mu_i x \phi(x) dx - C \right).$$

(A.10)
Taking the derivative against $\mu_i$, $t_i$ and $\lambda$ respectively, we obtain the KKT condition for optimal solution,

$$\frac{\partial L}{\partial \mu_i} = 2\mu_i \rho^2 \sum_{i=1}^{M} \int_{t_{i-1}}^{t_i} x^2 \phi(x) dx + 2(1 - \rho^2) \mu_i \sum_{i=1}^{M} \int_{t_{i-1}}^{t_i} \phi(x) dx - \lambda \sum_{i=1}^{M} \int_{t_{i-1}}^{t_i} x \phi(x) dx = 0,$$

$$\frac{\partial L}{\partial t_i} = \rho^2 t_i^2 (\mu_i^2 - \mu_{i-1}^2) + (1 - \rho^2)(\mu_i^2 - \mu_{i-1}^2) - \lambda t_i (\mu_i - \mu_{i-1}) = 0,$$

$$\frac{\partial L}{\partial \lambda} = \sum_{i=1}^{M} \int_{t_{i-1}}^{t_i} \mu_i x \phi(x) dx - C = 0.$$

The third equation is non-informative, since $C$ could be chosen arbitrarily. Consequently, by solving the first two equations with caution (careful when $\rho = 0, 1$) we get the stationary point as the desired result.

The fact that $\lambda$ does not affect the debiased variance is because for any $\lambda$, we can get a set of solution, denoted as $S_\lambda$. Hence, elements in $\{S_\lambda : \lambda > 0\}$ are all valid solutions that can achieve the minimum. Another way to justify this claim is to look at the final solutions. If we take $\mu_i$ into the expression of $t_i$, we see that $\lambda$ is either canceled out or serve as a scale parameter, which would eventually be canceled out in the formula of distortion $D_{\rho,\text{opt}}$. Therefore, $\lambda$ is independent of the debiased variance. Actually, we can associate $\lambda$ with different choice of constant $C$, which we do not choose explicitly. \hfill \qed

**Proofs of Scenario 2**

**Hermite polynomials.** First we introduce an important tool for our following analysis.

The probabilists’ Hermite polynomials are defined as

$$H_l(x) = (-1)^l \exp\left(\frac{x^2}{2}\right) \frac{d^l}{dx^l} \exp\left(-\frac{x^2}{2}\right),$$

which form an orthogonal basis of the Hilbert space $\mathcal{H}$ of all functions satisfying

$$\int |f(x)|^2 e^{-x^2} dx < \infty.$$
w.r.t. the $e^{-\frac{x^2}{2}}$ measure. The inner product is well-defined as

$$\langle f, g \rangle = \int f(x)g(x)e^{-\frac{x^2}{2}}dx.$$ 

As an example, the first several Hermite polynomials are

$$H_0(x) = 1, \quad H_1(x) = x, \quad H_2(x) = x^2 - 1, \quad H_3(x) = x^3 - 3x, \ldots,$$

and they can be derived via a recursion relationship: for $l = 0, 1, \ldots$,

$$H_{l+1}(x) = xH_l(x) - H'_l(x).$$

Hermite Polynomials admits **Orthogonality** in the sense that

$$\int H_m(x)H_n(x)e^{-\frac{x^2}{2}}dx = 0, \quad m \neq n,$$

$$\int H_n(x)H_n(x)e^{-\frac{x^2}{2}}dx = \sqrt{2\pi}n!, \quad m = n.$$

We can deduct some useful quantities from this property. Let $x \sim N(0, 1)$, then we have for all $l = 1, 2, \ldots$,

$$\mathbb{E}(H_l(x)) = \mathbb{E}(H_0(x)H_l(x)) = 0, \quad \text{Var}(H_l(x)) = \frac{1}{\sqrt{2\pi}} \int H_l(x)H_l(x)e^{-\frac{x^2}{2}}dx = l!.$$ 

Moreover, $H_n(x)$ is an odd function if $n$ is odd, and is symmetric about $y$ axis when $n$ is even. One important application of Hermite polynomials is that we can decompose the bivariate normal density as below Anderson (2003):

$$\phi_{\rho}(x, y) = \sum_{l=0}^{\infty} \frac{\rho^l}{l!} H_l(x)H_l(y)\phi(x)\phi(y),$$
where $H_l(x)$ is the $l$-th order probabilists’ Hermite polynomial, and $\phi(x)$ is the density function of standard normal distribution as defined before. This immediately implies that for any functions $f_1$ and $f_2$, we can write

$$E[f_1(x)f_2(y)] = \int \int f_1(x)f_2(y)\phi(x,y)dxdy$$

$$= \int \int f_1(x)f_2(y)\sum_{l=0}^{\infty} \rho^l l! H_l(x)H_l(y)\phi(x)\phi(y)dxdy$$

$$= \sum_{l=0}^{\infty} \rho^l l! \int \int f_1(x)f_2(y)H_l(x)H_l(y)\phi(x)\phi(y)dxdy$$

$$= \sum_{l=0}^{\infty} \rho^l l! \left( \int f_1(x)H_l(x)\phi(x)dx \int f_2(y)H_l(y)\phi(y)dy \right).$$

(A.11)

As we can see, the correlation coefficient $\rho$ is factored out in (A.11), which is beneficial for studying the dependence of the expected value on $\rho$.

Proof of Theorem 2.3.1 & Theorem 2.3.2

To prove the results, we will use the following lemma.

**Lemma A.2.** Suppose we have a sequence of positive constants $V = (v_1, v_2, \ldots)$. Let $W = \text{diag}(V)$ and $c_1 = (c_{11}, c_{12}, \ldots)$ and $c_2 = (c_{21}, c_{22}, \ldots)$ be vectors with same length as $V$. Then

$$\max_{\|c_1\|^2_2=L_1,\|c_2\|^2_2=L_2} c_1^T W c_2 = \sqrt{L_1L_2}\|V\|_{\infty},$$

where the infinite norm $\|\cdot\|_{\infty}$ is the maximum absolute value of a vector.

**Proof.** By the symmetry of this optimization problem, we know that the optimal solution of $c_1$ and $c_2$ is not unique. Hence, we may cast two more constraints $c_1 \geq 0$ and $c_2 \geq 0$ to get a unique solution. To proceed, we introduce Lagrangian multipliers $L$ with slack variables $\tilde{s} = (s_1, s_2, \ldots), \tilde{t} = (t_1, t_2, \ldots)$ as:

$$L = c_1^T W c_2 - \lambda_1 (c_1^T c_1 - L_1) - \lambda_2 (c_2^T c_2 - L_2) + \tilde{\lambda}_3 (c_1 - \tilde{s}^2) - \tilde{\lambda}_4 (c_2 + \tilde{t}^2),$$
where $\tilde{\lambda}_3 = (\lambda_{31}, \lambda_{32}, ...)$ and $\tilde{\lambda}_4 = (\lambda_{41}, \lambda_{42}, ...)$. The Karush-Kuhn-Tucker conditions are satisfied at minimal point, which gives

\[
\begin{align*}
Wc_2 - 2\lambda_1 c_1 + \tilde{\lambda}_3 &= 0 \quad \text{(A.12)} \\
Wc_1 - 2\lambda_2 c_2 - \tilde{\lambda}_4 &= 0 \quad \text{(A.13)} \\
c_1^T c_1 &= L_1 \\
c_2^T c_2 &= L_2 \\
c_1 - \bar{s}^2 &= 0 \\
c_2 + \bar{t}^2 &= 0 \\
2\tilde{\lambda}_3 \odot \bar{s} &= 0 \\
2\tilde{\lambda}_4 \odot \bar{t} &= 0
\end{align*}
\]

where $\odot$ denotes element-wise product. The equations leads to following observations:

- Any pair of values $(c_{1i}, c_{2i})$ must be zero or nonzero at the same time. To see this, suppose $c_{1i} = 0$ and $c_{2i} \neq 0$, then by (A.13) we have two situations:
  1) $\lambda_2 \neq 0$ and $\lambda_{4i} \neq 0$, which implies that $t_i = 0$ and thus $c_{2i} = 0$. A contradiction occurs.
  2) $\lambda_2 = 0$ and $\lambda_{4i} = 0$. Firstly, we note that there must exist at least one $j \neq i$ such that $c_{1j} \neq 0$. For a nonzero $c_{1j}$, $\lambda_2 = 0$ forces $\lambda_{4j} \neq 0$, and thus $c_{2j} = 0$ must be zero. Therefore, for $\forall i = 1, 2, ..., \lambda_{3i} = \lambda_{4i} = 0$ for $\forall i$. From (A.12) and (A.13) we deduce that $c_{1i} = \frac{\lambda_2 c_{2i}}{V_i} = \frac{V_i c_{2i}}{\lambda_1}$, from which we can further derive $V_i^2 = \lambda_1 \lambda_2$.

Based on above reasoning, we can consider 2 situations for the optimal solution. First, if
only one pair \((c_{1i}, c_{2i})\) is nonzero, then the maximum of \(c_1^T W c_2\) is trivially derived at

\[
c_1 = \sqrt{L_1 I_{\max}}, \quad c_2 = \sqrt{L_2 I_{\max}},
\]

with \(I_{\max}\) the indicator vector of where the maximum of \(V\) is located, e.g. in the form \((...,0,0,1,0,...)\). The maxima in this case equals to

\[
\max_{c_1,c_2} c_1^T W c_2 = \sqrt{L_1 L_2} \max V = \sqrt{L_1 L_2} \|V\|_\infty,
\]

subject to constraints \(\|c_1\|_2^2 = L_1, \|c_2\|_2^2 = L_2\).

Now consider the case where more than two pairs of values \((c_{1i}, c_{2i}), i \in S\) are nonzero, where \(S\) denotes the set of nonzero indices. Then \(\lambda_1 \lambda_2 = V_i^2 := V_i^* \forall i \in S\) must hold. By Cauchy-Schwartz inequality, we have

\[
c_1^T W c_2 = V^* c_1^T c_2 \leq V^* \|c_1\|_2 \|c_2\|_2 \leq \sqrt{L_1 L_2} V^* \leq \sqrt{L_1 L_2} \|V\|_\infty,
\]

and the bound is tight (i.e. equality holds when \(c_1\) and \(c_2\) have same direction).

Combining above analysis, we have shown that

\[
\max_{\|c_1\|_2^2 = L_1, \|c_2\|_2^2 = L_2} c_1^T W c_2 = \sqrt{L_1 L_2} \|V\|_\infty.
\]
Proof. (of Theorem 2.3.1 & Theorem 2.3.2) First, we have that

\[
\mathbb{E}(Q_{b_1}(x)Q_{b_2}(y)) = \sum_{l=0}^{\infty} \frac{\rho^l}{l!} \int Q_{b_1}(x)H_l(x)\phi(x)dx \int Q_{b_2}(y)H_l(y)\phi(y)dy
\]

\[
= \sum_{l=1,\text{odd}}^{\infty} \frac{\rho^l}{l!} \mathbb{E}[Q_{b_1}(x)H_l(x)]\mathbb{E}[Q_{b_2}(x)H_l(x)]
\]

\[
= (1 - D_{b_1} - D_{b_2} + D_{b_1}D_{b_2})\rho + \sum_{l=3,\text{odd}}^{\infty} \frac{\rho^l}{l!} \text{Cov}[Q_{b_1}(x), H_l(x)] \cdot \text{Cov}[Q_{b_2}(x), H_l(x)].
\]

Note that \( E_{-\rho}[Q_{b_1}(x)Q_{b_2}(y)] = -E_{\rho}[Q_{b_1}(x)Q_{b_2}(y)] \), so it suffices to consider the case where \( \rho \geq 0 \) in the remaining part of the proof.

From previous sections we know that for a fixed quantizer \( Q_{b}(\cdot) \) with distortion \( D_b \) and Hermite Polynomial \( H_k(\cdot) \) with \( k > 1 \),

\[
\text{Var}(H_k(x)) = \mathbb{E}(H_k(x)^2) = k!, \quad \text{Cov}(Q_{b}(x), x) = \mathbb{E}(Q_{b}(x)x) = 1 - D_b,
\]

\[
\text{Var}(Q_{b}(x)) = \mathbb{E}(Q_{b}(x)^2) = 1 - D_b, \quad \text{Cov}(H_k(x), x) = \mathbb{E}(H_k(x)x) = 0.
\]

We can compute the correlations:

\[
\text{Corr}(Q_{b}(x), x) = \sqrt{1 - D_b}, \quad \text{Corr}(H_k(x), x) = 0.
\]

By working with correlations between 3 random variables and using Cauchy-Schwartz inequality, we get

\[-\sqrt{D_b} \leq \text{Corr}(Q_{b}(x), H_k(x)) \leq \sqrt{D_b}.\]

Denote the correlation \( \text{Corr}(Q_{b}(x), H_k(x)) \) as \( c_k, k = 0, 1, 2, \ldots \), and \( \mathbf{C}=(c_0, c_1, c_2, \ldots) \).

Note that Hermite polynomials are infinite orthogonal basis of the function space \( \mathcal{H} \), and
thus we have the decomposition $Q_b(x) = \sum_{i=1}^{\infty} a_i H_i(x)$ for some $a_i, i = 1, 2, \ldots$. Simple calculation yields $Cov(Q, H_i) = a_i Var(H_i(x))$, $Var(Q) = \sum_{i=1}^{\infty} a_i^2 Var(H_i(x))$. So the correlations can be derived as

$$c_i = Corr(Q, H_i) = \frac{a_i Var(H_i(x))}{\sqrt{\sum_{j=1}^{\infty} a_j^2 Var(H_j(x))} \sqrt{Var(H_i(x))}}. $$

Consequently, we have $C^T C \equiv 1$. Given that $c_1 = Corr(Q_b(x), x) = \sqrt{1 - D_b}$ and $c_k = 0$ for all even $k$’s, we have $\sum_{k=3, odd}^{\infty} c_k^2 = D_b$.

The above argument holds for both $Q_{b_1}$ and $Q_{b_2}$. Denote $c_{1k} = Corr(Q_{b_1}, H_k)$ and $c_{2k} = Corr(Q_{b_2}, H_k)$ and notice that for $i = 1, 2$ and $k = 0, 1, 2, \ldots$,

$$Cov(Q_{b_i}(x), H_k(x)) = c_{ik} \sqrt{1 - D_i} \sqrt{k!},$$

because $Var[H_k(x)] = k!$. Continuing with (A.15) we obtain

$$E(\hat{\rho}_{b_1,b_2}) = E(Q_{b_1}(x)Q_{b_2}(y)) = (1-D_{b_1})(1-D_{b_2})\rho + \sqrt{1-D_{b_1}} \sqrt{1-D_{b_2}} \sum_{k=3, odd}^{\infty} c_{1k} c_{2k} \rho^k. $$

(A.16)

Now we seek to bound the last term in above equation. Applying Lemma A.2 with $V(\rho) = (\rho^3, \rho^5, \rho^7, \ldots)$ and constraints $\|c_1\|^2_2 = D_{b_1}, \|c_2\|^2_2 = D_{b_2}$, we get

$$-\sqrt{D_{b_1} D_{b_2}} |\rho|^3 \leq \sum_{k=3, odd}^{\infty} c_{1k} c_{2k} \rho^k \leq \sqrt{D_{b_1} D_{b_2}} |\rho|^3,$$

for $\rho \in [-1, 1]$ by symmetry, and this bound is tight in worst-case. Therefore, we have

$$|E(\hat{\rho}_{b_1,b_2}) - (1 - D_{b_1})(1 - D_{b_2})\rho| \leq \sqrt{D_{b_1} D_{b_2}} \sqrt{1-D_{b_1}} \sqrt{1-D_{b_2}} |\rho|^3. $$

(A.17)
To get the bound on absolute bias, note that for $\rho > 0$, by Eq. (A.17) we have

$$
\rho - \mathbb{E}(\hat{\rho}_{b_1,b_2}) \geq (D_{b_1} + D_{b_2} - D_{b_1}D_{b_2})\rho - \sqrt{D_{b_1}D_{b_2}}\sqrt{1 - D_{b_1}}(1 - D_{b_2})\rho^3. \quad (A.18)
$$

By computing

$$
(D_{b_1} + D_{b_2} - D_{b_1}D_{b_2})^2 - D_{b_1}D_{b_2}(1 - D_{b_1})(1 - D_{b_2})
= D_{b_1}^2 + D_{b_2}^2 + D_{b_1}D_{b_2}(1 - D_{b_1} - D_{b_2})
\geq 0,
$$

since for LM quantizers, $1 - D_{b_1} - D_{b_2} > 0$ always holds. Consequently, we know that $\rho - \mathbb{E}(\hat{\rho}_{b_1,b_2}) > 0$ for $\rho > 0$. Now by the symmetry of $\hat{\rho}_{b_1,b_2}$ and elementary inequalities, we have for $\rho \in [-1, 1]$,

$$
\Delta_2 - \Delta_1 \leq |E(\hat{\rho}_{b_1,b_2}) - \rho| \leq \Delta_1 + \Delta_2,
$$

where

$$
\Delta_1 = \sqrt{D_{b_1}D_{b_2}}\sqrt{1 - D_{b_1}}\sqrt{1 - D_{b_2}}|\rho|^3, \quad \Delta_2 = (D_{b_1} + D_{b_2} - D_{b_1}D_{b_2})|\rho|.
$$

To prove Theorem 2.3.2, notice that when $Q_1 = Q_2 := Q_b$ and $D_{b_1} = D_{b_2} := D_b$, we can modify (A.16) as

$$
\mathbb{E}(Q_b(x)Q_b(y)) = (1 - 2D_b + D_b^2)\rho + (1 - D_b)\sum_{l=3,\text{odd}}^{\infty} c_k\rho^k,
$$

where $c_k = Corr(Q_b(x), H_k(x))$ and $\sum_{k=3}^{\infty} c_k = D_b$. Obviously, the summation is lower bounded by 0 and upper bounded by $D_b\rho^3$. Similar calculation can be conducted to get the bound on absolute bias. This completes the proof.
Proof of Theorem 2.3.3

Proof. We recall some notations. The data vectors are LM quantized with different bits $b_1 < b_2$, and we denote two Lloyd-Max quantizers as $Q_{b_1}$ and $Q_{b_2}$ and distortion $D_{b_1}$ and $D_{b_2}$, respectively. With a little abuse of notation, in this section we re-define $\xi_{\alpha,\beta} = \mathbb{E}(Q_{b_1}(x)^\alpha x^\beta)$, $\gamma_{\alpha,\beta} = \mathbb{E}(Q_{b_2}(x)^\alpha x^\beta)$ and $\zeta_{\alpha,\beta} = \mathbb{E}(Q_{b_1}(x)^\alpha Q_{b_2}(y)^\beta)$. As $k \to \infty$, Taylor Expansion of $\frac{x}{y}$ at $x_0, y_0$ gives:

$$\frac{x}{y} = \frac{x_0}{y_0} + \frac{x - x_0}{y_0} \cdot \frac{y - y_0}{y_0^2} + O((x - x_0)^2) + O((y - y_0)^2).$$

We apply the expansion at expectations:

$$\mathbb{E}(\hat{\rho}_{b_1,b_2,n}) = \frac{\mathbb{E}(\frac{1}{k} \langle Q_{b_1}(x), Q_{b_2}(y) \rangle)}{\mathbb{E} \left( \sqrt{\frac{1}{k^2} \|Q_{b_1}(x)\|_2^2 \|Q_{b_2}(y)\|_2^2} \right)} + O\left(\frac{1}{k}\right) = \frac{\mathbb{E}(\hat{\rho}_{b_1,b_2})}{\mathbb{E}(\sqrt{T})} + O\left(\frac{1}{k}\right).$$

Let $T = \frac{1}{k^2} \|Q_{b_1}(x)\|_2^2 \|Q_{b_2}(y)\|_2^2$, $\mathbb{E}(T) = E_0$. Using Taylor Expansion again, we have:

$$\mathbb{E}(\sqrt{T}) = \mathbb{E} \left( \sqrt{E_0 + \frac{T - E_0}{2\sqrt{E_0}}} + O((T - E_0)^2) \right)$$

$$= \sqrt{E_0} + O\left(\frac{1}{k}\right), \text{ as } k \to \infty,$$

$$E_0 = \mathbb{E}(T) = \frac{1}{k^2} \mathbb{E} \left( \sum_{i=1}^{k} Q_{b_1}(x_i)^2 \sum_{i=1}^{k} Q_{b_2}(y_i)^2 \right)$$

$$= \frac{1}{k^2} \left[ \mathbb{E}(\sum_{i \neq j} Q_{b_1}(x_i)^2 Q_{b_2}(y_j)^2) + \mathbb{E}(\sum_{l=1}^{k} Q_{1}(x_l)^2 Q_{2}(y_l)^2) \right]$$

$$= \frac{k(k-1)}{k^2} \mathbb{E}[Q_1(x_1)^2] \mathbb{E}[Q_2(y_1)^2] + \frac{1}{k} \mathbb{E}[Q_1(x_1)^2 Q_2(y_1)^2]$$

$$= \frac{k - 1}{k} \xi_{2,0} \gamma_{2,0} + \frac{1}{k} \zeta_{2,2} + O\left(\frac{1}{k}\right), \text{ as } k \to \infty.$$
Combining parts together we have as $k \to \infty$,

$$
\mathbb{E}(\hat{\rho}_{b_1, b_2, n}) = \frac{\zeta_{1,1}}{\sqrt{\xi_{2,0} \gamma_{2,0}}} + O\left(\frac{1}{k}\right).
$$

Let $a = \frac{<Q_{b_1}(x), Q_{b_2}(y)>}{k}$, $b = \frac{||Q_{b_1}(x)||^2}{k}$, $c = \frac{||Q_{b_2}(y)||^2}{k}$, and thus $\hat{\rho}_{b_1, b_2, n} = \frac{a}{\sqrt{b \sqrt{c}}}$. We have:

$$
\mathbb{E}(a) = \zeta_{1,1}, \quad \text{Var}(a) = \frac{\zeta_{2,2} - \zeta_{1,1}^2}{k},
$$

$$
\mathbb{E}(b) = \xi_{2,0}, \quad \text{Var}(b) = \frac{\xi_{4,0} - \xi_{2,0}^2}{k},
$$

$$
\mathbb{E}(c) = \gamma_{2,0}, \quad \text{Var}(c) = \frac{\gamma_{4,0} - \gamma_{2,0}^2}{k},
$$

$$
\text{Cov}(a, b) = \mathbb{E}\left(\frac{1}{k^2} \left( \sum_{1}^{k} Q_{b_1}(x_i)Q_{b_2}(y_i) \right) \sum_{1}^{k} Q_{b_1}(x_i)^2 \right) - \mathbb{E}(a)\mathbb{E}(b)
$$

$$
= \frac{1}{k^2} [k(k-1)\zeta_{1,1}\xi_{2,0} + k\zeta_{3,1}] - \zeta_{1,1}\xi_{2,0}
$$

$$
= \frac{\zeta_{3,1} - \zeta_{1,1}\xi_{2,0}}{k},
$$

$$
\text{Cov}(a, c) = \frac{\zeta_{1,1} - \zeta_{1,1}\gamma_{2,0}}{k}, \quad \text{Cov}(b, c) = \frac{\zeta_{2,2} - \xi_{2,0}\gamma_{2,0}}{k}.
$$

Therefore,

$$
\text{Cov}(a, b, c) = \frac{1}{k} \begin{pmatrix}
\zeta_{2,2} - \zeta_{1,1}^2 & \zeta_{3,1} - \zeta_{1,1}\xi_{2,0} & \zeta_{1,3} - \zeta_{1,1}\gamma_{2,0} \\
\zeta_{3,1} - \zeta_{1,1}\xi_{2,0} & \xi_{4,0} - \xi_{2,0}^2 & \zeta_{2,2} - \xi_{2,0}\gamma_{2,0} \\
\zeta_{1,3} - \zeta_{1,1}\gamma_{2,0} & \zeta_{2,2} - \xi_{2,0}\gamma_{2,0} & \gamma_{4,0} - \gamma_{2,0}^2
\end{pmatrix},
$$

and

$$
\nabla (\mathbb{E}(a), \mathbb{E}(b), \mathbb{E}(c)) = \left(\frac{1}{\sqrt{\xi_{2,0}\gamma_{2,0}}}, -\frac{\zeta_{1,1}}{2\xi_{2,0}^2\sqrt{\gamma_{2,0}}}, -\frac{\zeta_{1,1}}{2\gamma_{2,0}^2\sqrt{\xi_{2,0}}} \right).
$$

Using Taylor expansion we have

$$
\text{Var}(\hat{\rho}_{b_1, b_2, n}) = \nabla (\mathbb{E}(a), \mathbb{E}(b), \mathbb{E}(c))^T \text{Cov}(a, b, c) \nabla (\mathbb{E}(a), \mathbb{E}(b), \mathbb{E}(c)) + O\left(\frac{1}{k^2}\right).
$$
The final result is derived by direct calculation and collecting terms:

\[
\text{Var}(\hat{\rho}_{b_1, b_2, n}) = \frac{1}{k} \left[ \frac{\zeta_{2,2} - \zeta_{1,1}^2}{\xi_{0,\gamma_{2,0}}} - \frac{\zeta_{1,1} \zeta_{3,1} - \zeta_{1,1}^2 \xi_{2,0}}{\xi_{2,0,\gamma_{2,0}}} - \frac{\zeta_{1,1} \zeta_{1,3} - \zeta_{1,1}^2 \gamma_{2,0}}{\xi_{2,0,\gamma_{2,0}}} \right] + \frac{\zeta_{1,1}^2 \zeta_{2,2} - \zeta_{1,1}^2 \xi_{2,0,\gamma_{2,0}}}{2\xi_{2,0,\gamma_{2,0}}^2} + \frac{\zeta_{1,1} \xi_{4,0} - \zeta_{1,1}^2 \xi_{2,0}}{4\xi_{2,0,\gamma_{2,0}}} + \frac{\zeta_{1,1} \gamma_{4,0} - \zeta_{1,1}^2 \gamma_{2,0}^2}{4\xi_{2,0,\gamma_{2,0}}^2} + O\left(\frac{1}{k^2}\right).
\]

This completes the proof.

\section*{Proofs of Monotonicity}

\textbf{Lemma A.1.} Assume \( \begin{pmatrix} x \\ y \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \). For \( 0 \leq s < t \) and \(-1 \leq \rho \leq 1\), \( \Pr(x \in [s, t], y \geq 0) \) is increasing in \( \rho \), \( \Pr(x \in [s, t], y < 0) \) is decreasing in \( \rho \).

\textbf{Proof.} We have

\[
P_{s,t,+} = \Pr(x \in [s, t], y \geq 0) = \int_0^\infty \int_s^t \frac{1}{2\pi \sqrt{1-\rho^2}} e^{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}} \, dx \, dy
\]

\[
= \int_0^\infty \frac{1}{2\pi \sqrt{1-\rho^2}} e^{-\frac{x^2}{2}} \int_s^t e^{-\frac{(y-\rho x)^2}{2(1-\rho^2)}} \, dy \, dx
\]

\[
= \int_0^\infty \frac{1}{2\pi \sqrt{1-\rho^2}} e^{-\frac{x^2}{2}} \int_s^t e^{-\frac{y^2}{2(1-\rho^2)}} \sqrt{1-\rho^2} \, e^{\frac{t-\rho x}{\sqrt{1-\rho^2}}} \, du \, dx
\]

\[
= \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left[ \Phi\left( \frac{t-\rho x}{\sqrt{1-\rho^2}} \right) - \Phi\left( \frac{s-\rho x}{\sqrt{1-\rho^2}} \right) \right] \, dx.
\]

It is easy to check that this integral meets the conditions of DCT. Hence, taking the derivative yields

\[
\frac{\partial P_{s,t,+}}{\partial \rho} := \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left[ \phi\left( \frac{t-\rho x}{\sqrt{1-\rho^2}} \right) - x + t \rho \right] - \phi\left( \frac{s-\rho x}{\sqrt{1-\rho^2}} \right) \left( \frac{3}{2} - \frac{x + s \rho}{(1-\rho^2)^{3/2}} \right) dx.
\]
For the first term we have

\[
\int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \phi\left( \frac{t - \rho x}{\sqrt{1 - \rho^2}} \right) - x + t \rho \, dx = \int_0^{\infty} \frac{1}{2\pi} e^{-\frac{(x-t\rho)^2}{2(1-\rho^2)}} e^{-\frac{t^2}{2(1-\rho^2)}} \left( -x + t \rho \right) \, dx
\]

\[
= \frac{1}{2\pi \sqrt{1 - \rho^2}} e^{-\frac{t^2}{2(1-\rho^2)}} \left|_{0}^{\infty} \right.
\]

\[
= -\frac{1}{2\pi \sqrt{1 - \rho^2}} e^{-\frac{t^2}{2(1-\rho^2)}}. \tag{A.19}
\]

Similarly we can compute

\[
\int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \phi\left( \frac{s - \rho x}{\sqrt{1 - \rho^2}} \right) - x + s \rho \, dx = -\frac{1}{2\pi \sqrt{1 - \rho^2}} e^{-\frac{s^2}{2(1-\rho^2)}}. \tag{A.20}
\]

Thus, we obtain

\[
\frac{\partial P_{s,t,+}}{\partial \rho} = \frac{1}{2\pi \sqrt{1 - \rho^2}} e^{-\frac{s^2}{2(1-\rho^2)}} - \frac{1}{2\pi \sqrt{1 - \rho^2}} e^{-\frac{t^2}{2(1-\rho^2)}}
\]

\[
= \frac{1}{2\pi \sqrt{1 - \rho^2}} \left( e^{-\frac{s^2}{2(1-\rho^2)}} - e^{-\frac{t^2}{2(1-\rho^2)}} \right)
\]

\[
> 0,
\]

due to the fact that \( s < t \). For \( P_{s,t,-} := Pr( x \in [s, t], y < 0 ) \), we proceed with similar calculation, which will change the sign in (A.19) and (A.20) and eventually gives

\[
\frac{\partial P_{s,t,-}}{\partial \rho} = \frac{1}{2\pi \sqrt{1 - \rho^2}} \left( e^{-\frac{t^2}{2(1-\rho^2)}} - e^{-\frac{s^2}{2(1-\rho^2)}} \right) < 0.
\]

The proof is complete. \( \square \)

Proof of Lemma 2.3.4

We prove a more detailed version of Lemma 2.3.4.

**Lemma A.3** (Detailed Lemma 2.3.4). Assume \( Q_{b_1} \) is a \( M \)-bit symmetric quantizer in the
sense that it divides the positive axis into $M$ intervals with cut point $t_0 = 0 < t_1 < \cdots < t_{M-1}$. The reconstruction levels are given by $Q_{b_1}(x) = \mu_i > 0$, $x \in [t_{i-1}, t_i]$ and $Q_{b_1}(x) = -\mu_i$, $x \in [-t_i, t_{i-1}]$, $i = 1, \ldots, M$. $Q_{b_2}$ is a 1-bit quantizer such that $Q_{b_2}(y) = \nu > 0$ when $y \geq 0$ and $Q_{b_2}(y) = -\nu$ when $y < 0$. Then $E[Q_{b_1}(x)Q_{b_2}(y)]$ is strictly increasing in $\rho$ on $[-1, 1]$.

**Proof.** Denote $P_{s,t,+} = Pr(x \in [s, t], y \geq 0)$ and $P_{s,t,-} = Pr(x \in [s, t], y < 0)$. We write explicitly

$$E[Q_{b_1}(x)Q_{b_2}(y)] = \nu \sum_{i=1}^{M} \mu_i Pr_{t_{i-1},t_i,+} - \nu \sum_{i=1}^{M} \mu_i Pr_{t_{i-1},t_i,-}$$

$$- \nu \sum_{i=1}^{M} \mu_i Pr_{-t_i,-t_{i-1},+} + \nu \sum_{i=1}^{M} \mu_i Pr_{-t_i,-t_{i-1},-}$$

$$= 2\nu \sum_{i=1}^{M} \mu_i (Pr_{t_{i-1},t_i,+} - Pr_{t_{i-1},t_i,-}),$$

due to the symmetry of bivariate normal density. Since $\nu > 0$ and $\mu_i > 0$, $i = 1, \ldots, M$, applying Lemma A.1 we prove the desired result. \qed

**Proof of Lemma 2.3.5**

In the following we prove a detailed version of Lemma 2.3.5.

**Lemma A.4** (Detailed Lemma 2.3.5). Consider two 2-bit symmetric quantizers $Q_{b_1}$ and $Q_{b_2}$. $Q_{b_1}$ has cut point at $(-t_1, 0, t_1)$ with distinct quantizing values $(-\mu_2, -\mu_1, \mu_1, \mu_2)$, $0 < \mu_1 < \mu_2$ on the 4 intervals separated by the cut points. Similarly, $Q_{b_2}$ has cut points $(-t_2, 0, t_2)$ and distinct codes $(-\xi_2, -\xi_1, \xi_1, \xi_2)$, $0 < \xi_1 < \xi_2$. Assume that both quantizers to be increasing, namely, $\mu_1 < \mu_2$, $\xi_1 < \xi_2$. Then $E[Q_{b_1}(x)Q_{b_2}(y)]$ is strictly increasing in $\rho$ on $[-1, 1]$. 


Proof. The expectations is computed as

$$\mathbb{E}[Q_{b_1}(x)Q_{b_2}(y)] = 2\mu_1\xi_1(P_{11} - p_{11}) + 2\mu_1\xi_2(P_{12} - p_{12}) + 2\mu_2\xi_1(P_{21} - p_{21}) + 2\mu_2\xi_2(P_{22} - p_{22}),$$

(A.21)

where

$$P_{11} = Pr(x \in [0, t_1], y \in [0, t_2]), P_{12} = Pr(x \in [0, t_1], y \in [t_2, +\infty]),$$
$$P_{21} = Pr(x \in [t_1, +\infty], y \in [0, t_2]), P_{22} = Pr(x \in [t_1, +\infty], y \in [t_2, +\infty]),$$
$$p_{11} = Pr(x \in [0, t_1], y \in [-t_2, 0]), p_{12} = Pr(x \in [0, t_1], y \in [t_2, 0]),$$
$$p_{21} = Pr(x \in [t_1, +\infty], y \in [-t_2, 0]), p_{22} = Pr(x \in [t_1, +\infty], y \in [-t_2, 0]).$$

We compute the derivative with respect to \( \rho \) for each probability using the procedure in proving lemma.

$$\frac{\partial P_{11}}{\partial \rho} = \frac{1}{2\pi \sqrt{1 - \rho^2}} \left[ e^{-\frac{t_1^2 + t_2^2 - 2\rho t_1 t_2}{2(1 - \rho^2)}} - e^{-\frac{t_1^2}{2(1 - \rho^2)}} - e^{-\frac{t_2^2}{2(1 - \rho^2)}} + 1 \right]$$
$$\frac{\partial P_{12}}{\partial \rho} = \frac{1}{2\pi \sqrt{1 - \rho^2}} \left[ -e^{-\frac{t_1^2 + t_2^2 - 2\rho t_1 t_2}{2(1 - \rho^2)}} + e^{-\frac{t_1^2}{2(1 - \rho^2)}} \right]$$
$$\frac{\partial P_{21}}{\partial \rho} = \frac{1}{2\pi \sqrt{1 - \rho^2}} \left[ -e^{-\frac{t_1^2 + t_2^2 - 2\rho t_1 t_2}{2(1 - \rho^2)}} + e^{-\frac{t_1^2}{2(1 - \rho^2)}} \right]$$
$$\frac{\partial P_{22}}{\partial \rho} = \frac{1}{2\pi \sqrt{1 - \rho^2}} \left[ e^{-\frac{t_1^2 + t_2^2 - 2\rho t_1 t_2}{2(1 - \rho^2)}} \right]$$
$$\frac{\partial p_{11}}{\partial \rho} = \frac{1}{2\pi \sqrt{1 - \rho^2}} \left[ -e^{-\frac{t_1^2 + t_2^2 + 2\rho t_1 t_2}{2(1 - \rho^2)}} + e^{-\frac{t_1^2}{2(1 - \rho^2)}} \right]$$
$$\frac{\partial p_{12}}{\partial \rho} = \frac{1}{2\pi \sqrt{1 - \rho^2}} \left[ e^{-\frac{t_1^2 + t_2^2 + 2\rho t_1 t_2}{2(1 - \rho^2)}} - e^{-\frac{t_1^2}{2(1 - \rho^2)}} \right]$$
$$\frac{\partial p_{21}}{\partial \rho} = \frac{1}{2\pi \sqrt{1 - \rho^2}} \left[ e^{-\frac{t_1^2 + t_2^2 + 2\rho t_1 t_2}{2(1 - \rho^2)}} - e^{-\frac{t_2^2}{2(1 - \rho^2)}} \right]$$
$$\frac{\partial p_{22}}{\partial \rho} = \frac{1}{2\pi \sqrt{1 - \rho^2}} \left[ -e^{-\frac{t_1^2 + t_2^2 + 2\rho t_1 t_2}{2(1 - \rho^2)}} \right].$$
Now, taking the derivative of (A.21) and collecting terms yields

\[
\frac{\partial \mathbb{E}[Q_{b_1}(x)Q_{b_2}(y)]}{\partial \rho} = \frac{1}{\pi} \frac{1}{\sqrt{1 - \rho^2}} \left[ \mu_1 \xi_1 (A + 2 - 2C_1 - 2C_2) + \mu_1 \xi_2 (2C_2 - A) + \mu_2 \xi_1 (2C_1 - A) + \mu_2 \xi_2 A \right],
\]

where \( A = e^{-t_1^2 + t_2^2 - 2t_1 t_2 - \frac{2}{2(1 - \rho^2)}} e^{-t_1^2 + t_2^2 - 2t_1 t_2} \), \( C_1 = e^{-t_1^2 - \frac{2}{2(1 - \rho^2)}} \), \( C_2 = e^{-t_1^2} \). Rearranging terms, we obtain

\[
\frac{\partial \mathbb{E}[Q_{b_1}(x)Q_{b_2}(y)]}{\partial \rho} \propto A (\mu_1 \xi_1 - \mu_1 \xi_2 - \mu_2 \xi_1 + \mu_2 \xi_2) + (2 - 2C_1 - 2C_2) \mu_1 \xi_1 + 2C_2 \mu_1 \xi_2 + 2C_1 \mu_2 \xi_1
\]

\[
= A (\mu_1 \xi_1 - \mu_1 \xi_2 - \mu_2 \xi_1 + \mu_2 \xi_2) + 2 \mu_1 \xi_1 + 2C_1 \xi_1 (\mu_2 - \mu_1) + 2C_2 \mu_1 (\xi_2 - \xi_1)
\]

\[
> 0.
\]

The last inequality holds due to \( 0 < \mu_1 < \mu_2, 0 < \xi_1 < \xi_2 \).

Lemma A.4 requires that both quantizers be “stair-shaped” (i.e., increasing) functions. Next, we extend the analysis to the general case based on this result.

**Proof of Lemma 2.3.6**

*Proof.* We show how to construct such decomposition. By symmetry, it suffices to consider the positive part. Suppose the cut point of \( Q_b \) is \((t_1 = 0, t_2, \ldots, t_k)\) with values \((\mu_1, \ldots, \mu_k)\), all greater than 0 and in an increasing order. Now choose a number \( 0 < \xi_1 < \min(\mu_1, \mu_k - \mu_{k-1}) \), and set the values of \( Q_{b-1} \) as \((\mu_1 - \xi_1, \mu_2 - \xi_1, \ldots, \mu_{k-1} - \xi_1)\), with cut points \((t_1 = 0, t_2, \ldots, t_{k-1})\). Let \( Q_{b_2} \) be cut at \( t_k \), with values \((\xi_1, \mu_k - \mu_{k-1} + \xi_1)\). It is easy to check that this procedure is valid in any case. This proves the lemma.
proof of Theorem 2.3.7

Proof. By Lemma 2.3.4, we know that the statement is valid for \( b_1 = 1 \), and arbitrary \( b_2 \). Now we look at the case where \( b_1 \geq 2 \), \( b_2 \geq 2 \). By Lemma 2.3.6, we can always write

\[
Q_{b_1}(x) = \sum_{i=1}^{b_1-1} \hat{Q}_2^{(i)}(x), \quad Q_{b_2}(y) = \sum_{j=1}^{b_2-1} \hat{Q}_2^{(j)}(y),
\]

where \( \hat{Q}_2^1, \ldots, \hat{Q}_2^{b_1-1} \) and \( \hat{Q}_2^1, \ldots, \hat{Q}_2^{b_2-1} \) are two sets of symmetric increasing 2-bit quantizers. Thus,

\[
\frac{\partial \mathbb{E}(Q_{b_1}(x)Q_{b_2}(y))}{\partial \rho} = \frac{\partial \mathbb{E}\left(\sum_{i=1}^{b_1-1} \hat{Q}_2^{(i)}(x) \sum_{j=1}^{b_2-1} \hat{Q}_2^{(j)}(y)\right)}{\partial \rho} = \sum_{i=1}^{b_1-1} \sum_{j=1}^{b_2-1} \frac{\partial \mathbb{E}(\hat{Q}_2^{(i)}(x)\hat{Q}_2^{(j)}(y))}{\partial \rho} > 0,
\]

where the last equality is due to linearity of expectation and derivative, and the inequality holds because of Lemma 2.3.5. Therefore, \( \mathbb{E}(Q_{b_1}(x)Q_{b_2}(y)) \) is increasing in \( \rho \) for any \( b_1 \geq 1 \) and \( b_2 \geq 1 \).
APPENDIX B
PROOFS OF CHAPTER 3

Proof of Theorem 3.1.1

Proof. The proof follows Murguia et al. (2018), and we present it for completeness. First define

\[ f_Y \triangleq - \sum_{j=2}^{2^{b+1} - 2} P_{Y,j} \log P_{Y,j}. \]

Thus we can write the entropy of \( Y \) as

\[
H[Y] = f_Y - P_{Z,1}P_{R,1} \log P_{Z,1} - P_{Z,1}P_{R,1} \log P_{R,1} \\
- P_{Z,2^b}P_{R,2^b} \log P_{Z,2^b} - P_{Z,2^b}P_{R,2^b} \log P_{R,2^b}.
\]

Moreover, we have

\[
H[Z] = -(P_{R,1} + \ldots + P_{R,2^b}) \sum_{j=1}^{2^b} P_{Z,j} \log P_{Z,j} \\
\triangleq f_Z - P_{Z,1}P_{R,1} \log P_{Z,1} - P_{Z,2^b}P_{R,2^b} \log P_{Z,2^b},
\]

\[
H[R] = -(P_{Z,1} + \ldots + P_{Z,2^b}) \sum_{j=1}^{2^b} P_{R,j} \log P_{R,j} \\
\triangleq f_R - P_{Z,1}P_{R,1} \log P_{R,1} - P_{Z,2^b}P_{R,2^b} \log P_{R,2^b},
\]
which implies \( J_R = H[Y] - H[R] = H[Z] + f_Y - f_Z - f_R \). By plugging in the expressions, we have

\[
f_Y - f_Z - f_R = \sum_{j=2}^{2^b+1-2} A_j,
\]

with

\[
A_j = \begin{cases} 
\sum_{i=1}^{j} P_{Z,i} P_{R,j+1-i} \log \left( \frac{P_{Z,i} P_{R,j+1-i}}{\sum_{k=1}^{2^b} P_{Z,k} P_{R,j+1-k}} \right), & j = 2, \ldots, 2^b, \\
\sum_{i=j+1-2^b}^{2^b} P_{Z,i} P_{R,j+1-i} \log \left( \frac{P_{Z,i} P_{R,j+1-i}}{\sum_{k=i+1-2^b}^{2^b} P_{Z,k} P_{R,j+1-k}} \right), & \text{otherwise.}
\end{cases}
\]

By the log sum inequality, we know that \( A_j \) is convex. Then \( f_Y - f_Z - f_R \) is convex. Since \( H[Z] \) is a constant free of \( P_R \), we know that the objective function \( J_R \) is convex. \( \square \)

**Proof of Theorem 3.2.2**

*Proof.* For a data sample \( u \) with \( \|u\|^2 = m \), the projected data, \( x = w^T u, w \sim \mathcal{N}(0, 1) \), follows \( \mathcal{N}(0, m) \). Denote a neighboring data vector \( u' \) with \( m' = \|u'\|^2 = m + 1 \), and its projected data \( x' = w'^T u \) where \( w', w \) are independent. Let \( f_x(z) \) and \( f_{x'}(z) \) be their p.d.f., respectively. Note that, here the characterization of adjacency is “looser” than requiring \( \|u' - u\| = 1 \), which means that the privacy guarantee would trivially hold for \( \|u' - u\| = 1 \) as in Definition 3.2.1. The proof finds the lower bound of \( m \) such that the required privacy is achieved. Then, by adding the Gaussian noise with variance matching the lower bound, we are assured to achieve same privacy guarantee. First, we have

\[
f_x(z) = \frac{1}{\sqrt{2\pi}\sqrt{m}} \exp\left\{ -\frac{z^2}{2m} \right\},
\]

\[
f_{x'}(z) = \frac{1}{\sqrt{2\pi}\sqrt{m+1}} \exp\left\{ -\frac{z^2}{2(m+1)} \right\}.
\]
Thus,

\[
\frac{f_x(z)}{f_x'(z)} = \sqrt{\frac{m+1}{m}} \exp\{-z^2\left(\frac{1}{2m} - \frac{1}{2(m+1)}\right)\}.
\]

When \( m \geq \frac{1}{2\epsilon'} \), we have for all \( z \in \mathcal{R} \),

\[
\frac{f_x(z)}{f_x'(z)} \leq \sqrt{1 + 2\epsilon'} = \sqrt{\exp\{\log(1 + 2\epsilon')\}} \leq e^{\epsilon'}, \quad (B.1)
\]
due to the inequality \( \log(1 + x) \leq x \) for \( x \geq 0 \). On the other hand,

\[
\frac{f_x'(z)}{f_x(z)} \leq \exp\left\{\frac{z^2}{2m(m+1)}\right\}.
\]

Denote the event \( A_\gamma = \{|Z| \leq \gamma \sqrt{m}\} \). In \( A_\gamma \), we have

\[
\frac{f_x'(z)}{f_x(z)} \leq \exp\left\{\frac{\gamma^2}{2(m+1)}\right\} \leq e^{\epsilon'}, \quad (B.2)
\]
when \( m \geq \frac{\gamma^2}{2\epsilon'} \). By the tail-probability of Gaussian, by controlling the probability of \( A_\gamma \) to be at least \( 1 - \delta' \), we have

\[
P(A_\gamma^c) \leq 2 \exp\{-\gamma^2/2\} = \delta', \quad (B.3)
\]
which implies

\[
\gamma^2 \geq 2 \log \frac{1}{\delta'}.
\]

Thus, \( m \) is required to be \( m \geq \frac{\log \frac{1}{\delta'}}{\epsilon'}. \) When \( \delta' \leq e^{-1/2} \), \( (B.1) \) holds for \( \forall z \in \mathbb{R} \). Thus, for \( \forall z \in \mathbb{R} \), it holds that

\[
f_x'(z) \leq e^{\epsilon'} f_x(z) + \delta'.
\]
In conclusion, when \( m \geq \frac{\log \frac{1}{\epsilon'}}{\epsilon'} \), a single projection preserves \((\epsilon', \delta')\) DP. Now, we need to determine the \( \epsilon' \) and \( \delta' \) to get \((\epsilon, \delta)\)-DP output of \( k \)-fold composition. According to Theorem 3.2.1,

\[
\epsilon = k\epsilon'(e'^r - 1) + \epsilon'\sqrt{2k \log(1/\delta')}
\geq \epsilon'\sqrt{2k \log(1/\delta')}
\]

Solving for \( \epsilon' \) by equation

\[
\epsilon'\sqrt{2k \log(1/\delta')} = \epsilon,
\]

and by setting \( \tilde{\delta} = \delta/2 \), we obtain

\[
\epsilon' = \frac{\epsilon}{\sqrt{2k \log \frac{2}{\delta}}}, \quad \delta' = \frac{\delta}{2k}.
\quad \text{(B.4)}
\]

Therefore, \( m \geq \frac{\log \frac{2k}{\delta}}{\epsilon} \sqrt{2k \log \frac{2}{\delta}} \). Therefore, adding Gaussian noise with variance \( \frac{\log \frac{2k}{\delta}}{\epsilon} \sqrt{2k \log \frac{2}{\delta}} \)
provides \((\epsilon, \delta)\)-DP.

\[\square\]

**Proof of Theorem 3.2.4**

**Proof.** As we can see, the key ingredient in Gaussian mechanism is to identify the sensitivity. For random projection, for two adjacent data vectors with \( \|u\|^2 - u\| = 1 \), we have

\[
\Delta^2 = \|u^TW - u^TW\|^2 \leq \|W_{i,:}\|^2.
\quad \text{(B.5)}
\]

Note that Definition 3.2.2 requires an deterministic bound on the maximal change in output, while for random projections, \( \text{(B.5)} \) is itself random. Thus, the first step is to develop a high probability bound for \( \Delta \), and then allocate the small probability as part of the \( \delta \)-DP budget.
Lemma B.1 (Laurent and Massart (2000)). Let $Y_1, \ldots, Y_n$ be $n$ i.i.d. standard Gaussian random variables. Denote $Z = \sum_{i=1}^n Y_i^2$. Then for any $t > 0$,

$$P[Z \geq n + 2\sqrt{nt} + 2t] \leq \exp\{-t\}.$$ 

By letting the RHS of Lemma B.1 equal to $\tilde{\delta}$, we deduce

$$P\left[\|W_{i,:}\|^2 \geq k + 2\sqrt{k \log \frac{1}{\delta} + 2 \log \frac{1}{\delta}}\right] \leq \tilde{\delta}, \quad (B.6)$$

which means that with probability at least $(1 - \tilde{\delta})$, the sensitivity $\triangle$ is bounded by $\sqrt{k + 2\sqrt{k \log \frac{1}{\tilde{\delta}} + 2 \log \frac{1}{\tilde{\delta}}}}$. Denoting this event as $A$. In this event, if we apply Theorem 3.2.3 with $\sigma = \frac{\sqrt{k + 2\sqrt{k \log \frac{1}{\tilde{\delta}} + 2 \log \frac{1}{\tilde{\delta}}}}}{\epsilon} \sqrt{2 \log \frac{1.25}{\delta}}$, the approach would be $(\epsilon, \delta')$-DP. Taking account the failure probability of event $A$, the final privacy guarantee is $(\epsilon, \delta' + \tilde{\delta})$-DP. Choosing $\delta' = \frac{5}{9} \delta$ and $\tilde{\delta} = \frac{4}{9} \delta$, we finally get a $(\epsilon, \delta)$-DP algorithm by adding $N(0, \sigma^2)$ Gaussian noise with

$$\sigma = \frac{\sqrt{k + 2\sqrt{k \log \frac{2.25}{\delta} + 2 \log \frac{2.25}{\delta}}} \sqrt{2 \log \frac{2.25}{\delta}}}{\epsilon}.$$ 

This completes the proof.

Proof of Theorem 3.2.5

Proof. By the largest borders of $Q$, we are assured to have

$$P[w^T u_i \notin Q] \leq \delta' = \delta/2k.$$
for $\forall i = 1, \ldots, n$. When $w^T \tilde{u}_i \in Q$, by Theorem 3.2, it holds that

$$\frac{f_{w^T \tilde{u}_i}(z)}{f_{w^T \tilde{u}_i}(z)} \leq e^\epsilon', \quad \frac{f_{w^T \tilde{u}_i'}(z)}{f_{w^T \tilde{u}_i}(z)} \leq e^\epsilon'.$$

Then we have for any data point $\tilde{u} \in \tilde{U}$,

$$\max_{i=1, \ldots, 2^b} P[Q(w^T \tilde{u}) = \mu_i] = \max_{i=1, \ldots, 2^b} \frac{\int_{t_{i-1}}^{t_i} f_{w^T \tilde{u}}(z)dz}{\int_{t_{i-1}}^{t_i} f_{w^T \tilde{u}'}(z)dz}$$

$$\leq e^\epsilon' \frac{\int_{t_{i-1}}^{t_i} f_{w^T \tilde{u}'}(z)dz}{\int_{t_{i-1}}^{t_i} f_{w^T \tilde{u}'}(z)dz} = e^\epsilon'.$$

Thus, publishing a single quantized random projection is $(\epsilon'_Q, \delta')$-DP, with $\epsilon'_Q < \epsilon$. By composition theorem, we know that publishing $k$ quantized random projections preserves $(\epsilon_Q, \delta)$-DP, with $\epsilon_Q < \epsilon$.

\[\square\]

**Proof of Theorem 3.2.7**

**Proof.** For DP-RP, we have

$$\mathbb{E}[d_{DP-RP}] = \frac{1}{k} \mathbb{E}[\|u_1^TW + Z_1 - u_2^TW - Z_2\|^2] - 2\sigma^2$$

$$= \frac{1}{k} \mathbb{E}[\langle A \rangle + 2\langle B \rangle + \langle C \rangle] - 2\sigma^2.$$

Since $W$, $Z_1$ and $Z_2$ are independent, the second term $B$ equals to 0. Moreover, $A \overset{d}{=} \|u_1 - u_2\|^2$ is the sum of $k$ i.i.d. squared $N(0, d)$ variables, thus $\|u_1 - u_2\|^2 \sim d\chi_k^2$, where $\chi_k^2$ is the chi-square distribution with $k$ degrees of freedom. Similarly, $C \overset{d}{=} \|Z_1 - Z_2\|^2$
is the sum of $k$ i.i.d. squared $N(0, 2\sigma^2)$, which follows $2\sigma \chi_k^2$. Therefore, we obtain

$$\mathbb{E}[\hat{d}_{DP-RP}] = \frac{dk + 2k\sigma^2}{k} - 2\sigma^2 = d.$$  

For the variance, we have

$$\text{Var}[\hat{d}_{DP-RP}] = \frac{1}{k^2} \mathbb{E} \left[ \left( A + B + C \right)^2 \right] - \left( d + 2\sigma^2 \right)^2.$$  

It is easy to verify that all the cross terms of $\mathbb{E}[AB] = \mathbb{E}[BC] = 0$ by independence. In addition,

$$\mathbb{E}[AC] = \mathbb{E}[A] \cdot \mathbb{E}[C] = 2k^2 d\sigma^2,$$

$$\mathbb{E}[A^2] = \text{Var}[d\chi_k^2] + \mathbb{E}[d\chi_k^2]^2 = 2kd^2 + k^2 d^2,$$

$$\mathbb{E}[B^2] = 8kd\sigma^2,$$

$$\mathbb{E}[C^2] = \text{Var}[2\sigma^2 \chi_k^2] + \mathbb{E}[2\sigma^2 \chi_k^2]^2 = 8k\sigma^4 + 4k^2 \sigma^2.$$  

Combining parts together yields the desired result. \hfill \-box

**Proof of Theorem 3.2.8**

**Proof.** Firstly, by standard Gaussian density transformation we can show that

$$\begin{pmatrix} u_1^T W + \Sigma_1 \\ u_2^T W + \Sigma_2 \end{pmatrix} \sim N^k(0, \begin{pmatrix} \bar{m}_1 & u_1^T u_2 \\ u_1^T u_2 & \bar{m}_2 \end{pmatrix}),$$

$$\begin{pmatrix} \bar{m}_1 & u_1^T u_2 \\ u_1^T u_2 & \bar{m}_2 \end{pmatrix}$$
which has same distribution as \((\sqrt{m_1}x, \sqrt{m_2}y)\), for a pair of random variables \((x, y) \sim N(0, \begin{pmatrix} 1 & \bar{\rho} \\ \bar{\rho} & 1 \end{pmatrix})\) with \(\bar{\rho} = \frac{u_1^T u_2}{\sqrt{m_1} \sqrt{m_2}}\). We can then write the mean estimate as

\[
\mathbb{E}[\hat{d}_{DP-QRP}] = \frac{1}{k} \mathbb{E}[\|Q(u_1^T W + \Sigma_1) - Q(u_2^T W + \Sigma_2)\|^2]
\]

\[
= \frac{1}{k} \mathbb{E}[\|Q(u_1^T W + \Sigma_1)\|^2 + \|Q(u_2^T W + \Sigma_2)\|^2 - 2Q(u_1^T W + \Sigma_1)^T Q(u_2^T W + \Sigma_2)]
\]

\[
= \frac{1}{k} \mathbb{E}[\|\sqrt{m_1} Q_{/\sqrt{m_1}(x)}\|^2 + \|\sqrt{m_2} Q_{/\sqrt{m_2}(y)}\|^2 - 2\sqrt{m_1} \sqrt{m_2} Q_{/\sqrt{m_1}(x)} Q_{/\sqrt{m_2}(y)}]
\]

\[
= \zeta_{2,0} m_1 + \zeta_{0,2} m_2 - 2\zeta_{1,1} \sqrt{m_1 m_2},
\]

where the third equation is due to the linearity of uniform quantizer. The variance can be formulated in a similar way by expanding the variance formula.
Technical Lemmas

Lemma C.1. (Györfi et al., 2002) Let the losses and estimators be defined as in Theorem 4.4.1. under fixed design setting. Let \( \gamma = \text{Var}[y_i], \) \( i = 1, \ldots, n, \) and \( X \) is fixed. Then the expected risk
\[
\mathbb{E}_Y[L(\hat{\beta}^*)] - L(\beta^*) \leq \gamma \frac{\text{rank}(X)}{n}.
\]

Lemma C.2. (Siegel, 1995) Let \( X = \sum_{i=1}^n X_i \) with possibly dependent \( X_i \)'s. \( Y = \sum_{i=1}^n Y_i \) where \( Y_i \)'s are independent copies of \( X_i \)'s (i.e. \( Y_i \) has same distribution as \( X_i, i = 1, \ldots, n \)). If \( B \) is a Chernoff bound on \( \Pr[Y - \mathbb{E}[Y] \geq \epsilon] \), then we have
\[
\Pr[X - \mathbb{E}[X] \geq \epsilon] \leq B^{\frac{\epsilon}{2}}.
\]

Proof of Theorem 4.2.3

Proof. Firstly, we have
\[
\mathbb{E}_{X,Y}[\mathcal{L}(h_Q(x))] = \mathbb{E}_{X \sim \mathcal{X}, Y \sim \eta(X)}[\mathbb{E}_{x \sim \mathcal{X}, y \sim \eta(x)}[\mathbb{1}\{h_Q(x) \neq y\} | X, Y]]
= \mathbb{E}_{X \sim \mathcal{X}, x \sim \mathcal{X}}[\Pr_{y \sim \eta(x), y_Q^{(1)} \sim \eta(x_Q^{(1)})}[y_Q^{(1)} \neq y | X, x]]. (C.1)
\]

We can bound the inner probability for any two points \( x, x' \sim \mathcal{X} \) as
\[
\Pr_{y \sim \eta(x), y' \sim \eta(x')}[y \neq y'|x, x'] = \eta(x)(1 - \eta(x')) + \eta(x')(1 - \eta(x))
= 2\eta(x)(1 - \eta(x)) + (\eta(x) - \eta(x'))(2\eta(x) - 1)
\leq 2\eta(x)(1 - \eta(x)) + \|\eta(x) - \eta(x')\|, (C.2)
\]
by the definition of \( \eta(x) \). Here we use the fact \( |2\eta(x) - 1| \leq 1 \). Combining (C.2) and (C.1) we have

\[
\mathbb{E}_{X,Y}[\mathcal{L}(h_Q(x))] \leq \mathbb{E}_{X\sim x, x\sim x}[2\eta(x)(1 - \eta(x)) + \|\eta(x) - \eta(x_Q^{(1)})\|].
\]

Notice from a classical result that

\[
\mathcal{L}(h^*(x)) = \min\{\eta(x), 1 - \eta(x)\},
\]

we obtain

\[
\mathbb{E}_{X,Y}[\mathcal{L}(h_Q(x))] \leq 2\mathcal{L}(h^*(x)) + \mathbb{E}_{X,Y,x\sim x}[\|\eta(x) - \eta(x_Q^{(1)})\|].
\]

The first term is the Bayes risk, and it remains to bound the second term. By Theorem 1, given that \( k = O(\omega^{-2}(\gamma(T))^2 + \log(2/\delta)) \), with probability \( 1 - \delta \) we have

\[
(1 - \omega)\|x - y\|^2 \leq \left\| \frac{1}{\sqrt{k}} R^T x - \frac{1}{\sqrt{k}} R^T y \right\|^2 \leq (1 + \omega)\|x - y\|^2, \forall x, y \in \mathcal{X}. \quad (C.3)
\]

Denote this event \( \Omega \). Now we proceed our analysis in \( \Omega \) (with high probability). The space of projected samples \( S_R = \left\{ \frac{1}{\sqrt{k}} R^T x_1, ..., \frac{1}{\sqrt{k}} R^T x_n \right\} \) is now bounded by

\[
u_R = \left\| \frac{1}{\sqrt{k}} R^T x \right\| \leq \sqrt{1 + \omega},
\]

by taking \( y = 0 \) in (C.3). Therefore, \( S_R \subset [-u_R, u_R]^k \). Now we cover \( [-u_R, u_R] \) by \( N_C = (2u_R/\epsilon)^k \) boxes with length \( \epsilon \). For the test sample \( x \), let \( B_\epsilon(x) \) be the box containing \( \frac{1}{\sqrt{k}} R^T x \). In the event \( \Omega \), we have

\[
\mathbb{E}_{X,Y,x}[\|\eta(x) - \eta(x_Q^{(1)})\||\Omega] = \mathbb{E}_{X,Y,x}[\|\eta(x) - \eta(x_Q^{(1)})\||\Omega, V] Pr(V)
+ \mathbb{E}_{X,Y,x}[\|\eta(x) - \eta(x_Q^{(1)})\||\Omega, V^c] Pr(V^c),
\]
where the event \( V = \{ B_\epsilon(x) \cap S_R(X) = \emptyset \} \), and \( V^c \) is its complement. Further, we denote \( S_R(X) = \{ \frac{1}{\sqrt{k}} R^T x_1, \ldots, \frac{1}{\sqrt{k}} R^T x_n \} \) as the projected samples. By Lemma 19.2 in Shalev-Shwartz and Ben-David (2014), we have

\[
Pr(V) \leq \frac{N_C}{ne}.
\]

Thus,

\[
\mathbb{E}_{X,Y,x}[\|\eta(x) - \eta(x^{(1)}_Q)\|_\Omega] \leq \frac{N_C}{ne} + \mathbb{E}_{X,Y,x}[\|\eta(x) - \eta(x^{(1)}_Q)\|_\Omega, V^c] \\
\leq \frac{N_C}{ne} + L \cdot \mathbb{E}_{X,Y,x}[\|x - x^{(1)}_Q\|_\Omega, V^c] \\
\leq \frac{N_C}{ne} + \frac{L}{\sqrt{1 - \omega}} \mathbb{E}_{X,Y,x}[\frac{1}{\sqrt{k}} R^T x - \frac{1}{\sqrt{k}} R^T x^{(1)}_Q\|_V^c],
\]

where the second line is because \( \eta(x) \) is \( L \)-Lipschitz and the last line is due to (C.3).

![Figure C.1: An illustration of bounding the distance in projected space \( S_R \) with two covers. For simplicity we omit the scaling term \( \frac{1}{\sqrt{k}} \). Boxes resulted from red dash lines are the \( \epsilon \)-cover constructed by hand, and boxes surrounded by blue solid lines are induced cover by \( Q \).](image)

To bound the second term, we consider another cover of \( S_R \) which is intrinsically induced by the borders of \( Q \). Denote \( x^{(1)}_R \) as the nearest point of \( x \) in the projected space. In this case, we know that \( \| \frac{1}{\sqrt{k}} R^T x - \frac{1}{\sqrt{k}} R^T x^{(1)}_R \| \leq \epsilon \sqrt{k} \). However, \( \| \frac{1}{\sqrt{k}} R^T x - \frac{1}{\sqrt{k}} R^T x^{(1)}_Q \| \) cannot
be bounded in this way, due to the discretization of quantizing function $Q$. In a simple example (Figure C.1), assume we only have 3 points in a 2-D case. Denote the centroids of these 3 points respectively as $\mu(\frac{1}{\sqrt{k}} R^T x)$, $\mu(\frac{1}{\sqrt{k}} R^T x^{(1)}_R)$ and $\mu(\frac{1}{\sqrt{k}} R^T x^{(1)}_Q)$. In this plot, $V^c$ is obviously satisfied since there are two points in the same $\epsilon$-box. Now the green point $\frac{1}{\sqrt{k}} R^T x^{(1)}_R$ is closer to black point $\frac{1}{\sqrt{k}} R^T x^{(1)}_R$ in the space of $\mathcal{S}_R$, but after quantization, the nearest neighbor returned changes to the pink point $\frac{1}{\sqrt{k}} R^T x^{(1)}_Q$, since $\mu(\frac{1}{\sqrt{k}} R^T x^{(1)}_Q)$ lies closer to $\mu(\frac{1}{\sqrt{k}} R^T x^{(1)}_R)$ than $\mu(\frac{1}{\sqrt{k}} R^T x^{(1)}_Q)$. However, $\|\frac{1}{\sqrt{k}} R^T x - \frac{1}{\sqrt{k}} R^T x^{(1)}_Q\|$ might be greater than $\epsilon \sqrt{2}$.

Note that for uniform quantizer, the distances between nearby reconstruction levels all equal to $\Delta$, and the distances between consecutive borders (view $-\sqrt{1-\omega}$ and $\sqrt{1-\omega}$ as borders too) are upper bounded by $g_Q(-\sqrt{1-\omega}, \sqrt{1-\omega})$. Using triangle inequality, we get

$$[\|\frac{1}{\sqrt{k}} R^T x - \frac{1}{\sqrt{k}} R^T x^{(1)}_Q\| V^c]$$

$$\leq [\|\frac{1}{\sqrt{k}} R^T x - \mu(\frac{1}{\sqrt{k}} R^T x)\| + \|\frac{1}{\sqrt{k}} R^T x^{(1)}_Q - \mu(\frac{1}{\sqrt{k}} R^T x^{(1)}_Q)\|$$

$$+ \|\mu(\frac{1}{\sqrt{k}} R^T x) - \mu(\frac{1}{\sqrt{k}} R^T x^{(1)}_Q)\| V^c]$$

$$\leq \frac{\Delta \sqrt{k}}{2} + \frac{\Delta \sqrt{k}}{2} + \|\mu(\frac{1}{\sqrt{k}} R^T x) - \mu(\frac{1}{\sqrt{k}} R^T x^{(1)}_Q)\| V^c].$$

To bound the last term, we need to find out that given two points $a, b$ in a same $\epsilon$-box, how large the distance between their quantized centroids $\mu(a), \mu(b)$ can be. We proceed by noticing that for a given $B_\epsilon(x)$ with any $\epsilon$, the maximum number of different $Q$-boxes that can be contained (perhaps partially) on the diagonal of $B_\epsilon(x)$ is equal to $\lfloor \frac{\epsilon}{g_Q} \rfloor + 2$. The largest distance between the centroids occurs when two points fall into the two regions on the diagonal endpoints (as black and green stars in Figure C.1), which equals to $(\lfloor \frac{\epsilon}{g_Q} \rfloor + 1) \Delta \sqrt{k}$. 
Therefore, we have

\[
\| \mu \left( \frac{1}{\sqrt{k}} R^T x \right) - \mu \left( \frac{1}{\sqrt{k}} R^T x^{(1)}_R \right) \| \leq \left( \frac{\epsilon}{g_Q} \right) + 1 \triangle \sqrt{k} \leq \frac{\epsilon \triangle}{g_Q} + \triangle \sqrt{k},
\]

where for simplicity we write \( g_Q \) instead of \( g_Q(-\sqrt{1-\omega}, \sqrt{1-\omega}) \). Hence, we get the worst case bound

\[
\mathbb{E}_{X,Y,x} \left[ \left\| \frac{1}{\sqrt{k}} R^T x - \frac{1}{\sqrt{k}} R^T x^{(1)}_Q \right\| \right] \leq \left( \frac{\epsilon \triangle g_Q}{g_Q} + 2 \triangle \sqrt{k} \right).
\]

Combining with previous result, we obtain

\[
\mathbb{E}_{X,Y,x} \left[ \left\| \eta(x) - \eta(x^{(1)}_Q) \right\| \right] \leq \frac{N C}{n e} + \frac{L \triangle \epsilon \sqrt{k}}{g_Q \sqrt{1-\omega}} + \frac{2L \triangle \sqrt{k}}{\sqrt{1-\omega}}.
\]

Now we choose \( \epsilon \) to minimize the RHS. Let \( f(\epsilon) = \frac{\epsilon^{1+\omega}}{n e} + \frac{L \epsilon \epsilon \sqrt{k}}{\sqrt{k}} \frac{\sqrt{1-\omega}}{\sqrt{1-\omega}} \). Following the standard technique, we take the derivative for \( f \) with respect to \( \epsilon \) and set it to zero, which yields

\[
\epsilon^* = \left( 2 \sqrt{\frac{1+\omega}{1-\omega}} \right)^{\frac{k}{k+1}} \frac{L \triangle \epsilon \sqrt{k}}{g_Q \sqrt{1-\omega}} \left( \frac{k}{n e} \right)^{\frac{1}{k+1}} \sqrt{k}^{\frac{k+1}{k+1}}.
\]

Plugging in the expression and after some calculation we get

\[
\mathbb{E}_{X,Y,x} \left[ \left\| \eta(x) - \eta(x^{(1)}_Q) \right\| \right] \leq \left( \frac{2L \triangle \sqrt{\frac{1+\omega}{1-\omega}} \frac{1}{k+1}}{g_Q} \right)^{\frac{k}{k+1}} \left( n e \right)^{\frac{1}{k+1}} \sqrt{k} \left( \frac{2k+1}{k+1} \right) + \sqrt{k} \left( \frac{1}{k+1} \right).
\]

Following Kabán (2015), we have \( 2^{\frac{k}{k+1}} \left( \frac{2k+1}{k+1} + \sqrt{k} \frac{1}{k+1} \right) \leq 2 \sqrt{2} \). Replacing the terms and combining all parts together, the proof is complete.
**Proof of Theorem 4.2.4**

*Proof.* The proof is based on the probability that $x_Q^{(1)}$ is different from $x^{(1)}$. First we have

$$
\mathbb{E}_{X,Y,R}[L(h_Q(x))] = \mathbb{E}_{X,Y}[\mathbb{E}_{x,y,R}[1\{h_Q(x) \neq y\}|X,Y]]
$$

$$
= \mathbb{E}_{X,Y}\{\mathbb{E}_{x,y,R}[1\{h_S(x) \neq y\}1\{h_Q(x) = h_S(x)\}] + 1\{h_S(x) = y\}1\{h_Q(x) \neq h_S(x)\}|X,Y]\}
$$

$$
\leq \mathbb{E}_{X,Y}\{\mathbb{E}_{x,y,R}[1\{h_S(x) \neq y\} + 1\{h_Q(x) \neq h_S(x)\}|X,Y]\}
$$

$$
\triangleq A + B.
$$

We recognize the term $A$ is simply the risk of data space NN classifier, $A = \mathbb{E}_{X,Y}[L(h_S(x))]$.

It suffices to study term $B$. Note that

$$
B = \mathbb{E}_{X\sim X, x\sim X,y^{(1)}\sim \eta(x^{(1)}), y_Q^{(1)}\sim \eta(x_Q^{(1)}), R}[y_Q^{(1)} \neq y^{(1)}]
$$

$$
= \mathbb{E}_{X,x}\{Pr_y^{(1)}\sim \eta(x^{(1)}), y_Q^{(1)}\sim \eta(x_Q^{(1)}), R}[x_Q^{(1)} \neq x^{(1)}, y_Q^{(1)} \neq y^{(1)}|X,x]\}
$$

$$
\leq \mathbb{E}_{X,x}\{Pr_R[x_Q^{(1)} \neq x^{(1)}|X,x]\}
$$

$$
\triangleq \mathbb{E}_{X,x}\{p_c\},
$$

where the second line is because $y_R^{(1)} \neq y^{(1)}$ implies that $x_Q^{(1)} \neq x^{(1)}$. For a fixed $X$ and $x$, we denote the set $G = X/x^{(1)}$. Then for the inner probability, we have

$$
P_c = \sum_{i:x_i \in G} Pr_R[x_Q^{(1)} = x_i|X,x]
$$

$$
= \sum_{i:x_i \in G} Pr[\bigcap_{x_j \neq x_i} \{\hat{\rho}_Q(x, x_i) \geq \hat{\rho}_Q(x, x_j)\}|X,x]
$$

$$
\leq \sum_{i:x_i \in G} Pr[\hat{\rho}_Q(x, x_i) \geq \hat{\rho}_Q(x, x^{(1)})|X,x]
$$

(C.4)

due to the equivalence of inner product and Euclidean distance estimation. Under the
asymptotic assumption $k \to \infty$, by Central Limit Theorem (CLT) we know that for any $x, y \in X$,

$$
\hat{\rho}_Q(x, y) \sim N(\alpha \rho_{x,y}, \frac{\sigma_{x,y}^2}{k}),
$$

for $\sigma_{x,y} = Var[Q(r^T x)^T Q(r^T y)]$ a fixed constant given $x, y$. Here $r$ is a column of $R$. Next, we obtain for $\forall i, j$,

$$
\hat{\rho}_Q(x, x_i) - \hat{\rho}_Q(x, x_j) \sim N(\alpha (\rho_{x,x_i} - \rho_{x,x_j}), \sigma^2),
$$

with $\sigma^2 = \sigma_{x,x_i}^2 + \sigma_{x,x_j}^2 - 2\text{Corr}(\hat{\rho}_Q(x, x_i) \hat{\rho}_Q(x, x_j) \sigma_{x,x_i} \sigma_{x,x_j})$. Therefore,

$$
\Pr[\hat{\rho}_Q(x, x_i) \geq \hat{\rho}_Q(x, x_j)] = \Pr[\hat{\rho}_Q(x, x_i) - \hat{\rho}_Q(x, x_j) \geq 0]
$$

$$
= \Phi\left( \frac{\sqrt{k} \alpha (\rho_{x,x_i} - \rho_{x,x_j})}{\sqrt{\sigma_{x,x_i}^2 + \sigma_{x,x_j}^2 - 2\text{Corr}(\hat{\rho}_Q(x, x_i) \hat{\rho}_Q(x, x_j) \sigma_{x,x_i} \sigma_{x,x_j})}} \right)
$$

$$
= \Phi\left( \frac{\sqrt{k} (\rho_{x,x_i} - \rho_{x,x_j})}{\sqrt{\xi_{x,x_i}^2 + \xi_{x,x_j}^2 - 2\text{Corr}(\hat{\rho}_Q(x, x_i) \hat{\rho}_Q(x, x_j) \xi_{x,x_i} \xi_{x,x_j})}} \right),
$$

since by the definition of debiased variance we have $\xi_{x,x_i}^2 = \frac{\sigma_{x,x_i}^2}{\alpha^2}$. Now plugging above equation into (C.4), we have

$$
B = \mathbb{E}_{X,x} \left[ \sum_{i:x_i \in G} \Phi\left( \frac{\sqrt{k} (\cos(x, x_i) - \cos(x, x^{(1)}))}{\sqrt{\xi_{x,x_i}^2 + \xi_{x,x^{(1)}}^2 - 2\text{Corr}(\hat{\rho}_Q(x, x_i) \hat{\rho}_Q(x, x^{(1)}) \xi_{x,x_i} \xi_{x,x^{(1)}})}} \right) \right],
$$

by noting that $\rho_{x,x_i} = \cos(x, x_i)$. Combining parts together, we get the result as required.

$\square$
Proof of Lemma 4.2.5

Proof. Denote the random projection matrix $R \in \mathbb{R}^{d \times k}$. Recall that the estimates of $\rho_{xy}$ and $\rho_{xz}$ are

$$\hat{\rho}_R(x, y) = \frac{x^T RR^T y}{k}, \quad \hat{\rho}_R(x, z) = \frac{x^T RR^T z}{k}.$$ 

Denote the columns of $R$ as $[r_1, \ldots, r_k]$, we have

$$\mathbb{E}[\hat{\rho}_R(x, y)\hat{\rho}_R(x, z)]$$

$$= \frac{1}{k^2} \mathbb{E}[xRR^T y^T xRR'^T z^T]$$

$$= \frac{1}{k^2} \left[ \langle x, r_1 \rangle, \ldots, \langle x, r_k \rangle \right] \left[ \langle y, r_1 \rangle \right]^T \left[ \langle x, r_1 \rangle, \ldots, \langle x, r_k \rangle \right] \left[ \langle z, r_1 \rangle \right]^T$$

$$= \frac{1}{k^2} \left( \sum_{i=1}^{k} \langle x, r_i \rangle \langle y, r_i \rangle \right) \cdot \left( \sum_{i=1}^{k} \langle x, r_i \rangle \langle z, r_i \rangle \right)$$

$$= \frac{1}{k^2} \left( \sum_{i=1}^{k} \sum_{p=1}^{d} x_p r_{ip} \left( \sum_{q=1}^{d} y_q r_{iq} \right) \right) \cdot \left( \sum_{i=1}^{k} \sum_{s=1}^{d} x_s r_{is} \left( \sum_{t=1}^{d} y_t r_{jt} \right) \right)$$

$$= \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{p=1}^{d} \sum_{q=1}^{d} \sum_{s=1}^{d} \sum_{t=1}^{d} x_p y_q x_s z_t \mathbb{E}[r_{ip} r_{iq} r_{is} r_{jt}]$$

$$= \frac{1}{k^2} \left( \sum_{i=1}^{k} \sum_{j \neq i} \sum_{p=1}^{d} \sum_{q=1}^{d} \sum_{s=1}^{d} \sum_{t=1}^{d} x_p y_q x_s z_t \mathbb{E}[r_{ip} r_{iq} r_{is} r_{jt}] \right)$$

$$+ \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{p=1}^{d} \sum_{q=1}^{d} \sum_{s=1}^{d} \sum_{t=1}^{d} x_p y_q x_s z_t \mathbb{E}[r_{ip} r_{iq} r_{is} r_{jt}].$$

$$= A + B.$$

For the first term $A$, since $i \neq j$ and all entries of $R$ are i.i.d. standard normal, the expectation is non-zero only when $p = q$ and $s = t$. Also note the each row vector $r_i$ and $r_j$
are independent. Consequently we obtain

\[ A = \frac{k(k-1)}{k^2} \left( \sum_{p=1}^d \sum_{s=1}^d x_p y_p x_s z_s \right) = \frac{k-1}{k} \langle x, y \rangle \cdot \langle x, z \rangle = \frac{k-1}{k} \rho_{xy} \rho_{xz}. \]

For term \( B \), we note that the expectation is non-zero when: (i) \( p = q \) and \( s = t \); (ii) \( p = s \) and \( q = t \); or (iii) \( p = t \) and \( q = s \). In these cases, when \( p, q, s, t \) are not all equal, the expected value is simply \( \mathbb{E}[r^2_{ip} r^2_{iq}] = 1 \). When \( p = q = s = t \), the expected value is \( \mathbb{E}[r^4_{ip}] = 3 \). Therefore we have

\[
B = \frac{k}{k^2} \left( \sum_{p=1}^d \sum_{s=1}^d x_p y_p x_s z_s + \sum_{p=1}^d \sum_{q=1}^d x_p y_q x_p z_q + \sum_{p=1}^d \sum_{q=1}^d x_p y_q x_q z_p - 2 \times 3 \sum_{p=1}^d x_p y_p x_p z_p \right)
\]

\[ = \frac{1}{k} [\langle x, y \rangle \cdot \langle x, z \rangle + \|x\|^2 \langle y, z \rangle + \langle x, y \rangle \cdot \langle x, z \rangle + 3 \times 2 \sum_{p=1}^d x_p y_p x_p z_p - 2 \times 3 \sum_{p=1}^d x_p y_p x_p z_p] \]

\[ = \frac{1}{k} (\rho_{yz} + 2 \rho_{xy} \rho_{xz}), \]

where the first line is due to the fact that we count the case \( p = q = s = t \) for three times.

Now putting parts together, we have

\[
\text{Cov}(\hat{\rho}_R(x, y), \hat{\rho}_R(x, z)) = \frac{1}{k^2} \mathbb{E}[x^T RR^T y x^T RR^T z] - \mathbb{E}[\hat{\rho}_R(x, y)] \mathbb{E}[\hat{\rho}_R(x, z)]
\]

\[ = \frac{(k-1) \rho_{xy} \rho_{xz} + \rho_{yz} + 2 \rho_{xy} \rho_{xz} - \rho_{xy} \rho_{xz}}{k} \]

\[ = \frac{1}{k} (\rho_{yz} + \rho_{xy} \rho_{xz}). \]

\[ \Box \]
Proof of Proposition 4.3.1

Proof. To start with, we notice that for \( x_i, y_i \), \( i = 1, ..., k \) all i.i.d. standard normal,

\[
F_{k,k}(\frac{1-\rho}{1+\rho}) = Pr[\sum_{i=1}^{k} x_i^2 \leq \frac{1-\rho}{1+\rho} \sum_{i=1}^{k} y_i^2] = Pr[\frac{1}{k} \sum_{i=1}^{k} x_i^2 \leq \frac{1-\rho}{1+\rho} (\frac{1}{k} \sum_{i=1}^{k} y_i^2)].
\]

By Central Limit Theorem we have \( w = \frac{1}{k} \sum_{i=1}^{k} x_i^2 \sim N(1, 2/k) \), \( z = \frac{1}{k} \sum_{i=1}^{k} y_i^2 \sim N(1, 2/k) \) and they are independent. Hence, when \( k \to \infty \), we have

\[
f_k(\rho) = F_{k,k}(\frac{1-\rho}{1+\rho}) = Pr[w \leq \frac{1-\rho}{1+\rho} z]
= \int_{-\infty}^{\infty} \frac{\sqrt{k}}{2\sqrt{\pi}} e^{-\frac{k(z-1)^2}{4}} dwdz
= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} \Phi\left(\frac{1-\rho}{\sqrt{2k\rho}}(s - 1)\right) dz
= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} \Phi\left(\frac{(1-\rho)s - \sqrt{2k\rho}}{1+\rho}\right) ds
= \mathbb{E}_s[\Phi\left(\frac{(1-\rho)s - \sqrt{2k\rho}}{1+\rho}\right)],
\]

where the second and third line are derived by simple change of variable, and \( s \sim N(0, 1) \).

For another \( v \sim N(0, 1) \) independent of \( s \), by law of total expectation we obtain

\[
\mathbb{E}_s[\Phi\left(\frac{(1-\rho)s - \sqrt{2k\rho}}{1+\rho}\right)] = \mathbb{E}_{s,v}[1 \{ v \leq \frac{(1-\rho)s - \sqrt{2k\rho}}{1+\rho}\}]
= Pr[v - \frac{1-\rho}{1+\rho} s \leq -\frac{\sqrt{2k\rho}}{1+\rho}]
= Pr\left[\frac{v - \frac{(1-\rho)s}{1+\rho}}{\sqrt{1 + (\frac{(1-\rho)s}{1+\rho})^2}} \leq -\frac{\sqrt{2k\rho}}{(1+\rho)\sqrt{1 + (\frac{(1-\rho)^2}{(1+\rho)^2})}}\right]
= \Phi\left(-\frac{\sqrt{k\rho}}{\sqrt{1 + \rho^2}}\right) = \tilde{f}_k(\rho).
\]

This completes the proof. \(\square\)
Proof of Theorem 4.3.2

Proof. The proof follows from Durrant and Kabán (2013). First by classical VC theory (Vapnik, 1995), with probability $1 - \delta$ we have

$$
Pr[\hat{H}_Q(x) \neq y] \leq \hat{L}_{(0,1)}(S_Q, \hat{h}_Q) + 2\sqrt{\frac{(k + 1) \log \frac{2en}{k+1} + \log \frac{1}{\delta}}{n}},
$$

where $\hat{L}_{(0,1)}(S_Q, \hat{h}_Q) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{(0,1)}(\hat{H}_Q(Q(R^T x_i)), y_i)$ the empirical loss in the quantized space (with optimal ERM quantizer $\hat{h}_Q$ in $S_Q$). Since $\hat{h}_Q$ is the minimizer of $\hat{L}_{(0,1)}(S_Q, \hat{h}_Q)$, we have

$$
\hat{L}_{(0,1)}(S_Q, \hat{h}_Q) \leq \hat{L}_{(0,1)}(S_Q, Q(R^T \hat{h}))
$$

$$
= \hat{L}_{(0,1)}(S, \hat{h}) + \mathcal{L}_{(0,1)}(S_Q, Q(R^T \hat{h})) - \hat{L}_{(0,1)}(S, \hat{h})
$$

$$
\leq \hat{L}_{(0,1)}(S, \hat{h}) + \frac{1}{n} \sum_{i=1}^{n} 1\{\text{sign}(Q(\hat{h}^T R)Q(R^T x_i)) \neq \text{sign}(\hat{h}^T x_i)\}
$$

$$
:= \hat{L}_{(0,1)}(S, \hat{h}) + M.
$$

We note that $M$ is a sum of dependent flipping probabilities because of the commonly used projection matrix $R$. Using Markov’s Inequality we have

$$
M \leq (1 + \frac{1 - \delta}{\delta})\mathbb{E}_R[M]
$$

with probability $1 - \delta$. To get a better bound with small $\delta$, we make use of Lemma C.2. By applying the lemma, if $M^*$ is an independent copy of $M$, standard Chernoff bound gives

$$
Pr[M^* \geq (1 + \epsilon)\mathbb{E}_R[M^*]] \leq \exp(-n\mathbb{E}_R[M^*]\epsilon^2/3).
$$
Then, Lemma C.2 yields

\[ Pr[M \geq (1 + \epsilon) \mathbb{E}_R[M]] \leq \exp(-n\mathbb{E}_R[M^*] \epsilon^2 / 3) \frac{1}{n} \]

\[ = \exp(-\mathbb{E}_R[M] \epsilon^2 / 3). \]

Transforming probability bound to expectation bound, we obtain with probability \(1 - \delta\),

\[ M \leq \mathbb{E}_R[M] + \sqrt{3\mathbb{E}_R[M] \log \frac{1}{\delta}}. \]

The proof is completed by noting that \( \mathbb{E}_R[M] = \sum_{i=1}^{n} \Phi(-\frac{\sqrt{k} |\rho_i|}{\xi_i}) \) as \(k \to \infty\), which could be easily derived from Central Limit Theorem and Proposition 1.

\[ \square \]

**Proof of Theorem 4.4.1**

**Proof.** By applying Lemma C.1 we have

\[ \mathbb{E}_{Y|R}[L_Q(\hat{\beta}_Q^*)] - L_Q(\beta_Q^*) \leq \frac{\gamma}{n}. \quad (C.5) \]

Since \( \beta_Q^* \) is the minimizer of the squared loss in the quantized space, by elementary algebra we have that

\[ L_Q(\beta_Q^*) \leq L_Q(\frac{1}{\sqrt{k(1 - D_Q)}} R^T \beta^*) \]

\[ = \frac{1}{n} \mathbb{E}_{Y|R}[||Y - \frac{1}{k(1 - D_Q)} Q(XR)R^T \beta^*||^2] \]

\[ = \frac{1}{n} \mathbb{E}_{Y|R}[||Y - X\beta^*||^2] + \frac{1}{n} ||X\beta^* - \frac{1}{k(1 - D_Q)} Q(XR)R^T \beta^*||^2 \]

\[ \overset{(a)}{=} \frac{1}{n} \mathbb{E}_{Y|R}[||Y - X\beta^*||^2] + \frac{1}{n} ||X\beta^* - \frac{1}{k(1 - D_Q)} Q(XR)R^T \beta^*||^2 \]

\[ = L(\beta^*) + (\beta^*)^T \Sigma \beta^* + \frac{2}{nk(1 - D_Q)} (\beta^*)^T RQ(XR)^T X\beta^* \]

\[ + \frac{1}{nk^2(1 - D_Q)^2} (\beta^*)^T RQ(XR)^T Q(XR)R^T \beta^*. \quad (C.6) \]
where (a) is due to $Y - X \beta^* = \epsilon$ is i.i.d zero-mean Gaussian independent of $R$. Here, the factor $\frac{1}{1 - D_Q}$ is again related to cosine estimation, and we will provide some discussions at the end of the proof. Recall the notation $X = [x_1, ..., x_n]^T$ with $x_i$ having unit norm, and $R = [r_1, ..., r_k]$. We denote $z_{ip} \Delta \langle x_i, r_p \rangle$. Hence, the quantized matrix $Q(XR)$ has $z_{ip}$ as the $(i, p)$-th entry, for $i = 1, ..., n$ and $p = 1, ..., k$. It is obvious that

$$\begin{pmatrix} z_{ip} \\ z_{jp} \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{ij} \\ \rho_{ij} & 1 \end{pmatrix} \right),$$

(C.7)

where $\rho_{ij} = \langle x_i, x_j \rangle$. Moreover, Theorem 2.2.1 then gives $E[z_{ip}Q(z_{jp})] = (1 - D_Q)(x_i, x_j)$.

Further denote $\tilde{\beta} = \beta^*/\|\beta^*\|$ the standardized true parameter vector. It follows that

$$E[(\beta^*)^T RQ(XR)^T X \beta^*] = E[(\beta^*)^T RQ(XR)^T X \beta^*]$$

$$= E \left[ (\tilde{\beta})^T RQ(XR)^T X \beta^* \| \beta^* \| \right]$$

$$= E \left[ \sum_{p=1}^{k} z_{\tilde{\beta},p} Q(z_{1p}), ..., \sum_{p=1}^{k} z_{\tilde{\beta},p} Q(z_{np}) \right] X \beta^* \| \beta^* \|$$

$$\stackrel{(b)}{=} k(1 - D_Q)(\beta^*)^T X^T X \beta^*$$

$$= nk(1 - D_Q)(\beta^*)^T \Sigma \beta^*. \quad \text{(C.8)}$$

Here, $z_{\tilde{\beta},p} = \langle \tilde{\beta}, r_p \rangle$, and (b) is due to Theorem 2.2.1. Note that for $(x, y)$ following distribution (C.7) with cosine $\rho$, we have

$$E[x^2 Q(y)^2] = E[(\rho y + \sqrt{1 - \rho^2 W})^2 Q(y)^2]$$

$$= \rho^2 \xi_{2,2} + (1 - \rho^2)(1 - D_Q), \quad \text{(C.9)}$$
where $W \sim N(0, 1)$ is independent of $x, y$, and $\xi_{2,2} \triangleq \mathbb{E}[y^2Q(y)^2]$ for $y \sim N(0, 1)$. Denote $\rho_{\beta,i} = \langle \tilde{\beta}, x_i \rangle$. Now we can obtain

$$
\mathbb{E}[(\beta^*)^T R Q(XR)^T Q(XR) R^T \beta^*] = \|\beta^*\|^2 \mathbb{E} \left[ \tilde{\beta}^T R Q(XR)^T Q(XR) R^T \tilde{\beta} \right] = \|\beta^*\|^2 \mathbb{E} \left[ \sum_{i=1}^{n} \left( \sum_{p=1}^{k} z_{\tilde{\beta},p} Q(z_{ip}) \right)^2 \right] = \|\beta^*\|^2 \sum_{i=1}^{n} \mathbb{E} \left[ k \sum_{p=1}^{k} z_{\tilde{\beta},p}^2 Q(z_{ip})^2 + k \sum_{p=1}^{k} \sum_{q \neq p} z_{\tilde{\beta},p} Q(z_{ip}) z_{\tilde{\beta},q} Q(z_{iq}) \right] = \|\beta^*\|^2 \sum_{i=1}^{n} \left[ k(\xi_{2,2} \rho_{\beta,i}^2 + (1 - \rho_{\beta,i}^2)(1 - D_Q)) + k(k - 1)(1 - D_Q)^2 \rho_{\beta,i}^2 \right]. \tag{C.10}
$$

In the above, (C.10) holds because of (C.9) and the fact that $z_{\tilde{\beta},p} Q(z_{ip})$ is independent of $z_{\tilde{\beta},q} Q(z_{iq})$ for any $p \neq q$. By noticing that

$$
\sum_{i=1}^{n} \rho_{\beta,i}^2 = \frac{(\beta^*)^T X^T X \beta^*}{\|\beta^*\|^2} = \frac{n(\beta^*)^T \Sigma \beta^*}{\|\beta^*\|^2},
$$

we can further have

$$
\mathbb{E}[(\beta^*)^T R Q(XR)^T Q(XR) R^T \beta^*] = \|\beta^*\|^2 \left[ k(\xi_{2,2} \sum_{i=1}^{n} \rho_{\beta,i}^2 + (n - \sum_{i=1}^{n} \rho_{\beta,i}^2)(1 - D_Q)) + k(k - 1)(1 - D_Q)^2 \sum_{i=1}^{n} \rho_{\beta,i}^2 \right] = nk(1 - D_Q)\|\beta^*\|^2 + n \left[ k(\xi_{2,2} - 1 + D_Q) + k(k - 1)(1 - D_Q)^2 \right] (\beta^*)^T \Sigma \beta^*. \tag{C.11}
$$
Now, taking expectation on both sides of (C.6) \( w.r.t. \) \( R \) and combining (C.8) and (C.11), we have

\[
\mathbb{E}_R[L_Q(\beta^*_Q)] \\
\leq L(\beta^*) + \left[ 1 - 2 + \frac{\xi_{2,2} - 1 + D_Q}{k(1 - D_Q)^2} + \frac{k - 1}{k} \right] (\beta^*)^T \Sigma \beta^* + \frac{1}{k(1 - D_Q)} \| \beta^* \|^2 \\
= L(\beta^*) + \frac{1}{k} \| \beta^* \|^2 \Omega,
\]

where \( \Omega = [\xi_{2,2}^{-1} + D_Q]^{-1} \Sigma + \frac{1}{1 - D_Q} I_d, \) with \( \| \beta^* \|_\Omega = \sqrt{(\beta^*)^T \Omega \beta^*} \), and \( I_d \) the identity matrix. Lastly, taking expectation \( w.r.t. \) \( R \) in (C.5), we obtain

\[
\mathbb{E}_{Y,R}[L_Q(\hat{\beta}_Q^*)] \\
\leq \mathbb{E}[L_Q(\beta^*_Q)] + \frac{k}{n} \\
\leq \frac{k}{n} + L(\beta^*) + \frac{1}{k} \| \beta^* \|^2 \Omega.
\]

This completes the proof. Now we briefly discuss the role of factor \( \frac{1}{1 - D_Q} \) in (C.6). Note that in our model, \( X \beta^* = \| \beta^* \| X \hat{\beta} = \| \beta^* \| [\rho_{\beta,1}, \ldots, \rho_{\beta,n}]^T \) can be regarded as the scaled cosine between data vectors and the true parameter, and \( Q(XR)R^T \beta^* \) is then a biased estimator of \( X \beta^* \) with mean equal to \( (1 - D_Q)X \beta^* \), according to Theorem 2.2.1. Therefore, the factor \( \frac{1}{1 - D_Q} \) acts as a debiasing operation—Similar in spirit to the previous analysis for classification problems. \( \square \)
Proof of Lemma 5.4.1

Proof. First we can show that for \( \forall j, r \leq v \), \( H^D_{jr} = -\frac{1}{vn}1_{\tilde{n}}1_{\tilde{n}}^T + \frac{c}{vn}I_c \), where \( I_c \) is a \( c \times c \) block matrix with diagonal matrices all equal to \( 1_{\tilde{n}}1_{\tilde{n}}^T \). The matrix \(-\frac{1}{vn}1_{\tilde{n}}1_{\tilde{n}}^T\) contains exactly one non-zero eigenvalue, which equals to \(-\frac{1}{v}\). Also, \( \frac{c}{vn}I_c \) has \( c \) positive eigenvalues equal to \( \frac{1}{v} \). Hence, we have rank(\( H^D_{jr} \)) = \( c - 1 \), and all \( c - 1 \) non-zero eigenvalues are equal to \( \frac{1}{v} \). By the definition of spectral norm is the largest magnitude of the eigenvalues, we obtain
\[
\|H^D_{jr}\| = \frac{1}{v}, \quad \forall j, r.
\]

Similar analysis could be applied to \( H^S \). According to fundamental linear algebra theories on block matrices, rank(\( H^S_{jj} \)) = \( n \), with \( \frac{\hat{a}}{c} \) eigenvalues equal to \( \frac{v-1}{v} \) and the rest \( \frac{c}{c}n \) eigenvalues being 1. In addition, rank(\( H^S_{jr} \)) = \( c \), and all eigenvalues equal \(-\frac{1}{v}\). Consequently, we obtain
\[
\|H^S_{jj}\| = 1, \quad \|H^S_{jr}\| = \frac{1}{v}.
\]

Spectrum of large matrices. \( H^D \) is a \( v \times v \) block matrix with repeating blocks \( H^D_{jr} \). Hence, it admits the form of Kronecker product,
\[
H^D = 1_v 1_v^T \otimes H^D_{jr}.
\]

Consequently, the spectrum of \( H_D \) consist of \( c - 1 \) eigenvalues equal to \( \frac{1}{v} \cdot v = 1 \), and the rest all equal to 0. Therefore, \( H^D \) is positive semi-definite (i.e \( H^D \succeq 0 \)). Recall the notation
\( K = \text{diag}(K_1, K_2, ..., K_v) \), we have

\[
D = K^T H_D K \succeq 0,
\]

since for \( \forall x \in \mathbb{R}^n \), \( x^T K^T H_D K x = \bar{x}^T H_D \bar{x} \geq 0 \). Define \( H_{off} = H_{jr}^S \) for \( j \neq r \) as the off-diagonal block matrix of \( H^S \). We have

\[
H^S = 1_v 1_v^T \otimes H_{off} + \text{diag}_{v \times v}(I_{n \times n}).
\]

The eigenvalues of \( 1_v 1_v^T \otimes H_{off} \), by previous analysis, are -1 with multiplicity \( c \) and 0 with multiplicity \( \tilde{v}n - c \). By adding diagonal block matrix of identities, \( H^S \) has \( c \) eigenvalues of 0 and all others equal to 1. Therefore, \( S \) is also positive semi-definite.

**Proof of Lemma 5.4.2**

**Proof.** We denote \( F_{w_i} = [\dot{f}_{w_i}(x_1) \ldots \dot{f}_{w_i}(x_n)]^T \), and define random matrices \( Z_i = \frac{1}{m} (F_{w_i} F_{w_i}^T - K) \). By the unbiasedness of RFFs, we know that \( EZ_i = 0 \). To bound \( \|X_i\| \), we have

\[
\|Z_i\| = \frac{1}{m} \|(F_{w_i} F_{w_i}^T - K)\| \leq \frac{2n}{m},
\]
due to triangle inequality and boundedness of \( K \). In addition, we have

\[
EZ_i^2 = \frac{1}{m^2} E[(F_{w_i} F_{w_i}^T - K)^2]
\]

\[
\leq \frac{1}{m^2} E[n F_{w_i} F_{w_i}^T - 2 F_{w_i} F_{w_i}^T K + K^2] \leq \frac{nK}{m^2}.
\]

The second line is due to the fact that \( \|F_{w_i}^T F_{w_i}\|^2 \leq n \). Thus,

\[
\sigma^2 = \| \sum_{i=1}^m EZ_i^2 \| \leq m \| EZ_i^2 \| \leq \frac{n \|K\|}{m}.
\]
Applying matrix Bernstein inequality (Theorem 5.4.1 in Vershynin (2018a)),

\[
P\{\|\hat{K} - K\| \geq t\} \leq 2n \exp\left( -\frac{t^2/2}{n\|K\|/m + 2nt/3m}\right).
\]

Now taking the right-hand-side to be equal to \(\eta\), we derive a quadratic equation of \(t\). Solving for this equation gives us the desired bound.

**Proof of Theorem 5.4.3**

**Proof.** By Lemma 5.4.2, with probability \(1 - \eta\), we have for \(\forall i = 1, \ldots, v\),

\[
\|\hat{K}_i - K_i\| \leq \frac{2n \log(\frac{2n}{1-(1-\eta)^{1/v}})}{3vm} + \sqrt{\frac{4(4/n)^2(\log(\frac{2n}{1-(1-\eta)^{1/v}}))^2 + \frac{18}{v}mn\|K^*\| \log(\frac{2n/v}{1-(1-\eta)^{1/v}})}{3m}}.
\]

Denote this event \(\Omega\). In this event, we have

\[
\|D - \hat{D}\| = \|KH^D K - \hat{K}H^D \hat{K}\| = \|KH^D K - \hat{K}H^D K + \hat{K}H^D K - \hat{K}H^D \hat{K}\|
\]

\[
= \|(K - \hat{K})H^D K + \hat{K}H^D(K - \hat{K})\|
\]

\[
\leq \|K - \hat{K}\||H^D||K|| + \|\hat{K}\||H^D||K - \hat{K}||
\]

\[
= \|K - \hat{K}\|(\max_{i=1,\ldots,v} \|K_i\| + \max_{i=1,\ldots,v} \|\hat{K}_i\|),
\]

where we recall that \(K = \text{diag}(K_1, \ldots, K_v)\) and \(\hat{K} = \text{diag}(\hat{K}_1, \ldots, \hat{K}_v)\). The last line holds because \(\|H^D\| = 1\) and \(K, \hat{K}\) are both diagonal block matrix. Therefore,

\[
\|K\| = \|K^*\|, \quad \|\hat{K}\| = \|\hat{K}^*\|.
\]

It is easy to check that \(\|S - \hat{S}\| \leq \|K - \hat{K}\|(\max_{i=1,\ldots,v} \|K_i\| + \max_{i=1,\ldots,v} \|\hat{K}_i\|)\) analogously using same argument. Moreover, by sub-multiplicity of operator norms, we have
\[\sqrt{\|D^2 + (S + \epsilon I)^2\|} \leq \sqrt{\|D^2\| + \|(S + \epsilon I)^2\|} = \sqrt{\|KH D^* K\|^2 + \|KH S^* K\|^2} \leq \sqrt{2\|K\|^4 + (\|K\|^2)^2} \leq \sqrt{2(\|K\|^2)^2} = \sqrt{2(\|K^*\|^2)},\]

since \(\|H^D\| = \|H^S\| = 1\). Because \(Z_1\) is orthogonal, we have

\[\|(D - \hat{D})Z_1\| \leq \|D - \hat{D}\||Z_1\| = \|D - \hat{D}\|,\]

and same inequality holds for \(S\). Hence we have

\[\sqrt{\|(D - \hat{D})Z_1\|^2 + \|(S - \hat{S})Z_1\|^2} \leq \xi_\eta(\|K^*\| + \|\hat{K}^*\|).\]

Putting all parts together and using Theorem 2.1 from Sun (1983), we get the desired bound.

**Proof of Theorem 5.4.4**

**Lemma D.1** (Wedin (1973)). Let \(A \in \mathbb{C}^{m \times n}, B = A + E\). Then

\[\|B^\dagger - A^\dagger\| \leq \frac{1 + \sqrt{5}}{2} \max\{\|A^\dagger\|^2, \|B^\dagger\|^2\} \|E\|,\]

where \(A^\dagger, B^\dagger\) are the Moore–Penrose pseudoinverse of \(A\) and \(B\).

**Proof.** Since \((S + \epsilon I)\) is invertible, we may consider the SEP \((S + \epsilon I)^{-1} Dw = \lambda w\). We
have

\[
\|(S + \epsilon I)^{-1} D - (\hat{S} + \epsilon I)^{-1} \hat{D}\|
\]

\[
= \|[(S + \epsilon I)^{-1} - (\hat{S} + \epsilon I)^{-1}] D + (\hat{S} + \epsilon I)^{-1}(D - \hat{D})\|
\]

\[
\leq \|[(S + \epsilon I)^{-1} - (\hat{S} + \epsilon I)^{-1}] D\|
\]

\[
+ \|(\hat{S} + \epsilon I)^{-1}(D - \hat{D})\| \tag{i}
\]

\[
\leq C \frac{\|K^*\|^2 (\|K^*\| + \|\hat{K}^*\|) \xi_2}{\epsilon^2} + \frac{(\|K^*\| + \|\hat{K}^*\|) \xi_1}{\epsilon},
\]

where \(C = \frac{1 + \sqrt{5}}{2}\). Here (i) is induced by Lemma D.1. Since \((S + \epsilon I)\) is positive definite and symmetric, \((S + \epsilon I)^{-1}\) is also symmetric and positive definite. Given that \(D\) is symmetric and positive semi-definite, we know that \((S + \epsilon I)^{-1} D\) is similar to a symmetric PSD matrix,

\[
(S + \epsilon I)^{1/2}[(S + \epsilon I)^{-1} D](S + \epsilon I)^{-1/2}
\]

\[
= (S + \epsilon I)^{-1/2} D(S + \epsilon I)^{-1/2}.
\]

Hence, the eigenvalues of \((S + \epsilon I)^{-1} D\) are all real and non-negative. Therefore, the eigenvalues is equivalent to singular values. The proof is then complete using the classic sin\(\Theta\) Theorem from Wedin (1972).

\[\square\]

**Proof of Theorem 5.4.5**

To prove the Theorem 5.4.5, we first introduce some technical lemmas.

**Lemma D.2.** Suppose \(A \in \mathbb{R}^{p \times q}\). For \(U \in \mathbb{R}^{p \times m}\) and \(V \in \mathbb{R}^{q \times n}\) both with orthonormal columns, where \(m \leq p\) and \(n \leq q\). We have

\[
\|U^T AV\| \leq \|A\|.
\]
Proof. Let $U_1 \in \mathbb{R}^{p \times (p-m)}$ and $V_1 \in \mathbb{R}^{q \times (q-n)}$ such that $\tilde{U} = (U \ U_1)$ and $\tilde{V} = (V \ V_1)$ are orthonormal. By Courant-Fisher’s min-max representation, we have

$$
\Vert AV \Vert^2 = \max_{\Vert x \Vert^2 = 1, x \in \mathbb{R}^m} \Vert AVx \Vert^2
$$

$$
= \max_{\Vert x \Vert^2 = 1, x \in \mathbb{R}^m} x^T V^T A^T A V x
$$

$$
= \max_{\Vert x \Vert^2 = 1, x \in \mathbb{R}^m} [x_1 0]^T \tilde{V}^T A^T A \tilde{V} [x_1 0]
$$

$$
\leq \max_{x_1, x_2 \in \mathbb{R}^m} [x_1 x_2]^T \tilde{V}^T A^T A \tilde{V} [x_1 x_2]
$$

$$
= \Vert AV \Vert^2.
$$

Therefore we have $\Vert AV \Vert \leq \Vert A \tilde{V} \Vert$. Moreover, we can obtain $\Vert U^T A \Vert \leq \Vert \tilde{U}^T A \Vert$ by similar argument which in short reads as

$$
\Vert U^T A \Vert = \Vert A^T U \Vert \leq \Vert A^T \tilde{U} \Vert = \Vert \tilde{U}^T A \Vert.
$$

By the rotational invariance of operator norm, we have

$$
\Vert A \Vert = \Vert \tilde{U}^T A \tilde{V} \Vert \geq \Vert U^T A \Vert.
$$

\[\square\]

Lemma D.3 (Symmetric case). Let $\Sigma, \tilde{\Sigma} \in \mathbb{R}^{p \times p}$ be two symmetric matrices with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_p$, respectively. For $l \leq \text{rank}(\Sigma) < p$, assume that $\lambda_l - \lambda_{l+1} = \delta > 0$. Let $V = (v_1, \ldots, v_l) \in \mathbb{R}^{p \times l}$ and $\tilde{V} = (\tilde{v}_1, \ldots, \tilde{v}_l) \in \mathbb{R}^{p \times l}$ have orthonormal columns such that $Av_j = \lambda_j v_j$ and $\tilde{A}\tilde{v}_j = \tilde{\lambda}_j \tilde{v}_j$, for $j = 1, \ldots, l$. Then the following holds,

$$
\Vert \sin \Theta(V, \tilde{V}) \Vert \leq \frac{2\Vert \Sigma - \tilde{\Sigma} \Vert}{\delta}.
$$
Proof. Denote $\Lambda = \text{diag}(\lambda_1, ..., \lambda_l)$, and $\tilde{\Lambda} = \text{diag}(\tilde{\lambda}_1, ..., \tilde{\lambda}_l)$. Then we have

$$\tilde{\Sigma} \tilde{V} - \tilde{V} \tilde{\Lambda} = \Sigma \tilde{V} - \tilde{V} \Lambda + (\tilde{\Sigma} - \Sigma) \tilde{V} - \tilde{V}(\tilde{\Lambda} - \Lambda) = 0,$$

which follows that by triangle inequality and Lemma D.2,

$$\|\Sigma \tilde{V} - \tilde{V} \Lambda\| \leq \|(\tilde{\Sigma} - \Sigma)\tilde{V}\| + \|\tilde{V}(\tilde{\Lambda} - \Lambda)\| \leq \|\tilde{\Sigma} - \Sigma\| + \|\tilde{\Lambda} - \Lambda\|.$$

By the well-known Weyl’s inequality (Horn and Johnson, 2012), we have for $\forall i = 1, ..., p$,

$$\lambda_i + \mu_p \leq \tilde{\lambda}_i \leq \lambda_i + \mu_1,$$

where $\mu_1 \geq \cdots \geq \mu_p$ are eigenvalues of $\tilde{\Sigma} - \Sigma$. Therefore we have

$$\|\Sigma \tilde{V} - \tilde{V} \Lambda\| \leq \|\tilde{\Sigma} - \Sigma\| + \max_i |\tilde{\lambda}_i - \lambda_i| \leq \|\tilde{\Sigma} - \Sigma\| + \max\{|\mu_1|, |\mu_p|\} = 2\|\tilde{\Sigma} - \Sigma\|. \quad (D.1)$$

Define $\Lambda_1 = (\lambda_{l+1}, ..., \lambda_p)$, $V_1 = [v_{l+1}, ..., v_p]$. Let $G = [V \ V_1]$ and we have $\Sigma = G \begin{pmatrix} \Lambda & 0 \\ 0 & \Lambda_1 \end{pmatrix} G^T$. Using Lemma D.2 again yields

$$\|\Sigma \tilde{V} - \tilde{V} \Lambda\| = \|VAV^T \tilde{V} + V_1 \Lambda_1 V_1^T \tilde{V} - VV^T \tilde{V} \Lambda - V_1 V_1^T \tilde{V} \Lambda\| \geq \|V_1^T \tilde{V} \Lambda - \Lambda_1 V_1^T \tilde{V}\|, \quad (D.2)$$

where the last inequality is a derived by left multiplying $V_1^T$ and noticing that $V_1^T V = 0,
$V_1^TV_1 = I_{p-1}$ and $\|A\| = \|-A\|$. Let $\alpha = \frac{\lambda_1 + \lambda_2}{2}$, $r = \frac{\lambda_1 - \lambda_2}{2}$. We have

$$\|V_1^T\tilde{V}A - \Lambda_1 V_1^T\tilde{V}\| = \|V_1^T\tilde{V}(\Lambda - \alpha I) - (\Lambda_1 - \alpha I)V_1^T\tilde{V}\|$$

$$\geq \|(\Lambda_1 - \alpha I)V_1^T\tilde{V}\| - \|V_1^T\tilde{V}(\Lambda - \alpha I)\|,$$

where the last line is due to triangle inequality. Note that the eigenvalues of $(\Lambda - \alpha I)$ all lie in the range $[-r, r]$, while the largest eigenvalue of $(\Lambda_1 - \alpha I)$ is $-r - \delta$, with (by definition) $\delta = \lambda_l - \lambda_{l+1}$. Therefore, we have

$$\|V_1^T\tilde{V}\| = \|(\Lambda_1 - \alpha I)^{-1}(\Lambda_1 - \alpha I)V_1^T\tilde{V}\|$$

$$\leq \|(\Lambda_1 - \alpha I)^{-1}\| \cdot \|(\Lambda_1 - \alpha I)V_1^T\tilde{V}\|,$$

and thus

$$\|(\Lambda_1 - \alpha I)V_1^T\tilde{V}\| \geq \frac{1}{\|(\Lambda_1 - \alpha I)^{-1}\|} \|V_1^T\tilde{V}\|$$

$$\geq (r + \delta)\|V_1^T\tilde{V}\|.$$

On the other hand, we have

$$\|V_1^T\tilde{V}(\Lambda - \alpha I)\| \leq r\|V_1^T\tilde{V}\|.$$

Consequently, we obtain the bound

$$\|V_1^T\tilde{V}\Lambda - \Lambda_1 V_1^T\tilde{V}\| \geq (r + \delta)\|V_1^T\tilde{V}\| - r\|V_1^T\tilde{V}\|$$

$$= \delta\|V_1^T\tilde{V}\|. $$
We notice that

\[ \| V_1^T \tilde{V} \| \geq \sigma_l(V_1^T \tilde{V}) \]

\[ = \sqrt{\sigma_l(\tilde{V}^T V_1^T \tilde{V})} \]

\[ \overset{(i)}{=} \sqrt{\sigma_l(\tilde{V}^T (I - V V^T) \tilde{V})} \]

\[ = \sqrt{\sigma_l(I - \tilde{V}^T V V^T \tilde{V})} \]

\[ = \sqrt{1 - \sigma_l^2(V \tilde{V})} \]

\[ \overset{(ii)}{=} \| \sin \Theta(V, \tilde{V}) \|, \quad (D.3) \]

where (i) is because the projection operator \( V V^T = I - V_1 V_1^T \), and (ii) is because by (5.5) and Lemma D.2, it holds that

\[ 1 - \sigma_l^2(V \tilde{V}) = 1 - \min_{1 \leq i \leq l} \cos^2(\theta) = \max_{1 \leq i \leq l} \sin^2(\theta). \]

Finally, we can deduce from (D.1), (D.2) and (D.3) that

\[ \| \sin \Theta(V, \tilde{V}) \| \leq \frac{2\| \Sigma - \tilde{\Sigma} \|}{\delta}, \]

which completes the proof.

**Proof.** (of Theorem 5.4.5) We know that \( A^T A, \tilde{A}^T \tilde{A} \in \mathbb{R}^{q \times q} \) are symmetric with eigenvalues equal to \( \sigma_1^2 \geq \cdots \geq 1 \) and \( \tilde{\sigma}_1^2 \geq \cdots \geq \tilde{\sigma}_1 \), respectively. Recalling the donation of SVD as \( A = U \Sigma V^T \), we know that

\[ A^T A = V \Sigma^T U^T U \Sigma V^T = V \Sigma^2 V^T. \]

Hence, the right singular vectors of \( A \) are the same as the eigenvectors of \( A^T A \). Adopting
Lemma D.3, we obtain

$$\| \sin \Theta(V, \tilde{V}) \| \leq \frac{2 \| A^T A - \tilde{A}^T \tilde{A} \|}{\sigma_i^2 - \sigma_{i-1}^2}.$$ 

Regarding the numerator, we have

$$\| A^T A - \tilde{A}^T \tilde{A} \| = \| A^T (A - \tilde{A}) + (A^T - \tilde{A}^T) \tilde{A} \|$$

$$\leq (\sigma_1 + \tilde{\sigma}_1) \| A - \tilde{A} \|.$$ 

This completes the proof. \qed
APPENDIX E
PROOFS OF CHAPTER 6

Proof of Lemma 6.1.1 & Theorem 6.1.2

Proof. (of Lemma 6.1.1) We have the convolution of uniform and Gaussian distribution as

\[ f_Y(y) = \int_{-\infty}^{\infty} P(b = u, \gamma X = y - u) du \]

\[ = \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{\sqrt{2\pi}\gamma} e^{-\frac{(y-u)^2}{2\gamma^2}} du \]

\[ = \frac{1}{2\pi} \left[ \Phi\left(\frac{2\pi - y}{\gamma}\right) - \Phi\left(-\frac{y}{\gamma}\right) \right]. \]

\[ \square \]

Proof. (of Theorem 6.1.2) Denote \( Y = \gamma X + \tau \). We have

\[ P(Z \leq z) = \sum_{k=-\infty}^{\infty} P\left(2k\pi + \cos^{-1} z \leq Y \leq 2(k+1)\pi - \cos^{-1} z\right) \]

\[ = \sum_{k=-\infty}^{\infty} \int_{2k\pi + \cos^{-1} z}^{2(k+1)\pi - \cos^{-1} z} f_Y(y) dy, \]

where \( f(y) \) is given by Lemma 6.1.1. Let the density of \( Z \) be \( g_Z \), and denote \( t^* = \cos^{-1} z \).
It follows that

\[
g_Z(z) = \sum_{k=-\infty}^{\infty} \frac{1}{\sqrt{1-z^2}} \left[ f_Y(2(k+1)\pi - t^*) + f_Y(2k\pi + t^*) \right]
\]

\[
= \frac{1}{2\pi \sqrt{1-z^2}} \sum_{k=-\infty}^{\infty} \left[ \Phi\left( \frac{t^* - 2k\pi}{\gamma} \right) - \Phi\left( \frac{t^* - 2(k+1)\pi}{\gamma} \right) 
- \Phi\left( \frac{-t^* - 2(k-1)\pi}{\gamma} \right) - \Phi\left( \frac{-t^* - 2k\pi}{\gamma} \right) \right]
\]

\[
= \frac{1}{\pi \sqrt{1-z^2}}.
\]  

(E.1)

To prove the last line, denote the term in the bracket as \( \alpha_k \). By cancellation, for any \( k_1, k_2 \), we have

\[
\sum_{k=k_1}^{k_2} \alpha_k = \left[ \Phi\left( \frac{t^* - 2k_1\pi}{\gamma} \right) + \Phi\left( \frac{-t^* - 2(k_1-1)\pi}{\gamma} \right) 
- \Phi\left( \frac{t^* - 2(k_2+1)\pi}{\gamma} \right) - \Phi\left( \frac{-t^* - 2k_2\pi}{\gamma} \right) \right],
\]

which equals to 2 in the limit \( k_1 \to -\infty, k_2 \to \infty \). Using a similar approach, we can show that Eq. (E.1) is exactly the density of the cosine of a uniform random variable on \([0, 2\pi]\).

For \( Z_2 = \cos(\gamma X + \tau) = Z^2 \), we have

\[
P[Z_2 \leq z] = P[|Z| \leq \sqrt{z}]
\]

\[
= \frac{1}{\pi} \int_{\sqrt{z}}^{\sqrt{z}} \frac{1}{\sqrt{1-z^2}} dz
\]

\[
= \frac{1}{\pi} \left[ \sin^{-1}(\sqrt{z}) - \sin^{-1}(\sqrt{z}) \right]
\]

\[
= \frac{2}{\pi} \sin^{-1}(\sqrt{z}).
\]

Taking the derivative we get the p.d.f. as

\[
f_{Z_2}(z) = \frac{1}{\pi \sqrt{z - z^2}}.
\]
The proof is now complete.

**Proof of Lemma 6.1.3 & Theorem 6.1.4**

**Proof.** (of Lemma 6.1.3) Similar to the proof of Lemma 6.1.1, we have

\[
 f(t_x, t_y) = \frac{1}{2\pi} \int_0^{2\pi} P(\gamma x = t_x - u, \gamma y = t_y - u) du 
\]

\[
 = \frac{1}{4\pi^2 \gamma^2 \sqrt{1 - \rho^2}} \int_0^{2\pi} e^{-\frac{(t_x-u)^2 - 2\rho(t_x-u)(t_y-u) + (t_y-u)^2}{2(1-\rho^2)\gamma^2}} du 
\]

\[
 = \frac{1}{4\pi^2 \gamma^2 \sqrt{1 - \rho^2}} \int_0^{2\pi} e^{-\frac{2(1-\rho)(u^2 - u(t_x + t_y)) + t_x^2 + t_y^2 - 2\rho t_x t_y}{2(1-\rho^2)\gamma^2}} du 
\]

\[
 = \frac{1}{4\pi^2 \gamma^2 \sqrt{1 - \rho^2}} \int_0^{2\pi} e^{-\frac{2(1-\rho)(u - t_x + t_y)^2 + 1 + \rho (t_x - t_y)^2}{2(1-\rho^2)\gamma^2}} du 
\]

\[
 = \frac{1}{2\pi} \phi \sqrt{2(1-\rho)\gamma} (t_x - t_y) \left[ \Phi\left(\frac{4\pi - (t_x + t_y)}{\gamma \sqrt{2(1+\rho)}}\right) - \Phi\left(-\frac{t_x + t_y}{\gamma \sqrt{2(1+\rho)}}\right) \right], 
\]

where \( \phi \sqrt{2(1-\rho)\gamma} \) is the density of \( N(0, 2(1-\rho)\gamma^2) \).

**Proof.** (of Theorem 6.1.4) Denote \( Z_x = \cos(t_x), Z_y = \cos(t_y) \). Let \( a_x^* = \cos^{-1}(z_x), a_y^* = \cos^{-1}(z_y) \). Denote \( \phi = \phi \sqrt{2(1-\rho)\gamma} \) for simplicity. We have

\[
 P(Z_x \leq z_x, Z_y \leq z_y) = \sum_{k_x = -\infty}^{\infty} \sum_{k_y = -\infty}^{\infty} \int_{2k_x \pi + a_x^*}^{2(k_x+1)\pi - a_x^*} \int_{2k_y \pi + a_y^*}^{2(k_y+1)\pi - a_y^*} f(t_x, t_y) dt_x dt_y. 
\]
By Lemma 6.1.3, it follows that

\[ f(z_x, z_y) \]

\[
= \frac{1}{2\pi} \sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} \int_{2k_x\pi + a_x^*}^{2(k_x+1)\pi - a_x^*} \frac{1}{\sqrt{1 - z_y^2}} \left\{ \phi(t_x - 2(k_y + 1)\pi + a_y^*) \right. \\
\left. - \Phi \left( \frac{4\pi - (t_x + 2(k_y + 1)\pi - a_y^*)}{\gamma \sqrt{2(1 + \rho)}} \right) - \Phi \left( \frac{t_x + 2(k_y + 1)\pi - a_y^*}{\gamma \sqrt{2(1 + \rho)}} \right) + \phi(t_x - 2k_y\pi - a_y^*) \right\} \; dt_x \\
= \frac{1}{2\pi \sqrt{1 - z_x^2} \sqrt{1 - z_y^2}} \sum_{k_x} \sum_{k_y} \left\{ \phi(-a_x^* + a_y^* + 2(k_x - k_y)\pi) \frac{\Phi(\frac{a_x^* + a_y^* - 2(k_x + k_y)\pi}{\gamma \sqrt{2(1 + \rho)}})}{\gamma \sqrt{2(1 + \rho)}} \\
- \Phi \left( \frac{a_x^* + a_y^* - 2(k_x + k_y + 2)\pi}{\gamma \sqrt{2(1 + \rho)}} \right) \right\} + \phi(-a_x^* - a_y^* + 2(k_x - k_y + 1)\pi) \\
\left. \frac{\Phi(\frac{a_x^* - a_y^* - 2(k_x + k_y - 1)\pi}{\gamma \sqrt{2(1 + \rho)}})}{\gamma \sqrt{2(1 + \rho)}} \right] \}
\]

\[
= \frac{1}{2\pi \sqrt{1 - z_x^2} \sqrt{1 - z_y^2}} \sum_{k_x} \sum_{k_y} \left\{ \phi(-a_x^* + a_y^* + 2k\pi) + \phi(-a_x^* - a_y^* + 2k\pi) \\
+ \phi(a_x^* + a_y^* + 2k\pi) + \phi(a_x^* - a_y^* + 2k\pi) \right\} \\
= \frac{1}{\pi \sqrt{1 - z_x^2} \sqrt{1 - z_y^2}} \sum_{k_x} \sum_{k_y} \left\{ \phi(a_x^* - a_y^* + 2k\pi) + \phi(a_x^* + a_y^* + 2k\pi) \right\},
\]

where (a) is derived by writing the summations \( \sum_{k_x=-\infty}^{\infty} \sum_{k_y=-\infty}^{\infty} \{ \cdot \} \) into \( \sum_{l=-\infty}^{\infty} \sum_{k_x=-\infty}^{\infty} \{ \cdot \} \) with \( l = k_x - k_y \) and canceling terms, along with the symmetry of \( \phi(\cdot) \). This gives the joint density of \( z_x \) and \( z_y \).

For the sine counterpart, with some abuse of notation, let us denote \( z_x = \sin(t_x) \) and
\( z_y = \sin(t_y) \) from now on. Using similar argument, we have

\[
P(Z_x \leq z_x, Z_y \leq z_y) = \sum_{k_x = -\infty}^{\infty} \sum_{k_y = -\infty}^{\infty} \int_{(2k_x+1)\pi - \sin^{-1}(z_x)}^{2(k_x+1)\pi + \sin^{-1}(z_x)} \int_{(2k_y+1)\pi - \sin^{-1}(z_y)}^{2(k_y+1)\pi + \sin^{-1}(z_y)} f(t_x, t_y) dt_x dt_y.
\]

After simplification, we finally arrive at

\[
f(z_x, z_y) = \frac{1}{\pi \sqrt{1 - z_x^2} \sqrt{1 - z_y^2}} \sum_{k = -\infty}^{\infty} \left[ \phi(\sin^{-1}(z_x) - \sin^{-1}(z_y) + 2k\pi) 
+ \phi(\sin^{-1}(z_x) + \sin^{-1}(z_y) + (2k + 1)\pi) \right]. \quad (E.2)
\]

Considering \( Z_x = \sin(t_x), Z_y = \sin(t_y) \). Since \( \sin^{-1}(x) = \frac{\pi}{2} - \cos^{-1}(x) \), we can substitute into the density to derive

\[
f(z_x, z_y) = \frac{1}{\pi \sqrt{1 - z_x^2} \sqrt{1 - z_y^2}} \sum_{k = -\infty}^{\infty} \left[ \phi(\cos^{-1}(z_x) - \cos^{-1}(z_y) + 2k\pi) 
+ \phi(\cos^{-1}(z_x) + \cos^{-1}(z_y) + (2k + 2)\pi) \right]
= \frac{1}{\pi \sqrt{1 - z_x^2} \sqrt{1 - z_y^2}} \sum_{k = -\infty}^{\infty} \left[ \phi(a_x^* - a_y^* + 2k\pi) + \phi(a_x^* + a_y^* + 2k\pi) \right],
\]

which is the same as the previous cosine transformation. This completes the proof. \( \Box \)

**Proof of Proposition 6.1.5**

*Proof.* Let us denote \( \sigma = \sqrt{2(1 - \rho)\gamma} \) for simplicity. By symmetry and exchangeability of \( f \), to prove the desired result, it suffices to consider the case where both \( z_x \) and \( z_y \) are positive, i.e., \((z_x, z_y) \in (0, 1]^2\). Define the notation \( a_x^* = \sin^{-1}(z_x) \geq 0, a_y^* = \sin^{-1}(z_y) \geq 0 \). From
(E.2), we deduce

\[ f(z_x, z_y) - f(z_x, -z_y) \]

\[ \propto \sum_{k=-\infty}^{\infty} \left[ \phi_\sigma(a_x^* - a_y^* + 2k\pi) + \phi_\sigma(a_x^* + a_y^* + (2k + 1)\pi) \\
- \phi_\sigma(a_x^* + a_y^* + 2k\pi) - \phi_\sigma(a_x^* - a_y^* + (2k + 1)\pi) \right] \]

\[ = \sum_{k=0}^{\infty} (-1)^k \left[ \phi_\sigma(k\pi + d) - \phi_\sigma(k\pi + s) + \phi_\sigma((k + 1)\pi - s) - \phi_\sigma((k + 1)\pi - d) \right], \]

\[ = \phi_\sigma(d) - \phi_\sigma(s) + \sum_{k=1}^{\infty} \left[ \phi_\sigma(k\pi - s) - \phi_\sigma(k\pi - d) - \phi_\sigma(k\pi + d) + \phi_\sigma(k\pi + s) \right], \]

\[ \triangleq \phi_\sigma(d) - \phi_\sigma(s) + \sum_{k=1}^{\infty} M_k, \quad \text{(E.3)} \]

where we let \( d = a_x^* - a_y^* \) and \( d = a_x^* + a_y^* \), and we use the fact that \( \phi_\sigma(-x) = \phi_\sigma(x) \).

Note that, we consider \( z_y > 0 \) so that \( d \neq s \), since when \( z_y = 0 \) we trivially have \( f(z_x, 0) = f(z_x, 0) \). For now, we assume that \( z_x \geq z_y > 0 \), such that \( d \) and \( s \) are defined on the domain \( 0 < s \leq \pi \) and \( 0 \leq d < \min\{s, \pi - s\} \). Since

\[ \phi'_\sigma(x) = -\frac{x}{\sqrt{2\pi\sigma^3}} e^{-\frac{x^2}{2\sigma^2}}, \quad \phi''_\sigma(x) = -\frac{x^2 - \sigma^2}{\sqrt{2\pi\sigma^5}} e^{-\frac{x^2}{2\sigma^2}}, \]

we know that \( \phi_\sigma \) is piecewise concave on \((0, \sigma)\) and piecewise convex on \((\sigma, \infty)\). Thus,

\[ \phi_\sigma(a) - \phi_\sigma(a + g) \geq \phi_\sigma(c) - \phi_\sigma(c + g) \quad \text{(E.4)} \]

for any \( \sigma \leq a \leq c \) and \( g \geq 0 \). The equality holds only when \( a = c \) or \( g = 0 \). Consequently, under the assumption that \( \sigma \leq \pi \), \( M_k \geq 0 \) for \( k \geq 2 \) since \( 2\pi - s \geq \sigma \), where the equality holds only when \( d = s \), i.e., \( z_y = 0 \). Furthermore, the piecewise convexity of \( \phi_\sigma(\cdot) \) and (E.4) imply that for \( \sigma \leq a < c \),

\[ \frac{(c - a)c}{\sigma^2} e^{-\frac{c^2}{2\sigma^2}} < e^{-\frac{a^2}{2\sigma^2}} - e^{-\frac{c^2}{2\sigma^2}} < \frac{(c - a)a}{\sigma^2} e^{-\frac{a^2}{2\sigma^2}}. \quad \text{(E.5)} \]
Also note that the function $e^{-x}$ is convex on the real line, which gives for $\forall a < c$,

$$\frac{(c - a)(c + a)}{2\sigma^2} e^{-\frac{c^2}{2\sigma^2}} < e^{-\frac{a^2}{2\sigma^2}} - e^{-\frac{c^2}{2\sigma^2}} < \frac{(c - a)(c + a)}{2\sigma^2} e^{-\frac{a^2}{2\sigma^2}}. \quad (E.6)$$

Now that $M_k > 0$ for $k \geq 2$, evaluating (E.3) we obtain

$$(E.3) > \phi_\sigma(d) - \phi_\sigma(s) + M_1
\geq a \frac{1}{\sqrt{2\pi\sigma^3}} \left[ \frac{(s - d)(s + d)}{2} e^{-\frac{s^2}{2\sigma^2}} + \frac{(s - d)(2\pi - s - d)}{2} e^{-\frac{(s-d)^2}{2\sigma^2}} - (s - d)(\pi + d) e^{-\frac{(\pi+d)^2}{2\sigma^2}} \right]
\geq b \frac{s - d}{\sqrt{2\pi\sigma^3}} \left[ \pi e^{-\frac{(s-d)^2}{2\sigma^2}} - (\pi + d) e^{-\frac{(x+d)^2}{2\sigma^2}} \right],$$

where (a) uses (E.5) and (E.6), and (b) is because $s \leq \pi - d$. It is easy to verify that the ratio

$$\left( \pi e^{-\frac{(s-d)^2}{2\sigma^2}} \right) \left( \frac{\pi + d}{2\pi \sigma^2} \right) \geq 1$$

for $\sigma \leq \pi$ and $0 \leq d < \min\{s, \pi - s\} < \frac{\pi}{2}$. Therefore, we have proved that $f(z_x, z_y) > f(z_x, -z_y)$, for $z_x \geq z_y > 0$. Now, by exchangeability and symmetry of $f$, we have

$$f(z_y, z_x) = f(z_x, z_y) > f(z_x, -z_y) = f(-z_x, z_y) = f(z_y, -z_x).$$

Therefore, our result also holds for $z_y \geq z_x > 0$. The proof is now complete. \qed
Proof of Theorem 6.3.1

Proof. Denote the StocQ quantizer as $Q$. For each RFF $z$, assume $z \in [t_{i-1}, t_i]$ for some $i$. We can then write $Q(z) = z + \epsilon$, where

$$E[\epsilon] = \frac{z - t_{i-1}}{t_i - t_{i-1}} + \frac{t_i - z}{t_i - t_{i-1}} - z = 0.$$

Thus, it follows that

$$Var[\epsilon] = E[\epsilon^2] = \frac{t_i^2 z - t_{i-1}^2}{t_i - t_{i-1}} + \frac{t_{i-1}^2 z}{t_i - t_{i-1}} - z^2 = (t_i - z)(z - t_{i-1}).$$

For two data vectors $u, v$, let $F_{\text{StocQ}}(u) = \sqrt{2}Q(z_u)$ and $F_{\text{StocQ}}(v) = \sqrt{2}Q(z_v)$, where $z_u = \cos(w^T u + \tau)$ and $z_v = \cos(w^T v + \tau)$ follows the distribution $f$ given by Theorem 6.1.4. We can write $Q(z_u) = z_u + \epsilon_u, Q(z_v) = z_v + \epsilon_v$ where $\epsilon_u$ and $\epsilon_v$ are independent. Let

$$\hat{k} = F_{\text{StocQ}}(u)F_{\text{StocQ}}(v).$$

We have

$$E[\hat{k}_{\text{StocQ}}] = 2E[(z_u + \epsilon_u)(z_v + \epsilon_v)] = 2E[z_u z_v] = K(u, v),$$

implying that StocQ estimate is unbiased. The variance factor can be computed as

$$Var[\hat{k}_{\text{StocQ}}] = 4E[(z_u + \epsilon_u)^2(z_v + \epsilon_v)^2] - K(u, v)^2 = 4E[z_u^2 \epsilon_v^2 + z_v^2 \epsilon_u^2 + \epsilon_u^2 \epsilon_v^2] + Var[\hat{k}] \triangleq A + Var[\hat{k}],$$

(E.7)

where $Var[\hat{k}]$ is the variance of full-precision RFF kernel estimator. Obviously, $A > 0$, thus StocQ estimator always has larger variance than full-precision RFF.
analysis,

\[
\mathbb{E}[z_u^2 \epsilon_v^2] = \mathbb{E}_{z_u,z_v} \mathbb{E}[\epsilon_u^2 | z_v]
\]

\[
= \int_{-1}^{1} dz_u \left( \sum_{j=1}^{2^{b-1} - 1} \int_{t_{j-1}}^{t_j} (t_j - z_v)(z_v - t_{j-1})z_u^2 f(z_u, z_v) dz_v \right)
\]

\[
= \sum_{i=1}^{2^{b-1} - 1} \sum_{j=1}^{2^{b-1} - 1} \int_{t_{j-1}}^{t_j} \int_{t_{i-1}}^{t_i} \left( (t_i - t_{i-1})(t_j - t_{i-1})z_u z_v \right.
\]

\[
- (t_{j-1} + t_j)z_u z_v^2 + z_u^2 z_v^2 - (t_i - t_{i-1})t_j z_u - (t_{j-1} + t_j)t_{i-1} t_i z_v
\]

\[
+ t_{j-1} t_j z_u^2 + t_{i-1} t_i z_v^2 + t_{i-1} t_i t_j z_u z_v \left) f(z_u, z_v) dz_u dz_v \right.
\]

\[
= \sum_{i=1}^{2^{b-1} - 1} \sum_{j=1}^{2^{b-1} - 1} \int_{t_{j-1}}^{t_j} \int_{t_{i-1}}^{t_i} \left( (t_i - t_{i-1})(t_j - t_{i-1})z_u z_v - 2(t_{j-1} + t_j)z_u z_v^2
\]

\[
+ z_u^2 z_v^2 + 2 t_j t_j z_u z_v + t_{i-1} t_i t_j z_u z_v \right) f(z_u, z_v) dz_u dz_v,
\]

where equation \( (a) \) is due to the symmetry of density \( f \) and the borders \( t_0 < \ldots < t_{2^b-1} \).

Substituting above expressions into (E.7) and cancelling terms, we obtain

\[
A = 4 \sum_{i=1}^{2^{b-1} - 1} \sum_{j=1}^{2^{b-1} - 1} \int_{t_{j-1}}^{t_j} \int_{t_{i-1}}^{t_i} \left( (t_i - t_{i-1})(t_j - t_{i-1})z_u z_v + t_{i-1} t_i t_{j-1} t_j - z_u^2 z_v^2 \right) f(z_u, z_v) dz_u dz_v
\]

\[
= 4 \sum_{i=1}^{2^{b-1} - 1} \sum_{j=1}^{2^{b-1} - 1} \left[ (t_{i-1} + t_i)(t_{j-1} + t_j) \kappa_{i,j} + t_{i-1} t_i t_{j-1} t_j p_{i,j} \right] - 4 \mathbb{E}[z_u^2 z_v^2].
\]
Therefore,

\[ \text{Var}[\hat{k}_{\text{StocQ}}] = 4 \sum_{i=1}^{2^b-1} \sum_{j=1}^{2^b-1} \left[ (t_{i-1} + t_i)(t_{j-1} + t_j)\kappa_{i,j} + t_{i-1}t_it_{j-1}t_jp_{i,j} \right] - K(u, v)^2. \]

The proof is completed by noting that StocQ estimator is the average of i.i.d. Bernoulli random variables.

**Proof of Theorem 6.3.2**

**Proof.** For simplicity, we prove the result specifically for LM-RFF quantization. Similar arguments holds for general quantizers. The Chebyshev polynomials (Borwein and Erdélyi, 1995) of the first kind are defined through trigonometric identities

\[ T_n(\cos(x)) = \cos(n \cos(x)), \]

where admit the following recursion,

\[ T_0(x) = 1, \quad T_1(x) = x, \]

\[ T_{i+1}(x) = 2xT_i(x) - T_{i-1}(x), \quad i \geq 2. \]

\{T_0, T_1, \ldots\} forms an orthogonal basis of the function space on \([-1, 1]\) with finite number of discontinuities. Precisely, define the inner product w.r.t. measure \(\frac{1}{\sqrt{1-x^2}}\) as

\[ \langle f(x), g(x) \rangle = \int_{-1}^{1} f(x)g(x) \frac{1}{\sqrt{1-x^2}} dx. \]
Then orthogonality holds:

\[
\int_{-1}^{1} T_i(x)T_j(x)\frac{1}{\sqrt{1-x^2}}dx = \begin{cases} 
0, & i \neq j, \\
\pi, & i = j = 0, \\
\frac{\pi}{2}, & i \neq j.
\end{cases}
\]

By Chebyshev functional decomposition, our LM quantizer can be written as

\[
Q(x) = \sum_{k=0}^{\infty} \alpha_k T_k(x),
\]

where \(\alpha_k\) are computed through the inner products,

\[
\alpha_0 = \frac{2}{\pi} \int_{-1}^{1} Q(x)T_0(x)\frac{dx}{\sqrt{1-x^2}} = 0, \\
\alpha_1 = \frac{2}{\pi} \int_{-1}^{1} Q(x)T_1(x)\frac{dx}{\sqrt{1-x^2}} = 1 - 2D, \\
\alpha_2 = \frac{2}{\pi} \int_{-1}^{1} Q(x)T_2(x)\frac{dx}{\sqrt{1-x^2}} = 0, \\
\alpha_3 = \frac{2}{\pi} \int_{-1}^{1} Q(x)T_3(x)\frac{dx}{\sqrt{1-x^2}}, \\
\ldots
\]

with \(D\) the distortion of \(Q\) given in equation (7) of the main paper. Firstly, it is easy to show that \(|E[T_i(z_x)T_j(z_y)]| \leq E[T_i(z_x)^2] = \frac{1}{2}\). Note that \(\alpha_k = 0\) when \(k\) is even because \(T_k(x)\) is even function and \(Q(x)\) is odd. Recall \(u, v\) are two normalized data vectors with correlation \(\rho\). Denote \(z_x = \cos(\gamma x + \tau)\) and \(z_y = \cos(\gamma y + \tau)\) with distribution \(f(z_x, z_y)\),
where \((x, y) \sim N(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})\), \(\tau \sim \text{uniform}(0, 2\pi)\). It follows that

\[
\mathbb{E}[\sqrt{2}Q(z_x)\sqrt{2}Q(z_y)] = 2 \int_{-1}^{1} \int_{-1}^{1} Q(z_x)Q(z_y)f(z_x, z_y)dz_xdz_y
\]

\[
= 2 \int_{-1}^{1} \int_{-1}^{1} (\sum_{i=1, \text{odd}}^{\infty} \alpha_i T_i(z_x)) (\sum_{j=1, \text{odd}}^{\infty} \alpha_j T_j(z_y)) f(z_x, z_y)dz_xdz_y
\]

\[
= (1 - 2D)^2 K(u, v) + 2 \sum_{i=1, \text{odd}}^{\infty} \sum_{j=3, \text{odd}}^{\infty} \alpha_i \alpha_j \int_{-1}^{1} \int_{-1}^{1} T_i(z_x)T_j(z_y) f(z_x, z_y)dz_xdz_y. \quad (E.8)
\]

This proves the first part. There is an intrinsic constraint on \(\alpha_i, i = 3, 5, \ldots\). First, we can compute the cosine of \(Q(x)\) and each \(T_i(x)\) as

\[
c_i = \frac{\int_{-1}^{1} Q(x)T_i(x) \frac{dx}{\sqrt{1-x^2}}}{\sqrt{\int_{-1}^{1} Q(x)^2 \frac{dx}{\sqrt{1-x^2}} \int_{-1}^{1} T_i(x)^2 \frac{dx}{\sqrt{1-x^2}}}}
\]

\[
= \frac{\pi \alpha_i}{\sqrt{(\frac{\pi}{2} - D)\pi \frac{\pi}{2}}}
\]

\[
= \frac{\alpha_i}{\sqrt{1 - 2D}}.
\]

Since the Chebyshev polynomials form an orthogonal basis of function space on \([-1, 1]\), it holds that \(\sum_{i=0}^{\infty} c_i^2 = 1\). Therefore, we have \(\sum_{i=0}^{\infty} \alpha_i^2 = 1 - 2D\). Now that \(\alpha_i = 0\) when \(i\) is even, and \(\alpha_1 = 1 - 2D\), we then have \(\sum_{i=3, \text{odd}}^{\infty} \alpha_i^2 = 1 - 2D - (1 - 2D)^2 = 2D(1 - 2D)\).

When \(\rho = 0\), from (E.8), it is easy to see that all the integrals would be zero by independence. Thus, the estimated kernel \(\mathbb{E}[\sqrt{2}Q(z_x)\sqrt{2}Q(z_y)] = (1 - 2D)^2 K(u, v)\).

When \(\rho = 1\) \((K(u, v) = 1)\), we have \(\int_{-1}^{1} T_i(z_x)T_j(z_x) f(z_x)dz_x = 0\) for \(i \neq j\) by orthogonality of Chebyshev polynomials, where \(f(z_x)\) is the marginal distribution of \(z_x\). It
follows that

\[
\mathbb{E}[\sqrt{2}Q(x)\sqrt{2}Q(y)] = (1 - 2D)^2 + \sum_{i=3,\text{odd}}^{\infty} \alpha_i^2
\]

\[
= (1 - 2D)^2 + 2D(1 - 2D)
\]

\[
= 1 - 2D.
\]

This completes the proof of the theorem. \(\square\)

**Proof of Theorem 6.3.3**

*Proof.* Denote \(w = \cos(\gamma x + \tau), z = \cos(\gamma y + \tau)\), with \((x, y)\) are random vectors with i.i.d. entries from \(N(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})\), and \(\tau \sim \text{uniform}(0, 2\pi)\) is also a vector with i.i.d. entries. Recall the notation \(\zeta_{s,t} = \mathbb{E}[Q(w_1)^sQ(z_1)^t]\), where \(Q\) is our LM-RFF quantizer. By Taylor expansion at the expectations, we have as \(m \to \infty\),

\[
\mathbb{E}[k_{m,Q}] = \frac{\mathbb{E}\left[ \frac{1}{m} \sum_{i=1}^{m} Q(w_i)Q(z_i) \right]}{\mathbb{E}\left[ \sqrt{\frac{1}{m^2}\|Q(w)\|^2\|Q(z)\|^2} \right]} + \mathcal{O}\left(\frac{1}{m}\right)
\]

\[
\triangleq \frac{\zeta_{1,1}}{\mathbb{E}[\sqrt{\Lambda}]} + \mathcal{O}\left(\frac{1}{m}\right).
\]

Applying Taylor expansion again,

\[
\mathbb{E}[\sqrt{\Lambda}] = \mathbb{E}\left[ \sqrt{\mathbb{E}[\Lambda] + \frac{\Lambda - \mathbb{E}[\Lambda]}{2\sqrt{\mathbb{E}[\Lambda]}} + \mathcal{O}((\Lambda - \mathbb{E}[\Lambda])^2)} \right]
\]

\[
= \mathbb{E}[\Lambda] + \mathcal{O}\left(\frac{1}{m}\right), \quad m \to \infty.
\]
Furthermore, we have the expectation of $\Lambda$ as

\[
\mathbb{E}[\Lambda] = \frac{1}{m^2} \mathbb{E} \left[ \left( \sum_{i=1}^{m} Q(w_i)^2 \right) \left( \sum_{i=1}^{m} Q(z_i)^2 \right) \right]
= \frac{1}{m^2} \left[ \sum_{i \neq j} Q(w_i)^2 Q(z_j)^2 + \sum_{i=1}^{m} Q(w_i)^2 Q(z_i)^2 \right]
= \frac{m-1}{m} \mathbb{E}[Q(w_1)^2 Q(z_2)^2] + \frac{1}{m} \mathbb{E}[Q(w_1)^2 Q(z_1)^2]
= \xi_{2,0}, \quad m \to \infty.
\]

Consequently, we obtain

\[
\mathbb{E}[\hat{k}_{n,Q}] = \frac{\xi_{1,1}}{\xi_{2,0}}, \quad m \to \infty.
\]

This completes the proof for asymptotic mean. With a little abuse of notation, let $\hat{k}_{n,Q} = \frac{a}{\sqrt{bc}}$, with

\[
a = \frac{\langle Q(w), Q(z) \rangle}{k}, \quad b = \frac{\|Q(w)\|^2}{k}, \quad c = \frac{\|Q(z)\|^2}{k}.
\]

We have

\[
\mathbb{E}[a] = \xi_{1,1}, \quad Var[a] = \frac{\xi_{2,2} - \xi_{1,1}^2}{m},
\]
\[
\mathbb{E}[b] = \xi_{2,0} = \mathbb{E}[c], \quad Var[b] = \frac{\xi_{4,0} - \xi_{2,0}^2}{m} = Var[c],
\]
\[
Cov(a, b) = \frac{1}{m^2} \mathbb{E} \left[ \left( \sum_{i=1}^{m} Q(w_i)Q(z_i) \right) \left( \sum_{i=1}^{m} Q(w_i)^2 \right) \right] - \xi_{1,1} \xi_{2,0}
= \frac{m \xi_{3,1} + m(m-1) \xi_{1,1} \xi_{2,0}}{m^2} - \xi_{1,1} \xi_{2,0}
= \frac{\xi_{3,1} - \xi_{1,1} \xi_{2,0}}{m} = Cov(a, c),
\]
\[
Cov(b, c) = \frac{\xi_{2,2} - \xi_{2,0}^2}{m}.
\]
We can formulate the covariance matrix

\[ \text{Cov}(a, b, c) = \frac{1}{m} \begin{pmatrix} 
\zeta_{2,2} - \zeta_{1,1}^2 & \zeta_{3,1} - \zeta_{1,1} \zeta_{2,0} & \zeta_{3,1} - \zeta_{1,1} \zeta_{2,0} \\
\zeta_{3,1} - \zeta_{1,1} \zeta_{2,0} & \zeta_{4,0} - \zeta_{2,0}^2 & \zeta_{2,2} - \zeta_{2,0}^2 \\
\zeta_{3,1} - \zeta_{1,1} \zeta_{2,0} & \zeta_{2,2} - \zeta_{2,0}^2 & \zeta_{4,0} - \zeta_{2,0}^2 
\end{pmatrix}. \]

The gradient vector at the expectations is

\[ \nabla \hat{k}_{n,Q}(\mathbb{E}[a], \mathbb{E}[b], \mathbb{E}[c]) = \left( \frac{1}{\zeta_{2,0}}, -\frac{\zeta_{1,1}}{2\zeta_{2,0}}, -\frac{\zeta_{1,1}}{2\zeta_{2,0}} \right). \]

By Taylor expansion, it holds that

\[ \text{Var}[\hat{k}_{n,Q}] = \nabla \hat{k}_{n,Q}(\mathbb{E}[a], \mathbb{E}[b], \mathbb{E}[c])^T \text{Cov}(a, b, c) \nabla \hat{k}_{n,Q}(\mathbb{E}[a], \mathbb{E}[b], \mathbb{E}[c]) + O\left(\frac{1}{m^2}\right). \]

The theorem is proved by plugging in the expressions. \(\square\)

\textbf{Proof of Theorem 6.3.4}

\textit{Proof.} Let \(z_x = \cos(\gamma X + \tau), z_y = \cos(\gamma Y + \tau)\) where \((X, Y) \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),\) \(\tau \sim \text{uniform}(0, 2\pi).\) Denote \(\zeta_{s,t} = \mathbb{E}[Q(z_x)^s Q(z_y)^t].\) Recalling Theorem 6.3.2 and Theorem 6.3.3, we have asymptotically (omitting lower order terms)

\[ \mathbb{E}[\hat{k}_Q] = 2\zeta_{1,1}, \quad \text{Var}[\hat{k}_Q] = \frac{4}{m}(\zeta_{2,2} - \zeta_{1,1}^2), \]

\[ \mathbb{E}[\hat{k}_{n,Q}] = \frac{\zeta_{1,1}}{\zeta_{2,0}}, \quad \text{Var}[\hat{k}_{n,Q}] = \frac{1}{m}\left(\frac{\zeta_{2,2}}{\zeta_{2,0}^2} - \frac{2\zeta_{1,1} \zeta_{3,1}}{\zeta_{2,0}^3} + \frac{\zeta_{1,1}^2 (\zeta_{4,0} + \zeta_{2,2})}{2\zeta_{2,0}^4}\right). \]
Thus, we can compute the debiased estimator variance as (after simplification)

\[
\text{Var}_{db}[\hat{k}_Q] = \frac{K(u, v)^2}{m} \left( \frac{2\zeta_2}{\zeta_{1,1}} - 1 \right),
\]

\[
\text{Var}_{db}[\hat{k}_{n,Q}] = \frac{K(u, v)^2}{m} \left( \frac{2\zeta_{3,1}}{2\zeta_{1,1} \zeta_{2,0}} + \frac{\zeta_{4,0}\zeta_{2,2}}{2\zeta_{2,0}} \right).
\]

Taking the difference, we obtain

\[
\text{Var}_{db}[\hat{k}_Q] - \text{Var}_{db}[\hat{k}_{n,Q}] \propto 4\zeta_{2,0}\zeta_{3,1} + \zeta_{1,1}(\zeta_{4,0} + \zeta_{2,2}) - 2\zeta_{1,1}\zeta_{2,0}^2 \geq 4\zeta_{2,0}\zeta_{3,1} + \zeta_{1,1}(\zeta_{2,2} - \zeta_{2,0}^2) \overset{\Delta}{=} M(\rho),
\]

where the inequality is due to the fact that \(\zeta_{4,0} - \zeta_{2,0}^2 = \text{Var}[Q^2(z_x)] \geq 0\). Here we denote \(M\) as a function of \(\rho\). At \(\rho = 0\), we have

\[
\zeta_{3,1} = 0, \quad \zeta_{2,2} = \zeta_{2,0}^2,
\]

so that \(M(0) = 0\). At \(\rho = 1\), it holds that

\[
\zeta_{3,1} = \zeta_{2,2} = \zeta_{4,0},
\]

hence \(M(1) > 0\). Notice that \(Q(\cdot)\) and \(Q^3(\cdot)\) are non-decreasing odd functions, and \(Q^2(\cdot)\) is an even function. For \(\rho \in [0, 1]\), since \(\sqrt{2(1-\rho)} \leq \sqrt{2\gamma} \leq \pi\) by assumption, it follows from Theorem 6.3.6 that \(\zeta_{1,1} \), \(\zeta_{2,2} \) and \(\zeta_{3,1} \) are all increasing in \(\rho\) on \([0, 1]\). Consequently, \(M(\rho) > 0\) for any \(\rho \in [0, 1]\). The desired result thus follows.

\[\square\]

**Proof of Lemma 6.3.5**

**Lemma E.1** (Stein’s Lemma). Suppose \(X \sim N(\mu, \sigma^2)\), and \(g\) is a differentiable function such that \(\mathbb{E}[g(X)(X - \mu)]\) and \(\mathbb{E}[g'(X)]\) exist. Then, \(\mathbb{E}[g(X)(X - \mu)] = \sigma^2 \mathbb{E}[g'(X)]\).

**Proof.** (of Lemma 6.3.5) We use the technique of Gaussian interpolation and Stein’s Lemma.
First, we formulate \( Y = \gamma \rho X + \gamma \sqrt{1 - \rho^2} Z \) where \( Z \sim N(0, 1) \) independent of \( X \). By continuity and boundedness of \( g_1 \) and \( g_2 \), it holds that

\[
\frac{\partial}{\partial \rho} \mathbb{E}[g_1(\cos(s_x))g_2(\cos(s_y))]
= \frac{\partial}{\partial \rho} \mathbb{E}_{X,Z,\tau}[g_1(\cos(\gamma X + \tau))g_2(\cos(\gamma \rho X + \gamma \sqrt{1 - \rho^2} Z + \tau))]
= -\mathbb{E}_{X,Z,\tau} \left[ (\gamma X - \frac{\gamma \rho Z}{\sqrt{1 - \rho^2}}) \right. \\
\left. \frac{g_1(\cos(\gamma X + \tau))g_2(\cos(\gamma \rho X + \gamma \sqrt{1 - \rho^2} Z + \tau)) \sin(\gamma \rho X + \gamma \sqrt{1 - \rho^2} Z + \tau)}{\gamma(X, Z, \rho)} \right].
\]

We analyze two parts respectively. By Lemma E.1 and law of total expectation, we have

\[
\mathbb{E}_{X,Z,\tau}[\Upsilon(X, Z; \rho) \gamma X]
= \mathbb{E}_{Z,\tau} \mathbb{E}_X \left[ -\gamma^2 g'_1(\cos(\gamma X + \tau)) \sin(\gamma X + \tau) \\
- \frac{g'_2(\cos(\gamma Y + \tau)) \sin(\gamma \rho X + \gamma \sqrt{1 - \rho^2} Z + \tau)}{\gamma(X, Z, \rho)} \\
- \gamma^2 \rho g_1(\cos(\gamma X + \tau))g'_2(\cos(\gamma \rho X + \gamma \sqrt{1 - \rho^2} Z + \tau)) \sin^2(\gamma \rho X + \gamma \sqrt{1 - \rho^2} Z + \tau) \\
+ \gamma^2 \rho g_1(\cos(\gamma X + \tau))g'_2(\cos(\gamma \rho X + \gamma \sqrt{1 - \rho^2} Z + \tau)) \cos(\gamma \rho X + \gamma \sqrt{1 - \rho^2} Z + \tau) \big| Z, \tau \right], \quad (E.9)
\]
Combining (E.9) and (E.10), we get

\[
\frac{\partial}{\partial \rho} \mathbb{E}[g_1(\cos(s_x))g_2(\cos(s_y))] \\
= \mathbb{E}_{X,Z,\tau} \left[ \gamma^2 g_1'(\cos(\gamma X + \tau)) \sin(\gamma X + \tau) g_2'(\cos(\gamma Y + \tau)) \sin(\gamma X + \gamma \sqrt{1 - \rho^2 Z + \tau}) \right] \\
= \gamma^2 \mathbb{E}_{X,Y,\tau} \left[ g_1'(\cos(s_x)) \sin(s_x) g_2'(\cos(s_y)) \sin(s_y) \right],
\]

which gives the desired expression.

To prove the monotonicity, suppose that \(g_1\) and \(g_2\) are increasing odd or non-constant even functions. So, \(g_1(-x)g_2(-x) = g_1(x)g_2'(x) > 0\), \(\forall x \in [-1, 1]\). Assume \(\sqrt{2(1 - \rho)} \gamma \leq \pi\),
and denote \( f(x, y) \) as the joint density given by Theorem 6.2. We can write

\[
\frac{\partial \mathbb{E}[g_1(\cos(s_x))g_2(\cos(s_y))]}{\partial \rho} = \gamma^2 \int_{-1}^{1} \int_{-1}^{1} z_x z_y g'_1(\sqrt{1 - z_x^2}) g'_2(\sqrt{1 - z_y^2}) f(z_x, z_y) dz_x dz_y
\]

\[
(a) = 2\gamma^2 \left( \int_0^1 \int_0^1 z_x z_y g'_1(\sqrt{1 - z_x^2}) g'_2(\sqrt{1 - z_y^2}) f(z_x, z_y) dz_x dz_y \right)

\]

\[
+ \int_{-1}^1 \int_{-1}^0 z_x z_y g'_1(\sqrt{1 - z_x^2}) g'_2(\sqrt{1 - z_y^2}) f(z_x, z_y) dz_x dz_y
\]

\[
= 2\gamma^2 \int_0^1 \int_0^1 z_x z_y g'_1(\sqrt{1 - z_x^2}) g'_2(\sqrt{1 - z_y^2}) \left[ f(z_x, z_y) - f(z_x, -z_y) \right] dz_x dz_y
\]

\[
(b) > 0,
\]

where (a) is due to the symmetry of \( f \) and \( g \), and (b) is a consequence of Proposition 6.1.5 that \( f(z_x, z_y) > f(z_x, -z_y) \) for all \( z_x, z_y \in (0, 1]^2 \), provided that \( \sqrt{2(1-\rho)}\gamma \leq \pi \). The proof is complete.

Proof of Theorem 6.3.6

Proof. Since \( Q_1 \) and \( Q_2 \) both are non-decreasing and have finite number of discontinuities, by Baire's Characterization Theorem we know that each of them is the point-wise limit of a sequence of continuous increasing functions. Suppose that \( \{g_{1,n}\} \) and \( \{g_{2,n}\} \) are two sequences of continuous increasing functions such that as \( n \to \infty \), \( g_{1,n} \to Q_1 \) and \( g_{2,n} \to Q_2 \) with point-wise convergence. By dominated convergence theorem, we have

\[
\frac{\partial \mathbb{E}[Q_1(z_x)Q_2(z_y)]}{\partial \rho} = \frac{\partial \mathbb{E}\left[ \lim_{n \to \infty} g_{1,n}(z_x) \lim_{n \to \infty} g_{2,n}(z_y) \right]}{\partial \rho} = \lim_{n \to \infty} \frac{\partial \mathbb{E}[g_{1,n}(z_x)g_{2,n}(z_y)]}{\partial \rho} > 0,
\]
where Lemma 6.3.5 is adopted for continuous $g_{1,n}$ and $g_{2,n}$ functions.
APPENDIX F
PROOFS OF CHAPTER 7

Proof of Theorem 7.3.2

Proof. First we look at the sine function. Let \((x, y) \sim \phi_{\gamma, \rho}\). \(Q\) is a general quantizer applied to the linear random projections. We have

\[
\mathbb{E}[\sin(Q(x)\sin(Q(y))] = \mathbb{E}\left[\left(\sin(Q(x)) - \sin(x) + \sin(x)\right)\right.
\]
\[
\left.\left(\sin(Q(y)) - \sin(y) + \sin(y)\right)\right]
\]
\[
= \mathbb{E}\left[(\sin(Q(x)) - \sin(x))(\sin(Q(y)) - \sin(y))\right]
\]
\[
+ 2\mathbb{E}[\sin(Q(x)) - \sin(x) \sin(y)] + E[\sin(x) \sin(y)]
\]
\[\triangleq T_1 + 2T_2 + E[\sin(x) \sin(y)]. \quad \text{(F.1)}\]

By Young’s inequality, we have

\[\quad -D_s \leq T_1 \leq D_s. \quad \text{(F.2)}\]

To bound the second term, the following identities would be useful. For \(x \sim N(0, \gamma^2)\),

\[
\mathbb{E}[\cos(x)] = e^{-\frac{\gamma^2}{2}},
\]
\[
\mathbb{E}[\sin(ax)\sin(bx)] = \frac{1}{2}\left[e^{-\frac{\gamma^2(a-b)^2}{2}} - e^{-\frac{\gamma^2(a+b)^2}{2}}\right],
\]
\[
\mathbb{E}[\cos(ax)\cos(bx)] = \frac{1}{2}\left[e^{-\frac{\gamma^2(a-b)^2}{2}} + e^{-\frac{\gamma^2(a+b)^2}{2}}\right].
\]
Since we can write \( y = \rho x + \sqrt{1 - \rho^2}Z \) with \( Z \sim N(0, \gamma^2) \) independent of \( x \), we have

\[
T_2 = \mathbb{E}\left[ (\sin(Q(x)) - \sin(x)) \sin(\rho x + \sqrt{1 - \rho^2}Z) \right] \\
= \mathbb{E}\left[ (\sin(Q(x)) - \sin(x)) \sin(\rho x) \cos(\sqrt{1 - \rho^2}Z) \right] \\
+ \mathbb{E}\left[ (\sin(Q(x)) - \sin(x)) \cos(\rho x) \sin(\sqrt{1 - \rho^2}Z) \right] \\
= e^{-\frac{\gamma^2(1-\rho^2)}{2}} \mathbb{E}\left[ (\sin(Q(x)) - \sin(x)) \sin(\rho x) \right]. \tag{F.3}
\]

By assumption,

\[
\text{Cov}[\sin(Q(x)) - \sin(x), \sin(x)] = \zeta_s, \\
\text{Var}[\sin(Q(x)) - \sin(x)] = D_s,
\]

Now we can compute

\[
\begin{align*}
\text{Var}[\sin(x)] &= \frac{1}{2} \left[ 1 - e^{-2\gamma^2} \right], \\
\text{Cov}[\sin(\rho x), \sin(x)] &= \frac{1}{2} \left[ e^{-\frac{\gamma^2(1-\rho^2)}{2}} - e^{-\frac{\gamma^2(1+\rho^2)}{2}} \right], \\
\text{Var}[\sin(\rho x)] &= \frac{1}{2} \left[ 1 - e^{-2\rho^2\gamma^2} \right] \triangleq V_s^*.
\end{align*}
\]

The correlation coefficients are

\[
C_1 \triangleq \text{Corr}[\sin(Q(x)) - \sin(x), \sin(x)] = \frac{\sqrt{2}\zeta_s}{\sqrt{D_s(1 - e^{-2\gamma^2})}},
\]

\[
C_2 \triangleq \text{Corr}[\sin(\rho x), \sin(x)] = \frac{e^{-\frac{\gamma^2(1-\rho^2)}{2}} - e^{-\frac{\gamma^2(1+\rho^2)}{2}}}{\sqrt{(1 - e^{-2\gamma^2})(1 - e^{-2\rho^2\gamma^2})}}.
\]

By Cauchy-Schwartz inequality, we know that \( \text{Corr}[\sin(Q(x)) - \sin(x), \sin(\rho x)] \) is bounded between

\[
C_{11} \triangleq C_1 C_2 - \sqrt{(1 - C_1^2)(1 - C_2^2)},
\]
\[ C_{u1} \triangleq C_1 C_2 + \sqrt{(1 - C_1^2)(1 - C_2^2)}. \]

Therefore, we have that
\[ C_{u1} \sqrt{D_s V_s^*} \leq E\left[ (\sin(Q(x)) - \sin(x)) \sin(\rho x) \right] \leq C_{u1} \sqrt{D_s V_s^*}. \]

Combining with (F.1), (F.2) and (F.3) gives the expression for the sine part. For the cosine, we can use similar approach. From now on, denote \( Q = Q_{c, \gamma} \). In particular, we can have
\[
\mathbb{E}[\cos(Q(x) \cos(Q(y))] = T_3 + 2T_4 + E[\cos(x) \cos(y)],
\]
where
\[ T_3 = \mathbb{E}[(\cos(Q(x)) - \cos(x))(\cos(Q(y)) - \cos(y))], \]
\[ T_4 = \mathbb{E}[(\cos(Q(x)) - \cos(x)) \cos(y)]. \]

Similarly, we have
\[ -D_c \leq T_3 \leq D_c, \]
and
\[ T_4 = e^{-\frac{\gamma^2(1-\rho^2)}{2}} \mathbb{E}\left[ (\cos(Q(x)) - \cos(x)) \cos(\rho x) \right]. \]
Similarly, we can obtain

\[
\begin{align*}
\text{Cov}[\cos(Q(x)) - \cos(x), \cos(x)] &= \zeta_c, \\
\text{Var}[\cos(Q(x)) - \cos(x)] &= \tilde{D}_c, \\
\text{Var}[\cos(x)] &= \frac{1}{2} \left[ 1 + e^{-2\gamma^2} \right] - e^{-\gamma^2}, \\
\text{Cov}[\cos(\rho x), \cos(x)] &= \frac{1}{2} \left[ e^{-\frac{\gamma^2(1-\rho)^2}{2}} + e^{-\frac{\gamma^2(1+\rho)^2}{2}} \right] - e^{-\frac{\gamma^2(1+\rho^2)}{2}}, \\
\text{Var}[\cos(\rho x)] &= \frac{1}{2} \left[ 1 + e^{-2\rho^2\gamma^2} \right] - e^{-\rho^2\gamma^2} \triangleq V^*_c.
\end{align*}
\]

The remaining part is similar, where we use Cauchy-Schwartz to bound the correlation of \(\cos(Q(x)) - \cos(x)\) and \(\cos(\rho x)\). We omit it for conciseness. The desired result is obtained by combining two parts and noticing that \(\mathbb{E}[\sin(x) \sin(y) + \cos(x) \cos(y)] = e^{-\gamma^2(1-\rho)} = k(x, y)\).

\[\blacksquare\]

**Proof of Theorem 7.3.3**

**Proof.** We denote the sample space (unit sphere) as \(\mathcal{S} = \mathbb{S}^{d-1}\). Let \(\tilde{\mathcal{S}}_\Delta\) be a \(\Delta\)-net placed on \(\mathcal{S}\). We then can express any \(x \in \mathcal{S}\) as \(x = \tilde{x} + r_x\), for the center \(\tilde{x} \in \tilde{\mathcal{S}}_\Delta\) and \(\|r_x\| \leq \Delta\).

Define

\[
\begin{align*}
k^s_Q(x, y) &= \mathbb{E}[\sin(Q(w^T x)) \sin(Q(w^T y))], \\
k^c_Q(x, y) &= \mathbb{E}[\cos(Q(w^T x)) \cos(Q(w^T y))].
\end{align*}
\]
We have

\[ |\hat{k}_Q(x, y) - k_Q(x, y)| = |\hat{k}_Q(x, y) + \hat{k}_Q(x, y) - k'_Q(x, y) - k'_Q(x, y)| \]
\[ \leq |\hat{k}_Q(x, y) - k'_Q(x, y)| + |\hat{k}_Q(x, y) - k'_Q(x, y)|. \]  (F.4)

As before, we mainly provide details on the sine part, and the reasoning applies to the cosine part similarly. We will drop the subscript of \(Q_s\) for simplicity. For any \(x, y \in S\), firstly we assume that the following two events hold:

\[ \Omega_1 : \sup_{\tilde{x} \in \tilde{S}_\Delta} \frac{1}{m} \sum_{i=1}^{m} \sup_{\|r\| \leq \triangle} |\sin(Q(w_i^T \tilde{x} + w_i^T r)) - \sin(Q(w_i^T \tilde{x}))| \leq L_s \gamma \Delta + \epsilon_1, \]

\[ \Omega_2 : \sup_{\tilde{x}, \tilde{y} \in \tilde{S}_\Delta} |\hat{k}_s^s(\tilde{x}, \tilde{y}) - k_s^s(\tilde{x}, \tilde{y})| \leq \epsilon_2. \]

For any \(x, y \in S\), we have the following bound by triangle inequality,

\[ |\hat{k}_Q(x, y) - k'_Q(x, y)| \leq |\hat{k}_Q(x, y) - \hat{k}_Q(\tilde{x}, y)| + |\hat{k}_Q(\tilde{x}, y) - \hat{k}_Q(\tilde{x}, \tilde{y})| + |k'_Q(x, y) - k'_Q(\tilde{x}, \tilde{y})| \]
\[ \triangleq T_1 + T_2 + T_3 + T_4. \]  (F.5)

We now bound these terms separately. We have

\[ T_1 = \frac{1}{m} \left| \sum_{i=1}^{m} \sin(Q(w_i^T \tilde{x} + w_i^T r_x)) \sin(Q(w_i^T \tilde{y} + w_i^T r_y)) - \sin(Q(w_i^T \tilde{x})) \sin(Q(w_i^T \tilde{y} + w_i^T r_y)) \right| \]
\[ = \frac{1}{m} \left| \sum_{i=1}^{m} \left[ \sin(Q(w_i^T \tilde{x} + w_i^T r_x)) - \sin(Q(w_i^T \tilde{x})) \right] \sin(Q(w_i^T \tilde{y} + w_i^T r_y)) \right| \]
\[ \leq L_s \gamma \Delta + \epsilon_1, \]
where the last line is due to event $\Omega_1$ and boundedness of sine function. Similarly,

$$T_2 = \frac{1}{m} \left| \sum_{i=1}^{m} \sin(Q(w_i^T \tilde{x})) \left[ \sin(Q(w_i^T \tilde{y} + w_i^T r_y)) - \sin(Q(w_i^T \tilde{y})) \right] \right|$$

$$\leq L_Q^s \gamma \Delta + \epsilon_1.$$

The event $\Omega_2$ directly implies that

$$T_3 \leq \epsilon_2.$$

For $T_4$, by mean smoothness assumption we have

$$T_4 = \left| E \left[ \sin(Q(w_i^T \tilde{x} + w_i^T r_x)) \sin(Q(w_i^T \tilde{y} + w_i^T r_y)) - \sin(Q(w_i^T \tilde{x})) \sin(Q(w_i^T \tilde{y})) \right] \right|$$

$$= \left| E \left[ \sin(Q(w_i^T \tilde{x} + w_i^T r_x)) \sin(Q(w_i^T \tilde{y} + w_i^T r_y)) - \sin(Q(w_i^T \tilde{x})) \sin(Q(w_i^T \tilde{y})) + \sin(Q(w_i^T \tilde{x} + w_i^T r_x)) \sin(Q(w_i^T \tilde{y})) - \sin(Q(w_i^T \tilde{x} + w_i^T r_x)) \sin(Q(w_i^T \tilde{y})) \right] \right|$$

$$\leq E \left| \sin(Q(w_i^T \tilde{x} + w_i^T r_x)) \left[ \sin(Q(w_i^T \tilde{y} + w_i^T r_y)) - \sin(Q(w_i^T \tilde{y})) \right] - \left[ \sin(Q(w_i^T \tilde{x} + w_i^T r_x)) - \sin(Q(w_i^T \tilde{x})) \right] \sin(Q(w_i^T \tilde{y})) \right|$$

$$\leq L_Q^s E \left[ \|w_i^T r_x\| + \|w_i^T r_y\| \right]$$

$$\leq 2 L_Q^s \gamma \Delta.$$

Summing up ingredients together in (F.5) we get that in event $\Omega_1$ and $\Omega_2$, we have

$$|k^s_Q(x, y) - k^s_Q(x, y)| \leq 2\epsilon_1 + \epsilon_2 + 4L_Q^s \gamma \Delta.$$

To derive a high probability bound, we now investigate the two events. First, we have the
complement

\[ P[\Omega_1^c] = P\left[ \sup_{\tilde{x} \in \tilde{S}_\Delta} \frac{1}{m} \sum_{i=1}^{m} \sup_{\|r\| \leq \Delta} |\sin(Q(w_i^T \tilde{x} + w_i^T r)) - \sin(Q(w_i^T \tilde{x}))| \geq L_Q \gamma \Delta + \epsilon_1 \right]. \]

Since the terms in the summation, \( \sup_{\|r\| \leq \Delta} |\sin(Q(w_i^T \tilde{x} + w_i^T r)) - \sin(Q(w_i^T \tilde{x}))| \), are i.i.d. random variables, for any \( \tilde{x} \in \tilde{S}_\Delta \) the expectation admits

\[ \mathbb{E}\left[ \sup_{\|r\| \leq \Delta} |\sin(Q(w_i^T \tilde{x} + w_i^T r)) - \sin(Q(w_i^T \tilde{x}))| \right] \leq L_Q^s \gamma \Delta, \]

due to mean smoothness of \( Q \). By Hoeffding’s inequality on bounded variables, we get \( \forall \tilde{x} \in \tilde{S}_\Delta \)

\[ P\left[ \frac{1}{m} \sum_{i=1}^{m} \sup_{\|r\| \leq \Delta} |\sin(Q(w_i^T \tilde{x} + w_i^T r)) - \sin(Q(w_i^T \tilde{x}))| \geq L_Q^s \gamma \Delta + \epsilon_1 \right] \leq e^{-2m \epsilon_1^2 / 4m} = e^{-m \epsilon_1^2 / 2}. \]

Applying union bound over all \( \tilde{x} \in \tilde{S}_\Delta \), we get

\[ P[\Omega_1^c] \leq |\tilde{S}_\Delta| e^{-m \epsilon_1^2 / 2} \leq \left( \frac{2}{\Delta} + 1 \right)^d e^{-m \epsilon_1^2 / 2}, \]

where the last inequality is due to the bound on covering number of the unit sphere (Corollary 4.2.13 in Vershynin (2018b)). When \( m \geq \frac{4d}{\epsilon_1^2} \log\left(\frac{2}{\Delta} + 1\right) \), we have \( P[\Omega_1^c] \leq e^{-m \epsilon_1^2 / 4} \). For \( \Omega_2 \), applying Hoeffding’s inequality yields a point-wise bound, where for \( \forall \tilde{x}, \tilde{y} \in \tilde{S}_\Delta \),

\[ P\left[ |k_Q^s(\tilde{x}, \tilde{y}) - k_Q^s(\tilde{x}, \tilde{y})| \geq \epsilon_2 \right] = P\left[ \frac{1}{m} \sum_{i=1}^{m} \sin(Q(\tilde{x})) \sin(Q(\tilde{y})) - K_Q^s(\tilde{x}, \tilde{y}) \geq \epsilon_2 \right] \leq 2e^{-m \epsilon_2^2 / 2}. \]
Casting an union bound over \((\tilde{x}, \tilde{y}) \in \tilde{S}_\triangle \times \tilde{S}_\triangle\) yields

\[
P\left[\Omega_2^c\right] = P\left[\sup_{\tilde{x}, \tilde{y} \in \tilde{S}_\triangle} |\hat{k}_Q^s(\tilde{x}, \tilde{y}) - k_Q^s(\tilde{x}, \tilde{y})| \geq \epsilon_2 \right] 
\leq 2 \left(\frac{|\tilde{S}_\triangle|}{2}\right) e^{-m\epsilon_2^2/2}
\leq \left(\frac{2}{\Delta} + 1\right)^{2d} e^{-m\epsilon_2^2/2}.
\]

Consequently, \(P[\Omega_2^c] \leq e^{-m\epsilon_2^2/4}\) when \(m \geq \frac{8d}{\epsilon_1^2} \log\left(\frac{2}{\Delta} + 1\right)\). Therefore, we obtain that when \(m \geq 4d \log\left(\frac{2}{\Delta} + 1\right) \max\{\epsilon_1^{-2}, 2\epsilon_2^{-2}\}\),

\[
P\left[\Omega_1^c \cup \Omega_2^c\right] \leq e^{-m\epsilon_1^2/4} + e^{-m\epsilon_2^2/4}.
\]

Now by letting \(\epsilon_1 = \epsilon_2 = \epsilon/8\), and choosing \(\Delta = \frac{\epsilon}{32L_Q^2\gamma}\), we have proved that when \(m \geq \frac{512d}{\epsilon^2} \log\left(\frac{64L_Q^2\gamma}{\epsilon} + 1\right)\), the error of sine part is bounded as

\[
|\hat{k}_Q^s(x, y) - k_Q^s(x, y)| \leq \epsilon/2,
\]

with probability at least \(1 - 2e^{-m\epsilon^2/256}\). Similarly analysis can be used to bound the cosine part. For conciseness we omit the detailed proof. It is true that when \(m \geq \frac{512d}{\epsilon^2} \log\left(\frac{64L_Q\gamma}{\epsilon} + 1\right)\), with probability \(1 - 2e^{-m\epsilon^2/256}\) we have

\[
|\hat{k}_Q^c(x, y) - k_Q^c(x, y)| \leq \epsilon/2.
\]

Therefore, by (F.4) and union bound we know that when \(m \geq \frac{512d}{\epsilon^2} \log\left(\frac{64\max\{L_Q^s, L_Q^c\}\gamma}{\epsilon} + 1\right)\), the kernel approximation error is uniformly bounded by

\[
|\hat{k}_Q(x, y) - k_Q(x, y)| \leq \epsilon,
\]

for \(\forall x, y \in S\), with probability \(1 - 4e^{-m\epsilon^2/256}\). This completes the proof. \(\square\)
Proof of Proposition 7.3.4

Proof. We present the analysis of sine function. Assume the quantizer has $b$ bits. Thus it contains $2^b - 1$ finite borders, denoted as $t^*_1, ..., t^*_{2^b-1}$. For a fixed point $t$, the value $\sup_{|r| \leq \delta} | \sin(Q(t+r)) - \sin(Q(t)) |$ equals to 0 if $t^*_i + \delta \leq t \leq t^*_i + 1 - \delta$ for some $i$. Otherwise, $\sup_{|r| \leq \delta} | \sin(Q(t+r)) - \sin(Q(t)) |$ would be bounded by 2. Therefore, integrating over the domain of $\Gamma$ gives

$$
\mathbb{E}_{t \sim N(0, \gamma^2)} \left[ \sup_{|r| \leq \delta} \left| f(Q(t+r)) - f(Q(t)) \right| \right] \\
\leq 2P \left[ t \in \bigcup_{i=1}^{2^b-1} [t^*_i - \delta, t^*_i + \delta] \right] \\
\leq \frac{4(2^b - 1)}{\sqrt{2\pi}\gamma} \delta.
$$

The last line is due to the fact that for $t \sim N(0, \gamma^2)$, the property of normal density implies $P[t^* - \delta \leq t \leq t^* + \delta] \leq 2\delta \cdot \frac{1}{\sqrt{2\pi}\gamma}$ for any $t^*$. Therefore, the mean smoothness constant $L^s_Q$ is at most $\frac{4(2^b - 1)}{\sqrt{2\pi}\gamma}$. Similar proof holds for cosine function. \qed