

SPARSE  $K$ -MEANS COMPRESSION FOR  
FEDERATED MACHINE LEARNING AND LINEAR REGRESSION  
USING SKETCHED AND QUANTIZED PREDICTORS

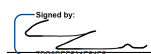
by

Daniel Hill  
A Dissertation  
Submitted to the  
Graduate Faculty  
of  
George Mason University  
In Partial fulfillment of  
The Requirements for the Degree  
of  
Doctor of Philosophy  
Statistics

Committee:

Signed by:  
  
279D324D050F474...

Dr. David Keppinger, Dissertation Director

Signed by:  
  
10B88E934E4F0...

Dr. Martin Slawski, Co-Director, External Examiner

Signed by:  
  
8C4203320F7B491...

Dr. Ben Seiyon Lee, Committee Member

DocuSigned by:  
  
F8E0B82293C403...

Dr. Anand Vidyashankar, Committee Member

Signed by:  
  
0993CF409326401...

Dr. Jiayang Sun, Department Chair

Date: \_\_\_\_\_

Spring 2025  
George Mason University  
Fairfax, VA

Sparse  $K$ -Means Compression for Federated Machine Learning and Linear Regression  
Using Sketched and Quantized Predictors

A dissertation submitted in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy at George Mason University

By

Daniel Hill  
Master of Science  
Air Force Institute of Technology, 2018  
Bachelor of Science  
Bethel College, 2011

Director: Dr. David Kepplinger, Professor  
Department of Statistics

Spring 2025  
George Mason University  
Fairfax, VA

This is a work of the U.S. Government and is not subject to copyright protection in the United States. Foreign copyrights may apply.

The views expressed in this dissertation are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

## Table of Contents

	Page
List of Tables . . . . .	vi
List of Figures . . . . .	vii
Abstract . . . . .	viii
1 Introduction . . . . .	1
2 Sparse K-means (SparK) . . . . .	3
2.1 Background . . . . .	3
2.1.1 Introduction . . . . .	3
2.1.2 Existing Compression Algorithms . . . . .	8
2.2 Sparse K-means (SparK) . . . . .	10
2.2.1 SparK Description . . . . .	11
2.2.2 SparK on Simulated Data . . . . .	15
2.2.3 SparK Theory . . . . .	17
2.2.4 Evaluating the Use of $\xi$ . . . . .	19
2.3 Sparse K-means in Simulation . . . . .	24
2.3.1 Simulation Description . . . . .	24
2.3.2 Results . . . . .	25
2.4 Discussion . . . . .	25
2.4.1 Future Work . . . . .	28
3 1-Bit Quantized Regression . . . . .	31
3.1 Introduction . . . . .	31
3.1.1 Background . . . . .	31
3.1.2 Quantization . . . . .	33
3.2 1-bit Quantized Regression, 1-Predictor Case . . . . .	33
3.2.1 Asymptotic Bias of the 1-Bit Quantized Estimator . . . . .	35
3.2.2 Asymptotic Variance of the 1-bit Quantized Estimator . . . . .	38
3.2.3 Alternative Derivation Using Estimating Equations . . . . .	41
3.2.4 MSE and Asymptotic Relative Efficiency . . . . .	43
3.3 1-Bit Quantized Regression, $d$ -Predictor Case . . . . .	43

3.3.1	Results . . . . .	44
3.3.2	Introduction . . . . .	45
3.3.3	Quantized Predictor of $\beta^0$ . . . . .	47
3.3.4	Asymptotic Variance of the Quantized Estimator . . . . .	47
3.3.5	MSE of the Estimator . . . . .	51
3.3.6	Asymptotic Relative Efficiency . . . . .	52
3.4	Conclusion . . . . .	53
3.4.1	Summary . . . . .	53
4	Bounding the Error from Sketched and Quantized Regression Parameters . . . . .	56
4.1	Introduction . . . . .	56
4.1.1	Introduction . . . . .	56
4.1.2	Background . . . . .	56
4.1.3	Problem Setup . . . . .	59
4.1.4	Assumptions and Definitions . . . . .	64
4.2	Quantized Regression Parameters . . . . .	67
4.2.1	Introduction . . . . .	67
4.2.2	Quantized Scenario with Fixed $\mathbf{Z}$ . . . . .	69
4.2.3	Quantized Scenario with Gaussian $\mathbf{Z}$ . . . . .	84
4.3	Sketched Regression Parameters . . . . .	102
4.3.1	Introduction . . . . .	103
4.3.2	Sketched Scenario with Fixed $\mathbf{Z}$ . . . . .	103
4.3.3	Sketched Scenario with Gaussian $\mathbf{Z}$ . . . . .	113
4.4	Sketched and Quantized Regression Parameters . . . . .	133
4.4.1	Introduction . . . . .	133
4.4.2	The Sketched and Quantized Scenario with Fixed $\mathbf{Z}$ . . . . .	135
4.4.3	The Sketched and Quantized Scenario with Gaussian $\mathbf{Z}$ . . . . .	153
4.4.4	Summary of the Sketched and Quantized Scenario with Gaussian $\mathbf{Z}$ . . . . .	166
4.5	Discussion . . . . .	167
4.5.1	Quantized Estimator . . . . .	169
4.5.2	Sketched Estimator . . . . .	173
4.5.3	Sketched and Quantized Estimator . . . . .	179
A	Derivations from Section 3.2.2 . . . . .	184
A.1	Preliminary Formulations . . . . .	184
A.2	$\mathbf{E} [\hat{U}]$ . . . . .	185
A.3	$\mathbf{E} [\hat{V}]$ . . . . .	186

A.4	$\mathbf{Var} \left[ \widetilde{X}_i^2 \right]$	186
A.5	$\mathbf{Var} \left[ \widetilde{X}_i \widetilde{Y}_i \right]$	186
A.6	$\mathbf{Cov} \left( \widetilde{X}_i \widetilde{Y}_i, \widetilde{X}_i^2 \right)$	186
B	Derivations from Section 3.3.4	187
B.1	Formulas for Calculating the Elements of $\mathbf{E} \left[ \widetilde{\psi}_{\beta^0} \widetilde{\psi}_{\beta^0}^T \right]$	187
B.1.1	Term(a)	189
B.1.2	Terms (b) and (c)	193
B.1.3	Term (d)	195
B.1.4	Combining Terms and Bounding	195
C	Supporting Work for Chapter 4	200
C.1	Quantized Scenario	200
C.1.1	Quantized Scenario with Fixed $\mathbf{Z}$	200
C.1.2	Quantized Scenario with Gaussian $\mathbf{Z}$	202
D	General Supporting Ideas	204
D.0.1	Bound on the Norm of the Difference of Matrices	204
	Bibliography	205

## List of Tables

Table	Page
2.1 Parameters to Create Simulated Data. . . . .	15
3.1 Table of Important Notation . . . . .	46

## List of Figures

Figure	Page
2.1 Diagram of Federated Learning [3] . . . . .	4
2.2 Sorted Simulated Data and the Means Used to Create It. . . . .	16
2.3 SparK Applied to Simulated Data at Varying $\tau$ . . . . .	17
2.4 Validation Accuracy across Training Rounds on EMNIST . . . . .	26
2.5 Average Bitrates and Final Accuracies for all Compression Algorithms. . . . .	27
4.1 MSE of the Quantized Estimator . . . . .	171
4.2 Relative Efficiency of Quantized vs OLS Estimators . . . . .	172
4.3 MSE of the Sketched Estimator with 10000 Samples . . . . .	176
4.4 MSE of the Sketched Estimator with 100000 Samples . . . . .	177
4.5 Relative Efficiency of Sketched Estimator for $n = 10000$ . . . . .	178
4.6 Relative Efficiency of Sketched Estimator for $n = 100000$ . . . . .	179
4.7 MSE of the Sketched and Quantized Estimator . . . . .	182
4.8 Relative Efficiency of Sketched and Quantized vs OLS Estimators . . . . .	183

## Abstract

### SPARSE $K$ -MEANS COMPRESSION FOR FEDERATED MACHINE LEARNING AND LINEAR REGRESSION USING SKETCHED AND QUANTIZED PREDICTORS

Daniel Hill

George Mason University, 2025

Dissertation Director: Dr. David Kepplinger

The Information Age has led to the generation of vast and unquantifiable amounts of data, but technology has struggled to keep pace with the growing demand for efficient storage and transmission. Compression algorithms provide a means to reduce storage and transmission costs while preserving essential information for learning and analysis. This dissertation makes two contributions in this area: a novel compression scheme for federated learning and a statistical analysis framework regarding the use of compressed data in linear regression.

Regarding the first contribution, we propose the Sparse  $k$ -Means (SparK) algorithm specifically designed for Federated Learning applications. SparK compresses model parameter updates between clients and a server by combining sparsification with  $k$ -means clustering. Using the desired inverse compression rate as its sole hyperparameter, SparK optimizes the degree of sparsification and the number of clusters in  $k$ -means for each model layer to achieve the desired compression with minimal distortion. Experimental results demonstrate that SparK performs comparably or better than similar sparsification and clustering methods on a standard test bed across various compression levels.

Regarding the second contribution, we examine dithered 1-bit compression of predictors and response variables in the context of linear regression. We propose an M-estimator of the associated regression coefficients and establish its asymptotic Normality and asymptotic mean squared error (MSE). This is complemented by a non-asymptotic analysis of the MSE for three compressors: 1-bit stochastic quantization, Gaussian sketching, and their combination. High-probability upper bounds are derived for each compressor under both fixed and random design assumptions. The relative efficiency in comparison to the ordinary least squares estimator with access to uncompressed data is studied as well.

## Chapter 1: Introduction

Data is the currency of society and Machine learning (ML) unlocks its actionable and profitable insights. These data come from hospitals, cell phones, environmental sensors, and a myriad of other internet-of-things devices, and the number of devices and the size of the data continue to grow. To train a ML model on this type of data, researchers must overcome the difficulty posed by both the sensitive nature of user data and the sheer amount that each user (or sensor) creates. Compression has been introduced as a means to both improve privacy and reduce the quantity of data during its transmission to and storage at centralized data centers.

This dissertation adds to the work on compression by providing solutions to the data transmission problem.

Chapter 2 introduces a new compression algorithm called SparK. We designed SparK for compressing the parameter updates sent by clients in a federated machine learning setting. SparK furthers the work done by [1] and [2] by combining sparsification and  $k$ -means quantization. It takes as a hyperparameter the inverse compression rate and uses that to adaptively determine the optimal level of sparsification and the number of bits used for quantization in each layer of a neural network. We show its performance compared to several comparable algorithms (top- $k$  sparsification, QSGD, and Google's Quantized-encoding) on the CIFAR-100 and EMNIST datasets. We show that at very low compression SparK outperforms these algorithms and achieves only a minor loss in accuracy.

Chapter 3 introduces 1-bit Quantized Regression (QR). QR applies the same assumptions and regression equation as in linear regression, but the predictor and response variables are quantized. We begin the chapter by theoretically building up the QR framework from a 1-predictor case to a  $p$ -predictor case. We then show that the asymptotic variance of the quantized estimator of  $\beta$  can be bounded using only standard assumptions on the predictors

and error terms. We then evaluate the asymptotic relative error of the quantized estimators to the ordinary least squares estimator.

Chapter 4 shifts to a non-asymptotic analysis of regression parameter estimators created by compressing the predictors and responses using sketching, quantization, and their combination. Our work provides upper bounds on the variance of the estimators, allowing us to draw conclusions on the upper bound of the expected error as functions of the number of samples.

## Chapter 2: Sparse K-means (SparK)

In this chapter we present a new algorithm called Sparse K-Means (SparK). Section 2.1 of this chapter presents an overview of federated learning and introduces several existing compression algorithms similar to SparK. Section 2.2 describes our new algorithm and compares theoretically its performance to the algorithms presented in section 2.1. Section 2.3 provides results of a federated learning simulation implementing SparK on the EMNIST dataset and compares those results to several similar compression algorithms. The final section 2.4 discusses the results and speculates on possible ways to improve them in future work.

### 2.1 Background

#### 2.1.1 Introduction

The buzz word of our time is Artificial Intelligence (AI). Businesses, governments, tech giants, and entrepreneurs race to create the machine learning (ML) model, and its AI application, that transforms society and generates profit. The training of these models requires vast amounts of data, energy, and computing power. Under one paradigm of distributed learning, models are trained at a centralized location where user data are stored, aggregated, and used to train a global model which is then made remotely accessible to user. This paradigm puts at risk the privacy of user data as it requires both the transmission and centralized storage of personal information, such as browsing history or health information. Federated Learning (FL) reduces this risk by removing the need to transmit user data.

Federated Learning uses a server to hold a global ML model that is broadcast to all clients participating in the learning process. Each client uses its own data to train the

model locally. Each client then sends its updated parameters back to the server which aggregates the client parameters and updates the global model. The new global model is then broadcast to clients. This process repeats until the global model converges on the server. Figure 2.1 shows a simplified representation of a FL environment with four clients.

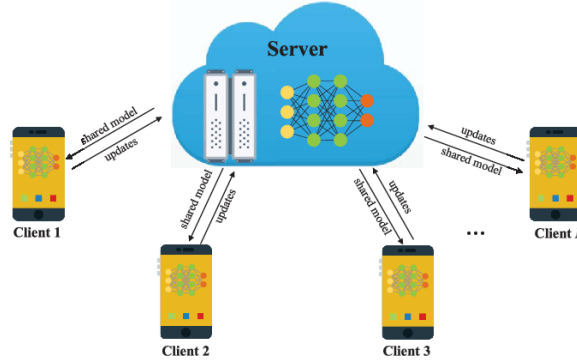


Figure 2.1: Diagram of Federated Learning [3]

While FL eliminates the need for transmitting and storing of personal user data, it requires instead the transmission of parameter updates to and from the clients and the server. Millions or even billions of parameters comprise many of today’s most advanced ML models. Such large models makes the communication of parameters in a FL paradigm a prohibitive factor to mass implementation, especially in resource constrained or bandwidth-limited environments. The focus of this chapter is the reduction of the communication requirement in a FL framework by compression of parameter updates from the clients to the server.

### Federated Learning

Before we discuss the existing communication reduction methods in FL, we will follow the lead of [4] and provide a more thorough description of FL. FL consists a system containing a server and  $K$  clients, where each  $k$ th client has a local dataset  $\mathcal{D}_k$  with  $|\mathcal{D}_k| = D_k$  data points. The purpose of the system is to train a model consisting of weights  $\mathbf{w} \in \mathbb{R}^{N_w}$  by

collaboratively optimizing

$$\min_{\mathbf{w} \in \mathbb{R}^{N_{\mathbf{w}}}} F(\mathbf{w}) = \frac{1}{\sum_{k=1}^K D_k} \sum_{k=1}^K \sum_{\ell=1}^{D_k} f(\mathbf{w}, \mathbf{z}_{k,\ell}) = \frac{1}{K} \sum_{k=1}^K \frac{1}{D_k} \sum_{\ell=1}^{D_k} f(\mathbf{w}, \mathbf{z}_{k,\ell}) \quad (2.1)$$

where  $\mathbf{z}_{k,\ell}$  is the  $\ell$ th data sample from the local data set  $\mathcal{D}_k$ . This optimization is often done in the context of mini-batch stochastic gradient descent (SGD). The process then iterates over four steps:

1. Server broadcasts the global model. The server broadcasts the parameters of an updated global model  $\mathbf{w}^t$  at the  $t$ -th training iteration.
2. Clients perform training on their local datasets. In the context of mini-batch SGD, the  $k$ -th client will compute its local gradient

$$\mathbf{g}_k^t = \frac{1}{|\mathcal{D}_k^t|} \nabla \sum_{\mathbf{z}_{k,\ell} \in \mathcal{D}_k^t} f(\mathbf{w}^t, \mathbf{z}_{k,\ell}) \quad (2.2)$$

where  $\mathcal{D}_k^t \subseteq \mathcal{D}_k$  is the randomly selected mini-batch selected during the  $t$ -th training round.

3. Clients upload local parameter updates to the server.
4. Server aggregates local parameter updates and updates global model. Based on the aggregation method, the server obtains an averaged gradient via

$$\mathbf{g}^t = \frac{1}{K} \sum_{k=1}^K \mathbf{g}_k^t \quad (2.3)$$

and updates the global model

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta_C^t \cdot \mathbf{g}^t \quad (2.4)$$

where  $\eta_s^t$  is the server learning rate at iteration  $t$ .

These four steps are repeated until some convergence criteria is met.

Using this framework, the clients' data remains on the client, thus improving data privacy and removing the need for a server to store the combined data of all clients. As mentioned previously, however, FL requires multiple rounds of client-server communication of a large number of gradient updates, often prohibiting its implementation in resource constrained or bandwidth-limited environments.

[1] and [5] provide an analysis of the communication requirement in a FL framework stating that

$$b_{total} \in \mathcal{O}(N \times freq \times b \times C) \quad (2.5)$$

is the order of the total number of bits required to achieve convergence in a FL framework. In this equation  $N$  represents the number of training rounds (completion of steps 1-4 provided above),  $freq$  is the frequency of communication of the client updates to the server,  $b$  is the total number of bits transferred in a client-to-server update, and  $C$  is the number of clients participating in each communication round.

To reduce the total required communication, one must reduce one of these four components, optimally without reducing model accuracy or convergence speed from the baseline.

### **Federated Averaging**

FedAvg [6] was the first to introduce effective means of conducting FL and remains a commonly used benchmark for new methods. FedAvg generalize FedSGD [7] while maintaining its computationally efficiency and good convergence properties. The paper [6] explains that FL differs from other forms of distributed learning in that FL removes three key assumptions:

1. independent and identically distributed (iid) data across clients
2. balanced data (number of observations at each client)

3. large data relative to the number of clients.

Furthermore, [6] introduced the now widely-used idea of replacing the transmission of the gradient of the parameters with the difference between the parameters of the client model and the global model as a proxy of the gradient.

At each global training round  $t$ , also called epochs, the FedAvg algorithm updates the global gradient  $w_{t+1}$  by taking a weighted average of the  $k = 1, \dots, K$  client updates  $w_{t+1}^k$ .

That is,

$$w_{t+1} \leftarrow \sum_{k=1}^K \frac{n_k}{n} w_{t+1}^k \quad (2.6)$$

where  $n_k$  are the number of samples at client  $k$  and  $n$  is the total number of samples across the enterprise. A description of the algorithm is provided in Algorithm 1

---

**Algorithm 1** Federated Averaging

---

- 1:  $K$  clients,  $\eta$  learning rate,  $B$  batchsize,  $\ell(w; b)$  client loss function,  $E$  local epochs of training,  $T$  max number of global epochs
  - 2:  $w_0 \leftarrow \mathbf{0}$
  - 3: **for**  $t = 1, \dots, T$  **do**
  - 4:   **for** each client  $k = 1, \dots, K$  in parallel **do**
  - 5:      $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
  - 6:      $w_{t+1} \leftarrow \sum_{k=1}^K \frac{n_k}{n} w_{t+1}^k$
  - 7: ClientUpdate( $k, w_t$ ):
  - 8:   **for**  $i = 1, \dots, E$  **do**
  - 9:     batches  $\leftarrow$  data split into batches of size  $B$
  - 10:    **for** batch  $b$  in batches **do**
  - 11:      $w^k \leftarrow w^k - \eta \nabla \ell(w^k; b)$
  - 12:  $w^k$  to Server
-

FedAvg is a commonly used benchmark used to compare new aggregation or compression algorithms against. In the next section we will provide a summary of some such algorithms.

### 2.1.2 Existing Compression Algorithms

We first present a few compression algorithms developed for deep neural networks in a conventional distributed learning model where iid assumptions hold. These methods showed performed well in these contexts and provided a foundation that subsequent research into FL used. Seide [8] presented 1-bit SGD with error feedback for the training of DNNs in 2014. This method sent the gradient to either -1 or 1, significantly reducing the communication cost with only a minor degradation to accuracy, with some assumptions. Strom’s [9] work in 2015 modified [8] by adding a thresholding, error feedback, sparsification, and loss-less encoding. Generally, quantization, sparsification, and encoding would form the foundational methods used by most future research into gradient compression.

Quantized SGD stochastically quantized the gradient, which preserved the statistical properties of the original gradient and ensured an unbiased estimate of the gradient. However, it required the hyperparameter selection of the number of bins [10] prior to training. While [10] admittedly did not take advantage of sparsity in the gradient in their method, Aji [11] sparsified the absolute smallest gradients and Wen [12] essentially performed the same quantization as [10], but added a degree of sparsification by performing ternary quantization, allowing the the gradient to be sent to -1, 0, or 1. Stich [13] provided a theoretical analysis of sparsification with SGD, showing it converges at the same rate as vanilla SGD when error feedback is equipped. Albasyoni [14] showed the theoretical variance bounds of sparse dithering in both the biased and unbiased cases (deterministic versus stochastic) of quantization.

To this point, the compression methods we have discussed assumed iid data, even if their evaluation was done in a federated framework. Sattler [5] pointed out that many of the previously mentioned methods perform poorly on non-iid data, thus making them inapplicable in a more realistic FL setting which almost inevitably has non-iid data. The

authors then presented an algorithm that combined top- $k$  sparsification, SignSGD [15] with error feedback, and ternary quantization [11], [12] and showed that it worked well on non-iid data. Furthermore, it outperformed the previous methods, making it ideal for use in a FL setting. Additionally, [5] enabled both server-to-client and client-to-server compression, further decreasing the required data transmission during learning.

After this point in time, most research into gradient compression combined a form of quantization with sparsification in a novel way. Several researchers investigated how to adaptively apply compression depending on characteristics of the gradient or of the learning process. Tsuzuku [16] analyzed the relative amplitude and variance of each gradient value and only relayed those values with high amplitude and low variance. Chen [17] analyzed the residuals relative to the maximum residual, sending to zero smaller ratios and quantizing the rest. Luo [18] adaptively selected compression rates for each layer based on the number of parameters in each layer, but the method required many hyperparameters, similar to AdaComp [17]. Yang [19] adjusted the compression rate by client depending on the non-iid and data distribution of the client relative to the overall system. Liu [20] adaptively varies the compression rate for each client based on their computational power and adjusts compression rate by round based on the changing gradient norms. Mitchel [21] introduced a computationally efficient method based on a hyperparameter step size that performed quantization and sparsification combined with a novel loss-less encoding.

In more recent years, researchers have introduced methods that consider the distribution of the gradients. The  $k$ -means algorithm, commonly used for clustering, has been successfully applied as a compression algorithm that better captures distribution than simple quantization. ClusterGrad [22], FedZIP [1], and FedACQ [23] apply the  $k$ -means to quantize gradients. ClusterGrad selects a subset of gradients determined by relative magnitude that are then quantized using  $k$ -means [22]. FedZIP sequentially performs top- $k$  sparsification and  $k$ -means clustering on each layer, but requires as hyperparameters the sparsification rate and the number of clusters [1]. FedACQ greedily finds the optimal number of clusters for each layer and applies  $k$ -means clustering [23].

FetchSGD [24] and Intrinsic Gradient Compression [25] apply sketching matrices to parameter updates to reduce dimensionality. FetchSGD reduces the dimensionality of parameter updates by applying a sketching matrix and then further compresses by applying top- $k$  sparsification. In order to maintain a client statelessness, the authors take advantage of the nature of the chosen sketching matrix to apply momentum and error accumulation at the server. They provide theoretical convergence guarantees of their method [24]. Intrinsic Gradient Compression determines the intrinsic dimension  $d$  of the minimization problem and applies a sketching matrix to gradient updates to compress to that smaller  $d$  dimensional space. They provide theoretical guarantees of performance and show their method outperforms methods such as FetchSGD [24], [25].

Researchers have created methods that apply a pre-training or warm-up step [26], [27]. There are methods that analyze the variance or relative magnitude of gradient update vectors as a means to select values to sparsify or how to quantize [9], [12], [16], [18]. Many methods (to include some of those mentioned previously) apply a state at the clients that allows for error accumulation or momentum [8], [9], [28].

## 2.2 Sparse K-means (SparK)

SparK follows the lead of [1], [5], and [21] by combining quantization, through  $k$ -means clustering, with sparsification. Unlike previous methods, SparK is unique in that it requires only a single hyperparameter,  $\tau$ , which is the inverse compression rate desired by the researcher. This is a unique feature of SparK because it allows its practitioners to dictate the amount of compression the algorithm will achieve prior to its implementation.

SparK compresses model parameter updates by balancing the number of quantization bits and sparsification rate to achieve the desired  $\tau$  compression rate at each layer. The use of  $k$ -means in the quantization step allows SparK to better capture the distribution of the gradient as compared to simple quantization based on bin or step sizes. Additionally, by performing the sparsification based on both a gradient’s distance from its assigned centroid

and its magnitude, SparK allows for the possibility of retaining (not sparsifying) small, but tightly clustered parameters if their retention would lead to a smaller distortion.

SparK takes much of its motivation from the algorithm offered by Mitchell [21], but it offers two main advantages over the algorithm: 1) the hyperparameter,  $\Delta$ , in the google-researchers' algorithm provides no insight into the amount of compression that will be achieved prior to its implementation. An empirical evaluation of multiple  $\Delta$  values must be performed to achieve a desired compression rate. SparK instead takes as input the desired compression rate (as the inverse compression rate  $\tau$ ) as its only hyperparameter. 2) Using step-size quantization with a single step-size across all layers results in the same sparsification threshold. By this we mean that given a step size  $\Delta$ , the quantized-encode algorithm in [21] will send to zero any value absolutely less than some  $|a|$  and this is applied across layers. SparK considers the distribution of values within each layer to identify the optimal clustering centroids for each layer. It then sparsifies values in each layer to maintain the overall compression budget.

### 2.2.1 SparK Description

A formal description of SparK is provided in Algorithm 2 below.

---

**Algorithm 2** Sparsified  $k$ -means

---

**Require:**  $\mathbf{x} \in \mathbb{R}^n$ , inverse compression rate  $0 < \tau \leq 1$

- 1:  $b \leftarrow 1$
  - 2: **while**  $b \leq b_{\max}$  **do**
  - 3:    $k \leftarrow 2^b$
  - 4:   Solve  $\left\{ \min_{\{\theta_\ell\}_{\ell=1}^k} \sum_{i=1}^n \min_{\theta \in \{0\} \cup \{\theta_\ell\}_{\ell=1}^k} |x_i - \theta|^2 \right\}$  subject to  $B \leq \left\lfloor \frac{32(n\tau - k)}{b} \right\rfloor$  where  $B$  denotes the number of indices  $\{i : 1 \leq i \leq n\}$  assigned to the non-zero centroids  $\{\theta_\ell\}_{\ell=0}^k$
  - 5:    $\tilde{x}_i^{(b)} \leftarrow \theta_{\ell_i^*}^*$  for  $1 \leq i \leq n$  where the  $\{\theta_\ell^*\}_{\ell=1}^k$  minimize (2.7), and  $\{\ell_i^*\}_{i=1}^n$  denotes the centroid assignments
  - 6:    $\text{MSE}_b \leftarrow \frac{1}{n} \sum_{i=1}^n |\tilde{x}_i^{(b)} - x_i|^2$
  - 7:    $b \leftarrow b + 1$
  - 8: **Return**  $\mathbf{x} \leftarrow \tilde{\mathbf{x}}^{(b^*)}$  with  $b^* = \arg \min_b \text{MSE}_b$
- 

Since the amount of sparsification is enforced by  $B$  and is dependent on  $b$ , then as  $b$  increases, and thus the number of  $k$ -means centroids increases, the rate of sparsification must increase to achieve the overall desired  $\tau$  inverse compression rate. The finer granularity of quantization from larger  $k$  must be balanced by more sparsity so that the overall rate of compression remains unchanged. Accordingly, the modified  $k$ -means criterion (2.7) is no longer monotonically decreasing in the number of centroids. We search through all possible values of  $b$  up to a  $b_{\max}$  and return the values associated with the  $b$  that minimizes the MSE.

We note the modified  $k$ -means loss function in step 4 in Algorithm 2:

$$\left\{ \min_{\{\theta_\ell\}_{\ell=1}^k} \sum_{i=1}^n \min_{\theta \in \{0\} \cup \{\theta_\ell\}_{\ell=1}^k} |x_i - \theta|^2 \right\} \quad \text{subject to} \quad B \leq \left\lfloor \frac{32(n\tau - k)}{b} \right\rfloor \quad (2.7)$$

where  $B$  denotes the number of indices  $\{i : 1 \leq i \leq n\}$  assigned to the non-zero centroids  $\{\theta_\ell\}_{\ell=1}^k$ . Unlike the standard  $k$ -means problem, (2.7) imposes a constraint requiring a

number of values to be mapped to a zero centroid. This enforces sparsification. This modified objective equation (2.7) can be solved by a modification of Lloyd’s algorithm [29] which we detail in Algorithm 3.

---

**Algorithm 3** Subroutine for solving (2.7)

---

```

1: procedure MODIFIED LLOYDS( $\mathbf{x}, b, \tau$ )
2:    $k \leftarrow 2^b, n \leftarrow |\mathbf{x}|$ 
3:   Initialize  $\{\theta_\ell^{(0)}\}_{\ell=1}^k$  by performing  $k$ -means clustering on  $\mathbf{x}$ ,  $\Delta_0 \leftarrow |\mathbf{0} - \mathbf{x}|^2$ 
4:   if  $n\tau > 2^b$  then ▷ Ensures the number of parameters is large enough
5:      $t \leftarrow 0, B \leftarrow \left\lfloor \frac{32(n\tau - k)}{b} \right\rfloor$  ▷ Establish the 'budget'
6:     repeat
7:        $\Delta \leftarrow \Delta_0$ 
8:       for  $i \in \{1, \dots, n\}$  do
9:          $(\ell_i^{(t)}, d_i^2) \leftarrow \left( \arg \min_{1 \leq \ell \leq k} |x_i - \theta_\ell^{(t)}|^2, \min_{1 \leq \ell \leq k} |x_i - \theta_\ell^{(t)}|^2 \right)$ 
10:         $\delta_i^2 \leftarrow |x_i|^2, \xi_i^2 \leftarrow \max\{\delta_i^2 - d_i^2, 0\}$ 
11:        Find a root  $\pi_*^{(t)}$  of the function  $\pi \mapsto \varphi(\pi) := \sum_{1 \leq i \leq n} I(\xi_i^2 > \pi) - B$ .
12:        for  $i \in \{1, \dots, n\}$  do
13:          if  $\xi_i^2 \leq \pi_*^{(t)}$  then  $\ell_i^{(t)} \leftarrow 0$ 
14:          Update objective function:  $\Delta_0 = \sum_{i: \ell_i^{(t)} \neq 0} d_i^2 + \sum_{i: \ell_i^{(t)} = 0} \delta_i^2$ 
15:          for  $\ell \in \{1, \dots, k\}$  do
16:            if  $m = \sum_i \mathbf{I}(\ell_i^{(t)} = \ell) \neq 0$  then  $\theta_\ell^{(t+1)} \leftarrow \frac{1}{m} \sum_{i: \ell_i^{(t)} = \ell} x_i$ 
17:            else  $\theta_\ell^{(t+1)} \leftarrow \text{RandomUnif}(\{x_i : \ell_i^{(t)} \neq 0\})$ 
18:             $t \leftarrow t + 1$ 
19:          until  $\Delta - \Delta_0 < \text{tolerance}$ 
20:           $\theta_\ell^* \leftarrow \theta_\ell^{(t)}, \ell = 1, \dots, k$ 
21:           $\ell_i^* \leftarrow \ell_i^{(t)}, 1 \leq i \leq n$ 
22:          Return  $(\{\theta_\ell^*\}_{\ell=0}^k, \{\ell_i^*\}_{i=1}^n)$ 
23:        else
24:          Return  $(\mathbf{x}, \{1, \dots, n\})$ 

```

---

To describe the algorithm in detail, we first note that SparK calculates a budget  $B$  based on the current  $b$ , the hyperparameter  $\tau$ , and the size  $n$  of the input  $\mathbf{x}$  parameter vector. This is done in line 5 of algorithm 3. We assume that an uncompressed vector would be

stored as a vector of 32-bit floating point values and thus require  $32 \times n$  bits to transmit to the server. SparK requires the transmission of two vectors: 1) the centroids (or code book) of the quantization and 2) the centroid assignments or mapping. We assume that the code book will be comprised of  $k = 2^b$  centroids stored as 32-bit floating point values. The assignments will consist of  $B$  non-zero values stored as  $b$ -bit integer values. Thus, SparK requires the transmission of  $32 \times k$  bits plus  $B \times b$  bits. By finding the difference between the uncompressed number of bits with SparK's required bits, we can determine the number of non-zero values allowed ( $B$ ) for each choice of  $b$  to achieve the desired  $\tau$ .

Once the budget is established, SparK then calculates  $\xi$ . The value  $\xi$  is the maximum of 0 and the difference between each squared value in  $\mathbf{x}$  and that value's squared distance from its assigned centroid. Thus,  $\xi$  is determining if each value is either close enough to its centroid or large enough in magnitude to be retained (not sparsified). By then ranking the values according to their  $\xi$  and sparsifying all but the  $B$  with the largest  $\xi$ , we take steps to minimize the distortion between the SparK compressed vector and the original  $\mathbf{x}$ .

SparK, like other clustering methods, introduces the possibility of creating clusters to which no values get assigned, especially if all values in a single cluster are sparsified. In standard  $k$ -means, when a centroid is not assigned any values, it is usually handled in one of three ways:

1. setting the quantization bin  $\theta_\ell^{(t)} \leftarrow 0$ , thereby reducing the number of bins.
2. leaving the quantization bin value unchanged, thereby keeping the number of bins, but may result in bins without any values assigned at the end.
3. randomly selecting a value from  $x_i \in \{x_i\}_{i=1}^n$  and assigning  $\theta_\ell^{(t)} \leftarrow x_i$ , keeping the number of bins and guaranteeing that all bins have at least one value assigned.

We chose to use a modification of 3. We uniformly randomly select a value from the set  $\{x_i : \ell_i^{(t)} \neq 0\}$ . That is, we uniformly randomly select a value from the values not assigned

to the zero-centroid. The selected value then becomes the centroid value of the formerly empty quantization bin  $\theta_\ell^{(t)}$ .

### 2.2.2 SparK on Simulated Data

To demonstrate SparK, we created a vector of 1250 values by generating random normal samples using the parameters in table 2.1. These parameters generated a vector similar to what we would expect to see in an update sent by a client in a federated learning framework: a few large values and many near-zero values.

Table 2.1: Parameters to Create Simulated Data.

Num Points	Mean	Std Dev
30	$\sqrt{0.2}$	0.005
70	$\sqrt{0.005}$	0.002
150	$\sqrt{0.0005}$	0.0005
900	0	0.004

Figure 2.2 shows the sorted data in blue and the associated means values are the horizontal green lines at the same location.

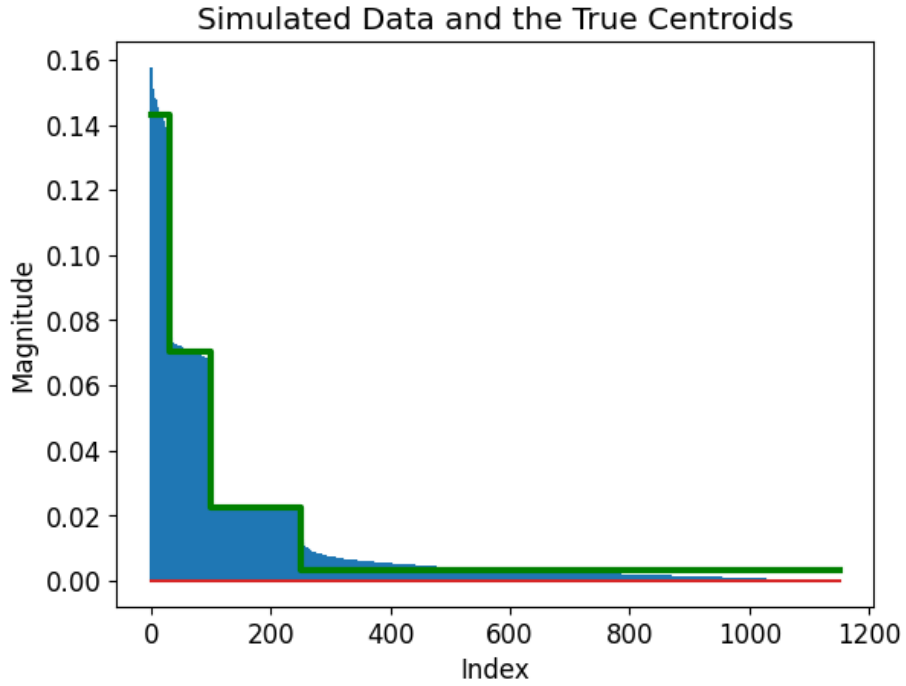


Figure 2.2: Sorted Simulated Data and the Means Used to Create It.

We then applied SparK to the simulated data with varying  $\tau$  values. We show four of the outputs in Figure 2.3. SparK chose to apply four centroids in subfigures 2.3a, 2.3b, and 2.3c, while in subfigure 2.3d it chose two centroids. In all cases, the smallest in magnitude values were sent to the zero centroid in combination with the quantization to achieve the desired compression rate. We also see that in subfigures 2.3a and 2.3c SparK determined it was more optimal to keep four centroids and split true centroids into two rather than reducing the number of centroids to two. It is this kind of choosing behavior that separates SparK from other methods such as FedZip [1] and [21] which must set either the number of centroids or a step size as hyperparameters.

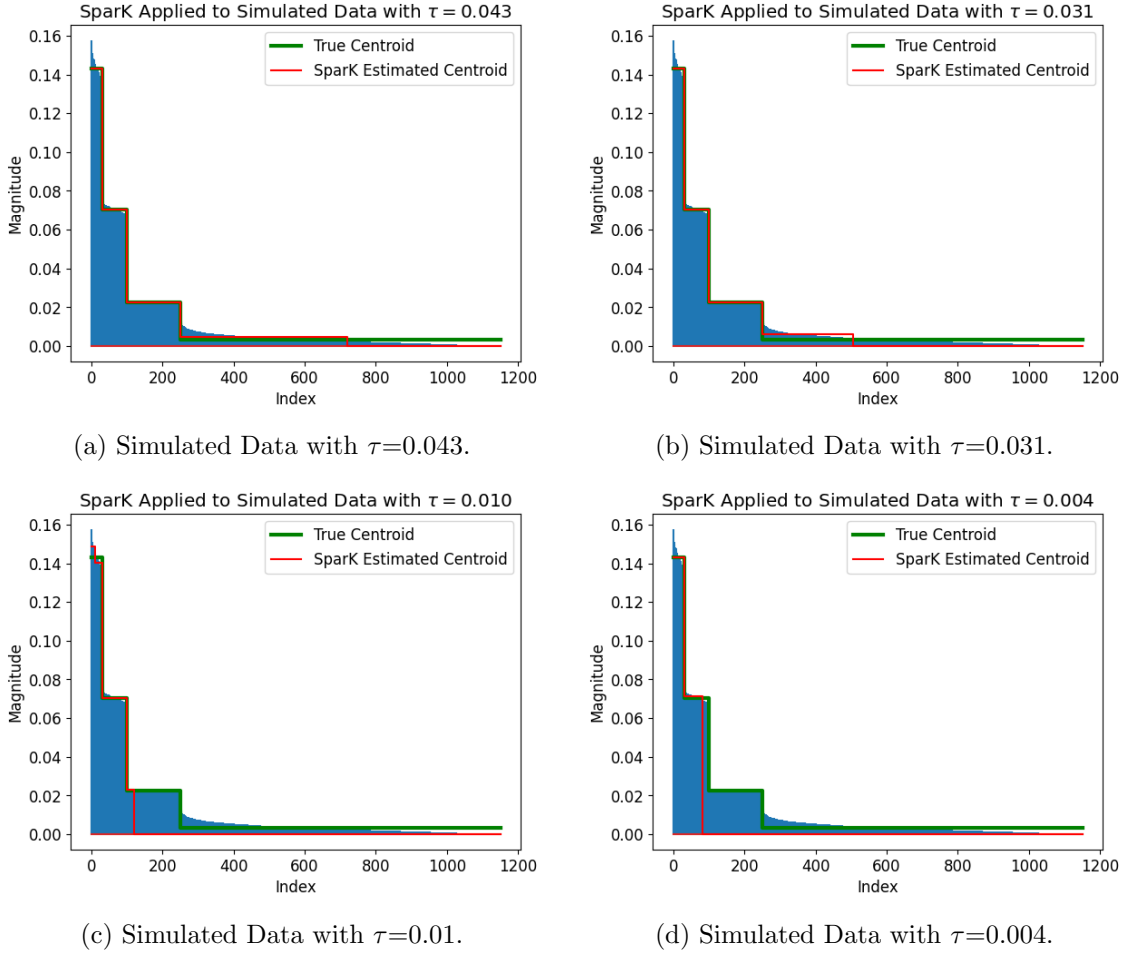


Figure 2.3: Spark Applied to Simulated Data at Varying  $\tau$ .

### 2.2.3 SparK Theory

We begin by stating a theorem:

**Theorem 2.2.1.** *Each iteration of Algorithm 3 reduces the objective (2.7), and the constraint in (2.7) remains satisfied. Furthermore, each iteration of Algorithm 3 can be run in  $\mathcal{O}(n)$  time.*

**Proof of Theorem 2.2.1**

We first note that the modified  $k$ -means problem (2.7) can be expressed as the optimization problem

$$\min_{\boldsymbol{\theta}, \boldsymbol{\ell}} f(\boldsymbol{\theta}, \boldsymbol{\ell}) \quad \text{subject to} \quad \sum_{i=1}^n \mathbf{I}(\ell_i \neq 0) \leq B$$

$$\text{where } f(\boldsymbol{\theta}, \boldsymbol{\ell}) := \sum_{i=1}^n |x_i - \theta_{\ell_i}|^2, \quad \boldsymbol{\theta} := \{\theta_{\ell}\}_{\ell=1}^n, \quad \theta_0 := 0,$$

$$\boldsymbol{\ell} := \{\ell_i\}_{i=1}^n \in \{0, 1, \dots, k\}^n, \quad B := \left\lfloor \frac{32(n\tau - k)}{b} \right\rfloor$$

We can then show that algorithm 3 can be viewed as a block coordinate descent for minimizing (2.7) by alternating between updates of  $\boldsymbol{\theta}$  given  $\boldsymbol{\ell}$  and vice versa:

$$\boldsymbol{\ell}^{(t+1)} \leftarrow \arg \min_{\boldsymbol{\ell}} f(\boldsymbol{\theta}^{(t)}, \boldsymbol{\ell}^{(t)}) \quad \text{subject to} \quad \sum_{i=1}^n \mathbf{I}(\ell_i \neq 0) \leq B \quad (2.8)$$

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \arg \min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(t)}, \boldsymbol{\ell}^{(t+1)}) \quad (2.9)$$

Let us first examine update (2.9). Letting  $\tilde{\ell}_i = \arg \min_{1 \leq \ell_i \leq k} |x_i - \theta_{\ell_i}|$ ,  $1 \leq i \leq n$ , observe that we can rewrite the loss function for a given round as

$$f(\boldsymbol{\theta}^{(t)}, \boldsymbol{\ell}) = \sum_{i=1}^n |x_i|^2 + \sum_{i=1}^n \left( |x_i - \theta_{\tilde{\ell}_i}^{(t)}|^2 - |x_i|^2 \right)_+ = \sum_{i=1}^n \delta_i^2 + \sum_{i=1}^n \xi_i^2,$$

$$\text{where } \delta_i := |x_i|, \quad \xi_i^2 := \{\delta_i^2 - d_i^2, 0\}_+, \quad d_i^2 := |x_i - \theta_{\tilde{\ell}_i}^{(t)}|^2, \quad 1 \leq i \leq n.$$

Thus, the optimization problem

$$\min_{\boldsymbol{\ell}} f(\boldsymbol{\theta}^{(t)}, \boldsymbol{\ell}) \quad \text{subject to} \quad \sum_{i=1}^n \mathbf{I}(\ell_i \neq 0) \leq B \quad (2.10)$$

is equivalent to

$$\min_{\ell} \sum_{i=1}^n \xi_i^2 \mathbf{I}(\ell_i = 0) \quad \text{subject to} \quad \sum_{i=1}^n \mathbf{I}(\ell_i \neq 0) \leq B, \quad \ell_i = \tilde{\ell}_i \text{ if } \ell_i \neq 0. \quad (2.11)$$

We can solve the last problem by finding the  $B$ -th largest value among  $\{\xi_i^2\}_{i=1}^n$  and assigning  $\ell_i^{(t)} = 0$  for all indices  $i$  whose  $\xi_i^2$  is below that value. This is equivalent to finding a root  $\pi_*$  for the equation

$$\varphi(\pi) = \sum_{i=1}^n \mathbf{I}(\xi_i^2 > \pi) - B \quad (2.12)$$

and setting  $\ell_i^{(t)} = 0$  whenever  $\xi_i^2 \leq \pi_*$ . This can be accomplished by bi-section search in time  $\mathcal{O}(n)$  even though  $\varphi$  is not continuous. This follows from  $\varphi$  being a non-increasing step function that has at least one root. All other operations in update (2.9) can also be performed in  $\mathcal{O}(n)$  time. The entirety of this update is done within the first half of the repeat loop of algorithm 3.

The update (2.8) can be solved by finding the average of the values assigned to each centroid. This operation can be done in  $\mathcal{O}(n)$  time and is performed in the last for loop within the repeat loop in algorithm 3. Note the loop accounts for the scenario when all values in a centroids are re-assigned to the zero centroid or when the centroid is not assigned any values initially.

Thus, Algorithm 3 is equivalent to a coordinate descent scheme and can be done in  $\mathcal{O}(n)$  time.

#### 2.2.4 Evaluating the Use of $\xi$

**Theorem 2.2.2.** *Let  $\mathbf{x} = \{x_i\}_{i=1}^n$  be a vector of real values. Define  $\{\theta_0\} \cup \{\theta_\ell^{(t)}\}_{\ell=1}^k$  to be the union of  $\theta_0 = 0$ , a fixed zero centroid, and  $\{\theta_\ell^{(t)}\}_{\ell=1}^k$ , the centroids from the  $t$ -th iteration of the  $k$ -means algorithm. Suppose  $\tilde{\ell}_i^{(t)}$  for  $i = 1, \dots, n$  are the assignments of the vector*

values to the centroids in the  $t$ -th iteration of the algorithm. Then for

$$\xi_i^2(x_i, \theta_\ell) := \max\{|x_i|^2 - |x_i - \theta_\ell|^2, 0\} \quad (2.13)$$

we have that  $\xi_i^2(x_i, \theta_{\tilde{\ell}_i^{(t)}}) > 0$  for all  $i = 1, \dots, n$  and, furthermore, that  $\xi_i^2 > \xi_j^2$  when  $|x_i|^2 > |x_j|^2$ .

### Proof of Theorem 2.2.2

We first show that  $\xi_i^2(x_i, \theta_{\tilde{\ell}_i^{(t)}}) > 0$  for all  $i = 1, \dots, n$ .

Let  $\mathbf{x} = \{x_i\}_{i=1}^n$  be a vector of real values. Define  $\{\theta_0\} \cup \{\theta_\ell^{(t)}\}_{\ell=1}^k$  to be the union of  $\theta_0 = 0$ , a fixed zero centroid, and  $\{\theta_\ell^{(t)}\}_{\ell=1}^k$ , the centroids from the  $t$ -th iteration of the  $k$ -means algorithm. Let  $\tilde{\ell}^{(t)} = \{\tilde{\ell}_i^{(t)}\}_{i=1}^n$  designate the assignments of each  $x_i$  to the  $\ell_i^{(t)}$  centroid  $\theta_\ell$  during the  $t$ -th iteration of the same  $k$ -means algorithm. In the proceeding, we drop  $t$  for simplicity of notation.

Let us first sort the centroids such that  $\theta_j^{(t)} = \theta_{(i)}^{(t)}$  where  $(i)$  represent the  $i$ -th order statistic for  $i, j = 1, \dots, k$ . Consider an  $x_i$  such that  $\tilde{\ell}_i = p + 1$  where  $p + 1$  is not the index of the zero centroid. Suppose without loss of generality that  $\theta_p \geq 0$  and  $\theta_p < \theta_{p+1} < \theta_{p+2}$ . Then by the  $k$ -means algorithm we know  $\theta_p < x_i < \theta_{p+2}$  and

$$|x_i - \theta_{p+1}|^2 < |x_i - \theta_p|^2 \implies x_i > \frac{\theta_{p+1} + \theta_p}{2} \text{ when } \theta_p < x_i < \theta_{p+1} \quad (2.14)$$

$$|x_i - \theta_{p+1}|^2 < |x_i - \theta_{p+2}|^2 \implies x_i < \frac{\theta_{p+1} + \theta_{p+2}}{2} \text{ when } \theta_{p+1} < x_i < \theta_{p+2}. \quad (2.15)$$

We consider these as two scenarios: scenario (2.14):  $\theta_p < x_i < \theta_{p+1}$  and scenario (2.15):  $\theta_{p+1} < x_i < \theta_{p+2}$ .

Now suppose  $\xi(x_i, \theta_{p+1}) := \max\{|x_i|^2 - |x_i - \theta_{p+1}|^2, 0\} = 0$ . This implies that

$$|x_i|^2 < |x_i - \theta_{p+1}|^2 \implies x_i < \frac{\theta_{p+1}}{2} \quad (2.16)$$

By this we can say that only values with magnitudes less than half the magnitudes of their assigned centroids can have  $\xi(x_i, \theta_{p+1}) = 0$ . This means that any value which is greater in magnitude than its assigned centroid will have a positive, non-zero  $\xi$ . We can then immediately see that  $x_i$  cannot be assigned to  $\theta_{p+1}$  and also be greater than  $\theta_{p+1}$ . Thus, we rule out scenario (2.15).

Now let us consider scenario (2.14). Combining (2.14) with (2.16) results in

$$\frac{\theta_{p+1} + \theta_p}{2} < x < \frac{\theta_{p+1}}{2}. \quad (2.17)$$

For  $\theta_p > 0$ , this creates a contradiction since  $\frac{\theta_{p+1} + \theta_p}{2} > \frac{\theta_{p+1}}{2}$ . For the case when  $\theta_p = 0$ , we have  $x_i = \frac{\theta_{p+1}}{2}$ .

Thus, the only way for  $\xi_i^2(x_i, \theta_{p+1}) = 0$  is for  $x_i = \frac{\theta_{p+1}}{2}$ . The same conclusion apply for negative centroids. While  $x_i = \frac{\theta_{p+1}}{2}$  is theoretically possible, this will not occur in practice. We can thus conclude that  $\xi_i^2 > 0$  for all  $i = 1, \dots, n$ .

Now let us examine the ordering of a vector according to its  $\xi_i^2$  values.

Consider two values  $x$  and  $y$  in the gradient vector  $\mathbf{x}$  such that  $|x| < |y|$ . Define  $\xi(z, \theta_z) = \max\{0, |z|^2 - |z - \theta_z|^2\}$ . Let us define  $d_1^2 = |x - \theta_1|^2$  and  $d_2^2 = |y - \theta_2|^2$ . Suppose an iteration of the  $k$ -means algorithm assigns  $x$  to  $\theta_1$  and  $y$  to  $\theta_2$  for centroids  $\theta_1$  and  $\theta_2$ . Let us assume  $\xi(x, \theta_1) > 0$  and  $\xi(y, \theta_2) > 0$ .

1. Consider the case  $0 < \theta_1 < x < y < \theta_2$ . Since we assume that  $y$  is assigned to the  $\theta_2$  centroid, then we have that  $d_2^2 < |y - \theta_1|^2$ , otherwise it would be assigned to the  $\theta_1$

centroid. For the sake of contradiction, let us assume that  $\xi(x, \theta_1) > \xi(y, \theta_2)$ . Then

$$\xi(x, \theta_1) > \xi(y, \theta_2)$$

$$x^2 - d_1^2 > y^2 - d_2^2 > y^2 - |y - \theta_1|^2$$

$$x^2 - |x - \theta_1|^2 > y^2 - |y - \theta_1|^2$$

$$x^2 - x^2 + 2\theta_1 x - \theta_1^2 > y^2 - y^2 + 2\theta_1 y - \theta_1^2$$

$$2\theta_1 x > 2\theta_1 y$$

$$x > y \tag{2.18}$$

This is a contradiction since we assumed  $0 < \theta_1 < x < y < \theta_2$ . Thus,  $\xi(x, \theta_1) < \xi(y, \theta_2)$ .

2. Consider the case  $\theta_2 < y < x < \theta_1 < 0$ . Since we assume that  $y$  is assigned to the  $\theta_2$  centroid, then we have that  $d_2^2 < |y - \theta_1|^2$ . For the sake of contradiction, let us assume that  $\xi(x, \theta_1) > \xi(y, \theta_2)$ .

$$\xi(x, \theta_1) > \xi(y, \theta_2)$$

$$x^2 - d_1^2 > y^2 - d_2^2 > y^2 - |y - \theta_1|^2$$

$$x^2 - |x - \theta_1|^2 > y^2 - |y - \theta_1|^2$$

$$x^2 - x^2 + 2\theta_1 x - \theta_1^2 > y^2 - y^2 + 2\theta_1 y - \theta_1^2$$

$$2\theta_1 x > 2\theta_1 y$$

$$x > y$$

This is a contradiction since we assumed  $\theta_2 < y < x < \theta_1 < 0$ . Thus,  $\xi(x, \theta_1) < \xi(y, \theta_2)$ .

3. For cases  $0 < \theta_1 < x < \theta_2 < y$ ,  $0 < x < \theta_1 < \theta_2 < y$ ,  $0 < x < \theta_1 < y < \theta_2$ , and each of their negative complements (maintaining  $|x| < |y|$ ), the same method of proof can be applied. By using (2.18), we can show a contradiction by assuming  $\xi(x, \theta_1) > \xi(y, \theta_2)$ . Thus, we conclude that the ordering of values by  $\xi$  across centroids assignments is the same as ordering them by their magnitude.

Let us now consider the case when  $x$  and  $y$  are both assigned to the same centroid  $\theta$ . Let us assume  $\xi(x, \theta) > 0$  and  $\xi(y, \theta) > 0$ . Consider the case when  $0 < x < \theta < y$ . Suppose for the sake of contradiction that  $\xi(x, \theta) > \xi(y, \theta)$ . Then

$$\xi(x, \theta) > \xi(y, \theta)$$

$$x^2 - d_1^2 > y^2 - d_2^2$$

$$x^2 - |x - \theta|^2 > y^2 - |y - \theta|^2$$

$$x^2 - x^2 + 2\theta x - \theta^2 > y^2 - y^2 + 2\theta y - \theta^2$$

$$2\theta x > 2\theta y$$

$$x > y$$

This contradicts our assumption that  $0 < x < \theta < y$ . Thus,  $\xi(x, \theta) > \xi(y, \theta)$ . The same arguments can be made for the cases  $0 < x < y < \theta$ ,  $0\theta < x < y$ , and their negative complements maintaining the  $|x| < |y|$  relationship. Thus, ordering using  $\xi$  within centroids is equivalent to ordering using magnitude.

## 2.3 Sparse K-means in Simulation

### 2.3.1 Simulation Description

Mitchell [21] performed their work using a federated simulation they created using the TensorFlow Federated (tff) package. Their simulation, combined with the tff-provided functions, enabled the creation of SparK as a plug-in aggregation factory in their simulation. As such, the experimental setup and this paper’s evaluation of SparK follows closely with their work.

#### Dataset and Model

We apply SparK to the EMNIST [30] dataset. The EMNIST dataset contains 671,585 training and 77,483 test datapoints consisting of images of written characters. The simulation consists of 3,400 clients who each have their own train and test portion of the data which has natural heterogeneity stemming from the data being partitioned by the author of the written digits. The number of samples each client has varies, but all have more than two training samples and one test sample.

The server and each client train a convolutional neural network with two convolutional layers (with  $3 \times 3$  kernels with a stride of 1), a max-pooling layer, a dropout layer ( $p = 0.25$ ), and two dense layers with dropout ( $p = 0.5$ ). This setup is the same as in [31].

#### Algorithms

FedAvg [6] is used in the training on the EMNIST dataset. Mini-batch SGD with client learning rate  $\eta_c = 0.1$  is performed for  $E$  epochs on a uniformly randomly selected  $m = 50$  clients during each global training round. The server model is updated with SGD with server learning rate  $\eta_s = 1.0$ . The value  $E = 1$  and a batch size of 32 is used on all clients throughout training.

## Benchmarks

We evaluate the performance of SparK against a baseline model without compression. We then compare its performance at varying compression levels against Top-K [11], QSGD [10], and the algorithm presented by [21], which we refer to as quantize-encode. The hyperparameters used for these algorithms match those chosen by Mitchell [21] in order to provide a fair comparison of SparK to the results from their work.

### 2.3.2 Results

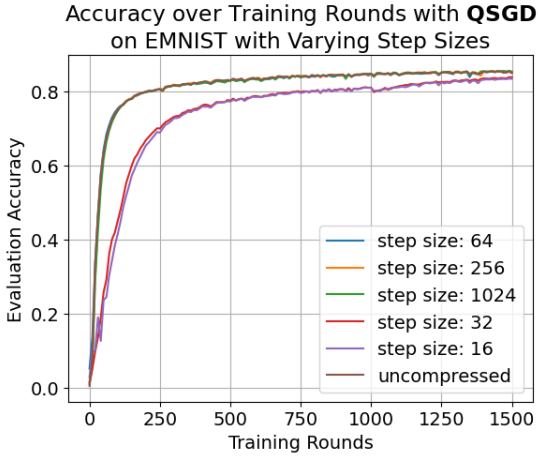
Figure 2.4 and its subfigures displays the validation accuracy across training rounds on the EMNIST dataset for all four compared compression algorithms. We note that SparK performs well, outperforming the other algorithms in many cases, but performs comparably otherwise.

We attempt to compare the algorithms directly in figure 2.5 by showing the average bitrate achieved over all training rounds on the x-axis and the evaluation accuracy achieved on the final round. A minimum of five simulations were performed at each set of hyperparameter values and their results are summarized together on this graph using averages.

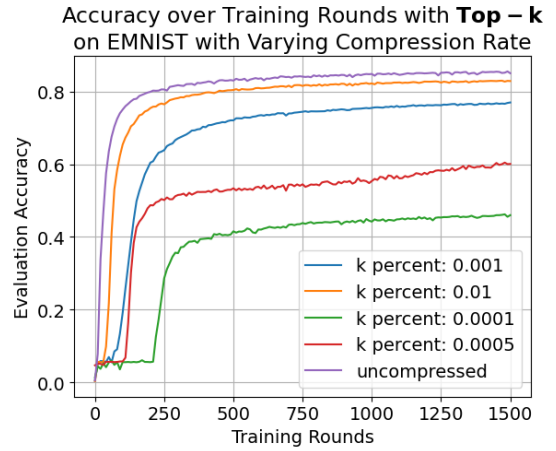
The points themselves represent the average of the average bitrates and the average final accuracies from simulations with the same parameters. The horizontal lines through points represent the range of average bitrates and the vertical lines through points represent the range of final accuracies achieved from simulations with the same hyperparameters. SparK outperforms other algorithms at almost all levels of compression, and it continues to perform adequately even at very low bitrates. We omit lower bitrates for other algorithms as they perform too poorly to display on this graph.

## 2.4 Discussion

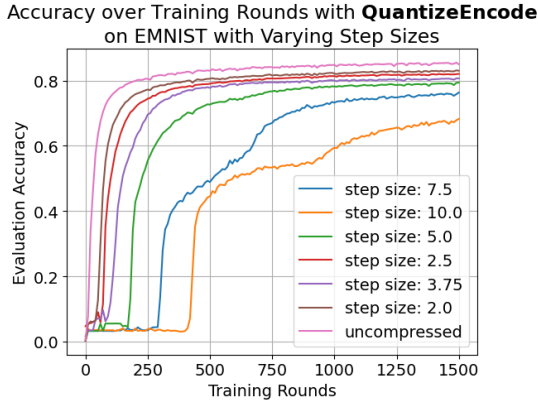
Figure 2.5 clearly shows that SparK performs or outperforms the other algorithms, especially at very high compression. Its current implementation in tensorflow federated requires very



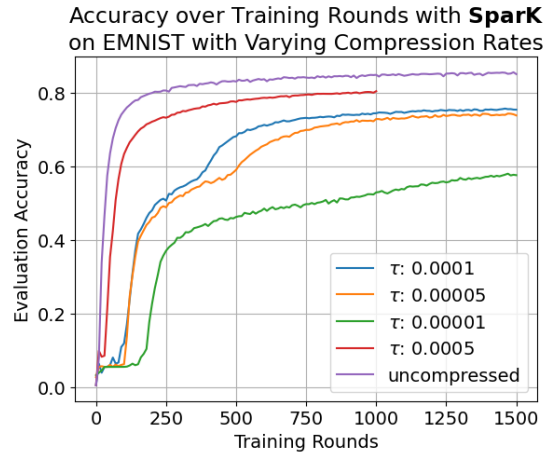
(a) QSGD Validation Accuracy on EMNIST



(b) Top- $k$  Validation Accuracy on EMNIST



(c) QE Validation Accuracy on EMNIST



(d) SparK Validation Accuracy on EMNIST

Figure 2.4: Validation Accuracy across Training Rounds on EMNIST

long computation time and system memory making its practical use cases limited. However, SparK offers an algorithm that performs well under extreme compression in situations where clients have high computational power and large amounts of memory, but the bandwidth to and from the server is extremely limited. In these cases, SparK provides the user the ability to compress the updates to meet the bandwidth requirement without having to perform an hyperparameter tuning.

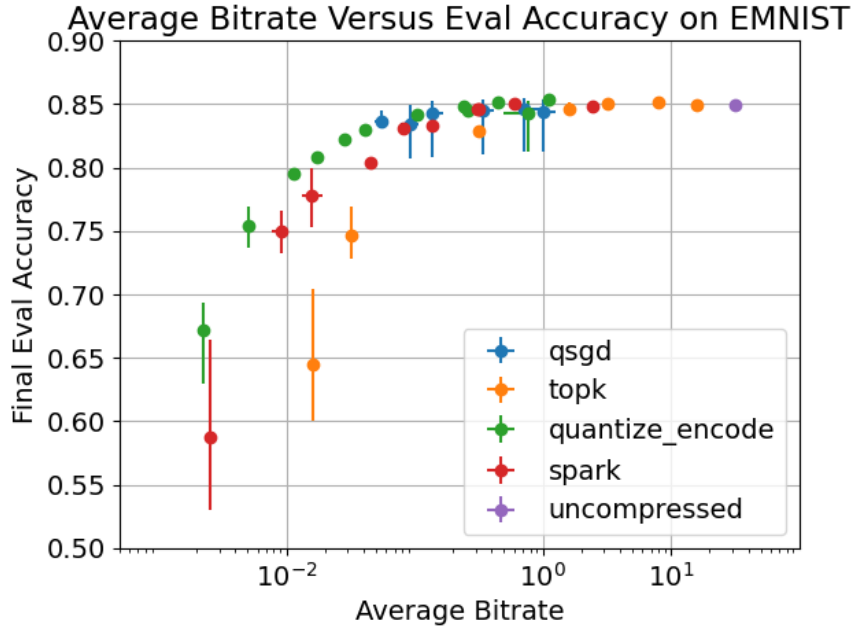


Figure 2.5: Average Bitrates and Final Accuracies for all Compression Algorithms.

We imagine the clients in a federated learning environment to be research vessels conducting research across the world in remote areas. These vessels have substantial computational abilities, but their connection to a centralized server is only through a periodic satellite link. The connection time to the satellite can be calculated fairly accurately, and the connection speed is also known. Using this information, the researcher can calculate the amount of data that can be sent in this connectivity window. Supposing that uploading the uncompressed updates would exceed the satellite connectivity time window, a compression algorithm would be required.

SparK is the sole compression algorithm that guarantees the required amount of compression is achieved without hyperparameter tuning. SparK only requires the  $\tau$  hyperparameter, which is the inverse compression rate. By using the known uncompressed update size and the calculated amount of data that can be sent in the satellite uplink window, the researcher can calculate the hyperparameter  $\tau$ . Expensive and time consuming hyperparameter tuning is not required. This provides the researcher with the immediate knowledge of how much compression will be achieved before implementation. SparK guarantees the

researcher will achieve the desired amount of compression and successfully transmit the update in the time window. To our knowledge, no other algorithm provides this benefit.

### 2.4.1 Future Work

- **Adaptive Compression Rates.** In this paper we assume that compression is required due to limited bandwidth between clients and the server in a federated environment. The same hyperparameter  $\tau$  is used by all clients regardless of their connectivity. In a real-world scenario, it is likely that clients would have different connectivity speeds. A client with better connectivity would be capable of transmitting a larger update than a client with a worse connection in the same amount of time. Future work should explore allowing the compression rate to vary based on the connectivity speed of each client and a maximum time allowed for transmitting an update. The hyperparameter  $\tau$  can then be the maximum inverse compression rate that a client must be able to achieve given its connection and the transmission time requirement. This permits each client to individually determine its optimal compression rate within the  $\tau$  maximum and maximum transmission time.

Additionally, future work can explore varying the  $\tau$  inverse compression rate across epochs. A smaller compression rate (larger  $\tau$ ) in the early stages of learning and larger compression rate later should result in faster and more robust convergence. This idea could be explored in both the current framework and in the by-client, transmission-speed-dependent compression described above.

Adaptive compression can also be performed within the model by varying the amount of compression by layer. By analyzing characteristics of each layer’s model updates (variance, magnitude, effect on loss function, etc.), one can vary the amount of compression applied to each layer. The  $\tau$  inverse compression rate can still be achieved by applying greater compression to less important layers and less compression to more important layers.

- **Learning Rate Optimization and Scheduling.** The current implementation of SparK uses a static learning rate across epochs and for all clients. Implementing a learning rate schedule that changes the learning rate for both clients and the server across epochs would likely improve the rate of convergence and the final accuracy of SparK. Future work should also explore how to determine the optimal learning rates for each client based on their individual characteristics (amount of data, heterogeneity, etc.).
- **Bidirectional Compression.** Some research has explored compression applied to both the upstream and downstream communications in federated learning; that is, both client-to-server and server-to-client [12], [15], [32]–[34]. It is unknown how compressing the server-to-client communication would affect SparK. Future research into this would be enlightening, especially as the upload and download speeds of a connection are often very different.
- **Error Feedback and Accumulation.** Research has shown that adding error accumulation and feedback in a distributed learning setting improves learning [13], [28]. However, it has only been recently that error feedback has been studied in a federated setting [32], [35]–[37]. Future work should explore incorporating error feedback to improve convergence and accuracy.
- **Algorithm Optimization and Generalization.** The current implementation of SparK is computationally expensive and requires large amounts of memory at the client. This is largely due to the algorithm being implemented in Tensorflow Federated and the framework built by [21]. At the time of this writing, Tensorflow Federated is still in pre-release. Future implementations will likely offer improved performance and customization that will likely improve the runtime and memory requirements of this algorithm if it is adapted to the updates. By decreasing its computational requirements it would allow the algorithm to be implemented on a broader range of clients.

In addition to coding optimization and Tensorflow Federated updates, SparK could also benefit from research into additional methods of performing both its quantization and sparsification. Currently, SparK uses a vanilla  $k$ -means quantization (with a novel budget restraint). Tian, et. al [23] provides a nice summary of alternative clustering algorithms that future work should explore as alternatives to  $k$ -means. As for sparsification, Top-k sparsification [38] uses only the magnitude to determine which values to sparsify. Future work should explore alternative methods of selecting the values to sparsify.

## Chapter 3: 1-Bit Quantized Regression

### 3.1 Introduction

#### 3.1.1 Background

The Internet of Things (IoT), often comprised of small, low-power, low-bandwidth, and low-computational-power devices; offers the opportunity to learn a great deal about the world from the data these devices collect. Because of the limited power, bandwidth, or computation power of many IoT devices, they often must reduce the data they store, transmit, and compute. This requires statistical methods capable of inferring meaningful insights from this "simplified" data.

Quantization is a common method employed in signal processing to turn a digital signal into an analog signal, or a continuous data value to an integer value. By applying quantization, the user reduces the size of the data from the standard 32 or 64 bit floating point precision to  $b$ -bits, called the bit rate or bit precision. This allows for a  $n \times d$  matrix of data stored in 32 bit precision to be reduced from  $32nd$  bits of storage or bandwidth requirement to  $bnd + 32b$ .

In more recent years, quantization has become popular to reduce the data storage requirement during neural network training [10], [23], [34], [39]–[41]. The research into the effects of quantization on linear models, and linear regression in particular, is more sparse.

In [42] we see a maximum likelihood method created from an infinite quantizer with  $L_n$  quantization levels to estimate the parameters in a linear model. The work provides an analysis of the error from using a uniform deterministic quantizer, and then derives a bias-compensated estimator under Gaussian assumptions that is asymptotically consistent and provide simulations showing the attainment of the expected asymptotic variance.

In [43], a randomized EM algorithm is used to estimate the parameter in a linear regression model given the quantized response. They provide simulations that show the parameter variance and estimated parameters, but do not provide formal analysis.

The work by Saha et.al. [44] takes an alternative approach. They seek to find the optimal bit rate of quantization given some budget constraint. They begin by imposing a budget constraint, which is the total number of bits that each sample is permitted to use to convey its information to the model. They then solve a least squares solution on the quantized data resulting from the enforced budget. This information-theoretic lens differs slightly from our approach, but the analysis in the work provides a lower bound on the minimax risk of estimating the parameters from a quantized model. The work provides several quantization methods, those methods' lower bound on their risk based on a provided budget, and the computation complexity of the algorithm [44].

In [45], they analyze the effect of both a deterministic and dithered 1-bit quantizer on the non-asymptotic bound of a masked estimator of the covariance matrix of a design matrix. We seek to perform linear regression on 1-bit quantized data using a stochastic quantizer. They then use numerical simulations to show the near-optimality of their estimators.

Our work resembles the work of [45] in that we seek a non-asymptotic error bound on the covariance matrix of a design matrix. However, our work differs in a few key ways. In [45], they use a 1-bit quantizer that simply returns the sign of the input value. Ours returns the right and left bound of the range of the provided data point. Also, their analysis is performed on masked covariance matrices. Our uses the original. We will provide full explanations of our methods and definitions in proceeding sections.

In summary, our work offers an estimator of the regression parameters  $\hat{\beta}$  derived from quantized predictors and responses. From this estimator, we provide an analysis of the estimator's asymptotic variance, and then compare this estimator's asymptotic variance to that of the Ordinary Least Squares (OLS) estimator based on uncompressed data. We then discuss an upper bound on the MSE of the estimator.

### 3.1.2 Quantization

We begin by defining a quantization method. We define a generic 1-bit stochastic quantizer for a bounded random variable  $U$  to be

$$Q_U(U) := \begin{cases} a & \text{with probability } \frac{b-U}{\Delta_U} \\ b & \text{with probability } \frac{U-a}{\Delta_U} \end{cases} \quad (3.1)$$

where  $Q_U(u)$  is defined on the range of  $U : [a, b]$  and  $\Delta_U := b - a$ . Equivalently, we could also define it

$$Q_U(U) := \begin{cases} a & \text{if } U + \xi < t \\ b & \text{if } U + \xi > t \end{cases} \quad (3.2)$$

where  $\xi \sim \text{Uniform}\left(\frac{-\Delta_U}{2}, \frac{\Delta_U}{2}\right)$  and  $t$  is the midpoint of the range of  $U$ .

For a quantizer of a random variable  $X$ , we note that both  $Q_X(X)$  and  $X$  are random variables. Thus, when we take the expectation omitting what the expectation is with respect to, we mean  $\mathbf{E}[Q_X(x)] = \mathbf{E}_X[\mathbf{E}_{Q_X}[Q_X(x)]]$ . We may exclude the subscript on  $Q$  for simplicity when it is understood in context so that the expectation will read  $\mathbf{E}[Q(x)] = \mathbf{E}_X[\mathbf{E}_Q[Q_X(x)]]$

Let us allow the notation

$$\tilde{X} := Q_X(X). \quad (3.3)$$

to represent the quantization of  $X$ .

## 3.2 1-bit Quantized Regression, 1-Predictor Case

We begin by establishing our regression scenario:

1. Let  $\mathbf{x} = \{X_i\}_{i=1}^n$  be a  $n \times 1$  vector of random variables such that  $|X_i| \leq R$  for all  $i$  and for some value  $R \in \mathbb{R}^+$ . We assume each observation pair  $(X_i, Y_i)$  are independent and that the  $\{X_i\}_{i=1}^n$  are independent and identically distributed (i.i.d.).
2. Let  $\mathbf{y} = \{Y_i\}_{i=1}^n$  be a  $n \times 1$  vector of random variables such that  $Y_i = X_i\beta^0 + \sigma\epsilon_i$  for some  $\beta^0 \in \mathbb{R}$ ,  $\sigma \in \mathbb{R}^+$ , and let  $\{\epsilon_i\}_{i=1}^n$  be a set of iid standard normal random variables. Assume  $|\epsilon_i| \leq L$  for all  $i$  so that  $Y_i \leq R|\beta^0| + \sigma L$  for all  $i$ . Define  $B := R|\beta^0| + \sigma L$ .
3. Let  $Q_X$ ,  $Q_{X^2}$ , and  $Q_Y$  be 1-bit quantizers as defined in section 3.1.2 which take values in  $\{X_i\}$ ,  $\{X_i^2\}$ , and  $\{Y_i\}$ , respectively. We note these quantizers are defined on  $[-R, R]$ ,  $[0, R^2]$ , and  $\{-B, B\}$ .

Rather than performing regression using the random variables  $X_i$  and  $Y_i$ , we wish to use the quantized variables  $\widetilde{X}_i := Q_X(X_i)$  and  $\widetilde{Y}_i := Q_Y(Y_i)$ . We propose the quantized regression estimator

$$\hat{\beta} = \frac{\frac{1}{n} \sum_{i=1}^n \widetilde{X}_i \widetilde{Y}_i}{\frac{1}{n} \sum_{i=1}^n \widetilde{X}_i^2} = \frac{\hat{U}}{\hat{V}} \quad (3.4)$$

since we know the Ordinary Least Squares estimator  $\beta^*$  is given by

$$\beta^* = \frac{\frac{1}{n} \sum_{i=1}^n X_i Y_i}{\frac{1}{n} \sum_{i=1}^n X_i^2} = \frac{U}{V}$$

which we know is an unbiased estimator of  $\beta^0$ . It is also important to note that

$$\mathbf{E}_Q[\hat{U}] = \mathbf{E}_Q\left[\frac{1}{n} \sum_{i=1}^n \widetilde{X}_i \widetilde{Y}_i\right] = \frac{1}{n} \sum_{i=1}^n X_i Y_i = U$$

and similarly

$$\mathbf{E}_Q[\hat{V}] = \mathbf{E}_Q\left[\frac{1}{n} \sum_{i=1}^n \widetilde{X}_i^2\right] = \frac{1}{n} \sum_{i=1}^n X_i^2 = V$$

since the stochastic 1-bit quantizer is unbiased. We then can say

$$\frac{\mathbf{E}_Q[\hat{U}]}{\mathbf{E}_Q[\hat{V}]} = \frac{U}{V} = \beta^*.$$

We will use this fact in the subsequent sections as we provide the basis to prove the theorem:

**Theorem 3.2.1.** *Given  $\mathbf{x} = \{X_i\}_{i=1}^n$  and  $\mathbf{y} = \{Y_i\}_{i=1}^n$  such that  $|X_i| \leq R$  for all  $i$  and for some value  $R \in \mathbb{R}^+$ . Let each observation pair  $(X_i, Y_i)$  be independent and each  $\{X_i\}_{i=1}^n$  be independent and identically distributed.*

*Let  $Y_i = X_i\beta^0 + \sigma\epsilon_i$  for some  $\beta^0 \in \mathbb{R}$ ,  $\sigma \in \mathbb{R}^+$ , and let  $\{\epsilon_i\}_{i=1}^n$  be a set of iid standard normal random variables. Assume  $|\epsilon_i| \leq L$  for all  $i$  so that  $Y_i \leq R|\beta^0| + \sigma L$  for all  $i$ . Define  $B := R|\beta^0| + \sigma L$ .*

*Let  $Q_X$  and  $Q_Y$  be 1-bit quantizers as defined in section 3.1.2 defined on  $[-R, R]$  and  $[-B, B]$ , respectively.*

*Then the estimator*

$$\hat{\beta} = \frac{\frac{1}{n} \sum_{i=1}^n \tilde{X}_i \tilde{Y}_i}{\frac{1}{n} \sum_{i=1}^n \tilde{X}_i^2}$$

*of  $\beta^0$  is asymptotically unbiased with a lower order  $\mathcal{O}(1/n^2)$  term and has a variance of order  $\mathcal{O}(1/n)$ . Furthermore, we can then say asymptotically that*

$$MSE(\hat{\beta}) = \frac{B^2 R^2 + \beta^{02} \left( R^2 \mathbf{E}_X[X_1^2] - 2 \mathbf{E}_X[X_1^4] \right)}{n \mathbf{E}_X[X_1^2]^2} = \mathcal{O}(1/n). \quad (3.5)$$

### 3.2.1 Asymptotic Bias of the 1-Bit Quantized Estimator

The work by Vershynin [46] is used throughout this section.

Let us write our quantized estimator as a function of  $\hat{U}$  and  $\hat{V}$ :  $\hat{\beta} = \phi(\hat{U}, \hat{V}) := \frac{\hat{U}}{\hat{V}}$ .

Now we use a Taylor Expansion about the point  $(\mathbf{E}_Q[\hat{U}], \mathbf{E}_Q[\hat{V}])$  to approximate the first moment of  $\phi(\hat{U}, \hat{V})$ . We will omit the  $Q$  in the internal expectation for brevity and understand that the external expectation is taken with respect to  $X$  and  $\epsilon$ .

$$\begin{aligned}
 \mathbf{E}[\hat{\beta}] &= \mathbf{E}[\phi(\hat{U}, \hat{V})] \\
 &= \mathbf{E}\left[\frac{\mathbf{E}[\hat{U}]}{\mathbf{E}[\hat{V}]} + \begin{pmatrix} \hat{U} - \mathbf{E}[\hat{U}] \\ \hat{V} - \mathbf{E}[\hat{V}] \end{pmatrix} \nabla\phi(\mathbf{E}[\hat{U}], \mathbf{E}[\hat{V}]) + \frac{1}{2} \begin{pmatrix} \hat{U} - \mathbf{E}[\hat{U}] \\ \hat{V} - \mathbf{E}[\hat{V}] \end{pmatrix}^T \nabla^2\phi(\bar{U}, \bar{V}) \begin{pmatrix} \hat{U} - \mathbf{E}[\hat{U}] \\ \hat{V} - \mathbf{E}[\hat{V}] \end{pmatrix}\right] \\
 &= \mathbf{E}\left[\frac{U}{\bar{V}} + \begin{pmatrix} \hat{U} - \mathbf{E}[\hat{U}] \\ \hat{V} - \mathbf{E}[\hat{V}] \end{pmatrix} \nabla\phi(\mathbf{E}[\hat{U}], \mathbf{E}[\hat{V}]) + \frac{1}{2} \begin{pmatrix} \hat{U} - \mathbf{E}[\hat{U}] \\ \hat{V} - \mathbf{E}[\hat{V}] \end{pmatrix}^T \nabla^2\phi(\bar{U}, \bar{V}) \begin{pmatrix} \hat{U} - \mathbf{E}[\hat{U}] \\ \hat{V} - \mathbf{E}[\hat{V}] \end{pmatrix}\right] \\
 &= \beta^0 + \mathbf{E}\left[\begin{pmatrix} \hat{U} - \mathbf{E}[\hat{U}] \\ \hat{V} - \mathbf{E}[\hat{V}] \end{pmatrix} \nabla\phi(\mathbf{E}[\hat{U}], \mathbf{E}[\hat{V}]) + \frac{1}{2} \begin{pmatrix} \hat{U} - \mathbf{E}[\hat{U}] \\ \hat{V} - \mathbf{E}[\hat{V}] \end{pmatrix}^T \nabla^2\phi(\bar{U}, \bar{V}) \begin{pmatrix} \hat{U} - \mathbf{E}[\hat{U}] \\ \hat{V} - \mathbf{E}[\hat{V}] \end{pmatrix}\right] \quad (3.6)
 \end{aligned}$$

where  $\bar{U} \in (\hat{U}, \mathbf{E}[\hat{U}])$  and  $\bar{V} \in (\hat{V}, \mathbf{E}[\hat{V}])$ . Note that by the Weak Law of Large numbers  $\sqrt{n}(\hat{U} - \mathbf{E}[\hat{U}]) \xrightarrow{p} 0$  and  $\sqrt{n}(\hat{V} - \mathbf{E}[\hat{V}]) \xrightarrow{p} 0$ . That is,  $\hat{U} - \mathbf{E}[\hat{U}] = \mathcal{O}_p(1/\sqrt{n})$  and  $\hat{V} - \mathbf{E}[\hat{V}] = \mathcal{O}_p(1/\sqrt{n})$ .

If we can now show that the maximum eigenvalue of  $\nabla^2\phi(\hat{U}, \hat{V})$  is upper bounded in a neighborhood around  $(\mathbf{E}[\hat{U}], \mathbf{E}[\hat{V}])$  containing  $\hat{U}$  and  $\hat{V}$ , then we can control the remainder term in the Taylor expansion.

The Hessian can be computed

$$\nabla^2\phi(\hat{U}, \hat{V}) = \begin{pmatrix} \frac{\partial^2}{\partial \hat{U}^2}\phi(\hat{U}, \hat{V}) & \frac{\partial}{\partial \hat{U}} \frac{\partial}{\partial \hat{V}}\phi(\hat{U}, \hat{V}) \\ \frac{\partial}{\partial \hat{V}} \frac{\partial}{\partial \hat{U}}\phi(\hat{U}, \hat{V}) & \frac{\partial^2}{\partial \hat{V}^2}\phi(\hat{U}, \hat{V}) \end{pmatrix} = \begin{pmatrix} 0 & -\frac{1}{\hat{V}^2} \\ -\frac{1}{\hat{V}^2} & 2\frac{\hat{U}}{\hat{V}^3} \end{pmatrix}$$

Then calculating the eigenvalues gives us

$$(-\lambda) \left( \frac{2\hat{U}}{\hat{V}^3} - \lambda \right) - \frac{1}{\hat{V}^4} = 0 \implies \lambda = \frac{\hat{U} \pm \sqrt{\hat{U}^2 + \hat{V}^2}}{\hat{V}^3}$$

Note then that  $\sqrt{\hat{U}^2 + \hat{V}^2} \leq \sqrt{\hat{U}^2} + \sqrt{\hat{V}^2} = |\hat{U}| + |\hat{V}|$  which implies  $\hat{U} + \sqrt{\hat{U}^2 + \hat{V}^2} \leq \hat{U} + |\hat{U}| + \hat{V}$ . Finally we can upper bound the maximum eigenvalue

$$\lambda = \frac{\hat{U} \pm \sqrt{\hat{U}^2 + \hat{V}^2}}{\hat{V}^3} \leq \frac{2|\hat{U}| + \hat{V}}{\hat{V}^3}$$

Then for any  $\bar{U} \in \left( \hat{U}, \mathbf{E} [\hat{U}] \right)$  and  $\bar{V} \in \left( \hat{V}, \mathbf{E} [\hat{V}] \right)$ , we can say

$$\frac{2|\hat{U}| + \hat{V}}{\hat{V}^3} = \frac{2|\hat{U}|}{\hat{V}^3} + \frac{1}{\hat{V}^2} \leq 2 \frac{\max \left\{ |\hat{U}|, \left| \mathbf{E} [\hat{U}] \right| \right\}}{\min \left\{ \hat{V}, \mathbf{E} [\hat{V}] \right\}^3} + \frac{1}{\min \left\{ \hat{V}^2, \mathbf{E} [\hat{V}]^2 \right\}} = \mathcal{O}_p(1)$$

Finally,

$$\mathbf{E} [\hat{\beta}] = \beta^0 + \mathcal{O} \left( \mathbf{E} \left[ \|T - \mathbf{E} [T]\|_2^2 \right] \right) = \beta^0 + \mathcal{O}(1/n) \quad \text{where } T := \begin{pmatrix} \hat{U} \\ \hat{V} \end{pmatrix}$$

and we can conclude that the squared bias will be a lower order  $\mathcal{O}(1/n^2)$  term.

### 3.2.2 Asymptotic Variance of the 1-bit Quantized Estimator

We begin noting that by again using a Taylor expansion and the CLT we can say that

$$\begin{aligned} & n^{1/2} \left( \phi(\hat{U}, \hat{V}) - \phi(\mathbf{E}[\hat{U}], \mathbf{E}[\hat{V}]) \right) \\ & \xrightarrow{d} N \left( 0, \nabla \phi(\mathbf{E}[\hat{U}], \mathbf{E}[\hat{V}])^T \text{Cov}(\tilde{X}_i \tilde{Y}_i, \tilde{X}_i^2) \nabla \phi(\mathbf{E}[\hat{U}], \mathbf{E}[\hat{V}]) \right) \end{aligned}$$

where by  $\nabla \phi(\mathbf{E}[\hat{U}], \mathbf{E}[\hat{V}])$  we mean  $\nabla \phi(\cdot, \cdot)$  evaluated at  $\mathbf{E}[\hat{U}]$  and  $\mathbf{E}[\hat{V}]$  and equivalently for the partial derivatives. The variance simplifies to

$$\begin{aligned} & \nabla \phi(\mathbf{E}[\hat{U}], \mathbf{E}[\hat{V}])^T \text{Cov}(\tilde{X}_i \tilde{Y}_i, \tilde{X}_i^2) \nabla \phi(\mathbf{E}[\hat{U}], \mathbf{E}[\hat{V}]) \\ & = \begin{pmatrix} \frac{\partial}{\partial \hat{U}} \phi(\mathbf{E}[\hat{U}], \mathbf{E}[\hat{V}]) \\ \frac{\partial}{\partial \hat{V}} \phi(\mathbf{E}[\hat{U}], \mathbf{E}[\hat{V}]) \end{pmatrix}^T \begin{pmatrix} \text{Var}[\tilde{X}_i \tilde{Y}_i] & \text{Cov}(\tilde{X}_i \tilde{Y}_i, \tilde{X}_i^2) \\ \text{Cov}(\tilde{X}_i \tilde{Y}_i, \tilde{X}_i^2) & \text{Var}[\tilde{X}_i^2] \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial \hat{U}} \phi(\mathbf{E}[\hat{U}], \mathbf{E}[\hat{V}]) \\ \frac{\partial}{\partial \hat{V}} \phi(\mathbf{E}[\hat{U}], \mathbf{E}[\hat{V}]) \end{pmatrix} \\ & = \begin{pmatrix} \frac{1}{\mathbf{E}[\hat{V}]} \\ \frac{-\mathbf{E}[\hat{U}]}{\mathbf{E}[\hat{V}]^2} \end{pmatrix}^T \begin{pmatrix} \text{Var}[\tilde{X}_i \tilde{Y}_i] & \text{Cov}(\tilde{X}_i \tilde{Y}_i, \tilde{X}_i^2) \\ \text{Cov}(\tilde{X}_i \tilde{Y}_i, \tilde{X}_i^2) & \text{Var}[\tilde{X}_i^2] \end{pmatrix} \begin{pmatrix} \frac{1}{\mathbf{E}[\hat{V}]} \\ \frac{-\mathbf{E}[\hat{U}]}{\mathbf{E}[\hat{V}]^2} \end{pmatrix} \\ & = \begin{pmatrix} \frac{1}{\mathbf{E}[\hat{V}]} \\ \frac{-\mathbf{E}[\hat{U}]}{\mathbf{E}[\hat{V}]^2} \end{pmatrix}^T \begin{pmatrix} \frac{\text{Var}[\tilde{X}_i \tilde{Y}_i]}{\mathbf{E}[\hat{V}]} + \frac{-\mathbf{E}[\hat{U}] \text{Cov}(\tilde{X}_i \tilde{Y}_i, \tilde{X}_i^2)}{\mathbf{E}[\hat{V}]^2} \\ \frac{\text{Cov}(\tilde{X}_i \tilde{Y}_i, \tilde{X}_i^2)}{\mathbf{E}[\hat{V}]} + \frac{-\mathbf{E}[\hat{U}] \text{Var}[\tilde{X}_i^2]}{\mathbf{E}[\hat{V}]^2} \end{pmatrix} \\ & = \frac{\text{Var}[\tilde{X}_i \tilde{Y}_i]}{\mathbf{E}[\hat{V}]^2} - 2 \text{Cov}(\tilde{X}_i \tilde{Y}_i, \tilde{X}_i^2) \frac{\mathbf{E}[\hat{U}]}{\mathbf{E}[\hat{V}]^3} + \text{Var}[\tilde{X}_i^2] \frac{\mathbf{E}[\hat{U}]^2}{\mathbf{E}[\hat{V}]^4} \end{aligned} \tag{3.7}$$

We calculate  $\mathbf{E} [\hat{U}]$ ,  $\mathbf{E} [\hat{V}]$ ,  $\mathbf{Var} [\tilde{X}_i^2]$ , and  $\mathbf{Var} [\tilde{X}_i \tilde{Y}_i]$  in Appendix A. Using them, we can then show:

$$\begin{aligned}
 & \frac{\mathbf{Var} [\tilde{X}_i \tilde{Y}_i]}{\mathbf{E} [\hat{V}]^2} - 2\text{Cov} (\tilde{X}_i \tilde{Y}_i, \tilde{X}_i^2) \frac{\mathbf{E} [\hat{U}]}{\mathbf{E} [\hat{V}]^3} + \mathbf{Var} [\tilde{X}_i^2] \frac{\mathbf{E} [\hat{U}]^2}{\mathbf{E} [\hat{V}]^4} \\
 &= \frac{B^2 R^2 - \beta^{02} \mathbf{E}_X [X_1^2]^2}{\mathbf{E}_X [X_1^2]^2} \\
 & - 2\beta^{02} \left( \mathbf{E}_X [X_1^4] - \mathbf{E}_X [X_1^2]^2 \right) \frac{1}{\mathbf{E}_X [X_1^2]^2} \\
 & + \left( R^2 \mathbf{E}_X [X_1^2] - \mathbf{E}_X [X_1^2]^2 \right) \frac{\beta^{02}}{\mathbf{E}_X [X_1^2]^2} \\
 &= \frac{B^2 R^2 - \beta^{02} \mathbf{E}_X [X_1^2]^2 - 2\beta^{02} \left( \mathbf{E}_X [X_1^4] - \mathbf{E}_X [X_1^2]^2 \right) + \beta^{02} \left( R^2 \mathbf{E}_X [X_1^2] - \mathbf{E}_X [X_1^2]^2 \right)}{\mathbf{E}_X [X_1^2]^2} \\
 &= \frac{B^2 R^2 + \beta^{02} \left( -\mathbf{E}_X [X_1^2]^2 - 2 \left( \mathbf{E}_X [X_1^4] - \mathbf{E}_X [X_1^2]^2 \right) + \left( R^2 \mathbf{E}_X [X_1^2] - \mathbf{E}_X [X_1^2]^2 \right) \right)}{\mathbf{E}_X [X_1^2]^2} \\
 &= \frac{B^2 R^2 + \beta^{02} \left( R^2 \mathbf{E}_X [X_1^2] - 2\mathbf{E}_X [X_1^4] \right)}{\mathbf{E}_X [X_1^2]^2}.
 \end{aligned}$$

Then the asymptotic variance of our estimator is given by

$$\mathbf{Var} [\hat{\beta}] = \frac{B^2 R^2 + \beta^{02} \left( R^2 \mathbf{E}_X [X_1^2] - 2\mathbf{E}_X [X_1^4] \right)}{n \mathbf{E}_X [X_1^2]^2} + o(1/n) = \mathcal{O} \left( \frac{1}{n} \right) \quad (3.8)$$

We will verify this result using estimating equations.

### Components of the Formulation of the Asymptotic Variance

In Appendix A we show:

$$\mathbf{E} [\hat{U}] = \beta^0 \mathbf{E}_X [X_1^2] \quad \mathbf{Var} [\tilde{X}_i \tilde{Y}_i] = B^2 R^2 - \beta^{02} \mathbf{E}_X [X_1^2]^2$$

$$\mathbf{E} [\hat{V}] = \mathbf{E}_X [X_1^2] \quad \mathbf{Var} [\tilde{X}_i^2] = R^2 \mathbf{E}_X [X_1^2] - \mathbf{E}_X [X_1^2]^2$$

$$\text{Cov} (\tilde{X}_i \tilde{Y}_i, \tilde{X}_i^2) = \beta^0 \left( \mathbf{E}_X [X_1^4] - \mathbf{E}_X [X_1^2]^2 \right)$$

Using these formulations, we can show the first component is:

$$\frac{\mathbf{Var} [\hat{U}]}{\mathbf{E} [\hat{V}]^2} = \frac{R^2 B^2 - \beta^{02} \mathbf{E}_X [X_1^2]^2}{\mathbf{E}_X [X_1^2]^2}.$$

For the second component:

$$\begin{aligned} 2\text{Cov} (\tilde{X}_i \tilde{Y}_i, \tilde{X}_i^2) \frac{\mathbf{E} [\hat{U}]}{\mathbf{E} [\hat{V}]^3} &= 2\beta^0 \left( \mathbf{E}_X [X_1^4] - \mathbf{E}_X [X_1^2]^2 \right) \frac{\beta^0 \mathbf{E}_X [X_1^2]}{\mathbf{E}_X [X_1^2]^3} \\ &= 2\beta^{02} \left( \mathbf{E}_X [X_1^4] - \mathbf{E}_X [X_1^2]^2 \right) \frac{1}{\mathbf{E}_X [X_1^2]^2}. \end{aligned}$$

And the last component:

$$\begin{aligned} \mathbf{Var} [\tilde{X}_i^2] \frac{\mathbf{E} [\hat{U}]^2}{\mathbf{E} [\hat{V}]^4} &= \left( R^2 \mathbf{E}_X [X_1^2] - \mathbf{E}_X [X_1^2]^2 \right) \frac{\beta^{02} \mathbf{E}_X [X_1^2]^2}{\mathbf{E}_X [X_1^2]^4} \\ &= \left( R^2 \mathbf{E}_X [X_1^2] - \mathbf{E}_X [X_1^2]^2 \right) \frac{\beta^{02}}{\mathbf{E}_X [X_1^2]^2}. \end{aligned}$$

### 3.2.3 Alternative Derivation Using Estimating Equations

To further verify our derivations from the previous section, we derive here the asymptotic variance of our quantized estimator using estimating equations. To begin, let us define an estimating equation

$$\Psi_n(\beta) := \frac{1}{n} \sum_{i=1}^n \psi_\beta(X_i, Y_i) = \frac{1}{n} \sum_{i=1}^n \widetilde{X}_i^2 \beta - \widetilde{X}_i \widetilde{Y}_i \quad (3.9)$$

where  $\psi_\beta(X_i, Y_i) := \widetilde{X}_i^2 \beta - \widetilde{X}_i \widetilde{Y}_i$ . Let  $\hat{\beta}$  be a zero of  $\Psi_n$ . Then if we expand  $\Psi_n(\hat{\beta})$  in a Taylor series around  $\beta^0$ , rearrange terms, and applying Slutsky's lemma, then we can show

$$\sqrt{n} (\hat{\beta} - \beta^0) \rightsquigarrow N \left( 0, \frac{\mathbf{E} [\psi_{\beta^0}^2]}{\mathbf{E} [\psi'_{\beta^0}]^2} \right) \quad (3.10)$$

where  $\psi'$  indicates the derivative with respect to  $\beta$  and  $\rightsquigarrow$  is convergence in distribution.

Examining the variance

$$\begin{aligned} \frac{\mathbf{E} [\psi_{\beta^0}^2]}{\mathbf{E} [\psi'_{\beta^0}]^2} &= \frac{\mathbf{E} \left[ \left( \widetilde{X}_1^2 \beta^0 - \widetilde{X}_1 \widetilde{Y}_1 \right)^2 \right]}{\mathbf{E} [\widetilde{X}_1^2]^2} = \frac{\mathbf{E} \left[ \beta^{02} \widetilde{X}_1^2 + \widetilde{X}_1^2 \widetilde{Y}_1^2 - 2\beta^0 \widetilde{X}_1^2 \widetilde{X}_1 \widetilde{Y}_1 \right]}{\mathbf{E} [\widetilde{X}_1^2]^2} \\ &= \frac{\beta^{02} \mathbf{E} [\widetilde{X}_1^2] + \mathbf{E} [\widetilde{X}_1^2 \widetilde{Y}_1^2] - 2\beta^0 \mathbf{E} [\widetilde{X}_1^2 \widetilde{X}_1 \widetilde{Y}_1]}{\mathbf{E} [\widetilde{X}_1^2]^2}. \end{aligned} \quad (3.11)$$

Examining each term

$$\mathbf{E} \left[ \widetilde{X}_1^2 \right] = \mathbf{E}_X \left[ (R^2)^2 \frac{X_1^2}{R^2} + (0)^2 \frac{R^2 - X_1^2}{R^2} \right] = R^2 \mathbf{E}_X \left[ X_1^2 \right] \quad (3.12)$$

$$\mathbf{E} \left[ \widetilde{X}_1^2 \widetilde{Y}_1^2 \right] = B^2 R^2 \quad (3.13)$$

$$\mathbf{E} \left[ \widetilde{X}_1^2 \widetilde{X}_1 \widetilde{Y}_1 \right] = \mathbf{E} \left[ X_1^3 Y_1 \right] = \beta^0 \mathbf{E}_X \left[ X_1^4 \right] \quad (3.14)$$

$$\mathbf{E} \left[ \widetilde{X}_1 \right]^2 = \mathbf{E}_X \left[ X_1^2 \right]^2 \quad (3.15)$$

Combining these and plugging them into our variance, we have

$$\frac{\mathbf{E} \left[ \psi_{\beta^0}^2 \right]}{\mathbf{E} \left[ \psi'_{\beta^0} \right]^2} = \frac{B^2 R^2 + \beta^{02} R^2 \mathbf{E}_X \left[ X_1^2 \right] - 2\beta^{02} \mathbf{E}_X \left[ X_1^4 \right]}{\mathbf{E}_X \left[ X_1^2 \right]^2}. \quad (3.16)$$

Thus, we can conclude that the asymptotic variance of our estimator is

$$\mathbf{Var} \left[ \hat{\beta} \right] = \frac{B^2 R^2 + \beta^{02} \left( R^2 \mathbf{E}_X \left[ X_1^2 \right] - 2\mathbf{E}_X \left[ X_1^4 \right] \right)}{n \mathbf{E}_X \left[ X_1^2 \right]^2} + o(1/n) = \mathcal{O} \left( \frac{1}{n} \right), \quad (3.17)$$

which agrees with our formulation in the previous section. If we wish to upper bound the numerator we can further say

$$\mathbf{Var} \left[ \hat{\beta} \right] \leq \frac{B^2 R^2 + \beta^{02} R^4}{n \mathbf{E}_X \left[ X_1^2 \right]^2} + o(1/n) = \mathcal{O} \left( \frac{1}{n} \right). \quad (3.18)$$

We do this to more easily see its agreement with the upper bound of the variance in the  $d$ -predictor case in Section 3.3.

### 3.2.4 MSE and Asymptotic Relative Efficiency

Having identified the asymptotic variance and bias of our estimator, we can now find its asymptotic MSE.

$$MSE(\hat{\beta}) = \mathbf{Var}[\hat{\beta}] + Bias(\hat{\beta})^2 = \mathcal{O}(1/n) + \mathcal{O}(1/n^2) = \mathcal{O}(1/n). \quad (3.19)$$

This implies that the rate of error between our estimator and the OLS estimator (which in the random design setting is equivalent to  $\beta^0$ ) decays at a rate of  $1/n$ .

We know that the asymptotic variance of the OLS estimator  $\beta^*$  is given by

$$\mathbf{Var}[\beta^*] = \frac{\sigma^2}{n\mathbf{E}_X[X_1^2]} + o(1/n). \quad (3.20)$$

Thus, the asymptotic relative efficiency compared to our quantized estimator is

$$\begin{aligned} ARE(\hat{\beta}, \beta^0) &= \frac{B^2R^2 + \beta^{02} \left( R^2\mathbf{E}_X[X_1^2] - 2\mathbf{E}_X[X_1^4] \right)}{n\mathbf{E}_X[X_1^2]^2} \cdot \frac{n\mathbf{E}_X[X_1^2]}{\sigma^2} \\ &= \frac{B^2R^2 + \beta^{02} \left( R^2\mathbf{E}_X[X_1^2] - 2\mathbf{E}_X[X_1^4] \right)}{\sigma^2\mathbf{E}_X[X_1^2]}. \end{aligned} \quad (3.21)$$

The tighter of a bound that can be achieved on the data, the closer it will be to achieving the optimal efficiency.

### 3.3 1-Bit Quantized Regression, $d$ -Predictor Case

We now move to extend our analysis of 1-bit quantized regression from the 1-predictor case, to a general  $d$ -dimensional case. As in the previous section, the work by Vershynin [46] was used throughout. In this section we prove the theorem:

### 3.3.1 Results

**Theorem 3.3.1.** *Let  $\mathbf{X}$  be a  $n \times d$  matrix of random variable predictors whose rows  $\{\mathbf{x}_i^T\}_{i=1}^n$  are  $1 \times d$  vectors of random variables  $X_{ij}$ ,  $j = 1, \dots, d$ . Let the rows of  $\mathbf{X}$  be independent and identically distributed. Let  $|X_{ij}| \leq R$  for all  $i, j$  for some  $R \in \mathbb{R}^+$ .*

*Let  $\mathbf{y} = \{Y_i\}_{i=1}^n$  be an  $n \times 1$  vector of random variable responses such that  $Y_i = \mathbf{x}_i^T \boldsymbol{\beta}^0 + \sigma \epsilon_i$  for  $\boldsymbol{\beta}^0 \in \mathbb{R}^d$  and  $\sigma \in \mathbb{R}^+$ . Let  $\{\epsilon_i\}_{i=1}^n$  be a set of iid standard normal random variables independent of the elements of  $\mathbf{x}_i^T$ . Assume  $|\epsilon_i| \leq L$  for all  $i$  so that we can say  $|Y_i| \leq R \|\boldsymbol{\beta}^0\|_1 + \sigma L$  for all  $i$ . Define  $B := R \|\boldsymbol{\beta}^0\|_1 + \sigma L$ .*

Define

$$\widetilde{X}_{ij} := Q_X(X_{ij}), \quad \widetilde{Y}_i := Q_Y(Y_i), \quad \widetilde{X}_{ij}^2 := Q_{X^2}(X_{ij}^2),$$

where  $Q_X$  and  $Q_Y$  are 1-bit quantizers for elements of  $\mathbf{X}$  and  $\mathbf{y}$  defined on  $[-R, R]$  and  $[-B, B]$ , respectively, as defined in section 3.1.2. Furthermore, let  $Q_{X^2}$  be a 1-bit quantizer for the squared elements of  $\mathbf{X}$ . Let  $\widetilde{\mathbf{x}}_i^T$  be the quantized  $i$ th row of  $\mathbf{X}$  and  $\widetilde{\mathbf{y}}$  be the quantized vector of  $\mathbf{y}$ .

Then for  $\widetilde{\Sigma}_{\mathbf{x}_i} := \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^T + \widetilde{\Delta}_i$  for  $\widetilde{\Delta}_i = \text{diag} \left( \widetilde{X}_{ij}^2 - \widetilde{X}_{ij}^2 \right)$  the estimator  $\hat{\boldsymbol{\beta}}$  that solves the estimating equation

$$\widetilde{\Psi}_n(\boldsymbol{\beta}) := \frac{1}{n} \sum_{i=1}^n \widetilde{\Sigma}_{\mathbf{x}_i} \boldsymbol{\beta} - \widetilde{\mathbf{x}}_i^T \widetilde{y}_i = 0$$

is unbiased and has a covariance matrix whose entries are bounded by a value of order  $\mathcal{O} \left( \frac{1}{n} \right)$ . Furthermore, if  $\lambda_{\min}(\boldsymbol{\Sigma})$  is the smallest magnitude eigenvalue of  $\boldsymbol{\Sigma}$ , we have

$$MSE(\hat{\boldsymbol{\beta}}) \leq n^{-1} d \left( B^2 R^2 + 7R^4 \|\boldsymbol{\beta}^0\|_1^2 \right) \frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})^2} = \mathcal{O} \left( \frac{d}{n} \right). \quad (3.22)$$

### 3.3.2 Introduction

We now extend to the multivariate case with  $d$  predictors. We begin by establishing our regression scenario:

1. Let  $\mathbf{X}$  be a  $n \times d$  matrix of random variable predictors whose rows  $\{\mathbf{x}_i^T\}_{i=1}^n$  are  $1 \times d$  vectors of random variables  $X_{ij}$  such that  $|X_{ij}| \leq R \in \mathbb{R}^+$  for all  $i, j$  where  $i = 1, \dots, n$  and  $j = 1, \dots, d$ . We assume that all rows of  $\mathbf{X}$  are independent and identically distributed, but that the columns are not necessarily so.
2. Let  $\mathbf{y} = \{Y_i\}_{i=1}^n$  be an  $n \times 1$  vector of random variable responses such that  $Y_i = \mathbf{x}_i^T \boldsymbol{\beta}^0 + \sigma \epsilon_i$  for  $\boldsymbol{\beta}^0 \in \mathbb{R}^d$  and  $\sigma \in \mathbb{R}$ . Let  $\{\epsilon_i\}_{i=1}^n$  be a set of iid standard normal random variable. Assume  $|\epsilon_i| \leq L$  for all  $i$  so that we can say  $|Y_i| \leq R \sum_{j=1}^d |\beta_j^0| + \sigma L$  for all  $i$ . Define  $B := R \sum_{j=1}^d |\beta_j^0| + \sigma L$
3. Let  $Q_X$ ,  $Q_{X^2}$ , and  $Q_Y$  be 1-bit quantizers as defined in section 3.1.2 which take values  $[-R, R]$ ,  $[0, R^2]$ , and  $[-B, B]$ , respectively.

#### Notation

We provide Table 3.1 as a reference for important notation. Throughout this section, we will consistently use  $i = 1, \dots, n$  and  $i' = 1, \dots, n$  as indices of observations and  $j = 1, \dots, d$  as indices for features. We will consistently use bold capital letters (both English and Greek) to indicate matrices, bold lowercase letters to denote vectors, and capital non-bold letters to indicate random variables.

Table 3.1: Table of Important Notation

$\mathbf{X}$	A $n \times d$ matrix of predictor variables.
$\mathbf{x}_i^T$	The $i$ th row vector of $\mathbf{X}$ .
$\boldsymbol{\beta}$	A $d \times 1$ vector of parameters.
$\beta_j$	The $j$ th entry of the vector $\boldsymbol{\beta}$ .
$x_{ij}, X_{ij}$	The value found in the $i$ th row and $j$ th column of the matrix $\mathbf{X}$ or, equivalently, the $j$ th element of the vector $\mathbf{x}_i^T$ .
$\widetilde{X}_{ij}$	= The output of the quantization function for predictor value $X_{ij}$ .
$Q_X(X_{ij})$	
$\widetilde{X}_{ij}^2$	= The output of the quantization function for squared predictor value $X_{ij}^2$ .
$Q_{X^2}(X_{ij}^2)$	
$\widetilde{Y}_i = Q_Y(Y_i)$	The output of the quantization function for response value $Y_i$ .
$\widetilde{\mathbf{x}}_i^T$	A vector whose values have been quantized using the $Q_X$ quantizer. e.g., given $\mathbf{x}_i^T = (X_{i1}, X_{i2}, X_{i3})$ , $\widetilde{\mathbf{x}}_i^T = (\widetilde{X}_{i1}, \widetilde{X}_{i2}, \widetilde{X}_{i3})$ .
$\widetilde{\mathbf{X}}$	A matrix $\mathbf{X}$ whose values have been quantized using the $Q_X$ quantizer.
$\widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}_i}$	The $d \times d$ quantized covariance matrix estimate defined as
$\widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}_i} = \begin{pmatrix} \widetilde{x}_{i1}^2 & \widetilde{x}_{i1}\widetilde{x}_{i2} & \widetilde{x}_{i1}\widetilde{x}_{i3} & \dots & \widetilde{x}_{i1}\widetilde{x}_{id} \\ \widetilde{x}_{i2}\widetilde{x}_{i1} & \widetilde{x}_{i2}^2 & \widetilde{x}_{i2}\widetilde{x}_{i3} & \dots & \widetilde{x}_{i2}\widetilde{x}_{id} \\ \widetilde{x}_{i3}\widetilde{x}_{i1} & \widetilde{x}_{i3}\widetilde{x}_{i2} & \widetilde{x}_{i3}^2 & \dots & \widetilde{x}_{i3}\widetilde{x}_{id} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \widetilde{x}_{id}\widetilde{x}_{i1} & \widetilde{x}_{id}\widetilde{x}_{i2} & \widetilde{x}_{id}\widetilde{x}_{i3} & \dots & \widetilde{x}_{id}^2 \end{pmatrix}$	
<p>Note <math>\widetilde{x}_{ij}^2 \neq \widetilde{x}_{ij}^2</math>. This is because <math>\widetilde{x}_{ij}^2 = Q_{X^2}(x_{ij}^2)</math>, which uses the <math>Q_{X^2}(\cdot)</math> quantizer, where as <math>\widetilde{x}_{ij}^2 = (Q_X(x_{ij}))^2</math> which uses the <math>Q_X(\cdot)</math> quantizer. This is a subtle but important distinction.</p>	
$\widetilde{\boldsymbol{\Sigma}}_{\mathbf{X}}$	The sum of the quantized covariance matrix estimates derived from each sample, that is: $\frac{1}{n} \sum_{i=1}^n \widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}_i}$ .

### 3.3.3 Quantized Predictor of $\beta^0$

We now proceed to estimate  $\beta^0$  using the quantized values of  $\mathbf{x}_i^T$  and  $Y_i$ , namely  $\tilde{\mathbf{x}}_i^T$  and  $\tilde{Y}_i$ . We define the estimating equation

$$\tilde{\Psi}_n(\beta) := \frac{1}{n} \sum_{i=1}^n \tilde{\psi}_\beta(\mathbf{x}_i^T, Y_i) = \frac{1}{n} \sum_{i=1}^n \tilde{\Sigma}_{\mathbf{x}_i} \beta - \tilde{\mathbf{x}}_i^T \tilde{y}_i, \quad (3.23)$$

where  $\tilde{\psi}_\beta(\mathbf{x}_i^T, Y_i) := \tilde{\Sigma}_{\mathbf{x}_i} \beta - \tilde{\mathbf{x}}_i^T \tilde{Y}_i$ . Let us assume  $\hat{\beta}$  is a root of  $\tilde{\Psi}_n(\beta)$ .

Note in this equation we use  $\tilde{\Sigma}_{\mathbf{x}_i}$ , as defined in Table 3.1. This is a subtle but important distinction, as defining it in this way results in

$$\mathbf{E}_Q \left[ \tilde{\Sigma}_{\mathbf{x}_i} \right] = \mathbf{x}_i \mathbf{x}_i^T, \quad (3.24)$$

taking advantage of the unbiasedness of the element-wise quantizers. Note that when evaluated at  $\beta^0$ :

$$\mathbf{E} \left[ \tilde{\psi}_{\beta^0} \right] = \mathbf{E} \left[ \mathbf{E}_Q \left[ \tilde{\Sigma}_{\mathbf{x}_i} \beta^0 - \tilde{\mathbf{x}}_i^T \tilde{Y}_i \right] \right] = \mathbf{E} \left[ \mathbf{x}_i \mathbf{x}_i^T \beta^0 - \mathbf{x}_i Y_i \right] = 0 \quad (3.25)$$

for all  $i$ , making  $\tilde{\Psi}_n$  an unbiased estimating equation.

### 3.3.4 Asymptotic Variance of the Quantized Estimator

Using estimating equations and the sandwich method, we move to establish the asymptotic variance of the quantized estimator  $\hat{\beta}$ , which is the estimator that solves

$$\tilde{\Psi}_n(\beta) := \frac{1}{n} \sum_{i=1}^n \tilde{\psi}_\beta(\mathbf{x}_i^T, Y_i) = \frac{1}{n} \sum_{i=1}^n \tilde{\Sigma}_{\mathbf{x}_i} \beta - \tilde{\mathbf{x}}_i^T \tilde{y}_i = \mathbf{0}.$$

Again using a Taylor expansion about  $\beta^0$ , we can show

$$\sqrt{n} \left( \hat{\beta} - \beta^0 \right) \rightsquigarrow N \left( \mathbf{0}, \mathbf{E} \left[ \tilde{\psi}'_{\beta^0} \right]^{-1} \mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right] \mathbf{E} \left[ \tilde{\psi}'_{\beta^0} \right]^{-1} \right). \quad (3.26)$$

where  $\tilde{\psi}'_{\beta}$  is the gradient with respect to  $\beta$ . We have that  $\tilde{\psi}'_{\beta^0} = \tilde{\Sigma}_{\mathbf{x}_i}$ . Then since  $\tilde{\Sigma}_{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \tilde{\Sigma}_{\mathbf{x}_i}$  is a consistent estimator of  $\Sigma := \mathbf{E} \left[ \mathbf{X}^T \mathbf{X} \right]$ , then by the continuous mapping theorem  $\tilde{\Sigma}^{-1}$  converges asymptotically to  $\Sigma^{-1}$ . Then we can refine the distribution

$$\sqrt{n} \left( \hat{\beta} - \beta^0 \right) \rightsquigarrow N \left( \mathbf{0}, \Sigma^{-1} \mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right] \Sigma^{-1} \right). \quad (3.27)$$

Let us now move to evaluate the asymptotic variance of our estimator.

**Bounding the Entries of  $\mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right]$  in the Asymptotic Variance.** We proceed by first looking at the meat portion of the formula:  $\mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right]$ . We note that  $\mathbf{E} \left[ \tilde{x}_{ij}^2 \right] \equiv R^2$  and  $\mathbf{E} \left[ \tilde{Y}_i^2 \right] \equiv B^2$ . Before we begin, let us recall that we have defined

$$\tilde{\Sigma}_{\mathbf{x}_i} := \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T + \Delta_i \text{ where } \Delta_i := \text{diag} \left( \tilde{x}_{ip}^2 - R^2 \right)_{p=1}^d.$$

To begin the evaluation of the entries of  $\mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right]$ , we note that we can express the matrix as:

$$\begin{aligned}
 \mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right] &= \mathbf{E} \left[ \left( \tilde{\Sigma}_{\mathbf{x}_i} \beta^0 - \tilde{\mathbf{x}}_i \tilde{Y}_i \right) \left( \tilde{\Sigma}_{\mathbf{x}_i} \beta^0 - \tilde{\mathbf{x}}_i \tilde{Y}_i \right)^T \right] \\
 &= \mathbf{E} \left[ \tilde{\Sigma}_{\mathbf{x}_i} \beta^0 \beta^{0T} \tilde{\Sigma}_{\mathbf{x}_i}^T - \tilde{\Sigma}_{\mathbf{x}_i} \beta^0 \tilde{Y}_i \tilde{\mathbf{x}}_i^T - \tilde{\mathbf{x}}_i \tilde{Y}_i \beta^{0T} \tilde{\Sigma}_{\mathbf{x}_i}^T + \tilde{\mathbf{x}}_i \tilde{Y}_i \tilde{Y}_i \tilde{\mathbf{x}}_i^T \right] \\
 &= \mathbf{E} \left[ \left( \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T + \Delta_i \right) \beta^0 \beta^{0T} \left( \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T + \Delta_i \right) - \left( \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T + \Delta_i \right) \beta^0 \tilde{Y}_i \tilde{\mathbf{x}}_i^T \right. \\
 &\quad \left. - \tilde{\mathbf{x}}_i \tilde{Y}_i \beta^{0T} \left( \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T + \Delta_i \right) + \tilde{\mathbf{x}}_i \tilde{Y}_i \tilde{Y}_i \tilde{\mathbf{x}}_i^T \right] \\
 &= \mathbf{E} \left[ \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T \beta^0 \beta^{0T} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T + \Delta_i \beta^0 \beta^{0T} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T + \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T \beta^0 \beta^{0T} \Delta_i + \Delta_i \beta^0 \beta^{0T} \Delta_i \right. \\
 &\quad \left. - \tilde{Y}_i \left( \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T \beta^0 \tilde{\mathbf{x}}_i^T + \Delta_i \beta^0 \tilde{\mathbf{x}}_i^T + \tilde{\mathbf{x}}_i \beta^{0T} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T + \tilde{\mathbf{x}}_i \beta^{0T} \Delta_i \right) + B^2 \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T \right]. \tag{3.28}
 \end{aligned}$$

Then the  $j, k$ -th element is given by

$$\begin{aligned}
 &\left( \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right)_{jk} \\
 &= \sum_{\ell=1}^d \sum_{m=1}^d \tilde{x}_{ij} \tilde{x}_{i\ell} \beta_{\ell}^0 \beta_m^0 \tilde{x}_{im} \tilde{x}_{ik} + \delta_j \beta_j^0 \sum_{m=1}^d \beta_m^0 \tilde{x}_{im} \tilde{x}_{ik} + \delta_k \beta_k^0 \sum_{m=1}^d \beta_m^0 \tilde{x}_{im} \tilde{x}_{ij} + \delta_j \beta_j^0 \beta_k^0 \delta_k \\
 &\quad - \tilde{Y}_i \left( \sum_{m=1}^d \beta_m^0 \tilde{x}_{im} \tilde{x}_{ij} \tilde{x}_{ik} + \beta_j^0 \delta_j \tilde{x}_{ik} + \sum_{m=1}^d \beta_m^0 \tilde{x}_{im} \tilde{x}_{ij} \tilde{x}_{ik} + \beta_k^0 \delta_k \tilde{x}_{ij} \right) + B^2 \tilde{x}_{ij} \tilde{x}_{ik} \tag{3.29}
 \end{aligned}$$

where we have let  $\delta_p$  be the  $p, p$ th element of  $\mathbf{\Delta}_i$ . Then taking the expectation with respect to the quantizers results in

$$\begin{aligned}
 \mathbf{E}_Q \left[ \left( \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right)_{jk} \right] &= \underbrace{\sum_{\ell=1}^d \sum_{m=1}^d \beta_{\ell}^0 \beta_m^0 \mathbf{E}_Q [ \widetilde{x}_{ij} \widetilde{x}_{i\ell} \widetilde{x}_{im} \widetilde{x}_{ik} ]}_{(a)} + \underbrace{\beta_j^0 \mathbf{E}_Q [ \delta_j ] \sum_{m=1}^d \beta_m^0 \mathbf{E}_Q [ \widetilde{x}_{im} \widetilde{x}_{ik} ]}_{(b)} \\
 &+ \underbrace{\mathbf{E}_Q [ \delta_k ] \beta_k^0 \sum_{m=1}^d \beta_m^0 \mathbf{E}_Q [ \widetilde{x}_{im} \widetilde{x}_{ij} ] + \delta_j \beta_j^0 \beta_k^0 \delta_k + B^2 \mathbf{E}_Q [ \widetilde{x}_{ij} \widetilde{x}_{ik} ]}_{(c)} \\
 &- \mathbf{E}_Q [ \tilde{Y}_i ] \left( \underbrace{2 \sum_{m=1}^d \beta_m^0 \mathbf{E}_Q [ \widetilde{x}_{im} \widetilde{x}_{ij} \widetilde{x}_{ik} ] + \beta_j^0 \mathbf{E}_Q [ \delta_j ] \mathbf{E}_Q [ \widetilde{x}_{ik} ] + \beta_k^0 \mathbf{E}_Q [ \delta_k ] \mathbf{E}_Q [ \widetilde{x}_{ij} ]}_{(d)} \right). \quad (3.30)
 \end{aligned}$$

It is important to note that  $\mathbf{E}_Q [ \widetilde{x}_{ip} \widetilde{x}_{iq} ] \neq \mathbf{E}_Q [ \widetilde{x}_{ip} ] \mathbf{E}_Q [ \widetilde{x}_{iq} ]$  when  $p = q$ . We have instead that  $\mathbf{E}_Q [ \widetilde{x}_{ip} \widetilde{x}_{iq} ] = R^2$  since  $\widetilde{x}_{ip} \widetilde{x}_{iq} \equiv R^2$  when  $p = q$ .

So, when calculating the  $j, k$ th term of  $\mathbf{E} [ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T ]$ , we must expand all possible combinations of the relationships of the iteration variables in (a), (b), (c), and (c). The full derivation of their formulas can be found in Appendix B.

Furthermore, we show in Appendix B that the entries of  $\mathbf{E} [ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T ]$  are upper bounded by

$$\mathbf{E} \left[ \left( \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right)_{jk} \right] \leq B^2 R^2 + R^4 \left( \left\| \boldsymbol{\beta}^0 \right\|_2^2 + 2 \left( \left| \beta_j^0 \right| + \left| \beta_k^0 \right| \right) \left\| \boldsymbol{\beta}^0 \right\|_1 + 2 \left\| \boldsymbol{\beta}^0 \right\|_1^2 \right) \quad (3.31)$$

when  $k \neq j$  and by

$$\mathbf{E} \left[ \left( \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right)_{jj} \right] \leq B^2 R^2 + R^4 \left( \left\| \boldsymbol{\beta}^0 \right\|_2^2 + 4 \left| \beta_j^0 \right| \left\| \boldsymbol{\beta}^0 \right\|_1 + 2 \left\| \boldsymbol{\beta}^0 \right\|_1^2 \right) \quad (3.32)$$

when  $k = j$ . Furthermore, we can then say that all entries of the matrix  $\mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right]$  are bounded by

$$\mathbf{E} \left[ \left( \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right)_{jk} \right] \leq B^2 R^2 + 7R^4 \left\| \beta^0 \right\|_1^2. \quad (3.33)$$

We now use this bound to upper bound the MSE of our estimator.

### 3.3.5 MSE of the Estimator

Let us begin by defining the notation  $\doteq$  which we define to be asymptotic equality up to  $o(1/n)$ . Using this notation allows us to omit the  $o(1/n)$  term. Then we can say that the asymptotic MSE of the estimator is given by

$$\begin{aligned} MSE(\hat{\beta}) &= tr \left( \mathbf{Var} \left[ \hat{\beta} \right] \right) \doteq tr \left( n^{-1} \mathbf{\Sigma}^{-1} \mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right] \mathbf{\Sigma}^{-1} \right) \\ &= n^{-1} \sum_j^d \sum_k^d \sum_\ell^d \mathbf{\Sigma}_{jk}^{-1} \mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right]_{k\ell} \mathbf{\Sigma}_{\ell j}^{-1}. \end{aligned}$$

Taking the absolute value of the trace gives us

$$\begin{aligned} \left| tr \left( \mathbf{\Sigma}^{-1} \mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right] \mathbf{\Sigma}^{-1} \right) \right| &= \left| n^{-1} \sum_j^d \sum_k^d \sum_\ell^d \mathbf{\Sigma}_{jk}^{-1} \mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right]_{k\ell} \mathbf{\Sigma}_{\ell j}^{-1} \right| \\ &\leq n^{-1} \sum_j^d \sum_k^d \sum_\ell^d \left| \mathbf{\Sigma}_{jk}^{-1} \right| \left| \mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right]_{k\ell} \right| \left| \mathbf{\Sigma}_{\ell j}^{-1} \right| \\ &\leq n^{-1} \left( B^2 R^2 + 7R^4 \left\| \beta^0 \right\|_1^2 \right) \sum_j^d \sum_k^d \sum_\ell^d \left| \mathbf{\Sigma}_{jk}^{-1} \right| \left| \mathbf{\Sigma}_{\ell j}^{-1} \right| \\ &= n^{-1} \left( B^2 R^2 + 7R^4 \left\| \beta^0 \right\|_1^2 \right) \left\| \mathbf{\Sigma}^{-1} \right\|_F^2 \end{aligned}$$

where  $\|\cdot\|_F$  is the Frobenius norm. Then we can refine our MSE bound by using the fact that  $\|\cdot\|_F \leq \sqrt{d} \|\cdot\|_2$ :

$$\begin{aligned} \left| \text{tr} \left( \boldsymbol{\Sigma}^{-1} \mathbf{E} \left[ \tilde{\boldsymbol{\psi}}_{\beta^0} \tilde{\boldsymbol{\psi}}_{\beta^0}^T \right] \boldsymbol{\Sigma}^{-1} \right) \right| &\leq n^{-1} \left( B^2 R^2 + 7R^4 \|\boldsymbol{\beta}^0\|_1^2 \right) \|\boldsymbol{\Sigma}^{-1}\|_F^2 \\ &\leq n^{-1} d \left( B^2 R^2 + 7R^4 \|\boldsymbol{\beta}^0\|_1^2 \right) \frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})^2}. \end{aligned} \quad (3.34)$$

Thus, we can bound our MSE by

$$MSE(\hat{\boldsymbol{\beta}}) \leq n^{-1} d \left( B^2 R^2 + 7R^4 \|\boldsymbol{\beta}^0\|_1^2 \right) \frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})^2} = \mathcal{O}\left(\frac{d}{n}\right). \quad (3.35)$$

### 3.3.6 Asymptotic Relative Efficiency

We wish to analyze the asymptotic relative efficiency (ARE) between the quantized estimator in the  $d$ -predictor case  $\hat{\boldsymbol{\beta}}$  and the OLS estimator  $\boldsymbol{\beta}$ . Having not yet examined the form of the asymptotic variance of the OLS estimator, we provide it here.

Assuming  $\mathbf{X}$  is full rank and each row of  $\mathbf{X}$ ,  $\mathbf{x}_i^T$ , is independent of  $\epsilon_i$ , we know the least squares estimator is given by  $\boldsymbol{\beta}^* = \left( \mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{y}$  and solves the estimating equation:

$$\boldsymbol{\Psi}_n(\boldsymbol{\beta}) := \frac{1}{n} \sum_{i=1}^n \psi_{\boldsymbol{\beta}}(\mathbf{x}_i, Y_i) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \left( \mathbf{x}_i^T \boldsymbol{\beta} - Y_i \right) = \mathbf{0}, \quad (3.36)$$

where we let  $\mathbf{0}$  denote a vector containing all zeros, in this case a  $d \times 1$  vector of zeros. We assume the pair  $(\mathbf{x}_i^T, Y_i)$  is independent of every other pair  $(\mathbf{x}_{i'}^T, Y_{i'})$  for all  $i' \neq i$ , then  $\psi_{\boldsymbol{\beta}}$  is independent of  $\psi_{i'}$ .

Then the asymptotic variance of the estimator  $\boldsymbol{\beta}^*$  is given by

$$\mathbf{Var}[\boldsymbol{\beta}^*] \doteq \boldsymbol{\Sigma}^{-1} \mathbf{E} \left[ \psi_{\beta^0} \psi_{\beta^0}^T \right] \boldsymbol{\Sigma}^{-1} = n^{-1} \sigma^2 \boldsymbol{\Sigma}^{-1}. \quad (3.37)$$

The ARE is then given by

$$ARE(\hat{\beta}, \beta^*) \doteq \frac{\text{tr} \left( n^{-1} \Sigma^{-1} \mathbf{E} \left[ \left( \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right)_{jk} \right] \Sigma^{-1} \right)}{\text{tr} \left( n^{-1} \sigma^2 \Sigma^{-1} \right)}. \quad (3.38)$$

We have upper bounded the numerator in the previous section. Now we lower bound the denominator by using properties of norms

$$\text{tr} \left( n^{-1} \sigma^2 \Sigma^{-1} \right) \geq n^{-1} d \sigma^2 \lambda_{\min}(\Sigma)^{-1}. \quad (3.39)$$

Combining these we can say

$$ARE(\hat{\beta}, \beta^*) \leq \frac{n^{-1} d \left( B^2 R^2 + 7R^4 \|\beta^0\|_1^2 \right) \frac{1}{\lambda_{\min}(\Sigma)^2}}{n^{-1} d \sigma^2 \lambda_{\min}(\Sigma)^{-1}} = \frac{B^2 R^2 + 7R^4 \|\beta^0\|_1^2}{\sigma^2 \lambda_{\min}(\Sigma)}. \quad (3.40)$$

## 3.4 Conclusion

### 3.4.1 Summary

We have shown in the 1-predictor case using two methods that the estimator

$$\hat{\beta} = \frac{\frac{1}{n} \sum_{i=1}^n \tilde{X}_i \tilde{Y}_i}{\frac{1}{n} \sum_{i=1}^n \tilde{X}_i^2} \quad (3.41)$$

is asymptotically unbiased and has asymptotic variance of

$$\text{Var} \left[ \hat{\beta} \right] \xrightarrow{p} \frac{B^2 R^2 + \beta^{02} \left( R^2 \mathbf{E}_X [X_1^2] - 2 \mathbf{E}_X [X_1^4] \right)}{n \mathbf{E}_X [X_1^2]^2} \leq \frac{B^2 R^2 + \beta^{02} R^4}{n \mathbf{E}_X [X_1^2]^2}.$$

This allows us to say the MSE of the estimator asymptotically approaches

$$MSE(\hat{\beta}) = \frac{B^2 R^2 + \beta^{02} \left( R^2 \mathbf{E}_X [X_1^2] - 2 \mathbf{E}_X [X_1^4] \right)}{n \mathbf{E}_X [X_1^2]^2} = \mathcal{O}(1/n). \quad (3.42)$$

Furthermore, we can say the asymptotic relative efficiency compared to the OLS estimator is

$$ARE(\hat{\beta}, \beta^0) = \frac{B^2 R^2 + \beta^{02} \left( R^2 \mathbf{E}_X [X_1^2] - 2 \mathbf{E}_X [X_1^4] \right)}{\sigma^2 \mathbf{E}_X [X_1^2]}. \quad (3.43)$$

In the  $d$ -predictor case, we showed that an estimator  $\hat{\beta}$ , that is the solution to the estimating equation

$$\tilde{\Psi}_n(\beta) := \frac{1}{n} \sum_{i=1}^n \tilde{\psi}_\beta(\mathbf{x}_i^T, Y_i) = \frac{1}{n} \sum_{i=1}^n \tilde{\Sigma}_{\mathbf{x}_i} \beta - \tilde{\mathbf{x}}_i^T \tilde{y}_i = \mathbf{0},$$

converges in distribution to

$$\hat{\beta} \rightsquigarrow N \left( \beta^0, n^{-1} \Sigma^{-1} \mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right] \Sigma^{-1} \right). \quad (3.44)$$

We then bounded the elements of  $\mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right]$  by

$$\mathbf{E} \left[ \left( \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right)_{jk} \right] \leq B^2 R^2 + 7R^4 \left\| \beta^0 \right\|_1^2. \quad (3.45)$$

This leads us to an upper bound on MSE of the estimator

$$MSE(\hat{\beta}) \leq n^{-1} d \left( B^2 R^2 + 7R^4 \left\| \beta^0 \right\|_1^2 \right) \frac{1}{\lambda_{\min}(\Sigma)^2} = \mathcal{O} \left( \frac{d}{n} \right). \quad (3.46)$$

Furthermore, we can say that the ARE is upper bounded by

$$ARE(\hat{\beta}, \beta^*) \leq \frac{B^2 R^2 + 7R^4 \|\beta^0\|_1^2}{\sigma^2 \lambda_{\min}(\Sigma)}. \quad (3.47)$$

## Chapter 4: Bounding the Error from Sketched and Quantized Regression Parameters

### 4.1 Introduction

#### 4.1.1 Introduction

In the era of big data, the extreme number of samples in a dataset frequently presents challenges for learning algorithms, even one so fundamental to machine learning and statistics as linear regression. Moreover, the substantial size of data can impede its transfer from its source to a processing location. Researchers have created many lossy compression algorithms that overcome these difficulties while still allowing learning algorithms to function with some error. This paper examines the error incurred in linear regression when using data compressed by three compression algorithms: 1-bit quantization, random sketching, and their combination, sketching and 1-bit quantization. It provides a theoretical upper bound on the difference between the regression parameters obtained from compressed data using each of these methods and those derived from standard least squares using the original dataset in both a fixed and random design setting. Finally, the paper provides a closed formula for the error's upper bound based on dimensions of the original dataset and the sketching matrix.

#### 4.1.2 Background

The size of datasets in the modern age requires that we invest in and research methods capable of compressing data while retaining our ability to learn from it. We focus on methods designed to compress data for use in a linear regression model.

Compressed Linear Regression (CLR) is often referred to as a method that reduces the features from  $d$  to some  $k \ll d \ll n$  where  $n$  is the number of samples. Maillard and Munos [47], Kaban [48], and Slawski [49] explore using a random projection matrix of size  $d \times k$  to reduce the feature set and analyzing the expected excess risk of OLS with the reduced data. Slawski provides an improved analysis over [47], [48] that shows CLR performs similarly to Principal Component Regression (PCR) and highlights CLR's advantages in computation, but not necessarily in the statistical sense. Principal Component Analysis in linear regression [50]–[54] and dimension reduction by a Johnson-Lindenstrauss transform or its like have been extensively studied [55]–[57]. Slawski [58] provides a comparison of both the computational and statistical performance of PCR compared to a Johnson-Lindenstrauss transform under mild assumptions.

While these methods are effective at reducing the size of the dataset by reducing the number of features, they do nothing to reduce the number of samples. There are many works that perform a reduction of the number of samples, which we call 'sketching' from now on [57], [59]–[67]. Each of these methods shows, under varying assumptions, how sketching can reduce the computational complexity and data storage requirements, and they provide theoretical upper bounds on the expected error, usually as expected excess risk from the hypothesized true relationship between  $\mathbf{X}$  and  $\mathbf{y}$  in a linear regression model.

Dobriban and Liu [62] perform an asymptotic analysis using various assumptions on the sketching matrix. They show the asymptotic increase in parameter estimation error, prediction efficiency, and out-of-sample prediction efficiency under different scenarios. While their work is asymptotic, our work similarly shows the increase in parameter estimation error.

Raskutti and Mahoney [66] provide a framework for evaluating the effectiveness of different sketching methods. They bin the types of sketching into two main categories: random selection and random projection. They then use the metrics *prediction efficiency* and *residual efficiency* to evaluate the usefulness of a sketching matrix instead of the usual ratio of variances of the estimators. They prove the following theorem

**Theorem 4.1.1.** [66] For any  $n \times d$  matrix  $\mathbf{X}$ , there exists a constant  $c$  such that if  $r \geq c \log(n)$ , then with probability greater than 0.7, it holds that  $\text{rank}(S_{SGP}) = d$  and that

$$C_{PE}(S_{SGP}) = \frac{\mathbf{E} \left[ \left\| \mathbf{X} (\boldsymbol{\beta}^0 - \boldsymbol{\beta}_{S_{SGP}}) \right\|_2^2 \right]}{\mathbf{E} \left[ \left\| \mathbf{X} (\boldsymbol{\beta}^0 - \boldsymbol{\beta}^*) \right\|_2^2 \right]} \leq 44 \left( 1 + \frac{n}{r} \right)$$

where  $r$  is the first dimension of the projection matrix  $S_{SGP}$  whose entries are iid standard gaussian, and  $\boldsymbol{\beta}_{S_{SGP}}$  is the solution to the LS equation using data projected using the  $S_{SGP}$  matrix,  $\boldsymbol{\beta}^*$  is the OLS solution, and  $\boldsymbol{\beta}^0$  are the true values in the LS equation [66].

Pilanci and Wainwright [65] perform sub-Gaussian and randomized orthogonal system projections. They provide high-probability guarantees in terms of the size of the sketching matrix on the optimality of the sketched estimators to the least squares problem. Their main result most pertinent to our work is given in their Theorem 1 which provides the value of a sub-Gaussian sketching matrix that will provide a high probability bound error between the OLS estimation error and a sketched estimator error. The theorem states

**Theorem 4.1.2.** [65] Let  $\mathbf{S} \in \mathbb{R}^{m \times n}$  be drawn from a  $\sigma$ -sub-Gaussian ensemble. Then there are universal constants  $(c_0, c_1, c_2)$  such that, for any tolerance parameter  $\delta \in (0, 1)$ , given a sketch size lower bounded as

$$m \geq \frac{c_0}{\delta^2} \mathcal{W}(\mathbf{X}\mathcal{K}),$$

the approximate solution  $\hat{\boldsymbol{\beta}}$  is guaranteed to be  $\delta$ -optimal for the original program with probability at least  $1 - c_1 \exp^{-c_2 m \delta^2}$ , where  $\mathbf{X}\mathcal{K}$  denotes the linearly transformed cone  $\{\mathbf{X}\Delta \in \mathbb{R}^n \mid \Delta \in \mathcal{K}\}$  and  $\mathcal{W}(\mathbf{X}\mathcal{K}) := \mathbf{E} \left[ \sup_{z \in \mathbf{X}\mathcal{K} \cap S^{n-1}} |\langle g, z \rangle| \right]$  for  $g \in \mathbb{R}^n$ .

This paper builds on Pilanci's [65] and Raskutti's [66] work. We focus on estimators of the covariance matrix and the product of the design matrix and response vector. In

doing this, we can perform analysis to show upper bounds on the parameter error in both the fixed and random design settings. Additionally, we examine the effect of stochastic quantization (or dithering) on linear regression and provide the same upper bounds. Lastly, we combine both sketching and quantization as a new method of compression with desirable error bounds.

### 4.1.3 Problem Setup

Suppose we are given pairs  $(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n)$  where each  $\mathbf{x}_i$  is a vector and each  $Y_i \in \mathbb{R}$ . Assume these pairs have a linear relationship described by

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta}^0 + \sigma \epsilon_i, \quad 1 \leq i \leq n \quad (4.1)$$

for some linear map  $\boldsymbol{\beta}^0$  and noise terms  $\epsilon_i$ . While we assume throughout that the noise terms are standard Gaussian, we expect that a more relaxed assumption of symmetry is possible using different proof techniques. In this chapter we will investigate a novel estimator  $\hat{\boldsymbol{\beta}}$  of the optimal solution  $\boldsymbol{\beta}^*$  under both a fixed and random design setting. We will evaluate an upper bound of the squared error between this estimator and the true linear map  $\boldsymbol{\beta}^*$ .

#### Bounding the Estimator Error

Let the matrix  $\mathbf{X}$  be comprised of vectors  $\{\mathbf{x}_i^T\}_{i=1}^n$ , and the vector  $\mathbf{y} = \{Y_i\}_{i=1}^n$  contain random variates where the relationship (4.1) holds for each  $i$ . The solution to the quadratic minimization problem corresponding to the least squares problem is given by

$$\boldsymbol{\beta}^* = \arg \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} - \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}} \boldsymbol{\beta} \right\}, \quad (4.2)$$

where in the fixed design setting

$$\boldsymbol{\Sigma} := \frac{\mathbf{X}^T \mathbf{X}}{n} \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}} := \frac{\mathbf{X}^T \mathbf{y}}{n} \quad (4.3)$$

and in the random design setting

$$\Sigma := \mathbf{E} \left[ \frac{\mathbf{X}^T \mathbf{X}}{n} \right] \quad \text{and} \quad \Sigma_{\mathbf{X}\mathbf{y}} := \mathbf{E} \left[ \frac{\mathbf{X}^T \mathbf{y}}{n} \right]. \quad (4.4)$$

We define an estimator  $\hat{\beta}$

$$\hat{\beta} := \arg \min_{\beta} \left\{ \frac{1}{2} \beta^T \hat{\Sigma} \beta - \hat{\Sigma}_{\mathbf{X}\mathbf{y}} \beta \right\}, \quad (4.5)$$

where  $\hat{\Sigma}$  and  $\hat{\Sigma}_{\mathbf{X}\mathbf{y}}$  are unbiased estimators of  $\Sigma$  and  $\Sigma_{\mathbf{X}\mathbf{y}}$  derived from transformed data, denoted as  $\hat{\mathbf{y}}$  and  $\hat{\mathbf{X}}$  that maintain conformability.

Then we can show that:

**Lemma 1.** *Given a matrix  $\mathbf{X}$  comprised of vectors  $\{\mathbf{x}_i^T\}_{i=1}^n$  and the vector  $\mathbf{y} = \{Y_i\}_{i=1}^n$  having a linear relationship  $\mathbf{y} = \mathbf{X}\beta^0 + \sigma\epsilon$  with standard regression assumptions. Let  $\beta^*$  be the solution that minimizes*

$$\beta^* = \arg \min_{\beta} \left\{ \frac{1}{2} \beta^T \Sigma \beta - \Sigma_{\mathbf{X}\mathbf{y}} \beta \right\} \quad (4.6)$$

and let  $\hat{\beta}$  be the solution that minimizes

$$\hat{\beta} = \arg \min_{\beta} \left\{ \frac{1}{2} \beta^T \hat{\Sigma} \beta - \hat{\Sigma}_{\mathbf{X}\mathbf{y}} \beta \right\} \quad (4.7)$$

where  $\hat{\mathbf{X}}$  and  $\hat{\mathbf{y}}$  are transformations of  $\mathbf{X}$  and  $\mathbf{y}$  that maintain the necessary conformability. In the fixed design setting define

$$\Sigma := \frac{\mathbf{X}^T \mathbf{X}}{n} \quad \text{and} \quad \Sigma_{\mathbf{X}\mathbf{y}} := \frac{\mathbf{X}^T \mathbf{y}}{n} \quad (4.8)$$

and in the random design setting define

$$\boldsymbol{\Sigma} := \mathbf{E} \left[ \frac{\mathbf{X}^T \mathbf{X}}{n} \right] \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}} := \mathbf{E} \left[ \frac{\mathbf{X}^T \mathbf{y}}{n} \right]. \quad (4.9)$$

Let  $\hat{\boldsymbol{\Sigma}}$  and  $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}}$  be unbiased estimators of  $\boldsymbol{\Sigma}$  and  $\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}}$  derived from the transformed data  $\hat{\mathbf{X}}$  and  $\hat{\mathbf{y}}$ .

Then if  $\lambda_{\min}(\boldsymbol{\Sigma})$  is the smallest magnitude eigenvalue of  $\boldsymbol{\Sigma}$ , we have

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{4\|\boldsymbol{\beta}^*\|_2 \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| + 4\|\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}}\|_2}{\lambda_{\min}(\boldsymbol{\Sigma})}.$$

### Proof of Lemma 1

*Proof.* Let

$$f(\boldsymbol{\beta}) := \frac{1}{2}\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} - \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}} \boldsymbol{\beta} \quad \text{and} \quad g(\boldsymbol{\beta}) := \frac{1}{2}\boldsymbol{\beta}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\beta} - \hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} \boldsymbol{\beta}. \quad (4.10)$$

We note that both  $\boldsymbol{\beta}^*$  and  $\hat{\boldsymbol{\beta}}$  are feasible solutions to both  $f$  and  $g$ , but  $\boldsymbol{\beta}^*$  is optimal for  $f$  and  $\hat{\boldsymbol{\beta}}$  is optimal for  $g$ . Thus, we can say

$$f(\boldsymbol{\beta}^*) \leq f(\hat{\boldsymbol{\beta}}) \quad (4.11)$$

$$g(\hat{\boldsymbol{\beta}}) \leq g(\boldsymbol{\beta}^*). \quad (4.12)$$

We now use ((4.12)) and the definition of  $g$  to say:

$$\begin{aligned} g(\hat{\boldsymbol{\beta}}) \leq g(\boldsymbol{\beta}^*) &\implies \frac{1}{2}\hat{\boldsymbol{\beta}}^T \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}}^T \hat{\boldsymbol{\beta}} \leq \frac{1}{2}\boldsymbol{\beta}^{*T} \hat{\boldsymbol{\Sigma}} \boldsymbol{\beta}^* - \hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}}^T \boldsymbol{\beta}^* \\ &\implies \frac{1}{2}\hat{\boldsymbol{\beta}}^T \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\beta}} \leq \frac{1}{2}\boldsymbol{\beta}^{*T} \hat{\boldsymbol{\Sigma}} \boldsymbol{\beta}^* + \hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}}^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*). \end{aligned} \quad (4.13)$$

Then, adding and subtracting  $\frac{1}{2}\hat{\beta}^T \hat{\Sigma} \beta^*$ ,  $\frac{1}{2}\beta^{*T} \hat{\Sigma} \hat{\beta}$ , and  $\beta^{*T} \hat{\Sigma} \beta^*$  from ((4.13)) and letting  $\hat{\delta} = \hat{\beta} - \beta^*$ , we get

$$\begin{aligned} \frac{1}{2}\hat{\delta}^T \hat{\Sigma} \hat{\delta} &\leq \beta^{*T} \hat{\Sigma} \beta^* - \beta^{*T} \hat{\Sigma} \hat{\beta} + \hat{\Sigma}_{\mathbf{X}_y}^T \hat{\delta} \\ &= \beta^{*T} \hat{\Sigma} (\beta^* - \hat{\beta}) - \hat{\Sigma}_{\mathbf{X}_y}^T (\beta^* - \hat{\beta}). \end{aligned} \quad (4.14)$$

Now adding and subtracting  $\beta^{*T} (\hat{\Sigma} - \Sigma) \hat{\delta}$  and  $(\hat{\Sigma}_{\mathbf{X}_y} - \Sigma_{\mathbf{X}_y})^T \hat{\delta}$  from the right side of (4.14) and using that  $\langle \Sigma \beta^* - \Sigma_{\mathbf{X}_y}, \beta^* - \hat{\beta} \rangle = (\Sigma_{\mathbf{X}_y} - \Sigma \beta^*)^T (\beta^* - \hat{\beta})$ , we get:

$$\frac{1}{2}\hat{\delta}^T \hat{\Sigma} \hat{\delta} \leq -\langle \Sigma \beta^* - \Sigma_{\mathbf{X}_y}, \beta^* - \hat{\beta} \rangle - \beta^{*T} (\hat{\Sigma} - \Sigma) \hat{\delta} + (\hat{\Sigma}_{\mathbf{X}_y} - \Sigma_{\mathbf{X}_y})^T \hat{\delta}.$$

Using the optimality of  $\beta^*$  for the original problem (4.11) and since  $\langle \Sigma \beta^* - \Sigma_{\mathbf{X}_y}, \beta^* - \hat{\beta} \rangle \geq 0$ , we have

$$\frac{1}{2}\hat{\delta}^T \hat{\Sigma} \hat{\delta} = -\beta^{*T} (\hat{\Sigma} - \Sigma) \hat{\delta} + (\hat{\Sigma}_{\mathbf{X}_y} - \Sigma_{\mathbf{X}_y})^T \hat{\delta} \quad (4.15)$$

Examining the right-hand side

$$\begin{aligned} -\beta^{*T} (\hat{\Sigma} - \Sigma) \hat{\delta} + (\hat{\Sigma}_{\mathbf{X}_y} - \Sigma_{\mathbf{X}_y})^T \hat{\delta} &= \left( -\beta^{*T} (\hat{\Sigma} - \Sigma) + (\hat{\Sigma}_{\mathbf{X}_y} - \Sigma_{\mathbf{X}_y})^T \right) \hat{\delta} \\ &\leq \|\hat{\delta}\|_2 \left( \|\beta^*\|_2 \|\hat{\Sigma} - \Sigma\| + \|(\hat{\Sigma}_{\mathbf{X}_y} - \Sigma_{\mathbf{X}_y})\|_2 \right) \end{aligned} \quad (4.16)$$

by Cauchy Schwarz, the triangle inequality, and properties of induced norms (see (4.31) for the definition of  $\|\cdot\|$ ). Now examining the left-hand side

$$\begin{aligned}
 \frac{1}{2} \hat{\boldsymbol{\delta}}^T \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\delta}} &= \frac{1}{2} \left( \hat{\boldsymbol{\delta}}^T \boldsymbol{\Sigma} \hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}^T (\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}) \hat{\boldsymbol{\delta}} \right) \\
 &\geq \frac{1}{2} \left( \lambda_{\min}(\boldsymbol{\Sigma}) \|\hat{\boldsymbol{\delta}}\|_2^2 - \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| \|\hat{\boldsymbol{\delta}}\|_2^2 \right) \\
 &= \frac{1}{2} \|\hat{\boldsymbol{\delta}}\|_2^2 \left( \lambda_{\min}(\boldsymbol{\Sigma}) - \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| \right) \\
 &\geq \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{4} \|\hat{\boldsymbol{\delta}}\|_2^2
 \end{aligned} \tag{4.17}$$

where we require  $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| \leq \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{2}$ . We will call this the lambda-min-requirement and reference it as requirement (4.17). This requirement can be shown to be met with high probability when the estimator is designed such that

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n \hat{\boldsymbol{\Sigma}}_i \quad \text{and} \quad \mathbf{E} [\hat{\boldsymbol{\Sigma}}_i] = \boldsymbol{\Sigma}, \tag{4.18}$$

for all  $i$ , which will be the case for all the estimators in this paper. Each section will provide a value of  $n$  such that with high probability we have

$$\begin{aligned}
 \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{4} \|\hat{\boldsymbol{\delta}}\|_2^2 &\leq \|\hat{\boldsymbol{\delta}}\|_2 \left( -\|\boldsymbol{\beta}^*\|_2 \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| + \|(\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}})\|_2 \right) \\
 \implies \|\hat{\boldsymbol{\delta}}\|_2 &\leq \frac{4\|\boldsymbol{\beta}^*\|_2 \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| + 4\|(\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}})\|_2}{\lambda_{\min}(\boldsymbol{\Sigma})}
 \end{aligned} \tag{4.19}$$

□

The inequality (4.19) is the general problem of this chapter and will be referenced repeatedly. We will separate problem (4.19) into two parts:

$$(a): \left\| \hat{\Sigma} - \Sigma \right\| \quad (b): \left\| \hat{\Sigma}_{\mathbf{X}\mathbf{Y}} - \Sigma_{\mathbf{X}\mathbf{Y}} \right\|_2 \quad (4.20)$$

We will use varying methods to control (a) and (b) based on the methods of each chapter and the assumptions made in each section. We will then combine their results to find an upper bound on  $\left\| \hat{\delta} \right\|_2$ .

#### 4.1.4 Assumptions and Definitions

##### Regression Model, Error Independence, and Scaling

We will construct a  $n \times d$  design matrix  $\mathbf{Z}$  and a  $n \times 1$  response vector  $\mathbf{w}$  such that  $\mathbf{w} = \mathbf{Z}\beta^0 + \sigma\epsilon$  where the elements of  $\epsilon$  are independent and standard normally distributed. Depending on the chapter, we will either perform analysis on  $\mathbf{Z}$  and  $\mathbf{w}$  or scale them and use  $\mathbf{X} = \frac{\mathbf{Z}}{\sqrt{n}}$  and  $\mathbf{y} = \frac{\mathbf{w}}{\sqrt{n}}$ . In the case of scaling, we will redefine  $\Sigma$  and  $\Sigma_{\mathbf{X}\mathbf{Y}}$  accordingly.

##### Independence of Samples and Errors

We assume the elements of  $\mathbf{Z}$  and the elements of  $\epsilon$  are independent.

##### Bounding (4.19) in the Fixed Design Setting

We recognize that in the random design setting  $\beta^* = \beta^0$ , but this is not the case in the fixed design setting. In the fixed design setting, we can use the triangle inequality to say

$$\left\| \beta^0 - \hat{\beta} \right\| \leq \left\| \beta^0 - \beta^* \right\| + \left\| \hat{\beta} - \beta^* \right\|. \quad (4.21)$$

It is well known that if  $d \leq n$  and  $\Sigma := \frac{\mathbf{Z}^T \mathbf{Z}}{n}$ , which we assume for fixed design, then

$$\|\beta^0 - \beta^*\|^2 \leq \frac{MSE(\mathbf{Z}\beta^*)}{\lambda_{\min}(\Sigma)} \quad (4.22)$$

where  $MSE(\mathbf{Z}\beta^*) := \frac{1}{n} \|\mathbf{Z}\beta^* - \mathbf{Z}\beta^0\|_2^2$ . By Theorem 2.2 in [68], we know that

$$MSE(\mathbf{Z}\beta^*) \leq \sigma^2 \frac{r + \log(1/\delta)}{n} \quad (4.23)$$

with probability  $1 - 1/\delta$  for any  $\delta > 0$  and where  $r = \text{Rank}(\mathbf{Z}^T \mathbf{Z})$ . Combining these, we can say

$$\|\beta^0 - \beta^*\|^2 \leq \frac{\sigma^2 (r + \log(1/\delta))}{\lambda_{\min}(\Sigma) n} \quad (4.24)$$

with the same probability  $1 - 1/\delta$ . Since we assume  $\Sigma$  has full rank, then  $r = d$  and we conclude

$$\|\beta^0 - \beta^*\| \leq \sqrt{\frac{\sigma^2 (d + \log(1/\delta))}{\lambda_{\min}(\Sigma) n}} \quad (4.25)$$

with high probability. Throughout this paper we will assume that  $\delta$  is chosen close enough to 1 that we can simply say

$$\|\beta^0 - \beta^*\| \leq \sqrt{\frac{\sigma^2 d}{\lambda_{\min}(\Sigma) n}} = \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right). \quad (4.26)$$

We will include this additional error in the error bounds in the fixed design cases.

### Sub-gaussian Behavior of Fixed $\mathbf{Z}$

We will make the assumption that a fixed design matrix  $\mathbf{Z}$  with bounded entries  $|z_{ij}| \leq R$  for all  $i, j$  behaves similarly to a bounded random design matrix such that, for arbitrary  $\boldsymbol{\beta}$ ,

$$\max_i \left| \mathbf{z}_i^T \boldsymbol{\beta} \right| \leq RC \sqrt{\log(n)} \|\boldsymbol{\beta}\|_2. \quad (4.27)$$

This is assumed as it eases calculations in the fixed design settings. We show here that this bound holds in the random setting. We can show using a Hoeffding bound that for a random design matrix with rows  $\mathbf{z}_i^T$  and bounded entries  $|z_{ij}| \leq r$  that

$$\mathbb{P} \left( \left| \mathbf{z}_i^T \boldsymbol{\beta} \right| \geq t \right) \leq 2 \exp \left( \frac{-t^2}{2R^2 \|\boldsymbol{\beta}\|_2^2} \right). \quad (4.28)$$

Then letting  $t = \sqrt{2R^2 \|\boldsymbol{\beta}\|_2^2 \log(n^2)}$  and using a union bound argument

$$\mathbb{P} \left( \max_i \left| \mathbf{z}_i^T \boldsymbol{\beta} \right| \geq t \right) \leq 2n \exp \left( \frac{-t^2}{2R^2 \|\boldsymbol{\beta}\|_2^2} \right) = \frac{2}{n}. \quad (4.29)$$

Thus, for a bounded random design matrix  $\mathbf{Z}$ , we can say

$$\max_i \left| \mathbf{z}_i^T \boldsymbol{\beta} \right| \leq \sqrt{2R^2 \|\boldsymbol{\beta}\|_2^2 \log(n^2)} = RC \sqrt{\log(n)} \|\boldsymbol{\beta}\|_2 \quad (4.30)$$

with high probability where  $C = 2$ , but more generally we can allow  $C > 0$ .

### Modified Big-O Notation

We introduce the notation  $\tilde{\mathcal{O}}(\cdot)$  to be a modified  $\mathcal{O}(\cdot)$  that simply removes constant and log terms.

## Indices

A note on indices: throughout this paper we will remain consistent with the use of indices. We will let  $i$  and  $i'$  range from 1 to  $n$ ,  $k$  range from 1 to  $m$ , and  $j$  and  $j'$  range from 1 to  $d$ .

## Operator Norm

Throughout the paper we will use the  $\ell^2 - \ell^2$  operator norm or spectral norm, denoted by  $\|\cdot\|$ , which yields the largest singular value of the matrix  $\mathbf{A}$ . That is:

$$\|\mathbf{A}\| = \sqrt{\max_i |\lambda_i(\mathbf{A}^* \mathbf{A})|}, \quad \text{for } \mathbf{A} \in \mathbb{R}^{n \times d} \quad (4.31)$$

where  $\lambda_i(\mathbf{A}^* \mathbf{A})$  denotes the  $i$ th eigenvalue of  $\mathbf{A}^* \mathbf{A}$  and  $\mathbf{A}^*$  denotes the conjugate transpose.

## 4.2 Quantized Regression Parameters

### 4.2.1 Introduction

Let  $\mathbf{Z}$  be a  $n \times d$  design matrix and let  $\mathbf{w}$  be a  $n \times 1$  response vector where it is assumed a linear relationship  $\mathbf{w} = \mathbf{Z}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\epsilon}$  exists. We assume the elements of  $\boldsymbol{\epsilon}$  are independent and standard normally distributed. Whether  $\mathbf{Z}$  is fixed or random will be stated at the beginning of each major section.

### Quantizers

We define a generic 1-bit stochastic quantizer for a bounded random variable  $U$  to be

$$Q_U(U) := \begin{cases} a & \text{with probability } \frac{b-U}{\Delta_U} \\ b & \text{with probability } \frac{U-a}{\Delta_U} \end{cases} \quad (4.32)$$

where  $Q_U(u)$  is defined on the range of  $U : [a, b]$  and  $\Delta_U := b - a$ .

Using this definition, we define three element-wise quantizers:

1.  $Q_{\mathbf{Z}}(z_{ij})$  is the quantizer defined on the range  $[-R, R]$  where  $R$  is the  $\max_{ij} |Z_{ij}| := R$  or is defined on the range of high-probability bounds of the elements of  $\mathbf{Z}$  as defined in Section 4.2.3 in the random design case. We define

$$\tilde{z}_{ij} := Q_{\mathbf{Z}}(z_{ij}). \quad (4.33)$$

We allow the notation  $\tilde{\mathbf{Z}}$  to be the matrix of quantized elements from  $\mathbf{Z}$  and  $\tilde{\mathbf{z}}_i^T$  to be the  $i$ th row of  $\tilde{\mathbf{Z}}$ .

2.  $Q_{\mathbf{Z}^2}(z_{ij}^2)$  is a quantizer defined on the interval  $[0, R^2]$  where  $R$  is again a bound (or high-probability bound) of the elements of  $\mathbf{Z}$ . We define

$$\tilde{z}_{ij}^2 := Q_{\mathbf{Z}^2}(z_{ij}^2). \quad (4.34)$$

3.  $Q_{\mathbf{w}}(w_i)$  is the quantizer defined on the high-probability range of  $w_i$ , which will be defined in Section 4.2.2 and 4.2.3. We define

$$\tilde{w}_i = Q_{\mathbf{w}}(w_i). \quad (4.35)$$

We allow the notation  $\tilde{\mathbf{w}}$  to be the vector of quantized  $\mathbf{w}$  values.

The exact ranges of these quantizers will be established in each section, as they depend on the assumptions made on  $\mathbf{Z}$ .

In section 4.2.2 we assume  $\mathbf{Z}$  to be fixed and in section 4.2.3 we assume  $\mathbf{Z}$  has standard normal independent elements.

### 4.2.2 Quantized Scenario with Fixed $\mathbf{Z}$

Throughout this section we will assume  $\mathbf{Z}$  is fixed and that the elements of  $\mathbf{Z}$ ,  $\{z_{ij}\}_{1 \leq i \leq n, 1 \leq j \leq d}$  are bounded by some  $R$ . That is,  $|z_{ij}| \leq R$  for all  $i, j$ . Thus, we use the definitions

$$\boldsymbol{\Sigma} := \frac{\mathbf{Z}^T \mathbf{Z}}{n} \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathbf{Z}\mathbf{w}} := \frac{\mathbf{Z}^T \mathbf{w}}{n}. \quad (4.36)$$

We define estimators of  $\boldsymbol{\Sigma}$  and  $\boldsymbol{\Sigma}_{\mathbf{Z}\mathbf{w}}$  in a 1-bit quantized setting with fixed  $\mathbf{Z}$  to be

$$\tilde{\boldsymbol{\Sigma}} := \frac{1}{n} \sum_{i=1}^n \tilde{\boldsymbol{\Sigma}}_i = \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T + \boldsymbol{\Delta}_i \quad (4.37)$$

$$\tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{w}} := \frac{\tilde{\mathbf{Z}}^T \tilde{\mathbf{w}}}{n} = \frac{1}{n} \sum_{i=1}^n \tilde{z}_{ij} \tilde{w}_i \quad \text{for } 1 \leq j \leq d \quad (4.38)$$

where we let  $\tilde{\boldsymbol{\Sigma}}_i = \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T + \boldsymbol{\Delta}_i$  where  $\boldsymbol{\Delta}_i = \text{diag} \left( \tilde{z}_{ij}^2 - \tilde{z}_{ij}^2 \right)_{j=1}^d$  be a modified sample estimate of  $\boldsymbol{\Sigma}$  based on the  $i$ th row of  $\tilde{\boldsymbol{\Sigma}}$ . These definition allow  $\tilde{\boldsymbol{\Sigma}}$  and  $\tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{w}}$  to be unbiased estimators, as shown in equations (C.1) and (C.2) in the appendix.

From these assumption and definitions we will prove

**Theorem 4.2.1.** *Let  $\mathbf{Z}$  be a design matrix whose elements  $z_{ij}$  are fixed and  $|z_{ij}| \leq R$  for some  $R \in \mathbb{R}^+$  for  $i = 1, \dots, n$  and  $j = 1, \dots, d$  and let  $\mathbf{w} = \{w_i\}_{i=1}^n$  be a vector of response variables. Suppose  $\mathbf{Z}$  and  $\mathbf{w}$  have a relationship such that  $\mathbf{w} = \mathbf{Z}\boldsymbol{\beta}^0 + \boldsymbol{\epsilon}$  where the elements of  $\boldsymbol{\epsilon}$  are standard normally distributed and independent and independent of the elements of  $\mathbf{Z}$ .*

*Then for  $\boldsymbol{\Sigma}$ ,  $\tilde{\boldsymbol{\Sigma}}$ ,  $\boldsymbol{\Sigma}_{\mathbf{Z}\mathbf{w}}$ , and  $\tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{w}}$ , as defined in (4.36), (4.37), and (4.38); for*

$$\begin{aligned} \tau_0 &= R^2 \|\boldsymbol{\Sigma}\| + \frac{R^4 (1 + 3d)}{d}, & b_0 &= 2R^2 + d^{-1} \|\boldsymbol{\Sigma}\|, \\ \ell &= RC\sqrt{\log(n)} \left\| \boldsymbol{\beta}^0 \right\|_2 + \sigma\sqrt{2\log(2n^2)}, & \eta &= \frac{2}{\lambda_{\min}(\boldsymbol{\Sigma})}; \end{aligned}$$

for some constant  $C > 0$ , and when  $n$  is chosen such that

$$n \geq \max \left\{ 2^3 d \log(2d) \eta^2 \left( 3r^4 + R^2 \|\Sigma\| + \frac{R^4}{d} \right), 2^2 d \log(2d) \eta b_0 \right\};$$

then with probability at least  $1 - 3/n$ ,

$$\begin{aligned} \|\beta^0 - \hat{\beta}\|_2 &\leq \frac{4\|\beta^*\|_2 \left( \sqrt{\frac{2d \log(2nd)\tau_0}{n}} + \frac{2d \log(2nd)b_0}{n} \right) + 4\sqrt{\frac{8R^2 \ell^2 d \log(2nd)}{n}}}{\lambda_{\min}(\Sigma)} + \sqrt{\frac{\sigma^2 d}{\lambda_{\min}(\Sigma) n}} \\ &= \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{n}} \right). \end{aligned}$$

### Proof of Theorem 4.2.1

*Proof.* We wish to bound

$$\|\hat{\beta} - \beta^0\|_2 \leq \frac{4\|\beta^*\|_2 \|\tilde{\Sigma} - \Sigma\| + 4\|(\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}})\|_2}{\lambda_{\min}(\Sigma)} + \|\beta^0 - \beta^*\|$$

by controlling

$$(a): \|\tilde{\Sigma} - \Sigma\| \quad (b): \|\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}}\|_2.$$

**Analyzing Part (a) in the Quantized Scenario with Fixed  $\mathbf{Z}$ .** By Lemma 3 we have

$$\|\tilde{\Sigma} - \Sigma\| \leq \sqrt{\frac{2 \log(2nd)\tau}{n}} + \frac{2 \log(2nd)b}{n} \quad (4.39)$$

with probability  $1 - 1/n$  where

$$\tau = d \left( R^2 \|\Sigma\| + \frac{R^4 (1 + 3d)}{d} \right) \quad \text{and} \quad b \geq 2dR^2 + \|\Sigma\|.$$

Recall in the formulation of the general problem (4.19) that we required

$$\|\tilde{\Sigma} - \Sigma\| \leq \frac{\lambda_{\min}(\Sigma)}{2}. \quad (4.40)$$

Thus, we wish to find a  $n$  such that

$$\sqrt{\frac{2 \log(2nd)\tau}{n}} + \frac{2 \log(2nd)b}{n} \leq \frac{\lambda_{\min}(\Sigma)}{2}. \quad (4.41)$$

It is sufficient to show that each term is less than  $\frac{\lambda_{\min}(\Sigma)}{4}$ . Let us consider each term individually. For the first term:

$$\begin{aligned} \sqrt{\frac{2 \log(2nd)\tau}{n}} &\leq \frac{\lambda_{\min}(\Sigma)}{4} \\ \implies \frac{\log(2nd)d \left( R^2 \|\Sigma\| + \frac{R^4(1+3d)}{d} \right)}{n} &\leq \frac{\lambda_{\min}^2(\Sigma)}{2^5} \\ \implies \frac{\log(2d)d \left( R^2 \|\Sigma\| + \frac{R^4(1+3d)}{d} \right)}{n} &\leq \frac{\lambda_{\min}^2(\Sigma)}{2^5} \\ \implies n &\geq \frac{2^5 \log(2d)d \left( R^2 \|\Sigma\| + \frac{R^4(1+3d)}{d} \right)}{\lambda_{\min}^2(\Sigma)} \\ \implies n &\geq 2^3 d \log(2d) \eta^2 \left( 3R^4 + R^2 \|\Sigma\| + \frac{R^4}{d} \right) \end{aligned} \quad (4.42)$$

where

$$\eta = \frac{2}{\lambda_{\min}(\Sigma)}. \quad (4.43)$$

The second term:

$$\begin{aligned} \frac{2 \log(2nd)b}{n} &\leq \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{4} \\ \frac{\log(2d)}{n} &\leq \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{2^3 b} \\ \implies n &\geq 2^2 \log(2d)\eta b. \end{aligned} \tag{4.44}$$

Thus, if we allow

$$n \geq \max \left\{ 2^3 d \log(2d)\eta^2 \left( 3R^4 + R^2 \|\boldsymbol{\Sigma}\| + \frac{R^4}{d} \right), 2^2 \log(2d)\eta b \right\}, \tag{4.45}$$

then we have guaranteed

$$\|\tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| \leq \sqrt{\frac{2 \log(2nd)\tau}{n}} + \frac{2 \log(2nd)b}{n} \leq \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{2}. \tag{4.46}$$

We have thus satisfied the requirement and we have bound part (a) with probability at least  $1 - 1/n$ .

**Analyzing Part (b) in the 1-bit Quantization Setting.** We now upper bound

$$\|\tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}_w} - \boldsymbol{\Sigma}_{\mathbf{Z}_w}\|_2 \leq \sqrt{d} \|\tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}_w} - \boldsymbol{\Sigma}_{\mathbf{Z}_w}\|_\infty.$$

Recall that we have defined  $\tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}_w}$  to be

$$\tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}_w} = \frac{\tilde{\mathbf{Z}}^T \tilde{\mathbf{w}}}{n} = \frac{1}{n} \sum_{i=1}^n \tilde{z}_{ij} \tilde{w}_i \quad \text{for } 1 \leq j \leq d. \tag{4.47}$$

Let us define

$$\psi_j = \sum_{i=1}^n \tilde{z}_{ji} \tilde{w}_i - \sum_{i'=1}^n z_{ji'} w_{i'} \quad (4.48)$$

so that

$$\frac{1}{n} \psi_j = \left( \tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}} \right)_j = \frac{1}{n} \sum_{i=1}^n \tilde{z}_{ji} \tilde{w}_i - \frac{1}{n} \sum_{i'=1}^n z_{ji'} w_{i'} \quad (4.49)$$

is the  $j$ th term of the difference between the quantized estimator and the term to be estimated.

Let us define the event

$$E : \max_i |w_i| \leq \ell \quad (4.50)$$

for  $\ell = RC\sqrt{\log(n)} \|\beta^0\|_2 + \sigma\sqrt{2\log(2n^2)}$ . We recognize under the condition of  $E$ ,  $\psi_j$  is the sum of zero-mean, bounded random variables in the region  $[-2r\ell, 2r\ell]$ . We can thus apply a Hoeffding bound and say

$$\mathbb{P}(|\psi_j| \geq t | E) \leq 2 \exp\left(\frac{-2t^2}{n(2R\ell + 2R\ell)^2}\right) = 2 \exp\left(\frac{-t^2}{8nR^2\ell^2}\right) \quad (4.51)$$

for all  $t > 0$  (see Wainwright Exercise 2.4, Example 2.4, and Equation 2.11) [69]. Using a union bound argument and letting  $t = \sqrt{8R^2\ell^2 n \log(2nd)}$  we have

$$\mathbb{P}\left(\max_j |\psi_j| \geq t \mid E\right) \leq 2d \exp\left(-\frac{t^2}{8nR^2\ell^2}\right) = \frac{1}{n}. \quad (4.52)$$

Applying the law of total probability, we recognize that

$$\mathbb{P}(A) = \mathbb{P}(A | E) \mathbb{P}(E) + \mathbb{P}(A | E^c) \mathbb{P}(E^c) \leq \mathbb{P}(A | E) + \mathbb{P}(E^c). \quad (4.53)$$

By Lemma 2 in section 4.2.2 we established that  $E$  occurs with probability  $1 - 1/n$ . Thus we can say

$$\mathbb{P}\left(\max_j \frac{1}{n} |\psi_j| \geq \frac{t}{n}\right) = \mathbb{P}\left(\max_j \left(\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}}\right)_j \geq \sqrt{\frac{8R^2\ell^2 \log(2nd)}{n}}\right) \leq 1/n + 1/n = 2/n. \quad (4.54)$$

Thus, we have bounded our desired quantity:

$$\left\|\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}}\right\|_2 \leq \sqrt{d} \left\|\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}}\right\|_\infty \leq \sqrt{\frac{8R^2\ell^2 d \log(2nd)}{n}} \quad (4.55)$$

with probability at least  $1 - 2/n$ .

Combining (4.39) and (4.55) results in

$$\left\|\hat{\beta} - \beta^*\right\|_2 \leq \frac{4\|\beta^*\|_2 \left(\sqrt{\frac{4\log(2nd)\tau}{n}} + \frac{4\log(2nd)b}{n}\right) + 4\sqrt{\frac{8R^2\ell^2 d \log(2nd)}{n}}}{\lambda_{\min}(\Sigma)}.$$

with probability at least  $1 - (1/n + 2/n)$ . We now add on the additional error from  $\left\|\hat{\beta} - \beta^*\right\|$  from the inequality

$$\left\|\beta^0 - \hat{\beta}\right\| \leq \left\|\hat{\beta} - \beta^*\right\| + \left\|\beta^0 - \beta^*\right\|. \quad (4.56)$$

Then with probability  $1 - (1/n + 2/n)$

$$\left\|\beta^0 - \hat{\beta}\right\|_2 \leq \frac{4\|\beta^*\|_2 \left(\sqrt{\frac{4\log(2nd)\tau}{n}} + \frac{4\log(2nd)b}{n}\right) + 4\sqrt{\frac{8R^2\ell^2 d \log(2nd)}{n}}}{\lambda_{\min}(\Sigma)} + \sqrt{\frac{\sigma^2 d}{\lambda_{\min}(\Sigma) n}}. \quad (4.57)$$

□

The following sections provide the lemmas used to prove Theorem 4.2.1.

### Establishing a High-Probability Bound on the Elements of $\mathbf{w}$ with Fixed $\mathbf{Z}$

We will establish a high-probability bound for the elements of  $\mathbf{w} = \mathbf{Z}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\epsilon}$ . Using this, we define

$$\ell := RC\sqrt{\log(n)}\left\|\boldsymbol{\beta}^0\right\|_2 + \sigma\sqrt{2\log(2n^2)} \quad (4.58)$$

for some constant  $C > 0$  so we can say that  $|w_i| \leq \ell$  for all  $i$  with probability at least  $1 - 1/n$ .

**Lemma 2.** *Let  $\mathbf{w} = \mathbf{Z}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\epsilon}$  for independent and standard normally distributed  $\boldsymbol{\epsilon}$  and  $\mathbf{Z}$  is defined as in 4.1. Then with probability at least  $1 - 1/n$*

$$\max_i |w_i| \leq RC\sqrt{\log(n)}\left\|\boldsymbol{\beta}^0\right\|_2 + \sigma\sqrt{2\log(2n^2)}. \quad (4.59)$$

*Proof.* We assumed the elements of  $\boldsymbol{\epsilon}$  to be independent and standard normally distributed. We can thus apply a Gaussian tail bound to say

$$\mathbb{P}(|\epsilon_i| \geq t) \leq 2\exp\left(-\frac{t^2}{2}\right) \quad (4.60)$$

for each  $i$ . Letting  $t = \sqrt{2\log(2n^2)}$  and using a union bound argument yields

$$\mathbb{P}\left(\max_i |\epsilon_i| \geq t\right) \leq 2n\exp\left(-\frac{t^2}{2}\right) = \frac{1}{n}. \quad (4.61)$$

Letting  $\mathbf{z}_i^T$  be the  $i$ th row of  $\mathbf{Z}$ , we assumed (see 4.1.4) that

$$\max_i \left|\mathbf{z}_i^T \boldsymbol{\beta}^0\right| \leq RC\sqrt{\log(n)}\left\|\boldsymbol{\beta}^0\right\|_2 \quad (4.62)$$

for some constant  $C > 0$ .

Then using the triangle inequality, we can say

$$|w_i| = \left| \mathbf{z}_i^T \boldsymbol{\beta}^0 + \sigma \epsilon_i \right| \leq \left| \mathbf{z}_i^T \boldsymbol{\beta}^0 \right| + \sigma |\epsilon_i| \leq RC \sqrt{\log(n)} \left\| \boldsymbol{\beta}^0 \right\|_2 + \sigma \sqrt{2 \log(2n^2)} \quad (4.63)$$

for all  $i$  with probability at least  $1 - 1/n$ . □

### Establishing the Bounds of the Quantizers with Fixed $\mathbf{Z}$

We must establish the bounds of our quantizers. We will use our assumption that  $|z_{ij}| \leq R$  and the bound for  $w_i$  established using Lemma 2 to define our bounds for our quantizers  $Q_{\mathbf{Z}}$  and  $Q_{\mathbf{Z}^2}$ . We have from Lemma 2 that

$$|w_i| \leq \ell = RC \sqrt{\log(n)} \left\| \boldsymbol{\beta}^0 \right\|_2 + \sigma \sqrt{2 \log(2n^2)} \quad (4.64)$$

for all  $i$  with probability  $1 - 1/n$ . Then our quantizers can be defined on:

	$\alpha^-$	$\alpha^+$	$\Delta$
$Q_{\mathbf{Z}}$	$-R$	$R$	$2R$
$Q_{\mathbf{Z}^2}$	$0$	$R^2$	$R^2$
$Q_{\mathbf{w}}$	$-\ell$	$\ell$	$2\ell$

where the bounds of  $Q_{\mathbf{w}}$  are with probability  $1 - 1/n$ .

### Matrix Bernstein Inequality for the Quantized Estimator with Fixed $\mathbf{Z}$

We apply the Matrix Bernstein Inequality to bound  $\left\| \tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \right\|$ .

**Lemma 3.** *Given the same assumptions on  $\mathbf{Z}$  and  $\mathbf{w}$  given in 4.1 and 4.2 and the definitions of  $\Sigma$  and  $\tilde{\Sigma}$  in (4.36) and (4.37). For*

$$\tau = d \left( R^2 \|\Sigma\| + \frac{R^4 (1 + 3d)}{d} \right) \quad \text{and} \quad b \geq 2dR^2 + \|\Sigma\|, \quad (4.65)$$

*we have with probability at least  $1 - 1/n$  that*

$$\|\tilde{\Sigma} - \Sigma\| \leq \sqrt{\frac{2 \log(2nd)\tau}{n}} + \frac{2 \log(2nd)b}{n}. \quad (4.66)$$

### Proof of Lemma 3

*Proof.* We begin the proof by noting a few important facts. First, we note that  $\tilde{z}_{ij}^2 \equiv R$  for all  $i, j$ , since  $\tilde{z}_{ij} \in \{-R, R\}$ . Second,  $\tilde{z}_{ij}^2 \in \{0, R^2\}$  for all  $i, j$ , since  $z_{ij}^2 \in [0, R^2]$ . Lastly, we see that  $\tilde{z}_{ij}^2 - z_{ij}^2 \leq 0$  for all  $i, j$ , allowing us to note that  $\Delta_i = \text{diag} \left( \tilde{z}_{ij}^2 - z_{ij}^2 \right)_{j=1}^d$  is a vector consisting of all non-positive values.

Recognizing that

$$\|\tilde{\Sigma} - \Sigma\| = \frac{1}{n} \left\| \sum_{i=1}^n (\tilde{\Sigma}_i - \Sigma) \right\|, \quad (4.67)$$

we can then apply a Matrix Bernstein inequality (Theorem 6.17 in Wainwright) [69] to say

$$\mathbb{P} \left( \frac{1}{n} \left\| \sum_{i=1}^n (\tilde{\Sigma}_i - \Sigma) \right\| \geq \theta \right) \leq 2 \text{rank} \left( \sum_{i=1}^n \mathbf{Var} [\tilde{\Sigma}_i] \right) \exp \left( \frac{-n\theta^2}{2(\sigma_{\tilde{\Sigma}}^2 + b\theta)} \right) \quad (4.68)$$

for some  $b > \|\tilde{\Sigma}_i - \Sigma\|$  and  $\theta \geq 0$  and where

$$\sigma_{\tilde{\Sigma}}^2 = \frac{1}{n} \left\| \sum_{i=1}^n \mathbf{Var} [\tilde{\Sigma}_i] \right\| = \frac{1}{n} \left\| n \mathbf{Var} [\tilde{\Sigma}_i] \right\| = \left\| \mathbf{Var} [\tilde{\Sigma}_i] \right\|. \quad (4.69)$$

In order to use the Matrix Bernstein inequality, we must show that  $\tilde{\Sigma}_i - \Sigma$  satisfies the condition

$$\left\| \tilde{\Sigma}_i - \Sigma \right\|_2 \leq b \quad (4.70)$$

almost surely. Using the boundedness of  $\mathbf{Z}$ , we first show that for all  $i$  and for any element of  $\tilde{\Sigma}_i$

$$\begin{aligned} \left( \tilde{\Sigma}_i \right)_{j,j'} &= \left( \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right)_{j,j'} - \left( \text{diag} \left( \tilde{z}_{i_j}^2 - R^2 \right)_{j=1}^d \right)_{j,j'} \\ &\leq \max \left\{ -R^2, R^2 \right\} - \min \left\{ 0, -R^2 \right\} \leq R^2 + R^2 = 2R^2. \end{aligned} \quad (4.71)$$

Using this, we can then show

$$\begin{aligned} \left\| \tilde{\Sigma}_i - \Sigma \right\|_2^2 &\leq \left\| \tilde{\Sigma}_i \right\|_2^2 + \left\| \Sigma \right\|_2^2 = \sup_{\|\mathbf{v}\|_2=1} \mathbf{v}^T \tilde{\Sigma}_i^T \tilde{\Sigma}_i \mathbf{v} + \left\| \Sigma \right\|_2^2 \\ &= \sup_{\|\mathbf{v}\|_2=1} \left| \sum_{j,j'} v_j v_{j'} \left( \sum_{\ell} \left( \tilde{\Sigma}_i \right)_{\ell j} \left( \tilde{\Sigma}_i \right)_{\ell j'} \right) \right| + \left\| \Sigma \right\|_2^2 \\ &\leq \sup_{\|\mathbf{v}\|_2=1} \sum_{j,j'} |v_j| |v_{j'}| \left| \sum_{\ell} \left( \tilde{\Sigma}_i \right)_{\ell j} \left( \tilde{\Sigma}_i \right)_{\ell j'} \right| + \left\| \Sigma \right\|_2^2 \\ &\leq d \left( 2r^2 \right)^2 \sup_{\|\mathbf{v}\|_2=1} \|\mathbf{v}\|_1^2 + \left\| \Sigma \right\|_2^2 = d^2 \left( 2r^2 \right)^2 + \left\| \Sigma \right\|_2^2 \\ &\implies \left\| \tilde{\Sigma}_i - \Sigma \right\|_2 \leq 2dR^2 + \left\| \Sigma \right\|_2, \end{aligned} \quad (4.72)$$

satisfying the Bernstein condition.

We proceed by bounding  $\sigma_{\tilde{\Sigma}}^2$ , which we defined previously as

$$\sigma_{\tilde{\Sigma}}^2 = \left\| \mathbf{Var} \left[ \tilde{\Sigma}_i \right] \right\|. \quad (4.73)$$

Using the definition of variance and the unbiasedness of our estimator, we can show that

$\mathbf{Var} \left[ \tilde{\boldsymbol{\Sigma}}_i \right]$  can be written as

$$\begin{aligned}
 \mathbf{Var} \left[ \tilde{\boldsymbol{\Sigma}}_i \right] &= \mathbf{E} \left[ \tilde{\boldsymbol{\Sigma}}_i^2 \right] - \mathbf{E} \left[ \tilde{\boldsymbol{\Sigma}}_i \right]^2 = \mathbf{E} \left[ \tilde{\boldsymbol{\Sigma}}_i^2 \right] - \boldsymbol{\Sigma}^2 \\
 &= \mathbf{E} \left[ (\tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T + \boldsymbol{\Delta}_i)^T (\tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T + \boldsymbol{\Delta}_i) \right] - \boldsymbol{\Sigma}^2 \\
 &= \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right] + \mathbf{E} \left[ \boldsymbol{\Delta}_i^2 \right] + \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \boldsymbol{\Delta}_i \right] + \mathbf{E} \left[ \boldsymbol{\Delta}_i^T \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right] - \boldsymbol{\Sigma}^2 \\
 &= \mathbf{E} \left[ \tilde{\mathbf{z}}_i \|\tilde{\mathbf{z}}_i\|_2^2 \tilde{\mathbf{z}}_i^T \right] + \mathbf{E} \left[ \boldsymbol{\Delta}_i^2 \right] + \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \boldsymbol{\Delta}_i \right] + \mathbf{E} \left[ \boldsymbol{\Delta}_i \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right] - \boldsymbol{\Sigma}^2 \\
 &= dr^2 \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right] + \mathbf{E} \left[ \boldsymbol{\Delta}_i^2 \right] + \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \boldsymbol{\Delta}_i \right] + \mathbf{E} \left[ \boldsymbol{\Delta}_i \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right] - \boldsymbol{\Sigma}^2 \tag{4.74}
 \end{aligned}$$

We can represent  $\mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right]$  as

$$\begin{aligned}
 \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right] &= \begin{bmatrix} \mathbf{E} \left[ \tilde{z}_{i1}^2 \right] & \mathbf{E} \left[ \tilde{z}_{i1} \tilde{z}_{i2} \right] & \dots & \mathbf{E} \left[ \tilde{z}_{i1} \tilde{z}_{id} \right] \\ \mathbf{E} \left[ \tilde{z}_{i2} \tilde{z}_{i1} \right] & \mathbf{E} \left[ \tilde{z}_{i2}^2 \right] & \dots & \mathbf{E} \left[ \tilde{z}_{i2} \tilde{z}_{id} \right] \\ \vdots & \vdots & \ddots & \dots \\ \mathbf{E} \left[ \tilde{z}_{id} \tilde{z}_{i1} \right] & \mathbf{E} \left[ \tilde{z}_{id} \tilde{z}_{i2} \right] & \dots & \mathbf{E} \left[ \tilde{z}_{id}^2 \right] \end{bmatrix} = \begin{bmatrix} R^2 & z_{i1} z_{i2} & \dots & z_{i1} z_{id} \\ z_{i2} z_{i1} & R^2 & \dots & z_{i2} z_{id} \\ \vdots & \vdots & \ddots & \dots \\ z_{id} z_{i1} & z_{id} z_{i2} & \dots & R^2 \end{bmatrix} \\
 &= \mathbf{z}_i \mathbf{z}_i^T + \begin{bmatrix} R^2 - z_{i1}^2 & 0 & \dots & 0 \\ 0 & R^2 - z_{i1}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & R^2 - z_{id}^2 \end{bmatrix} = \boldsymbol{\Sigma} + \begin{bmatrix} R^2 - z_{i1}^2 & 0 & \dots & 0 \\ 0 & R^2 - z_{i1}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & R^2 - z_{id}^2 \end{bmatrix} \\
 &= \boldsymbol{\Sigma} + \mathbf{D} \tag{4.75}
 \end{aligned}$$

where we added and subtracted a diagonal matrix with entries  $\left\{ z_{ij}^2 \right\}_{j=1}^d$ , and we defined

$\mathbf{D} = \text{diag} \left( R^2 - z_{ij}^2 \right)_{j=1}^d$ . Viewing  $\mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right]$  this way allows us to more easily bound its

spectral norm as:

$$\left\| \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right] \right\| = \left\| \boldsymbol{\Sigma} + \mathbf{D} \right\| \leq \left\| \boldsymbol{\Sigma} \right\| + \left\| \mathbf{D} \right\| \leq \left\| \boldsymbol{\Sigma} \right\| + R^2. \quad (4.76)$$

Now we are able to bound  $\sigma_{\tilde{\boldsymbol{\Sigma}}}^2$ :

$$\begin{aligned} \sigma_{\tilde{\boldsymbol{\Sigma}}}^2 &= \left\| \mathbf{Var} \left[ \tilde{\boldsymbol{\Sigma}}_i \right] \right\| = \left\| dr^2 \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right] + \mathbf{E} \left[ \boldsymbol{\Delta}_i^2 \right] + 2 \mathbf{E} \left[ \boldsymbol{\Delta}_i \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right] - \boldsymbol{\Sigma}^2 \right\| \\ &\leq \left\| dr^2 \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right] + \mathbf{E} \left[ \boldsymbol{\Delta}_i^2 \right] + 2 \mathbf{E} \left[ \boldsymbol{\Delta}_i \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right] \right\| \\ &\leq \left\| dr^2 \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right] \right\| + \left\| \mathbf{E} \left[ \boldsymbol{\Delta}_i^2 \right] \right\| + 2 \left\| \mathbf{E} \left[ \boldsymbol{\Delta}_i \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right] \right\| \\ &\leq dr^2 \left\| \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right] \right\| + R^4 + 2 \mathbf{E} \left[ \left\| \boldsymbol{\Delta}_i \right\| \left\| \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right\| \right] \\ &\leq dr^2 \left( \left\| \boldsymbol{\Sigma} \right\| + R^2 \right) + R^4 + 2dR^4 \\ &= dr^2 \left\| \boldsymbol{\Sigma} \right\| + R^4(1 + 3d) \\ &= d \left( R^2 \left\| \boldsymbol{\Sigma} \right\| + \frac{R^4(1 + 3d)}{d} \right) \end{aligned} \quad (4.77)$$

where we used Jensen's inequality and the convexity of the spectral norm. We were able to drop  $\left\| \boldsymbol{\Sigma}^2 \right\|$  in the second step using the fact that  $\mathbf{Var} \left[ \tilde{\boldsymbol{\Sigma}}_i \right] \geq 0$  and  $\boldsymbol{\Sigma} \geq 0$  and applying (D.1).

Let us define the value  $\tau$  to be the upper bound established for  $\sigma_{\tilde{\boldsymbol{\Sigma}}}^2$ :

$$\tau := d \left( R^2 \left\| \boldsymbol{\Sigma} \right\| + \frac{R^4(1 + 3d)}{d} \right). \quad (4.78)$$

Recognizing that  $\mathbf{Var} [\tilde{\Sigma}_i - \Sigma] = \mathbf{Var} [\tilde{\Sigma}_i]$ , and since  $\sigma_{\tilde{\Sigma}}^2 \leq \tau$ , then we can say

$$\mathbb{P} \left( \frac{1}{n} \left\| \sum_{i=1}^n (\tilde{\Sigma}_i - \Sigma) \right\| \geq \theta \right) \leq 2d \exp \left( \frac{-n\theta^2}{2(\sigma_{\tilde{\Sigma}}^2 + b\theta)} \right) \leq 2d \exp \left( \frac{-n\theta^2}{2(\tau + b\theta)} \right). \quad (4.79)$$

In order to further refine this probability, we examine possible choices of  $\theta$  and  $b$ . We wish to find a  $\theta$  and  $b$  such that

$$\frac{n\theta^2}{2(\tau + b\theta)} \geq \min \left\{ \frac{n\theta^2}{2\tau}, \frac{n\theta}{2b} \right\} \geq \log(2nd) \quad (4.80)$$

Let us choose

$$\theta = \sqrt{\frac{2\log(2nd)\tau}{n}} + \frac{2\log(2nd)b}{n}. \quad (4.81)$$

We can then show

$$\begin{aligned} \frac{n\theta^2}{2\sigma^2} &\geq \frac{n \left( \sqrt{\frac{2\log(2nd)\tau}{n}} + \frac{2\log(2nd)b}{n} \right)^2}{2\tau} \\ &= \frac{n \left( \frac{2\log(2nd)\tau}{n} + \frac{4\log^2(2nd)b^2}{n^2} + \frac{4\log(2nd)b}{n} \sqrt{\frac{2\log(2nd)\tau}{n}} \right)}{2\tau} \\ &= \log(2nd) + \frac{2\log^2(2nd)b^2}{n\tau} + \frac{b\sqrt{4\log^3(2nd)\tau}}{\tau\sqrt{n}} \\ &\geq \log(2nd), \end{aligned} \quad (4.82)$$

and similarly

$$\frac{n\theta}{2b} = \frac{n \left( \sqrt{\frac{2\log(2nd)\tau}{n}} + \frac{2\log(2nd)b}{n} \right)}{2b} = \log(2nd) + \frac{\sqrt{2\log(2nd)\tau}}{2b} \geq \log(2nd). \quad (4.83)$$

Now we can say

$$\mathbb{P}\left(\frac{1}{n}\left\|\sum_{i=1}^n(\tilde{\Sigma}_i - \Sigma)\right\|\geq\theta\right)\leq 2d\exp\left(\frac{-n\theta^2}{2(\tau+b\theta)}\right)\leq 2d\exp(-\log(2nd))=\frac{1}{n}\quad(4.84)$$

for  $\theta=\sqrt{\frac{2\log(2nd)\tau}{n}}+\frac{2\log(2nd)b}{n}$ ,  $\tau=d\left(R^2\|\Sigma\|+\frac{R^4(1+3d)}{d}\right)$ , and some  $b\geq 2dR^2+\|\Sigma\|$ .  $\square$

Thus, we have bound the spectral norm of the error between our estimator and its true value, namely:

$$\|\tilde{\Sigma}-\Sigma\|\leq\sqrt{\frac{2\log(2nd)\tau}{n}}+\frac{2\log(2nd)b}{n}.\quad(4.85)$$

with probability at least  $1-1/n$ .

### Summary of the Quantized Scenario with Fixed $\mathbf{Z}$

This section summarizes the definitions and conclusions made throughout this chapter with the addition of

For fixed  $\mathbf{Z}$  with bounded elements  $|z_{ij}|\leq R$  for all  $i, j$  and definitions

$$\Sigma=\frac{\mathbf{Z}^T\mathbf{Z}}{n}\qquad\tilde{\Sigma}=\frac{1}{n}\sum_{i=1}^n\tilde{\Sigma}_i=\frac{1}{n}\sum_{i=1}^n\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^T+\mathbf{\Delta}_i\quad(4.86)$$

$$\Sigma_{\mathbf{Z}\mathbf{w}}=\frac{\mathbf{Z}^T\mathbf{w}}{n}\qquad\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}}=\frac{\tilde{\mathbf{Z}}^T\tilde{\mathbf{w}}}{n}=\frac{1}{n}\sum_{i=1}^n\tilde{z}_{ij}\tilde{w}_i\quad\text{for }1\leq j\leq d,\quad(4.87)$$

we showed that

$$\|\tilde{\Sigma}-\Sigma\|\leq\sqrt{\frac{2\log(2nd)\tau}{n}}+\frac{2\log(2nd)b}{n}\quad\text{w.p. }1-1/n\quad(4.88)$$

$$\|\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}}-\Sigma_{\mathbf{Z}\mathbf{w}}\|_2\leq\sqrt{\frac{8R^2\ell^2d\log(2nd)}{n}}\quad\text{w.p. }1-2/n\quad(4.89)$$

for  $b \geq 2dR^2 + \|\Sigma\|$  and  $C \geq 0$  and where

$$\tau = d \left( R^2 \|\Sigma\| + \frac{R^4(1+3d)}{d} \right) \quad \ell = RC\sqrt{\log(n)} \|\beta^0\|_2 + \sigma\sqrt{2\log(2n^2)}.$$

We can summarize by stating

$$\|\tilde{\Sigma} - \Sigma\| = \mathcal{O} \left( \sqrt{\frac{(dR^2 + R^4) \log(nd)}{n}} \right) = \mathcal{O} \left( \sqrt{\frac{d \log(nd)}{n}} \right) \quad (4.90)$$

$$\|\tilde{\Sigma}_{\mathbf{Zw}} - \Sigma_{\mathbf{Zw}}\|_2 = \mathcal{O} \left( \sqrt{\frac{R^2 \ell^2 d \log(nd)}{n}} \right) = \mathcal{O} \left( \sqrt{\frac{d \log(nd) \log(n)}{n}} \right). \quad (4.91)$$

We include here an interim step where we retain the  $r$  and  $\ell$  to illustrate the error bound's dependence on these values. Ultimately, these values are either a constant or dependent on  $n$  and  $d$  and are thus removed in the final  $\mathcal{O}$  formulation.

We introduce a new notation that removes the constants and log terms  $\tilde{\mathcal{O}}(\cdot)$ :

$$\|\tilde{\Sigma} - \Sigma\| \leq \sqrt{\frac{2 \log(2nd) \tau}{n}} + \frac{2 \log(2nd) b}{n} = \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{n}} \right) \quad (4.92)$$

$$\|\tilde{\Sigma}_{\mathbf{Zw}} - \Sigma_{\mathbf{Zw}}\|_2 \leq \sqrt{\frac{8R^2 \ell^2 d \log(2nd)}{n}} = \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{n}} \right). \quad (4.93)$$

We identified that

$$n \geq \max \left\{ 2^3 d \log(2d) \eta^2 \left( 3R^4 + R^2 \|\Sigma\| + \frac{R^4}{d} \right), 2^2 \log(2d) \eta b \right\} \quad (4.94)$$

for  $\eta = \frac{2}{\lambda_{\min}(\boldsymbol{\Sigma})}$  satisfies the requirement for our general problem allowing us to conclude

$$\begin{aligned} \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 &\leq \frac{4\|\boldsymbol{\beta}^*\|_2 \|\tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| + 4\|(\tilde{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{Y}})\|_2}{\lambda_{\min}(\boldsymbol{\Sigma})} \\ &\leq \frac{4\|\boldsymbol{\beta}^*\|_2 \left( \sqrt{\frac{2d \log(2nd)\tau_0}{n}} + \frac{2d \log(2nd)b_0}{n} \right) + 4\sqrt{\frac{8R^2 \ell^2 d \log(2nd)}{n}}}{\lambda_{\min}(\boldsymbol{\Sigma})} \end{aligned} \quad (4.95)$$

$$= \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right) \quad (4.96)$$

with probability at least  $1 - (1/n + 2/n)$ . We define  $\tau_0 = R^2 \|\boldsymbol{\Sigma}\| + \frac{R^4(1+3d)}{d}$  and  $b_0 = 2R^2 + d^{-1} \|\boldsymbol{\Sigma}\|$  to allow for more easily recognized dependence on  $d$  in the final formulation.

We now add on the additional error from  $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|$  from the inequality

$$\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\| \leq \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^*\| + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \quad (4.97)$$

and say that,

$$\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\|_2 \leq \frac{4\|\boldsymbol{\beta}^*\|_2 \left( \sqrt{\frac{2d \log(2nd)\tau_0}{n}} + \frac{2d \log(2nd)b_0}{n} \right) + 4\sqrt{\frac{8R^2 \ell^2 d \log(2nd)}{n}}}{\lambda_{\min}(\boldsymbol{\Sigma})} + \sqrt{\frac{\sigma^2 d}{\lambda_{\min}(\boldsymbol{\Sigma}) n}} \quad (4.98)$$

$$= \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right) + \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right) = \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right) \quad (4.99)$$

with probability  $1 - 3/n$ .

### 4.2.3 Quantized Scenario with Gaussian $\mathbf{Z}$

Let us assume the rows of  $\mathbf{Z}$ , namely  $\mathbf{z}_i^T$ , are drawn independently from a multivariate normal distribution with mean  $\mathbf{0}$  and covariance  $\mathbf{I}_{d \times d}$ . Let  $\mathbf{w} = \mathbf{Z}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\epsilon}$  for independent

and standard normally distributed  $\epsilon_i$  elements of  $\epsilon$ . We thus use the definitions

$$\Sigma = \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{Z}}{n} \right] = \mathbf{I}_d \quad (4.100)$$

$$\Sigma_{\mathbf{Z}\mathbf{w}} = \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{w}}{n} \right]. \quad (4.101)$$

and we define their estimators

$$\tilde{\Sigma} = \frac{1}{n} \sum_{i=1}^n \tilde{\Sigma}_i = \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T + \Delta_i \quad (4.102)$$

$$\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} = \frac{\tilde{\mathbf{Z}}^T \tilde{\mathbf{w}}}{n} = \frac{1}{n} \sum_{i=1}^n \tilde{z}_{ij} \tilde{w}_i \quad \text{for } 1 \leq j \leq d \quad (4.103)$$

where  $\Delta_i = \text{diag} \left( \tilde{z}_{ij}^2 - z_{ij}^2 \right)_{j=1}^d$  and the  $\tilde{\cdot}$  notation denotes the quantization as defined in Section 4.1. We show in the the appendix in equations (C.3) and (C.4) that these estimators are unbiased.

**Theorem 4.2.2.** *Let  $\mathbf{Z}$  be a design matrix with standard normally distributed and independent elements  $z_{ij}$  for  $i = 1, \dots, n$  and  $j = 1, \dots, d$ . Let  $\mathbf{w} = \mathbf{Z}\beta^0 + \sigma\epsilon$  where the elements of  $\epsilon$  are standard normally distributed and independent of each other and of the elements of  $\mathbf{Z}$ . Let*

$$\Sigma := \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{Z}}{n} \right] = \mathbf{I}_d \quad \text{and} \quad \Sigma_{\mathbf{Z}\mathbf{w}} := \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{w}}{n} \right].$$

and their estimators

$$\tilde{\Sigma} := \frac{1}{n} \sum_{i=1}^n \tilde{\Sigma}_i = \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T + \Delta_i$$

$$\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} := \frac{\tilde{\mathbf{Z}}^T \tilde{\mathbf{w}}}{n} = \frac{1}{n} \sum_{i=1}^n \tilde{z}_{ij} \tilde{w}_i \quad \text{for } 1 \leq j \leq d.$$

Then for

$$\eta = \frac{2}{\lambda_{\min}(\boldsymbol{\Sigma})}, \quad \tau_0 = \log^2(2n^2d) \left(12 + \frac{4}{d}\right), \quad b_0 = 2R_g^2 + d^{-1} \|\boldsymbol{\Sigma}\|,$$

$$\ell_g = \sqrt{2 \left( \|\boldsymbol{\beta}^0\|_2^2 + \sigma^2 \right) \log(2n^2)}, \quad R_g = \sqrt{2 \log(2n^2d)},$$

and ensuring  $n$  is selected such that

$$n \geq \max \left\{ 2^3 d \log^3(2d) \eta^2 \left(12 + \frac{4}{d}\right), 2^2 d \log(2d) b_0 \eta \right\} \quad (4.104)$$

then with probability at least  $1 - 5/n$

$$\begin{aligned} \|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\|_2 &\leq \frac{4\|\boldsymbol{\beta}^*\|_2 \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| + 4\|(\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{Y}})\|_2}{\lambda_{\min}(\boldsymbol{\Sigma})} \\ &\leq \frac{4\|\boldsymbol{\beta}^*\|_2 \left( \sqrt{\frac{2d \log(2nd)\tau_0}{n}} + \frac{2d \log(2nd)b_0}{n} \right) + 4\sqrt{\frac{8dR_g^2 \ell_g^2 \log(2nd)}{n}}}{\lambda_{\min}(\boldsymbol{\Sigma})} = \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{n}} \right) \end{aligned} \quad (4.105)$$

where  $\tilde{\mathcal{O}}(\cdot)$  denotes the order with log terms removed.

### Proof of Theorem 4.2.2

*Proof.* We wish to bound

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2 \leq \frac{4\|\boldsymbol{\beta}^*\|_2 \|\tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| + 4\|(\tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{W}} - \boldsymbol{\Sigma}_{\mathbf{Z}\mathbf{W}})\|_2}{\lambda_{\min}(\boldsymbol{\Sigma})}$$

by controlling

$$(a): \|\tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| \quad (b): \|\tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{W}} - \boldsymbol{\Sigma}_{\mathbf{Z}\mathbf{W}}\|_2.$$

**Analyzing Part (a) in the Quantized Scenario with Gaussian  $\mathbf{Z}$**  We will establish a high-probability bound for  $\left\| \tilde{\Sigma} - \Sigma \right\|$  under Gaussian assumptions on  $\mathbf{Z}$ .

We can apply the results established in section 4.2.3 directly to conclude that

$$\frac{1}{n} \left\| \sum_{i=1}^n (\tilde{\Sigma}_i - \Sigma) \right\| = \left\| \tilde{\Sigma} - \Sigma \right\| \leq \sqrt{\frac{2 \log(2nd)\tau}{n}} + \frac{2 \log(2nd)b}{n} \quad (4.106)$$

for  $\tau = d \log^2(2n^2d) \left(12 + \frac{4}{d}\right)$ , and for any  $b \geq d \left(2R_g^2 + \frac{r^2}{n}\right)$  with probability  $1 - 2/n$ . To satisfy the lambda min requirement (4.17), it is sufficient to show that

$$\sqrt{\frac{2 \log(2nd)\tau}{n}} \leq \frac{\lambda_{\min}(\Sigma)}{4} \quad \text{and} \quad \frac{2 \log(2nd)b}{n} \leq \frac{\lambda_{\min}(\Sigma)}{4}. \quad (4.107)$$

We will examine each terms individually beginning with the first:

$$\begin{aligned} & \sqrt{\frac{2 \log(2nd)\tau}{n}} \leq \frac{\lambda_{\min}(\Sigma)}{4} \\ \implies & \sqrt{\frac{2 \log(2nd)d \log^2(2n^2d) \left(12 + \frac{4}{d}\right)}{n}} \leq \frac{\lambda_{\min}(\Sigma)}{4} \\ \implies & \frac{2d \log^3(2d) \left(12 + \frac{4}{d}\right)}{n} \leq \frac{\lambda_{\min}^2(\Sigma)}{2^4} \\ \implies & n \geq \frac{2^5 d \log^3(2d) \left(12 + \frac{4}{d}\right)}{\lambda_{\min}^2(\Sigma)} \\ \implies & n \geq 2^3 d \log^3(2d) \left(12 + \frac{4}{d}\right) \frac{4}{\lambda_{\min}^2(\Sigma)} \\ \implies & n \geq 2^3 d \log^3(2d) \eta^2 \left(12 + \frac{4}{d}\right) \end{aligned} \quad (4.108)$$

where

$$\eta = \frac{2}{\lambda_{\min}(\boldsymbol{\Sigma})}. \quad (4.109)$$

The second term:

$$\begin{aligned} \frac{2 \log(2nd)b}{n} &\leq \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{4} \\ \implies \frac{2 \log(2d)b}{n} &\leq \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{4} \\ \implies n &\geq 4 \log(2d)b\eta. \end{aligned} \quad (4.110)$$

Thus, by selecting

$$n \geq \max \left\{ 2^3 d \log^3(2d)\eta^2 \left( 12 + \frac{4}{d} \right), 2^2 \log(2d)b\eta \right\} \quad (4.111)$$

we guarantee we satisfy the requirement.

**Analyzing Part (b) in the Quantized Scenario with Gaussian  $\mathbf{Z}$**  We now upper bound

$$\left\| \tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{w}} - \boldsymbol{\Sigma}_{\mathbf{Z}\mathbf{w}} \right\|_2 \leq \sqrt{d} \left\| \tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{w}} - \boldsymbol{\Sigma}_{\mathbf{Z}\mathbf{w}} \right\|_{\infty}. \quad (4.112)$$

Recall that we have defined  $\tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{w}}$  to be

$$\tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{w}} = \frac{\tilde{\mathbf{Z}}^T \tilde{\mathbf{w}}}{n} = \frac{1}{n} \sum_{i=1}^n \tilde{z}_{ij} \tilde{w}_i \quad \text{for } 1 \leq j \leq d. \quad (4.113)$$

Let us define

$$\psi_j = \sum_{i=1}^n \tilde{z}_{ji} \tilde{w}_i - \sum_{i'}^n \mathbf{E} [z_{ji'} w_{i'}] \quad (4.114)$$

so that

$$\frac{1}{n}\psi_j = \left(\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}}\right)_j = \frac{1}{n}\sum_{i=1}^n \tilde{z}_{ji}\tilde{w}_i - \frac{1}{n}\sum_{i'}^n \mathbf{E}[z_{ji'}w_{i'}] \quad (4.115)$$

is the  $j$ th term of the difference between the quantized estimator and the term to be estimated. We recognize that by conditioning on the event

$$E : \left(\max_i |w_i| \leq \ell_g \text{ AND } \max_{ij} |z_{ij}| \leq R_g\right) \quad (4.116)$$

the random variable  $\psi_j | E$  is the sum zero-mean, bounded random variables in the region  $[-2R_g\ell_g, 2R_g\ell_g]$ . Thus, we can apply a Hoeffding bound and say

$$\mathbb{P}(|\psi_j| \geq t | E) \leq 2 \exp\left(-\frac{t^2}{8nR_g^2\ell_g^2}\right) \quad (4.117)$$

for all  $t > 0$  (see Wainwright Exercise 2.4, Example 2.4, and Equation 2.11) [69]. Using a union bound argument and letting  $t = \sqrt{8nR_g^2\ell_g^2 \log(2nd)}$  we have

$$\mathbb{P}\left(\max_j |\psi_j| \geq t \mid E\right) \leq 2d \exp\left(-\frac{t^2}{8nR_g^2\ell_g^2}\right) = \frac{1}{n}, \quad (4.118)$$

which allows us to say

$$\mathbb{P}\left(\max_j \left(\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}}\right)_j \geq \frac{t}{n} \mid E\right) = \mathbb{P}\left(\max_j \left(\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}}\right)_j \geq \sqrt{\frac{8R_g^2\ell_g^2 \log(2nd)}{n}} \mid E\right) \leq \frac{1}{n}. \quad (4.119)$$

Applying the law of total probability with events

$$A : \max_j \left( \tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}} \right)_j \geq \sqrt{\frac{8R_g^2 \ell_g^2 \log(2nd)}{n}} \quad (4.120)$$

$$E : \left( \max_i |w_i| \leq \ell_g \text{ AND } \max_{ij} |z_{ij}| \leq R_g \right), \quad (4.121)$$

we recognize that

$$\mathbb{P}(A) = \mathbb{P}(A|E)\mathbb{P}(E) + \mathbb{P}(A|E^c)\mathbb{P}(E^c) \leq \mathbb{P}(A|E) + \mathbb{P}(E^c). \quad (4.122)$$

In section 4.2.3 we established that  $\max_i |w_i| \leq \ell_g$  with probability  $1 - 1/n$ , and in section 4.2.3 we established that  $\max_{ij} |z_{ij}| \leq R_g$ , also with probability  $1 - 1/n$ . Thus we can say

$$\begin{aligned} & \mathbb{P} \left( \max_j \left( \tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}} \right)_j \geq \sqrt{\frac{8R_g^2 \ell_g^2 \log(2nd)}{n}} \right) \\ & \leq \mathbb{P} \left( \max_j \left( \tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}} \right)_j \geq \sqrt{\frac{8R_g^2 \ell_g^2 \log(2nd)}{n}} \mid E \right) + \mathbb{P} \left( \max_i |w_i| \geq \ell_g \text{ OR } \max_{ij} |z_{ij}| \geq R_g \right) \\ & = 3/n. \end{aligned} \quad (4.123)$$

Thus, we have bounded our desired quantity:

$$\left\| \tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}} \right\|_2 \leq \sqrt{d} \left\| \tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}} \right\|_\infty \leq \sqrt{\frac{8dR_g^2 \ell_g^2 \log(2nd)}{n}} \quad (4.124)$$

with probability at least  $1 - 3/n$ .

Combining (4.106) and (4.124) we can show

$$\left\| \hat{\beta} - \beta^0 \right\|_2 \leq \frac{4 \left\| \beta^* \right\|_2 \left( \sqrt{\frac{2 \log(2nd)\tau}{n}} + \frac{2 \log(2nd)b}{n} \right) + \sqrt{\frac{8dR_g^2 \ell_g^2 \log(2nd)}{n}}}{\lambda_{\min}(\Sigma)} = \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{n}} \right).$$

with probability at least  $1 - 5/n$ . □

### Establishing a High-Probability Bound on the Elements of a Gaussian $\mathbf{Z}$

We will establish a high-probability bound for the elements of  $\mathbf{Z}$  where the rows of  $\mathbf{Z}$  are independent and multivariate standard normally distributed. Then we define

$$R_g := \sqrt{2 \log(2n^2d)} \tag{4.125}$$

so we can then say  $|z_{ij}| \leq R_g$  for all  $i, j$  with probability at least  $1 - 1/n$ .

**Lemma 4.** *Let  $\mathbf{Z}$  have independent and multivariate standard normally distributed elements  $z_{ij}$ . Then*

$$|z_{ij}| \geq \sqrt{2 \log(2n^2d)}. \tag{4.126}$$

*with probability at least  $1 - 1/n$ .*

#### Proof of Lemma 4

*Proof.* For an element  $z_{ij}$  of  $\mathbf{Z}$ ,  $z_{ij} \sim N(0, 1)$ . Thus, we can use a Gaussian tail bound to state

$$\mathbb{P}(|z_{ij}| \geq t) \leq 2 \exp\left(-\frac{t^2}{2}\right). \tag{4.127}$$

Letting  $t = \sqrt{2 \log(2n^2d)}$  and using a union bound we can say

$$\mathbb{P}\left(\max_{ij} |z_{ij}| \geq t\right) \leq 2nd \exp\left(-\frac{t^2}{2}\right) = \frac{1}{n}. \tag{4.128}$$

□

### Establishing a High-Probability Bound on the Elements of $\mathbf{w}$

We will establish a high-probability bound for the elements of  $\mathbf{w} = \mathbf{Z}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\epsilon}$ . Then we define

$$\ell_g := \sqrt{2\sigma_{\mathbf{w}}^2 \log(2n^2)} \quad (4.129)$$

$\sigma_{\mathbf{w}} = \left\| \boldsymbol{\beta}^0 \right\|_2^2 + \sigma^2$ , so we can say that  $|w_i| \leq \ell_g$  for all  $i$  with probability at least  $1 - 1/n$ .

**Lemma 5.** *Let  $\mathbf{w} = \mathbf{Z}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\epsilon}$  for independent and standard normally distributed  $\boldsymbol{\epsilon}$  and  $\mathbf{Z}$  is defined as in 4.1. Then with probability at least  $1 - 1/n$*

$$\max_i |w_i| \leq \sqrt{2\sigma_{\mathbf{w}}^2 \log(2n^2)}. \quad (4.130)$$

where  $\sigma_{\mathbf{w}}^2 = \left\| \boldsymbol{\beta}^0 \right\|_2^2 + \sigma^2$ .

#### Proof of Lemma 5

*Proof.* Having established  $\mathbf{w} = \mathbf{Z}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\epsilon}$  and having assumed the rows of  $\mathbf{Z}$  to be independent and multivariate standard normally distributed, we wish to establish a high-probability bound for the elements of  $\mathbf{w}$ . With  $\boldsymbol{\epsilon}$  having independent and standard normally distributed elements, we recognize that the  $i$ th element of  $\mathbf{w}$  is normally distributed  $w_i \sim N\left(0, \left\| \boldsymbol{\beta}^0 \right\|_2^2 + \sigma^2\right)$ . We can apply a Gaussian tail bound to say

$$\mathbb{P}(|w_i| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\sigma_{\mathbf{w}}^2}\right) \quad (4.131)$$

where  $\sigma_{\mathbf{w}}^2 = \|\beta^0\|_2^2 + \sigma^2$ . Applying a union bound argument and letting  $t = \sqrt{2\sigma_{\mathbf{w}}^2 \log(2n^2)}$ , we can then say

$$\mathbb{P}\left(\max_i |w_i| \geq t\right) \leq 2n \exp\left(-\frac{t^2}{2\sigma_{\mathbf{w}}^2}\right) = \frac{1}{n}. \quad (4.132)$$

□

### Establishing the Bounds of the Quantizers with Gaussian $\mathbf{Z}$

We must establish the bounds of our quantizers under Gaussian assumptions on  $\mathbf{Z}$ . We showed in section 4.2.3 that

$$|z_{ij}| \leq R_g = \sqrt{2 \log(2n^2 d)} \quad (4.133)$$

for all  $i, j$  with probability  $1 - 1/n$ . We also showed in section 4.2.3 that

$$|w_i| \leq \ell_g = \sqrt{2\sigma_{\mathbf{w}}^2 \log(2n^2)} \quad (4.134)$$

for all  $i$  with probability  $1 - 1/n$  where  $\sigma_{\mathbf{w}} = \|\beta^0\|_2^2 + \sigma^2$ . Using these results, we define the bounds of our quantizers as:

	$\alpha^-$	$\alpha^+$	$\Delta$
$Q_{\mathbf{Z}}$	$-R_g$	$R_g$	$2R_g$
$Q_{\mathbf{Z}^2}$	0	$R_g^2$	$R_g^2$
$Q_{\mathbf{w}}$	$-\ell_g$	$\ell_g$	$2\ell_g$

where the bounds of each quantizer are with probability  $1 - 1/n$ .

### Matrix Bernstein Inequality in the Quantized Scenario with Gaussian $\mathbf{Z}$

We apply the Matrix Bernstein Inequality to bound  $\|\tilde{\Sigma} - \Sigma\|$ .

**Lemma 6.** *Given the same assumptions on  $\mathbf{Z}$  and  $\mathbf{w}$  given in 4.1 and 4.2 and the definitions of  $\Sigma$  and  $\tilde{\Sigma}$  in Section 4.2.3. For*

$$\tau = d \log^2(2n^2d) \left(12 + \frac{4}{d}\right) \quad b \geq 2dR_g^2 + \|\Sigma\|, \quad R_g = \sqrt{2 \log(2n^2d)}, \quad (4.135)$$

then with probability at least  $1 - 2/n$

$$\|\tilde{\Sigma} - \Sigma\| \leq \sqrt{\frac{2 \log(2nd)\tau}{n}} + \frac{2 \log(2nd)b}{n}. \quad (4.136)$$

**Proof of Lemma 6**

*Proof.* Let us define the event

$$E : \max_{ij} |z_{ij}| \leq R_g \quad (4.137)$$

where  $R_g = \sqrt{2 \log(2n^2d)}$  By Lemma 4 we know event  $E$  occurs with probability  $1 - 1/n$ .

Conditioning on event  $E$  allows us to state a few important facts that will be helpful throughout this section. Conditioned on  $E$ , we have  $\tilde{z}_{ij}^2 = R_g^2$  for all  $i, j$ . Also, we can say  $\tilde{z}_{ij}^2 \in \{0, R_g^2\}$ . Lastly, we can see that  $\tilde{z}_{ij}^2 - z_{ij}^2 \leq 0$  for all  $i, j$  resulting in  $\Delta_i = \text{diag} \left( \tilde{z}_{ij}^2 - z_{ij}^2 \right)_{j=1}^d$  to consist of all non-positive values. Recall that we defined

$$\tilde{\Sigma}_i := \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T + \Delta_i. \quad (4.138)$$

We now apply the Matrix Bernstein inequality (Theorem 6.17 in Wainwright)[69] to bound the error between our estimator  $\tilde{\Sigma}$  and its true value  $\Sigma$ . We will use the fact that  $\mathbf{Var} \left[ \tilde{\Sigma}_i - \Sigma \right] = \mathbf{Var} \left[ \tilde{\Sigma}_i \right]$ . We apply a Bernstein result to the zero-mean matrices

$\{\tilde{\Sigma}_i - \Sigma\}_{i=1}^n$  conditioned on  $E$ :

$$\mathbb{P} \left( \frac{1}{n} \left\| \sum_{i=1}^n (\tilde{\Sigma}_i - \Sigma) \right\| \geq \delta \mid E \right) \leq 2 \operatorname{rank} \left( \sum_{i=1}^n \mathbf{Var} [\tilde{\Sigma}_i \mid E] \right) \exp \left( \frac{-n\delta^2}{2(\sigma_{\tilde{\Sigma}}^2 + b\delta)} \right) \quad (4.139)$$

for some  $\delta \geq 0$  and where

$$\sigma_{\tilde{\Sigma}}^2 = \frac{1}{n} \left\| \sum_{i=1}^n \mathbf{Var} [\tilde{\Sigma}_i \mid E] \right\| = \frac{1}{n} \left\| n \mathbf{Var} [\tilde{\Sigma}_i \mid E] \right\| = \left\| \mathbf{Var} [\tilde{\Sigma}_i \mid E] \right\|. \quad (4.140)$$

We recognize that the matrices  $\{\tilde{\Sigma}_i - \Sigma\}_{i=1}^n$  conditioned on  $E$  satisfy the Bernstein condition when  $b$  is chosen such that  $b \geq 2dR_g^2 + \|\Sigma\|$  following the same argument as in section 4.2.2.

We proceed by bounding  $\sigma_{\tilde{\Sigma}}^2$ . We first note that  $\mathbf{Var} [\tilde{\Sigma}_i \mid E]$  can be written as

$$\begin{aligned} \mathbf{Var} [\tilde{\Sigma}_i \mid E] &= \mathbf{E} [\tilde{\Sigma}_i^2 \mid E] - \mathbf{E} [\tilde{\Sigma}_i \mid E]^2 = \mathbf{E} [\tilde{\Sigma}_i^2 \mid E] - \Sigma^2 \\ &= \mathbf{E} [(\tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T + \Delta_i)^T (\tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T + \Delta_i) \mid E] - \Sigma^2 \\ &= \mathbf{E} [\tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \mid E] + \mathbf{E} [\Delta_i^2 \mid E] + \mathbf{E} [\tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \Delta_i \mid E] + \mathbf{E} [\Delta_i^T \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \mid E] - \Sigma^2 \\ &= \mathbf{E} [\tilde{\mathbf{z}}_i \|\tilde{\mathbf{z}}_i\|_2^2 \tilde{\mathbf{z}}_i^T \mid E] + \mathbf{E} [\Delta_i^2 \mid E] + \mathbf{E} [\tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \Delta_i \mid E] + \mathbf{E} [\Delta_i \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \mid E] - \Sigma^2 \\ &= dR_g^2 \mathbf{E} [\tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \mid E] + \mathbf{E} [\Delta_i^2 \mid E] + \mathbf{E} [\tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \Delta_i \mid E] + \mathbf{E} [\Delta_i \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \mid E] - \Sigma^2 \quad (4.141) \end{aligned}$$

We then find a bound for  $\sigma_{\tilde{\Sigma}}^2$ :

$$\begin{aligned}
 \sigma_{\tilde{\Sigma}}^2 &= \left\| \mathbf{Var} \left[ \tilde{\Sigma}_i \right] \middle| E \right\| \\
 &= \left\| dR_g^2 \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \middle| E \right] + \mathbf{E} \left[ \Delta_i^2 \middle| E \right] + \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \Delta_i \middle| E \right] + \mathbf{E} \left[ \Delta_i \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \middle| E \right] - \Sigma^2 \right\| \\
 &\leq \left\| dR_g^2 \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \middle| E \right] + \mathbf{E} \left[ \Delta_i^2 \middle| E \right] + 2\mathbf{E} \left[ \Delta_i \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \middle| E \right] \right\| \\
 &\leq dR_g^2 \left\| \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \middle| E \right] \right\| + \left\| \mathbf{E} \left[ \Delta_i^2 \middle| E \right] \right\| + \left\| 2\mathbf{E} \left[ \Delta_i \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \middle| E \right] \right\| \\
 &\leq dR_g^4 + R_g^4 + 2\mathbf{E} \left[ \left\| \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right\| \left\| \Delta_i \right\| \middle| E \right] \\
 &\leq dR_g^4 + R_g^4 + 2dR_g^4 \\
 &= (3d + 1)R_g^4. \tag{4.142}
 \end{aligned}$$

Here we used Jensen's inequality and the convexity of the spectral norm. Additionally, we used the facts:

1. Since  $\mathbf{Var} \left[ \tilde{\Sigma}_i \right] \geq 0$  and  $\Sigma \geq 0$  we applied (D.1) to drop  $\Sigma$  from the second step.
2. We recognize that  $z_{ij}$  conditioned on  $E$  is distributed according to a truncated normal distribution, giving us

$$\mathbf{E} \left[ z_{ij} \middle| E \right] = \mu_{z_k} + \frac{\varphi(\alpha) - \varphi(\beta)}{\Phi(\beta) - \Phi(\alpha)} \sigma_{z_j} \tag{4.143}$$

$$\mathbf{Var} \left[ z_{ij} \middle| E \right] = \sigma_{z_j}^2 \left[ 1 - \frac{\beta\varphi(\beta) - \alpha\varphi(\alpha)}{\Phi(\beta) - \Phi(\alpha)} - \left( \frac{\varphi(\alpha) - \varphi(\beta)}{\Phi(\beta) - \Phi(\alpha)} \right)^2 \right] \tag{4.144}$$

where

$$\alpha = \frac{-r - \mu_{z_j}}{\sigma_{z_j}} \quad \beta = \frac{r - \mu_{z_j}}{\sigma_{z_j}} \quad (4.145)$$

$$\mu_{z_j} \text{ is the mean of } z_{ij} \quad \sigma_{z_j}^2 \text{ is the variance of } z_{ij} \quad (4.146)$$

$$\varphi(\cdot) \text{ is the probability density function of } \cdot \quad (4.147)$$

$$\Phi(\cdot) \text{ is the cumulative distribution function of } \cdot \cdot \quad (4.148)$$

In the scenario where the left and right bounds are equal in absolute value, then we have

$$\mathbf{E} [z_{ij} | E] = \mu_{z_j} \quad \forall i, j \quad (4.149)$$

$$\mathbf{E} [z_{ij}^2 | E] = \mathbf{Var} [z_{ij} | E] + \mathbf{E} [z_{ij} | E]^2 = \sigma_{\mathbf{Z}}^2 \left[ 1 - \frac{2r\varphi(r)}{\Phi(r) - \Phi(-r)} \right] + \mu_{z_j}^2 \quad (4.150)$$

Let us briefly consider the scenario where  $\mathbf{Z}$  is distributed according to a multivariate normal distribution with mean vector  $\boldsymbol{\mu} = (\mu_{z_1}, \mu_{z_2}, \dots, \mu_{z_d})^T$  and arbitrary covariance

matrix  $\Sigma$ . Then calculating  $\mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \mid E \right]$ :

$$\begin{aligned}
 \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \mid E \right] &= \mathbf{E} \left[ \begin{array}{c} \left[ \begin{array}{cccc} \tilde{z}_{i1}^2 & \tilde{z}_{i1}\tilde{z}_{i2} & \dots & \tilde{z}_{i1}\tilde{z}_{id} \\ \tilde{z}_{i2}\tilde{z}_{i1} & \tilde{z}_{i2}^2 & \dots & \tilde{z}_{i2}\tilde{z}_{id} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{z}_{id}\tilde{z}_{i1} & \tilde{z}_{id}\tilde{z}_{i2} & \dots & \tilde{z}_{id}^2 \end{array} \right] \mid E \end{array} \right] \\
 &= \text{diag} \left( \mathbf{E} \left[ \tilde{z}_{ij}^2 \mid E \right] - \mathbf{E} \left[ \tilde{z}_{ij} \mid E \right]^2 \right)_{j=1}^d + \mathbf{E} \left[ \begin{array}{c} \left[ \begin{array}{cccc} \tilde{z}_{i1} & \tilde{z}_{i1}\tilde{z}_{i2} & \dots & \tilde{z}_{i1}\tilde{z}_{id} \\ \tilde{z}_{i2}\tilde{z}_{i1} & \tilde{z}_{i2} & \dots & \tilde{z}_{i2}\tilde{z}_{id} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{z}_{id}\tilde{z}_{i1} & \tilde{z}_{id}\tilde{z}_{i2} & \dots & \tilde{z}_{id} \end{array} \right] \mid E \end{array} \right] \\
 &= \text{diag} \left( R_g^2 - \left( \sigma_{z_j}^2 (1 - C) + \mu_{z_j}^2 \right) \right)_{j=1}^d \\
 &+ \begin{array}{c} \left[ \begin{array}{cccc} \sigma_{z_1}^2 (1 - C) + \mu_{z_1}^2 & \mathbf{E}_{\mathbf{Z}} [z_{i1}z_{i2} \mid E] & \dots & \mathbf{E}_{\mathbf{Z}} [z_{i1}z_{id} \mid E] \\ \mathbf{E}_{\mathbf{Z}} [z_{i2}z_{i1} \mid E] & \sigma_{z_2}^2 (1 - C) + \mu_{z_2}^2 & \dots & \mathbf{E}_{\mathbf{Z}} [z_{i2}z_{id} \mid E] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{E}_{\mathbf{Z}} [z_{id}z_{i1} \mid E] & \mathbf{E}_{\mathbf{Z}} [z_{id}z_{i2} \mid E] & \dots & \sigma_{z_d}^2 (1 - C) + \mu_{z_d}^2 \end{array} \right] \end{array} \\
 &= \text{diag} \left( R_g^2 \right)_{j=1}^d + \begin{array}{c} \left[ \begin{array}{cccc} 0 & \mathbf{E}_{\mathbf{Z}} [z_{i1}z_{i2} \mid E] & \dots & \mathbf{E}_{\mathbf{Z}} [z_{i1}z_{id} \mid E] \\ \mathbf{E}_{\mathbf{Z}} [z_{i2}z_{i1} \mid E] & 0 & \dots & \mathbf{E}_{\mathbf{Z}} [z_{i2}z_{id} \mid E] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{E}_{\mathbf{Z}} [z_{id}z_{i1} \mid E] & \mathbf{E}_{\mathbf{Z}} [z_{id}z_{i2} \mid E] & \dots & 0 \end{array} \right] \end{array} \\
 &= \text{diag} \left( R_g^2 \right)_{j=1}^d + \mathbf{D} \tag{4.151}
 \end{aligned}$$

where we let  $\mathbf{D}$  be the off-diagonal matrix. We recognize that, conditioned on  $E$ ,  $z_{ij}z_{ij'} \leq R_g^2$ . Thus, using the Gershgorin circle theorem we can bound  $\|\mathbf{D}\| \leq dR_g^2$ .

Then we can bound the operator norm of  $\mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \mid E \right]$  by

$$\begin{aligned} \left\| \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \mid E \right] \right\| &= \left\| \text{diag} \left( R_g^2 \right)_{j=1}^d + \mathbf{D} \right\| \\ &\leq \left\| \text{diag} \left( R_g^2 \right)_{j=1}^d \right\| + \|\mathbf{D}\| \\ &\leq r_g^2 + dR_g^2 \end{aligned} \tag{4.152}$$

In the case when  $\mathbf{Z}$  has a zero mean vector and covariance matrix  $\Sigma = \mathbf{I}_{d \times d}$ , then  $\mathbf{D}$  becomes a zero matrix and the bound becomes

$$\left\| \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \mid E \right] \right\| \leq R_g^2 \tag{4.153}$$

3. Recognizing that all the elements in  $\tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T$  are bounded in absolute value by  $R_g^2$ , then a bound of  $\left\| \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right\|$  conditioned on  $E$  is given by

$$\begin{aligned} \left\| \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right\|^2 &= \sup_{\|\mathbf{v}\|=1} \mathbf{v}^T \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \mathbf{v} = \sup_{\|\mathbf{v}\|=1} \left| \sum_{j,j'} v_j v_{j'} \left( \sum_{\ell} \left( \tilde{\mathbf{z}}_i^T \tilde{\mathbf{z}}_i \right)_{\ell j} \left( \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right)_{\ell j'} \right) \right| \\ &\leq \sup_{\|\mathbf{v}\|=1} \sum_{j,j'} |v_j| |v_{j'}| \left| \sum_{\ell} \left( \tilde{\mathbf{z}}_i^T \tilde{\mathbf{z}}_i \right)_{\ell j} \left( \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right)_{\ell j'} \right| \\ &\leq dR_g^4 \sup_{\|\mathbf{v}\|=1} \|\mathbf{v}\|_1^2 = d^2 R_g^4 \\ &\implies \left\| \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T \right\| \leq dR_g^2 \end{aligned} \tag{4.154}$$

4. We use  $\tilde{z}_{ij}^2 = R_g^2$  for all  $i, j$  and  $\tilde{z}_{ij}^2 \in \{0, R_g^2\}$  when conditioned on  $E$  to say

$$\mathbf{E} [\|\Delta_i\| \mid E] = \mathbf{E} \left[ \left\| \text{diag} \left( \tilde{z}_{ij}^2 - z_{ij}^2 \right)_{j=1}^d \right\| \mid E \right] = \mathbf{E} \left[ \left\| \text{diag} \left( R_g^2 - z_{ij}^2 \right)_{j=1}^d \right\| \mid E \right] \leq R_g^2. \quad (4.155)$$

We proceed by defining  $\tau$  to be the bound established for  $\sigma_{\tilde{\Sigma}}^2$ :

$$\tau := (3d + 1)R_g^4 = d \left( \sqrt{2 \log(2n^2d)} \right)^4 \left( 3 + \frac{1}{d} \right) = d \log^2(2n^2d) \left( 12 + \frac{4}{d} \right). \quad (4.156)$$

Since  $\sigma_{\tilde{\Sigma}}^2 = \left\| \mathbf{Var} \left[ \tilde{\Sigma}_i \mid E \right] \right\| \leq \tau$  by (4.142) we can say

$$\mathbb{P} \left( \frac{1}{n} \left\| \sum_{i=1}^n \left( \tilde{\Sigma}_i - \Sigma \right) \right\| \geq \delta \mid E \right) \leq 2d \exp \left( \frac{-n\delta^2}{2 \left( \sigma_{\tilde{\Sigma}}^2 + b\delta \right)} \right) \leq 2d \exp \left( \frac{-n\delta^2}{2 \left( \tau + b\delta \right)} \right). \quad (4.157)$$

Let us choose

$$\delta = \sqrt{\frac{2 \log(2nd)\tau}{n} + \frac{2 \log(2nd)b}{n}} \quad (4.158)$$

so that, following the same logic as in section 4.2.2

$$\frac{n\delta^2}{2(\tau + b\delta)} \geq \min \left\{ \frac{n\delta^2}{2\tau}, \frac{n\delta}{2b} \right\} \geq \log(2nd). \quad (4.159)$$

Then we have

$$\mathbb{P} \left( \frac{1}{n} \left\| \sum_{i=1}^n \left( \tilde{\Sigma}_i - \Sigma \right) \right\| \geq \delta \mid E \right) \leq 2d \exp \left( \frac{-n\delta^2}{2(\tau + b\delta)} \right) \leq 2d \exp \left( -\log(2nd) \right) = \frac{1}{n} \quad (4.160)$$

for  $\delta = \sqrt{\frac{2 \log(2nd)\tau}{n} + \frac{2 \log(2nd)b}{n}}$ , for  $\tau = d \log^2(2n^2d) \left( 12 + \frac{4}{d} \right)$ , and for any  $b \geq 2dR_g^2 + \|\Sigma\|$ .

Applying the law of total probability, we conclude

$$\mathbb{P}\left(\frac{1}{n}\left\|\sum_{i=1}^n(\tilde{\Sigma}_i - \Sigma)\right\|\geq\delta\right)\leq\mathbb{P}\left(\frac{1}{n}\left\|\sum_{i=1}^n(\tilde{\Sigma}_i - \Sigma)\right\|\geq\delta\mid E\right)+\mathbb{P}(|z_{ij}|\geq R_g)=\frac{2}{n}. \quad (4.161)$$

□

### Summary of Quantized Scenario with Gaussian $\mathbf{Z}$

For a random  $\mathbf{Z}$  whose rows  $\mathbf{z}_i^T$  are drawn independently from a multivariate normal distribution with mean  $\mathbf{0}$  and covariance  $\mathbf{I}_{d\times d}$  and definitions

$$\Sigma = \mathbf{E}\left[\frac{\mathbf{Z}^T\mathbf{Z}}{n}\right] = \mathbf{I}_d \quad \tilde{\Sigma} = \frac{1}{n}\sum_{i=1}^n\tilde{\Sigma}_i = \frac{1}{n}\sum_{i=1}^n\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^T + \Delta_i \quad (4.162)$$

$$\Sigma_{\mathbf{Z}\mathbf{w}} = \mathbf{E}\left[\frac{\mathbf{Z}^T\mathbf{w}}{n}\right] \quad \tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} = \frac{\tilde{\mathbf{Z}}^T\tilde{\mathbf{w}}}{n} = \frac{1}{n}\sum_{i=1}^n\tilde{z}_{ij}\tilde{w}_i \quad \text{for } 1 \leq j \leq d, \quad (4.163)$$

we showed that

$$\left\|\tilde{\Sigma} - \Sigma\right\| \leq \sqrt{\frac{2\log(2nd)\tau}{n}} + \frac{2\log(2nd)b}{n} \quad \text{w.p. } 1 - 2/n \quad (4.164)$$

$$\left\|\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}}\right\|_2 \leq \sqrt{\frac{8dR_g^2\ell_g^2\log(2nd)}{n}} \quad \text{w.p. } 1 - 3/n \quad (4.165)$$

where

$$\tau = d\log^2(2n^2d)\left(12 + \frac{4}{d}\right) \quad b \geq 2dR_g^2 + \|\Sigma\|$$

$$\ell_g = \sqrt{2\left(\|\beta^0\|_2^2 + \sigma^2\right)\log(2n^2)} \quad R_g = \sqrt{2\log(2n^2d)}.$$

We can then say the order of the two terms are

$$\left\| \tilde{\Sigma} - \Sigma \right\| = \mathcal{O} \left( \sqrt{\frac{d \log^3(nd)}{n}} \right) = \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{n}} \right) \quad (4.166)$$

$$\left\| \tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}} \right\|_2 = \mathcal{O} \left( \sqrt{\frac{d \log(n) \log^2(nd)}{n}} \right) = \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{n}} \right) \quad (4.167)$$

where we again use the notation  $\tilde{\mathcal{O}}(\cdot)$  to denote the order of the terms with constants and log terms removed. Then ensuring that  $n$  is selected such that

$$n \geq \max \left\{ 2^3 d \log^3(2d) \eta^2 \left( 12 + \frac{4}{d} \right), 2^2 \log(2d) b \eta \right\} \quad (4.168)$$

for  $\eta = \frac{2}{\lambda_{\min}(\Sigma)}$  satisfies the lambda-min-requirement and allows us to say

$$\begin{aligned} \left\| \hat{\beta} - \beta^* \right\|_2 &\leq \frac{4 \|\beta^*\|_2 \left\| \hat{\Sigma} - \Sigma \right\| + 4 \left\| (\hat{\Sigma}_{\mathbf{X}\mathbf{y}} - \Sigma_{\mathbf{X}\mathbf{y}}) \right\|_2}{\lambda_{\min}(\Sigma)} \\ &\leq \frac{4 \|\beta^*\|_2 \left( \sqrt{\frac{2d \log(2nd) \tau_0}{n}} + \frac{2d \log(2nd) b_0}{n} \right) + 4 \sqrt{\frac{8d R_g^2 \ell_g^2 \log(2nd)}{n}}}{\lambda_{\min}(\Sigma)} = \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{n}} \right) \end{aligned} \quad (4.169)$$

with probability at least  $1 - 5/n$  where  $\tau_0 = \log^2(2n^2d) \left( 12 + \frac{4}{d} \right)$  and  $b_0 = 2R_g^2 + d^{-1} \|\Sigma\|$ .

### 4.3 Sketched Regression Parameters

This section examines the effect of transforming the scaled design matrix  $\mathbf{X} = \frac{\mathbf{Z}}{\sqrt{n}}$  and scaled response vector  $\mathbf{y} = \frac{\mathbf{w}}{\sqrt{n}}$  by left multiplying by a sketching matrix  $\mathbf{S}$  consisting of independent standard gaussian rows.

### 4.3.1 Introduction

Throughout this section we will perform our analysis on a scaled design matrix and response vector:

$$\mathbf{X} := \frac{\mathbf{Z}}{\sqrt{n}} \quad \text{and} \quad \mathbf{y} := \frac{\mathbf{w}}{\sqrt{n}}, \quad (4.170)$$

where  $\mathbf{Z}$  and  $\mathbf{w}$  have the linear relationship  $\mathbf{w} = \mathbf{Z}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\epsilon}$ . We transform our data using this sketching matrix and define

$$\hat{\mathbf{X}} := \mathbf{S}\mathbf{X}, \quad \hat{\mathbf{y}} := \mathbf{S}\mathbf{y}. \quad (4.171)$$

In the sections where  $\mathbf{Z}$  is assumed to be fixed, let us define

$$\boldsymbol{\Sigma} = \mathbf{X}^T \mathbf{X} = \frac{\mathbf{Z}^T \mathbf{Z}}{n} \quad (4.172)$$

$$\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}} = \mathbf{X}^T \mathbf{y} = \frac{\mathbf{Z}^T \mathbf{w}}{n}. \quad (4.173)$$

When we assume  $\mathbf{Z}$  consists of random variables, we will define

$$\boldsymbol{\Sigma} := \mathbf{E} \left[ \mathbf{X}^T \mathbf{X} \right] = \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{Z}}{n} \right] = \mathbf{I}_{d \times d} \quad (4.174)$$

$$\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}} := \mathbf{E} \left[ \mathbf{X}^T \mathbf{y} \right] = \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{w}}{n} \right]. \quad (4.175)$$

Section 4.3.2 will assume  $\mathbf{Z}$  to be fixed and section 4.3.3 will assume the elements of  $\mathbf{Z}$  to be independent and standard normally distributed.

### 4.3.2 Sketched Scenario with Fixed $\mathbf{Z}$

This section will assume that the  $n \times d$  matrix  $\mathbf{Z}$  is fixed and that the elements of  $\mathbf{Z}$ ,  $z_{ij}$  are bounded in the range  $[-R, R]$ . That is  $|z_{ij}| \leq R$  for all  $i = 1, \dots, n$  and  $j = 1, \dots, d$ .

We then define our estimators of  $\Sigma$  and  $\Sigma_{\mathbf{X}\mathbf{y}}$ , as defined in (4.172) and (4.173), as

$$\hat{\Sigma} := \frac{1}{m} \mathbf{X}^T \mathbf{S}^T \mathbf{S} \mathbf{X} = \frac{1}{m} \hat{\mathbf{X}}^T \hat{\mathbf{X}} \quad (4.176)$$

$$\hat{\Sigma}_{\mathbf{X}\mathbf{y}} := \frac{1}{m} \mathbf{X}^T \mathbf{S}^T \mathbf{S} \mathbf{y} = \frac{1}{m} \hat{\mathbf{X}}^T \hat{\mathbf{y}}. \quad (4.177)$$

We will prove

**Theorem 4.3.1.** *Let  $\mathbf{Z}$  be a fixed  $n \times d$  matrix whose elements  $|z_{ij}| \leq R$  for all  $i = 1, \dots, n$  and  $j = 1, \dots, d$ . Let  $\mathbf{w} = \mathbf{Z}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\epsilon}$  where  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I})$ . Let  $\mathbf{X} := \frac{\mathbf{Z}}{\sqrt{n}}$  and  $\mathbf{y} := \frac{\mathbf{w}}{\sqrt{n}}$ . Let  $\mathbf{S}$  be a  $m \times n$  matrix, for some  $m < n$ , whose elements  $s_{ij}$  are drawn independently from a standard normal distribution. Let  $\Sigma$ ,  $\Sigma_{\mathbf{X}\mathbf{y}}$ ,  $\hat{\Sigma}$ , and  $\hat{\Sigma}_{\mathbf{X}\mathbf{y}}$  be defined as in (4.172), (4.173), (4.174), and (4.175), respectively. Then for*

$$\eta = \frac{2}{\lambda_{\min}(\Sigma)}, \quad \epsilon_0 = \left(1 + \sqrt{\frac{2 \log(2n)}{d}}\right),$$

and  $m$  chosen such that

$$m \geq 2d\eta \|\Sigma\| \left(1 + \frac{2\sqrt{2 \log(2n)}}{d} + \frac{2 \log(2n)}{d}\right) \max\{2^3 \eta \|\Sigma\|, 1\}$$

then with probability at least  $1 - 3/n$  and some  $C > 0$

$$\begin{aligned} \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 &= \frac{4 \left(2\sqrt{\frac{d}{m}}\epsilon_0 + \frac{d}{m}\epsilon_0^2\right) \left(\|\boldsymbol{\beta}^*\|_2 \|\Sigma\| + \left(RC\sqrt{\log(n)}\|\boldsymbol{\beta}^0\|_2 + \sigma \left(1 + \sqrt{\frac{2 \log(n)}{n}}\right)\right) \|\mathbf{X}\| \right)}{\lambda_{\min}(\Sigma)} \\ &= \tilde{O}\left(\sqrt{\frac{d}{m}}\right). \end{aligned}$$

Adding in the additional error from  $\|\hat{\beta} - \beta^*\|$  from the inequality

$$\|\beta^0 - \hat{\beta}\| \leq \|\beta^0 - \beta^*\| + \|\hat{\beta} - \beta^*\|. \quad (4.178)$$

Then with probability  $1 - 3/n$

$$\begin{aligned} & \|\beta^0 - \hat{\beta}\|_2 \\ & \leq \frac{4 \left( 2\sqrt{\frac{d}{m}}\epsilon_0 + \frac{d}{m}\epsilon_0^2 \right) \left( \|\beta^*\|_2 \|\Sigma\| + \left( RC\sqrt{\log(n)}\|\beta^0\|_2 + \sigma \left( 1 + \sqrt{\frac{2\log(n)}{n}} \right) \right) \|\mathbf{X}\| \right)}{\lambda_{\min}(\Sigma)} + \sqrt{\frac{\sigma^2 d}{\lambda_{\min}(\Sigma) n}} \\ & = \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{m}}\right) + \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right). \end{aligned}$$

### Proof of Theorem 4.3.1

*Proof.* We wish to bound

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{4\|\beta^*\|_2 \|\tilde{\Sigma} - \Sigma\| + 4\|(\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}})\|_2}{\lambda_{\min}(\Sigma)}$$

by controlling

$$(a): \|\tilde{\Sigma} - \Sigma\| \quad (b): \|\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}}\|_2.$$

**Analyzing Part (a) in the Sketched Scenario with Fixed  $\mathbf{Z}$**  Let us examine part

(a), that is, let us find a high-probability bound for  $\|\hat{\Sigma} - \Sigma\|$ . Recognizing that  $\hat{\mathbf{X}} = \mathbf{S}\mathbf{X}$

is a Gaussian matrix with covariance matrix  $\Sigma = \mathbf{X}^T\mathbf{X}$  and having defined  $\hat{\Sigma} = \frac{1}{m}\hat{\mathbf{X}}^T\hat{\mathbf{X}}$ ,

we can directly apply Theorem 6.1 and example 6.3 in Wainwright [69] to get

$$\mathbb{P} \left( \frac{\|\hat{\Sigma} - \Sigma\|}{\|\Sigma\|} \geq 2\sqrt{\frac{d}{m}} + 2\delta + \left( \sqrt{\frac{d}{m}} + \delta \right)^2 \right) \leq 2e^{-m\delta^2/2} \quad (4.179)$$

for all  $\delta > 0$ . Letting  $\delta = \sqrt{\frac{2\log(2n)}{m}}$ , then

$$\mathbb{P} \left( \frac{\|\hat{\Sigma} - \Sigma\|}{\|\Sigma\|} \geq 2\sqrt{\frac{d}{m}} + 2\delta + \left( \sqrt{\frac{d}{m}} + \delta \right)^2 \right) \leq \frac{1}{n} \quad (4.180)$$

Then we can say with probability at least  $1 - 1/n$  that

$$\|\hat{\Sigma} - \Sigma\| \leq \|\Sigma\| (2\epsilon + \epsilon^2) \quad (4.181)$$

where we have let  $\epsilon = \sqrt{\frac{d}{m}} + \sqrt{\frac{2\log(2n)}{m}}$ .

To satisfy the lambda min requirement (4.17), we require

$$(2\epsilon + \epsilon^2) \leq 2\sqrt{\frac{d}{m}} + 2\sqrt{\frac{2\log(2n)}{m}} + \left( \sqrt{\frac{d}{m}} + \sqrt{\frac{2\log(2n)}{m}} \right)^2 \leq \frac{\lambda_{\min}(\Sigma)}{2\|\Sigma\|}. \quad (4.182)$$

It is sufficient to find an  $m$  such that the inequalities hold:

$$2\sqrt{\frac{d}{m}} + 2\sqrt{\frac{2\log(2n)}{m}} \leq \frac{\lambda_{\min}(\Sigma)}{4\|\Sigma\|} \quad \text{and} \quad \left( \sqrt{\frac{d}{m}} + \sqrt{\frac{2\log(2n)}{m}} \right)^2 \leq \frac{\lambda_{\min}(\Sigma)}{4\|\Sigma\|}. \quad (4.183)$$

We examine each inequality separately beginning with the first

$$\begin{aligned}
 2\sqrt{\frac{d}{m}} + 2\sqrt{\frac{2\log(2n)}{m}} &\leq \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{4\|\boldsymbol{\Sigma}\|} \\
 \implies \frac{(\sqrt{d} + \sqrt{2\log(2n)})^2}{m} &\leq \frac{\lambda_{\min}^2(\boldsymbol{\Sigma})}{2^6\|\boldsymbol{\Sigma}\|^2} \\
 \implies m &\geq \frac{2^6\|\boldsymbol{\Sigma}\|^2(\sqrt{d} + \sqrt{2\log(2n)})^2}{\lambda_{\min}^2(\boldsymbol{\Sigma})} \\
 \implies m &\geq \frac{2^6\|\boldsymbol{\Sigma}\|^2(d + 2\sqrt{2d\log(2n)} + 2\log(2n))}{\lambda_{\min}^2(\boldsymbol{\Sigma})} \\
 \implies m &\geq 2^4d\eta^2\|\boldsymbol{\Sigma}\|^2\left(1 + \frac{2\sqrt{2\log(2n)}}{d} + \frac{2\log(2n)}{d}\right) \tag{4.184}
 \end{aligned}$$

where

$$\eta = \frac{2}{\lambda_{\min}(\boldsymbol{\Sigma})}. \tag{4.185}$$

Examining the second term:

$$\begin{aligned}
 \left(\sqrt{\frac{d}{m}} + \sqrt{\frac{2\log(2n)}{m}}\right)^2 &\leq \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{4\|\boldsymbol{\Sigma}\|} \\
 \implies \frac{(\sqrt{d} + \sqrt{2\log(2n)})^2}{m} &\leq \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{4\|\boldsymbol{\Sigma}\|} \\
 \implies m &\geq \frac{4\|\boldsymbol{\Sigma}\|}{\lambda_{\min}(\boldsymbol{\Sigma})}(d + 2\sqrt{2d\log(2n)} + 2\log(2n)) \\
 \implies m &\geq 2d\eta\|\boldsymbol{\Sigma}\|\left(1 + \frac{2\sqrt{2\log(2n)}}{\sqrt{d}} + \frac{2\log(2n)}{d}\right). \tag{4.186}
 \end{aligned}$$

Selecting  $m$  such that

$$m \geq 2d\eta \|\Sigma\| \left( 1 + \frac{2\sqrt{2\log(2n)}}{d} + \frac{2\log(2n)}{d} \right) \max \left\{ 2^3\eta \|\Sigma\|, 1 \right\} \quad (4.187)$$

ensures the requirement is met.

**Analyzing Part (b) in the Sketched Scenario with Fixed  $\mathbf{Z}$**  We wish to now upper bound

$$\left\| \hat{\Sigma}_{\mathbf{X}\mathbf{y}} - \Sigma_{\mathbf{X}\mathbf{y}} \right\|_2.$$

We note that the normed difference can be bounded by Cauchy Schwarz

$$\begin{aligned} \left\| \hat{\Sigma}_{\mathbf{X}\mathbf{y}} - \Sigma_{\mathbf{X}\mathbf{y}} \right\|_2 &= \left\| \frac{1}{m} \mathbf{X}^T \mathbf{S}^T \mathbf{S} \mathbf{y} - \mathbf{X}^T \mathbf{y} \right\|_2 = \left\| \mathbf{X}^T \left( \frac{1}{m} \mathbf{S}^T \mathbf{S} - \mathbf{I} \right) \mathbf{y} \right\|_2 \\ &\leq \|\mathbf{X}\| \left\| \frac{1}{m} \mathbf{S}^T \mathbf{S} - \mathbf{I} \right\| \|\mathbf{y}\|_2. \end{aligned} \quad (4.188)$$

Then we apply Theorem 6.1 and Example 6.2 from Wainwright [69] to the middle term to get

$$\mathbb{P} \left( \left\| \frac{1}{m} \mathbf{S}^T \mathbf{S} - \mathbf{I} \right\| \geq 2\epsilon + \epsilon^2 \right) \leq 2e^{-m\delta^2/2}$$

where  $\epsilon = \sqrt{\frac{d}{m}} + \delta$  for all  $\delta > 0$  and assuming  $m \geq d$ . Letting  $\delta = \sqrt{\frac{2\log(2n)}{m}}$ , then

$$\left\| \frac{1}{m} \mathbf{S}^T \mathbf{S} - \mathbf{I} \right\| \leq 2\epsilon + \epsilon^2 \quad (4.189)$$

with probability at least  $1 - 1/n$ .

Thus, we can say

$$\left\| \hat{\Sigma}_{\mathbf{X}\mathbf{y}} - \Sigma_{\mathbf{X}\mathbf{y}} \right\|_2 \leq \|\mathbf{y}\|_2 \left( 2\epsilon + \epsilon^2 \right) \|\mathbf{X}\| \quad (4.190)$$

with probability at least  $1 - 1/n$  for  $\epsilon = \sqrt{\frac{d}{m}} + \sqrt{\frac{2\log(2n)}{m}}$ . Combining this with the result from section 4.3.2 Lemma 7 which provided a bound on  $\|\mathbf{y}\|_2$ , and using union bounds, we conclude

$$\left\| \hat{\Sigma}_{\mathbf{X}\mathbf{y}} - \Sigma_{\mathbf{X}\mathbf{y}} \right\|_2 \leq \left( RC\sqrt{\log(n)}\|\beta^*\|_2 + \sigma \left( 1 + \sqrt{\frac{2\log(n)}{n}} \right) \right) (2\epsilon + \epsilon^2) \|\mathbf{X}\| \quad (4.191)$$

with probability  $1 - (1/n + 1/n)$ .

Combining (4.181) and (4.191) yields with probability at least  $1 - (1/n + 1/n + 1/n)$

$$\begin{aligned} \|\beta^* - \hat{\beta}\|_2 &\leq \frac{4(2\epsilon + \epsilon^2) \left( \|\beta^*\|_2 \|\Sigma\| + \left( RC\sqrt{\log(n)}\|\beta^*\|_2 + \sigma \left( 1 + \sqrt{\frac{2\log(n)}{n}} \right) \right) \|\mathbf{X}\| \right)}{\lambda_{\min}(\Sigma)} \\ &= \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{m}} \right). \end{aligned}$$

We now add on the additional error from  $\|\hat{\beta} - \beta^*\|$  from the inequality

$$\|\beta^0 - \hat{\beta}\| \leq \|\beta^0 - \beta^*\| + \|\hat{\beta} - \beta^*\|. \quad (4.192)$$

Then with probability  $1 - 3/n$

$$\|\beta^0 - \hat{\beta}\|_2 \quad (4.193)$$

$$\leq \frac{4(2\epsilon + \epsilon^2) \left( \|\beta^*\|_2 \|\Sigma\| + \left( RC\sqrt{\log(n)}\|\beta^*\|_2 + \sigma \left( 1 + \sqrt{\frac{2\log(n)}{n}} \right) \right) \|\mathbf{X}\| \right)}{\lambda_{\min}(\Sigma)} + \sqrt{\frac{\sigma^2 d}{\lambda_{\min}(\Sigma) n}} \quad (4.194)$$

$$= \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{m}} \right) + \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{n}} \right). \quad (4.195)$$

The subsequent sections will provide the lemmas for this proof.  $\square$

### Establishing a High-Probability Bound on the Norm of $\mathbf{y}$ with Fixed $\mathbf{Z}$

**Lemma 7.** *Let  $\mathbf{Z}$  be a fixed  $n \times d$  matrix whose elements  $|z_{ij}| \leq R$  for all  $i = 1, \dots, n$  and  $j = 1, \dots, d$ . Let  $\mathbf{w} = \mathbf{Z}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\epsilon}$  where  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I})$ . Let  $\mathbf{X} := \frac{\mathbf{Z}}{\sqrt{n}}$  and  $\mathbf{y} := \frac{\mathbf{w}}{\sqrt{n}}$ . Then for some  $C > 0$  and with probability at least  $1 - 1/n$*

$$\|\mathbf{y}\|_2 \leq \|\mathbf{X}\boldsymbol{\beta}^0\|_2 + \sigma \frac{\|\boldsymbol{\epsilon}\|_2}{\sqrt{n}} \leq RC\sqrt{\log(n)}\|\boldsymbol{\beta}^0\|_2 + \sigma \left(1 + \sqrt{\frac{2\log(n)}{n}}\right).$$

#### Proof of Lemma 7

*Proof.* We will bound  $\|\mathbf{y}\|_2$ . We analyze  $\|\mathbf{y}\|_2$  in two steps: 1)  $\|\mathbf{X}\boldsymbol{\beta}^0\|_2$ , and 2)  $\|\boldsymbol{\epsilon}/\sqrt{n}\|_2$ .

We assumed in section 4.1.4 that

$$|\mathbf{z}_i^T \boldsymbol{\beta}^0| \leq RC\sqrt{\log(n)}\|\boldsymbol{\beta}^0\|_2, \quad (4.196)$$

for all  $i$  and some  $C > 0$ . This allows us to say

$$\|\mathbf{Z}\boldsymbol{\beta}^0\|_2 \leq RC\sqrt{n\log(n)}\|\boldsymbol{\beta}^0\|_2, \quad (4.197)$$

which implies

$$\|\mathbf{X}\boldsymbol{\beta}^0\|_2 = \frac{\|\mathbf{Z}\boldsymbol{\beta}^0\|_2}{\sqrt{n}} \leq RC\sqrt{\log(n)}\|\boldsymbol{\beta}^0\|_2. \quad (4.198)$$

We now examine  $\boldsymbol{\epsilon}/\sqrt{n}$ . We follow Example 2.28 in Wainwright [69]. We first recognize that  $\|\boldsymbol{\epsilon}\|_2^2 = \sum_{i=1}^n \epsilon_i^2$  follows a  $\chi^2$ -distribution with  $n$  degrees of freedom. We define

$$\mathbf{V} = \frac{\|\boldsymbol{\epsilon}\|_2}{\sqrt{n}}. \quad (4.199)$$

Using that the Euclidean norm is a 1-Lipschitz function, Theorem 2.26 in Wainwright[69] implies that

$$\mathbb{P}(\mathbf{V} \geq \mathbf{E}[\mathbf{V}] + \delta) \leq e^{-n\delta^2/2} \quad \text{for all } \delta \geq 0. \quad (4.200)$$

Using concavity of the square root function and Jensen's inequality

$$\mathbf{E}[\mathbf{V}] \leq \sqrt{\mathbf{E}[\mathbf{V}^2]} = \left[ \frac{1}{n} \sum_{i=1}^n \mathbf{E}[\epsilon_i^2] \right]^{1/2} = 1. \quad (4.201)$$

Using that  $\mathbf{V} = \sqrt{\|\boldsymbol{\epsilon}\|_2^2}/\sqrt{n}$  and by combining these pieces

$$\mathbb{P}\left(\frac{\|\boldsymbol{\epsilon}\|_2}{\sqrt{n}} \geq 1 + \delta\right) \leq e^{-n\delta^2/2} \quad \text{for all } \delta \geq 0. \quad (4.202)$$

Letting  $\delta = \sqrt{\frac{2\log(n)}{n}}$ , then

$$\frac{\|\boldsymbol{\epsilon}\|_2}{\sqrt{n}} \leq 1 + \delta \quad (4.203)$$

with probability at least  $1 - 1/n$ . Using the triangle inequality, we can then say

$$\|\mathbf{y}\|_2 \leq \|\mathbf{X}\boldsymbol{\beta}^0\|_2 + \sigma \frac{\|\boldsymbol{\epsilon}\|_2}{\sqrt{n}} \leq RC\sqrt{\log(n)}\|\boldsymbol{\beta}^0\|_2 + \sigma \left(1 + \sqrt{\frac{2\log(n)}{n}}\right) \quad (4.204)$$

with probability at least  $1 - 1/n$ . □

### Summary of the Sketched Scenario with Fixed $\mathbf{Z}$

For fixed  $\mathbf{Z}$  with bounded elements  $|z_{ij}| \leq R$  for all  $i, j$  and definitions

$$\begin{aligned}\boldsymbol{\Sigma} &= \frac{\mathbf{Z}^T \mathbf{Z}}{n} & \hat{\boldsymbol{\Sigma}} &= \frac{1}{m} \hat{\mathbf{X}}^T \hat{\mathbf{X}} = \frac{1}{m} \mathbf{X}^T \mathbf{S}^T \mathbf{S} \mathbf{X} \\ \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}} &= \frac{\mathbf{Z}^T \mathbf{w}}{n} & \hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} &= \frac{1}{m} \hat{\mathbf{X}}^T \hat{\mathbf{y}} = \frac{1}{m} \mathbf{X}^T \mathbf{S}^T \mathbf{S} \mathbf{y}.\end{aligned}$$

we showed that

$$\left\| \hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \right\| \leq \left\| \boldsymbol{\Sigma} \right\| (2\epsilon + \epsilon^2) \quad \text{w.p. } 1 - 1/n$$

$$\left\| \hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}} \right\|_2 \leq \left( RC\sqrt{\log(n)} \left\| \boldsymbol{\beta}^0 \right\|_2 + \sigma \left( 1 + \sqrt{\frac{2\log(n)}{n}} \right) \right) (2\epsilon + \epsilon^2) \left\| \mathbf{X} \right\| \quad \text{w.p. } (1 - 1/n)^2$$

for  $\epsilon = \sqrt{\frac{d}{m}} + \sqrt{\frac{2\log(2n)}{m}}$ . We can summarize by stating

$$\left\| \tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \right\| = \mathcal{O} \left( \sqrt{\frac{d}{m}} \right) = \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{m}} \right) \quad (4.205)$$

$$\left\| \tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}\mathbf{w}} - \boldsymbol{\Sigma}_{\mathbf{Z}\mathbf{w}} \right\|_2 = \mathcal{O} \left( R\sqrt{\frac{d\log(n)}{m}} \right) = \mathcal{O} \left( \sqrt{\frac{d\log(n)}{m}} \right) = \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{m}} \right). \quad (4.206)$$

Ensuring that  $m$  is chosen such that

$$m \geq 2d\eta \left\| \boldsymbol{\Sigma} \right\| \left( 1 + \frac{2\sqrt{2\log(2n)}}{d} + \frac{2\log(2n)}{d} \right) \max \left\{ 2^3 \eta \left\| \boldsymbol{\Sigma} \right\|, 1 \right\}$$

for  $\eta = \frac{2}{\lambda_{\min}(\boldsymbol{\Sigma})}$  satisfies the lambda min requirement and allows us to say

$$\begin{aligned} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 &\leq \frac{4\|\boldsymbol{\beta}^*\|_2 \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| + 4\|(\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{Y}})\|_2}{\lambda_{\min}(\boldsymbol{\Sigma})} \\ &\leq \frac{4\left(2\sqrt{\frac{d}{m}}\epsilon_0 + \frac{d}{m}\epsilon_0^2\right) \left(\|\boldsymbol{\beta}^*\|_2 \|\boldsymbol{\Sigma}\| + \left(RC\sqrt{\log(n)}\|\boldsymbol{\beta}^0\|_2 + \sigma\left(1 + \sqrt{\frac{2\log(n)}{n}}\right)\right) \|\mathbf{X}\|\right)}{\lambda_{\min}(\boldsymbol{\Sigma})} \end{aligned}$$

with probability at least  $1 - 3/n$  where  $\epsilon_0 := \left(1 + \sqrt{\frac{2\log(2n)}{d}}\right)$  to more easily see the dependence on  $d$  and  $m$ .

We now add on the additional error from  $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|$  from the inequality

$$\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\| \leq \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^*\| + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|.$$

Then with probability  $1 - 3/n$ ,

$$\begin{aligned} &\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\|_2 \\ &\frac{4\left(2\sqrt{\frac{d}{m}}\epsilon_0 + \frac{d}{m}\epsilon_0^2\right) \left(\|\boldsymbol{\beta}^*\|_2 \|\boldsymbol{\Sigma}\| + \left(RC\sqrt{\log(n)}\|\boldsymbol{\beta}^0\|_2 + \sigma\left(1 + \sqrt{\frac{2\log(n)}{n}}\right)\right) \|\mathbf{X}\|\right)}{\lambda_{\min}(\boldsymbol{\Sigma})} + \sqrt{\frac{\sigma^2 d}{\lambda_{\min}(\boldsymbol{\Sigma}) n}} \\ &= \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{m}}\right) + \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right) \end{aligned}$$

where again  $\epsilon_0 := \left(1 + \sqrt{\frac{2\log(2n)}{d}}\right)$  is defined to more easily see the dependence on  $d$  and  $m$ .

### 4.3.3 Sketched Scenario with Gaussian $\mathbf{Z}$

Let us assume the rows of  $\mathbf{Z}$ , namely  $\mathbf{z}_i^T$ , are drawn independently from a multivariate normal distribution with mean  $\mathbf{0}$  and covariance  $\mathbf{I}_{d \times d}$ . Assume  $\mathbf{Z}$  and  $\mathbf{w}$  have the linear

relationship  $\mathbf{w} = \mathbf{Z}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\epsilon}$  where the elements of  $\boldsymbol{\epsilon}$  are drawn independently from  $N(0, 1)$  distribution. Let the  $m \times n$  sketching matrix  $\mathbf{S}$  consisting of rows  $\{\mathbf{s}_{i'}^T\}_{i'=1}^m$  drawn IID from a multivariate  $N(\mathbf{0}, \mathbf{I}_{n \times n})$  distribution for some  $m < n$ . Then define

$$\mathbf{X} := \frac{\mathbf{Z}}{\sqrt{n}} \quad \text{and} \quad \mathbf{y} := \frac{\mathbf{w}}{\sqrt{n}} \quad (4.207)$$

and

$$\hat{\mathbf{X}} := \mathbf{S}\mathbf{X}, \quad \hat{\mathbf{y}} := \mathbf{S}\mathbf{y}. \quad (4.208)$$

We use the definitions

$$\boldsymbol{\Sigma} = \mathbf{E} \left[ (\mathbf{X} - \mathbf{E}[\mathbf{X}])^T (\mathbf{X} - \mathbf{E}[\mathbf{X}]) \right] = \mathbf{E} \left[ \mathbf{X}^T \mathbf{X} \right] = \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{Z}}{n} \right] = \mathbf{I}_d \quad (4.209)$$

$$\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}} = \mathbf{E} \left[ \mathbf{X}^T \mathbf{y} \right] = \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{w}}{n} \right], \quad (4.210)$$

and we define their estimators

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{m} \hat{\mathbf{X}}^T \hat{\mathbf{X}} = \frac{1}{m} \mathbf{X}^T \mathbf{S}^T \mathbf{S} \mathbf{X} \quad (4.211)$$

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} = \frac{1}{m} \hat{\mathbf{X}}^T \hat{\mathbf{y}} = \frac{1}{m} \mathbf{X}^T \mathbf{S}^T \mathbf{S} \mathbf{y}. \quad (4.212)$$

Then we show

**Theorem 4.3.2.** *Let  $\mathbf{Z}$  be a  $n \times d$  matrix whose elements are drawn independently from a standard normal distribution. Let  $\mathbf{S}$  be a  $m \times n$  sketching matrix  $\mathbf{S}$  consisting of rows  $\{\mathbf{s}_{i'}^T\}_{i'=1}^m$  drawn IID from a multivariate  $N(\mathbf{0}, \mathbf{I}_{n \times n})$  distribution for some  $m < n$ . Let  $\mathbf{w} = \mathbf{Z}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\epsilon}$  where the elements of  $\boldsymbol{\epsilon}$  are drawn independently from  $N(0, 1)$  distribution. Define*

$$\mathbf{X} := \frac{\mathbf{Z}}{\sqrt{n}} \quad \text{and} \quad \mathbf{y} := \frac{\mathbf{w}}{\sqrt{n}}$$

and

$$\hat{\mathbf{X}} := \mathbf{S}\mathbf{X}, \quad \hat{\mathbf{y}} := \mathbf{S}\mathbf{y}. \quad (4.213)$$

Then define  $\boldsymbol{\Sigma}$ ,  $\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}}$ ,  $\hat{\boldsymbol{\Sigma}}$ , and  $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}}$  as in (4.209), (4.210), (4.211), and (4.212), respectively.

Then for

$$\begin{aligned} \epsilon_0 &= 1 + \sqrt{\frac{2 \log(2n)}{d}}, & \eta &= \frac{2}{\lambda_{\min}(\boldsymbol{\Sigma})}, & b_0 &= 2\hat{R}_g^2 + d^{-1} \|\boldsymbol{\Sigma}\| \\ \hat{R}_g &= \left(1 + \sqrt{2 \log(nd)}\right) \sqrt{2 \log(2nmd)}, & C &= \|\boldsymbol{\beta}^0\|_2 + \sigma, \\ D &= \left(1 + \sqrt{\frac{2 \log(n)}{n}} + \sqrt{\frac{d}{n}}\right) \left(1 + \sqrt{\frac{2 \log(n)}{n}}\right) C, \end{aligned}$$

and  $m$  chosen such that

$$m \geq \max \left\{ 2^4 d \log^2(nd) \log(2nd) \eta^2 \|\boldsymbol{\Sigma}\| \left(2 + \frac{1}{\log(nd)} + \frac{2\sqrt{2}}{\sqrt{\log(nd)}}\right), 2^2 \log(nd) b \eta \right\},$$

then with probability at least  $1 - 8/n$

$$\begin{aligned} &\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 \\ &\leq \frac{4}{\lambda_{\min}(\boldsymbol{\Sigma})} \sqrt{\frac{d}{m}} \left( \|\boldsymbol{\beta}^*\|_2 \left( \sqrt{2 \log(nd) \hat{R}_g^2 \|\boldsymbol{\Sigma}\|} + 2 \log(nd) b_0 \sqrt{\frac{d}{m}} \right) + \left( 2\epsilon_0 + \epsilon_0^2 \sqrt{\frac{d}{m}} \right) D \right) \\ &= \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{m}} \right). \end{aligned}$$

**Proof of 4.3.2**

*Proof.* We wish to bound

$$\left\| \hat{\beta} - \beta^* \right\|_2 \leq \frac{4 \|\beta^*\|_2 \left\| \tilde{\Sigma} - \Sigma \right\| + 4 \left\| (\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}}) \right\|_2}{\lambda_{\min}(\Sigma)}$$

by controlling

$$(a): \left\| \tilde{\Sigma} - \Sigma \right\| \quad (b): \left\| \tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}} \right\|_2.$$

**Analyzing Part (a) in the Sketched Scenario with Gaussian  $\mathbf{Z}$**  The results from section 4.3.3 allow us to immediately say with probability at least  $1 - 4/n$

$$\left\| \hat{\Sigma} - \Sigma \right\| \leq \sqrt{\frac{2d \log(nd) \hat{R}_g^2 \|\Sigma\|}{m}} + \frac{2 \log(nd)b}{m} \tag{4.214}$$

where  $\hat{R}_g = \left(1 + \sqrt{2 \log(nd)}\right) \sqrt{2 \log(2nmd)}$ . To ensure the lambda min requirement is met, we must ensure that

$$\sqrt{\frac{2d \log(nd) \hat{R}_g^2 \|\Sigma\|}{m}} + \frac{2 \log(nd)b}{m} \leq \frac{\lambda_{\min}(\Sigma)}{2}. \tag{4.215}$$

It is sufficient to show that for a given  $m$  each term is less than  $\frac{\lambda_{\min}(\Sigma)}{4}$ . Examining the first term:

$$\begin{aligned}
 & \sqrt{\frac{2d \log(nd) \left( \left(1 + \sqrt{2 \log(nd)}\right) \sqrt{2 \log(2nmd)} \right)^2 \|\Sigma\|}{m}} \leq \frac{\lambda_{\min}(\Sigma)}{4} \\
 \implies & \sqrt{\frac{4d \log(nd) \left(1 + \sqrt{2 \log(nd)}\right)^2 \log(2nmd) \|\Sigma\|}{m}} \leq \frac{\lambda_{\min}(\Sigma)}{4} \\
 \implies & \frac{4d \log(nd) \left(1 + \sqrt{2 \log(nd)}\right)^2 \log(2nd) \|\Sigma\|}{m} \leq \frac{\lambda_{\min}^2(\Sigma)}{2^4} \\
 \implies & m \geq \frac{2^6 d \log(nd) \left(1 + \sqrt{2 \log(nd)}\right)^2 \log(2nd) \|\Sigma\|}{\lambda_{\min}^2(\Sigma)} \\
 \implies & m \geq d \log(nd) \log(2nd) \left(1 + 2\sqrt{2 \log(nd)} + 2 \log(nd)\right) \frac{2^6 \|\Sigma\|}{\lambda_{\min}^2(\Sigma)} \\
 \implies & m \geq 2^4 d \log^2(nd) \log(2nd) \eta^2 \|\Sigma\| \left(2 + \frac{1}{\log(nd)} + \frac{2\sqrt{2}}{\sqrt{\log(nd)}}\right) \tag{4.216}
 \end{aligned}$$

where

$$\eta = \frac{2}{\lambda_{\min}(\Sigma)}. \tag{4.217}$$

Examining the second term:

$$\begin{aligned}
 \frac{2 \log(nd)b}{m} & \leq \frac{\lambda_{\min}(\Sigma)}{4} \\
 \implies m & \geq 2^2 \log(nd)b\eta. \tag{4.218}
 \end{aligned}$$

Thus, choosing  $m$  such that

$$m \geq \max \left\{ 2^4 d \log^2(nd) \log(2nd) \eta^2 \|\Sigma\| \left( 2 + \frac{1}{\log(nd)} + \frac{2\sqrt{2}}{\sqrt{\log(nd)}} \right), 2^2 \log(nd) b \eta \right\} \quad (4.219)$$

ensures the condition is met.

**Analyzing Part (b) in the Sketched Scenario with Gaussian  $\mathbf{Z}$**  We restate the inequality from section 4.3.2:

$$\left\| \hat{\Sigma}_{\mathbf{X}\mathbf{y}} - \Sigma_{\mathbf{X}\mathbf{y}} \right\|_2 = \left\| \frac{1}{m} \mathbf{X}^T \mathbf{S}^T \mathbf{S} \mathbf{y} - \mathbf{X}^T \mathbf{y} \right\|_2 = \left\| \mathbf{X}^T \left( \frac{1}{m} \mathbf{S}^T \mathbf{S} - \mathbf{I} \right) \mathbf{y} \right\|_2 \leq \|\mathbf{X}\| \left\| \frac{1}{m} \mathbf{S}^T \mathbf{S} - \mathbf{I} \right\| \|\mathbf{y}\|_2. \quad (4.220)$$

Then applying Lemmas 8 and 10, which upper bounded  $|x_{kj}|$  with probability  $1 - 2/n$  and  $\|\mathbf{y}\|$  with probability  $1 - 2/n$ , using union bounds, we can refine this bound and say

$$\left\| \hat{\Sigma}_{\mathbf{X}\mathbf{y}} - \Sigma_{\mathbf{X}\mathbf{y}} \right\|_2 \leq \left( 1 + \delta + \sqrt{\frac{d}{n}} \right) \left\| \frac{1}{m} \mathbf{S}^T \mathbf{S} - \mathbf{I} \right\| (1 + \delta) \left( \|\beta^0\|_2 + \sigma \right) \quad (4.221)$$

with probability at least  $1 - (2/n + 2/n)$  for  $\delta = \sqrt{\frac{2 \log(n)}{n}}$ . Now we reuse the result from section 4.3.2, which states

$$\left\| \frac{1}{m} \mathbf{S}^T \mathbf{S} - \mathbf{I} \right\| \leq 2\epsilon + \epsilon^2 \quad (4.222)$$

with probability at least  $1 - 1/n$  for  $\epsilon = \sqrt{\frac{d}{m}} + \sqrt{\frac{2\log(2n)}{m}}$ . Combining the elements, we conclude by stating that

$$\begin{aligned} \left\| \hat{\Sigma}_{\mathbf{X}_Y} - \Sigma_{\mathbf{X}_Y} \right\|_2 &\leq \left( 1 + \delta + \sqrt{\frac{d}{n}} \right) (2\epsilon + \epsilon^2) (1 + \delta) C \\ &= \left( 1 + \sqrt{\frac{2\log(n)}{n}} + \sqrt{\frac{d}{n}} \right) (2\epsilon + \epsilon^2) \left( 1 + \sqrt{\frac{2\log(n)}{n}} \right) C \end{aligned} \quad (4.223)$$

with probability at least  $1 - (2/n + 2/n + 1/n) = 1 - 5/n$  for  $C = \left\| \beta^0 \right\|_2 + \sigma$ .

Combining (4.214) and (4.223) we have with probability at least  $1 - (5/n + 3/n) = 1 - 8/n$

$$\left\| \beta^* - \hat{\beta} \right\|_2 \quad (4.224)$$

$$\leq \frac{4 \left\| \beta^* \right\|_2 \left( \sqrt{\frac{2d \log(nd) \hat{R}_g^2 \left\| \Sigma \right\|}{m}} + \frac{2 \log(nd) b}{m} \right) + 4 \left( 1 + \sqrt{\frac{2\log(n)}{n}} + \sqrt{\frac{d}{n}} \right) (2\epsilon + \epsilon^2) \left( 1 + \sqrt{\frac{2\log(n)}{n}} \right) C}{\lambda_{\min}(\Sigma)}. \quad (4.225)$$

□

The following sections show the lemmas used in the proof.

## Establishing a High-Probability Bound on the Sketched Data with Gaussian Z

We will establish a high-probability bound for the elements of  $\hat{\mathbf{X}}$ . Then we define

$$\hat{R}_g = \left( 1 + \sqrt{2 \log(nd)} \right) \sqrt{2 \log(2nmd)} \quad (4.226)$$

so we can say  $|\hat{x}_{kj}| \leq \hat{R}_g$  for all  $k, j$  with probability at least  $1 - 2/n$ .

**Lemma 8.** *Let  $\mathbf{Z}$  be a  $n \times d$  matrix whose elements are drawn independently from a standard normal distribution. Let  $\mathbf{S}$  be a  $m \times n$  sketching matrix  $\mathbf{S}$  consisting of rows  $\{\mathbf{s}_{i'}^T\}_{i'=1}^m$  drawn*

*IID from a multivariate  $N(\mathbf{0}, \mathbf{I}_{n \times n})$  distribution for some  $m < n$ . Define  $\mathbf{X} := \frac{\mathbf{Z}}{\sqrt{n}}$  and  $\hat{\mathbf{X}} := \mathbf{S}\mathbf{X}$  where  $\hat{x}_{kj}$  are the elements of  $\hat{\mathbf{X}}$  for  $k = 1, \dots, m$  and  $j = 1, \dots, m$ . Then with probability at least  $1 - 2/n$*

$$|\hat{x}_{kj}| \leq \left(1 + \sqrt{2 \log(nd)}\right) \sqrt{2 \log(2nmd)}$$

*for all  $k, j$ .*

**Proof of 8**

*Proof.* We begin by examining the columns of  $\mathbf{X}$  where we will denote the  $j$ th column of  $\mathbf{X}$  as  $\mathbf{x}_{\cdot j}$ . Using the fact that the euclidean norm is 1-Lipschitz, we can use Theorem 2.26 in Wainwright [69] to say

$$\mathbb{P} \left( \|\mathbf{x}_{\cdot j}\|_2 - \mathbf{E} \left[ \|\mathbf{x}_{\cdot j}\|_2 \right] \geq \delta \right) \leq \exp \left( -\delta^2/2 \right) \tag{4.227}$$

for all  $\delta \geq 0$ . Then, using the concavity of the square root function and Jensen's inequality, we have that  $\mathbf{E} \left[ \|\mathbf{x}_{\cdot j}\|_2 \right] \leq 1$ . Then we can say

$$\mathbb{P} \left( \|\mathbf{x}_{\cdot j}\|_2^2 \geq (1 + \delta)^2 \right) \leq \exp \left( -\delta^2/2 \right) \tag{4.228}$$

for all  $\delta \geq 0$ . We use a union bound argument to say

$$\mathbb{P} \left( \max_j \|\mathbf{x}_{\cdot j}\|_2^2 \geq (1 + \delta)^2 \right) \leq d \exp \left( -\delta^2/2 \right) \tag{4.229}$$

for all  $\delta \geq 0$ . Then if we let  $\delta = \sqrt{2 \log(nd)}$  we can say

$$\mathbb{P} \left( \max_j \|\mathbf{x}_{\cdot j}\|_2^2 \geq (1 + \delta)^2 \right) \leq \frac{1}{n}. \quad (4.230)$$

Let us define the event

$$E : \max_j \|\mathbf{x}_{\cdot j}\|_2^2 \leq \left(1 + \sqrt{2 \log(nd)}\right)^2, \quad (4.231)$$

which is true with probability  $1 - 1/n$ . We will use this event as we examine the elements of  $\hat{\mathbf{X}}$ , denoted as  $\hat{x}_{kj} = \mathbf{s}_k^T \mathbf{x}_{\cdot j} = \sum_{i=1}^n x_{ij} s_{ki}$  where  $\mathbf{x}_{\cdot j}$  is the  $j$ th column of  $\mathbf{X}$  and  $\mathbf{s}_k^T$  is the  $k$ th row of  $\mathbf{S}$ . Since  $s_{ki} \sim N(0, 1)$  for all  $k, i$ , then we know that  $\hat{x}_{kj} \mid \mathbf{X} \sim N\left(0, \|\mathbf{x}_{\cdot j}\|_2^2\right)$ .

This also means that

$$\mathbf{Vars} [\hat{x}_{kj} \mid E] = \mathbf{E}_{\mathbf{S}} \left[ \mathbf{x}_{\cdot j}^T \left( \mathbf{s}_k^T \right)^T \mathbf{s}_k^T \mathbf{x}_{\cdot j} \mid E \right] = \|\mathbf{x}_{\cdot j}\|_2^2 \leq \left(1 + \sqrt{2 \log(nd)}\right)^2. \quad (4.232)$$

Here we used that  $\left( \mathbf{s}_k^T \right)^T \mathbf{s}_k^T$  is a  $n \times n$  diagonal matrix whose entries are each distributed according to a  $\chi$ -squared distribution with 1 degree of freedom. Thus,  $\mathbf{E} \left[ \left( \mathbf{s}_k^T \right)^T \mathbf{s}_k^T \right] = \mathbf{I}_{n \times n}$ .

Now we can apply a Gaussian tail bound to our conditioned sketched data and use a union bound to say

$$\mathbb{P} \left( \max_{k,j} |\hat{x}_{kj}| \geq t \mid E \right) \leq 2md \exp \left( \frac{-t^2}{2\sigma_{\hat{x}}^2} \right) \leq 2md \exp \left( -\frac{t^2}{2 \left(1 + \sqrt{2 \log(nd)}\right)^2} \right) \quad (4.233)$$

where  $\sigma_{\hat{x}}^2 = \mathbf{Var} [\hat{x}_{kj} | E]$ . Letting

$$t = \sqrt{2 \left(1 + \sqrt{2 \log(nd)}\right)^2 \log(2nmd)} = \left(1 + \sqrt{2 \log(nd)}\right) \sqrt{2 \log(2nmd)} \quad (4.234)$$

then

$$\mathbb{P} \left( \max_{k,j} |\hat{x}_{kj}| \geq t \mid E \right) \leq \frac{1}{n}. \quad (4.235)$$

Then applying the law of total probability

$$\begin{aligned} \mathbb{P} \left( \max_{k,j} |\hat{x}_{kj}| \geq t \right) &\leq \mathbb{P} \left( \max_{k,j} |\hat{x}_{kj}| \geq t \mid E \right) + \mathbb{P}(E^c) \\ &\leq \frac{1}{n} + \frac{1}{n} = \frac{2}{n}. \end{aligned} \quad (4.236)$$

Then we have

$$|\hat{x}_{kj}| \leq \left(1 + \sqrt{2 \log(nd)}\right) \sqrt{2 \log(2nmd)} \quad (4.237)$$

for all  $k, j$  with probability at least  $1 - 2/n$ . □

### Matrix Bernstein Inequality for the Sketched Scenario with Gaussian $\mathbf{Z}$

**Lemma 9.** *Let  $\mathbf{Z}$  be a  $n \times d$  matrix whose elements are drawn independently from a standard normal distribution. Let  $\mathbf{S}$  be a  $m \times n$  sketching matrix  $\mathbf{S}$  consisting of rows  $\{\mathbf{s}_j^T\}_{j=1}^m$  drawn IID from a multivariate  $N(\mathbf{0}, \mathbf{I}_{n \times n})$  distribution for some  $m < n$ . Define  $\mathbf{X} := \frac{\mathbf{Z}}{\sqrt{n}}$  and  $\hat{\mathbf{X}} := \mathbf{S}\mathbf{X}$  where  $\hat{x}_{kj}$  are the elements of  $\hat{\mathbf{X}}$  for  $k = 1, \dots, m$  and  $j = 1, \dots, n$ . Then for  $\Sigma$  and  $\hat{\Sigma}$  as defined in (4.172) and (4.211) and*

$$b \geq 2d\hat{R}_g^2 + \|\Sigma\| \quad \text{and} \quad \hat{R}_g = \left(1 + \sqrt{2 \log(nd)}\right) \sqrt{2 \log(2nmd)},$$

then with probability at least  $1 - 3/n$

$$\|\hat{\Sigma} - \Sigma\| \leq \sqrt{\frac{2d \log(2nd) \hat{R}_g^2 \|\Sigma\|}{m}} + \frac{2 \log(2nd)b}{m}.$$

**Proof of 9**

*Proof.* Let us begin by defining

$$\hat{\Sigma}_k = \hat{\mathbf{x}}_k \hat{\mathbf{x}}_k^T \tag{4.238}$$

where  $\hat{\mathbf{x}}_k^T$  is the  $k$ th row of  $\hat{\mathbf{X}}$  and  $\hat{\mathbf{x}}_k$  is the transpose of the  $k$ th row of  $\hat{\mathbf{X}}$ . Then we can say

$$\hat{\Sigma} = \frac{1}{m} \sum_{k=1}^m \hat{\Sigma}_k. \tag{4.239}$$

Let us define the event

$$E : \max_{kj} |\hat{x}_{kj}| \leq \hat{R}_g \tag{4.240}$$

which we know is true with probability at least  $1 - 2/n$  by Lemma 8. We now apply the Matrix Bernstein inequality (Theorem 6.17 in Wainwright)[69]. We use the fact that  $\mathbf{Var} [\hat{\Sigma}_k - \Sigma] = \mathbf{Var} [\hat{\Sigma}_k]$ , and apply the Bernstein result to the zero-mean matrices  $\{\hat{\Sigma}_k - \Sigma\}_{k=1}^m$ . Then we can say

$$\mathbb{P} \left( \frac{1}{m} \left\| \sum_{k=1}^m (\hat{\Sigma}_k - \Sigma) \right\| \geq \delta \mid E \right) \leq 2 \text{rank} \left( \sum_{k=1}^m \mathbf{Var} [\hat{\Sigma}_k \mid E] \right) \exp \left( \frac{-m\delta^2}{2(\sigma_{\hat{\Sigma}}^2 + b\delta)} \right) \tag{4.241}$$

for some  $b \geq \left\| \hat{\Sigma}_k - \Sigma \right\|$  and  $\delta \geq 0$  and where

$$\sigma_{\hat{\Sigma}}^2 = \frac{1}{m} \left\| \sum_{k=1}^m \mathbf{Var} \left[ \hat{\Sigma}_k \mid E \right] \right\| = \frac{1}{m} \left\| m \mathbf{Var} \left[ \hat{\Sigma}_k \mid E \right] \right\| = \left\| \mathbf{Var} \left[ \hat{\Sigma}_k \mid E \right] \right\|. \quad (4.242)$$

In order to use the Matrix Bernstein inequality, we must show that  $\hat{\Sigma}_k - \Sigma$  conditioned on the event  $E$  satisfies the condition

$$\left\| \hat{\Sigma}_k - \Sigma \right\|_2 \leq b \quad (4.243)$$

almost surely. Given event  $E$ , we know immediately that any element of  $\hat{\Sigma}_k$  is bounded in absolute value by  $\hat{R}_g^2$ . Using this, we can then show

$$\begin{aligned} \left\| \hat{\Sigma}_k - \Sigma \right\|_2^2 &\leq \left\| \hat{\Sigma}_k \right\|_2^2 + \left\| \Sigma \right\|_2^2 \\ &= \sup_{\|\mathbf{v}\|_2=1} \mathbf{v}^T \hat{\Sigma}_k^T \hat{\Sigma}_k \mathbf{v} + \left\| \Sigma \right\|_2^2 \\ &= \sup_{\|\mathbf{v}\|_2=1} \left| \sum_{j,j'}^d v_j v_{j'} \left( \sum_{\ell} \left( \hat{\Sigma}_k \right)_{\ell j} \left( \hat{\Sigma}_k \right)_{\ell j'} \right) \right| + \left\| \Sigma \right\|_2^2 \\ &\leq \sup_{\|\mathbf{v}\|_2=1} \sum_{j,j'}^d |v_j| |v_{j'}| \left| \sum_{\ell} \left( \hat{\Sigma}_k \right)_{\ell j} \left( \hat{\Sigma}_k \right)_{\ell j'} \right| + \left\| \Sigma \right\|_2^2 \\ &\leq d \left( 2\hat{R}_g^2 \right)^2 \sup_{\|\mathbf{v}\|_2=1} \left\| \mathbf{v} \right\|_1^2 + \left\| \Sigma \right\|_2^2 \\ &= d^2 \left( 2\hat{R}_g^2 \right)^2 + \left\| \Sigma \right\|_2^2 \\ \implies \left\| \hat{\Sigma}_k - \Sigma \right\|_2 &\leq 2d\hat{R}_g^2 + \left\| \Sigma \right\|_2, \end{aligned} \quad (4.244)$$

satisfying the Bernstein condition. Bounding  $\sigma_{\hat{\Sigma}}^2$  yields

$$\begin{aligned}
 \sigma_{\hat{\Sigma}}^2 &= \left\| \mathbf{Var} \left[ \hat{\Sigma}_k \mid E \right] \right\| = \left\| \mathbf{E} \left[ \hat{\Sigma}_k^2 \mid E \right] - \mathbf{E} \left[ \hat{\Sigma}_k \mid E \right]^2 \right\| = \left\| \mathbf{E} \left[ \hat{\Sigma}_k^2 \mid E \right] - \Sigma^2 \right\| \\
 &\leq \left\| \mathbf{E} \left[ \hat{\mathbf{x}}_k \hat{\mathbf{x}}_k^T \hat{\mathbf{x}}_k \hat{\mathbf{x}}_k^T \mid E \right] \right\| \\
 &= \left\| \mathbf{E} \left[ \hat{\mathbf{x}}_k \|\hat{\mathbf{x}}_k\|_2^2 \hat{\mathbf{x}}_k^T \mid E \right] \right\| \\
 &\leq d \hat{R}_g^2 \left\| \mathbf{E} \left[ \hat{\mathbf{x}}_k \hat{\mathbf{x}}_k^T \mid E \right] \right\| \\
 &= d \hat{R}_g^2 \|\Sigma\|. \tag{4.245}
 \end{aligned}$$

We were able to drop  $\|\Sigma^2\|$  in the second step using the fact that  $\mathbf{Var} \left[ \hat{\Sigma}_k \right] \geq 0$  and  $\Sigma \geq 0$  and applying (D.1). In the last step we used that the  $k, j$ th element of  $\hat{\mathbf{x}}_k \hat{\mathbf{x}}_k^T$  can be written as  $\hat{x}_{kj} = \sum_{i=1}^n x_{ij} s_{ki}$ . Then for the  $j, j'$  element of  $\hat{\mathbf{x}}_k \hat{\mathbf{x}}_k^T$  where  $j \neq j'$

$$\begin{aligned}
 \mathbf{E} \left[ \left( \hat{\mathbf{x}}_k \hat{\mathbf{x}}_k^T \right)_{j,j'} \mid E \right] &= \mathbf{E} \left[ \left( \sum_{i=1}^n x_{ij} s_{ki} \right) \left( \sum_{i'=1}^n x_{i'j'} s_{ki'} \right) \mid E \right] \\
 &= \mathbf{E} \left[ \sum_{i=1}^n x_{ij} s_{ki} x_{i'j'} s_{ki'} \mid E \right] \\
 &= \mathbf{E} \left[ \sum_{i=1}^n \sum_{i'=i} x_{ij} x_{i'j'} s_{ki}^2 \mid E \right] + \mathbf{E} \left[ \sum_{i=1}^n \sum_{i' \neq i} x_{ij} x_{i'j'} s_{ki}^2 \mid E \right] \\
 &= \sum_{i=1}^n \mathbf{E}_{\mathbf{X}} [x_{ij} x_{ij'} \mid E] \mathbf{E}_{\mathbf{S}} [s_{ki}^2 \mid E] + \sum_{i=1}^n \sum_{i' \neq i} \mathbf{E}_{\mathbf{X}} [x_{ij} x_{i'j'} \mid E] \mathbf{E}_{\mathbf{S}} [s_{ki}^2 \mid E] \\
 &= \sum_{i=1}^n \mathbf{E}_{\mathbf{X}} [x_{ij} x_{ij'}] \tag{4.246}
 \end{aligned}$$

by the independence and zero mean of the rows of  $\mathbf{Z}$  and that the variance of the elements of  $\mathbf{S}$  is 1. In the last step, the conditioning is dropped because it has no effect on the

distribution of the elements of  $\mathbf{X}$ . For the case when  $j = j'$ :

$$\begin{aligned}
 \mathbf{E} \left[ \left( \hat{\mathbf{x}}_k \hat{\mathbf{x}}_k^T \right)_{j,j'} \middle| E \right] &= \mathbf{E} \left[ \left( \sum_{i=1}^n x_{ij} s_{ki} \right) \left( \sum_{i'=1}^n x_{i'j} s_{ki'} \right) \middle| E \right] \\
 &= \mathbf{E} \left[ \sum_{i=1}^n x_{ij} s_{ki} x_{i'j} s_{ki'} \middle| E \right] \\
 &= \mathbf{E} \left[ \sum_{i=1}^n \sum_{i'=1} x_{ij}^2 s_{ki}^2 \middle| E \right] + \mathbf{E} \left[ \sum_{i=1}^n \sum_{i' \neq i} x_{ij} x_{i'j} s_{ki}^2 \middle| E \right] \\
 &= \sum_{i=1}^n \mathbf{E} \left[ x_{ij}^2 \right] \tag{4.247}
 \end{aligned}$$

using the independence of the rows of  $\mathbf{Z}$  and that the variance of each element of  $\mathbf{S}$  is 1. In the last step, the conditioning is dropped because it has no effect on the distribution of the elements of  $\mathbf{X}$ . Combining these facts,

$$\begin{aligned}
 &\mathbf{E} \left[ \left( \hat{\mathbf{x}}_k \hat{\mathbf{x}}_k^T \right)_{j,j'} \middle| E \right] \\
 &= \sum_{i=1}^n \mathbf{E} \left[ \begin{bmatrix} x_{i1}^2 & x_{i1}x_{i2} & \dots & x_{i1}x_{id} \\ x_{i2}x_{i1} & x_{i2}^2 & \dots & x_{i2}x_{id} \\ \vdots & \vdots & \ddots & \vdots \\ x_{id}x_{i1} & x_{id}x_{i2} & \dots & x_{id}^2 \end{bmatrix} \right] = \sum_{i=1}^n \mathbf{E} \left[ \mathbf{x}_i \mathbf{x}_i^T \right] = \mathbf{E} \left[ \mathbf{X}^T \mathbf{X} \right] = \Sigma \tag{4.248}
 \end{aligned}$$

Proceeding, since  $\sigma_{\hat{\Sigma}}^2 = \left\| \mathbf{Var} \left[ \hat{\Sigma}_k \middle| E \right] \right\| \leq d \hat{R}_g^2 \|\Sigma\|$ , we can say

$$\mathbb{P} \left( \frac{1}{m} \left\| \sum_{k=1}^m (\hat{\Sigma}_k - \Sigma) \right\| \geq \delta \middle| E \right) \leq 2d \exp \left( \frac{-m\delta^2}{2(\sigma_{\hat{\Sigma}}^2 + b\delta)} \right) \leq 2d \exp \left( \frac{-m\delta^2}{2(d\hat{R}_g^2 \|\Sigma\| + b\delta)} \right). \tag{4.249}$$

Following the techniques in section 4.2.3, let us choose

$$\delta = \sqrt{\frac{2 \log(2nd) d \hat{R}_g^2 \|\Sigma\|}{m}} + \frac{2 \log(2nd) b}{m} \quad (4.250)$$

so that

$$\frac{m\delta^2}{2(d\hat{R}_g^2\|\Sigma\| + b\delta)} \geq \min \left\{ \frac{m\delta^2}{2d\hat{R}_g^2\|\Sigma\|}, \frac{m\delta}{2b} \right\} \geq \log(2nd). \quad (4.251)$$

Then we have

$$\mathbb{P} \left( \frac{1}{m} \left\| \sum_{k=1}^m (\hat{\Sigma}_k - \Sigma) \right\| \geq \delta \mid E \right) \leq 2d \exp \left( \frac{-m\delta^2}{2(d\hat{R}_g^2\|\Sigma\| + b\delta)} \right) \leq 2d \exp(-\log(2nd)) = \frac{1}{n} \quad (4.252)$$

for  $\delta = \sqrt{\frac{2d \log(2nd) \hat{R}_g^2 \|\Sigma\|}{m}} + \frac{2 \log(2nd) b}{m}$ , and for any  $b \geq 2d\hat{R}_g^2 + \|\Sigma\|$ .

We showed in section 4.3.3 that  $E$  occurs with probability  $1 - 2/n$ . Then using the law of total probability

$$\begin{aligned} \mathbb{P} \left( \frac{1}{m} \left\| \sum_{k=1}^m (\hat{\Sigma}_k - \Sigma) \right\| \geq \delta \right) &\leq \mathbb{P} \left( \frac{1}{m} \left\| \sum_{k=1}^m (\hat{\Sigma}_k - \Sigma) \right\| \geq \delta \mid E \right) + \mathbb{P} \left( \max_{kj} |\hat{x}_{kj}| \geq \hat{R}_g \right) \\ &= \frac{1}{n} + \frac{2}{n} \\ &= \frac{3}{n} \end{aligned} \quad (4.253)$$

Thus we can conclude that

$$\frac{1}{m} \left\| \sum_{k=1}^m (\hat{\Sigma}_k - \Sigma) \right\| = \|\hat{\Sigma} - \Sigma\| \leq \sqrt{\frac{2d \log(2nd) \hat{R}_g^2 \|\Sigma\|}{m}} + \frac{2 \log(2nd) b}{m} \quad (4.254)$$

with probability at least  $1 - 3/n$ . □

### Bounding the Spectral Norm of $\mathbf{X}$ with Gaussian $\mathbf{Z}$

We bound  $\|\mathbf{X}\| = \frac{\|\mathbf{Z}\|}{\sqrt{n}}$  by directly applying a theorem from [69]. We note that  $\mathbf{Z}$  is a  $n \times d$  matrix drawn from a  $\mathbf{I}_{d \times d}$ -Gaussian ensemble. Then using Theorem 6.1 in Wainwright [69]

$$\begin{aligned} \mathbb{P}\left(\frac{\|\mathbf{Z}\|}{\sqrt{n}} \geq \gamma_{\max}(\sqrt{\mathbf{I}})(1 + \delta) + \sqrt{\frac{\text{tr}(\mathbf{I})}{n}}\right) &\leq 2e^{-n\delta^2/2} \\ \mathbb{P}\left(\frac{\|\mathbf{Z}\|}{\sqrt{n}} \geq 1 + \delta + \sqrt{\frac{d}{n}}\right) &\leq 2e^{-n\delta^2/2} \\ \mathbb{P}\left(\|\mathbf{X}\| \geq 1 + \delta + \sqrt{\frac{d}{n}}\right) &\leq 2e^{-n\delta^2/2} \end{aligned} \tag{4.255}$$

where  $\gamma(\mathbf{A})$  indicates the vector of eigenvalues of  $\mathbf{A}$ . Now if we let  $\delta = \sqrt{\frac{2 \log(2n)}{n}}$ , we can say

$$\|\mathbf{X}\| \leq 1 + \delta + \sqrt{\frac{d}{n}} \tag{4.256}$$

with probability at least  $1 - 1/n$ .

### Bounding the Norm of $\mathbf{y}$ with Gaussian $\mathbf{Z}$

**Lemma 10.** *Let  $\mathbf{Z}$  be a  $n \times d$  matrix whose elements are drawn independently from a standard normal distribution. Assume  $\mathbf{Z}$  and  $\mathbf{w}$  have the linear relationship  $\mathbf{w} = \mathbf{Z}\boldsymbol{\beta}^* + \sigma\boldsymbol{\epsilon}$  where the elements of  $\boldsymbol{\epsilon}$  are drawn independently from  $N(0, 1)$  distribution. Then define*

$$\mathbf{X} := \frac{\mathbf{Z}}{\sqrt{n}} \quad \text{and} \quad \mathbf{y} := \frac{\mathbf{w}}{\sqrt{n}}.$$

Then with probability at least  $1 - 2/n$

$$\|\mathbf{y}\|_2 \leq \left(1 + \sqrt{\frac{2 \log(n)}{n}}\right) (\|\boldsymbol{\beta}^*\|_2 + \sigma). \quad (4.257)$$

**Proof of Lemma 10**

*Proof.* We analyze  $\|\mathbf{y}\|_2$  in two steps: 1)  $\|\mathbf{X}\boldsymbol{\beta}^*\|_2$ , and 2)  $\|\boldsymbol{\epsilon}/\sqrt{n}\|_2$ . For both analyses we follow Example 2.28 in Wainwright [69]. Letting

$$D = \frac{n\|\mathbf{X}\boldsymbol{\beta}^*\|_2^2}{\|\boldsymbol{\beta}^*\|_2^2} \quad (4.258)$$

and recognizing that  $D$  is distributed according to a  $\chi^2$  distribution with  $n$  degrees of freedom, we can then apply Theorem 2.26 in Wainwright [69] and say

$$\mathbb{P}\left(\frac{\sqrt{D}}{\sqrt{n}} \geq \mathbf{E}\left[\frac{\sqrt{D}}{\sqrt{n}}\right] + \delta\right) \leq \exp(-n\delta^2/2) \quad (4.259)$$

for all  $\delta \geq 0$ . Using the concavity of the square root function and Jensen's inequality we get

$$\mathbf{E}\left[\frac{\sqrt{D}}{\sqrt{n}}\right] = \mathbf{E}\left[\frac{\sqrt{D}}{\sqrt{n}}\right] \leq \sqrt{\mathbf{E}\left[\frac{D}{n}\right]} = \left(\frac{1}{n}\mathbf{E}[D]\right)^{1/2} = 1, \quad (4.260)$$

where  $\mathbf{E}[D] = n$  since  $D$  is distributed  $\chi^2$  with  $n$  degrees of freedom, as stated before.

Then we can refine our probability

$$\begin{aligned}
 & \mathbb{P}\left(\frac{\sqrt{D}}{\sqrt{n}} \geq 1 + \delta\right) \leq \exp(-n\delta^2/2) \\
 \implies & \mathbb{P}\left(D \geq n(1 + \delta)^2\right) \leq \exp(-n\delta^2/2) \\
 \implies & \mathbb{P}\left(\frac{n\|\mathbf{X}\boldsymbol{\beta}^*\|_2^2}{\|\boldsymbol{\beta}^*\|_2^2} \geq n(1 + \delta)^2\right) \leq \exp(-n\delta^2/2) \\
 \implies & \mathbb{P}\left(\|\mathbf{X}\boldsymbol{\beta}^*\|_2^2 \geq \|\boldsymbol{\beta}^*\|_2^2(1 + \delta)^2\right) \leq \exp(-n\delta^2/2) \\
 \implies & \mathbb{P}\left(\|\mathbf{X}\boldsymbol{\beta}^*\|_2 \geq \|\boldsymbol{\beta}^*\|_2(1 + \delta)\right) \leq \exp(-n\delta^2/2). \tag{4.261}
 \end{aligned}$$

Letting  $\delta = \sqrt{\frac{2\log(n)}{n}}$ , then we can say

$$\|\mathbf{X}\boldsymbol{\beta}^*\|_2 \leq \|\boldsymbol{\beta}^*\|_2(1 + \delta) \tag{4.262}$$

with probability at least  $1 - 1/n$ .

We now examine  $\boldsymbol{\epsilon}/\sqrt{n}$ . We showed in section 4.3.2 that

$$\frac{\|\boldsymbol{\epsilon}\|_2}{\sqrt{n}} \leq 1 + \delta \tag{4.263}$$

with probability at least  $1 - 1/n$  for  $\delta = \sqrt{\frac{2\log(n)}{n}}$ .

Using the triangle inequality and union bounds, we can then say

$$\|\mathbf{y}\|_2 \leq \|\mathbf{X}\boldsymbol{\beta}^*\|_2 + \sigma \frac{\|\boldsymbol{\epsilon}\|_2}{\sqrt{n}} \leq \|\boldsymbol{\beta}^*\|_2(1 + \delta) + \sigma(1 + \delta) = (1 + \delta)(\|\boldsymbol{\beta}^*\|_2 + \sigma) \tag{4.264}$$

with probability at least  $1 - 2/n$  where we have again let  $\delta = \sqrt{\frac{2\log(n)}{n}}$ .

□

### Summary of Sketched Scenario with Gaussian $\mathbf{Z}$

For a random  $\mathbf{Z}$  whose rows  $\mathbf{z}_i^T$  are drawn independently from a multivariate normal distribution with mean  $\mathbf{0}$  and covariance  $\mathbf{I}_{d \times d}$  and definitions

$$\boldsymbol{\Sigma} = \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{Z}}{n} \right] = \mathbf{I}_d \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{m} \hat{\mathbf{X}}^T \hat{\mathbf{X}} = \frac{1}{m} \mathbf{X}^T \mathbf{S}^T \mathbf{S} \mathbf{X} \quad (4.265)$$

$$\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}} = \mathbf{E} \left[ \mathbf{X}^T \mathbf{y} \right] = \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{w}}{n} \right] \quad \hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} = \frac{1}{m} \hat{\mathbf{X}}^T \hat{\mathbf{y}} = \frac{1}{m} \mathbf{X}^T \mathbf{S}^T \mathbf{S} \mathbf{y}, \quad (4.266)$$

we showed

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| \leq \sqrt{\frac{2d \log(nd) \hat{R}_g^2 \|\boldsymbol{\Sigma}\|}{m}} + \frac{2 \log(nd) b}{m} \quad \text{w.p. } 1 - 3/n \quad (4.267)$$

$$\|\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}}\|_2 \leq \left(1 + \sqrt{\frac{2 \log(n)}{n}} + \sqrt{\frac{d}{n}}\right) (2\epsilon + \epsilon^2) \left(1 + \sqrt{\frac{2 \log(n)}{n}}\right) C \quad \text{w.p. } 1 - 5/n \quad (4.268)$$

for

$$\epsilon = \sqrt{\frac{d}{m}} + \sqrt{\frac{2 \log(2n)}{m}} \quad b \geq 2d \hat{R}_g^2 + \|\boldsymbol{\Sigma}\|$$

$$\hat{R}_g = \left(1 + \sqrt{2 \log(nd)}\right) \sqrt{2 \log(2nmd)} \quad C = \|\boldsymbol{\beta}^*\|_2 + \sigma$$

$$D := \left(1 + \sqrt{\frac{2 \log(n)}{n}} + \sqrt{\frac{d}{n}}\right) \left(1 + \sqrt{\frac{2 \log(n)}{n}}\right) C$$

We can summarize by stating

$$\left\| \tilde{\Sigma} - \Sigma \right\| = \mathcal{O} \left( \sqrt{\frac{d \log(nd) \log(nmd)}{m}} \right) = \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{m}} \right) \quad (4.269)$$

$$\left\| \tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} - \Sigma_{\mathbf{Z}\mathbf{w}} \right\|_2 = \mathcal{O} \left( \sqrt{\frac{d}{m}} \right) = \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{m}} \right). \quad (4.270)$$

Ensuring  $m$  is chosen such that

$$m \geq \max \left\{ 2^4 d \log^2(nd) \log(2nd) \eta^2 \left\| \Sigma \right\| \left( 2 + \frac{1}{\log(nd)} + \frac{2\sqrt{2}}{\sqrt{\log(nd)}} \right), 2^2 \log(nd) b \eta \right\} \quad (4.271)$$

where  $\eta = \frac{2}{\lambda_{\min}(\Sigma)}$  satisfies the lambda-min-requirement. Then

$$\begin{aligned} \left\| \hat{\beta} - \beta^0 \right\|_2 &\leq \frac{4 \left\| \beta^* \right\|_2 \left\| \hat{\Sigma} - \Sigma \right\| + 4 \left\| (\hat{\Sigma}_{\mathbf{X}\mathbf{y}} - \Sigma_{\mathbf{X}\mathbf{y}}) \right\|_2}{\lambda_{\min}(\Sigma)} \\ &\leq \frac{4 \left\| \beta^* \right\|_2 \left( \sqrt{\frac{2d \log(nd) \hat{R}_g^2 \left\| \Sigma \right\|}{m}} + \frac{2 \log(nd) b}{m} \right) + 4 \left( 1 + \sqrt{\frac{2 \log(n)}{n}} + \sqrt{\frac{d}{n}} \right) (2\epsilon + \epsilon^2) \left( 1 + \sqrt{\frac{2 \log(n)}{n}} \right) C}{\lambda_{\min}(\Sigma)} \\ &= \frac{4 \left\| \beta^* \right\|_2}{\lambda_{\min}(\Sigma)} \left( \sqrt{\frac{2d \log(nd) \hat{R}_g^2 \left\| \Sigma \right\|}{m}} + \frac{2d \log(nd) b_0}{m} \right) \\ &\quad + \frac{4}{\lambda_{\min}(\Sigma)} \left( 2\sqrt{\frac{d}{m}} \epsilon_0 + \frac{d}{m} \epsilon_0^2 \right) \left( 1 + \sqrt{\frac{2 \log(n)}{n}} + \sqrt{\frac{d}{n}} \right) \left( 1 + \sqrt{\frac{2 \log(n)}{n}} \right) C \\ &= \frac{4}{\lambda_{\min}(\Sigma)} \sqrt{\frac{d}{m}} \left( \left\| \beta^* \right\|_2 \left( \sqrt{2 \log(nd) \hat{R}_g^2 \left\| \Sigma \right\|} + 2 \log(nd) b_0 \sqrt{\frac{d}{m}} \right) + \left( 2\epsilon_0 + \epsilon_0^2 \sqrt{\frac{d}{m}} \right) D \right) \quad (4.272) \\ &= \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{m}} \right) \quad (4.273) \end{aligned}$$

with probability at least  $1 - 8/n$  where  $b_0 := 2\hat{R}_g^2 + d^{-1} \left\| \Sigma \right\|$  and  $\epsilon_0 := 1 + \sqrt{\frac{2 \log(2n)}{d}}$  are defined to more easily see the dependence on  $d$  and  $m$  in the final formulation.

## 4.4 Sketched and Quantized Regression Parameters

This section examines the effect of transforming the scaled design matrix  $\mathbf{X} = \frac{\mathbf{Z}}{\sqrt{n}}$  and scaled response vector  $\mathbf{y} = \frac{\mathbf{w}}{\sqrt{n}}$  by applying a sketching transformation followed by a 1-bit quantization.

In the sections where  $\mathbf{Z}$  is assumed to be fixed, let us define

$$\Sigma = \mathbf{X}^T \mathbf{X} = \frac{\mathbf{Z}^T \mathbf{Z}}{n} \quad (4.274)$$

$$\Sigma_{\mathbf{X}\mathbf{y}} = \mathbf{E}_\epsilon \left[ \mathbf{X}^T \mathbf{y} \right] = \mathbf{E}_\epsilon \left[ \mathbf{X}^T (\mathbf{X}\beta^* + \sigma\epsilon) \right] = \mathbf{X}^T \mathbf{X}\beta^* = \frac{\mathbf{Z}^T \mathbf{Z}\beta^*}{n}. \quad (4.275)$$

When we assume  $\mathbf{Z}$  consists of random variables, let us define

$$\Sigma = \mathbf{E} \left[ \mathbf{X}^T \mathbf{X} \right] = \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{Z}}{n} \right] \quad (4.276)$$

$$\Sigma_{\mathbf{X}\mathbf{y}} = \mathbf{E} \left[ \mathbf{X}^T \mathbf{y} \right] = \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{w}}{n} \right]. \quad (4.277)$$

### 4.4.1 Introduction

Throughout this chapter we perform our analysis on a scaled design matrix and response vector:

$$\mathbf{X} = \frac{\mathbf{Z}}{\sqrt{n}} \quad \text{and} \quad \mathbf{y} = \frac{\mathbf{w}}{\sqrt{n}} \quad (4.278)$$

where, as before,  $\mathbf{Z}$  is a  $n \times d$  matrix and  $\mathbf{w}$  has a linear relationship with  $\mathbf{Z}$  such that  $\mathbf{w} = \mathbf{Z}\beta^0 + \sigma\epsilon$  for  $\epsilon \sim N(\mathbf{0}, \mathbf{I}_{n \times n})$ . Now let the  $m \times n$  sketching matrix  $\mathbf{S}$  consist of rows  $\{\mathbf{s}_i^T\}_{i=1}^m$  drawn IID as random samples from  $Normal(\mathbf{0}, \mathbf{I}_{m \times m})$  for some  $m < n$ . Then we transform our scaled design matrix and response vectors and define

$$\hat{\mathbf{X}} = \mathbf{S}\mathbf{X}, \quad \hat{\mathbf{y}} = \mathbf{S}\mathbf{y}. \quad (4.279)$$

Let  $\hat{x}_{kj}$  be the elements of  $\hat{\mathbf{X}}$  and  $\hat{y}_k$  be the elements of  $\hat{\mathbf{y}}$ , noting that  $1 \leq k \leq m$  and  $1 \leq j \leq d$ .

Using the definition of a 1-bit quantizer from section 4.2.1, we will define three element-wise quantizers:

1.  $Q_{\hat{\mathbf{X}}}(\hat{x}_{kj})$  is a quantizer defined on the high-probability range of  $\hat{x}_{kj}$ , which will be defined in Sections 4.4.2 and 4.4.3. We allow the notation  $\tilde{\mathbf{X}}$  to be the matrix of quantized elements from  $\hat{\mathbf{X}}$  and  $\tilde{\mathbf{x}}_k^T$  to be the  $k$ th row of  $\tilde{\mathbf{X}}$ . We define

$$\tilde{x}_{kj} := Q_{\hat{\mathbf{X}}}(\hat{x}_{kj}). \quad (4.280)$$

2.  $Q_{\hat{\mathbf{X}}^2}(\hat{x}_{kj}^2)$  is a quantizer defined on the high-probability range of squared elements of the matrix  $\hat{\mathbf{X}}$ , which will be defined in future sections. We define

$$\tilde{x}_{kj}^2 := Q_{\hat{\mathbf{X}}^2}(\hat{x}_{kj}^2). \quad (4.281)$$

3.  $Q_{\hat{\mathbf{y}}}(\hat{y}_k)$  is a quantizer defined on the high-probability range of the elements of  $\hat{\mathbf{y}}$ . We allow the notation  $\tilde{\mathbf{y}}$  to be the vector of the quantized elements of  $\hat{\mathbf{y}}$ . We define

$$\tilde{y}_k := Q_{\hat{\mathbf{y}}}(\hat{y}_k). \quad (4.282)$$

The endpoints of the quantizers will be established in section 4.4.2 and 4.4.3, as they depend on the assumptions made on  $\mathbf{Z}$  which will vary by section.

We separate this chapter into two major sections. Section 4.4.2 assumes  $\mathbf{Z}$  to be fixed, while 4.4.3 assumes a standard Gaussian distribution on  $\mathbf{Z}$ . The results will be summarized in section 4.4.4.

#### 4.4.2 The Sketched and Quantized Scenario with Fixed $\mathbf{Z}$

Let us assume  $\mathbf{Z}$  is fixed and that the elements of  $\mathbf{Z}$  are bounded, that is  $|z_{ij}| \leq R$  for all  $i, j$  for some  $R \in \mathbb{R}^+$ . We will thus use the definitions

$$\boldsymbol{\Sigma} := \mathbf{X}^T \mathbf{X} = \frac{\mathbf{Z}^T \mathbf{Z}}{n} \quad (4.283)$$

$$\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}} := \frac{\mathbf{Z}^T \mathbf{w}}{n}, \quad (4.284)$$

and we define sketched and quantized estimators of  $\boldsymbol{\Sigma}$  and  $\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}}$  as

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{m} \sum_{k=1}^m \tilde{\boldsymbol{\Sigma}}_k = \frac{1}{m} \sum_{k=1}^m \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T + \tilde{\boldsymbol{\Delta}}_k \quad (4.285)$$

$$\tilde{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} = \frac{1}{m} \tilde{\mathbf{X}}^T \tilde{\mathbf{y}} = \frac{1}{m} \sum_{k=1}^m \tilde{x}_{kj} \tilde{y}_k \quad \text{for } 1 \leq j \leq d \quad (4.286)$$

where we define  $\tilde{\boldsymbol{\Sigma}}_k = \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T + \tilde{\boldsymbol{\Delta}}_k$  and  $\tilde{\boldsymbol{\Delta}}_k = \text{diag} \left( \tilde{x}_{kj}^2 - \tilde{x}_{kj}^2 \right)_{j=1}^d$ . We note that, as in other chapters, we define the estimators in such a ways as to allow for their unbiasedness, that is  $\mathbf{E} \left[ \tilde{\boldsymbol{\Sigma}} \right] = \boldsymbol{\Sigma}$  and  $\mathbf{E} \left[ \tilde{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} \right] = \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}}$ . With these definitions, we prove the theorem

**Theorem 4.4.1** (High Probability Error Bound for Sketched and Quantized Parameters with Fixed  $\mathbf{Z}$ ). *Let  $\mathbf{Z}$  be a  $n \times d$  matrix whose elements satisfy  $|z_{ij}| \leq R$  for all  $i, j$  for some  $R > 0$ . Define a scaled design matrix  $\mathbf{X} = \frac{\mathbf{Z}}{\sqrt{n}}$  and scaled response vector  $\mathbf{y} = \frac{\mathbf{w}}{\sqrt{n}}$  such that  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^0 + \sigma^2 \frac{\boldsymbol{\epsilon}}{\sqrt{n}}$  with  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}_n)$ .*

*Then given a  $m \times n$  matrix  $\mathbf{S}$  for some  $m < n$  whose rows  $\{\mathbf{s}_i^T\}_{i=1}^m$  are drawn IID as random samples from  $\text{Normal}(\mathbf{0}, \mathbf{I}_{m \times m})$ , define the sketched design matrix  $\hat{\mathbf{X}} = \mathbf{S}\mathbf{X}$  and sketched response vector  $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ . Let  $\hat{x}_{kj}$  be the elements of  $\hat{\mathbf{X}}$  and  $\hat{y}_k$  be the elements of  $\hat{\mathbf{y}}$  for  $k = 1, \dots, m$  and  $j = 1, \dots, d$ .*

Let  $\tilde{x}_{kj} = Q_{\tilde{\mathbf{X}}}(\hat{x}_{kj})$ ,  $\tilde{x}_{kj}^2 = Q_{\tilde{\mathbf{X}}^2}(\hat{x}_{kj}^2)$ , and  $\tilde{y}_k = Q_{\tilde{\mathbf{Y}}}(\hat{y}_k)$  be quantizers defined on the intervals  $[-\hat{R}, \hat{R}]$ ,  $[0, \hat{R}^2]$ , and  $[-\hat{L}, \hat{L}]$ , respectively, where  $|\tilde{x}_{kj}| \leq \hat{R}$  and  $|\tilde{y}_k| \leq \hat{L}$  with high-probability where

$$\hat{R} := \sqrt{2R^2 \log(2nmd)}, \quad \hat{L} := \sqrt{2\ell^2 \log(2nm)},$$

$$\ell = RC\sqrt{\log(n)} \|\boldsymbol{\beta}^0\|_2 + \sigma\sqrt{2\log(2n^2)}, \quad b_0 = 2\hat{R}^2 + d^{-1} \|\boldsymbol{\Sigma}\|.$$

Define unbiased estimators of  $\boldsymbol{\Sigma} := \frac{\mathbf{Z}^T \mathbf{Z}}{n}$  and  $\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{Y}} := \frac{\mathbf{Z}^T \mathbf{w}}{n}$  as

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{m} \sum_{k=1}^m \tilde{\boldsymbol{\Sigma}}_k = \frac{1}{m} \sum_{k=1}^m \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T + \tilde{\boldsymbol{\Delta}}_k$$

$$\tilde{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{Y}} = \frac{1}{m} \tilde{\mathbf{X}}^T \tilde{\mathbf{y}} = \frac{1}{m} \sum_{k=1}^m \tilde{x}_{kj} \tilde{y}_k \quad \text{for } 1 \leq j \leq d.$$

Then for  $\tau_0 = \log^2(2nmd)2R^2 \left(3 + \frac{2}{d} + \frac{\|\boldsymbol{\Sigma}\|}{\log(2nmd)}\right)$ ,  $\eta = \frac{2}{\lambda_{\min}(\boldsymbol{\Sigma})}$ , and provided that

$$m \geq \max \left\{ 2^4 d \log^3(2nd) \eta^2 R^2 \left(6 + \frac{4}{d} + \frac{2 \|\boldsymbol{\Sigma}\|}{\log(2nd)}\right), 2^2 d \log(2nd) b_0 \eta \right\},$$

Then with probability  $1 - 6/n$

$$\begin{aligned} & \|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\|_2 \\ & \leq \frac{4\|\boldsymbol{\beta}^*\|_2 \left( \sqrt{\frac{2d \log(2nd) \tau_0}{m}} + \frac{2d \log(2nd) b_0}{m} \right) + 4\sqrt{\frac{8d \hat{R}^2 \hat{L}^2 \log(2nd)}{m}}}{\lambda_{\min}(\boldsymbol{\Sigma})} + \sqrt{\frac{\sigma^2 d}{\lambda_{\min}(\boldsymbol{\Sigma}) n}} \\ & = \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{m}} \right) + \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{n}} \right). \end{aligned} \tag{4.287}$$

**Proof of 4.4.1**

*Proof.* We wish to bound

$$\left\| \hat{\beta} - \beta^* \right\|_2 \leq \frac{4 \|\beta^*\|_2 \left\| \tilde{\Sigma} - \Sigma \right\| + 4 \left\| (\tilde{\Sigma}_{\mathbf{Z}_w} - \Sigma_{\mathbf{Z}_w}) \right\|_2}{\lambda_{\min}(\Sigma)}$$

by controlling

$$(a): \left\| \tilde{\Sigma} - \Sigma \right\| \quad (b): \left\| \tilde{\Sigma}_{\mathbf{Z}_w} - \Sigma_{\mathbf{Z}_w} \right\|_2.$$

**Analyzing Part (a) in the Sketched and Quantized Scenario with Fixed Data**

We establish in Lemma 12 the following result

$$\frac{1}{m} \left\| \sum_{k=1}^m (\tilde{\Sigma}_k - \Sigma) \right\| = \left\| \tilde{\Sigma} - \Sigma \right\| \leq \sqrt{\frac{2 \log(2nd)\tau}{m}} + \frac{2 \log(2nd)b}{m} \quad (4.288)$$

with probability  $1 - 2/n$  where

$$\tau_0 = \log^2(2nmd)2R^2 \left( 3 + \frac{2}{d} + \frac{\|\Sigma\|}{\log(2nmd)} \right) \quad \text{and} \quad b_0 = 2\hat{R}^2 + d^{-1} \|\Sigma\|.$$

Thus, we can say for part (a) of problem (4.19)

$$\left\| \tilde{\Sigma} - \Sigma \right\| \leq \sqrt{\frac{2d \log(2nd)\tau_0}{m}} + \frac{2d \log(2nd)b_0}{m} \quad (4.289)$$

with probability at least  $1 - 2/n$  once we ensure the lambda min requirement is met. To do so, we must find a value for  $m$  such that

$$\sqrt{\frac{2d \log(2nd)\tau_0}{m}} + \frac{2d \log(2nd)b_0}{m} \leq \frac{\lambda_{\min}(\Sigma)}{2}. \quad (4.290)$$

It is sufficient to show that each term is less than  $\frac{\lambda_{\min}(\Sigma)}{4}$ . We examine the first term:

$$\begin{aligned}
 \sqrt{\frac{2d \log(2nd)\tau_0}{m}} &\leq \frac{\lambda_{\min}(\Sigma)}{4} \\
 \implies \sqrt{\frac{2 \log(2nd)d \log^2(2nmd)2R^2 \left(3 + \frac{2}{d} + \frac{\|\Sigma\|}{\log(2nmd)}\right)}{m}} &\leq \frac{\lambda_{\min}(\Sigma)}{4} \\
 \implies \sqrt{\frac{2 \log(2nd)d \log^2(2nd)2R^2 \left(3 + \frac{2}{d} + \frac{\|\Sigma\|}{\log(2nd)}\right)}{m}} &\leq \frac{\lambda_{\min}(\Sigma)}{4} \\
 \implies \frac{4d \log^3(2nd)R^2 \left(3 + \frac{2}{d} + \frac{\|\Sigma\|}{\log(2nd)}\right)}{m} &\leq \frac{\lambda_{\min}^2(\Sigma)}{2^4} \\
 \implies m &\geq \frac{2^6 d \log^3(2nd)R^2 \left(3 + \frac{2}{d} + \frac{\|\Sigma\|}{\log(2nd)}\right)}{\lambda_{\min}^2(\Sigma)} \\
 \implies m &\geq 2^4 d \log^3(2nd)\eta^2 R^2 \left(6 + \frac{4}{d} + \frac{2\|\Sigma\|}{\log(2nd)}\right) \tag{4.291}
 \end{aligned}$$

where

$$\eta = \frac{2}{\lambda_{\min}(\Sigma)}. \tag{4.292}$$

Examining the second term:

$$\begin{aligned}
 \frac{2d \log(2nd)b_0}{m} &\leq \frac{\lambda_{\min}(\Sigma)}{4} \\
 \implies m &\geq 2^2 d \log(2nd)b_0\eta. \tag{4.293}
 \end{aligned}$$

Selecting  $m$  such that

$$m \geq \max \left\{ 2^4 d \log^3(2nd)\eta^2 R^2 \left(6 + \frac{4}{d} + \frac{2\|\Sigma\|}{\log(2nd)}\right), 2^2 d \log(2nd)b_0\eta \right\} \tag{4.294}$$

ensures the condition is met.

### Analyzing Part (b) in the Sketched and Quantized Scenario with Fixed Data

We now upper bound

$$\left\| \tilde{\Sigma}_{\mathbf{X}\mathbf{Y}} - \Sigma_{\mathbf{X}\mathbf{Y}} \right\|_2$$

recalling that we have defined  $\tilde{\Sigma}_{\mathbf{X}\mathbf{Y}}$  to be

$$\tilde{\Sigma}_{\mathbf{X}\mathbf{Y}} = \frac{1}{m} \tilde{\mathbf{X}}^T \tilde{\mathbf{y}} = \frac{1}{m} \sum_{k=1}^m \tilde{x}_{kj} \tilde{y}_k \quad \text{for } 1 \leq j \leq d. \quad (4.295)$$

Let us define

$$\psi_j = \sum_{k=1}^m \tilde{x}_{jk} \tilde{y}_k - \sum_{i=1}^n x_{ji} y_i \quad (4.296)$$

so that

$$\frac{1}{m} \psi_j = \left( \tilde{\Sigma}_{\mathbf{X}\mathbf{Y}} - \Sigma_{\mathbf{X}\mathbf{Y}} \right)_j = \frac{1}{m} \sum_{k=1}^m \tilde{x}_{jk} \tilde{y}_k - \frac{1}{m} \sum_{i=1}^n x_{ji} y_i \quad (4.297)$$

is the  $j$ th term of the difference between the sketched and quantized estimator and  $\Sigma_{\mathbf{X}\mathbf{Y}}$ .

Let us define the event

$$E : \left( \max_k |\tilde{y}_k| \leq \hat{L} \right) \text{ AND } \left( \max_{kj} |\hat{x}_{kj}| \leq \hat{R} \right) \quad (4.298)$$

We then recognize that  $\psi_j$  conditioned on  $E$  is a sum of zero-mean, bounded random variables in the region  $[-2\hat{R}\hat{L}, 2\hat{R}\hat{L}]$ . Thus, we can apply a Hoeffding bound and say

$$\mathbb{P} \left( |\psi_j| \geq t \mid E \right) \leq 2 \exp \left( -\frac{t^2}{8m\hat{R}^2\hat{L}^2} \right) \quad (4.299)$$

for all  $t > 0$  (see Wainwright Exercise 2.4, Example 2.4, and Equation 2.11)[69]. Using a union bound argument and letting  $t = \sqrt{8m\hat{R}^2\hat{L}^2 \log(2nd)}$  we have

$$\begin{aligned} \mathbb{P}\left(\max_j |\psi_j| \geq \sqrt{8m\hat{R}^2\hat{L}^2 \log(2nd)} \mid E\right) &= \mathbb{P}\left(\max_j \left(\tilde{\Sigma}_{\mathbf{X}_Y} - \Sigma_{\mathbf{X}_Y}\right)_j \geq \sqrt{\frac{8\hat{R}^2\hat{L}^2 \log(2nd)}{m}} \mid E\right) \\ &\leq 2d \exp\left(-\frac{t^2}{8m\hat{R}^2\hat{L}^2}\right) = \frac{1}{n}. \end{aligned} \quad (4.300)$$

In section 4.4.2 we established that  $\max_k |\hat{y}_k| \leq \hat{L}$  with probability  $1 - 2/n$  and that  $\max_{kj} |\hat{x}_{kj}| \leq \hat{R}$  with probability  $1 - 1/n$ . Thus, the union of the two events, which is  $E$ , occurs with probability  $1 - (2/n + 1/n)$ . So we can say

$$\begin{aligned} &\mathbb{P}\left(\max_j \left(\tilde{\Sigma}_{\mathbf{X}_Y} - \Sigma_{\mathbf{X}_Y}\right)_j \geq t\right) \\ &\leq \mathbb{P}\left(\max_j \left(\tilde{\Sigma}_{\mathbf{X}_Y} - \Sigma_{\mathbf{X}_Y}\right)_j \geq t \mid \left(\max_k |\tilde{y}_k| \leq \hat{L}\right) \text{ AND } \left(\max_{kj} |\hat{x}_{kj}| \leq \hat{R}\right)\right) \\ &+ \mathbb{P}\left(\left(\max_k |\tilde{y}_k| \geq \hat{L}\right) \text{ OR } \left(\max_{kj} |\hat{x}_{kj}| \geq \hat{R}\right)\right) = \frac{1}{n} + \frac{2}{n} + \frac{1}{n} \\ &= \frac{4}{n} \end{aligned} \quad (4.301)$$

Thus, we have bounded our desired quantity:

$$\left\|\tilde{\Sigma}_{\mathbf{X}_Y} - \Sigma_{\mathbf{X}_Y}\right\|_2 \leq \sqrt{d} \left\|\tilde{\Sigma}_{\mathbf{X}_Y} - \Sigma_{\mathbf{X}_Y}\right\|_\infty \leq \sqrt{\frac{8d\hat{R}^2\hat{L}^2 \log(2nd)}{m}} \quad (4.302)$$

with probability at least  $1 - 4/n$ .

Combining (4.289) and (4.302) results into problem (4.19)

$$\begin{aligned} \left\| \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} \right\|_2 &\leq \frac{4 \left\| \boldsymbol{\beta}^* \right\|_2 \left\| \hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \right\| + 4 \left\| (\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}}) \right\|_2}{\lambda_{\min}(\boldsymbol{\Sigma})} \\ &\leq \frac{4 \left\| \boldsymbol{\beta}^* \right\|_2 \left( \sqrt{\frac{2d \log(2nd)\tau_0}{m}} + \frac{2d \log(2nd)b_0}{m} \right) + 4 \sqrt{\frac{8d\hat{R}^2\hat{L}^2 \log(2nd)}{m}}}{\lambda_{\min}(\boldsymbol{\Sigma})} \end{aligned} \quad (4.303)$$

$$= \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{m}} \right) \quad (4.304)$$

with probability at least  $1 - (2/n + 4/n)$ . We now add on the additional error from  $\left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \right\|$  from the inequality

$$\left\| \boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}} \right\| \leq \left\| \boldsymbol{\beta}^0 - \boldsymbol{\beta}^* \right\| + \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \right\|. \quad (4.305)$$

Then for some  $\delta > 0$ ,

$$\begin{aligned} \left\| \boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}} \right\|_2 &\leq \frac{4 \left\| \boldsymbol{\beta}^* \right\|_2 \left( \sqrt{\frac{2d \log(2nd)\tau_0}{m}} + \frac{2d \log(2nd)b_0}{m} \right) + 4 \sqrt{\frac{8d\hat{R}^2\hat{L}^2 \log(2nd)}{m}}}{\lambda_{\min}(\boldsymbol{\Sigma})} + \sqrt{\frac{\sigma^2 d}{\lambda_{\min}(\boldsymbol{\Sigma}) n}} \\ &= \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{m}} \right) + \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{n}} \right) \end{aligned} \quad (4.306)$$

with probability  $1 - 6/n$ . □

The subsequent sections provide the lemmas and proofs supporting this theorem.

### Establishing a High-Probability Bound on the Sketched Data with Fixed $\mathbf{Z}$

As a requisite to defining the bounds of our quantizers, we must establish high-probability bounds on the sketched data.

**Lemma 11.** *Let  $\mathbf{Z}$  be a  $n \times d$  matrix whose elements satisfy  $|z_{ij}| \leq R$  for all  $i, j$  for some  $R > 0$ . Define a scaled design matrix  $\mathbf{X} = \frac{\mathbf{Z}}{\sqrt{n}}$  and scaled response vector  $\mathbf{y} = \frac{\mathbf{w}}{\sqrt{n}}$  such that  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^0 + \sigma^2 \frac{\boldsymbol{\epsilon}}{\sqrt{n}}$  with  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}_n)$ .*

*Then given a  $m \times n$  matrix  $\mathbf{S}$  for some  $m < n$  whose rows  $\{\mathbf{s}_{i'}^T\}_{i'=1}^m$  are drawn IID as random samples from  $\text{Normal}(\mathbf{0}, \mathbf{I}_{m \times m})$ , define the sketched design matrix  $\hat{\mathbf{X}} = \mathbf{S}\mathbf{X}$  and sketched response vector  $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ . Let  $\hat{x}_{kj}$  be the elements of  $\hat{\mathbf{X}}$  and  $\hat{y}_k$  be the elements of  $\hat{\mathbf{y}}$  for  $k = 1, \dots, m$  and  $j = 1, \dots, d$ . Then*

$$|\hat{x}_{kj}| \leq \sqrt{2R^2 \log(2nmd)}$$

*for all  $k, j$  with probability  $1 - 1/n$ , and*

$$|\hat{y}_k| \leq \sqrt{2\ell^2 \log(2nm)}$$

*for all  $k$  with probability  $1 - 2/n$  where  $\ell = RC\sqrt{\log(n)} \|\boldsymbol{\beta}^0\|_2 + \sigma\sqrt{2\log(2n^2)}$  is a high-probability bound on  $\max_i |y_i|$  from Lemma 2.*

Lemma 11 allows us to define

$$\hat{R} := \sqrt{2R^2 \log(2nmd)} \tag{4.307}$$

and

$$\hat{L} := \sqrt{2\ell^2 \log(2nm)} \tag{4.308}$$

as high-probability bounds of  $\hat{x}_{kj}$  and  $\hat{y}_k$  for all  $k, j$ .

### Proof of Lemma 11

**Establishing a High-Probability Bound on the Elements of  $\hat{\mathbf{X}}$**  We first examine the elements of  $\hat{\mathbf{X}}$ , that is,  $\hat{x}_{kj} = \sum_{i=1}^n x_{ij}s_{ki}$ . Since  $s_{ki} \sim N(0, 1)$  for all  $k, i$ , then we know that  $\hat{x}_{kj} \sim N\left(0, \|\hat{x}_{\cdot j}\|_2^2\right)$  where  $\hat{x}_{\cdot j}$  is the  $j$ th column of  $\mathbf{X}$ . This also means that

$$\mathbf{Var} [\hat{x}_{kj}] = \|\hat{x}_{\cdot j}\|_2^2 = \left\| \frac{\mathbf{z}_{\cdot j}}{\sqrt{n}} \right\|_2^2 = \frac{1}{n} \|\mathbf{z}_{\cdot j}\|_2^2 \leq R^2 \quad (4.309)$$

where  $\mathbf{z}_{\cdot j}$  is the  $j$ th row of  $\mathbf{Z}$ . We apply a Gaussian tail bound and state that

$$\mathbb{P} (|\hat{x}_{kj}| \geq t) \leq 2 \exp\left(\frac{-t^2}{2\sigma_{\hat{\mathbf{X}}}^2}\right) \leq 2 \exp\left(\frac{-t^2}{2R^2}\right) \quad (4.310)$$

for all  $t \geq 0$  where  $\sigma_{\hat{\mathbf{X}}}^2 = \mathbf{Var} [\hat{x}_{kj}]$ . Letting  $t = \sqrt{2R^2 \log(2nmd)}$  and using a union bound argument we can say

$$\mathbb{P} \left( \max_{k,j} |\hat{x}_{kj}| \geq t \right) \leq 2md \exp\left(\frac{-t^2}{2R^2}\right) = \frac{1}{n} \quad (4.311)$$

Then

$$|\hat{x}_{kj}| \leq \sqrt{2R^2 \log(2nmd)} \quad (4.312)$$

for all  $k, j$  with probability  $1 - 1/n$ .

**Establishing a High-Probability Bound on the Elements of  $\hat{\mathbf{y}}$**  Now let us examine  $\hat{y}_k = \sum_{i=1}^n s_{ki}y_i$ . We define the event

$$E : \max_i |w_i| \leq \ell \quad \text{for} \quad \ell = RC\sqrt{\log(n)} \|\boldsymbol{\beta}^0\|_2 + \sigma\sqrt{2\log(2n^2)} \quad (4.313)$$

which we showed in section 4.2.2 occurs with probability  $1 - 1/n$ . Examining the conditioned variance we see

$$\mathbf{Var} [\hat{y}_k | E] = \mathbf{E} \left[ \mathbf{y}^T \mathbf{s}_k \mathbf{s}_k^T \mathbf{y} \mid E \right] = \mathbf{y}^T \mathbf{y} = \frac{1}{n} \|\mathbf{w}\|_2^2 \leq \ell^2. \quad (4.314)$$

We can again apply a Gaussian tail bound and say

$$\mathbb{P} (|\hat{y}_k| \geq t \mid E) \leq 2 \exp \left( \frac{-t^2}{2\sigma_{\hat{y}}^2} \right) \leq 2 \exp \left( \frac{-t^2}{2\ell^2} \right) \quad (4.315)$$

where  $\sigma_{\hat{y}}^2 = \mathbf{Var} [\hat{y}_k | E]$ . Using a union bound and letting  $t = \sqrt{2\ell^2 \log(2nm)}$  results in

$$\mathbb{P} \left( \max_k |\hat{y}_k| \geq \sqrt{2\ell^2 \log(2nm)} \mid E \right) \leq 2m \exp \left( \frac{-t^2}{2\ell^2} \right) \leq \frac{1}{n} \quad (4.316)$$

Now using the law total of probability

$$\mathbb{P} \left( \max_k |\hat{y}_k| \geq t \right) \leq \mathbb{P} \left( |\hat{y}_k| \geq t \mid \max_i |w_i| \leq \ell \right) + \mathbb{P} \left( \max_i |w_i| \geq \ell \right) = \frac{1}{n} + \frac{1}{n} = \frac{2}{n} \quad (4.317)$$

for  $t = \sqrt{2\ell^2 \log(2nm)}$  and  $\ell = RC\sqrt{\log(n)} \|\beta^0\|_2 + \sigma\sqrt{2\log(2n^2)}$  for some constant  $C > 0$ .

### Establishing the Bounds of the Quantizers in the Sketched and Quantized Scenario with Fixed $\mathbf{Z}$

We established in section 4.4.2 that

$$|\hat{x}_{jk}| \leq \hat{R} = \sqrt{2R^2 \log(2nmd)} \quad \text{w.p. } 1 - 1/n \quad (4.318)$$

$$|\hat{y}_k| \leq \hat{L} = \sqrt{2\ell^2 \log(2nm)} \quad \text{w.p. } 1 - 2/n. \quad (4.319)$$

for all  $j, k$  for  $\ell = RC\sqrt{\log(n)}\|\beta^0\|_2 + \sigma\sqrt{2\log(2n^2)}$  for some constant  $C > 0$ . Using these results, we establish our bounds for our three quantizers:

	$\alpha^-$	$\alpha^+$	$\Delta$
$Q_{\hat{\mathbf{x}}}$	$-\hat{R}$	$\hat{R}$	$2\hat{R}$
$Q_{\hat{\mathbf{x}}^2}$	0	$\hat{R}^2$	$\hat{R}^2$
$Q_{\hat{\mathbf{y}}}$	$-\hat{L}$	$\hat{L}$	$2\hat{L}$

### Matrix Bernstein Inequality for the Sketched and Quantized Scenario with Fixed $\mathbf{Z}$

**Lemma 12.** *Let  $\mathbf{Z}$  be a  $n \times d$  matrix whose elements satisfy  $|z_{ij}| \leq R$  for all  $i, j$  for some  $R > 0$ . Define a scaled design matrix  $\mathbf{X} = \frac{\mathbf{Z}}{\sqrt{n}}$  and scaled response vector  $\mathbf{y} = \frac{\mathbf{w}}{\sqrt{n}}$  such that  $\mathbf{y} = \mathbf{X}\beta^0 + \sigma\frac{\epsilon}{\sqrt{n}}$  with  $\epsilon \sim N(\mathbf{0}, \mathbf{I}_n)$ .*

*Then given a  $m \times n$  matrix  $\mathbf{S}$  for some  $m < n$  whose rows  $\{\mathbf{s}_i^T\}_{i=1}^m$  are drawn IID as random samples from  $\text{Normal}(\mathbf{0}, \mathbf{I}_{m \times m})$ , define the sketched design matrix  $\hat{\mathbf{X}} = \mathbf{S}\mathbf{X}$  and sketched response vector  $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ . Let  $\hat{x}_{kj}$  be the elements of  $\hat{\mathbf{X}}$  and  $\hat{y}_k$  be the elements of  $\hat{\mathbf{y}}$  for  $k = 1, \dots, m$  and  $j = 1, \dots, d$ .*

*Let  $\tilde{x}_{kj} = Q_{\hat{\mathbf{X}}}(\hat{x}_{kj})$ ,  $\tilde{x}_{kj}^2 = Q_{\hat{\mathbf{X}}^2}(\hat{x}_{kj}^2)$ , and  $\tilde{y}_k = Q_{\hat{\mathbf{y}}}(\hat{y}_k)$  be quantizers defined on the intervals  $[-\hat{R}, \hat{R}]$ ,  $[0, \hat{R}^2]$ , and  $[-\hat{L}, \hat{L}]$ , respectively, where  $|\tilde{x}_{kj}| \leq \hat{R}$  and  $|\tilde{y}_k| \leq \hat{L}$  with high-probability where*

$$\hat{R} := \sqrt{2R^2 \log(2nmd)}, \quad \hat{L} := \sqrt{2\ell^2 \log(2nm)}, \quad \ell = RC\sqrt{\log(n)}\|\beta^0\|_2 + \sigma\sqrt{2\log(2n^2)}.$$

Define unbiased estimators of  $\Sigma := \frac{\mathbf{Z}^T \mathbf{Z}}{n}$  and  $\Sigma_{\mathbf{X}\mathbf{y}} := \frac{\mathbf{Z}^T \mathbf{w}}{n}$  as

$$\begin{aligned}\tilde{\Sigma} &= \frac{1}{m} \sum_{k=1}^m \tilde{\Sigma}_k = \frac{1}{m} \sum_{k=1}^m \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T + \tilde{\Delta}_k \\ \tilde{\Sigma}_{\mathbf{X}\mathbf{y}} &= \frac{1}{m} \tilde{\mathbf{X}}^T \tilde{\mathbf{y}} = \frac{1}{m} \sum_{k=1}^m \tilde{x}_{kj} \tilde{y}_k \quad \text{for } 1 \leq j \leq d.\end{aligned}$$

For

$$\tau_0 = \log^2(2nmd)2R^2 \left( 3 + \frac{2}{d} + \frac{\|\Sigma\|}{\log(2nmd)} \right), \quad \text{and} \quad b_0 = 2\hat{R}^2 + d^{-1} \|\Sigma\|,$$

then with probability  $1 - 2/n$

$$\|\tilde{\Sigma} - \Sigma\| \leq \sqrt{\frac{2d \log(2nd)\tau_0}{m}} + \frac{2d \log(2nd)b_0}{m}.$$

### Proof of Lemma 12

*Proof.* Let us begin by defining the event

$$E : \max_{j,k} |\hat{x}_{jk}| \leq \hat{R} \tag{4.320}$$

which we showed is true with probability  $1 - 1/n$  in section 4.4.2.

Conditioned on this event, we establish a few helpful facts. First, we note that  $\tilde{x}_{kj}^2 = \hat{R}^2$  for all  $k, j$ . Second,  $\tilde{x}_{kj}^2 \in \{0, \hat{R}^2\}$ . Thus, we can see that  $\tilde{x}_{kj}^2 - \hat{x}_{kj}^2 \leq 0$  for all  $k, j$  resulting in  $\tilde{\Delta}_k = \text{diag} \left( \tilde{x}_{kj}^2 - \hat{x}_{kj}^2 \right)_{j=1}^d$  to consist of all non-positive values.

Recall that we defined

$$\tilde{\Sigma}_k = \tilde{x}_k \tilde{x}_k^T + \tilde{\Delta}_k \tag{4.321}$$

We now apply the Matrix Bernstein inequality (Theorem 6.17 in Wainwright)[69]. We use the fact that  $\mathbf{Var} [\tilde{\Sigma}_k - \Sigma] = \mathbf{Var} [\tilde{\Sigma}_k]$ , and apply a Bernstein result to the zero-mean matrices  $\{\tilde{\Sigma}_k - \Sigma\}_{k=1}^m$ . Then we can say

$$\mathbb{P} \left( \frac{1}{m} \left\| \sum_{k=1}^m (\tilde{\Sigma}_k - \Sigma) \right\| \geq \delta \mid E \right) \leq 2 \operatorname{rank} \left( \sum_{k=1}^m \mathbf{Var} [\tilde{\Sigma}_k \mid E] \right) \exp \left( \frac{-m\delta^2}{2(\sigma_{\tilde{\Sigma}}^2 + b\delta)} \right) \quad (4.322)$$

for some  $b \geq \left\| \tilde{\Sigma}_k - \Sigma \right\|$  and  $\delta \geq 0$  and where

$$\sigma_{\tilde{\Sigma}}^2 = \frac{1}{m} \left\| \sum_{k=1}^m \mathbf{Var} [\tilde{\Sigma}_k \mid E] \right\| = \frac{1}{m} \left\| m \mathbf{Var} [\tilde{\Sigma}_k \mid E] \right\| = \left\| \mathbf{Var} [\tilde{\Sigma}_k \mid E] \right\|. \quad (4.323)$$

Using arguments similar to those made in section 4.2.2 and 4.3.3, choosing  $b \geq 2d\hat{R}^2 + \|\Sigma\|$  satisfies the Bernstein condition.

We then proceed by bounding the spectral norm of the variance of  $\tilde{\Sigma}_k$  conditioned on  $E$ :

$$\begin{aligned} \sigma_{\tilde{\Sigma}}^2 &= \left\| \mathbf{Var} [\tilde{\Sigma}_k \mid E] \right\| = \left\| d\hat{R}^2 \mathbf{E} [\tilde{x}_k \tilde{x}_k^T \mid E] + \mathbf{E} [\tilde{\Delta}_k^2 \mid E] + 2\mathbf{E} [\tilde{\Delta}_k \tilde{x}_k \tilde{x}_k^T \mid E] - \Sigma^2 \right\| \\ &\leq \left\| d\hat{R}^2 \mathbf{E} [\tilde{x}_k \tilde{x}_k^T \mid E] + \mathbf{E} [\tilde{\Delta}_k^2 \mid E] + 2\mathbf{E} [\tilde{\Delta}_k \tilde{x}_k \tilde{x}_k^T \mid E] \right\| \\ &\leq d\hat{R}^2 \left\| \mathbf{E} [\tilde{x}_k \tilde{x}_k^T \mid E] \right\| + \left\| \mathbf{E} [\tilde{\Delta}_k^2 \mid E] \right\| + \left\| 2\mathbf{E} [\tilde{\Delta}_k \tilde{x}_k \tilde{x}_k^T \mid E] \right\| \\ &\leq d\hat{R}^2 (\hat{R}^2 + \|\Sigma\|) + 2\hat{R}^4 + 2\mathbf{E} \left[ \left\| \tilde{x}_k \tilde{x}_k^T \right\| \left\| \tilde{\Delta}_k \right\| \mid E \right] \\ &\leq d\hat{R}^2 \|\Sigma\| + d\hat{R}^4 + 2\hat{R}^4 + 2d\hat{R}^4 \\ &= d\hat{R}^2 \|\Sigma\| + (3d + 2)\hat{R}^4. \end{aligned} \quad (4.324)$$

Here we used Jensen's inequality and the convexity of the spectral norm. Additionally, we used the facts

1. Since  $\mathbf{Var} \left[ \widetilde{\boldsymbol{\Sigma}}_k \right] \geq 0$  and  $\boldsymbol{\Sigma} \geq 0$  we applied (D.1) to drop  $\boldsymbol{\Sigma}$  from the second step.

2. We can bound  $\left\| \mathbf{E} \left[ \widetilde{\boldsymbol{x}}_k \widetilde{\boldsymbol{x}}_k^T \mid E \right] \right\|$  by rewriting  $\mathbf{E} \left[ \widetilde{\boldsymbol{x}}_k \widetilde{\boldsymbol{x}}_k^T \mid E \right]$  as

$$\text{diag} \left( \widehat{R}^2 - \mathbf{E} \left[ \widetilde{x}_{kj}^2 \mid E \right] \right)_{j=1}^d + \mathbf{E} \left[ \begin{array}{cccc} \widetilde{x}_{k1}^2 & \widetilde{x}_{k1}\widetilde{x}_{k2} & \dots & \widetilde{x}_{k1}\widetilde{x}_{kd} \\ \widetilde{x}_{k2}\widetilde{x}_{k1} & \widetilde{x}_{k2}^2 \dots & \widetilde{x}_{k2}\widetilde{x}_{kd} & \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{x}_{kd}\widetilde{x}_{k1} & \widetilde{x}_{kd}\widetilde{x}_{k2} \dots & \widetilde{x}_{kd}^2 & \end{array} \mid E \right] \quad (4.325)$$

by adding and subtracting  $\text{diag} \left( \mathbf{E} \left[ \widetilde{x}_{kj}^2 \mid E \right] \right)_{j=1}^d$ . Then using that

$$\mathbf{E} \left[ \widetilde{x}_{kj}^2 \mid E \right] = \sum_{i=1}^n x_{ij}^2 \quad \text{and} \quad \mathbf{E} \left[ \widetilde{x}_{kj}\widetilde{x}_{kj'} \mid E \right] = \sum_{i=1}^n x_{ij}x_{ij'} \quad (4.326)$$

where  $j \neq j'$  (see appendix A for formulations), we can then say

$$\mathbf{E} \left[ \widetilde{\boldsymbol{x}}_k \widetilde{\boldsymbol{x}}_k^T \mid E \right] = \text{diag} \left( \widehat{R}^2 - \sum_{i=1}^n x_{ij}^2 \right)_{j=1}^d + \boldsymbol{\Sigma}. \quad (4.327)$$

We can then bound its spectral norm using the triangle inequality:

$$\begin{aligned}
 \left\| \mathbf{E} \left[ \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T \mid E \right] \right\| &= \left\| \text{diag} \left( \hat{R}^2 - \sum_{i=1}^n x_{ij}^2 \right)_{j=1}^d + \boldsymbol{\Sigma} \right\| \\
 &\leq \left\| \text{diag} \left( \hat{R}^2 - \sum_{i=1}^n x_{ij}^2 \right)_{j=1}^d \right\| + \|\boldsymbol{\Sigma}\| \\
 &\leq \hat{R}^2 - nR^2 + \|\boldsymbol{\Sigma}\| \\
 &\leq \hat{R}^2 + \|\boldsymbol{\Sigma}\|
 \end{aligned} \tag{4.328}$$

3. Recognizing that all the elements in  $\tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T$  are bounded in absolute value by  $\hat{R}^2$ , then a bound of  $\left\| \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T \right\|$  conditioned on  $E$  is

$$\begin{aligned}
 \left\| \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T \right\|^2 &= \sup_{\|\mathbf{v}\|=1} \mathbf{v}^T \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T \mathbf{v} \\
 &= \sup_{\|\mathbf{v}\|=1} \left| \sum_{j,j'} v_j v_{j'} \left( \sum_{\ell} \left( \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k \right)_{\ell j} \left( \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T \right)_{\ell j'} \right) \right| \\
 &\leq \sup_{\|\mathbf{v}\|=1} \sum_{j,j'} |v_j| |v_{j'}| \left| \sum_{\ell} \left( \tilde{\mathbf{x}}_k^T \tilde{\mathbf{x}}_k \right)_{\ell j} \left( \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T \right)_{\ell j'} \right| \\
 &\leq d \hat{R}^4 \sup_{\|\mathbf{v}\|=1} \|\mathbf{v}\|_1^2 \\
 &= d^2 \hat{R}^4 \\
 \implies \left\| \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T \right\| &\leq d \hat{R}^2
 \end{aligned} \tag{4.329}$$

4. A bound of  $\|\tilde{\Delta}_k\|$  conditioned on  $E$  is

$$\|\tilde{\Delta}_k\| = \left\| \text{diag} \left( \hat{R}^2 - \tilde{x}_{kj}^2 \right)_{j=1}^d \right\| \leq \hat{R}^2 \quad (4.330)$$

since conditioned on  $E$ ,  $\tilde{x}_{kj}^2 \in \{0, \hat{R}^2\}$  and  $\tilde{x}_{kj}^2 = \hat{R}^2$ .

We proceed by defining

$$\begin{aligned} \tau &:= d\hat{R}^2 \|\Sigma\| + (3d+2)\hat{R}^4 \\ &= d \left( \sqrt{2R^2 \log(2nmd)} \right)^2 \|\Sigma\| + (3d+2) \left( \sqrt{2R^2 \log(2nmd)} \right)^4 \\ &= d \left( 2R^2 \log(2nmd) \|\Sigma\| + \left( 3 + \frac{2}{d} \right) \left( 2R^2 \log(2nmd) \right)^2 \right) \\ &= d \log^2(2nmd) 2R^2 \left( 3 + \frac{2}{d} + \frac{\|\Sigma\|}{\log(2nmd)} \right). \end{aligned} \quad (4.331)$$

Now, since  $\sigma_{\Sigma}^2 \leq \tau$ , then we can say

$$\mathbb{P} \left( \frac{1}{m} \left\| \sum_{k=1}^m (\tilde{\Sigma}_k - \Sigma) \right\| \geq \delta \mid E \right) \leq 2d \exp \left( \frac{-m\delta^2}{2(\sigma_{\Sigma}^2 + b\delta)} \right) \leq 2d \exp \left( \frac{-m\delta^2}{2(\tau + b\delta)} \right). \quad (4.332)$$

Let us choose

$$\delta = \sqrt{\frac{2 \log(2nd)\tau}{m} + \frac{2 \log(2nd)b}{m}} \quad (4.333)$$

so that

$$\frac{m\delta^2}{2(\tau + b\delta)} \geq \min \left\{ \frac{m\delta^2}{2\tau}, \frac{m\delta}{2b} \right\} \geq \log(2nd). \quad (4.334)$$

We can now say

$$\mathbb{P} \left( \frac{1}{m} \left\| \sum_{k=1}^m (\tilde{\Sigma}_k - \Sigma) \right\| \geq \delta \mid E \right) \leq 2d \exp \left( \frac{-m\delta^2}{2(\tau + b\delta)} \right) \leq 2d \exp(-\log(2nd)) = \frac{1}{n} \quad (4.335)$$

for  $\delta = \sqrt{\frac{2\log(2nd)\tau}{m} + \frac{2\log(2nd)b}{m}}$ , for  $\tau = d \log^2(2nmd) 2R^2 \left( 3 + \frac{2}{d} + \frac{\|\Sigma\|}{\log(2nmd)} \right)$ , and for any  $b \geq 2d\hat{R}^2 + \|\Sigma\|$ .

Then using the law of total probability

$$\begin{aligned} \mathbb{P} \left( \frac{1}{m} \left\| \sum_{k=1}^m (\tilde{\Sigma}_k - \Sigma) \right\| \geq \delta \right) &\leq \mathbb{P} \left( \frac{1}{m} \left\| \sum_{k=1}^m (\tilde{\Sigma}_k - \Sigma) \right\| \geq \delta \mid E \right) + \mathbb{P} \left( \max_{jk} |\hat{x}_{jk}| \geq \hat{R} \right) \\ &= \frac{1}{n} + \frac{1}{n} \\ &= \frac{2}{n} \end{aligned} \quad (4.336)$$

□

### Summary of the Sketched and Quantized Scenario with Fixed $\mathbf{Z}$

For fixed  $\mathbf{Z}$  with bounded elements  $|z_{ij}| \leq R$  for all  $i, j$  we defined

$$\Sigma = \mathbf{X}^T \mathbf{X} = \frac{\mathbf{Z}^T \mathbf{Z}}{n} \quad \tilde{\Sigma} = \frac{1}{m} \sum_{k=1}^m \tilde{\Sigma}_k \quad (4.337)$$

$$\Sigma_{\mathbf{X}\mathbf{y}} = \mathbf{X}^T \mathbf{y} = \frac{\mathbf{Z}^T \mathbf{w}}{n} \quad \tilde{\Sigma}_{\mathbf{X}\mathbf{y}} = \frac{1}{m} \tilde{\mathbf{X}}^T \tilde{\mathbf{y}} \quad \text{for } 1 \leq j \leq d. \quad (4.338)$$

where we let  $\tilde{\Sigma}_k = \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T + \tilde{\Delta}_k$  and  $\tilde{\Delta}_k = \text{diag} \left( \tilde{x}_{kj}^2 - \tilde{x}_{kj}^2 \right)_{j=1}^d$ . We showed that

$$\left\| \tilde{\Sigma} - \Sigma \right\| \leq \sqrt{\frac{4d \log(2nd) \tau_0}{m}} + \frac{4d \log(2nd) b_0}{m} \quad \text{w.p. } 1 - 3/n \quad (4.339)$$

$$\left\| \tilde{\Sigma}_{\mathbf{X}\mathbf{Y}} - \Sigma_{\mathbf{X}\mathbf{Y}} \right\|_2 \leq \sqrt{\frac{8d \hat{R}^2 \hat{L}^2 \log(2nd)}{m}} \quad \text{w.p. } 1 - 4/n \quad (4.340)$$

for

$$\tau_0 = \log^2(2nmd) 2R^2 \left( 3 + \frac{2}{d} + \frac{\|\Sigma\|}{\log(2nmd)} \right) \quad b_0 = 2\hat{R}^2 + d^{-1} \|\Sigma\|$$

$$\hat{R} = \sqrt{2R^2 \log(2nmd)} \quad \hat{L} = \sqrt{2\ell^2 \log(2nm)}$$

$$\ell = RC \sqrt{\log(n)} \left\| \beta^0 \right\|_2 + \sigma \sqrt{2 \log(2n^2)} \quad \text{for } C > 0.$$

We can summarize the results by stating

$$\left\| \tilde{\Sigma} - \Sigma \right\| = \mathcal{O} \left( \sqrt{\frac{R^2 d \log(nd) \log^2(nmd)}{m}} \right) = \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{m}} \right) \quad (4.341)$$

$$\begin{aligned} \left\| \tilde{\Sigma}_{\mathbf{Z}\mathbf{W}} - \Sigma_{\mathbf{Z}\mathbf{W}} \right\|_2 &= \mathcal{O} \left( \sqrt{\frac{R^2 \ell^2 d \log(nmd) \log(nm) \log(nd)}{m}} \right) \\ &= \mathcal{O} \left( \sqrt{\frac{d \log(nmd) \log(nm) \log(nd) \log(n)}{m}} \right) \\ &= \tilde{\mathcal{O}} \left( \sqrt{\frac{d}{m}} \right) \end{aligned} \quad (4.342)$$

Choosing  $m$  such that

$$m \geq \max \left\{ 2^4 d \log^3(2nd) \eta^2 R^2 \left( 6 + \frac{4}{d} + \frac{2 \|\Sigma\|}{\log(2nd)} \right), 2^2 d \log(2nd) b_0 \eta \right\} \quad (4.343)$$

for  $\eta = \frac{2}{\lambda_{\min}(\boldsymbol{\Sigma})}$  satisfies the lambda-min-requirement and allows us to say

$$\begin{aligned} \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \right\|_2 &\leq \frac{4 \left\| \boldsymbol{\beta}^* \right\|_2 \left\| \hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \right\| + 4 \left\| (\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{Y}}) \right\|_2}{\lambda_{\min}(\boldsymbol{\Sigma})} \\ &\leq \frac{4 \left\| \boldsymbol{\beta}^* \right\|_2 \left( \sqrt{\frac{2d \log(2nd)\tau_0}{m}} + \frac{2d \log(2nd)b_0}{m} \right) + 4 \sqrt{\frac{8d\hat{R}^2\hat{L}^2 \log(2nd)}{m}}}{\lambda_{\min}(\boldsymbol{\Sigma})} \end{aligned} \quad (4.344)$$

with probability at least  $1 - 6/n$ . We now add on the additional error from  $\left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \right\|$  from the inequality

$$\left\| \boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}} \right\| \leq \left\| \boldsymbol{\beta}^0 - \boldsymbol{\beta}^* \right\| + \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \right\|. \quad (4.345)$$

resulting in

$$\begin{aligned} &\left\| \boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}} \right\|_2 \\ &\leq \frac{4 \left\| \boldsymbol{\beta}^* \right\|_2 \left( \sqrt{\frac{2d \log(2nd)\tau_0}{m}} + \frac{2d \log(2nd)b_0}{m} \right) + 4 \sqrt{\frac{8d\hat{R}^2\hat{L}^2 \log(2nd)}{m}}}{\lambda_{\min}(\boldsymbol{\Sigma})} + \sqrt{\frac{\sigma^2 d}{\lambda_{\min}(\boldsymbol{\Sigma}) n}} \\ &= \tilde{O} \left( \sqrt{\frac{d}{m}} \right) + \tilde{O} \left( \sqrt{\frac{d}{n}} \right) \end{aligned} \quad (4.346)$$

with probability  $1 - 6/n$ .

#### 4.4.3 The Sketched and Quantized Scenario with Gaussian $\mathbf{Z}$

Let us assume the rows of  $\mathbf{Z}$ , namely  $\mathbf{z}_i^T$ , are drawn independently from a multivariate normal distribution with mean  $\mathbf{0}$  and covariance  $\mathbf{I}_{d \times d}$ . Furthermore, suppose  $\mathbf{y} = \frac{\mathbf{Z}}{\sqrt{n}} \boldsymbol{\beta}^0 + \sigma \frac{\boldsymbol{\epsilon}}{\sqrt{n}}$  where  $\boldsymbol{\epsilon}$  is a  $n \times 1$  vector whose entries are independent and standard normally distributed.

Then we define:

$$\mathbf{X} := \frac{\mathbf{Z}}{\sqrt{n}}, \quad \mathbf{y} := \frac{\mathbf{w}}{\sqrt{n}} = \mathbf{X}\boldsymbol{\beta}^0 + \sigma \frac{\boldsymbol{\epsilon}}{\sqrt{n}} \quad (4.347)$$

$$\boldsymbol{\Sigma} := \mathbf{E} \left[ (\mathbf{X} - \mathbf{E}[\mathbf{X}])^T (\mathbf{X} - \mathbf{E}[\mathbf{X}]) \right] = \mathbf{E} \left[ \mathbf{X}^T \mathbf{X} \right] = \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{Z}}{n} \right] \quad (4.348)$$

$$\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}} := \mathbf{E} \left[ \mathbf{X}^T \mathbf{y} \right] = \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{w}}{n} \right]. \quad (4.349)$$

We define  $\tilde{x}_{kj}$  and  $\tilde{y}_k$  as in the introduction (Section 4.1). We then define sketched and quantized estimators of  $\boldsymbol{\Sigma}$  and  $\boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}}$  as

$$\tilde{\boldsymbol{\Sigma}} := \frac{1}{m} \sum_{k=1}^m \tilde{\boldsymbol{\Sigma}}_k = \frac{1}{m} \sum_{k=1}^m \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T + \tilde{\boldsymbol{\Delta}}_k \quad (4.350)$$

$$\tilde{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} := \frac{1}{m} \tilde{\mathbf{X}}^T \tilde{\mathbf{y}} = \frac{1}{m} \sum_{k=1}^m \tilde{x}_{kj} \tilde{y}_k \quad \text{for } 1 \leq j \leq d \quad (4.351)$$

where we define  $\tilde{\boldsymbol{\Sigma}}_k := \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T + \tilde{\boldsymbol{\Delta}}_k$  and  $\tilde{\boldsymbol{\Delta}}_k := \text{diag} \left( \tilde{x}_{kj}^2 - \tilde{x}_{kj}^2 \right)_{j=1}^d$ . We note that, as in other chapters, we define the estimators so they are unbiased, that is  $\mathbf{E} \left[ \tilde{\boldsymbol{\Sigma}} \right] = \boldsymbol{\Sigma}$  and  $\mathbf{E} \left[ \tilde{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} \right] = \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}}$ .

**Theorem 4.4.2.** *Let the rows of a  $n \times d$  matrix  $\mathbf{Z}$ , namely  $\mathbf{z}_i^T$ , be drawn independently from a multivariate normal distribution with mean  $\mathbf{0}$  and covariance  $\mathbf{I}_{d \times d}$ . Define a scaled design matrix  $\mathbf{X} = \frac{\mathbf{Z}}{\sqrt{n}}$  and scaled response vector  $\mathbf{y} = \frac{\mathbf{w}}{\sqrt{n}}$  such that  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^0 + \sigma \frac{\boldsymbol{\epsilon}}{\sqrt{n}}$  with  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}_n)$ .*

*Then given a  $m \times n$  matrix  $\mathbf{S}$  for some  $m < n$  whose rows  $\{\mathbf{s}_{i'}^T\}_{i'=1}^m$  are drawn IID as random samples from  $\text{Normal}(\mathbf{0}, \mathbf{I}_{m \times m})$ , define the sketched design matrix  $\hat{\mathbf{X}} = \mathbf{S}\mathbf{X}$  and*

sketched response vector  $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ . Let  $\hat{x}_{kj}$  be the elements of  $\hat{\mathbf{X}}$  and  $\hat{y}_k$  be the elements of  $\hat{\mathbf{y}}$  for  $k = 1, \dots, m$  and  $j = 1, \dots, d$ .

Let  $\tilde{x}_{kj} = Q_{\hat{\mathbf{X}}}(\hat{x}_{kj})$ ,  $\tilde{x}_{kj}^2 = Q_{\hat{\mathbf{X}}^2}(\hat{x}_{kj}^2)$ , and  $\tilde{y}_k = Q_{\hat{\mathbf{y}}}(\hat{y}_k)$  be quantizers defined on the intervals  $[-\hat{R}, \hat{R}]$ ,  $[0, \hat{R}^2]$ , and  $[-\hat{L}, \hat{L}]$ , respectively, where  $|\tilde{x}_{kj}| \leq \hat{R}$  and  $|\tilde{y}_k| \leq \hat{L}$  with high-probability where

$$\hat{R}_g = \left(1 + \sqrt{2 \log(nd)}\right) \sqrt{2 \log(2nmd)}, \quad \hat{L}_g = \sqrt{2c \log(2mn)} \left(1 + \sqrt{\frac{2 \log(n)}{n}}\right),$$

for  $c = \left(\sigma + \|\beta^0\|_2\right)^2$ . Define unbiased estimators of  $\Sigma := \mathbf{E}\left[\frac{\mathbf{Z}^T \mathbf{Z}}{n}\right]$  and  $\Sigma_{\mathbf{X}\mathbf{y}} := \mathbf{E}\left[\frac{\mathbf{Z}^T \mathbf{w}}{n}\right]$

as

$$\begin{aligned} \tilde{\Sigma} &= \frac{1}{m} \sum_{k=1}^m \tilde{\Sigma}_k = \frac{1}{m} \sum_{k=1}^m \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T + \tilde{\Delta}_k \\ \tilde{\Sigma}_{\mathbf{X}\mathbf{y}} &= \frac{1}{m} \tilde{\mathbf{X}}^T \tilde{\mathbf{y}} = \frac{1}{m} \sum_{k=1}^m \tilde{x}_{kj} \tilde{y}_k \quad \text{for } 1 \leq j \leq d. \end{aligned}$$

For

$$\tau_0 = \log^2(2nmd) \left(1 + \sqrt{2 \log(nd)}\right)^4 \left(12 + \frac{2}{\log(2nmd) \left(1 + \sqrt{2 \log(nd)}\right)^2} + \frac{8}{d}\right),$$

$$b_0 = 2\hat{R}_g^2 + d^{-1} \|\Sigma\|, \quad \eta = \frac{2}{\lambda_{\min}(\Sigma)},$$

and choosing  $m$  such that

$$m \geq \max \left\{ 2^4 d \log^3(2nd) \left(1 + \sqrt{2 \log(nd)}\right)^4 \eta^2 \left(6 + \frac{1}{\log(2nd) \left(1 + \sqrt{4 \log(nd)}\right)^2} + \frac{4}{d}\right), 2^2 d \log(2nd) b_0 \eta \right\},$$

then with probability at least  $1 - 10/n$

$$\begin{aligned}
 \|\beta^* - \hat{\beta}\|_2 &\leq \frac{4\|\beta^*\|_2 \|\hat{\Sigma} - \Sigma\| + 4\|(\hat{\Sigma}_{\mathbf{X}\mathbf{Y}} - \Sigma_{\mathbf{X}\mathbf{Y}})\|_2}{\lambda_{\min}(\Sigma)} \\
 &\leq \frac{4\|\beta^*\|_2 \left( \sqrt{\frac{2d \log(2nd)\tau_0}{m}} + \frac{2d \log(2nd)b_0}{m} \right) + 4\sqrt{\frac{8d\hat{R}_g^2 \hat{L}_g^2 \log(2nd)}{m}}}{\lambda_{\min}(\Sigma)} \\
 &= \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{m}}\right).
 \end{aligned}$$

### Proof of Theorem 4.4.2

*Proof.* We wish to bound

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{4\|\beta^*\|_2 \|\tilde{\Sigma} - \Sigma\| + 4\|(\tilde{\Sigma}_{\mathbf{Z}\mathbf{W}} - \Sigma_{\mathbf{Z}\mathbf{W}})\|_2}{\lambda_{\min}(\Sigma)}$$

by controlling

$$\text{(a): } \|\tilde{\Sigma} - \Sigma\| \quad \text{(b): } \|\tilde{\Sigma}_{\mathbf{Z}\mathbf{W}} - \Sigma_{\mathbf{Z}\mathbf{W}}\|_2.$$

**Analyzing Part (a) in the Sketched and Quantized Scenario with Gaussian Assumptions** We use Lemma 12 modified for the Gaussian scenario to say:

$$\frac{1}{m} \left\| \sum_{k=1}^m (\tilde{\Sigma}_k - \Sigma) \right\| = \|\tilde{\Sigma} - \Sigma\| \leq \sqrt{\frac{2d \log(2nd)\tau_0}{m}} + \frac{2d \log(2nd)b_0}{m} \quad (4.352)$$

with probability  $1 - 4/n$  where  $b_0 \geq 2\hat{R}_g^2 + d^{-1} \|\Sigma\|$

$$\tau_0 = \log^2(2nmd) \left(1 + \sqrt{2\log(nd)}\right)^4 \left(12 + \frac{2}{\log(2nmd) \left(1 + \sqrt{2\log(nd)}\right)^2} + \frac{8}{d}\right).$$

This is further explained in a subsection below. To ensure the lambda min requirement is met,  $m$  must be chosen such that

$$\sqrt{\frac{2d\log(2nd)\tau_0}{m}} + \frac{2d\log(2nd)b_0}{m} \leq \frac{\lambda_{\min}(\Sigma)}{2}. \quad (4.353)$$

It is sufficient for each term to be less than  $\frac{\lambda_{\min}(\Sigma)}{4}$ . We examine the first term:

$$\begin{aligned} \sqrt{\frac{2d\log(2nd)\tau_0}{m}} &\leq \frac{\lambda_{\min}(\Sigma)}{4} \implies \\ \sqrt{\frac{2\log(2nd)d\log^2(2nmd) \left(1 + \sqrt{2\log(nd)}\right)^4 \left(12 + \frac{2}{\log(2nmd) \left(1 + \sqrt{2\log(nd)}\right)^2} + \frac{8}{d}\right)}{m}} &\leq \frac{\lambda_{\min}(\Sigma)}{4} \\ \implies \frac{4\log(2nd)d\log^2(2nd) \left(1 + \sqrt{2\log(nd)}\right)^4 \left(12 + \frac{2}{\log(2nd) \left(1 + \sqrt{2\log(nd)}\right)^2} + \frac{8}{d}\right)}{m} &\leq \frac{\lambda_{\min}^2(\Sigma)}{2^4} \\ \implies m \geq \frac{2^6 d \log^3(2nd) \left(1 + \sqrt{2\log(nd)}\right)^4 \left(6 + \frac{1}{\log(2nd) \left(1 + \sqrt{2\log(nd)}\right)^2} + \frac{4}{d}\right)}{\lambda_{\min}^2(\Sigma)} \\ \implies m \geq 2^4 d \log^3(2nd) \left(1 + \sqrt{2\log(nd)}\right)^4 \eta^2 \left(6 + \frac{1}{\log(2nd) \left(1 + \sqrt{4\log(nd)}\right)^2} + \frac{4}{d}\right) \end{aligned} \quad (4.354)$$

where

$$\eta = \frac{2}{\lambda_{\min}(\Sigma)}. \quad (4.355)$$

Examining the second term:

$$\begin{aligned} \frac{2d \log(2nd)b_0}{m} &\leq \frac{\lambda_{\min}(\Sigma)}{4} \\ \implies m &\geq 2^2 d \log(2nd)b_0\eta. \end{aligned} \quad (4.356)$$

Choosing  $m$  to be the maximum value of these two values will ensure the condition is met.

**Analyzing Part (b) in the Sketched and Quantized Scenario with Gaussian Assumptions** We wish to now upper bound

$$\left\| \tilde{\Sigma}_{\mathbf{X}_Y} - \Sigma_{\mathbf{X}_Y} \right\|_2$$

where we have defined  $\tilde{\Sigma}_{\mathbf{X}_Y}$  to be

$$\tilde{\Sigma}_{\mathbf{X}_Y} = \frac{1}{m} \tilde{\mathbf{X}}^T \tilde{\mathbf{y}} = \frac{1}{m} \sum_{k=1}^m \tilde{x}_{kj} \tilde{y}_k \quad \text{for } 1 \leq j \leq d. \quad (4.357)$$

Let us define

$$\psi_j = \sum_{k=1}^m \tilde{x}_{jk} \tilde{y}_k - \sum_{i=1}^n x_{ji} y_i \quad (4.358)$$

such that

$$\frac{1}{m} \psi_j = \left( \tilde{\Sigma}_{\mathbf{X}_Y} - \Sigma_{\mathbf{X}_Y} \right)_j = \frac{1}{m} \sum_{k=1}^m \tilde{x}_{jk} \tilde{y}_k - \frac{1}{m} \sum_{i=1}^n x_{ji} y_i \quad (4.359)$$

is the  $j$ th term of the difference between the sketched and quantized estimator and  $\Sigma_{\mathbf{X}_Y}$ .

With the event

$$E : \left( \max_k |\tilde{y}_k| \leq \hat{L}_g \right) \text{ AND } \left( \max_{kj} |\hat{x}_{kj}| \leq \hat{R}_g \right) \quad (4.360)$$

which we showed with Lemmas 11 and 13 to occur with probability  $1 - 3/n$  and  $1 - 2/n$ , respectively. Then we can directly apply the arguments and conclusions from the proof of Theorem 4.4.1 to say

$$\begin{aligned}
 & \mathbb{P} \left( \max_j \left( \tilde{\Sigma}_{\mathbf{X}\mathbf{y}} - \Sigma_{\mathbf{X}\mathbf{y}} \right)_j \geq t \right) \\
 & \leq \mathbb{P} \left( \max_j \left( \tilde{\Sigma}_{\mathbf{X}\mathbf{y}} - \Sigma_{\mathbf{X}\mathbf{y}} \right)_j \geq t \mid \left( \max_k |\tilde{y}_k| \leq \hat{L}_g \right) \text{ AND } \left( \max_{kj} |\hat{x}_{kj}| \leq \hat{R}_g \right) \right) \\
 & \quad + \mathbb{P} \left( \left( \max_k |\tilde{y}_k| \geq \hat{L}_g \right) \text{ OR } \left( \max_{kj} |\hat{x}_{kj}| \geq \hat{R}_g \right) \right) \\
 & = \frac{1}{n} + \frac{3}{n} + \frac{2}{n} \\
 & = \frac{6}{n} \tag{4.361}
 \end{aligned}$$

for  $t = \sqrt{\frac{8\hat{R}_g^2 \hat{L}_g^2 \log(2nd)}{m}}$  and  $\hat{R}_g$  and  $\hat{L}_g$  as defined in sections in sections 4.3.3 and 4.4.2, respectively. Thus, we have bounded our desired quantity:

$$\left\| \tilde{\Sigma}_{\mathbf{X}\mathbf{y}} - \Sigma_{\mathbf{X}\mathbf{y}} \right\|_2 \leq \sqrt{d} \left\| \tilde{\Sigma}_{\mathbf{X}\mathbf{y}} - \Sigma_{\mathbf{X}\mathbf{y}} \right\|_\infty \leq \sqrt{\frac{8d\hat{R}_g^2 \hat{L}_g^2 \log(2nd)}{m}} \tag{4.362}$$

with probability at least  $1 - 6/n$ .

□

### Establishing a High-Probability Bound on the Sketched Data with Gaussian Z

With Gaussian assumptions on our data, we must establish high-probability upper and lower bounds on the sketched data. In section 4.4.2, we relied heavily on our assumptions that bounded  $|z_{ij}|$  for all  $i, j$ . Since in this section we no longer have these bounds, we will

proceed by using a Gaussian tail bound to establish high-probability bounds on  $|z_{ij}|$  and  $|w_i|$ .

**Establishing a High-Probability Bound on the Elements of  $\hat{\mathbf{X}}$**  We reuse results from Section 4.3.3 to say

$$|\hat{x}_{kj}| \leq \hat{R}_g = \left(1 + \sqrt{2 \log(nd)}\right) \sqrt{2 \log(2nmd)} \quad (4.363)$$

for all  $k, j$  with probability  $1 - 2/n$ .

**Establishing a High-Probability Bound on the Elements of  $\hat{\mathbf{y}}$**

**Lemma 13.** *Let the rows of a  $n \times d$  matrix  $\mathbf{Z}$ , namely  $\mathbf{z}_i^T$ , be drawn independently from a multivariate normal distribution with mean  $\mathbf{0}$  and covariance  $\mathbf{I}_{d \times d}$ . Define a scaled design matrix  $\mathbf{X} = \frac{\mathbf{Z}}{\sqrt{n}}$  and scaled response vector  $\mathbf{y} = \frac{\mathbf{w}}{\sqrt{n}}$  such that  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^0 + \sigma \frac{\boldsymbol{\epsilon}}{\sqrt{n}}$  with  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}_n)$ .*

*Then given a  $m \times n$  matrix  $\mathbf{S}$  for some  $m < n$  whose rows  $\{\mathbf{s}_{i'}^T\}_{i'=1}^m$  are drawn IID as random samples from  $Normal(\mathbf{0}, \mathbf{I}_{m \times m})$ , define the sketched design matrix  $\hat{\mathbf{X}} = \mathbf{S}\mathbf{X}$  and sketched response vector  $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ . Let  $\hat{x}_{kj}$  be the elements of  $\hat{\mathbf{X}}$  and  $\hat{y}_k$  be the elements of  $\hat{\mathbf{y}}$  for  $k = 1, \dots, m$  and  $j = 1, \dots, d$ . Then for  $c = \left(\sigma + \|\boldsymbol{\beta}^0\|_2\right)^2$ ,*

$$|\hat{y}_k| \leq \sqrt{2c \log(2mn)} \left(1 + \sqrt{\frac{2 \log(n)}{n}}\right)$$

*for all  $k$  with probability at least  $1 - 3/n$ .*

We use Lemma 13 to define

$$\hat{L}_g := \sqrt{2c \log(2mn)} \left( 1 + \sqrt{\frac{2 \log(n)}{n}} \right) \quad (4.364)$$

so that we can say  $|\hat{y}_k| \leq \hat{L}_g$  for all  $k$  with probability at least  $1 - 3/n$  for  $c = \left( \sigma + \|\boldsymbol{\beta}^0\|_2 \right)^2$ .

**Proof of Lemma 13**

*Proof.* Recognizing that  $\hat{y}_k = \sum_{i=1}^n s_{ki} y_i$ , we can say that  $\hat{y}_k | \mathbf{y} \sim N \left( 0, \|\mathbf{y}\|_2^2 \right)$ . We showed in section 4.3.3 that

$$\|\mathbf{y}\|_2^2 \leq (1 + \delta)^2 \left( \sigma + \|\boldsymbol{\beta}^0\|_2 \right)^2 = c(1 + \delta)^2 \quad (4.365)$$

with probability at least  $(1 - 1/n)$  where we let  $\delta = \sqrt{\frac{2 \log(n)}{n}}$  and we let  $c = \left( \sigma + \|\boldsymbol{\beta}^0\|_2 \right)^2$ .

Let us define the events

$$A : \max_k |\hat{y}_k| \geq t \quad (4.366)$$

$$E : \|\mathbf{y}\|_2^2 \leq c(1 + \delta)^2 \quad (4.367)$$

for the same  $\delta$  and  $c$ . We can now use event  $E$  to say that

$$\mathbf{Var} [\hat{y}_k | E] = \|\mathbf{y}\|_2^2 \leq c(1 + \delta)^2. \quad (4.368)$$

Then we apply a Gaussian tail bound to our conditioned sketched data and use a union bound to say

$$\mathbb{P}\left(\max_k |\hat{y}_k| \geq t \mid E\right) \leq 2m \exp\left(\frac{-t^2}{2\sigma_{\hat{y}}^2}\right) \leq 2m \exp\left(-\frac{t^2}{2c(1+\delta)^2}\right) \quad (4.369)$$

where  $\mathbf{Var}[\hat{y}_k \mid E] = \sigma_{\hat{y}}^2$ . Letting

$$\begin{aligned} t &= \sqrt{2c(1+\delta)^2 \log(2mn)} \\ &= \sqrt{2c \left(1 + \sqrt{\frac{2\log(n)}{n}}\right)^2 \log(2mn)} \\ &= \sqrt{2c \log(2mn)} \left(1 + \sqrt{\frac{2\log(n)}{n}}\right) \end{aligned} \quad (4.370)$$

then

$$\mathbb{P}\left(\max_k |\hat{y}_k| \geq t \mid E\right) \leq \frac{1}{n}. \quad (4.371)$$

We recognize that the probability of event  $E$  is the union of the probability of two events:

$$B : \left\| \mathbf{X}\beta^0 \right\|_2 \leq \left\| \beta^0 \right\|_2 (1 + \delta) \quad (4.372)$$

$$C : \frac{\|\epsilon\|_2}{\sqrt{n}} \leq 1 + \delta \quad (4.373)$$

which both occur with probability  $1 - 1/n$  from Section 4.3.3. Then we use the law of total probability to say

$$\begin{aligned}
 \mathbb{P}\left(\max_k |\hat{y}_k| \geq t\right) &= \mathbb{P}(A|E)\mathbb{P}(E) + \mathbb{P}(A|E^c)\mathbb{P}(E^c) \\
 &\leq \mathbb{P}(A|E) + \mathbb{P}(E^c) \\
 &= \mathbb{P}(A|E) + \mathbb{P}(B^c) + \mathbb{P}(C^c) \\
 &\leq \frac{1}{n} + \frac{1}{n} + \frac{1}{n} = \frac{3}{n}.
 \end{aligned} \tag{4.374}$$

So we have shown

$$|\hat{y}_k| \leq \sqrt{2c \log(2mn)} \left(1 + \sqrt{\frac{2 \log(n)}{n}}\right) \tag{4.375}$$

for all  $k$  with probability at least  $1 - 3/n$ . □

### Establishing the Bounds of the Quantizers in the Sketched and Quantized Scenario with Gaussian $\mathbf{Z}$

We established in sections 4.3.3 and 4.4.2 that

$$|\hat{x}_{jk}| \leq \hat{R}_g = \left(1 + \sqrt{2 \log(nd)}\right) \sqrt{2 \log(2nmd)} \quad \text{w.p. } 1 - 2/n \tag{4.376}$$

$$|\hat{y}_k| \leq \hat{L}_g = \sqrt{2c \log(2nm)} \left(1 + \sqrt{\frac{2 \log(n)}{n}}\right) \quad \text{w.p. } 1 - 3/n. \tag{4.377}$$

for all  $j, k$  and for  $c = \left(\sigma + \|\beta^0\|_2\right)^2$ . Using these results, we establish our bounds for our three quantizers:

	$\alpha^-$	$\alpha^+$	$\Delta$	w.p.
$Q_{\hat{\mathbf{x}}}$	$-\hat{R}_g$	$\hat{R}_g$	$2\hat{R}_g$	$1 - 2/n$
$Q_{\hat{\mathbf{x}}^2}$	0	$\hat{R}_g^2$	$\hat{R}_g^2$	$1 - 2/n$
$Q_{\hat{\mathbf{y}}}$	$-\hat{L}_g$	$\hat{L}_g$	$2\hat{L}_g$	$1 - 3/n$

### Matrix Bernstein Inequality for the Sketched and Quantized Estimator with Gaussian $\mathbf{Z}$

With the event

$$E : \max_{jk} |\hat{x}_{jk}| \leq \hat{R}_g, \tag{4.378}$$

we can apply Lemma 12 to say

$$\begin{aligned} \mathbb{P} \left( \frac{1}{m} \left\| \sum_{k=1}^m (\tilde{\Sigma}_k - \Sigma) \right\| \geq \delta \right) &\leq \mathbb{P} \left( \frac{1}{m} \left\| \sum_{k=1}^m (\tilde{\Sigma}_k - \Sigma) \right\| \geq \delta \mid E \right) + \mathbb{P} \left( \max_{jk} |\hat{x}_{jk}| \geq \hat{R}_g \right) \\ &= \frac{2}{n} + \frac{2}{n} \\ &= \frac{4}{n} \end{aligned} \tag{4.379}$$

for  $\delta = \sqrt{\frac{2\log(2nd)\tau}{m} + \frac{2\log(2nd)b}{m}}$ , for any  $b \geq 2d\hat{R}_g^2 + \|\Sigma\|$ , and for

$$\begin{aligned}
 \tau &= d\hat{R}_g^2 \|\Sigma\| + (3d+2)\hat{R}_g^4 \\
 &= d \left( (1 + \sqrt{2\log(nd)}) \sqrt{2\log(2nmd)} \right)^2 \|\Sigma\| \\
 &\quad + (3d+2) \left( (1 + \sqrt{2\log(nd)}) \sqrt{2\log(2nmd)} \right)^4 \\
 &= d4\log^2(2nmd) \left( \frac{(1 + \sqrt{2\log(nd)})^2}{2\log(2nmd)} + \left(3 + \frac{2}{d}\right) (1 + \sqrt{2\log(nd)})^4 \right) \\
 &= d\log^2(2nmd) \left(1 + \sqrt{2\log(nd)}\right)^4 \left( 12 + \frac{2}{\log(2nmd) \left(1 + \sqrt{2\log(nd)}\right)^2} + \frac{8}{d} \right). \tag{4.380}
 \end{aligned}$$

This implies that

$$\|\tilde{\Sigma} - \Sigma\| \leq \sqrt{\frac{2\log(2nd)\tau}{m} + \frac{2\log(2nd)b}{m}} \tag{4.381}$$

with probability at least  $1 - 4/n$  where  $b \geq 2d\hat{R}_g^2 + \|\Sigma\|$  and

$$\tau = d\log^2(2nmd) \left(1 + \sqrt{2\log(nd)}\right)^4 \left( 12 + \frac{2}{\log(2nmd) \left(1 + \sqrt{2\log(nd)}\right)^2} + \frac{8}{d} \right).$$

#### 4.4.4 Summary of the Sketched and Quantized Scenario with Gaussian $\mathbf{Z}$

For a random  $\mathbf{Z}$  whose rows  $\mathbf{z}_i^T$  are drawn independently from a multivariate normal distribution with mean  $\mathbf{0}$  and covariance  $\mathbf{I}_{d \times d}$ . we defined

$$\begin{aligned} \boldsymbol{\Sigma} &= \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{Z}}{n} \right] = \mathbf{I}_d & \tilde{\boldsymbol{\Sigma}} &= \frac{1}{m} \sum_{k=1}^m \tilde{\boldsymbol{\Sigma}}_k = \frac{1}{m} \sum_{k=1}^m \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T + \tilde{\boldsymbol{\Delta}}_k \\ \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}} &= \mathbf{E} \left[ \mathbf{X}^T \mathbf{y} \right] = \mathbf{E} \left[ \frac{\mathbf{Z}^T \mathbf{w}}{n} \right] & \tilde{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} &= \frac{1}{m} \tilde{\mathbf{X}}^T \tilde{\mathbf{y}} = \frac{1}{m} \sum_{k=1}^m \tilde{x}_{kj} \tilde{y}_k \quad \text{for } 1 \leq j \leq d, \end{aligned}$$

where we let  $\tilde{\boldsymbol{\Sigma}}_k = \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T + \tilde{\boldsymbol{\Delta}}_k$  and  $\tilde{\boldsymbol{\Delta}}_k = \text{diag} \left( \tilde{x}_{kj}^2 - \tilde{x}_{kj}^2 \right)_{j=1}^d$ . We then showed

$$\left\| \tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \right\| \leq \sqrt{\frac{2d \log(2nd) \tau_0}{m}} + \frac{2d \log(2nd) b_0}{m} \quad \text{w.p. } 1 - 4/n \quad (4.382)$$

$$\left\| \tilde{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{y}} \right\|_2 \leq \sqrt{\frac{8d \hat{R}_g^2 \hat{L}_g^2 \log(2nd)}{m}} \quad \text{w.p. } 1 - 6/n \quad (4.383)$$

where

$$\tau_0 = \log^2(2nmd) \left( 1 + \sqrt{2 \log(nd)} \right)^4 \left( 12 + \frac{2}{\log(2nmd) \left( 1 + \sqrt{2 \log(nd)} \right)^2} + \frac{8}{d} \right)$$

$$\hat{R}_g = \left( 1 + \sqrt{2 \log(nd)} \right) \sqrt{2 \log(2nmd)} \quad b_0 = 2\hat{R}_g^2 + d^{-1} \left\| \boldsymbol{\Sigma} \right\|$$

$$\hat{L}_g = \sqrt{2c \log(2nm)} \left( 1 + \sqrt{\frac{2 \log(n)}{n}} \right) \quad c = \left( \sigma + \left\| \boldsymbol{\beta}^0 \right\|_2 \right)^2$$

We can summarize by stating

$$\|\tilde{\Sigma} - \Sigma\| = \mathcal{O}\left(\sqrt{\frac{d \log^5(nd) \log^2(nmd)}{m}}\right) \quad (4.384)$$

$$\|\tilde{\Sigma}_{\mathbf{Zw}} - \Sigma_{\mathbf{Zw}}\|_2 = \mathcal{O}\left(\sqrt{\frac{d \log^2(nd) \log(nm) \log(nmd)}{m}}\right). \quad (4.385)$$

Choosing  $m$  such that

$$m \geq \max \left\{ 2^4 d \log^3(2nd) \left(1 + \sqrt{2 \log(nd)}\right)^4 \eta^2 \left(6 + \frac{1}{\log(2nd) \left(1 + \sqrt{4 \log(nd)}\right)^2 + \frac{4}{d}}, 2^2 d \log(2nd) b_0 \eta \right) \right\} \quad (4.386)$$

for  $\eta = \frac{2}{\lambda_{\min}(\Sigma)}$  satisfies the lambda min requirement and allows us to say

$$\begin{aligned} \|\hat{\beta} - \beta^0\|_2 &\leq \frac{4\|\beta^*\|_2 \|\hat{\Sigma} - \Sigma\| + 4\|(\hat{\Sigma}_{\mathbf{Xy}} - \Sigma_{\mathbf{Xy}})\|_2}{\lambda_{\min}(\Sigma)} \\ &\leq \frac{4\|\beta^*\|_2 \left(\sqrt{\frac{2d \log(2nd)\tau_0}{m}} + \frac{2d \log(2nd)b_0}{m}\right) + 4\sqrt{\frac{8d\hat{R}_g^2 \hat{L}_g^2 \log(2nd)}{m}}}{\lambda_{\min}(\Sigma)} \end{aligned} \quad (4.387)$$

with probability at least  $1 - 10/n$ .

## 4.5 Discussion

Omitting the definitions and explanations that are provided in each chapter, we provide here a brief summary of the results from each section. Then, we analyze the results of each

section by calculating the relative efficiency of the estimator to the OLS estimator, where

$$RE(\hat{\beta}, \beta^*) := \frac{\mathbf{Var}[\hat{\beta}]}{\mathbf{Var}[\beta^*]}. \quad (4.388)$$

Let us define the estimator MSE as

$$MSE(\hat{\beta}) := \text{tr} \left( \mathbf{E} \left[ \left( \hat{\beta} - \mathbf{E}[\hat{\beta}] \right)^2 \right] \right) + \left( \hat{\beta} - \mathbf{E}[\hat{\beta}] \right)^2. \quad (4.389)$$

Since we have defined  $\hat{\beta}$  to be an unbiased estimator throughout, then

$$MSE(\hat{\beta}) := \text{tr} \left( \mathbf{E} \left[ \left( \hat{\beta} - \mathbf{E}[\hat{\beta}] \right)^2 \right] \right) = \mathbf{E} \left[ \left\| \hat{\beta} - \beta^0 \right\|_2^2 \right]. \quad (4.390)$$

In each section we have found an upper bound to  $\left\| \beta^0 - \hat{\beta} \right\|_2$  and thus an upper bound to its square, which we know is greater than its expectation. We can thus use this upper bound to find a worse case scenario bound on the relative efficiency (RE) compared to the OLS estimator.

$$\frac{\left\| \hat{\beta} - \beta^0 \right\|_2^2}{\sigma^2 \lambda_{\max}(\Sigma)} \leq RE(\hat{\beta}, \beta^*) \leq \frac{\left\| \hat{\beta} - \beta^0 \right\|_2^2}{\sigma^2 \lambda_{\min}(\Sigma)} \quad (4.391)$$

While these bounds are not tight, especially the lower bound, it does provide a worse case scenario of the performance of our estimators compared to the OLS estimator assuming the estimator's variance achieves its upper bound. While we know our estimator will never be more efficient than the OLS estimator, values closer to 1 indicate comparable efficiency.

### 4.5.1 Quantized Estimator

#### Quantized Estimator in Fixed Design

We showed that the difference between the estimator and the true  $\beta$  can be bound by

$$\|\beta^0 - \hat{\beta}\|_2 \leq \frac{2\|\beta^*\|_2 \left( \sqrt{\frac{2 \log(2nd)\tau}{n}} + \frac{4 \log(2nd)b}{n} \right) + 4\sqrt{\frac{8r^2 \ell^2 d \log(2nd)}{n}}}{\lambda_{\min}(\Sigma)} + \sqrt{\frac{\sigma^2 d}{\lambda_{\min}(\Sigma) n}}$$

where

$$\begin{aligned} \tau = d\tau_0 &= \left( r^2 \|\Sigma\| + \frac{r^4(1+3d)}{d} \right), & b = b_0 &= 2r^2 + d^{-1} \|\Sigma\|, \\ \ell &= rC\sqrt{\log(n)}\|\beta^0\|_2 + \sigma\sqrt{2 \log(2n^2)}, & \eta &= \frac{2}{\lambda_{\min}(\Sigma)}. \end{aligned}$$

Since we are interested in the relative efficiency of our estimator compared to the OLS estimator, we examine

$$\begin{aligned} RE(\hat{\beta}, \beta^*) &\leq \left( \frac{2\|\beta^*\|_2 \left( \sqrt{2 \log(2nd)\tau} + 4 \log(2nd)b \right) + 4\sqrt{8r^2 \ell^2 d \log(2nd)}}{\sqrt{\lambda_{\min}(\Sigma)} \sigma^2 d} \right)^2 \\ &= \frac{16}{\sigma^2 \lambda_{\min}(\Sigma)} \left( \|\beta^*\|_2 \left( \sqrt{2 \log(2nd)\tau_0} + 2\sqrt{d} \log(2nd)b_0 \right) + 4\sqrt{8r^2 \ell^2 \log(2nd)} \right)^2 \quad (4.392) \end{aligned}$$

From this we can see that the RE is dependent on  $r$ ,  $\ell$ , the nature of the design matrix, and the magnitude of the terms in  $\beta^*$  and  $\beta^0$ . All of these will contribute to the variance of our estimator and thus to the RE.

### Quantized Estimator in Random Design

We showed that the difference between the quantized estimator and the OLS estimator can be bounded by

$$\begin{aligned} \|\beta^* - \hat{\beta}\|_2 &\leq \frac{4\|\beta^*\|_2 \|\hat{\Sigma} - \Sigma\| + 4\|(\hat{\Sigma}\mathbf{x}_y - \Sigma\mathbf{x}_y)\|_2}{\lambda_{\min}(\Sigma)} \\ &\leq \frac{4\|\beta^*\|_2 \left( \sqrt{\frac{2\log(2nd)\tau}{n}} + \frac{2\log(2nd)b}{n} \right) + 4\sqrt{\frac{8dr_g^2\ell_g^2\log(2nd)}{n}}}{\lambda_{\min}(\Sigma)} \end{aligned} \quad (4.393)$$

where

$$\begin{aligned} \eta &= \frac{2}{\lambda_{\min}(\Sigma)}, \quad \tau = d\tau_0 = \log^2(2n^2d) \left( 12 + \frac{4}{d} \right), \quad b = b_0 = 2r_g^2 + d^{-1} \|\Sigma\|, \\ \ell_g &= \sqrt{2 \left( \|\beta^0\|_2^2 + \sigma^2 \right) \log(2n^2)}, \quad r_g = \sqrt{2\log(2n^2d)}. \end{aligned}$$

We can now calculate our the relative efficiency of our estimator compared to the OLS estimator as:

$$\begin{aligned} RE(\hat{\beta}, \beta^*) &\leq \left( \frac{2\|\beta^*\|_2 \left( \sqrt{2\log(2nd)\tau} + 4\log(2nd)b \right) + 4\sqrt{8r_g^2\ell_g^2d\log(2nd)}}{\sqrt{\lambda_{\min}(\Sigma)}\sigma^2d} \right)^2 \\ &= \frac{16}{\sigma^2\lambda_{\min}(\Sigma)} \left( \|\beta^*\|_2 \left( \sqrt{2\log(2nd)\tau_0} + 2\sqrt{d}\log(2nd)b_0 \right) + 4\sqrt{8r^2\ell^2\log(2nd)} \right)^2 \end{aligned} \quad (4.394)$$

This formulation's dependencies are the same as that of the fixed case. This is illustrated below in the figures of simulated data in the next section. Thus we expect the RE to increase at a rate of  $d$ .

### Quantized Estimator in Simulation

In a simulation using toy data created under the assumptions of this section, the average MSE of the estimator and the average RE were calculated for each combination of design matrix with dimensions  $n = 10^4, 10^5, 10^6$  and  $d = 3, 10, 20$  and whose entries were absolutely bounded by 1. The MSE is shown in Figure 4.1 and the RE is shown in Figure 4.2.

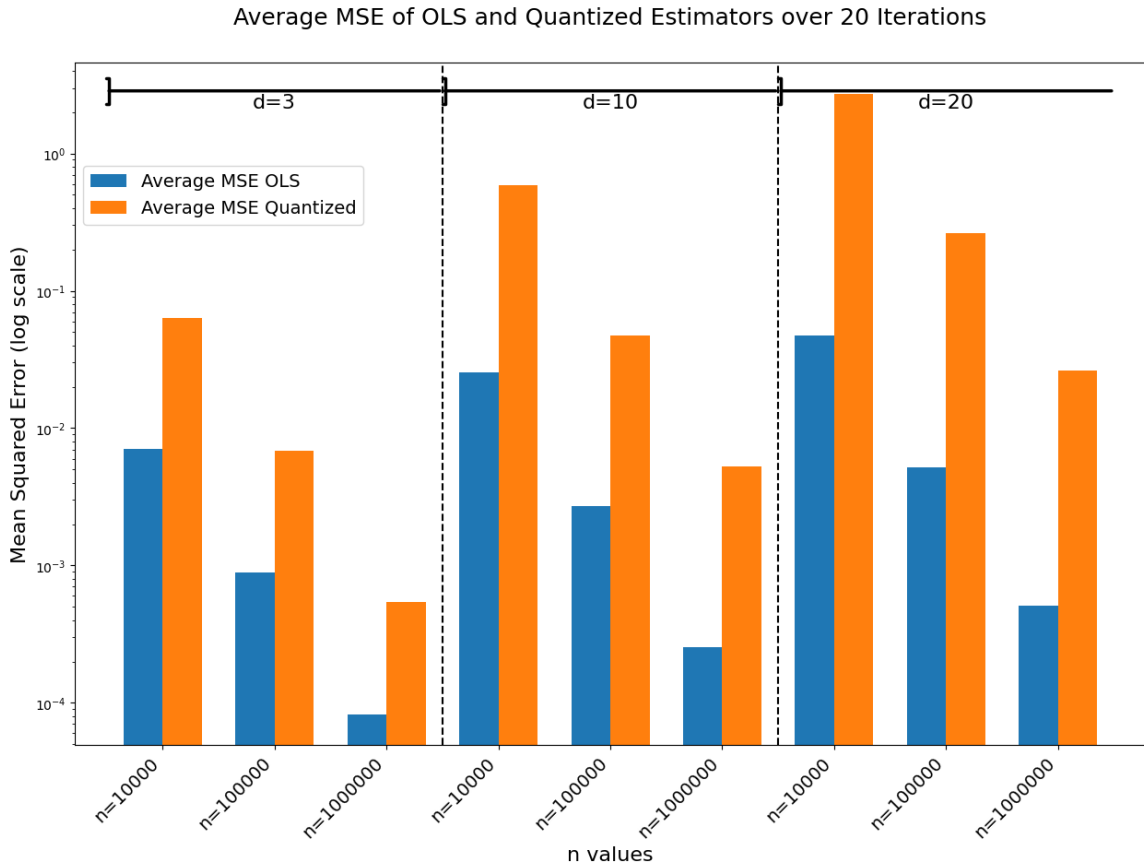


Figure 4.1: MSE of the Quantized Estimator

Figure 4.1 clearly shows the decay of the MSE as the number of samples increases. As expected, the MSE of the OLS estimator also decreases with the number of samples. As the quantizer is a lossy compression, we expect an inherent amount of error to exist between these two.

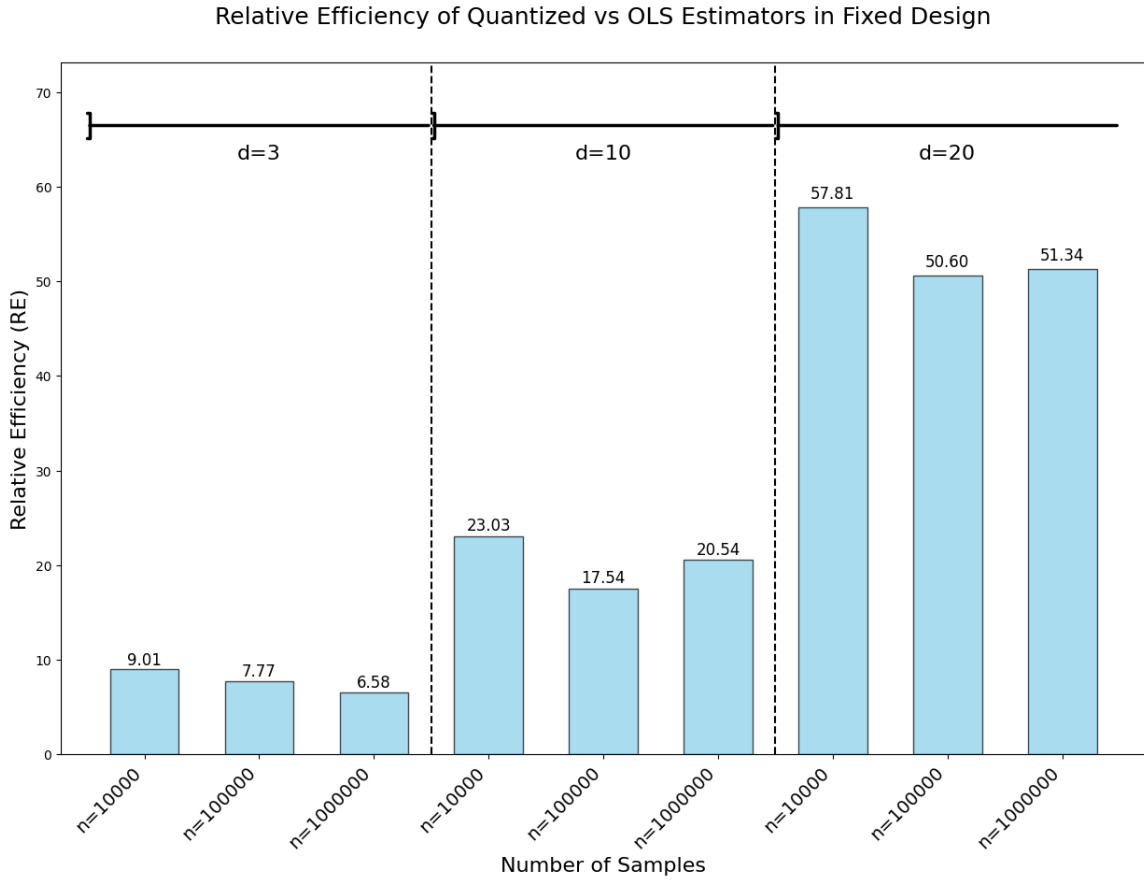


Figure 4.2: Relative Efficiency of Quantized vs OLS Estimators

In Figure 4.2, we see that as the number of features and samples increases, so does the RE; however, within each category we see the RE remains relatively consistent across the number of samples. This is consistent with our formulation in (4.392), as we expect the  $d$  term to have the greatest affect on the RE. The other terms dependent on  $n$  and  $d$  are  $\log(\cdot)$  or  $\sqrt{\log(\cdot)}$  which will have minimal effect.

## 4.5.2 Sketched Estimator

### Sketched Estimator in Fixed Design

We showed that the difference between the sketched estimator and the true parameter can be bounded by

$$\|\beta^0 - \hat{\beta}\|_2 \tag{4.395}$$

$$\leq \frac{4(2\epsilon + \epsilon^2) \left( \|\beta^*\|_2 \|\Sigma\| + \left( rC\sqrt{\log(n)} \|\beta^0\|_2 + \sigma \left( 1 + \sqrt{\frac{2\log(n)}{n}} \right) \|\mathbf{X}\| \right) \right)}{\lambda_{\min}(\Sigma)} + \sqrt{\frac{\sigma^2 d}{\lambda_{\min}(\Sigma) n}} \tag{4.396}$$

where

$$\eta = \frac{2}{\lambda_{\min}(\Sigma)}, \quad \epsilon = \sqrt{\frac{d}{m}} + \sqrt{\frac{2\log(2n)}{m}}.$$

Then the relative efficiency compared to the OLS estimator is given by

$$\begin{aligned}
 & RE(\hat{\beta}, \beta^*) \\
 & \leq \left( \frac{4(2\epsilon + \epsilon^2) \left( \|\beta^*\|_2 \|\Sigma\| + \left( rC\sqrt{\log(n)} \|\beta^0\|_2 + \sigma \left( 1 + \sqrt{\frac{2\log(n)}{n}} \right) \right) \|\mathbf{X}\| \right)}{\lambda_{\min}(\Sigma)} \right)^2 \frac{\lambda_{\min}(\Sigma) n}{\sigma^2 d} \\
 & = \frac{16n(2\epsilon + \epsilon^2)^2 \left( \|\beta^*\|_2 \|\Sigma\| + \left( rC\sqrt{\log(n)} \|\beta^0\|_2 + \sigma \left( 1 + \sqrt{\frac{2\log(n)}{n}} \right) \right) \|\mathbf{X}\| \right)^2}{\lambda_{\min}(\Sigma) \sigma^2 d} \\
 & = \frac{16n \left( \sqrt{\frac{d}{m}} + \sqrt{\frac{2\log(2n)}{m}} + \left( \sqrt{\frac{d}{m}} + \sqrt{\frac{2\log(2n)}{m}} \right)^2 \right)^2 C^2}{\lambda_{\min}(\Sigma) \sigma^2 d} \\
 & = \frac{16n \left( \frac{\sqrt{d} + \sqrt{2\log(2n)}}{\sqrt{m}} + \frac{d}{m} + \sqrt{\frac{8d\log(2n)}{m^2}} + \frac{2\log(2n)}{m} \right)^2 C^2}{\lambda_{\min}(\Sigma) \sigma^2 d} \\
 & = \frac{16n \left( \frac{\sqrt{d} + \sqrt{2\log(2n)}}{\sqrt{m}} + \frac{d + \sqrt{8d\log(2n)} + 2\log(2n)}{m} \right)^2 C^2}{\lambda_{\min}(\Sigma) \sigma^2 d} \\
 & = \frac{16n \left( \frac{d + \sqrt{8d\log(2n)} + 2\log(2n)}{m} + \frac{2(\sqrt{d} + \sqrt{2\log(2n)})(d + \sqrt{8d\log(2n)} + 2\log(2n))}{m^{3/2}} + \left( \frac{d + \sqrt{8d\log(2n)} + 2\log(2n)}{m} \right)^2 \right) C^2}{\lambda_{\min}(\Sigma) \sigma^2 d} \\
 & = \frac{16n \left( \frac{d}{m} + \tilde{O}\left(\frac{d^{3/2} + d + \sqrt{d}}{m^{3/2}}\right) + \tilde{O}\left(\frac{d^2 + d^{3/2} + d}{m^2}\right) \right) C^2}{\lambda_{\min}(\Sigma) \sigma^2 d} \\
 & = \tilde{O}\left(\frac{n}{m} + \frac{n\sqrt{d}}{m^{3/2}} + \frac{n}{m^{3/2}} + \frac{n}{m^{3/2}\sqrt{d}} + \frac{nd}{m^2} + \frac{n\sqrt{d}}{m^2} + \frac{n}{m^2}\right) \\
 & = \tilde{O}\left(\frac{n}{m}\right). \tag{4.397}
 \end{aligned}$$

Thus we expect the RE to grow at a rate of  $\frac{n}{m}$ . In all cases, the RE will be affected by  $r$ , the nature of  $\mathbf{X}$  and  $\Sigma$ , and  $\beta^*$ .

### Sketched Estimator in Random Design

We showed that the difference between the sketched estimator and the OLS estimator can be bounded by

$$\begin{aligned}
 & \|\beta^* - \hat{\beta}\|_2 \\
 & \leq \frac{4\|\beta^*\|_2 \left( \sqrt{\frac{2d \log(nd) \hat{R}_g^2 \|\Sigma\|}{m}} + \frac{2 \log(nd)b}{m} \right) + 4 \left( 1 + \sqrt{\frac{2 \log(n)}{n}} + \sqrt{\frac{d}{n}} \right) (2\epsilon + \epsilon^2) \left( 1 + \sqrt{\frac{2 \log(n)}{n}} \right) C}{\lambda_{\min}(\Sigma)} \\
 & = \frac{4\|\beta^*\|_2 \left( \sqrt{\frac{d}{m}} \sqrt{2 \log(nd) \hat{R}_g^2 \|\Sigma\|} + \frac{d}{m} (2 \log(nd)b_0) \right) + 4 (2\epsilon + \epsilon^2) \left( 1 + \sqrt{\frac{2 \log(n)}{n}} + \sqrt{\frac{d}{n}} \right) \left( 1 + \sqrt{\frac{2 \log(n)}{n}} \right) C}{\lambda_{\min}(\Sigma)} \\
 & = \tilde{O} \left( \sqrt{\frac{d}{m}} \right)
 \end{aligned}$$

where

$$\begin{aligned}
 \epsilon &= \sqrt{\frac{d}{m}} + \sqrt{\frac{2 \log(2n)}{m}}, & \eta &= \frac{2}{\lambda_{\min}(\Sigma)}, & b &= db_0 = 2\hat{R}_g^2 + d^{-1} \|\Sigma\| \\
 \hat{R}_g &= \left( 1 + \sqrt{2 \log(nd)} \right) \sqrt{2 \log(2nmd)}, & C &= \|\beta^*\|_2 + \sigma.
 \end{aligned}$$

Writing out these calculations would be cumbersome and space-consuming. Thus we simplify them using the  $\tilde{O}$  notation. Then we can show the RE decays at a rate of

$$RE(\hat{\beta}, \beta^*) \leq \frac{\tilde{O} \left( \sqrt{\frac{d}{m}} \right)^2}{\tilde{O} \left( \frac{d}{n} \right)} = \tilde{O} \left( \frac{n}{m} \right)$$

Thus, we have a term that grows at a rate of  $\frac{n}{m}$ , as in the fixed design case. We illustrate this behavior in the graphs of simulated data in the next section.

### Sketched Estimator in Simulation

We simulate data of dimensions  $n = 10^4, 10^5$  and  $d = 3, 10, 20$  according to the assumptions made in the random design scenario. In Figure 4.3 we display the results from the scenario with  $n = 10^4$  and in Figure 4.4 we show the results from the scenario with  $n = 10^5$ .

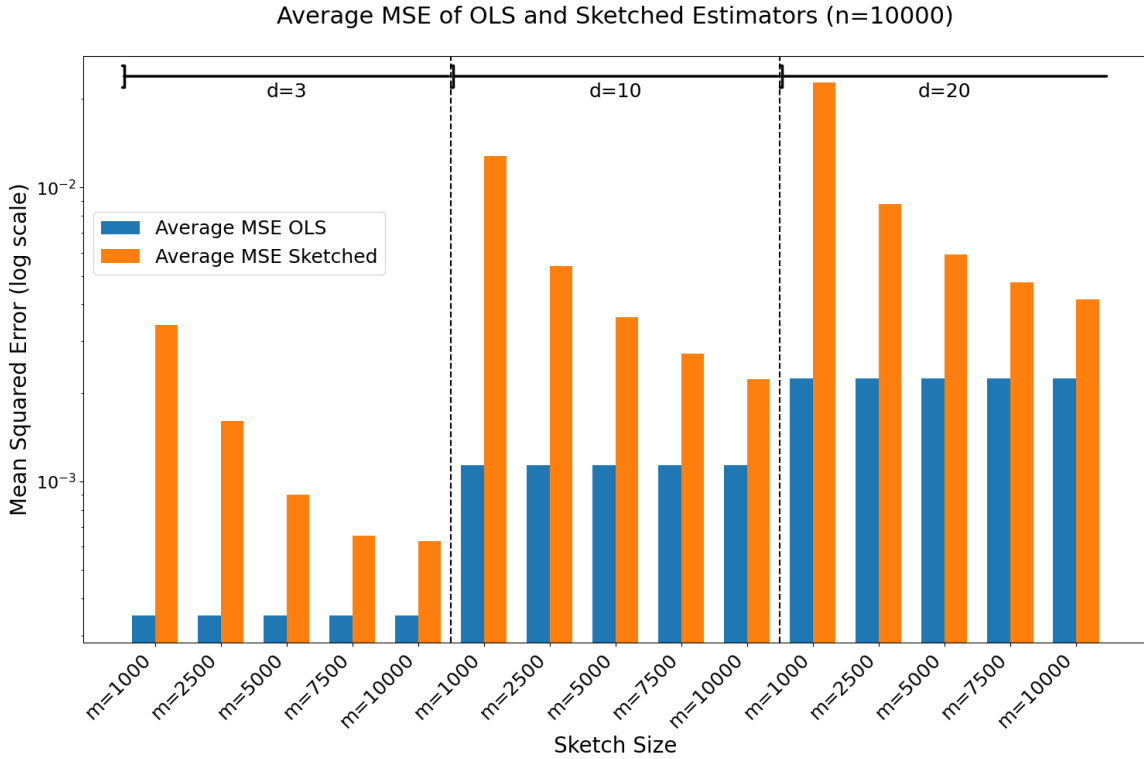


Figure 4.3: MSE of the Sketched Estimator with 10000 Samples

Comparing the two graphs, we can see that the values in the scenario with  $n = 10^5$  are substantially smaller in their relative case in the  $n = 10^4$  scenario. This is consistent with the rate of decay we expect to see based on our analysis, namely:  $\tilde{O}\left(\sqrt{\frac{d}{m}}\right) + \tilde{O}\left(\sqrt{\frac{d}{n}}\right)$ . We can also see that in all scenarios, the MSE decreases as  $m$  increases and increases as  $d$  increases. Again, this is as expected.

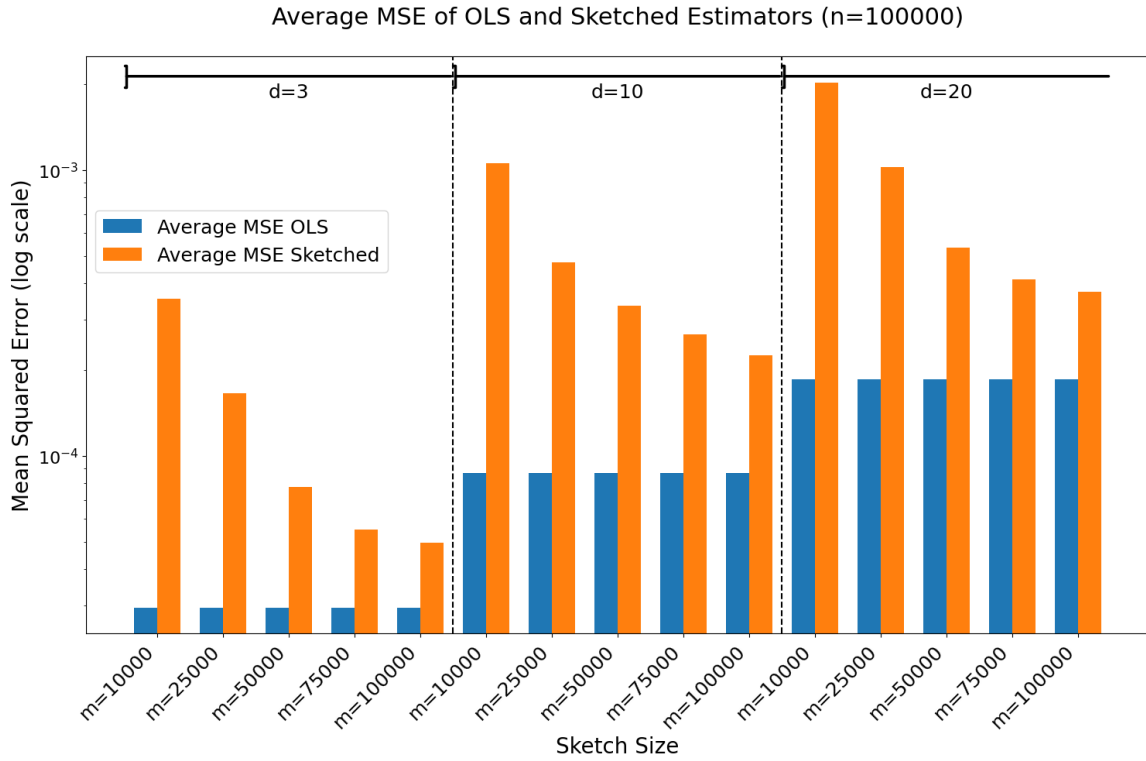


Figure 4.4: MSE of the Sketched Estimator with 100000 Samples

We also looked at the RE in simulation. Using the same data and scenarios, Figure 4.5 shows the scenario with  $n = 10^4$  and Figure 4.6 shows the scenario with  $n = 10^5$ . Comparing these graphs, we can see that the RE is almost unchanged between them for all parameter combinations. This is because the  $m$  that was chosen in each bar of the graph was the same proportion of  $n$ . Thus, the graphs demonstrate that the RE scales proportional to the ratio  $n/m$ , as expected.

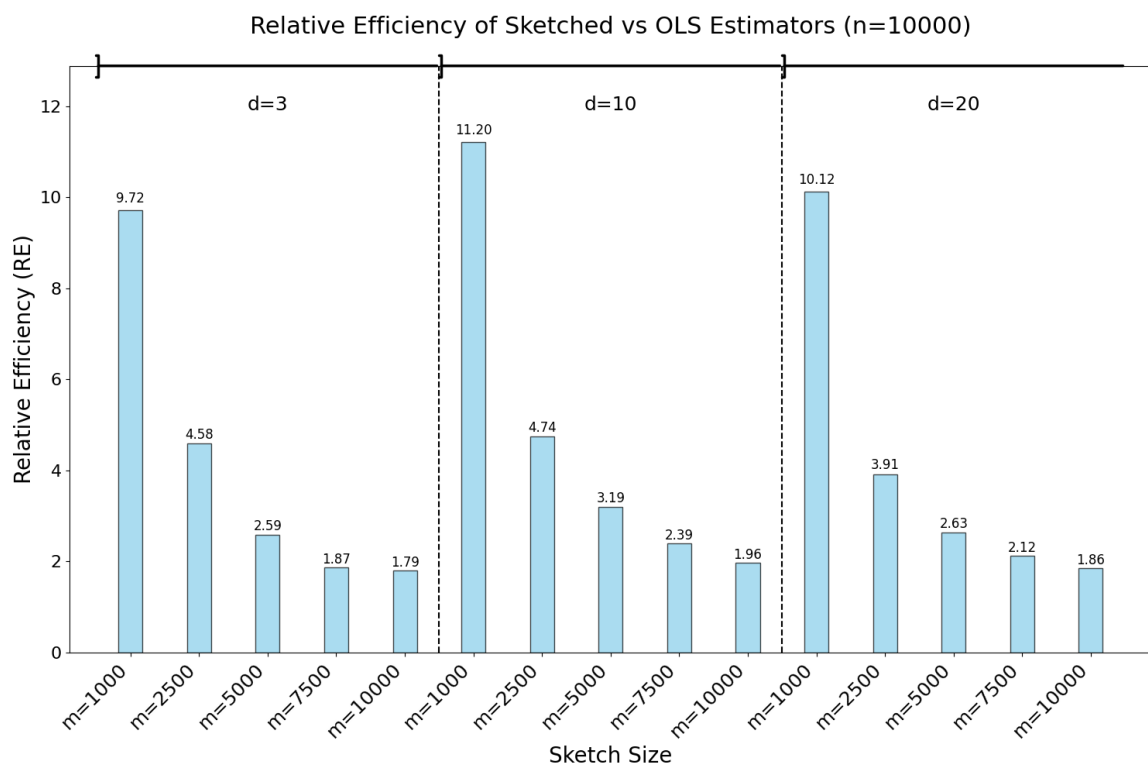


Figure 4.5: Relative Efficiency of Sketched Estimator for  $n = 10000$

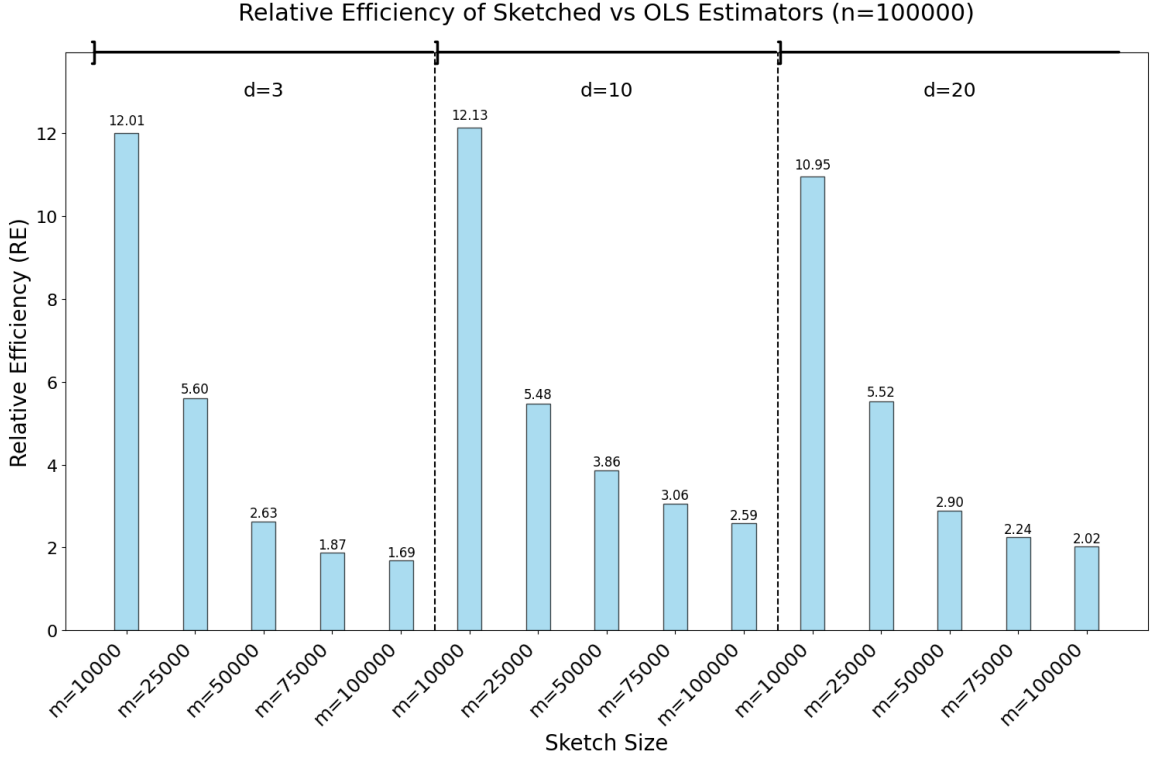


Figure 4.6: Relative Efficiency of Sketched Estimator for  $n = 100000$

### 4.5.3 Sketched and Quantized Estimator

#### Sketched and Quantized Estimator in Fixed Design

We showed that the difference between the sketched and quantized estimator and the true parameter is bounded by

$$\begin{aligned}
 & \|\beta^0 - \hat{\beta}\|_2 \\
 & \leq \frac{4\|\beta^*\|_2 \left( \sqrt{\frac{2 \log(2nd)\tau}{m}} + \frac{2 \log(2nd)b}{m} \right) + 4\sqrt{\frac{8d\hat{R}^2 \hat{L}^2 \log(2nd)}{m}}}{\lambda_{\min}(\Sigma)} + \sqrt{\frac{\sigma^2 d}{\lambda_{\min}(\Sigma) n}} \\
 & = \tilde{O}\left(\sqrt{\frac{d}{m}}\right) + \tilde{O}\left(\sqrt{\frac{d}{n}}\right)
 \end{aligned} \tag{4.398}$$

where

$$\begin{aligned}\hat{R} &:= \sqrt{2r^2 \log(2nmd)}, & \hat{L} &:= \sqrt{2\ell^2 \log(2nm)}, \\ \ell &= rC\sqrt{\log(n)} \|\beta^0\|_2 + \sigma\sqrt{2\log(2n^2)}, \\ \tau &= d\tau_0 = \log^2(2nmd)2r^2 \left( 3 + \frac{2}{d} + \frac{\|\Sigma\|}{\log(2nmd)} \right), \\ \eta &= \frac{2}{\lambda_{\min}(\Sigma)}, & b &= db_0 = 2\hat{R}^2 + d^{-1} \|\Sigma\|.\end{aligned}$$

Then we can calculate the RE as

$$\begin{aligned}RE(\hat{\beta}, \beta^*) &\leq \frac{\left( 4\|\beta^*\|_2 \left( \sqrt{\frac{2\log(2nd)\tau}{m}} + \frac{2\log(2nd)b}{m} \right) + 4\sqrt{\frac{8d\hat{R}^2\hat{L}^2\log(2nd)}{m}} \right)^2}{\lambda_{\min}(\Sigma)^2 \frac{\sigma^2 d}{\lambda_{\min}(\Sigma)n}} \\ &= \frac{n \left( 4\|\beta^*\|_2 \left( \sqrt{\frac{2d\log(2nd)\tau_0}{m}} + \frac{2d\log(2nd)b_0}{m} \right) + 4\sqrt{\frac{8d\hat{R}^2\hat{L}^2\log(2nd)}{m}} \right)^2}{\lambda_{\min}(\Sigma) \sigma^2}\end{aligned}\tag{4.399}$$

$$= \frac{n \left( 4\|\beta^*\|_2 \left( \sqrt{\frac{2\log(2nd)\tau_0}{m}} + \frac{2\sqrt{d}\log(2nd)b_0}{m} \right) + 4\sqrt{\frac{8\hat{R}^2\hat{L}^2\log(2nd)}{m}} \right)^2}{\lambda_{\min}(\Sigma) d\sigma^2}\tag{4.400}$$

$$= \tilde{\mathcal{O}}\left(\frac{n}{m}\right).\tag{4.401}$$

Thus the RE decays at a rate comparable to that of the sketched estimator, albeit with different dependencies on the constant terms.

### Sketched and Quantized Estimator in Random Design

We showed that the difference between the sketched and quantized estimator and the OLS estimator in a random design setting can be bounded by

$$\begin{aligned} \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 &\leq \frac{4\|\boldsymbol{\beta}^*\|_2 \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| + 4\|(\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{Y}})\|_2}{\lambda_{\min}(\boldsymbol{\Sigma})} \\ &\leq \frac{4\|\boldsymbol{\beta}^*\|_2 \left( \sqrt{\frac{2\log(2nd)\tau}{m}} + \frac{2\log(2nd)b}{m} \right) + 4\sqrt{\frac{8d\hat{R}_g^2 \hat{L}_g^2 \log(2nd)}{m}}}{\lambda_{\min}(\boldsymbol{\Sigma})} \\ &= \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{m}}\right) \end{aligned}$$

where

$$\tau = d\tau_0 = \log^2(2nmd) \left(1 + \sqrt{2\log(nd)}\right)^4 \left(12 + \frac{2}{\log(2nmd) \left(1 + \sqrt{2\log(nd)}\right)^2} + \frac{8}{d}\right),$$

$$b = db_0 = 2\hat{R}_g^2 + d^{-1} \|\boldsymbol{\Sigma}\|, \quad \eta = \frac{2}{\lambda_{\min}(\boldsymbol{\Sigma})},$$

$$\hat{R}_g = \left(1 + \sqrt{2\log(nd)}\right) \sqrt{2\log(2nmd)}, \quad \hat{L}_g = \sqrt{2c\log(2mn)} \left(1 + \sqrt{\frac{2\log(n)}{n}}\right).$$

Then we can calculate the RE as

$$RE(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \leq \frac{\tilde{\mathcal{O}}\left(\sqrt{\frac{d}{m}}\right)^2}{\tilde{\mathcal{O}}\left(\frac{d}{n}\right)} = \tilde{\mathcal{O}}\left(\frac{n}{m}\right). \quad (4.402)$$

Thus the RE decays at a rate equivalent to its fixed counterpart at  $\frac{n}{m}$ . We show this using simulated data in the graphs in the following section.

### Sketched and Quantized Estimator in Simulation

In a simulation using toy data created under the assumptions of this section, the average MSE of the estimator and the average RE were calculated for each combination of design matrix with dimensions  $n = 10^5, 10^6, 10^7$  and  $d = 3$ . The MSE is shown in Figure 4.7 and the RE is shown in Figure 4.8.

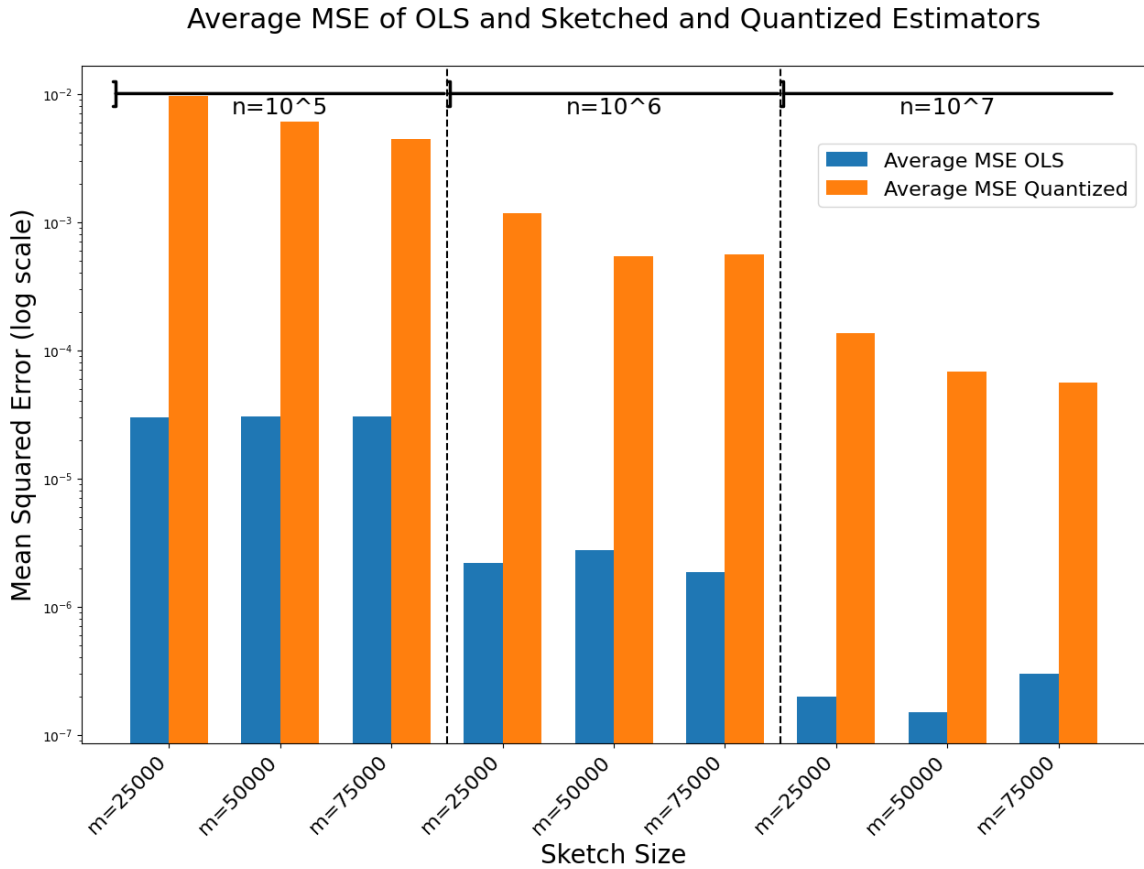


Figure 4.7: MSE of the Sketched and Quantized Estimator

Figure 4.7 clearly shows the decay of the MSE as the number of samples and sketched size increase. This agrees with the expected rate of decay of the MSE as calculated in the fixed section, which yielded the MSE would decay at a rate of  $\tilde{O}\left(\sqrt{\frac{d}{m}}\right) + \tilde{O}\left(\sqrt{\frac{d}{n}}\right)$ . As

also expected, the MSE of the OLS estimator decreases with the number of samples. As the quantizer is a lossy compression, we expect an inherent amount of error to exist between these two.

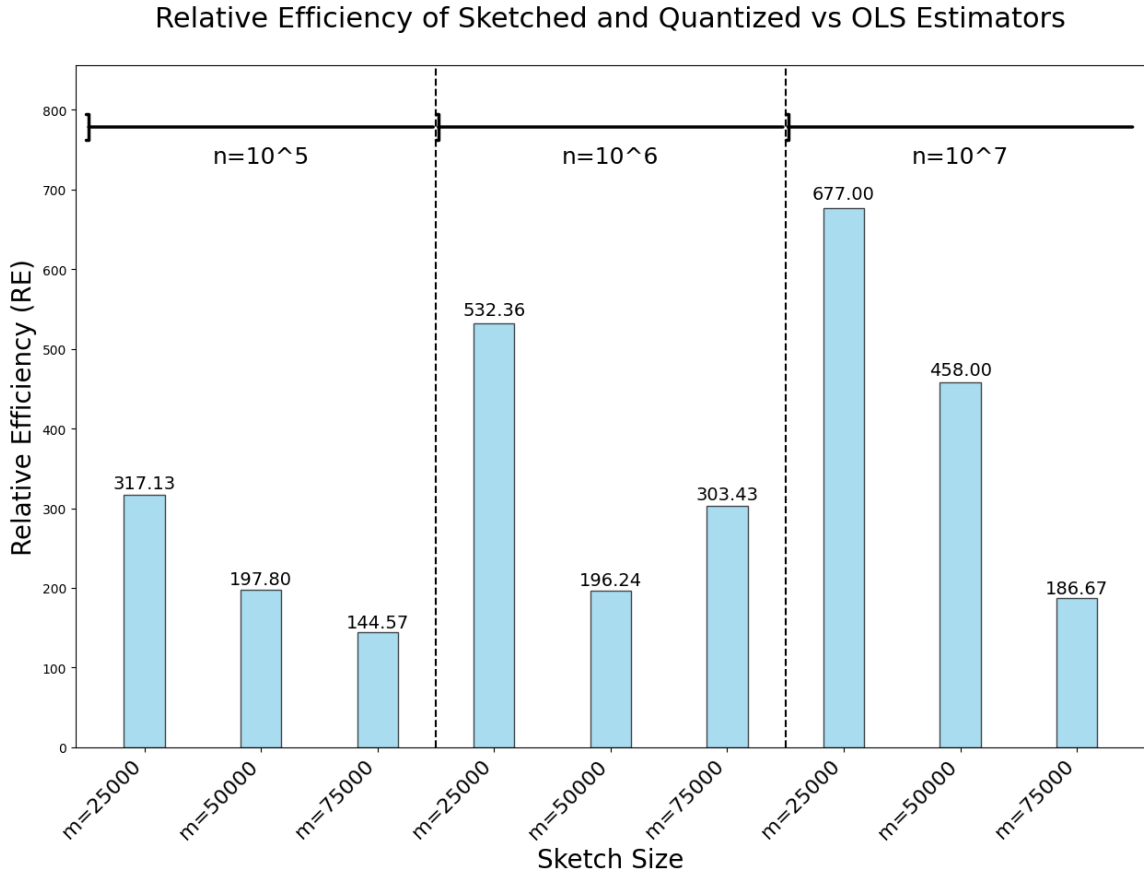


Figure 4.8: Relative Efficiency of Sketched and Quantized vs OLS Estimators

In Figure 4.2, we see that as the number of samples increases, the RE does not decrease. However, within each  $n$  value, we see that the RE decreases as  $m$  increases. Since the  $m$  values do not change between  $n$  values, the ratio  $n/m$  increases. The graph shows that the RE follows this relationship, which is as we calculated. We calculated that the RE would follow a rate of decay  $\tilde{O}\left(\frac{n}{m}\right)$ .

## Appendix A: Derivations from Section 3.2.2

### A.1 Preliminary Formulations

Let us assume a random variable  $X$  is bounded on  $[-R, R]$  and  $\Delta := R - (-R) = 2R$ . Then we can show a quantizer  $Q_X$  is unbiased:

$$\mathbf{E}_Q[\tilde{X}] = R \frac{X - (-R)}{\Delta} + (-R) \frac{R - X}{\Delta} = \frac{X(R - (-R))}{\Delta} = \frac{X\Delta}{\Delta} = X$$

From unbiasedness, we can then show the following three results which are used extensively in the formulations throughout this section:

i.  $\mathbf{E}_Q[\tilde{X}X] = X^2$  Indeed, we have:

$$\mathbf{E}_Q[\tilde{X}X] = \mathbf{E}_Q[\tilde{X}X] = \mathbf{E}_Q[\tilde{X}]X = X^2$$

ii.  $\mathbf{E}_Q[X(\tilde{X} - X)] = 0$  Using part i.:

$$\mathbf{E}_Q[X(\tilde{X} - X)] = \mathbf{E}_Q[X\tilde{X}] - \mathbf{E}_Q[X^2] = X^2 - X^2 = 0$$

iii.  $\mathbf{Var}_Q [\tilde{X}] \leq \frac{\Delta^2}{4}$ . Noting that  $\Delta = R - (-R) = 2R$  then

$$\begin{aligned} \mathbf{Var}_Q [\tilde{X}] &= \mathbf{E}_Q [\tilde{X}^2] - \mathbf{E}_Q [\tilde{X}]^2 \\ &= R^2 \frac{X - (-R)}{\Delta} + (-R)^2 \frac{R - X}{\Delta} - X^2 \\ &= \frac{1}{\Delta} (XR^2 + R^3 + R^3 - XR^2 - \Delta X^2) \\ &= \frac{1}{2R} (2R^3 - 2RX^2) \\ &= R^2 - X^2 \\ &= (R - X)(X - (-R)) \end{aligned}$$

We note that minimizing the first operand maximizes the second, and vice versa.

Thus, the maximum value is achieved when  $\frac{\Delta}{2} = R - X = X - (-R)$ . Thus, we can

conclude  $\mathbf{Var}_Q [\tilde{X}] \leq \frac{\Delta^2}{4}$

## A.2 $\mathbf{E} [\hat{U}]$

$$\begin{aligned} \mathbf{E} [\hat{U}] &= n^{-1} \sum_{i=1}^n \mathbf{E} [\tilde{X}_i \tilde{Y}_i] = n^{-1} \sum_{i=1}^n \mathbf{E}_X \left[ \mathbf{E}_Q [\tilde{X}_i \tilde{Y}_i \mid X_i, Y_i] \right] \\ &= n^{-1} \sum_{i=1}^n \mathbf{E} [X_i Y_i] \\ &= n^{-1} \beta^0 \sum_{i=1}^n \mathbf{E}_X [X_i^2] \\ &= \beta^0 \mathbf{E}_X [X_1^2] \end{aligned}$$

where  $\beta^0$  is the true value of  $\beta$  in the regression equation.

### A.3 $\mathbf{E} \left[ \hat{V} \right]$

Using the unbiasedness of the quantizer, we can see,

$$\mathbf{E} \left[ \hat{V} \right] = n^{-1} \sum_{i=1}^n \mathbf{E} \left[ \widetilde{X}_i^2 \right] = n^{-1} \sum_{i=1}^n \mathbf{E}_X \left[ X_i^2 \right] = \mathbf{E}_X \left[ X_1^2 \right]$$

### A.4 $\mathbf{Var} \left[ \widetilde{X}_i^2 \right]$

$$\mathbf{Var} \left[ \widetilde{X}_i^2 \right] = \mathbf{E} \left[ \widetilde{X}_i^2 \right]^2 - \mathbf{E} \left[ \widetilde{X}_i^4 \right] = \mathbf{E}_X \left[ (R^2)^2 \frac{X_i^2}{R^2} \right] - \mathbf{E}_X \left[ X_i^2 \right]^2 = R^2 \mathbf{E}_X \left[ X_i^2 \right] - \mathbf{E}_X \left[ X_i^2 \right]^2$$

### A.5 $\mathbf{Var} \left[ \widetilde{X}_i \widetilde{Y}_i \right]$

$$\mathbf{Var} \left[ \widetilde{X}_i \widetilde{Y}_i \right] = \mathbf{E} \left[ \left( \widetilde{X}_i \widetilde{Y}_i \right)^2 \right] - \mathbf{E} \left[ \widetilde{X}_i \widetilde{Y}_i \right]^2 = B^2 R^2 - \mathbf{E} \left[ X_i Y_i \right]^2 = B^2 R^2 - \beta^{0^2} \mathbf{E}_X \left[ X_i^2 \right]^2$$

### A.6 $\mathbf{Cov} \left( \widetilde{X}_i \widetilde{Y}_i, \widetilde{X}_i^2 \right)$

$$\begin{aligned} \mathbf{Cov} \left( \widetilde{X}_i \widetilde{Y}_i, \widetilde{X}_i^2 \right) &= \mathbf{E} \left[ \widetilde{X}_i \widetilde{Y}_i \widetilde{X}_i^2 \right] - \mathbf{E} \left[ \widetilde{X}_i \widetilde{Y}_i \right] \mathbf{E} \left[ \widetilde{X}_i^2 \right] \\ &= \mathbf{E} \left[ X_i Y_i X_i^2 \right] - \mathbf{E} \left[ X_i Y_i \right] \mathbf{E}_X \left[ X_i^2 \right] \\ &= \mathbf{E}_X \left[ X_i^3 \mathbf{E}_\epsilon \left[ (X_i \beta^0 + \sigma \epsilon) \right] \right] - \beta^0 \mathbf{E}_X \left[ X_i^2 \right]^2 \\ &= \beta^0 \left( \mathbf{E}_X \left[ X_i^4 \right] - \mathbf{E}_X \left[ X_i^2 \right]^2 \right) \end{aligned}$$

## Appendix B: Derivations from Section 3.3.4

### B.1 Formulas for Calculating the Elements of $\mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right]$

$$\begin{aligned}
 \mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right] &= \mathbf{E} \left[ \left( \tilde{\Sigma}_{\mathbf{x}_i} \beta^0 - \tilde{\mathbf{x}}_i \tilde{Y}_i \right) \left( \tilde{\Sigma}_{\mathbf{x}_i} \beta^0 - \tilde{\mathbf{x}}_i \tilde{Y}_i \right)^T \right] \\
 &= \mathbf{E} \left[ \tilde{\Sigma}_{\mathbf{x}_i} \beta^0 \beta^{0T} \tilde{\Sigma}_{\mathbf{x}_i}^T - \tilde{\Sigma}_{\mathbf{x}_i} \beta^0 \tilde{Y}_i \tilde{\mathbf{x}}_i^T - \tilde{\mathbf{x}}_i \tilde{Y}_i \beta^{0T} \tilde{\Sigma}_{\mathbf{x}_i}^T + \tilde{\mathbf{x}}_i \tilde{Y}_i \tilde{Y}_i \tilde{\mathbf{x}}_i^T \right] \\
 &= \mathbf{E} \left[ \left( \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T + \Delta_i \right) \beta^0 \beta^{0T} \left( \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T + \Delta_i \right) - \left( \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T + \Delta_i \right) \beta^0 \tilde{Y}_i \tilde{\mathbf{x}}_i^T \right. \\
 &\quad \left. - \tilde{\mathbf{x}}_i \tilde{Y}_i \beta^{0T} \left( \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T + \Delta_i \right) + \tilde{\mathbf{x}}_i \tilde{Y}_i \tilde{Y}_i \tilde{\mathbf{x}}_i^T \right] \\
 &= \mathbf{E} \left[ \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T \beta^0 \beta^{0T} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T + \Delta_i \beta^0 \beta^{0T} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T + \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T \beta^0 \beta^{0T} \Delta_i + \Delta_i \beta^0 \beta^{0T} \Delta_i \right. \\
 &\quad \left. - \tilde{Y}_i \left( \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T \beta^0 \tilde{\mathbf{x}}_i^T + \Delta_i \beta^0 \tilde{\mathbf{x}}_i^T + \tilde{\mathbf{x}}_i \beta^{0T} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T + \tilde{\mathbf{x}}_i \beta^{0T} \Delta_i \right) + B^2 \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T \right]. \quad (\text{B.1})
 \end{aligned}$$

Then the  $j, k$ th element is given by

$$\begin{aligned}
 &\left( \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right)_{jk} \\
 &= \sum_{\ell=1}^d \sum_{m=1}^d \tilde{x}_{ij} \tilde{x}_{i\ell} \beta_{\ell}^0 \beta_m^0 \tilde{x}_{im} \tilde{x}_{ik} + \delta_j \beta_j^0 \sum_{m=1}^d \beta_m^0 \tilde{x}_{im} \tilde{x}_{ik} + \delta_k \beta_k^0 \sum_{m=1}^d \beta_m^0 \tilde{x}_{im} \tilde{x}_{ij} + \delta_j \beta_j^0 \beta_k^0 \delta_k \\
 &\quad - \tilde{Y}_i \left( \sum_{m=1}^d \beta_m^0 \tilde{x}_{im} \tilde{x}_{ij} \tilde{x}_{ik} + \beta_j^0 \delta_j \tilde{x}_{ik} + \sum_{m=1}^d \beta_m^0 \tilde{x}_{im} \tilde{x}_{ij} \tilde{x}_{ik} + \beta_k^0 \delta_k \tilde{x}_{ij} \right) + B^2 \tilde{x}_{ij} \tilde{x}_{ik} \quad (\text{B.2})
 \end{aligned}$$

where we have let  $\delta_p$  be the  $p, p$ th element of  $\mathbf{\Delta}_i$ . Then taking the expectation with respect to the quantizers of the  $j, k$ th term is

$$\begin{aligned}
 \mathbf{E}_Q \left[ \left( \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right)_{jk} \right] &= \underbrace{\sum_{\ell=1}^d \sum_{m=1}^d \beta_{\ell}^0 \beta_m^0 \mathbf{E}_Q [\tilde{x}_{ij} \tilde{x}_{i\ell} \tilde{x}_{im} \tilde{x}_{ik}]}_{(a)} + \underbrace{\beta_j^0 \mathbf{E}_Q [\delta_j] \sum_{m=1}^d \beta_m^0 \mathbf{E}_Q [\tilde{x}_{im} \tilde{x}_{ik}]}_{(b)} \\
 &+ \underbrace{\mathbf{E}_Q [\delta_k] \beta_k^0 \sum_{m=1}^d \beta_m^0 \mathbf{E} [\tilde{x}_{im} \tilde{x}_{ij}]}_{(c)} + \delta_j \beta_j^0 \beta_k^0 \delta_k + B^2 \mathbf{E} [\tilde{x}_{ij} \tilde{x}_{ik}] \\
 &- \mathbf{E}_Q [\tilde{Y}_i] \left( \underbrace{2 \sum_{m=1}^d \beta_m^0 \mathbf{E}_Q [\tilde{x}_{im} \tilde{x}_{ij} \tilde{x}_{ik}]}_{(d)} + \beta_j^0 \mathbf{E}_Q [\delta_j] \mathbf{E}_Q [\tilde{x}_{ik}] + \beta_k^0 \mathbf{E}_Q [\delta_k] \mathbf{E}_Q [\tilde{x}_{ij}] \right). \quad (\text{B.3})
 \end{aligned}$$

When we take the expectation with respect to the quantizer, we know that the quantization of the squared terms are independent of the quantization of their original terms. This allows us to separate the expectation of the  $\delta_p$  terms from the  $x$  terms.

We derive the formula for the terms (a), (b), (c), and (d). We must compare the possible values of the iterable variables to each other and to  $j$  and  $k$ , as their different combinations will result in different dependencies of the variables. These dependencies will affect how the expectations are calculated. We will use brackets below terms to denote how those terms have changed from previous steps based on the comparison of iterable variables.

For example, if we have a term with a double sum, the first sum with respect to  $m$  and the second sum with respect to  $\ell$ . Our first step might be to compare  $\ell$  to  $j$  by splitting the term into two terms, the first assumes  $\ell = j$  and the second  $\ell \neq j$ . In the new term that assumes  $\ell = j$ , all variables with a  $\ell$  subscript now become  $j$  and the sum over  $\ell$  is eliminated. In the new second term, the sum over  $\ell$  changes from a sum from 1 to  $d$  to a sum 1 to  $d$ , but not  $j$ . We denote this by simply putting  $\ell \neq j$  as a subscript.

To aid readability, we use a numbering system. If the original sum is labeled (a1), then when (a1) is split, the resulting two terms will have labels (a11) and (a12). If (a11) is then split, the resulting two terms will have labels (a111) and (a112). Each time a sum is split, the resulting terms will have its parent's label plus an additional digit.

The splitting and labeling are depicted in the first line of the formulation of term (a):

### B.1.1 Term(a)

$$\begin{aligned}
 & \underbrace{\sum_{\ell=1}^d \sum_{m=1}^d \beta_{\ell}^0 \beta_m^0 \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{i\ell} \widetilde{x}_{im} \widetilde{x}_{ik}]}_{(a)} = \underbrace{\sum_{m=1}^d \beta_j^0 \beta_m^0 \mathbf{E} [\widetilde{x}_{ij}^2 \widetilde{x}_{im} \widetilde{x}_{ik}]}_{\substack{\ell=j \\ (a1)}} + \underbrace{\sum_{\ell \neq j}^d \sum_{m=1}^d \beta_{\ell}^0 \beta_m^0 \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{i\ell} \widetilde{x}_{im} \widetilde{x}_{ik}]}_{\substack{\ell \neq j \\ (a2)}} \\
 & = \underbrace{\beta_j^{02} \mathbf{E} [\widetilde{x}_{ij}^3 \widetilde{x}_{ik}]}_{\substack{\ell=j, m=j \\ (a11)}} + \underbrace{\sum_{m \neq j}^d \beta_j^0 \beta_m^0 \mathbf{E} [\widetilde{x}_{ij}^2 \widetilde{x}_{im} \widetilde{x}_{ik}]}_{\substack{\ell=j, m \neq j \\ (a12)}} \\
 & + \underbrace{\sum_{\ell \neq j}^d \beta_{\ell}^0 \beta_j^0 \mathbf{E} [\widetilde{x}_{ij}^2 \widetilde{x}_{i\ell} \widetilde{x}_{ik}]}_{\substack{\ell \neq j, m=j \\ (a21)}} + \underbrace{\sum_{\ell \neq j}^d \sum_{m \neq j}^d \beta_{\ell}^0 \beta_m^0 \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{i\ell} \widetilde{x}_{im} \widetilde{x}_{ik}]}_{\substack{\ell \neq j, m \neq j \\ (a22)}} \\
 & = \underbrace{\beta_j^{02} \mathbf{E} [\widetilde{x}_{ij}^3 \widetilde{x}_{ik}]}_{\substack{\ell=j, m=j \\ (a11)}} + \underbrace{\beta_j^0 \beta_k^0 \mathbf{E} [\widetilde{x}_{ij}^2 \widetilde{x}_{ik}^2]}_{\substack{\ell=j, m \neq j, m=k \\ (a121)}} + \underbrace{\sum_{\substack{m \neq j \\ m \neq k}}^d \beta_j^0 \beta_m^0 \mathbf{E} [\widetilde{x}_{ij}^2 \widetilde{x}_{im} \widetilde{x}_{ik}]}_{\substack{\ell=j, m \neq j, m \neq k \\ (a122)}} + \underbrace{\beta_k^0 \beta_j^0 \mathbf{E} [\widetilde{x}_{ij}^2 \widetilde{x}_{ik}^2]}_{\substack{\ell \neq j, m=j, \ell=k \\ (a211)}} \\
 & + \underbrace{\sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_{\ell}^0 \beta_j^0 \mathbf{E} [\widetilde{x}_{ij}^2 \widetilde{x}_{i\ell} \widetilde{x}_{ik}]}_{\substack{\ell \neq j, m=j, \ell \neq k \\ (a212)}} + \underbrace{\sum_{\ell \neq j}^d \beta_{\ell}^0 \beta_k^0 \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{i\ell} \widetilde{x}_{ik}^2]}_{\substack{\ell \neq j, m \neq j, m=k \\ (a221)}} + \underbrace{\sum_{\ell \neq j}^d \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_{\ell}^0 \beta_m^0 \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{i\ell} \widetilde{x}_{im} \widetilde{x}_{ik}]}_{\substack{\ell \neq j, m \neq j, m \neq k \\ (a222)}}. \quad (\text{B.4})
 \end{aligned}$$

Two additional comparisons are required for (a221) and (a222):

$$\begin{aligned}
 & \underbrace{\sum_{\ell=1}^d \sum_{m=1}^d \beta_{\ell}^0 \beta_m^0 \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{i\ell} \widetilde{x}_{im} \widetilde{x}_{ik}]}_{(a)} \\
 &= \underbrace{\beta_j^{02} \mathbf{E} [\widetilde{x}_{ij}^3 \widetilde{x}_{ik}]}_{\substack{\ell=j, m=j \\ (a11)}} + \underbrace{\beta_j^0 \beta_k^0 \mathbf{E} [\widetilde{x}_{ij}^2 \widetilde{x}_{ik}^2]}_{\substack{\ell=j, m \neq j, m=k \\ (a121)}} + \underbrace{\sum_{\substack{m \neq j \\ m \neq k}}^d \beta_j^0 \beta_m^0 \mathbf{E} [\widetilde{x}_{ij}^2 \widetilde{x}_{im} \widetilde{x}_{ik}]}_{\substack{\ell=j, m \neq j, m \neq k \\ (a122)}} + \underbrace{\beta_k^0 \beta_j^0 \mathbf{E} [\widetilde{x}_{ij}^2 \widetilde{x}_{ik}^2]}_{\substack{\ell \neq j, m=j, \ell=k \\ (a211)}} \\
 &+ \underbrace{\sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_{\ell}^0 \beta_j^0 \mathbf{E} [\widetilde{x}_{ij}^2 \widetilde{x}_{i\ell} \widetilde{x}_{ik}]}_{\substack{\ell \neq j, m=j, \ell \neq k \\ (a212)}} + \underbrace{\beta_k^{02} \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{ik}^3]}_{\substack{\ell \neq j, m \neq j, m=k, \ell=k \\ (a2211)}} + \underbrace{\sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_{\ell}^0 \beta_k^0 \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{i\ell} \widetilde{x}_{ik}^2]}_{\substack{\ell \neq j, m \neq j, m=k, \ell \neq k \\ (a2212)}} + \underbrace{\sum_{\substack{m \neq j \\ m \neq k}}^d \beta_k^0 \beta_m^0 \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{im} \widetilde{x}_{ik}^2]}_{\substack{\ell \neq j, m \neq j, m \neq k, \ell=k \\ (a2221)}} \\
 &+ \underbrace{\sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_{\ell}^0 \beta_m^0 \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{i\ell} \widetilde{x}_{im} \widetilde{x}_{ik}]}_{\substack{\ell \neq j, m \neq j, m \neq k, \ell \neq k \\ (a2222)}} \\
 &= \underbrace{\beta_j^{02} \mathbf{E} [\widetilde{x}_{ij}^3 \widetilde{x}_{ik}]}_{\substack{\ell=j, m=j \\ (a11)}} + \underbrace{\beta_j^0 \beta_k^0 \mathbf{E} [\widetilde{x}_{ij}^2 \widetilde{x}_{ik}^2]}_{\substack{\ell=j, m \neq j, m=k \\ (a121)}} + \underbrace{\sum_{\substack{m \neq j \\ m \neq k}}^d \beta_j^0 \beta_m^0 \mathbf{E} [\widetilde{x}_{ij}^2 \widetilde{x}_{im} \widetilde{x}_{ik}]}_{\substack{\ell=j, m \neq j, m \neq k \\ (a122)}} + \underbrace{\beta_k^0 \beta_j^0 \mathbf{E} [\widetilde{x}_{ij}^2 \widetilde{x}_{ik}^2]}_{\substack{\ell \neq j, m=j, \ell=k \\ (a211)}} \\
 &+ \underbrace{\sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_{\ell}^0 \beta_j^0 \mathbf{E} [\widetilde{x}_{ij}^2 \widetilde{x}_{i\ell} \widetilde{x}_{ik}]}_{\substack{\ell \neq j, m=j, \ell \neq k \\ (a212)}} + \underbrace{\beta_k^{02} \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{ik}^3]}_{\substack{\ell \neq j, m \neq j, m=k, \ell=k \\ (a2211)}} + \underbrace{\sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_{\ell}^0 \beta_k^0 \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{i\ell} \widetilde{x}_{ik}^2]}_{\substack{\ell \neq j, m \neq j, m=k, \ell \neq k \\ (a2212)}} + \underbrace{\sum_{\substack{m \neq j \\ m \neq k}}^d \beta_k^0 \beta_m^0 \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{im} \widetilde{x}_{ik}^2]}_{\substack{\ell \neq j, m \neq j, m \neq k, \ell=k \\ (a2221)}} \\
 &+ \underbrace{\sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_{\ell}^{02} \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{i\ell}^2 \widetilde{x}_{ik}]}_{\substack{\ell \neq j, m \neq j, m \neq k, \ell \neq k, m=\ell \\ (a22221)}} + \underbrace{\sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \sum_{\substack{m \neq j \\ m \neq k \\ m \neq \ell}}^d \beta_{\ell}^0 \beta_m^0 \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{i\ell} \widetilde{x}_{im} \widetilde{x}_{ik}]}_{\substack{\ell \neq j, m \neq j, m \neq k, \ell \neq k, m \neq \ell \\ (a22222)}}. \tag{B.5}
 \end{aligned}$$

Now we consider the case when  $k \neq j$ , that is, the off-diagonals of (a). Then all terms in the summands are independent and we can calculate the expectation with respect to the

quantizers as:

$$\begin{aligned}
 & \underbrace{\sum_{\ell=1}^d \sum_{m=1}^d \beta_{\ell}^0 \beta_m^0 \mathbf{E}_Q [\widetilde{x}_{ij} \widetilde{x}_{i\ell} \widetilde{x}_{im} \widetilde{x}_{ik}]}_{(a)} \\
 &= \underbrace{R^2 \beta_j^{0^2} x_{ij} x_{ik}}_{\substack{\ell=j, m=j \\ (a11)}} + \underbrace{R^4 \beta_j^0 \beta_k^0}_{\substack{\ell=j, m \neq j, m=k \\ (a121)}} + \underbrace{R^2 \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_j^0 \beta_m^0 x_{im} x_{ik}}_{\substack{\ell=j, m \neq j, m \neq k \\ (a122)}} + \underbrace{R^4 \beta_k^0 \beta_j^0}_{\substack{\ell \neq j, m=j, \ell=k \\ (a211)}} \\
 &+ \underbrace{R^2 \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_{\ell}^0 \beta_j^0 x_{i\ell} x_{ik}}_{\substack{\ell \neq j, m=j, \ell \neq k \\ (a212)}} + \underbrace{R^2 \beta_k^{0^2} x_{ij} x_{ik}}_{\substack{\ell \neq j, m \neq j, m=k, \ell=k \\ (a2211)}} + \underbrace{R^2 \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_{\ell}^0 \beta_k^0 x_{ij} x_{i\ell}}_{\substack{\ell \neq j, m \neq j, m=k, \ell \neq k \\ (a2212)}} + \underbrace{R^2 \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_k^0 \beta_m^0 x_{ij} x_{im}}_{\substack{\ell \neq j, m \neq j, m \neq k, \ell=k \\ (a2221)}} \\
 &+ \underbrace{R^2 \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_{\ell}^{0^2} x_{ij} x_{ik}}_{\substack{\ell \neq j, m \neq j, m \neq k, \ell \neq k, m=\ell \\ (a22221)}} + \underbrace{\sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \sum_{\substack{m \neq j \\ m \neq k \\ m \neq \ell}}^d \beta_{\ell}^0 \beta_m^0 x_{ij} x_{i\ell} x_{im} x_{ik}}_{\substack{\ell \neq j, m \neq j, m \neq k, \ell \neq k, m \neq \ell \\ (a22222)}} \\
 &= R^2 \left( x_{ij} x_{ik} \left( \beta_j^{0^2} + \beta_k^{0^2} + \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_m^{0^2} \right) + 2 \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_j^0 \beta_m^0 x_{im} x_{ik} + 2 \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_{\ell}^0 \beta_k^0 x_{ij} x_{i\ell} \right) \\
 &+ R^4 \beta_k^0 \beta_j^0 + x_{ij} x_{ik} \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \sum_{\substack{m \neq j \\ m \neq k \\ m \neq \ell}}^d \beta_{\ell}^0 \beta_m^0 x_{i\ell} x_{im} \\
 &= R^4 \beta_k^0 \beta_j^0 + R^2 \left( x_{ij} x_{ik} \|\beta^0\|_2^2 + 2 \left( \beta_j^0 x_{ik} + \beta_k^0 x_{ij} \right) \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_{\ell}^0 x_{i\ell} \right) + x_{ij} x_{ik} \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_{\ell}^0 \beta_m^0 x_{i\ell} x_{im} \quad (\text{B.6})
 \end{aligned}$$

When we let  $k = j$ , then the result becomes

$$\begin{aligned}
 & \underbrace{\sum_{\ell=1}^d \sum_{m=1}^d \beta_{\ell}^0 \beta_m^0 \mathbf{E} [\widetilde{x}_{ij} \widetilde{x}_{i\ell} \widetilde{x}_{im} \widetilde{x}_{ik}]}_{(a)} \\
 &= \underbrace{R^4 \beta_j^{02}}_{\substack{\ell=j, m=j \\ (a11)}} + \underbrace{R^4 \beta_j^{02}}_{\substack{\ell=j, m \neq j, m=k \\ (a121)}} + R^2 \underbrace{\sum_{\substack{m \neq j \\ m \neq k}}^d \beta_j^0 \beta_m^0 x_{ij} x_{im}}_{\substack{\ell=j, m \neq j, m \neq k \\ (a122)}} + \underbrace{R^4 \beta_j^{02}}_{\substack{\ell \neq j, m=j, \ell=k \\ (a211)}} \\
 &+ R^2 \underbrace{\sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_{\ell}^0 \beta_j^0 x_{i\ell} x_{ij}}_{\substack{\ell \neq j, m=j, \ell \neq k \\ (a212)}} + \underbrace{R^4 \beta_j^{02}}_{\substack{\ell \neq j, m \neq j, m=k, \ell=k \\ (a2211)}} + R^2 \underbrace{\sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_{\ell}^0 \beta_j^0 x_{ij} x_{i\ell}}_{\substack{\ell \neq j, m \neq j, m=k, \ell \neq k \\ (a2212)}} + R^2 \underbrace{\sum_{\substack{m \neq j \\ m \neq k}}^d \beta_j^0 \beta_m^0 x_{ij} x_{im}}_{\substack{\ell \neq j, m \neq j, m \neq k, \ell=k \\ (a2221)}} \\
 &+ \underbrace{R^4 \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_{\ell}^0}_{\substack{\ell \neq j, m \neq j, m \neq k, \ell \neq k, m=\ell \\ (a22221)}} + R^2 \underbrace{\sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \sum_{\substack{m \neq j \\ m \neq k \\ m \neq \ell}}^d \beta_{\ell}^0 \beta_m^0 x_{i\ell} x_{im}}_{\substack{\ell \neq j, m \neq j, m \neq k, \ell \neq k, m \neq \ell \\ (a22222)}} \\
 &= R^4 \left( 4\beta_j^{02} + \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_m^{02} \right) + 4R^2 \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_j^0 \beta_m^0 x_{ij} x_{im} + R^2 \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_{\ell}^0 \beta_m^0 x_{i\ell} x_{im} \\
 &= R^4 \left( 2\beta_j^{02} + \|\beta^0\|_2^2 \right) + 4R^2 \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_j^0 \beta_m^0 x_{ij} x_{im} + R^2 \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_{\ell}^0 \beta_m^0 x_{i\ell} x_{im} \tag{B.7}
 \end{aligned}$$

### B.1.2 Terms (b) and (c)

We now move to examining terms (b) and (c). We note that when  $k = j$  that these two terms are identical. Thus, we will derive their formulas together.

$$\underbrace{\beta_j^0 \mathbf{E}_Q [\delta_j] \sum_{m=1}^d \beta_m^0 \mathbf{E}_Q [\widetilde{x}_{im} \widetilde{x}_{ik}]}_{(b)} \quad \underbrace{\mathbf{E}_Q [\delta_k] \beta_k^0 \sum_{m=1}^d \beta_m^0 \mathbf{E} [\widetilde{x}_{im} \widetilde{x}_{ij}]}_{(c)} \quad (\text{B.8})$$

For both (b) and (c), we only have to compare  $m$  with  $k$  and  $m$  with  $j$ . For term (b)

$$\begin{aligned} \underbrace{\beta_j^0 \mathbf{E}_Q [\delta_j] \sum_{m=1}^d \beta_m^0 \mathbf{E}_Q [\widetilde{x}_{im} \widetilde{x}_{ik}]}_{(b)} &= \underbrace{\beta_j^0 \mathbf{E}_Q [\delta_j] \beta_k^0 \mathbf{E}_Q [\widetilde{x}_{ik}^2]}_{m=k} + \underbrace{\beta_j^0 \mathbf{E}_Q [\delta_j] \sum_{m \neq k}^d \beta_m^0 \mathbf{E}_Q [\widetilde{x}_{im} \widetilde{x}_{ik}]}_{m \neq k} \\ &= \beta_j^0 \beta_k^0 R^2 \mathbf{E}_Q [\delta_j] + \beta_j^0 \mathbf{E}_Q [\delta_j] \sum_{m \neq k}^d \beta_m^0 \mathbf{E}_Q [\widetilde{x}_{im} \widetilde{x}_{ik}]. \end{aligned} \quad (\text{B.9})$$

Similarly for term (c):

$$\begin{aligned} \underbrace{\mathbf{E}_Q [\delta_k] \beta_k^0 \sum_{m=1}^d \beta_m^0 \mathbf{E} [\widetilde{x}_{im} \widetilde{x}_{ij}]}_{(c)} &= \underbrace{\mathbf{E}_Q [\delta_k] \beta_k^0 \beta_j^0 \mathbf{E} [\widetilde{x}_{ij}^2]}_{m=j} + \underbrace{\mathbf{E}_Q [\delta_k] \beta_k^0 \sum_{m \neq j}^d \beta_m^0 \mathbf{E} [\widetilde{x}_{im} \widetilde{x}_{ij}]}_{m \neq j} \\ &= \beta_k^0 \beta_j^0 R^2 \mathbf{E}_Q [\delta_k] + \beta_k^0 \mathbf{E}_Q [\delta_k] \sum_{m \neq j}^d \beta_m^0 \mathbf{E} [\widetilde{x}_{im} \widetilde{x}_{ij}] \end{aligned} \quad (\text{B.10})$$

When  $k \neq j$ , and we take the expectation with respect to the quantizer, then we have for term (b)

$$\begin{aligned} \underbrace{\beta_j^0 \mathbf{E}_Q [\delta_j] \sum_{m=1}^d \beta_m^0 \mathbf{E}_Q [\widetilde{x}_{im} \widetilde{x}_{ik}]}_{(b)} &= \beta_j^0 \beta_k^0 R^2 \mathbf{E}_Q [\delta_j] + \beta_j^0 \mathbf{E}_Q [\delta_j] \sum_{m \neq k}^d \beta_m^0 \mathbf{E}_Q [\widetilde{x}_{im} \widetilde{x}_{ik}] \\ &= \beta_j^0 (x_{ij}^2 - R) \left( \beta_k^0 R^2 + \sum_{m \neq k}^d \beta_m^0 x_{im} x_{ik} \right). \end{aligned} \quad (\text{B.11})$$

and for term (c):

$$\begin{aligned} \underbrace{\mathbf{E}_Q [\delta_k] \beta_k^0 \sum_{m=1}^d \beta_m^0 \mathbf{E} [\widetilde{x}_{im} \widetilde{x}_{ij}]}_{(c)} &= \beta_k^0 \beta_j^0 R^2 \mathbf{E}_Q [\delta_k] + \beta_k^0 \mathbf{E}_Q [\delta_k] \sum_{m \neq j}^d \beta_m^0 \mathbf{E} [\widetilde{x}_{im} \widetilde{x}_{ij}] \\ &= \beta_k^0 (x_{ik}^2 - R^2) \left( \beta_j^0 R^2 + \sum_{m \neq j}^d \beta_m^0 x_{im} x_{ij} \right) \end{aligned} \quad (\text{B.12})$$

When  $k = j$ , then both terms (b) and (c) become:

$$\begin{aligned} \underbrace{\beta_j^0 \mathbf{E}_Q [\delta_j] \sum_{m=1}^d \beta_m^0 \mathbf{E}_Q [\widetilde{x}_{im} \widetilde{x}_{ik}]}_{(b),(c)} &= \beta_j^{0^2} R^2 \mathbf{E}_Q [\delta_j] + \beta_j^0 \mathbf{E}_Q [\delta_j] \sum_{m \neq j}^d \beta_m^0 \mathbf{E}_Q [\widetilde{x}_{im} \widetilde{x}_{ij}] \\ &= \beta_j^0 (x_{ij}^2 - R^2) \left( \beta_j^0 R^2 + \sum_{m \neq j}^d \beta_m^0 x_{im} x_{ij} \right) \end{aligned} \quad (\text{B.13})$$

### B.1.3 Term (d)

Now we examine term (d), which requires us to compare  $m$  with both  $j$  and  $k$ .

$$\begin{aligned}
 \underbrace{2 \sum_{m=1}^d \beta_m^0 \mathbf{E}_Q [\widetilde{x}_{im} \widetilde{x}_{ij} \widetilde{x}_{ik}]}_{(d)} &= \underbrace{2\beta_j^0 \mathbf{E}_Q [\widetilde{x}_{ij}^2 \widetilde{x}_{ik}]}_{m=j} + \underbrace{2 \sum_{m \neq j}^d \beta_m^0 \mathbf{E}_Q [\widetilde{x}_{im} \widetilde{x}_{ij} \widetilde{x}_{ik}]}_{m \neq j} \\
 &= \underbrace{2\beta_j^0 \mathbf{E}_Q [\widetilde{x}_{ij}^2 \widetilde{x}_{ik}]}_{m=j} + \underbrace{2\beta_k^0 \mathbf{E}_Q [\widetilde{x}_{ij} \widetilde{x}_{ik}^2]}_{m \neq j, m=k} + \underbrace{2 \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_m^0 \mathbf{E}_Q [\widetilde{x}_{im} \widetilde{x}_{ij} \widetilde{x}_{ik}]}_{m \neq j, m \neq k}. \tag{B.14}
 \end{aligned}$$

Then when  $k \neq j$ , term (d) taken with respect to the quantizer becomes

$$\underbrace{2 \sum_{m=1}^d \beta_m^0 \mathbf{E}_Q [\widetilde{x}_{im} \widetilde{x}_{ij} \widetilde{x}_{ik}]}_{(d)} = 2\beta_j^0 R^2 x_{ik} + 2\beta_k^0 R^2 x_{ij} + 2x_{ij} x_{ik} \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_m^0 x_{im}. \tag{B.15}$$

Then when  $k = j$ , term (d) evaluates to

$$\begin{aligned}
 \underbrace{2 \sum_{m=1}^d \beta_m^0 \mathbf{E}_Q [\widetilde{x}_{im} \widetilde{x}_{ij} \widetilde{x}_{ik}]}_{(d)} &= \underbrace{2\beta_j^0 \mathbf{E}_Q [\widetilde{x}_{ij}^3]}_{m=j} + \underbrace{2\beta_k^0 \mathbf{E}_Q [\widetilde{x}_{ij}^3]}_{m \neq j, m=k} + \underbrace{2 \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_m^0 \mathbf{E}_Q [\widetilde{x}_{im} \widetilde{x}_{ij}^2]}_{m \neq j, m \neq k} \\
 &= 2\beta_j^0 R^2 x_{ij} + 2\beta_k^0 R^2 x_{ij} + 2R^2 x_{ij} \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_m^0 \\
 &= 2R^2 x_{ij} \sum_{m=1}^d \beta_m^0. \tag{B.16}
 \end{aligned}$$

### B.1.4 Combining Terms and Bounding

Now we have the components necessary to calculate the  $j, k$ th element of of  $\mathbf{E} [\widetilde{\psi}_{\beta^0} \widetilde{\psi}_{\beta^0}^T]$ .

We begin with the case when  $k \neq j$ . Taking the expectation with respect to the quantizer and combining the components yields:

$$\begin{aligned}
 \mathbf{E}_Q \left[ \left( \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right)_{jk} \right] &= \left( x_{ij}^2 - R^2 \right) \beta_j^0 \beta_k^0 \left( x_{ik}^2 - R^2 \right) + B^2 x_{ij} x_{ik} \\
 &+ \underbrace{R^4 \beta_k^0 \beta_j^0 + R^2 \left( x_{ij} x_{ik} \left\| \beta^0 \right\|_2^2 + 2 \left( \beta_j^0 x_{ik} + \beta_k^0 x_{ij} \right) \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_\ell^0 x_{i\ell} \right) + x_{ij} x_{ik} \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_\ell^0 \beta_m^0 x_{i\ell} x_{im}}}_{(a)} \\
 &+ \underbrace{\beta_j^0 \left( x_{ij}^2 - R^2 \right) \left( \beta_k^0 R^2 + \sum_{m \neq k}^d \beta_m^0 x_{im} x_{ik} \right)}_{(b)} + \underbrace{\beta_k^0 \left( x_{ik}^2 - R^2 \right) \left( \beta_j^0 R^2 + \sum_{m \neq j}^d \beta_m^0 x_{im} x_{ij} \right)}_{(c)} \\
 &- Y_i \left( \underbrace{2\beta_j^0 R^2 x_{ik} + 2\beta_k^0 R^2 x_{ij} + 2x_{ij} x_{ik} \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_m^0 x_{im}}_{(d)} + \beta_j^0 \left( x_{ij}^2 - R^2 \right) x_{ik} + \beta_k^0 \left( x_{ik}^2 - R^2 \right) x_{ij} \right) \\
 &= \beta_j^0 \beta_k^0 x_{ij}^2 x_{ik}^2 - R^2 \beta_j^0 \beta_k^0 x_{ik}^2 - R^2 \beta_j^0 \beta_k^0 x_{ij}^2 + R^4 \beta_j^0 \beta_k^0 + B^2 x_{ij} x_{ik} + R^4 \beta_k^0 \beta_j^0 + R^2 x_{ij} x_{ik} \left\| \beta^0 \right\|_2^2 \\
 &+ 2R^2 \beta_j^0 x_{ik} \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_\ell^0 x_{i\ell} + 2R^2 \beta_k^0 x_{ij} \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_\ell^0 x_{i\ell} + x_{ij} x_{ik} \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_\ell^0 \beta_m^0 x_{i\ell} x_{im} \\
 &+ R^2 \beta_j^0 \beta_k^0 x_{ij}^2 + \beta_j^0 x_{ij}^2 \sum_{m \neq k}^d \beta_m^0 x_{im} x_{ik} - \beta_j^0 \beta_k^0 R^4 - \beta_j^0 R^2 \sum_{m \neq k}^d \beta_m^0 x_{im} x_{ik} \\
 &+ R^2 \beta_k^0 \beta_j^0 x_{ik}^2 + \beta_k^0 x_{ik}^2 \sum_{m \neq j}^d \beta_m^0 x_{im} x_{ij} - \beta_k^0 \beta_j^0 R^4 - \beta_k^0 R^2 \sum_{m \neq j}^d \beta_m^0 x_{im} x_{ij} \\
 &- Y_i \left( 2\beta_j^0 R^2 x_{ik} + 2\beta_k^0 R^2 x_{ij} + 2x_{ij} x_{ik} \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_m^0 x_{im} + \beta_j^0 x_{ij}^2 x_{ik} - \beta_j^0 x_{ik} R^2 + \beta_k^0 x_{ik}^2 x_{ij} - \beta_k^0 x_{ij} R^2 \right)
 \end{aligned} \tag{B.17}$$

We use colors here to identify like terms. Now we combine like terms and reduce, while also taking the expectation  $\mathbf{E}_{\mathbf{y}} [Y_i] = \mathbf{x}_i^T \boldsymbol{\beta}^0$ , where the expectation is taken with respect to the variance in the error term in  $Y_i$ .

$$\begin{aligned}
 \mathbf{E}_{Q,\mathbf{y}} \left[ \left( \tilde{\boldsymbol{\psi}}_{\beta^0} \tilde{\boldsymbol{\psi}}_{\beta^0}^T \right)_{jk} \right] &= \beta_j^0 \beta_k^0 x_{ij}^2 x_{ik}^2 + B^2 x_{ij} x_{ik} + R^2 x_{ij} x_{ik} \left\| \boldsymbol{\beta}^0 \right\|_2^2 + R^2 \beta_j^0 x_{ik} \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_\ell^0 x_{i\ell} \\
 &+ R^2 \beta_k^0 x_{ij} \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_\ell^0 x_{i\ell} + x_{ij} x_{ik} \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_\ell^0 \beta_m^0 x_{i\ell} x_{im} + \beta_j^0 x_{ij}^2 \sum_{m \neq k}^d \beta_m^0 x_{im} x_{ik} + \beta_k^0 x_{ik}^2 \sum_{m \neq j}^d \beta_m^0 x_{im} x_{ij} \\
 &- \left( \sum_{\ell=1}^d x_{i\ell} \beta_\ell^0 \right) \left( R^2 \beta_j^0 x_{ik} + R^2 \beta_k^0 x_{ij} + 2x_{ij} x_{ik} \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_m^0 x_{im} + \beta_j^0 x_{ij}^2 x_{ik} + \beta_k^0 x_{ik}^2 x_{ij} \right) \quad (\text{B.18})
 \end{aligned}$$

When  $k = j$  the calculation becomes:

$$\begin{aligned}
 \mathbf{E}_{Q,\mathbf{y}} \left[ \left( \tilde{\boldsymbol{\psi}}_{\beta^0} \tilde{\boldsymbol{\psi}}_{\beta^0}^T \right)_{jj} \right] &= \beta_j^0 \beta_j^0 x_{ij}^2 x_{ij}^2 + B^2 x_{ij} x_{ij} + R^2 x_{ij} x_{ij} \left\| \boldsymbol{\beta}^0 \right\|_2^2 + R^2 \beta_j^0 x_{ij} \sum_{\substack{\ell \neq j \\ \ell \neq j}}^d \beta_\ell^0 x_{i\ell} \\
 &+ R^2 \beta_j^0 x_{ij} \sum_{\substack{\ell \neq j \\ \ell \neq j}}^d \beta_\ell^0 x_{i\ell} + x_{ij} x_{ij} \sum_{\substack{\ell \neq j \\ \ell \neq j}}^d \sum_{\substack{m \neq j \\ m \neq j}}^d \beta_\ell^0 \beta_m^0 x_{i\ell} x_{im} + \beta_j^0 x_{ij}^2 \sum_{m \neq j}^d \beta_m^0 x_{im} x_{ij} + \beta_j^0 x_{ij}^2 \sum_{m \neq j}^d \beta_m^0 x_{im} x_{ij} \\
 &- \left( \sum_{\ell=1}^d x_{i\ell} \beta_\ell^0 \right) \left( R^2 \beta_j^0 x_{ij} + R^2 \beta_j^0 x_{ij} + 2x_{ij} x_{ij} \sum_{\substack{m \neq j \\ m \neq j}}^d \beta_m^0 x_{im} + \beta_j^0 x_{ij}^2 x_{ij} + \beta_j^0 x_{ij}^2 x_{ij} \right) \\
 &= \beta_j^0^2 x_{ij}^4 + B^2 x_{ij}^2 + R^2 x_{ij}^2 \left\| \boldsymbol{\beta}^0 \right\|_2^2 + 2R^2 \beta_j^0 x_{ij} \sum_{\substack{\ell \neq j \\ \ell \neq j}}^d \beta_\ell^0 x_{i\ell} + x_{ij}^2 \sum_{\substack{\ell \neq j \\ \ell \neq j}}^d \sum_{\substack{m \neq j \\ m \neq j}}^d \beta_\ell^0 \beta_m^0 x_{i\ell} x_{im} \\
 &+ 2\beta_j^0 x_{ij}^2 \sum_{m \neq j}^d \beta_m^0 x_{im} x_{ij} - \left( \sum_{\ell=1}^d x_{i\ell} \beta_\ell^0 \right) \left( 2R^2 \beta_j^0 x_{ij} + 2\beta_j^0 x_{ij}^3 + 2x_{ij}^2 \sum_{\substack{m \neq j \\ m \neq j}}^d \beta_m^0 x_{im} \right) \quad (\text{B.19})
 \end{aligned}$$

Now we wish to bound the entries of  $\mathbf{E} \left[ \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right]$  using these formulas and the boundedness of the entries of  $\mathbf{X}$ . For the  $k \neq j$  case:

$$\begin{aligned}
 \mathbf{E}_{Q,\mathbf{y}} \left[ \left( \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right)_{jk} \right] &= \beta_j^0 \beta_k^0 x_{ij}^2 x_{ik}^2 + B^2 x_{ij} x_{ik} + R^2 x_{ij} x_{ik} \left\| \boldsymbol{\beta}^0 \right\|_2^2 + R^2 \beta_j^0 x_{ik} \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_\ell^0 x_{i\ell} \\
 &+ R^2 \beta_k^0 x_{ij} \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \beta_\ell^0 x_{i\ell} + x_{ij} x_{ik} \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_\ell^0 \beta_m^0 x_{i\ell} x_{im} + \beta_j^0 x_{ij}^2 \sum_{m \neq k}^d \beta_m^0 x_{im} x_{ik} + \beta_k^0 x_{ik}^2 \sum_{m \neq j}^d \beta_m^0 x_{im} x_{ij} \\
 &- \left( \sum_{\ell=1}^d x_{i\ell} \beta_\ell^0 \right) \left( R^2 \beta_j^0 x_{ik} + R^2 \beta_k^0 x_{ij} + 2x_{ij} x_{ik} \sum_{\substack{m \neq j \\ m \neq k}}^d \beta_m^0 x_{im} + \beta_j^0 x_{ij}^2 x_{ik} + \beta_k^0 x_{ik}^2 x_{ij} \right) \\
 &\leq R^4 \left| \beta_j^0 \right| \left| \beta_k^0 \right| + B^2 R^2 + R^4 \left\| \boldsymbol{\beta}^0 \right\|_2^2 + R^4 \left| \beta_j^0 \right| \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \left| \beta_\ell^0 \right| + R^4 \left| \beta_k^0 \right| \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \left| \beta_\ell^0 \right| + R^4 \sum_{\substack{\ell \neq j \\ \ell \neq k}}^d \sum_{\substack{m \neq j \\ m \neq k}}^d \left| \beta_\ell^0 \right| \left| \beta_m^0 \right| \\
 &+ R^4 \left| \beta_j^0 \right| \sum_{m \neq k}^d \left| \beta_m^0 \right| + R^4 \left| \beta_k^0 \right| \sum_{m \neq j}^d \left| \beta_m^0 \right| \\
 &+ R \left\| \boldsymbol{\beta}^0 \right\|_1 \left( R^3 \left| \beta_j^0 \right| + R^3 \left| \beta_k^0 \right| + 2R^3 \sum_{\substack{m \neq j \\ m \neq k}}^d \left| \beta_m^0 \right| + R^3 \left| \beta_j^0 \right| + R^3 \left| \beta_k^0 \right| \right) \\
 &\leq B^2 R^2 + R^4 \left( \left\| \boldsymbol{\beta}^0 \right\|_2^2 + 2 \left( \left| \beta_j^0 \right| + \left| \beta_k^0 \right| \right) \left\| \boldsymbol{\beta}^0 \right\|_1 + 2 \left\| \boldsymbol{\beta}^0 \right\|_1^2 \right) \\
 &\leq B^2 R^2 + 7R^4 \left\| \boldsymbol{\beta}^0 \right\|_1^2 \tag{B.20}
 \end{aligned}$$

Now when  $k = j$ , our upper bound becomes:

$$\begin{aligned}
 \mathbf{E}_{Q,\mathbf{y}} \left[ \left( \tilde{\psi}_{\beta^0} \tilde{\psi}_{\beta^0}^T \right)_{jj} \right] &= \beta_j^{0^2} x_{ij}^4 + B^2 x_{ij}^2 + R^2 x_{ij}^2 \left\| \boldsymbol{\beta}^0 \right\|_2^2 + 2R^2 \beta_j^0 x_{ij} \sum_{\substack{\ell \neq j \\ \ell \neq j}}^d \beta_\ell^0 x_{i\ell} \\
 &+ x_{ij}^2 \sum_{\substack{\ell \neq j \\ \ell \neq j}}^d \sum_{\substack{m \neq j \\ m \neq j}}^d \beta_\ell^0 \beta_m^0 x_{i\ell} x_{im} + 2\beta_j^0 x_{ij}^2 \sum_{m \neq j}^d \beta_m^0 x_{im} x_{ij} \\
 &- \left( \sum_{\ell=1}^d x_{i\ell} \beta_\ell^0 \right) \left( 2R^2 \beta_j^0 x_{ij} + 2\beta_j^0 x_{ij}^3 + 2x_{ij}^2 \sum_{\substack{m \neq j \\ m \neq j}}^d \beta_m^0 x_{im} \right) \\
 &\leq B^2 R^2 + R^4 \beta_j^{0^2} + R^4 \left\| \boldsymbol{\beta}^0 \right\|_2^2 + 2R^4 \left| \beta_j^0 \right| \sum_{\ell \neq j}^d \left| \beta_\ell^0 \right| + R^4 \sum_{\ell \neq j}^d \sum_{m \neq j}^d \left| \beta_\ell^0 \right| \left| \beta_m^0 \right| + 2R^4 \left| \beta_j^0 \right| \sum_{m \neq j}^d \left| \beta_m^0 \right| \\
 &+ R \left\| \boldsymbol{\beta}^0 \right\|_1 \left( 2R^3 \left| \beta_j^0 \right| + 2R^3 \left| \beta_j^0 \right| + 2R^3 \sum_{m \neq j}^d \left| \beta_m^0 \right| \right) \\
 &\leq B^2 R^2 + R^4 \left( \left\| \boldsymbol{\beta}^0 \right\|_2^2 + 4 \left| \beta_j^0 \right| \left\| \boldsymbol{\beta}^0 \right\|_1 + 2 \left\| \boldsymbol{\beta}^0 \right\|_1^2 \right) \\
 &\leq B^2 R^2 + 7R^4 \left\| \boldsymbol{\beta}^0 \right\|_1^2 \tag{B.21}
 \end{aligned}$$

## Appendix C: Supporting Work for Chapter 4

### C.1 Quantized Scenario

#### C.1.1 Quantized Scenario with Fixed $\mathbf{Z}$

Unbiasedness of the estimator  $\tilde{\Sigma}$

$$\begin{aligned}
 \mathbf{E} [\tilde{\Sigma}] &= \frac{1}{n} \sum_{i=1}^n \mathbf{E} [\tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T + \Delta_i] \\
 &= \frac{1}{n} \sum_{i=1}^n \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T + \text{diag} \left( \tilde{z}_{ij}^2 - z_{ij}^2 \right)_{j=1}^d \right] \\
 &= \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} \mathbf{E} [z_{i1}^2] & \mathbf{E} [\tilde{z}_{i1} \tilde{z}_{i2}] & \dots & \mathbf{E} [\tilde{z}_{i1} \tilde{z}_{id}] \\ \mathbf{E} [\tilde{z}_{i2} \tilde{z}_{i1}] & \mathbf{E} [z_{i2}^2] & \dots & \mathbf{E} [\tilde{z}_{i2} \tilde{z}_{id}] \\ \vdots & \vdots & \ddots & \dots \\ \mathbf{E} [\tilde{z}_{id} \tilde{z}_{i1}] & \mathbf{E} [\tilde{z}_{id} \tilde{z}_{i2}] & \dots & \mathbf{E} [z_{id}^2] \end{bmatrix} \\
 &= \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} z_{i1}^2 & z_{i1} z_{i2} & \dots & z_{i1} z_{id} \\ z_{i2} z_{i1} & z_{i2}^2 & \dots & z_{i2} z_{id} \\ \vdots & \vdots & \ddots & \dots \\ z_{id} z_{i1} & z_{id} z_{i2} & \dots & z_{id}^2 \end{bmatrix} \\
 &= \frac{\mathbf{Z}^T \mathbf{Z}}{n} \\
 &= \Sigma
 \end{aligned} \tag{C.1}$$

Unbiasedness of the estimator  $\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}}$

$$\begin{aligned}
 \mathbf{E} \left[ \tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} \right] &= \frac{1}{n} \sum_{i=1}^n \mathbf{E} \left[ \tilde{z}_{ij} \tilde{w}_i \right] \quad \text{for } 1 \leq j \leq d \\
 &= \frac{1}{n} \sum_{i=1}^n \mathbf{E} \left[ \tilde{z}_{ij} \right] \mathbf{E} \left[ \tilde{w}_i \right] \quad \text{for } 1 \leq j \leq d \\
 &= \frac{1}{n} \sum_{i=1}^n z_{ij} \mathbf{E}_\epsilon \left[ w_i \right] \quad \text{for } 1 \leq j \leq d \\
 &= \frac{1}{n} \sum_{i=1}^n z_{ij} \mathbf{E}_\epsilon \left[ \mathbf{z}_i \boldsymbol{\beta}^* + \sigma \boldsymbol{\epsilon} \right] \quad \text{for } 1 \leq j \leq d \\
 &= \frac{1}{n} \sum_{i=1}^n z_{ij} \mathbf{z}_i \boldsymbol{\beta}^* \quad \text{for } 1 \leq j \leq d \\
 &= \frac{\mathbf{Z}^T \mathbf{Z} \boldsymbol{\beta}^*}{n} \\
 &= \tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} \tag{C.2}
 \end{aligned}$$

### C.1.2 Quantized Scenario with Gaussian $\mathbf{Z}$

Unbiasedness of the estimator  $\tilde{\Sigma}$

$$\begin{aligned}
 \mathbf{E} \left[ \tilde{\Sigma} \right] &= \frac{1}{n} \sum_{i=1}^n \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T + \Delta_i \right] \\
 &= \frac{1}{n} \sum_{i=1}^n \mathbf{E} \left[ \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^T + \text{diag} \left( \tilde{z}_{ij}^2 - z_{ij}^2 \right)_{j=1}^d \right] \\
 &= \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} \mathbf{E} \left[ \tilde{z}_{i1}^2 \right] & \mathbf{E} \left[ \tilde{z}_{i1} \tilde{z}_{i2} \right] & \dots & \mathbf{E} \left[ \tilde{z}_{i1} \tilde{z}_{id} \right] \\ \mathbf{E} \left[ \tilde{z}_{i2} \tilde{z}_{i1} \right] & \mathbf{E} \left[ \tilde{z}_{i2}^2 \right] & \dots & \mathbf{E} \left[ \tilde{z}_{i2} \tilde{z}_{id} \right] \\ \vdots & \vdots & \ddots & \dots \\ \mathbf{E} \left[ \tilde{z}_{id} \tilde{z}_{i1} \right] & \mathbf{E} \left[ \tilde{z}_{id} \tilde{z}_{i2} \right] & \dots & \mathbf{E} \left[ \tilde{z}_{id}^2 \right] \end{bmatrix} \\
 &= \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} \mathbf{E}_{\mathbf{Z}} \left[ z_{i1}^2 \right] & \mathbf{E}_{\mathbf{Z}} \left[ z_{i1} z_{i2} \right] & \dots & \mathbf{E}_{\mathbf{Z}} \left[ z_{i1} z_{id} \right] \\ \mathbf{E}_{\mathbf{Z}} \left[ z_{i2} z_{i1} \right] & \mathbf{E}_{\mathbf{Z}} \left[ z_{i2}^2 \right] & \dots & \mathbf{E}_{\mathbf{Z}} \left[ z_{i2} z_{id} \right] \\ \vdots & \vdots & \ddots & \dots \\ \mathbf{E}_{\mathbf{Z}} \left[ z_{id} z_{i1} \right] & \mathbf{E}_{\mathbf{Z}} \left[ z_{id} z_{i2} \right] & \dots & \mathbf{E}_{\mathbf{Z}} \left[ z_{id}^2 \right] \end{bmatrix} \\
 &= \mathbf{E}_{\mathbf{Z}} \left[ \frac{\mathbf{Z}^T \mathbf{Z}}{n} \right] \\
 &= \Sigma
 \end{aligned} \tag{C.3}$$

Unbiasedness of the estimator  $\tilde{\Sigma}_{\mathbf{Z}\mathbf{w}}$

$$\begin{aligned}
 \mathbf{E} \left[ \tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} \right] &= \frac{1}{n} \sum_{i=1}^n \mathbf{E} \left[ \tilde{z}_{ij} \tilde{w}_i \right] \quad \text{for } 1 \leq j \leq d \\
 &= \frac{1}{n} \sum_{i=1}^n \mathbf{E} \left[ \tilde{z}_{ij} \right] \mathbf{E} \left[ \tilde{w}_i \right] \quad \text{for } 1 \leq j \leq d \\
 &= \frac{1}{n} \sum_{i=1}^n \mathbf{E}_{\mathbf{Z}} \left[ z_{ij} \right] \mathbf{E}_{\mathbf{Z}} \left[ w_i \right] \quad \text{for } 1 \leq j \leq d \\
 &= \mathbf{E}_{\mathbf{Z}} \left[ \frac{\mathbf{Z}^T \mathbf{Z} \boldsymbol{\beta}^*}{n} \right] \\
 &= \tilde{\Sigma}_{\mathbf{Z}\mathbf{w}} \tag{C.4}
 \end{aligned}$$

## Appendix D: General Supporting Ideas

### D.0.1 Bound on the Norm of the Difference of Matrices

While not an assumption, we will repeatedly use the fact that for conformable matrices  $\mathbf{A}$ ,  $\mathbf{B}$ , and  $\mathbf{C}$  where  $\mathbf{A} = \mathbf{B} - \mathbf{C}$  and  $\mathbf{A} \geq 0$  and  $\mathbf{C} \geq 0$ , we can show

$$\begin{aligned}
 \|\mathbf{A}\| &= \|\mathbf{B} - \mathbf{C}\| \\
 &= \max_{\|\hat{x}\|_2 \leq 1} \hat{x}^T \mathbf{B} \hat{x} - \hat{x}^T \mathbf{C} \hat{x} \\
 &\leq \max_{\|\hat{x}\|_2 \leq 1} \hat{x}^T \mathbf{B} \hat{x} - \min_{\|\hat{x}\|_2 \leq 1} \hat{x}^T \mathbf{C} \hat{x} \\
 &\leq \lambda_{\max}(\mathbf{B}) - \lambda_{\min}(\mathbf{C}) \\
 &\leq \|\mathbf{B}\|
 \end{aligned} \tag{D.1}$$

## Bibliography

- [1] A. Malekijoo, M. J. Fadaeieslam, H. Malekijou, M. Homayounfar, F. Alizadeh-Shabdiz, and R. Rawassizadeh, “FEDZIP: A Compression Framework for Communication-Efficient Federated Learning,” Feb. 2021. [Online]. Available: <http://arxiv.org/abs/2102.01593>.
- [2] Y. Oh, Y.-S. Jeon, M. Chen, and W. Saad, “FedVQCS: Federated Learning via Vector Quantized Compressed Sensing,” Apr. 2022. [Online]. Available: <http://arxiv.org/abs/2204.07692>.
- [3] M. Song, Z. Wang, Z. Zhang, *et al.*, “Analyzing User-Level Privacy Attack against Federated Learning,” *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 10, pp. 2430–2444, Oct. 2020, ISSN: 15580008. DOI: 10.1109/JSAC.2020.3000372.
- [4] Y. Xue, L. Su, and V. K. Lau, “FedOComp: Two-Timescale Online Gradient Compression for Over-the-Air Federated Learning,” *IEEE Internet of Things Journal*, vol. 9, no. 19, pp. 19 330–19 345, Oct. 2022, ISSN: 23274662. DOI: 10.1109/JIOT.2022.3165268.
- [5] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, “Robust and Communication-Efficient Federated Learning from Non-IID Data,” *arXiv:1903.02891v1*, Mar. 2019. [Online]. Available: <http://arxiv.org/abs/1903.02891>.
- [6] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in *20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, Feb. 2017. [Online]. Available: <http://arxiv.org/abs/1602.05629>.

- [7] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, “Federated Learning: Strategies for Improving Communication Efficiency,” Oct. 2017. [Online]. Available: <http://arxiv.org/abs/1610.05492>.
- [8] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, “1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs,” in *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech 2014*, International Speech and Communication Association, 2014, pp. 1058–1062. DOI: 10.21437/interspeech.2014-274.
- [9] N. Strom, “Scalable Distributed DNN Training Using Commodity GPU Cloud Computing,” in *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech 2015*, Dresden, Germany: International Speech Communication Association, Sep. 2015. DOI: 10.21437/Interspeech.2015.
- [10] D. Alistarh, D. Grubic, J. Z. Li, R. Tomioka, and M. Vojnovic, “QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding,” Tech. Rep., 2017.
- [11] A. F. Aji and K. Heafield, “Sparse Communication for Distributed Gradient Descent,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 440–445. DOI: 10.18653/v1/D17-1045. [Online]. Available: <http://dx.doi.org/10.18653/v1/D17-1045>.
- [12] W. Wen, C. Xu, F. Yan, *et al.*, “TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning,” in *31st Conference on Neural Information Processing Systems*, Long Beach, CA, Dec. 2017. [Online]. Available: <http://arxiv.org/abs/1705.07878>.
- [13] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, “Sparsified SGD with Memory,” Sep. 2018. [Online]. Available: <http://arxiv.org/abs/1809.07599>.

- [14] A. Albasyoni, M. Safaryan, L. Condat, and P. Richtárik, “Optimal Gradient Compression for Distributed and Federated Learning,” Oct. 2020. [Online]. Available: <http://arxiv.org/abs/2010.03246>.
- [15] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, “signSGD: Compressed Optimisation for Non-Convex Problems,” in *35th International Conference on Machine Learning*, Stockholm, Sweden, Feb. 2018. [Online]. Available: <http://arxiv.org/abs/1802.04434>.
- [16] Y. Tsuzuku, H. Imachi, and T. Akiba, “Variance-based Gradient Compression for Efficient Distributed Deep Learning,” Feb. 2018. [Online]. Available: <http://arxiv.org/abs/1802.06058>.
- [17] C.-Y. Chen, J. Choi, D. Brand, A. Agrawal, W. Zhang, and K. Gopalakrishnan, “AdaComp: Adaptive Residual Gradient Compression for Data-Parallel Distributed Training,” in *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, Yorktown Heights, 2018, pp. 2827–2835. [Online]. Available: [www.aaai.org](http://www.aaai.org).
- [18] P. Luo, F. R. Yu, J. Chen, J. Li, and V. C. Leung, “A Novel Adaptive Gradient Compression Scheme: Reducing the Communication Overhead for Distributed Deep Learning in the Internet of Things,” *IEEE Internet of Things Journal*, vol. 8, no. 14, pp. 11 476–11 486, Jul. 2021, ISSN: 23274662. DOI: 10.1109/JIOT.2021.3051611.
- [19] W. Yang, Y. Yang, X. Dang, H. Jiang, Y. Zhang, and W. Xiang, “A Novel Adaptive Gradient Compression Approach for Communication-Efficient Federated Learning,” in *Proceeding - 2021 China Automation Congress, CAC 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 674–678, ISBN: 9781665426473. DOI: 10.1109/CAC53003.2021.9728013.
- [20] H. Liu, F. He, and G. Cao, “Communication-Efficient Federated Learning for Heterogeneous Edge Devices Based on Adaptive Gradient Quantization,” Dec. 2022. [Online]. Available: <http://arxiv.org/abs/2212.08272>.

- [21] N. Mitchell, J. Ballé, Z. Charles, and J. Konečný, “Optimizing the Communication-Accuracy Trade-off in Federated Learning with Rate-Distortion Theory,” Jan. 2022. [Online]. Available: <http://arxiv.org/abs/2201.02664>.
- [22] L. Cui, X. Su, Y. Zhou, and L. Zhang, “ClusterGrad: Adaptive Gradient Compression by Clustering in Federated Learning,” in *2020 IEEE Global Communications Conference, GLOBECOM 2020 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, ISBN: 9781728182988. DOI: 10.1109/GLOBECOM42002.2020.9322527.
- [23] T. Tian, H. Shi, R. Ma, and Y. Liu, “FedACQ: adaptive clustering quantization of model parameters in federated learning,” *International Journal of Web Information Systems*, vol. 20, no. 1, pp. 88–110, Feb. 2024, ISSN: 17440092. DOI: 10.1108/IJWIS-08-2023-0128.
- [24] D. Rothchild, A. Panda, E. Ullah, *et al.*, “FetchSGD: Communication-Efficient Federated Learning with Sketching,” in *Proceedings of the 37th International Conference on Machine Learning*, H. Daume and A. Singh, Eds., PMLR, Jul. 2020, pp. 8253–8265. [Online]. Available: <https://proceedings.mlr.press/v119/rothchild20a.htm>.
- [25] L. Melas-Kyriazi and F. Wang, “Intrinsic Gradient Compression for Federated Learning,” Dec. 2021. [Online]. Available: <http://arxiv.org/abs/2112.02656>.
- [26] J. Hamer, M. Mohri, and A. T. Suresh, “FedBoost: Communication-Efficient Algorithms for Federated Learning,” in *Proceedings of the 37th International Conference on Machine Learning*, H. Daume and A. Singh, Eds., PMLR, Jul. 2020, pp. 3973–3983. [Online]. Available: <https://proceedings.mlr.press/v119/hamer20a.htm>.
- [27] L. Abrahamyan, Y. Chen, G. Bekoulis, and N. Deligiannis, “Learned Gradient Compression for Distributed Deep Learning,” Mar. 2021. [Online]. Available: <http://arxiv.org/abs/2103.08870>.
- [28] S. P. Karimireddy, Q. Rebjock, S. U. Stich, and M. Jaggi, “Error Feedback Fixes SignSGD and other Gradient Compression Schemes,” in *36th International Conference*

- on Machine Learning*, Long Beach, CA, 2019. [Online]. Available: <https://arxiv.org/abs/1901.09847>.
- [29] S. P. Lloyd, “Least Squares Quantization in PCM,” *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982. DOI: 10.1109/TIT.1982.1056489. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1056489>.
- [30] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, “EMNIST: an extension of MNIST to handwritten letters,” Feb. 2017. [Online]. Available: <http://arxiv.org/abs/1702.05373>.
- [31] S. Reddi, Z. Charles, M. Zaheer, *et al.*, “Adaptive Federated Optimization,” Feb. 2020. DOI: 10.48550/arxiv.2003.00295. [Online]. Available: <http://arxiv.org/abs/2003.00295>.
- [32] Y. Zhang, W. Lin, S. Chen, *et al.*, “Fed2Com: Towards Efficient Compression in Federated Learning,” in *2024 International Conference on Computing, Networking and Communications, ICNC 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 560–566, ISBN: 9798350370997. DOI: 10.1109/ICNC59896.2024.10556165.
- [33] H.-P. Wang, S. U. Stich, Y. He, and M. Fritz, “ProgFed: Effective, Communication, and Computation Efficient Federated Learning by Progressive Training,” in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., Baltimore, USA: PMLR, Jul. 2022, pp. 23 034–23 054. [Online]. Available: <https://proceedings.mlr.press/v162/wang22y.html>.
- [34] Y. Ji and L. Chen, “FedQNN: A Computation-Communication-Efficient Federated Learning Framework for IoT With Low-Bitwidth Neural Network Quantization,” *IEEE Internet of Things Journal*, vol. 10, no. 3, pp. 2494–2507, Feb. 2023, ISSN: 23274662. DOI: 10.1109/JIOT.2022.3213650.

- [35] Y. Wang, L. Lin, and J. Chen, “Communication-Efficient Adaptive Federated Learning,” in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., Baltimore, USA: PMLR, Jul. 2022, pp. 22 802–22 838. [Online]. Available: <https://proceedings.mlr.press/v162/wang22o.html>.
- [36] X. Li and P. Li, “Analysis of Error Feedback in Federated Non-Convex Optimization with Biased Compression: Fast Convergence and Partial Participation,” in *40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, B. Engelhardt, and J. Scarlett, Eds., Honolulu, Hawaii: PMLR, Jul. 2023, pp. 19 638–19 688. [Online]. Available: <https://proceedings.mlr.press/v202/li23o.html><https://proceedings.mlr.press/v202/li23o/li23o.pdf>.
- [37] X. Zhou, L. Chang, and J. Cao, “Communication-Efficient Nonconvex Federated Learning With Error Feedback for Uplink and Downlink,” *IEEE Transactions on Neural Networks and Learning Systems*, 2023, ISSN: 21622388. DOI: 10.1109/TNNLS.2023.3333804.
- [38] S. Shi, X. Chu, K. C. Cheung, and S. See, “Understanding Top-k Sparsification in Distributed Deep Learning,” Nov. 2019. [Online]. Available: <http://arxiv.org/abs/1911.08772>.
- [39] S. I. Young, W. Zhe, D. Taubman, and B. Girod, “Transform Quantization for CNN Compression,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, May 2021. DOI: 10.1109/TPAMI.2020. [Online]. Available: <https://www.ieee.org/publications/rights/index.html>.
- [40] S. Chen, C. Shen, L. Zhang, and Y. Tang, “Dynamic Aggregation for Heterogeneous Quantization in Federated Learning,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6804–6819, Oct. 2021, ISSN: 15582248. DOI: 10.1109/TWC.2021.3076613.

- [41] Y. Du, S. Yang, and K. Huang, “High-Dimensional Stochastic Gradient Quantization for Communication-Efficient Edge Learning,” *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1–5, 2019. DOI: 10.1109/GlobalSIP45357.2019.8969082. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8969082>.
- [42] R. S. Risuleo, G. Bottegal, and H. Hjalmarsson, “Identification of Linear Models from Quantized Data: A Midpoint-Projection Approach,” *IEEE Transactions on Automatic Control*, vol. 65, no. 7, pp. 2801–2813, Jul. 2020, ISSN: 15582523. DOI: 10.1109/TAC.2019.2933134.
- [43] L. Finessot and G. I. Kmcst, “A randomized EM-algorithm for estimating quantized linear Gaussian regression,” in *Proceedings of the 38th Conference on Decision & Control*, Phoenix, AZ: Conference on Decision & Control, Dec. 1999, pp. 5100–5101.
- [44] R. Saha, M. Pilanci, and A. J. Goldsmith, “Minimax Optimal Quantization of Linear Models: Information-Theoretic Limits and Efficient Algorithms,” Feb. 2022. [Online]. Available: <http://arxiv.org/abs/2202.11277>.
- [45] S. Dirksen, J. Maly, and H. Rauhut, “Covariance estimation under one-bit quantization,” Apr. 2021. [Online]. Available: <http://arxiv.org/abs/2104.01280>.
- [46] R. Vershynin, *High-Dimensional Probability*. Cambridge University Press, Sep. 2018, ISBN: 9781108231596. DOI: 10.1017/9781108231596.
- [47] O.-A. Maillard and R. Munos, “Compressed Least-Squares Regression,” in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds., Curran Associates, Inc., 2009, ISBN: 9781615679119. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2009/file/01882513d5fa7c329e940dda99b12147-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2009/file/01882513d5fa7c329e940dda99b12147-Paper.pdf).
- [48] A. Kabán, “New Bounds on Compressive Linear Least Squares Regression,” in *17th International Conference on Artificial Intelligence and Statistics*, Reykjavik, Iceland:

- JMLR: W&CP, 2014. [Online]. Available: [https://proceedings.mlr.press/v33/kaban14.html?utm\\_source=chatgpt.com](https://proceedings.mlr.press/v33/kaban14.html?utm_source=chatgpt.com).
- [49] M. Slawski, “Compressed Least Squares Regression revisited,” in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, 2017.
- [50] J.-C. Jong and S. Kotz, “On a Relation between Principal Components and Regression Analysis,” *The American Statistician*, vol. 53, no. 4, pp. 349–351, Nov. 1999, ISSN: 0003-1305. DOI: 10.1080/00031305.1999.10474488.
- [51] S. Boneh and G. R. Mendieta, “Variable selection in regression models using principal components,” *Communications in Statistics - Theory and Methods*, vol. 23, no. 1, pp. 197–213, Jan. 1994, ISSN: 0361-0926. DOI: 10.1080/03610929408831247.
- [52] I. T. Jolliffe, “Principal Components in Regression Analysis,” in 1986, pp. 129–155. DOI: 10.1007/978-1-4757-1904-8{\\_}8.
- [53] S. Maitra and J. Yan, “Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression,” *Casualty Actuarial Society, Tech. Rep.*, 2008, pp. 79–90.
- [54] F. S. Kurnaz, “Some Regression Methods Based on Principal Components,” *Gümüşhane Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, Jul. 2020, ISSN: 2146-538X. DOI: 10.17714/gumusfenbil.641791.
- [55] W. B. Johnson and J. Lindenstrauss, “Extensions of Lipschitz mappings into a Hilbert space,” in *Contemporary Mathematics*, R. Beals, A. Beck, A. Bellow, and A. Hajian, Eds., vol. 26, American Mathematical Society, 1984, pp. 189–206. DOI: 10.1090/conm/026/737400.
- [56] C. B. Freksen, “An Introduction to Johnson-Lindenstrauss Transforms,” Feb. 2021. [Online]. Available: <http://arxiv.org/abs/2103.00564>.

- [57] S. Kpotufe and B. K. Sriperumbudur, “Gaussian Sketching yields a J-L Lemma in RKHS,” in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 108, Palermo, Italy: PMLR, 2020.
- [58] M. Slawski, “On Principal Components Regression, Random Projections, and Column Subsampling,” Oct. 2017. [Online]. Available: <http://arxiv.org/abs/1709.08104>.
- [59] S. Becker, B. Kavas, and M. Petrik, “Robust Partially-Compressed Least-Squares,” in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence, Association for the Advancement of Artificial Intelligence, 2017, pp. 1742–1748. [Online]. Available: [www.aaai.org](http://www.aaai.org).
- [60] R. P. Browne and J. L. Andrews, “Statistical inference for sketching algorithms,” Jun. 2023. [Online]. Available: <http://arxiv.org/abs/2306.03593>.
- [61] J. T. Chi and I. C. F. Ipsen, “A Projector-Based Approach to Quantifying Total and Excess Uncertainties for Sketched Linear Regression,” Aug. 2018. [Online]. Available: <http://arxiv.org/abs/1808.05924>.
- [62] E. Dobriban and S. Liu, “Asymptotics for Sketching in Least Squares Regression,” Oct. 2018. [Online]. Available: <http://arxiv.org/abs/1810.06089>.
- [63] D. Homrighausen and D. J. McDonald, “Compressed and Penalized Linear Regression,” May 2017. DOI: 10.1080/10618600.2019.1660179. [Online]. Available: <http://arxiv.org/abs/1705.08036><http://dx.doi.org/10.1080/10618600.2019.1660179>.
- [64] M. W. Mahoney and P. Drineas, “CUR matrix decompositions for improved data analysis,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 3, pp. 697–702, Jan. 2009. DOI: 10.1073/pnas.0803205105. [Online]. Available: [www.pnas.org/cgi/content/full/](http://www.pnas.org/cgi/content/full/).

- [65] M. Pilanci and M. J. Wainwright, “Randomized Sketches of Convex Programs with Sharp Guarantees,” Apr. 2014. [Online]. Available: <http://arxiv.org/abs/1404.7203>.
- [66] G. Raskutti and M. W. Mahoney, “A Statistical Perspective on Randomized Sketching for Ordinary Least-Squares,” *Journal of Machine Learning Research*, vol. 17, pp. 1–31, 2016. [Online]. Available: <https://jmlr.csail.mit.edu/papers/volume17/15-440/15-440.pdf>.
- [67] J. Wong, E. Forsell, R. Lewis, T. Mao, and M. Wardrop, “You Only Compress Once: Optimal Data Compression for Estimating Linear Models,” Feb. 2021. [Online]. Available: <http://arxiv.org/abs/2102.11297>.
- [68] P. Rigollet, “Linear Regression Model,” in *High Dimensional Statistics*, MIT OpenCourseWare, 2015, ch. 2. [Online]. Available: <https://ocw.mit.edu/courses/18-s997-high-dimensional-statistics-spring-2015>.
- [69] M. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

## Biography

Daniel E. Hill graduated from Greenwood High School in Greenwood, IN in 2006. He received a Bachelor of Science in Mathematics and a Bachelor of Arts in Spanish from Bethel University, Indiana in 2011. He received his Masters of Science in Applied Mathematics from the Air Force Institute of Technology in 2018.