Quantized Low-Rank Multivariate Regression With Random Dithering

Junren Chen , Yueqi Wang , and Michael K. Ng , *Senior Member, IEEE*

Abstract—**Low-rank multivariate regression (LRMR) is an important statistical learning model that combines highly correlated tasks as a multiresponse regression problem with low-rank priori on the coefficient matrix. In this paper, we study quantized LRMR, a practical setting where the responses and/or the covariates are discretized to finite precision. We focus on the estimation of the underlying coefficient matrix. To make consistent estimator that could achieve arbitrarily small error possible, we employ uniform quantization with random dithering, i.e., we add appropriate random noise to the data before quantization. Specifically, uniform dither and triangular dither are used for responses and covariates, respectively. Based on the quantized data, we propose the constrained Lasso and regularized Lasso estimators, and derive the non-asymptotic error bounds. With the aid of dithering, the estimators achieve minimax optimal rate, while quantization only slightly worsens the multiplicative factor in the error rate. Moreover, we extend our results to a low-rank regression model with matrix responses. We corroborate and demonstrate our theoretical results via simulations on synthetic data, image restoration, as well as a real data application.**

Index Terms—**Multiresponse regression, quantization, M-estimator, low-rankness, dithering.**

I. INTRODUCTION

QUANTIZATION is the process of mapping continuous input to a discrete form (e.g. a finite dictionary or a finite number of bits) [25]. Quantization of signals or data recently has received considerable attention in the communities of signal processing, statistics and machine learning. In some signal processing problems, power consumption, manufacturing cost and chip area of analog-to-digital devices grow exponentially with their resolution [34]. In this situation, it is infeasible to use high-precision data or signals, and quantization with relatively low resolution is preferable, e.g., see the distributed machine

Junren Chen and Yueqi Wang are with the Department of Mathematics, The University of Hong Kong, Hong Kong 852, China (e-mail: chenjr58@connect.hku.hk; u3007895@connect.hku.hk).

Michael K. Ng is with the Department of Mathematics, Hong Kong Baptist University, Hong Kong 852, China (e-mail: michael-ng@hkbu.edu.hk).

learning system described in [18]. Besides, in modern machine learning problems extremely huge datasets and highly complex models are ubiquitous, which often lead to distributed learning systems [38], i.e., a setting involving repeatedly communication among multiple compute nodes that are oftentimes GPUs linked processors within a single machine or even multiple machines. When the participating workers are typically large in number and have slow or unstable internet connections (e.g., low-power or low-bandwidth device such as a mobile device), the communication cost would become prohibitive [38], [51], and recent works have studied how to send a small number of bits by quantization to overcome the bottleneck [3], [31], [32], [36], [51], [56], [70]. More specifically, working with low-precision training data has proven useful in reducing computation cost when training linear model, as shown by the experimental results in [68]. Additionally, while sending the quantized gradient is the mainstream in machine learning, it may be inefficient in distributed learning with a huge number of parameters to learn; in this case, transmitting some important quantized data samples could provably reduce the communication cost [28]. Thus, it is of particular interest to theoretically investigate the interplay between parameter learning and data quantization in some fundamental statistical learning or estimation problems, e.g., [14], [15], [21].

Departing momentarily from quantization, low-rank multivariate regression (LRMR), also known as multi-task learning and reduced-rank regression [2], [12], [50], is undoubtedly a widely used statistical machine learning model. For clarity we first provide its mathematical formulation:

$$\boldsymbol{y}_k = \boldsymbol{\Theta}_0^\top \boldsymbol{x}_k + \boldsymbol{\epsilon}_k, \ k = 1, ..., n, \tag{1}$$

and the main goal is to learn the underlying parameter $\boldsymbol{\Theta}_0 \in \mathbb{R}^{d_1 \times d_2}$ from the covariate-response pairs $(\boldsymbol{x}_k, \boldsymbol{y}_k) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. Compared to the canonical regression problem with scalar response (e.g., linear regression), the core spirit of LRMR is to combine and jointly solve d_2 highly correlated tasks. In particular, the coefficient vectors of the d_2 tasks are merged into $\boldsymbol{\Theta}_0$ in (1), and the low-rankness of $\boldsymbol{\Theta}_0$ is often assumed to exploit the "intrinsic relatedness" of the d_2 learning problems (e.g., [23], [24], [44]). This model can capture many natural phenomena and hence has a broad range of applications. For example, in genomics study [8], the gene expression profiles ($\boldsymbol{y}_k$) and the genetic markers ($\boldsymbol{x}_k$) can be approximately associated through only a few linear combinations of

highly-correlated genetic markers. Therefore, recovering a low-rank and sometimes also sparse coefficient matrix holds the key to reveal such connections between the responses and predictors. In addition, in the study of functional magnetic resonance imaging (fMRI) [30], each voxel within the brain is represented by a time series of neurophysiological activity. Combining with the multivariate voxel-based time series, researchers use a linear model to describe the underlying large-scale network connectivities among functionally specialized regions in the brain. A practical way is to use a suitable matrix to identify these complex interconnections in the brain; while aiming at modelling the connections via only a small subset of the given data, one often imposes appropriate structures (e.g., low-rankness, sparsity) on the coefficient matrix. Besides, other applications include analysis of electroencephalography (EEG) data decoding [1], neural response modeling [7], analysis of financial data [50], chemometrics, psychometrics and econometrics [67], to name just a few.

Note that in real applications, data are inevitably quantized to bit streams for the purpose of storage, processing and transmission. Also, the problem of LRMR can possibly arise in a distributed learning regime where quantization is necessary to render lower communication cost. Thus, a natural question is to study quantized LRMR, i.e., LRMR where data are quantized by some mechanism, and one can only access the quantized data for subsequent learning procedure. In signal and image processing, the most natural quantization method is arguably the uniform quantizer that discretizes data in a uniform manner [25], [26], [40]. More precisely, given quantization level δ, a real scalar a is quantized to $\mathcal{Q}_\delta(a) := \delta\left(\lfloor \frac{a}{\delta} \rfloor + \frac{1}{2}\right) \in \delta \cdot (\mathbb{Z} + \frac{1}{2})$. For instance, under $\mathcal{Q}_1(.)$, the data would be discretized to $\{..., -\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, ...\}$, depending on which cell they belong to. Unfortunately, directly applying the uniform quantizer $\mathcal{Q}_\delta(.)$ to LRMR will limit our ability to learn the desired $\boldsymbol{\Theta}_0$ — specifically, consistent estimator[1] is in general impossible from $(\boldsymbol{x}_k, \mathcal{Q}_\delta(\boldsymbol{y}_k))$. This is true even if we only have one task ($d_2 = 1$): consider a problem with binary features (i.e., $\boldsymbol{x}_k \in \{-1, 1\}^{d_1}$) and without noise; if we quantize $\boldsymbol{y}_k$ to $\mathcal{Q}_1(\boldsymbol{y}_k)$, then we can never distinguish $\boldsymbol{\Theta}_{01} = [0.5, 0.4, 0, ..., 0]^\top$ and $\boldsymbol{\Theta}_{02} = [0.5, -0.25, 0.2, ..., 0]^\top$ because $\mathcal{Q}_1(\boldsymbol{\Theta}_{01}^\top \boldsymbol{x}) = \mathcal{Q}_1(\boldsymbol{\Theta}_{02}^\top \boldsymbol{x})$ holds for any $\boldsymbol{x} \in \{-1, 1\}^{d_1}$. Later, this issue will be complemented by the numerical result in Fig. 4.

To address the issue, in this paper, we study LRMR under dithered quantization that involves random dithering — a process that adds random noise to the signal before quantization. The benefit of dithering for image or speech signals was empirically observed quite early [33], [41], [52], while the theoretical results for quantization error/noise were established in [54], see also a cleaner proof provided by [26]. In a nutshell, the benefit of dithering is to whiten the quantization noise. Even more surprisingly, the quantization errors follow i.i.d. uniform distribution (Lemma 1(a)). While we focus on the dithered uniform

quantizer, interested readers may consult [64] for an extensive treatment of quantization noise under various quantizers.

We deal with the quantization of both response and covariate. We propose to use uniform dither for $\boldsymbol{y}_k$, triangular dither for $\boldsymbol{x}_k$ (see precise definition later), and then apply the uniform quantizer. Note that the quantization method is memoryless and thus well suited to hardware implementation. Our main contributions are as follows:

- Based on the quantized data, we develop an empirical ℓ_2 loss, which coupled with either nuclear norm constraint or regularization leads to Lasso estimators. We establish minimax optimal non-asymptotic error bounds for the estimators in the cases of "partial quantization" (i.e., only quantize $\boldsymbol{y}_k$) and "complete quantization" (i.e., quantize both $\boldsymbol{x}_k, \boldsymbol{y}_k$). The bounds also characterize how quantization resolution affects the estimation error.
- We show that our quantization method is also applicable to a low-rank linear regression model with matrix response recently studied in [39]. Our Lasso estimators based on quantized data could still achieve error rate comparable to the full-data regime in [39].

A. Related Works

There has been rapidly growing literature on quantized compressed sensing [5], [14], [15], [20], [22], [47], [57], [59], [66], quantized matrix completion [4], [10], [14], [15], [19], [35], and more recently quantized covariance estimation [14], [15], [21], but we are not aware of any earlier work on quantized LRMR (or more generally put, quantized multiresponse regression). Closest to this paper are prior developments on compressed sensing (CS) under dithered uniform quantization [14], [57], [59], [66], which we briefly review here. Recall that the (noiseless) CS problem is to recover a structured (e.g., sparse/low-rank) signal $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ from the data of $(\boldsymbol{x}_k, y_k := \boldsymbol{x}_k^\top \boldsymbol{\theta}_0)_{k=1}^n$, where $\boldsymbol{x}_k$ is the sensing vector, y_k is the measurement, and the high-dimensional setting $n \ll d$ is of primary interest. It was shown that, while quantizing y_k via $\mathcal{Q}_\delta(.)$ with uniform dithering, recovery with near optimal error rate can still be achieved by constrained Lasso [59] or the *Projected Back Projection* (PBP) estimator [66]. In [57], Sun et al. extended [59] to corrupted sensing that aims at separating signal and corruption. While [57], [59], [66] only considered the quantization of y_k, a recent work [14] developed the quantization method for $\boldsymbol{x}_k$, i.e., via the same dithered uniform quantizer but with uniform dither substituted with triangular dither.

Although we adopt a similar dithered quantization scheme (specifically, similar to [14]), the estimation problem in this paper totally differs from CS. In particular, we will study regression models with multivariate response that can be a vector with considerably large dimension (LRMR in Section III) or even a huge matrix (see Section IV), as in sharp contrast to the scalar measurement y_k in CS. A different point of view is to consider each response scalar of (1). Let $\boldsymbol{\Theta}_0 = [\boldsymbol{\theta}_{0,1}, ..., \boldsymbol{\theta}_{0,d_2}]$, then the i-th entry of $\boldsymbol{y}_k$ in (1) can be expressed as $y_{ki} = \boldsymbol{x}_k^\top \boldsymbol{\theta}_{0,i} + \epsilon_{ki}$. Because y_{ki} only involves the i-th column of the desired signal $\boldsymbol{\Theta}_0$, it is often referred to as a local measurement and considered to be less informative than the global measurement used in CS

---

[1]In regression problems, an estimator is consistent if its estimation error vanishes when sample size tends to infinity.

(see, e.g., [61]). As a consequence, the technical ingredients in this work, especially the technique to bound various random terms arising in the proof, significantly deviate from those in quantized CS.

From the more statistical side, without considering any data quantization procedure, a lot of statistical procedures have been developed for estimation and prediction in multivariate regression. Among them the most relevant are the regularized ones that minimize an objective constituted by a loss function and a suitable regularizer, see [16], [37], [39], [44], [53], [67] for instance. Indeed, the key theoretical achievement of this work is to show the compatibility between the dithered uniform quantizer and the Lasso estimator. That is, Lasso estimator can still achieve near optimal estimation error from data quantized by a uniform quantizer with appropriate random dither.

B. Outline

The remainder of this paper is organized as follows: we provide the notational conventions and preliminaries in Section II; we propose our Lasso estimators for quantized LRMR and present the theoretical results in Section III; the main results are then extended to low-rank linear model with matrix response in Section IV; we provide experimental results in Section V to validate our theory; we give some remarks to conclude this work in Section VI.

II. PRELIMINARIES

(Notation). We denote matrices and vectors by boldface letters, while scalars by regular letters. We write $[m] = \{1, ..., m\}$ for positive integer m. For vector $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we work with the ℓ_p norm $\|\boldsymbol{x}\|_p = (\sum_{i \in [d]} |x_i|^p)^{1/p}$, max norm $\|\boldsymbol{x}\|_\infty = \max_{i \in [d]} |x_i|$, and inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y}$. For matrices $\boldsymbol{A}, \boldsymbol{B}$, we work with the transpose $\boldsymbol{A}^\top$, the operator norm $\|\boldsymbol{A}\|_{op}$, Frobenius norm $\|\boldsymbol{A}\|_F$, nuclear norm $\|\boldsymbol{A}\|_{nu}$ (sum of singular values), max norm $\|\boldsymbol{A}\|_\infty = \max_{i,j} |a_{ij}|$, and the inner product $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \text{Tr}(\boldsymbol{A}^\top \boldsymbol{B})$. The standard Euclidean sphere of $\mathbb{R}^d$ is denoted by $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$. For a random variable X, we let $\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E} \exp(\frac{X^2}{t^2}) \leq 2\}$ (resp. $\|X\|_{\psi_1} = \inf\{t > 0 : \mathbb{E} \exp(\frac{|X|}{t}) \leq 2\}$) be the sub-Gaussian norm (resp. sub-exponential norm), $\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p}$ be the L^p norm. We represent universal constants by C, c, C_i or c_i, whose value may vary from line to line. We write $T_1 \lesssim T_2$ or $T_1 = O(T_2)$ if $T_1 \leq CT_2$; Conversely, $T_1 \gtrsim T_2$ or $T_1 = \Omega(T_2)$ if $T_1 \geq cT_2$. Note that $T_1 \asymp T_2$ if $T_1 = O(T_2)$ and $T_2 = \Omega(T_1)$ simultaneously hold. We use $\mathscr{U}(W)$ to denote the uniform distribution over W. We use $\text{vec}(\boldsymbol{A}) \in \mathbb{R}^{mn \times 1}$ to vectorize a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, while $\text{mat}(.)$ denotes the inverse operator.

A. High-Dimensional Probability

A random variable X with finite $\|X\|_{\psi_2}$ is said to be sub-Gaussian. Note that sub-Gaussian X exhibits exponentially decaying probability tail, i.e., for any $t > 0$,

$$\mathbb{P}(|X| \geq t) \leq 2 \exp\left(-\frac{ct^2}{\|X\|_{\psi_2}^2}\right). \tag{2}$$

Similarly, X with finite $\|X\|_{\psi_1}$ is sub-exponential and has the following tail bound for any $t > 0$

$$\mathbb{P}(|X| \geq t) \leq 2 \exp\left(-\frac{ct}{\|X\|_{\psi_1}}\right). \tag{3}$$

Conversely, both properties in (2) (resp. (3)) can characterize the norm $\|.\|_{\psi_2}$ (resp. $\|.\|_{\psi_1}$) up to multiplicative constant, see [62, Proposition 2.5.2, 2.7.1] for instance. To relate sub-Gaussian norm and sub-exponential norm, one has (see, e.g., [62, Lem. 2.7.7])

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}. \tag{4}$$

For n-dimensional random vector $\boldsymbol{X}$ we let $\|\boldsymbol{X}\|_{\psi_2} = \sup_{\boldsymbol{v} \in \mathbb{S}^{n-1}} \|\boldsymbol{v}^\top \boldsymbol{X}\|_{\psi_2}$.

B. Dithered Uniform Quantization

First, we describe the dithered uniform quantization with $\delta > 0$ for an input signal $\boldsymbol{x} \in \mathbb{R}^N$ as follows:

- Independent of $\boldsymbol{x}$, we i.i.d. draw the entries of the random dither $\boldsymbol{\tau} \in \mathbb{R}^N$ from some suitable distribution;
- Then, we quantize $\boldsymbol{x}$ to $\dot{\boldsymbol{x}} = \mathcal{Q}_\delta(\boldsymbol{x} + \boldsymbol{\tau})$, with $\mathcal{Q}_\delta(a) := \delta(\lfloor \frac{a}{\delta} \rfloor + \frac{1}{2})$ $(a \in \mathbb{R})$ applied element-wisely.

We adopt the following conventions (as in [25], [26]): $\boldsymbol{w} := \dot{\boldsymbol{x}} - (\boldsymbol{x} + \boldsymbol{\tau})$ is the **quantization error**, and $\boldsymbol{\xi} := \dot{\boldsymbol{x}} - \boldsymbol{x}$ is the **quantization noise**.

The principal properties of the dithered quantization that underlie our analysis are provided in Lemma 1.

Lemma 1: (Theorems 1–2 in [26]). We consider the above dithered uniform quantization: $\boldsymbol{x} = [x_i]$ is the input signal, $\boldsymbol{\tau} = [\tau_i]$ is the random dither whose entries are i.i.d. copies of random variable Y. We use $\mathbf{i}$ to denote the complex unit.

(a) (Quantization Error). Let $\boldsymbol{w} := \dot{\boldsymbol{x}} - (\boldsymbol{x} + \boldsymbol{\tau}) = [w_i]$ be the quantization error. If $f(u) := \mathbb{E}(\exp(\mathbf{i} u Y))$ satisfies $f\left(\frac{2\pi l}{\delta}\right) = 0$ for all non-zero integer l, then x_i and w_j are independent for all $i, j \in [N]$. Moreover, $\{w_j : j \in [N]\}$ are i.i.d. distributed as $\mathscr{U}\left([-\frac{\delta}{2}, \frac{\delta}{2}]\right)$.

(b) (Quantization Noise). Let $\boldsymbol{\xi} := \dot{\boldsymbol{x}} - \boldsymbol{x} = [\xi_i]$ be the quantization noise. Assume $Z \sim \mathscr{U}\left[-\frac{\delta}{2}, \frac{\delta}{2}\right]$ is independent of Y. Let $g(u) := \mathbb{E}(\exp(\mathbf{i} u Y))\mathbb{E}(\exp(\mathbf{i} u Z))$. Given positive integer p, if the p-th order derivative $g^{(p)}(u)$ satisfies $g^{(p)}\left(\frac{2\pi l}{\delta}\right) = 0$ for all non-zero integer l, then the p-th conditional moment of ξ_i does not depend on $\boldsymbol{x}$. More precisely, we have $\mathbb{E}[\xi_i^p | \boldsymbol{x}] = \mathbb{E}(Y + Z)^p$.

Given quantization level $\delta > 0$, in this work we focus on uniform dither $\tau_i \sim \mathscr{U}\left([-\frac{\delta}{2}, \frac{\delta}{2}]\right)$ and triangular dither[2] $\tau_i \sim \mathscr{U}\left([-\frac{\delta}{2}, \frac{\delta}{2}]\right) + \mathscr{U}\left([-\frac{\delta}{2}, \frac{\delta}{2}]\right)$ (i.e., the sum of two independent uniform distribution). From Lemma 1, the following properties are immediate. The proof can be found in the Appendix.

Corollary 1: In the setting of Lemma 1, if $\tau_i \sim \mathscr{U}\left([-\frac{\delta}{2}, \frac{\delta}{2}]\right)$ or $\tau_i \sim \mathscr{U}\left([-\frac{\delta}{2}, \frac{\delta}{2}]\right) + \mathscr{U}\left([-\frac{\delta}{2}, \frac{\delta}{2}]\right)$, then x_i and w_j are independent $(\forall i, j)$, and $\{w_j : j\}$ are i.i.d. copies of $\mathscr{U}\left([-\frac{\delta}{2}, \frac{\delta}{2}]\right)$; In addition, for the triangular dither $\tau_i \sim \mathscr{U}\left([-\frac{\delta}{2}, \frac{\delta}{2}]\right) + \mathscr{U}\left([-\frac{\delta}{2}, \frac{\delta}{2}]\right)$, the variance of ξ_i is independent of signal; more precisely it holds that $\mathbb{E}\xi_i^2 = \frac{\delta^2}{4}$.

---

[2]This is the term used in prior work, e.g., [26].

The benefit of using proper dither (e.g., uniform dither) is now clear, i.e., to whiten the quantization noise. For instance, under $\boldsymbol{\tau} \sim \mathscr{U}\left([-\frac{\delta}{2}, \frac{\delta}{2}]^N\right)$, one has $\mathbb{E}\dot{\boldsymbol{x}} = \mathbb{E}(\boldsymbol{x} + \boldsymbol{\tau} + \boldsymbol{w}) = \mathbb{E}\boldsymbol{x} + \mathbb{E}\boldsymbol{\tau} + \mathbb{E}\boldsymbol{w} = \mathbb{E}\boldsymbol{x}$.

III. QUANTIZED LOW-RANK MULTIVARIATE REGRESSION

The low-rank multivariate regression (LRMR) model is

$$\boldsymbol{y}_k = \boldsymbol{\Theta}_0^\top \boldsymbol{x}_k + \boldsymbol{\epsilon}_k, \ k = 1, ..., n, \tag{5}$$

where $\boldsymbol{x}_k \in \mathbb{R}^{d_1}$ is the covariate, $\boldsymbol{y}_k \in \mathbb{R}^{d_2}$ is the response perturbed by random noise $\boldsymbol{\epsilon}_k$, and $\boldsymbol{\Theta}_0 \in \mathbb{R}^{d_1 \times d_2}$ is the desired parameter. Our goal is to estimate the $\boldsymbol{\Theta}_0$ from $(\boldsymbol{x}_k, \boldsymbol{y}_k)$'s. We make the following sub-Gaussian assumption. Note that we assume $\mathbb{E}\boldsymbol{x}_k = 0$ for simplicity, and the case of "$\mathbb{E}\boldsymbol{x}_k \neq 0$" can be addressed by data centering or including an intercept term in (5). It should be noted that these distributional assumptions are standard and commonly adopted for analysing regularized M-estimators (defined in (6) shortly) in multiresponse regression problems, see [44, Coro. 3], [27], [49] for instance. Indeed, our Assumption 1 slightly relaxes the assumptions made in these prior works from Gaussian data to sub-Gaussian data, and note that this relaxation is important for certain cases, e.g., when we work with binary data that cannot be captured by Gaussian distribution.

Assumption 1: The covariates $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are i.i.d., zero-mean and sub-Gaussian with $\|\boldsymbol{x}_k\|_{\psi_2} \leq K$; The covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{xx}} = \mathbb{E}(\boldsymbol{x}_k \boldsymbol{x}_k^\top)$ satisfies $\kappa_0 \leq \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{xx}}) \leq \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{xx}}) \leq \kappa_1$ for some $\kappa_1 \geq \kappa_0 > 0$; Independent of $\{\boldsymbol{x}_k : k \in [n]\}$, the noise vectors $\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_n$ are i.i.d., zero-mean and sub-Gaussian with $\|\boldsymbol{\epsilon}_k\|_{\psi_2} \leq E$; $\boldsymbol{y}_k$ is generated from (5) for some low-rank $\boldsymbol{\Theta}_0$ satisfying $\text{rank}(\boldsymbol{\Theta}_0) \leq r$.

Although this multivariate regression model was already intensively studied in the literature (e.g., [24], [44], [53]), the novelty of this work lies in the quantization that is inevitable in the era of digital signal processing. In particular, we study "partial quantization" where only the response is quantized, as well as a more tricky setting of "complete quantization" where the entire covariate-response pair $(\boldsymbol{x}_k, \boldsymbol{y}_k)$ is quantized to finite precision. We propose the dithered quantization scheme as follows:

- (Covariate quantization). Independent of $(\boldsymbol{x}_k, \boldsymbol{y}_k)$, we i.i.d. draw triangular dither $\boldsymbol{\phi}_k \sim \mathscr{U}\left([-\frac{\delta_1}{2}, \frac{\delta_1}{2}]^{d_1}\right) + \mathscr{U}\left([-\frac{\delta_1}{2}, \frac{\delta_1}{2}]^{d_1}\right)$, and then quantize $\boldsymbol{x}_k$ to $\dot{\boldsymbol{x}}_k = \mathcal{Q}_{\delta_1}(\boldsymbol{x}_k + \boldsymbol{\phi}_k)$.
- (Response quantization). Independent of $(\boldsymbol{x}_k, \boldsymbol{y}_k)$, we i.i.d. draw uniform dither $\boldsymbol{\tau}_k \sim \mathscr{U}\left([-\frac{\delta_2}{2}, \frac{\delta_2}{2}]^{d_2}\right)$, and then quantize $\boldsymbol{y}_k$ to $\dot{\boldsymbol{y}}_k = \mathcal{Q}_{\delta_2}(\boldsymbol{y}_k + \boldsymbol{\tau}_k)$.

Note that $\delta_1 = 0$ means no quantization on $\boldsymbol{x}_k$, thus corresponding to "partial quantization" that only involves response quantization. While almost all works related works studied response quantization (as reviewed in Section I-A), we comment on the necessity of also studying covariate quantization ($\delta_1 > 0$). For instance, when LRMR appears as a distributed learning problem where the features are transmitted among multiple parties, quantization is often needed for reducing communication cost. Also note that, a mode direct benefit is the lower memory load.

A. The Empirical Loss Under Quantization

Using the vector ℓ_1 norm as regularizer to promote sparsity, Lasso is viewed as a benchmark procedure for recovering sparse vector [60]. The efficacy of Lasso has extended to the recovery of low-rank matrix by replacing the ℓ_1 norm with the nuclear norm of a matrix, see [11], [44] for instance. Having assumed $\boldsymbol{\Theta}_0$ to be low-rank, one can apply similar idea to LRMR and formulate the regularized Lasso recovery program as

$$\underset{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}}{\arg\min} \underbrace{\frac{1}{n} \sum_{k=1}^n \left\|\boldsymbol{y}_k - \boldsymbol{\Theta}^\top \boldsymbol{x}_k\right\|_2^2}_{\mathcal{L}(\boldsymbol{\Theta})} + \lambda\|\boldsymbol{\Theta}\|_{nu}, \tag{6}$$

where $\mathcal{L}(\boldsymbol{\Theta})$ is the ℓ_2 loss function for data fitting purpose, $\lambda\|\boldsymbol{\Theta}\|_{nu}$ is the regularization part for low-rank structure, and λ should be tuned to balance data fidelity and the low-rankness. Note that (6) also falls into the range of M-estimator [44], [45]. When a good estimate on $\|\boldsymbol{\Theta}_0\|_{nu}$ is available, one can also consider the constrained Lasso

$$\underset{\|\boldsymbol{\Theta}\|_{nu} \leq R}{\arg\min} \frac{1}{n} \sum_{k=1}^n \left\|\boldsymbol{y}_k - \boldsymbol{\Theta}^\top \boldsymbol{x}_k\right\|_2^2 := \mathcal{L}(\boldsymbol{\Theta}). \tag{7}$$

We note that Lasso is known to achieve minimax rate in LRMR, see [44] for instance.

However, under data quantization one can only access $(\dot{\boldsymbol{x}}_k, \dot{\boldsymbol{y}}_k)$ for recovery; while the regularizer $\|\boldsymbol{\Theta}\|_{nu}$ is unproblematic, one evidently lacks full data for constructing the empirical ℓ_2 loss $\mathcal{L}(\boldsymbol{\Theta})$; modification of $\mathcal{L}(\boldsymbol{\Theta})$ is thus needed. To draw some inspiration, the quite instructive first step is to calculate the expected ℓ_2 loss:

$$\begin{aligned} \mathbb{E}\mathcal{L}(\boldsymbol{\Theta}) &= \mathbb{E}\|\boldsymbol{y}_k - \boldsymbol{\Theta}^\top \boldsymbol{x}_k\|_2^2 \\ &\stackrel{(i)}{=} \mathbb{E}\|\boldsymbol{\Theta}^\top \boldsymbol{x}_k\|^2 - 2\mathbb{E}(\boldsymbol{y}_k^\top \boldsymbol{\Theta}^\top \boldsymbol{x}_k) \\ &= \langle \boldsymbol{\Theta}\boldsymbol{\Theta}^\top, \mathbb{E}(\boldsymbol{x}_k \boldsymbol{x}_k^\top) \rangle - 2\langle \boldsymbol{\Theta}, \mathbb{E}(\boldsymbol{x}_k \boldsymbol{y}_k^\top) \rangle \\ &\stackrel{(ii)}{:=} \langle \boldsymbol{\Theta}\boldsymbol{\Theta}^\top, \boldsymbol{\Sigma}_{\boldsymbol{xx}} \rangle - 2\langle \boldsymbol{\Theta}, \boldsymbol{\Sigma}_{\boldsymbol{xy}} \rangle, \end{aligned}$$

where (i) holds up to constant that has no effect on the optimization, and in (ii) we introduce the shorthand for the covariance, $\boldsymbol{\Sigma}_{\boldsymbol{xx}} = \mathbb{E}(\boldsymbol{x}_k \boldsymbol{x}_k^\top)$ and $\boldsymbol{\Sigma}_{\boldsymbol{xy}} = \mathbb{E}(\boldsymbol{x}_k \boldsymbol{y}_k^\top)$. Therefore, in order to construct a suitable empirical ℓ_2 loss, we need to find surrogates for $\boldsymbol{\Sigma}_{\boldsymbol{xx}}, \boldsymbol{\Sigma}_{\boldsymbol{xy}}$ based on $(\dot{\boldsymbol{x}}_k, \dot{\boldsymbol{y}}_k)$.

To facilitate the exposition, we reserve the following notation in subsequent developments: for quantization of $\boldsymbol{x}_k$ with dither $\boldsymbol{\phi}_k$, $\boldsymbol{w}_{k1} = \dot{\boldsymbol{x}}_k - (\boldsymbol{x}_k + \boldsymbol{\phi}_k)$ is the quantization error, $\boldsymbol{\xi}_{k1} := \dot{\boldsymbol{x}}_k - \boldsymbol{x}_k$ is the quantization noise; for quantization of $\boldsymbol{y}_k$ with dither $\boldsymbol{\tau}_k$, $\boldsymbol{w}_{k2} := \dot{\boldsymbol{y}}_k - (\boldsymbol{y}_k + \boldsymbol{\tau}_k)$ stands for the quantization error, while $\boldsymbol{\xi}_{k2} := \dot{\boldsymbol{y}}_k - \boldsymbol{y}_k$ the quantization noise. We use $\xi_{kj,i}$ to denote the i-th entry of $\boldsymbol{\xi}_{kj}$, and the meanings of notation like $w_{kj,i}, \phi_{k,i}$ are similar. Now we are ready to present a Lemma that indicates the suitable surrogates of $\boldsymbol{\Sigma}_{\boldsymbol{xx}}, \boldsymbol{\Sigma}_{\boldsymbol{xy}}$.

Lemma 2: Based on the quantized data $(\dot{\boldsymbol{x}}_k, \dot{\boldsymbol{y}}_k)$, we let $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} := \frac{1}{n} \sum_{k=1}^n \dot{\boldsymbol{x}}_k \dot{\boldsymbol{x}}_k^\top - \frac{\delta_1^2}{4} \boldsymbol{I}_{d_1}$, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} = \frac{1}{n} \sum_{k=1}^n \dot{\boldsymbol{x}}_k \dot{\boldsymbol{y}}_k^\top$, then we have $\mathbb{E}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} = \boldsymbol{\Sigma}_{\boldsymbol{xx}}, \mathbb{E}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} = \boldsymbol{\Sigma}_{\boldsymbol{xy}}$.

Proof: We first calculate the easier $\mathbb{E}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} = \mathbb{E}(\dot{\boldsymbol{x}}_k \dot{\boldsymbol{y}}_k^\top)$:

$$\mathbb{E}(\dot{\boldsymbol{x}}_k \dot{\boldsymbol{y}}_k^\top) = \mathbb{E}(\boldsymbol{x}_k + \boldsymbol{\phi}_k + \boldsymbol{w}_{k1})(\boldsymbol{y}_k + \boldsymbol{\tau}_k + \boldsymbol{w}_{k2})^\top$$

$$= \mathbb{E}(\boldsymbol{x}_k \boldsymbol{y}_k^\top) + \mathbb{E}(\boldsymbol{x}_k \boldsymbol{\tau}_k^\top) + \mathbb{E}(\boldsymbol{x}_k \boldsymbol{w}_{k2}^\top)$$
$$+ \mathbb{E}(\boldsymbol{\phi}_k \boldsymbol{y}_k^\top) + \mathbb{E}(\boldsymbol{\phi}_k \boldsymbol{\tau}_k^\top) + \mathbb{E}(\boldsymbol{\phi}_k \boldsymbol{w}_{k2}^\top)$$
$$+ \mathbb{E}(\boldsymbol{w}_{k1} \boldsymbol{y}_k^\top) + \mathbb{E}(\boldsymbol{w}_{k1} \boldsymbol{\tau}_k^\top) + \mathbb{E}(\boldsymbol{w}_{k1} \boldsymbol{w}_{k2}^\top) \overset{(i)}{=} \boldsymbol{\Sigma}_{\boldsymbol{xy}},$$

where (i) is because in the previous step, all terms but $\mathbb{E}(\boldsymbol{x}_k \boldsymbol{y}_k^\top)$ vanish, due to the nice property that $\boldsymbol{w}_{k1}, \boldsymbol{w}_{k2}$ are independent of $(\boldsymbol{x}_k, \boldsymbol{y}_k)$, $\boldsymbol{w}_{k1} \overset{iid}{\sim} \mathscr{U}\left(\left[-\frac{\delta_1}{2}, \frac{\delta_1}{2}\right]\right)$, $\boldsymbol{w}_{k2} \overset{iid}{\sim} \mathscr{U}\left(\left[-\frac{\delta_2}{2}, \frac{\delta_2}{2}\right]\right)$, see Corollary 1. Similarly, to evaluate $\mathbb{E}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} = \mathbb{E}(\dot{\boldsymbol{x}}_k \dot{\boldsymbol{x}}_k^\top) - \frac{\delta_1^2}{4}\boldsymbol{I}_{d_1}$, we calculate $\mathbb{E}(\dot{\boldsymbol{x}}_k \dot{\boldsymbol{x}}_k^\top)$ as follows:

$$\mathbb{E}(\dot{\boldsymbol{x}}_k \dot{\boldsymbol{x}}_k^\top) = \mathbb{E}(\boldsymbol{x}_k + \boldsymbol{\xi}_{k1})(\boldsymbol{x}_k + \boldsymbol{\xi}_{k1})^\top$$
$$\overset{(i)}{=} \mathbb{E}(\boldsymbol{x}_k \boldsymbol{x}_k^\top) + \mathbb{E}(\boldsymbol{\xi}_{k1} \boldsymbol{\xi}_{k1}^\top) \overset{(ii)}{=} \mathbb{E}(\boldsymbol{x}_k \boldsymbol{x}_k^\top) + \frac{\delta_1^2}{4}\boldsymbol{I}_{d_1},$$
$$(8)$$

where (i) is due to $\mathbb{E}(\boldsymbol{x}_k \boldsymbol{\xi}_{k1}^\top) = \mathbb{E}(\boldsymbol{x}_k \boldsymbol{\phi}_k^\top) + \mathbb{E}(\boldsymbol{x}_k \boldsymbol{w}_{k1}^\top) = 0$, (ii) is because the diagonal entry equals $\mathbb{E}\xi_{k1,i}^2 = \frac{\delta_1^2}{4}$ (Corollary 1), and for $i \neq j$, $\mathbb{E}(\xi_{k1,i}\xi_{k1,j}) = \mathbb{E}(\phi_{k,i} + w_{k1,i})(\phi_{k,j} + w_{k1,j}) = 0$, again due to the properties in Corollary 1. The proof is complete. □

Remark 1: (Triangular dither) While the uniform dither is a quite standard choice in the literature, we comment on the necessity of using triangular dither for $\boldsymbol{x}_k$. In essence, this is because in the estimation of $\boldsymbol{\Sigma}_{\boldsymbol{xx}}$, the quantized sample covariance contains the bias $\mathbb{E}(\boldsymbol{\xi}_{k1}\boldsymbol{\xi}_{k1}^\top)$ (see (8)), which must be removed. However, the diagonal entry $\mathbb{E}\xi_{k1,i}^2$ remains unknown under the dither of $\mathscr{U}\left(\left[-\frac{\delta_1}{2}, \frac{\delta_1}{2}\right]\right)$, see [26, Page 3]. Fortunately, by Lemma 1(b), the direct remedy is to use a dither that enjoys quantization noise with signal-independent variance, e.g., $\mathscr{U}\left(\left[-\frac{\delta_1}{2}, \frac{\delta_1}{2}\right]\right) + \mathscr{U}\left(\left[-\frac{\delta_1}{2}, \frac{\delta_1}{2}\right]\right)$. Such triangular dither was also adopted in [14] when studying covariate quantization in compressed sensing.

With all these preparations, we are in a position to specify the empirical loss

$$\dot{\mathcal{L}}(\boldsymbol{\Theta}) = \langle \boldsymbol{\Theta}\boldsymbol{\Theta}^\top, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} \rangle - 2\langle \boldsymbol{\Theta}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} \rangle. \qquad (9)$$

Note that $\dot{\mathcal{L}}(\boldsymbol{\Theta})$ reduces to the ordinary ℓ_2 loss $\mathcal{L}(\boldsymbol{\Theta})$ (up to additive constant) if $\delta_1 = \delta_2 = 0$. Further combined with the regularizer, the Lasso recovery procedure for our quantized setting can be proposed. The remainder of this section is devoted to the theoretical analysis of Lasso.

B. Constrained Lasso

First, we study the constrained Lasso where the sparsity is promoted by a "hard" constraint. Indeed, we simply substitute the unknown $\mathcal{L}(\boldsymbol{\Theta})$ in (7) with $\dot{\mathcal{L}}(\boldsymbol{\Theta})$, and to focus on estimation problem per se we ideally assume the prior estimate is precise, i.e., $R := \|\boldsymbol{\Theta}_0\|_{nu}$.[3] Hence, we formulate the constrained Lasso estimator as

$$\widehat{\boldsymbol{\Theta}}_c = \underset{\|\boldsymbol{\Theta}\|_{nu} \leq \|\boldsymbol{\Theta}^\star\|_{nu}}{\arg\min} \dot{\mathcal{L}}(\boldsymbol{\Theta}), \qquad (10)$$

where $\dot{\mathcal{L}}(\boldsymbol{\Theta})$ is defined as (9). For convenience we define the estimation error $\widehat{\boldsymbol{\Delta}}_c := \widehat{\boldsymbol{\Theta}}_c - \boldsymbol{\Theta}_0$.

[3] One may relax this via more localized arguments as in [48], which we do not pursue here.

We begin with two Lemmas that will support the proof of our main Theorem.

Lemma 3: Assume $\boldsymbol{a}_1, ..., \boldsymbol{a}_n \in \mathbb{R}^{d_1}$ are independent and satisfy $\max_k \|\boldsymbol{a}_k\|_{\psi_2} \leq E_1$; $\boldsymbol{b}_1, ..., \boldsymbol{b}_n \in \mathbb{R}^{d_2}$ are independent and satisfy $\max_k \|\boldsymbol{b}_k\|_{\psi_2} \leq E_2$. Assume $n \geq d_1 + d_2$, then it holds with probability at least $1 - 2\exp(-c_2(d_1 + d_2))$ that,

$$\left\| \frac{1}{n} \sum_{k=1}^n \left\{ \boldsymbol{a}_k \boldsymbol{b}_k^\top - \mathbb{E}(\boldsymbol{a}_k \boldsymbol{b}_k^\top) \right\} \right\|_{op} \leq C_1 E_1 E_2 \sqrt{\frac{d_1 + d_2}{n}}.$$

Lemma 3 follows similar courses as [44, Lemma 3] and involves a standard covering argument. We defer the proof to the Appendix.

Lemma 4: Under Assumption 1 and our quantization scheme, recall that $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} = \frac{1}{n} \sum_{k=1}^n \dot{\boldsymbol{x}}_k \dot{\boldsymbol{x}}_k^\top - \frac{\delta_1^2}{4}\boldsymbol{I}_{d_1}$, then the event

$$\left\| \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} - \boldsymbol{\Sigma}_{\boldsymbol{xx}} \right\|_{op} \leq C_1 A_1 \left(\sqrt{\frac{d_1 + t}{n}} + \frac{d_1 + t}{n} \right) \qquad (11)$$

holds with probability at least $1 - 2\exp(-t)$, where the multiplicative factor is $A_1 := \frac{(K + c_2\delta_1)^2}{\kappa_0 + \delta_1^2/4}$.

Proof: By (8) we first note that

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} - \boldsymbol{\Sigma}_{\boldsymbol{xx}} = \frac{1}{n} \sum_{k=1}^n \dot{\boldsymbol{x}}_k \dot{\boldsymbol{x}}_k^\top - \frac{\delta_1^2}{4}\boldsymbol{I}_{d_1} - \mathbb{E}(\boldsymbol{x}_k \boldsymbol{x}_k^\top)$$
$$= \frac{1}{n} \sum_{k=1}^n \dot{\boldsymbol{x}}_k \dot{\boldsymbol{x}}_k^\top - \mathbb{E}(\dot{\boldsymbol{x}}_k \dot{\boldsymbol{x}}_k^\top).$$

We then verify the sub-Gaussianity of $\dot{\boldsymbol{x}}_k$. Since $\boldsymbol{\phi}_k \sim \mathscr{U}\left(\left[-\frac{\delta_1}{2}, \frac{\delta_1}{2}\right]^{d_1}\right) + \mathscr{U}\left(\left[-\frac{\delta_1}{2}, \frac{\delta_1}{2}\right]^{d_1}\right)$, $\boldsymbol{w}_{k1} \sim \mathscr{U}\left(\left[-\frac{\delta_1}{2}, \frac{\delta_1}{2}\right]^{d_1}\right)$, we have

$$\|\dot{\boldsymbol{x}}_k\|_{\psi_2} \leq \|\boldsymbol{x}_k\|_{\psi_2} + \|\boldsymbol{\phi}_k\|_{\psi_2} + \|\boldsymbol{w}_{k1}\|_{\psi_2} \leq K + c_2\delta_1$$

for some c_2. Moreover, $\lambda_{\min}\left(\mathbb{E}(\dot{\boldsymbol{x}}_k \dot{\boldsymbol{x}}_k^\top)\right) = \lambda_{\min}\left(\boldsymbol{\Sigma}_{\boldsymbol{xx}}\right) + \frac{\delta_1^2}{4} \geq \kappa_0 + \frac{\delta_1^2}{4}$, thus giving

$$\|\boldsymbol{v}^\top \dot{\boldsymbol{x}}_k\|_{L^2} = \sqrt{\mathbb{E}(\boldsymbol{v}^\top \dot{\boldsymbol{x}}_k \dot{\boldsymbol{x}}_k^\top \boldsymbol{v})} \geq \sqrt{\kappa_0 + \frac{\delta_1^2}{4}}$$

for any $\boldsymbol{v} \in \mathbb{S}^{d_1 - 1}$. Therefore, it holds that

$$\|\boldsymbol{v}^\top \dot{\boldsymbol{x}}_k\|_{\psi_2} \leq K + c_2\delta_1 \leq \frac{K + c_2\delta_1}{\sqrt{\kappa_0 + \delta_1^2/4}} \|\boldsymbol{v}^\top \dot{\boldsymbol{x}}_k\|_{L^2},$$

which is just $\sqrt{A_1}\|\boldsymbol{v}^\top \dot{\boldsymbol{x}}_k\|_{L^2}$. Finally, we can invoke [62, Exercise 4.7.3], which is a well-known estimate in covariance estimation, to arrive at the desired claim. □

We are now in a position to present our first main Theorem on error bound of (10). The proof follows standard lines for analysing regularized M-estimator (e.g., [44]), but there are additional random terms to bound due to quantization noise/error, e.g., $(\boldsymbol{\xi}_{k1}, \boldsymbol{\xi}_{k2})$ in $\mathscr{T}_1$ and $\mathscr{T}_2$ in (18).

Theorem 1: (Constrained Lasso). We consider LRMR under Assumption 1 and the quantization procedure described above. We assume the sample complexity $n \gtrsim \left(\frac{A_1}{\kappa_0}\right)^2 (d_1 + d_2)$, where A_1 is the multiplicative factor in (11). Then for the estimator $\widehat{\boldsymbol{\Theta}}_c$ in (10) we have the following guarantees.

(a) (Partial Quantization). If $\delta_1 = 0$, we let $A_2 := \frac{K(E+\delta_2)}{\kappa_0}$, then with probability at least $1 - c_2 \exp(-c_3(d_1 + d_2))$ it holds that

$$\|\widehat{\boldsymbol{\Delta}}_c\|_F \leq C_1 A_2 \sqrt{\frac{r(d_1 + d_2)}{n}}. \tag{12}$$

(b) (Complete Quantization). If $\delta_1 > 0$, $\|\boldsymbol{\Theta}_0\|_{op} \leq R$, we let $A_3 := \frac{(K+\delta_1)(E+\delta_2+R\delta_1)}{\kappa_0}$, then with probability at least $1 - c_5 \exp(-c_6 d_1)$ it holds that

$$\|\widehat{\boldsymbol{\Delta}}_c\|_F \leq C_4 A_3 \sqrt{\frac{r(d_1 + d_2)}{n}}. \tag{13}$$

Proof: We begin with the optimality of $\widehat{\boldsymbol{\Theta}}_c$

$$\langle \widehat{\boldsymbol{\Theta}}_c \widehat{\boldsymbol{\Theta}}_c^\top, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\rangle - 2\langle \boldsymbol{\Theta}_c, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}}\rangle \leq \langle \boldsymbol{\Theta}_0 \boldsymbol{\Theta}_0^\top, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\rangle - 2\langle \boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}}\rangle.$$

Then we use $\widehat{\boldsymbol{\Theta}}_c = \boldsymbol{\Theta}_0 + \widehat{\boldsymbol{\Delta}}_c$ and perform some algebra to arrive at

$$\langle \widehat{\boldsymbol{\Delta}}_c \widehat{\boldsymbol{\Delta}}_c^\top, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\rangle \leq 2\langle \widehat{\boldsymbol{\Delta}}_c, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\boldsymbol{\Theta}_0\rangle, \tag{14}$$

and the remainder of the proof is essentially to bound both sides of (14).

Step 1. Bound the left-hand side from below.

Due to the scaling $n \gtrsim \left(\frac{A_1}{\kappa_0}\right)^2 (d_1 + d_2)$, we can invoke Lemma 4 with $t = d_1 + d_2$, then $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} - \boldsymbol{\Sigma}_{\boldsymbol{xx}}\|_{op} \leq \frac{\kappa_0}{2}$ holds with probability at least $1 - 2\exp(-(d_1 + d_2))$. Combined with $\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{xx}}) \geq \kappa_0$, it implies $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}) \geq \frac{\kappa_0}{2}$. Therefore, with high probability we have

$$\langle \widehat{\boldsymbol{\Delta}}_c \widehat{\boldsymbol{\Delta}}_c^\top, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\rangle = \sum_{j=1}^{d_2} (\widehat{\boldsymbol{\Delta}}_c)_{:,j}^\top \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} (\widehat{\boldsymbol{\Delta}}_c)_{:,j}$$
$$\geq \frac{\kappa_0}{2} \sum_{j=1}^{d_2} \|(\widehat{\boldsymbol{\Delta}}_c)_{:,j}\|_2^2 = \frac{\kappa_0}{2} \|\widehat{\boldsymbol{\Delta}}_c\|_F^2. \tag{15}$$

Step 2. Bound the right-hand side from above.

Note that

$$\langle \widehat{\boldsymbol{\Delta}}_c, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\boldsymbol{\Theta}_0\rangle \leq \|\widehat{\boldsymbol{\Delta}}_c\|_{nu}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\boldsymbol{\Theta}_0\|_{op} \tag{16}$$

To bound $\|\widehat{\boldsymbol{\Delta}}_c\|_{nu}$, we let $\boldsymbol{\Theta}_0 = \boldsymbol{U}_1\boldsymbol{\Sigma}\boldsymbol{V}_1^\top$ be the (compact) singular value decomposition, where $\boldsymbol{U}_1 \in \mathbb{R}^{d_1 \times r}$, $\boldsymbol{V}_1 \in \mathbb{R}^{d_2 \times r}$. Also, let $\boldsymbol{U}_2 \in \mathbb{R}^{d_1 \times (d_1-r)}$ (resp. $\boldsymbol{V}_2 \in \mathbb{R}^{d_2 \times (d_2-r)}$) be the orthogonal complement of $\boldsymbol{U}_1$ (resp. $\boldsymbol{V}_1$). Following [46], we define a pair of subspaces as

$$\mathcal{M} = \{\boldsymbol{U}_1\boldsymbol{A}\boldsymbol{V}_1^\top : \forall \boldsymbol{A} \in \mathbb{R}^{r \times r}\},$$
$$\overline{\mathcal{M}} = \{\boldsymbol{U}_1\boldsymbol{A} + \boldsymbol{B}\boldsymbol{V}_1^\top : \forall \boldsymbol{A} \in \mathbb{R}^{r \times d_2}, \boldsymbol{B} \in \mathbb{R}^{d_1 \times r}\}.$$

For subspace $\mathcal{V}$, we let $\mathcal{V}^\perp$ be its orthogonal complement, and $\mathcal{P}_\mathcal{V}(.)$ be the projection onto $\mathcal{V}$. Then it is not hard to see the decomposibility [46]:

$$\|\boldsymbol{A} + \boldsymbol{B}\|_{nu} = \|\boldsymbol{A}\|_{nu} + \|\boldsymbol{B}\|_{nu}$$

if $\boldsymbol{A} \in \mathcal{M}$, $\boldsymbol{B} \in \overline{\mathcal{M}}^\perp$. Now we can deduce that

$$\|\boldsymbol{\Theta}_0 + \widehat{\boldsymbol{\Delta}}_c\|_{nu} = \|\boldsymbol{\Theta}_0 + \mathcal{P}_{\overline{\mathcal{M}}}\widehat{\boldsymbol{\Delta}}_c + \mathcal{P}_{\overline{\mathcal{M}}^\perp}\widehat{\boldsymbol{\Delta}}_c\|_{nu}$$
$$\geq \|\boldsymbol{\Theta}_0 + \mathcal{P}_{\overline{\mathcal{M}}^\perp}\widehat{\boldsymbol{\Delta}}_c\|_{nu} - \|\mathcal{P}_{\overline{\mathcal{M}}}\widehat{\boldsymbol{\Delta}}_c\|_{nu}$$

$$\geq \|\boldsymbol{\Theta}_0\|_{nu} + \|\mathcal{P}_{\overline{\mathcal{M}}^\perp}\widehat{\boldsymbol{\Delta}}_c\|_{nu} - \|\mathcal{P}_{\overline{\mathcal{M}}}\widehat{\boldsymbol{\Delta}}_c\|_{nu}. \tag{17}$$

Combined with the constraint $\|\widehat{\boldsymbol{\Theta}}_0 + \widehat{\boldsymbol{\Delta}}_c\|_{nu} = \|\widehat{\boldsymbol{\Theta}}_c\|_{nu} \leq \|\boldsymbol{\Theta}_0\|_{nu}$, we obtain $\|\mathcal{P}_{\overline{\mathcal{M}}^\perp}\widehat{\boldsymbol{\Delta}}_c\|_{nu} \leq \|\mathcal{P}_{\overline{\mathcal{M}}}\widehat{\boldsymbol{\Delta}}_c\|_{nu}$. Thus,

$$\|\widehat{\boldsymbol{\Delta}}_c\|_{nu} \leq \|\mathcal{P}_{\overline{\mathcal{M}}^\perp}\widehat{\boldsymbol{\Delta}}_c\|_{nu} + \|\mathcal{P}_{\overline{\mathcal{M}}}\widehat{\boldsymbol{\Delta}}_c\|_{nu}$$
$$\leq 2\|\mathcal{P}_{\overline{\mathcal{M}}}\widehat{\boldsymbol{\Delta}}_c\|_{nu} \leq 2\sqrt{2r}\|\widehat{\boldsymbol{\Delta}}_c\|_F.$$

The last inequality is because $\mathrm{rank}(\boldsymbol{A}) \leq 2r$ if $\boldsymbol{A} \in \overline{\mathcal{M}}$, and it always holds that $\|\boldsymbol{A}\|_{nu} \leq \mathrm{rank}(\boldsymbol{A})\|\boldsymbol{A}\|_F$.

It remains to bound $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\boldsymbol{\Theta}_0\|_{op}$. We first plug in $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}$, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}}$, and further

$$\dot{\boldsymbol{x}}_k = \boldsymbol{x}_k + \boldsymbol{\xi}_{k1}, \ \dot{\boldsymbol{y}}_k = \boldsymbol{y}_k + \boldsymbol{\xi}_{k2} = \boldsymbol{\Theta}_0^\top \boldsymbol{x}_k + \boldsymbol{\epsilon}_k,$$

some algebra yields

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\boldsymbol{\Theta}_0 = \frac{1}{n}\sum_{k=1}^n (\boldsymbol{x}_k + \boldsymbol{\xi}_{k1})(\boldsymbol{\epsilon}_k + \boldsymbol{\xi}_{k2})^\top$$
$$- \frac{1}{n}\sum_{k=1}^n \left(\boldsymbol{x}_k\boldsymbol{\xi}_{k1}^\top + (\boldsymbol{\xi}_{k1}\boldsymbol{\xi}_{k1}^\top - \frac{\delta_1^2}{4}\boldsymbol{I}_{d_1})\right)\boldsymbol{\Theta}_0 := \mathscr{T}_1 - \mathscr{T}_2. \tag{18}$$

Thus, $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\boldsymbol{\Theta}_0\|_{op} \leq \|\mathscr{T}_1\|_{op} + \|\mathscr{T}_2\|_{op}$.

(a) We consider the case of partial quantization ($\delta_1 = 0$). In this case $\boldsymbol{\xi}_{k1} = 0$, so $\mathscr{T}_2 = 0$ and we only need to bound $\|\mathscr{T}_1\|_{op}$ with $\mathscr{T}_1 = \frac{1}{n}\sum_k \boldsymbol{x}_k(\boldsymbol{\epsilon}_k + \boldsymbol{\xi}_{k2})^\top$. Note that

$$\|\boldsymbol{\epsilon}_k + \boldsymbol{\xi}_{k2}\|_{\psi_2} \leq \|\boldsymbol{\epsilon}_k\|_{\psi_2} + \|\boldsymbol{\xi}_{k2}\|_{\psi_2} \lesssim E + \delta_2,$$

$\|\boldsymbol{x}_k\|_{\psi_2} \leq K$, $\mathbb{E}(\boldsymbol{x}_k\boldsymbol{\epsilon}_k^\top) = \mathbb{E}(\boldsymbol{x}_k\boldsymbol{\xi}_{k2}^\top) = 0$, Lemma 3 guarantees the following to hold with probability at least $1 - 2\exp(-c_1(d_1 + d_2))$

$$\|\mathscr{T}_1\|_{op} \lesssim K(E + \delta_2)\sqrt{\frac{d_1 + d_2}{n}}. \tag{19}$$

Overall, we have

$$\langle \widehat{\boldsymbol{\Delta}}_c, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\boldsymbol{\Theta}_0\rangle$$
$$\lesssim K(E + \delta_2)\|\widehat{\boldsymbol{\Delta}}_c\|_F \sqrt{\frac{r(d_1 + d_2)}{n}}. \tag{20}$$

The result of part (a) follows by putting (15) and (20) into (14).

(b) We then consider the complete quantization case ($\delta_1 > 0$). Similarly to (a), we have the bound

$$\|\mathscr{T}_1\|_{op} \lesssim (K + \delta_1)(E + \delta_2)\sqrt{\frac{d_1 + d_2}{n}}.$$

So it remains to bound $\|\mathscr{T}_2\|_{op}$. Since $\|\boldsymbol{\Theta}_0\|_{op} \leq R$, and by Lemma 3, with the promised probability we have

$$\|\mathscr{T}_2\|_{op} \leq R\left\|\frac{1}{n}\sum_{k=1}^n \left(\boldsymbol{\xi}_{k1}\boldsymbol{\xi}_{k1}^\top - \frac{\delta_1^2}{4}\boldsymbol{I}_{d_1}\right)\right\|_{op}$$
$$+ R\left\|\frac{1}{n}\sum_{k=1}^n \boldsymbol{x}_k\boldsymbol{\xi}_{k1}^\top\right\|_{op} \lesssim R\delta_1(K + \delta_1)\sqrt{\frac{d_1 + d_2}{n}}. \tag{21}$$

By putting pieces similarly, we conclude the proof. $\square$

Several remarks are in order.

Remark 2: (Prediction error) As presented in Theorem 1, we will focus on the estimation of $\mathbf{\Theta}_0$ in this work, whereas in regression one may also be interested in the prediction performance. From $\|\widehat{\mathbf{\Delta}}_c\|_F$, the bound for prediction error is indeed immediate. For instance, when $\delta_1 = 0$, because with high probability $\|\widehat{\mathbf{\Sigma}}_{xx} - \mathbf{\Sigma}_{xx}\|_{op} \leq \frac{\kappa_0}{2}$ and $\lambda_{\max}(\widehat{\mathbf{\Sigma}}_{xx}) \leq \kappa_1$, one has

$$\frac{1}{n}\sum_{k=1}^{n}\|\widehat{\mathbf{\Delta}}_c^\top \boldsymbol{x}_k\|_2^2 = \langle \widehat{\mathbf{\Sigma}}_{xx}, \widehat{\mathbf{\Delta}}_c \widehat{\mathbf{\Delta}}_c^\top \rangle = O(\kappa_1 \|\widehat{\mathbf{\Delta}}_c\|^2).$$

Remark 3: (Compared to the least squares estimation) The ordinary least squares (OLS) estimator $\widehat{\mathbf{\Theta}}_{LS}$ is to minimize $\dot{\mathcal{L}}(\mathbf{\Theta})$ over $\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}$ without the nuclear norm constraint. This amounts to estimating d_2 columns of $\mathbf{\Theta}_0$ separately without utilizing their correlations. Under similar assumptions on covariate and noise, one can easily show $\|\widehat{\mathbf{\Theta}}_{LS} - \mathbf{\Theta}_0\|_F$ scales as $O\left(\sqrt{\frac{d_1 d_2}{n}}\right)$, which is essentially inferior to $O\left(\sqrt{\frac{r(d_1+d_2)}{n}}\right)$ in the case of $r \ll \min\{d_1, d_2\}$. This illustrates the benefit of incorporating the low-rank priori on $\mathbf{\Theta}_0$, which will be complemented by numerical example later (Fig. 5).

Remark 4: (Minimax optimality and the role of quantization) The non-asymptotic error bound $O\left(\sqrt{\frac{r(d_1+d_2)}{n}}\right)$ is minimax optimal compared to the information-theoretic lower bound in [53, Theorem 5] (also see [23, Remark 11], [24, Fact 1], [9, Page 12] for alternative statements). In fact, the quantization does not affect the order of (n, r, d_1, d_2) in the sample complexity and error bounds but only slightly worsens the multiplicative factors, i.e., A_1 in $n \gtrsim \left(\frac{A_1}{\kappa_0}\right)^2(d_1+d_2)$, A_2 in (12) and A_3 in (13). Thus, in a regime where the quantization levels δ_1, δ_2 are fixed, our result matches the one in a case without quantization up to multiplicative constant. In addition, δ_1 and E are on equal footing in A_2, hence the role of partial quantization can be nicely interpreted as additional sub-Gaussian noise. This extends similar findings in [57], [59], [66] from compressed sensing to LRMR. Further, a useful perspective is that the result for the setting without quantization can be recovered by letting $\delta_1, \delta_2 = 0$. For instance, when $\delta_2 = 0$ the bound in Theorem 1 reads as $O\left(\frac{KE}{\kappa_0}\sqrt{\frac{r(d_1+d_2)}{n}}\right)$, thus agreeing with the bound in [44, Coro. 3]. The above discussions regarding the role of quantization remain valid for our subsequent results.

C. Regularized Lasso

Since prior estimate on $\|\mathbf{\Theta}_0\|_{nu}$ is often unavailable, a more practically appealing recovery procedure is the following regularized Lasso given by

$$\widehat{\mathbf{\Theta}}_p = \underset{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}}{\arg\min}\ \dot{\mathcal{L}}(\mathbf{\Theta}) + \lambda\|\mathbf{\Theta}\|_{nu}, \qquad (22)$$

and we let $\widehat{\mathbf{\Delta}}_p$ be the estimation error. By properly tuning λ, the Regularized Lasso estimator $\widehat{\mathbf{\Theta}}_p$ achieves the same error rate as the previous $\widehat{\mathbf{\Theta}}_c$.

Theorem 2: (Regularized Lasso). We consider LRMR under Assumption 1 and the quantization procedure described above. We assume the scaling $n \gtrsim \left(\frac{A_1}{\kappa_0}\right)^2(d_1+d_2)$, where A_1 is the multiplicative factor in (11). Then for the estimator $\widehat{\mathbf{\Theta}}_p$ in (22) we have the following guarantees.

(a) (Partial Quantization). If $\delta_1 = 0$, we let $A_4 := K(E + \delta_2)$. Set $\lambda = C_1 A_4 \sqrt{\frac{d_1+d_2}{n}}$ with sufficiently large C_1, then with probability at least $1 - c_3 \exp(c_4(d_1 + d_2))$ it holds that

$$\|\widehat{\mathbf{\Delta}}_p\|_F \leq C_2 \kappa_0^{-1} A_4 \sqrt{\frac{r(d_1+d_2)}{n}}. \qquad (23)$$

(b) (Complete Quantization). If $\delta_2 > 0$, $\|\mathbf{\Theta}_0\|_{op} \leq R$, we let $A_5 := (K + \delta_1)(E + \delta_2 + R\delta_1)$. Set $\lambda = C_5 A_5 \sqrt{\frac{d_1+d_2}{n}}$ with sufficiently large C_5, then with probability at least $1 - c_7 \exp(-c_8 d_1)$ it holds that

$$\|\widehat{\mathbf{\Delta}}_p\|_F \leq C_6 \kappa_0^{-1} A_5 \sqrt{\frac{r(d_1+d_2)}{n}}.$$

By using some standard analyses for regularizer M-estimator (e.g., see [44]), the proof of Theorem 2 follows similar lines of Theorem 1. We defer the proof to the Appendix.

It is clear that we need an additional constraint on $\|\mathbf{\Theta}_0\|_{op}$ for the cases of complete quantization in Theorems 1–2, while this is not needed when we have access to the full-precision covariate. The following remark elaborates on this point.

Remark 5: (The norm constraint of $\mathbf{\Theta}_0$) When there is error in covariate, a norm constraint on the true parameter $\mathbf{\Theta}_0$ seems indispensable rather than an artifact from the proof technique. The main reason is that the error in covariate propagates along the true parameter, and hence its overall contribution to the response is proportional to $\mathbf{\Theta}_0$. Note that similar observation was also made in [42, Section 3.2] for corrected linear regression where the covariates suffer from zero-mean random noise with known covariance matrix.

IV. QUANTIZED LOW-RANK LINEAR REGRESSION MODEL WITH MATRIX RESPONSE

The proposed quantization scheme enjoys broader applicability, as we will show in this section that the dithered quantizer can be similarly applied to the problem of low-rank linear regression (L2RM) with matrix response [39]. In particular, such regression model finds application in imaging genetics, with matrix responses representing the weighted or binary adjacency matrix of a finite graph that characterizes structural or functional connectivity pattern, while the covariates are a set of genetic markers [43], [58], [63]. We would also like to note some recent advances on variable selection [29] and covariance estimation [69] for matrix-valued data.

Following the notation in [39], L2RM with matrix response can be formulated as

$$\boldsymbol{Y}_k = \sum_{i=1}^{s} x_{ki}\mathbf{\Theta}_0^{(i)} + \boldsymbol{E}_k,\ k = 1, ..., n, \qquad (24)$$

where $\boldsymbol{x}_k = [x_{k1}, ..., x_{ks}]^\top$ is the covariate, $\mathbf{\Theta}_0^{(i)} \in \mathbb{R}^{p \times q}$ are the true coefficient matrices, $\boldsymbol{E}_k, \boldsymbol{Y}_k \in \mathbb{R}^{p \times q}$ are respectively the noise matrix and response. Our goal is to estimate $\mathbf{\Theta}_0 = [\mathbf{\Theta}_0^{(1)}, ..., \mathbf{\Theta}_0^{(s)}] \in \mathbb{R}^{p \times (sq)}$ under moderately large s but p, q that

can be extremely huge.[4] Analogously to Assumption 1, for analysing the nuclear norm regularized M-estimator (see (27) below), we make the following sub-Gaussian data assumptions that relax the Gaussian ones in [39], (A9)-(A11).

Assumption 2: The assumptions on covariates $\boldsymbol{x}_k$'s are the same as Assumption 1; Independent of $\boldsymbol{x}_k$'s, the noise matrices $\boldsymbol{E}_1, ..., \boldsymbol{E}_n$ are i.i.d., zero-mean and sub-Gaussian with $\|\boldsymbol{E}_k\|_{\psi_2} := \sup_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \sup_{\boldsymbol{v} \in \mathbb{S}^{q-1}} \|\boldsymbol{u}^\top \boldsymbol{E}_k \boldsymbol{v}\|_{\psi_2} \le E$; the matrix responses $\boldsymbol{Y}_k$'s are generated from (24) for some $\boldsymbol{\Theta}_0$ satisfying $\sum_{i=1}^s \mathrm{rank}(\boldsymbol{\Theta}_0^{(i)}) \le r$.

Similarly, the dithered quantization for $(\boldsymbol{x}_k, \boldsymbol{Y}_k)$ is as follows: $\dot{\boldsymbol{x}}_k = \mathcal{Q}_{\delta_1}(\boldsymbol{x}_k + \boldsymbol{\phi}_k)$ for triangular random dither $\boldsymbol{\phi}_k$; $\dot{\boldsymbol{Y}}_k = \mathcal{Q}_{\delta_2}(\boldsymbol{Y}_k + \boldsymbol{\tau}_k)$ with uniform random dither $\boldsymbol{\tau}_k \sim \mathscr{U}([-\frac{\delta_2}{2}, \frac{\delta_2}{2}]^{p \times q})$. To be concise we only consider the more practical regularized Lasso. Based on the full data $(\boldsymbol{x}_k, \boldsymbol{Y}_k)$, [39] proposed the unconstrained convex program that minimizes

$$\underbrace{\frac{1}{n} \sum_{k=1}^n \left\| \boldsymbol{Y}_k - \sum_{i=1}^s x_{ki} \boldsymbol{\Theta}^{(i)} \right\|_F^2}_{\mathcal{L}_1(\boldsymbol{\Theta})} + \lambda \cdot \sum_{i=1}^s \|\boldsymbol{\Theta}^{(i)}\|_{nu}$$

over $\boldsymbol{\Theta} = [\boldsymbol{\Theta}^{(1)}, ..., \boldsymbol{\Theta}^{(s)}]$, where $\sum_i \|\boldsymbol{\Theta}^{(i)}\|_{nu}$ is the regularizer that incorporates the low-rankness structures of $\boldsymbol{\Theta}^{(i)}$'s. However, in our quantized regime one only observes $(\dot{\boldsymbol{x}}_k, \dot{\boldsymbol{Y}}_k)$ ($\dot{\boldsymbol{x}}_k = \boldsymbol{x}_k$ in partial quantization with $\delta_1 = 0$), modification of $\mathcal{L}_1(\boldsymbol{\Theta})$ is needed. By vectorization we first reformulate (24) as $\mathrm{vec}(\boldsymbol{Y}_k) = \sum_{i=1}^s x_{ki} \cdot \mathrm{vec}(\boldsymbol{\Theta}_0^{(i)}) + \mathrm{vec}(\boldsymbol{E}_k)$. Here, for $\boldsymbol{\Theta} = [\boldsymbol{\Theta}^{(1)}, ..., \boldsymbol{\Theta}^{(s)}]$ we define the rearrangement $\widetilde{\boldsymbol{\Theta}}$ as

$$\widetilde{\boldsymbol{\Theta}} = \begin{bmatrix} \mathrm{vec}(\boldsymbol{\Theta}^{(1)})^\top \\ \vdots \\ \mathrm{vec}(\boldsymbol{\Theta}^{(s)})^\top \end{bmatrix} \in \mathbb{R}^{s \times pq}, \qquad (25)$$

then we have

$$\mathrm{vec}(\boldsymbol{Y}_k) = \widetilde{\boldsymbol{\Theta}}_0^\top \boldsymbol{x}_k + \mathrm{vec}(\boldsymbol{E}_k) \qquad (26)$$

that agrees with (5). Now we can employ the prior developments — similar to (9) we let $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} := \frac{1}{n} \sum_{k=1}^n \dot{\boldsymbol{x}}_k \dot{\boldsymbol{x}}_k^\top - \frac{\delta_1^2}{4} \boldsymbol{I}_s$, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} = \frac{1}{n} \sum_{k=1}^n \dot{\boldsymbol{x}}_k \mathrm{vec}(\dot{\boldsymbol{Y}}_k)^\top$, and then change $\mathcal{L}_1(\boldsymbol{\Theta})$ to

$$\dot{\mathcal{L}}_1(\boldsymbol{\Theta}) = \langle \widetilde{\boldsymbol{\Theta}}\widetilde{\boldsymbol{\Theta}}^\top, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} \rangle - 2 \langle \widetilde{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} \rangle,$$

which can be constructed from the quantized data. Combining these pieces, we are in a position to define the Lasso estimator:

$$\widehat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \ \dot{\mathcal{L}}_1(\boldsymbol{\Theta}) + \lambda \cdot \sum_{i=1}^s \|\boldsymbol{\Theta}^{(i)}\|_{nu}$$
$$\text{subject to } \boldsymbol{\Theta} = [\boldsymbol{\Theta}^{(1)}, ..., \boldsymbol{\Theta}^{(s)}] \in \mathbb{R}^{p \times sq} \qquad (27)$$

We have the following theoretical guarantee for $\widehat{\boldsymbol{\Theta}}$.

Theorem 3: (Regularized Lasso). We consider L2RM with matrix response under Assumption 2 and the quantization procedure described above. We assume the scaling $n \gtrsim$

$\max\{s, p, q\}$ for some sufficiently large hidden constant and $\log s = O(p + q)$. Then for estimator $\widehat{\boldsymbol{\Theta}}$ in (27) we have the following guarantees.

(a) (Partial Quantization). If $\delta_1 = 0$, we let $A_6 := K(E + \delta_2)$. Set $\lambda = C_1 A_6 \sqrt{\frac{p+q}{n}}$ with sufficiently large C_1, then with probability at least $1 - \exp(-s) - c_3 \exp(-c_4(p + q))$ it holds that

$$\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F \le C_2 \kappa_0^{-1} A_6 \sqrt{\frac{r(p+q)}{n}}. \qquad (28)$$

(b) (Complete Quantization). If $\delta_1 > 0$, we further assume $\sum_{i=1}^s \|\boldsymbol{\Theta}_0^{(i)}\|_{op}^2 \le R^2$ for some $R > 0$, $s = O(p + q)$, and then let $A_7 := (K + \delta_1)(E + \delta_2 + R\delta_1)$. Set $\lambda = C_2 A_7 \sqrt{\frac{p+q}{n}}$ with sufficiently large C_2, then with probability at least $1 - \exp(-s) - c_1 \exp(-c_2(p + q))$ it holds that

$$\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F \le C_3 \kappa_0^{-1} A_7 \sqrt{\frac{r(p+q)}{n}}. \qquad (29)$$

Setting $\delta_2 = 0$ in Theorem 3(a) exactly recovers [39, Theorem 5]. While beyond the range of [39], our results clearly display how the dithered quantization affects the error bounds, i.e., slightly worse multiplicative factors (A_6, A_7). Specifically, when δ_1 and δ_2 are chosen and then fixed, the estimation error still scales as $O(\sqrt{\frac{r(p+q)}{n}})$, which matches the case without quantization up to multiplicative constant.

There are some technical differences between our proof and the one for [39, Theorem 5]. First, because we assume sub-Gaussian $(\boldsymbol{x}_k, \boldsymbol{E}_k)$ rather than the Gaussian ones as in [39], different arguments are required to proceed the proof. More specifically, Gaussian $(\boldsymbol{x}_k, \boldsymbol{E}_k)$ enables [39, Theorem 5] to use techniques from [49] like Anderson's comparison inequality (see [39, Lemma 4]) and tail bound of χ^2 random variable to bound $\|\frac{1}{n} \sum_{k=1}^n x_{ki} \boldsymbol{E}_k\|_{op}$. In contrast, this term is bounded via Lemma 5 in (32); besides handling sub-Gaussian $(\boldsymbol{x}_k, \boldsymbol{E}_k)$, Lemma 5 itself represents a cleaner way to bound this random term compared to the arguments in [39]. Second, in the "complete quantization" case, due to error in the covariate, there appears an additional random term in (33), (34), and to bound it we need to further assume $\sum_{i=1}^s \|\boldsymbol{\Theta}_0^{(i)}\|_{op}^2 \le R^2$ (as explained in Remark 5). We defer the detailed proof to the Appendix.

We give the following remark to compare (27) with the ordinary least squares method and the Lasso for LRMR based on the reformulation (26).

Remark 6: (Compared to OLS and LRMR via vectorization) For the estimator $\widehat{\boldsymbol{\Theta}}_{LS}$ defined by minimizing the empirical ℓ_2 loss over $\boldsymbol{\Theta} \in \mathbb{R}^{p \times sq}$, the error $\|\widehat{\boldsymbol{\Theta}}_{LS} - \boldsymbol{\Theta}_0\|_F$ would scale as $O(\sqrt{\frac{spq}{n}})$ even without quantization. By contrast, the deduced $O(\sqrt{\frac{r(p+q)}{n}})$ can be essentially better when $r \ll s \min\{p, q\}$. This illustrates the benefit of incorporating the low-rank structure. Moreover, if we impose low-rankness on $\widetilde{\boldsymbol{\Theta}}_0$ after vectorization (26), then by Theorem 2 the estimation error scales as $O(\sqrt{\frac{r_1 pq}{n}})$ (here, $r_1 = \mathrm{rank}(\widetilde{\boldsymbol{\Theta}}_0)$), which still suffers from the extremely large pq. Thus, the method in this section (also, as in [39]) achieves more effective dimension reduction in the case of matrix response.

---

[4]In fact, s in real applications can also be very large. For dimension reduction, [39] assumed $\boldsymbol{\Theta}_0^{(i)} = 0$ for most i's and developed a screening method to estimate those i's with non-zero $\boldsymbol{\Theta}_0^{(i)}$. We focus on the estimation after this screening step.
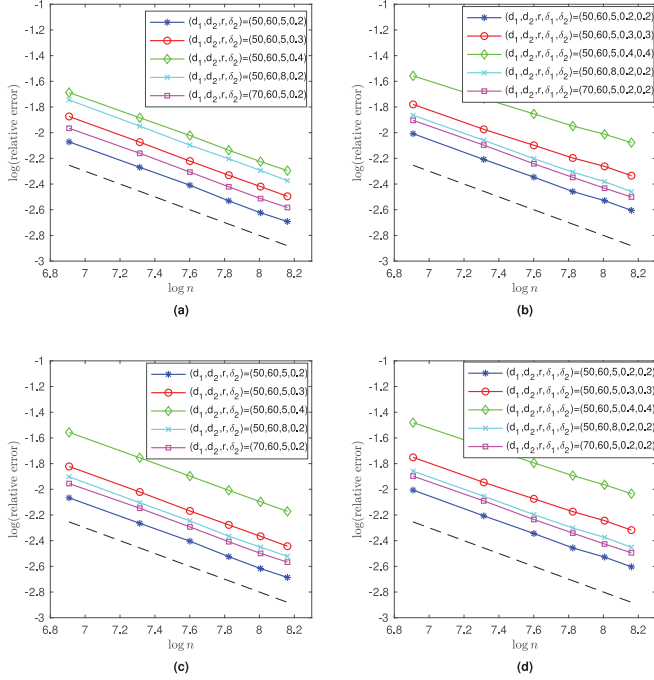
Fig. 1. (a) Constrained Lasso (partial quantization); (b) constrained Lasso (complete quantization); (c) regularized Lasso (partial quantization); (d) regularized Lasso (complete quantization).

V. EXPERIMENTAL RESULTS

In this section we provide experimental results to support and demonstrate our theoretical results. Otherwise specified, each data point is set to be the mean value of 50 independent trials.

A. Simulations With Synthetic Data

We first present simulation results on synthetic data. Our main purpose is to verify the established error rates, specifically $O\big(A\sqrt{\frac{r(d_1+d_2)}{n}}\big)$ in Theorems 1–2 and $O\big(A\sqrt{\frac{r(p+q)}{n}}\big)$ in Theorem 3, are in the correct order for characterizing the Lasso estimation errors. In particular, the dithered quantization only results in slightly larger multiplicative factor A. We will also demonstrate the important role played by the random dithering.

1) Constrained Lasso for Quantized LRMR: To simulate the setting of quantized LRMR we generate the low-rank underlying $\boldsymbol{\Theta}_0 \in \mathbb{R}^{d_1 \times d_2}$ as follows: we first generate $\boldsymbol{\Theta}_1 \in \mathbb{R}^{d_1 \times r}$, $\boldsymbol{\Theta}_2 \in \mathbb{R}^{r \times d_2}$ with i.i.d. standard Gaussian entries, and then use a rescaled version of $\boldsymbol{\Theta}_1\boldsymbol{\Theta}_2$ (with unit Frobenius norm) as $\boldsymbol{\Theta}_0$. To simulate the sub-Gaussian data in Assumption 1, for simplicity, we use $\boldsymbol{x}_k \sim \mathcal{N}(0, \boldsymbol{I}_{d_1})$ and $\boldsymbol{\epsilon}_k \sim \mathcal{N}(0, 0.1 \cdot \boldsymbol{I}_{d_2})$. The constrained Lasso is fed with $R = \|\boldsymbol{\Theta}_0\|_{nu}$ and optimized by an algorithm based on alternating direction method of multipliers (ADMM) [6]. To verify and demonstrate the error rate of $O\big(A\sqrt{\frac{r(d_1+d_2)}{n}}\big)$, we test different choices of $(d_1, d_2, r, \delta_1, \delta_2)$ under $n = 1000 : 500 : 3500$, with the log-log error plots displayed in Fig. 1. Firstly, the experimental curves are aligned with the dashed line that represents the decreasing rate of $n^{-1/2}$, thus confirming the order regarding the sample size. Then, to illustrate that quantization merely affects multiplicative factors, we compare the curves of $\delta_2 = 0.2, 0.3, 0.4$ in Fig. 1(a)
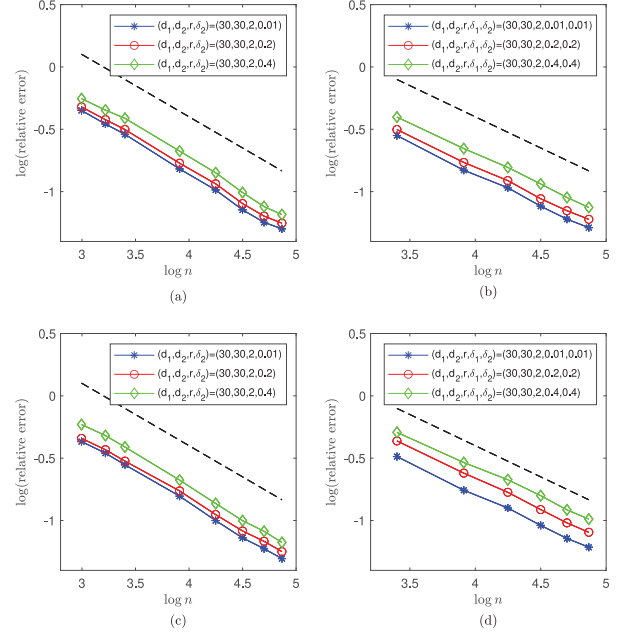

Fig. 2. (a) Constrained Lasso (partial quantization); (b) constrained Lasso (complete quantization); (c) regularized Lasso (partial quantization); (d) regularized Lasso (complete quantization).

(partial quantization) and the curves for $\delta_1 = \delta_2 = 0.2, 0.3, 0.4$ in Fig. 1(b) (complete quantization). Note that these curves are still parallel to each other, while the ones with larger δ_i are higher, which is consistent with our theory. Moreover, we note that increasing d_1 (from 50 to 70) or r (from 5 to 8) also leads to larger estimation errors. This is also predicted by the theoretical bound $O\big(\sqrt{\frac{r(d_1+d_2)}{n}}\big)$, that is, LRMR with more coefficients or weaker low-rank structure is harder.

2) Regularized Lasso for Quantized LRMR: We switch to the Regularized Lasso estimator, which is more practically appealing in that it does not requires a pre-estimate on $\|\boldsymbol{\Theta}\|_{nu}$. The choices of parameters, data generation and quantization are exactly the same as before. We follow the instruction in Theorem 2 for choosing λ in (22). That is, for each curve we slightly tune $C(\lambda)$ and then set $\lambda = C(\lambda)\sqrt{\frac{r(d_1+d_2)}{n}}$. We solve the regularized Lasso with ADMM algorithm and show the results in Fig. 1(c) and 1(d). Note that these results have implications similar to the previous ones for constrained Lasso, in terms of the $O(n^{-1/2})$ decreasing rate, the effect of quantization, problem size, low-rank structure. Thus, we do not repeat the demonstrations.

As suggested by an anonymous reviewer, we simulate quantized LRMR under a sample size n closer or even smaller than d_1, d_2. Specifically, we generate the low-rank $\boldsymbol{\Theta}_0 \in \mathbb{R}^{30 \times 30}$ using the same mechanism, and then test the constrained/regularized Lasso estimators under sample size $n = [20, 25, 30, 50, 70, 90, 110, 130]$ for partial quantization, or under $n = [30, 50, 70, 90, 110, 130]$ for complete quantization.[5] The results in Fig. 2 indicate that, using sample size close to

---

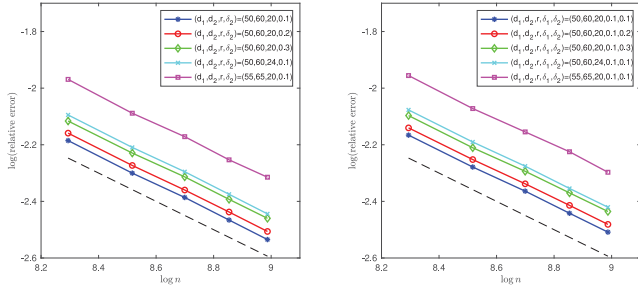[5]We do not test complete quantization under $n < d_1$ because this leads to non-convex program, see (9).

Fig. 3. (Left) Partial quantization; (right) complete quantization; we denote $(d_1, d_2) = (p, q)$ in the labels.

d_1 and d_2, the theoretical error bounds still characterize the estimation errors of our Lasso estimators fairly well.

3) Lasso for Quantized L2RM With Matrix Response: Now we move to the problem of low-rank linear regression with matrix response. Specifically, we set $s = 4$ in (24) and thus there are $\boldsymbol{\Theta}_0^{(1)}, ..., \boldsymbol{\Theta}_0^{(4)}$ as underlying coefficients matrices. We generate each $\boldsymbol{\Theta}_0^{(i)} \in \mathbb{R}^{p \times q}$ with rank $\frac{r}{4}$ as before. To fulfill Assumption 2, we adopt covariates $\boldsymbol{x}_k \sim \mathcal{N}(0, \boldsymbol{I}_s)$ and the noise matrices $\boldsymbol{E}_k \sim \mathcal{N}^{p \times q}(0, 0.01)$. We simulate different choices of $(p, q, r, \delta_1, \delta_2)$ under $n = 4000 : 1000 : 8000$. We note the following facts from the results in Fig. 3 that can support our theoretical error rate $O\left(\sqrt{\frac{r(p+q)}{n}}\right)$: all experimental curves decrease with n in a rate of $O(n^{-1/2})$; coarser quantization only lifts the curve a little bit; larger (p, q, r) results in larger estimation error.

4) The Importance of Dithering: As already analysed in Section I, under a direct uniform quantization without dithering, it is in general not possible to estimate the low-rank parameter matrix to arbitrarily small error. To demonstrate this, we use covariates with entries i.i.d. drawn from $\{\pm 1\}$-valued Bernoulli distribution to simulate LRMR with 50×60 underlying low-rank matrix given by

$$\boldsymbol{\Theta}_0 = \begin{bmatrix} \boldsymbol{A} & 0 \\ 0 & 0 \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.4 \end{bmatrix}, \boldsymbol{A} = \mathrm{diag}\big(\underbrace{\boldsymbol{B}, ..., \boldsymbol{B}}_{\text{ten } \boldsymbol{B}\text{'s}}\big).$$

Also, we simulate (24) with $s = 4$ and 50×60 true matrices

$$\boldsymbol{\Theta}_0^{(1)} = \begin{bmatrix} \frac{1}{2}\boldsymbol{C} & 0 \\ 0 & 0 \end{bmatrix}, \boldsymbol{\Theta}_0^{(2)} = \begin{bmatrix} \frac{2}{5}\boldsymbol{C} & 0 \\ 0 & 0 \end{bmatrix},$$

$$\boldsymbol{\Theta}_0^{(3)} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{2}\boldsymbol{C} \end{bmatrix}, \boldsymbol{\Theta}_0^{(4)} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{2}{5}\boldsymbol{C} \end{bmatrix},$$

where $\boldsymbol{C} = \mathrm{diag}(\boldsymbol{D}, ..., \boldsymbol{D}) \in \mathbb{R}^{10 \times 10}$ has five $\boldsymbol{D}$'s, and $\boldsymbol{D}$ is the 2×2 all-ones matrix. Under Gaussian noise, we quantize the responses with $\delta_2 = 1$ either under the uniform dither $\mathscr{U}[-\frac{1}{2}, \frac{1}{2}]$, or directly without dithering. Then we estimate the parameters via regularized Lasso under different sample sizes, the results are shown in Fig. 4. We find that, compared to a direct quantization, using dithering significantly reduces estimation errors; more prominently, the errors under dithering decrease at a sharp rate, whereas the curves without dithering reach some error floor where more data can no longer improve the estimation. We refer to [57, Fig. 1], [14, Fig. 5] for similar
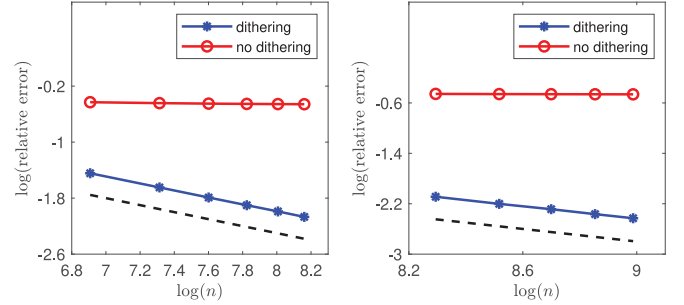

Fig. 4. We compare the quantized setting with or without random dithering. (Left) Quantized LRMR; (right) quantized L2RM with matrix response.

Fig. 5. Row 1: $n = 300$; row 2: $n = 400$. Column 1: The true image (peppers) and the singular values of its red channel (which indicates approximate low-rankness); column 2: Lasso with noisy unquantized data; column 3: Lasso with noisy quantized data; column 4: Least squares estimation with noisy quantized data.

experimental results in the contexts of compressed sensing, matrix completion, and covariance estimation.

B. Simulations of Image Restoration

Note that natural images are approximately low-rank[6] (e.g., [13], [17]), and our theoretical results can be easily extended to approximately low-rank case by slightly more work (e.g., [13], [23], [44]). To better visualize the effect of quantization, following prior work like [39], we conduct simulations with images as underlying low-rank matrices in this part.

1) Quantized LRMR: This numerical example simulates (5) with each channel of "Peppers" as $\boldsymbol{\Theta}_0$, aiming to test the effect of quantization in a relatively high-noise setting. We also demonstrate the advantage of LRMR over the ordinary least squares (OLS) estimation (see Remark 3). In the experiment, we separately deal with each channel, which is a 256×256 approximately low-rank matrix (see the left bottom of Fig. 5). Specifically, we draw entries of $\boldsymbol{x}_k$ from $\mathcal{N}(0, 1)$; let e be the average magnitude of the signal part $(\boldsymbol{\Theta}_0^\top \boldsymbol{x}_k)_{k=1}^n$, we use $\boldsymbol{\epsilon}_k \sim \frac{2e}{5} \cdot \mathcal{N}(0, \boldsymbol{I}_{256})$ to simulate a relatively large noise (signal-to-noise ratio less than 7); in the quantized setting, we use uniform dithering and quantize $\boldsymbol{y}_k$ with $\delta_2 = \frac{e}{8}$. Under $n = 300$ or $n = 400$, we test regularized Lasso with noisy unquantized/quantized $\boldsymbol{y}_k$, as well as OLS with noisy quantized $\boldsymbol{y}_k$. The results in Fig. 5 indicate that, quantization does not notably

[6]This means that its singular values decrease rapidly and only the first few are dominant.
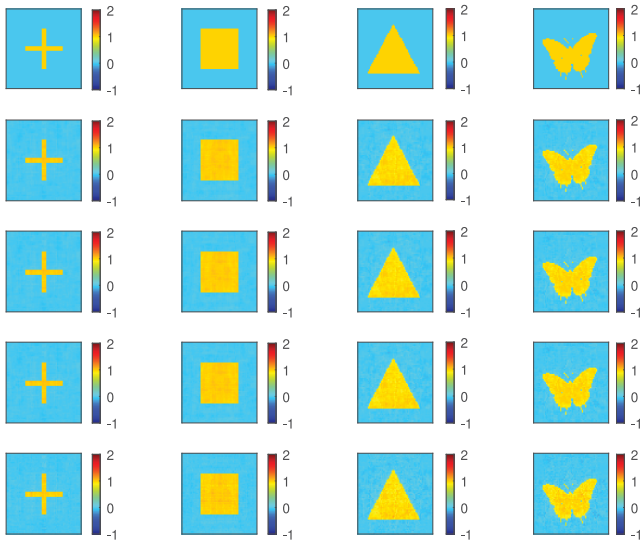
Fig. 6. $n = 2000$. (Row 1): The true images $\boldsymbol{\Theta}_0^{(i)}$; (rows 2–5): images reconstructed by regularized Lasso (27) under response quantization with $\delta_2 = 0.0$ (no quantization), $\delta_2 = 0.5$, $\delta_2 = 1.0$ and $\delta_2 = 3.0$, respectively.

TABLE I
$n = 2000$. MEAN RELATIVE FROBENIUS NORM ERRORS (STANDARD DEVIATION ($\times 10^{-3}$))

δ_2	$\boldsymbol{\Theta}_0^{(1)}$	$\boldsymbol{\Theta}_0^{(2)}$	$\boldsymbol{\Theta}_0^{(3)}$	$\boldsymbol{\Theta}_0^{(4)}$
0.0	0.0532(7.91)	0.0263(5.53)	0.0791(4.71)	0.0801(5.90)
0.5	0.0536(7.86)	0.0265(5.48)	0.0795(4.67)	0.0804(5.86)
1.0	0.0551(7.67)	0.0272(5.39)	0.0805(4.62)	0.0813(5.77)
3.0	0.0769(5.96)	0.0376(4.23)	0.0920(4.12)	0.0923(5.00)

harm the restoration (comparing columns 2 and 3). Moreover, in such a noisy and quantized setting, Lasso estimator significantly outperforms the OLS estimation that is ignorant of the low-rank structure (comparing columns 3 and 4).

2) Quantized L2RM With Matrix Response: We follow the experiment in [39, Figs. 1 and 2]. Specifically, we simulate (24) with $s = 4$ where $\boldsymbol{\Theta}_0^{(i)}$'s are 64×64 0-1 matrices and shown as images in the first row of Fig. 6. It can be verified that they are approximately low-rank. We also adopt the method of generating $(\boldsymbol{x}_k, \boldsymbol{E}_k)$ in [39]. While the experiment in [39] aims at comparing different methods of recovering $\boldsymbol{\Theta}_0^{(i)}$, however, our main goal here is to exhibit how quantization resolution affects the recovery. Thus, we simulate the regularized Lasso (27) under response quantization with $\delta_2 = 0.0, 0.5, 1.0, 3.0$. Under the sample size of $n = 2000$, the reconstructed images are shown in rows two through five in Fig. 6. We also run 100 independent trials and report the mean (relative) Frobenius norm error and standard deviation for each $\boldsymbol{\Theta}_0^{(i)}$ in Table I. It is clear both visually and on the mean error that, under quantization with relatively high resolution ($\delta_2 = 0.5, 1.0$), Lasso returns estimations fairly close to the ones obtained in a full-data regime. In fact, even if we quantize $\boldsymbol{Y}_k$ with $\delta_2 = 3$,[7] the Lasso estimator still delivers quite acceptable results. Therefore, we conclude that the dithered quantization will not significantly deteriorate one's

TABLE II
$n = 400$. MEAN RELATIVE FROBENIUS NORM ERRORS (STANDARD DEVIATION ($\times 10^{-3}$))

δ_2	$\boldsymbol{\Theta}_0^{(1)}$	$\boldsymbol{\Theta}_0^{(2)}$	$\boldsymbol{\Theta}_0^{(3)}$	$\boldsymbol{\Theta}_0^{(4)}$
0.0	0.1470(8.60)	0.0775(9.57)	0.1233(10.29)	0.1249(10.50)
0.5	0.1496(8.34)	0.0787(9.48)	0.1245(10.22)	0.1261(10.36)
1.0	0.1575(8.08)	0.0825(9.09)	0.1284(9.81)	0.1300(10.09)
3.0	0.2507(6.05)	0.1278(6.30)	0.1699(7.67)	0.1723(7.36)

ability to recover the underlying low-rank parameters; rather, the dithered uniform quantizer preserves the information fairly well. Generally speaking, there should be a trade-off between quantization resolution and recovery accuracy in practice. Note that the smaller sample size $n = 400$ is also simulated, see Table II for the results with similar implications.

C. A Real Data Application

To confirm the efficacy of the proposed method, we perform the quantization and estimation in a genetic association study for examining the regulatory control mechanisms in gene networks for isoprenoids in Arabidopsis thaliana [55], [65]. We adopt the LRMR model (5) with $\boldsymbol{x}_k$ being the expression levels of $d_1 = 39$ genes from the two isoprenoid biosynthesis pathways, $\boldsymbol{y}_k$ being the expression levels of $d_2 = 62$ genes from four downstream pathways, and we use $n = 115$ samples in total.[8] Besides, the mean magnitudes of the entries of $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ are 2160 and 3707, respectively.

We will focus on how the dithered quantization of $(\boldsymbol{x}_k, \boldsymbol{y}_k)$ affects the estimation and prediction of regularized Lasso (6). Note that the two major differences between this real data application and the previous simulations are that the data here may not be nicely captured by the sub-Gaussian distributions (Assumption 1), and that the relation between $\boldsymbol{x}_k, \boldsymbol{y}_k$ may not be perfectly modeled by LRMR (5). Thus, there is not an underlying $\boldsymbol{\Theta}_0$ serving as the ground truth. Alternatively, since the emphasis is on the effect of quantization, we use the Lasso estimator with suitable λ from unquantized data as $\boldsymbol{\Theta}_0$.

For partial quantization, we quantize $\boldsymbol{y}_k$ to $\dot{\boldsymbol{y}}_k$ under $\delta_2 = 0$: $100 : 1000$ and obtain $\boldsymbol{\Theta}_\delta$ from $(\boldsymbol{x}_k, \dot{\boldsymbol{y}}_k)$ as in (22), where the parameter λ increases with δ_2, as instructed by Theorem 2. The relative estimation error $\frac{\|\boldsymbol{\Theta}_\delta - \boldsymbol{\Theta}_0\|_F}{\|\boldsymbol{\Theta}_0\|_F}$ and relative prediction error $\frac{\|\boldsymbol{Y} - \boldsymbol{\Theta}_\delta^\top \boldsymbol{X}\|_F}{\|\boldsymbol{Y}\|_F}$ are reported as their mean values in 50 independent trials in Fig. 7(a) and 7(b). Specifically, the curves slowly increase with δ_2; compared to the unquantized case $\delta_2 = 0$, the estimation and prediction under the coarse quantization $\delta_2 = 1000$ are still acceptable. We also test the complete quantization setting where $\boldsymbol{x}_k$ is quantized to $\dot{\boldsymbol{x}}_k$ with $\delta_1 = 0 : 5 : 50$, $\boldsymbol{y}_k$ is quantized to $\dot{\boldsymbol{y}}_k$ with $\delta_2 = 0 : 50 : 500$. Similar results are reported in Fig. 7(c) and 7(d), but comparing Fig. 7(c) and 7(a), we also note that $\boldsymbol{\Theta}_\delta$ deviates from $\boldsymbol{\Theta}_0$ more significantly in complete quantization (even though $\delta_1 = 0 : 5 : 50$ is relatively small compared to the mean magnitude of $\boldsymbol{x}_k$); that is, the quantization of $\boldsymbol{x}_k$ affects the estimation more severely. Finally,

---

[7] This represents rather low resolution because in the simulation, entries of $\boldsymbol{Y}_k$ have magnitude about 1.6 in average.

[8] There are originally 118 samples in this real data study, but we remove 3 samples that are detected as (potential) outliers in [55], see Fig 1 therein.
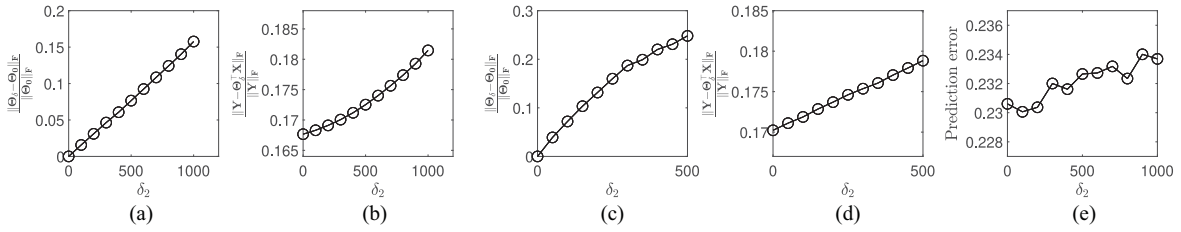
Fig. 7. (a, b) Estimation/prediction errors, partial quantization; (c, d) estimation/prediction errors, complete quantization; (e) prediction error over "testing data", partial quantization.

we conduct a more practical learning and prediction setting as follows: randomly dividing the columns of $\boldsymbol{X} \in \mathbb{R}^{39 \times 115}, \boldsymbol{Y} \in \mathbb{R}^{62 \times 115}$ into the "training data" $\boldsymbol{X}_1 \in \mathbb{R}^{39 \times 95}, \boldsymbol{Y}_1 \in \mathbb{R}^{62 \times 95}$ and the "testing data" $\boldsymbol{X}_2 \in \mathbb{R}^{39 \times 20}, \boldsymbol{Y}_2 \in \mathbb{R}^{62 \times 20}$, we quantize $\boldsymbol{Y}_1$ to $\dot{\boldsymbol{Y}}_1$ with $\delta_2 = 0 : 100 : 1000$ and use $(\boldsymbol{X}_1, \dot{\boldsymbol{Y}}_1)$ to obtain the estimator $\boldsymbol{\Theta}_\delta$ defined in (22), then we track the relative prediction error over the testing data, i.e., $\frac{\|\boldsymbol{Y}_2 - \boldsymbol{\Theta}_\delta^\top \boldsymbol{X}_2\|_F}{\|\boldsymbol{Y}_2\|_F}$, whose mean value in 50 independent trials is reported in Fig. 7(e). Compared to Fig. 7(b), the prediction error increases even more slowly with δ_2. In conclusion, our quantization scheme well preserves the data information for subsequent estimation and prediction procedures.

VI. CONCLUSION

This paper, for the first time, studied low-rank multivariate regression (LRMR) in a realistic setting that involves data quantization. We proposed to use the dithered uniform quantizer, associated with uniform dither for the response, or with triangular dither for the covariate. We proposed the Lasso estimators based on quantized data in a constrained or regularized manner. With the aid of random dithering, albeit losing information in quantization, our estimators achieve minimax optimal error rate. In fact, the derived error bounds demonstrate that the quantization only results in slightly worse multiplicative factors, which is reminiscent of similar results in quantized CS (Remark 4) and has been clearly observed in our simulations (e.g., Fig. 1). Moreover, we similarly applied the quantization scheme to a low-rank regression problem with matrix response and established the theoretical results accordingly. Experimental results were reported to complement our theoretical developments.

For future work, our first direction is to study LRMR under the more extreme 1-bit quantization, which only retains the sign of the data. Secondly, while we separately worked on LRMR and L2RM with matrix response in this paper, it would be of interest to attempt to unify their analyses, and ideally build a general theoretical framework for quantized multiresponse regression. Last but not least, it is desired to investigate whether our quantization method and theoretical results could be extended to a high-dimensional setting where $n < d_1$, which probably requires new machinery in the technical proofs and structure on $\boldsymbol{\Theta}_0$ beyond low-rankness.

APPENDIX

A. The Proof of Corollary 1

Proof: Following Lemma 1, the proof can be done by some elementary algebraic manipulation. For the first part of the

claim, we only need to verify both choices of τ_i satisfy the condition in Lemma 1(a): If $\tau_i \sim \mathscr{U}\left([-\frac{\delta}{2}, \frac{\delta}{2}]\right)$, then

$$
\begin{aligned}
\mathbb{E} \exp(\mathbf{i} u \tau_i) &= \int_{-\frac{\delta}{2}}^{\frac{\delta}{2}} \frac{1}{\delta}\left(\cos(ux) + \mathbf{i} \sin(ux)\right) \mathrm{d}x \\
&= \frac{2}{\delta u} \sin\left(\frac{\delta u}{2}\right),
\end{aligned}
$$

which obviously vanishes at $u = \frac{2\pi l}{\delta}$ for non-zero integer l; It is similar for triangular dither. For the second part of the claim, let us show the triangular dither satisfies the condition in Lemma 1(b). Let $Z \sim \mathscr{U}\left([-\frac{\delta}{2}, \frac{\delta}{2}]\right)$ be independent of τ_i, then

$$
\begin{aligned}
g(u) &:= \mathbb{E}(\exp(\mathbf{i} u Z)) \mathbb{E}(\exp(\mathbf{i} u \tau_i)) \\
&= \left[\mathbb{E}(\exp(\mathbf{i} u Z))\right]^3 = \left[\frac{2}{\delta u} \sin\left(\frac{\delta u}{2}\right)\right]^3.
\end{aligned}
$$

It is evident that $g''(u)$ contains a common factor $\sin\left(\frac{\delta u}{2}\right)$, thus $g''(\frac{2\pi l}{\delta}) = 0$ holds for any non-zero integer l. Hence,

$$
\mathbb{E}\xi_i^2 = \mathbb{E}(Z + \tau_i)^2 = \mathbb{E}Z^2 + \mathbb{E}\tau_i^2 = 3 \int_{-\delta/2}^{\delta/2} \frac{x^2}{\delta} \, \mathrm{d}x = \frac{\delta^3}{4},
$$

the proof is complete. $\qquad \square$

B. The Proof of Theorem 2

Proof: We start with the optimality of $\widehat{\boldsymbol{\Theta}}_p$

$$
\dot{\mathcal{L}}(\widehat{\boldsymbol{\Theta}}_p) + \lambda\|\widehat{\boldsymbol{\Theta}}_p\|_{nu} \leq \dot{\mathcal{L}}(\boldsymbol{\Theta}_0) + \lambda\|\boldsymbol{\Theta}_0\|_{nu}.
$$

Recall that $\widehat{\boldsymbol{\Delta}}_p = \widehat{\boldsymbol{\Theta}}_p - \boldsymbol{\Theta}_0$, by some algebra we arrive at

$$
\begin{aligned}
\langle \widehat{\boldsymbol{\Delta}}_p \widehat{\boldsymbol{\Delta}}_p^\top, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} \rangle \leq\; & 2\langle \widehat{\boldsymbol{\Delta}}_p, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\boldsymbol{\Theta}_0 \rangle \\
&+ \lambda\big(\|\boldsymbol{\Theta}_0\|_{nu} - \|\widehat{\boldsymbol{\Theta}}_p\|_{nu}\big). \quad (30)
\end{aligned}
$$

Note that the left-hand side is always non-negative (this holds deterministically when $\delta_1 = 0$, and holds within the promised probability when $\delta_1 > 0$, see *step 1* in the proof of Theorem 1). By (18), (19), (21) in the proof of Theorem 1, in both "partial quantization" and "complete quantization", our choices of λ can guarantee $\lambda \geq 4\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\boldsymbol{\Theta}_0\|_{op}$ holds under the promised probability. Under the same probability, we thus obtain

$$
0 \leq \frac{\lambda}{2}\|\widehat{\boldsymbol{\Delta}}_p\|_{nu} + \lambda\big(\|\boldsymbol{\Theta}_0\|_{nu} - \|\widehat{\boldsymbol{\Theta}}_p\|_{nu}\big),
$$

i.e., $\frac{1}{2}\|\widehat{\boldsymbol{\Delta}}_p\|_{nu} + \|\boldsymbol{\Theta}_0\|_{nu} \geq \|\widehat{\boldsymbol{\Theta}}_p\|_{nu}$. Recall that (17) can provide $\|\widehat{\boldsymbol{\Theta}}_p\|_{nu} - \|\boldsymbol{\Theta}_0\|_{nu} \geq \|\mathcal{P}_{\overline{\mathcal{M}}^\perp}\widehat{\boldsymbol{\Delta}}_p\|_{nu} - \|\mathcal{P}_{\overline{\mathcal{M}}}\widehat{\boldsymbol{\Delta}}_p\|_{nu}$, we deduce

$$
\big\|\mathcal{P}_{\overline{\mathcal{M}}^\perp}\widehat{\boldsymbol{\Delta}}_p\big\|_{nu} \leq \big\|\mathcal{P}_{\overline{\mathcal{M}}}\widehat{\boldsymbol{\Delta}}_p\big\|_{nu} + \frac{1}{2}\|\widehat{\boldsymbol{\Delta}}_p\|_{nu},
$$

where the involved subspaces and projections are defined in the proof of Theorem 1. Thus,

$$\|\widehat{\boldsymbol{\Delta}}_p\|_{nu} \le \|\mathcal{P}_{\overline{\mathcal{M}}}\widehat{\boldsymbol{\Delta}}_p\|_{nu} + \|\mathcal{P}_{\overline{\mathcal{M}}^\perp}\widehat{\boldsymbol{\Delta}}_p\|_{nu}$$
$$\le 2\|\mathcal{P}_{\overline{\mathcal{M}}}\widehat{\boldsymbol{\Delta}}_p\|_{nu} + \frac{1}{2}\|\widehat{\boldsymbol{\Delta}}_p\|_{nu},$$

which further gives

$$\|\widehat{\boldsymbol{\Delta}}_p\|_{nu} \le 4\|\mathcal{P}_{\overline{\mathcal{M}}}\widehat{\boldsymbol{\Delta}}_p\|_{nu} \le 4\sqrt{2r}\|\widehat{\boldsymbol{\Delta}}_p\|_F,$$

and the last inequality is because $\operatorname{rank}(\boldsymbol{A}) \le 2r$ if $\boldsymbol{A} \in \overline{\mathcal{M}}$.

Having deduced $\|\widehat{\boldsymbol{\Delta}}_p\|_{nu} \lesssim \sqrt{r}\|\widehat{\boldsymbol{\Delta}}_p\|_F$, we can upper bound the right-hand side of (30) by

$$\frac{\lambda}{2}\|\widehat{\boldsymbol{\Delta}}_p\|_{nu} + \lambda\big(\|\boldsymbol{\Theta}_0\|_{nu} - \|\widehat{\boldsymbol{\Theta}}_p\|_{nu}\big)$$
$$\le \frac{3}{2}\lambda\|\widehat{\boldsymbol{\Delta}}_p\|_{nu} \lesssim \lambda\sqrt{r}\|\widehat{\boldsymbol{\Delta}}_p\|_F.$$

As deduced in (15), within the promised probability the left-hand side of (30) is lower bounded by $\frac{\kappa_0}{2}\|\widehat{\boldsymbol{\Delta}}_p\|_F^2$. Thus, we arrive at $\|\widehat{\boldsymbol{\Delta}}_p\|_F \lesssim \kappa_0^{-1}\lambda\sqrt{r}$. To complete the proof, we only need to plug in the value of λ in both cases. $\qquad\square$

C. The Proof of Theorem 3

Proof: We define $\boldsymbol{\Delta} = \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0 \in \mathbb{R}^{p \times sq}$, $\boldsymbol{\Delta}^{(i)} = \widehat{\boldsymbol{\Theta}}^{(i)} - \boldsymbol{\Theta}_0^{(i)} \in \mathbb{R}^{p \times q}$, recall the rearrangement $\widetilde{\boldsymbol{\Theta}} \in \mathbb{R}^{s \times pq}$ defined in (25). We continue to use prior notation for quantization noise/error:

$$\dot{\boldsymbol{x}}_k = \boldsymbol{x}_k + \boldsymbol{\xi}_{k1} = \boldsymbol{x}_k + \boldsymbol{\phi}_k + \boldsymbol{w}_{k1}$$

and

$$\dot{\boldsymbol{Y}}_k = \boldsymbol{Y}_k + \boldsymbol{\xi}_{k2} = \boldsymbol{Y}_k + \boldsymbol{\tau}_k + \boldsymbol{w}_{k2}.$$

Now we use the definition and obtain $\dot{\mathcal{L}}_1(\widehat{\boldsymbol{\Theta}}) + \lambda \cdot \sum_{i=1}^s \|\widehat{\boldsymbol{\Theta}}\|_{nu} \le \dot{\mathcal{L}}_1(\boldsymbol{\Theta}_0) + \lambda \cdot \sum_{i=1}^s \|\boldsymbol{\Theta}_0\|_{nu}$. Then we perform some algebra to arrive at

$$\langle \widetilde{\boldsymbol{\Delta}}\widetilde{\boldsymbol{\Delta}}^\top, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\rangle \le 2\langle \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\widetilde{\boldsymbol{\Theta}}_0, \widetilde{\boldsymbol{\Delta}}\rangle$$
$$+ \lambda \cdot \sum_{i=1}^s \big(\|\boldsymbol{\Theta}_0^{(i)}\|_{nu} - \|\widehat{\boldsymbol{\Theta}}^{(i)}\|_{nu}\big). \quad (31)$$

Step 1. Bound the left-hand side from below.

This is exactly the same as *Step 1* in the proof of Theorem 1. In more detail, because $n \gtrsim s$, one can invoke Lemma 4 to show that $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}) \ge \frac{\kappa_0}{2}$ holds with probability at least $1 - 2\exp(-s)$. Assume that we are on this event, then evidently we have

$$\langle \widetilde{\boldsymbol{\Delta}}\widetilde{\boldsymbol{\Delta}}^\top, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\rangle \ge \frac{\kappa_0}{2}\|\widetilde{\boldsymbol{\Delta}}\|_F^2 = \frac{\kappa_0}{2}\|\boldsymbol{\Delta}\|_F^2.$$

Step 2. Bound $\mathcal{T} := \langle \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\widetilde{\boldsymbol{\Theta}}_0, \widetilde{\boldsymbol{\Delta}}\rangle$ *from above.*

Using $\operatorname{vec}(\boldsymbol{Y}_k) = \widetilde{\boldsymbol{\Theta}}_0^\top \boldsymbol{x}_k + \operatorname{vec}(\boldsymbol{E}_k)$ and the meaning of $\boldsymbol{\xi}_{k1}, \boldsymbol{\xi}_{k2}$, we can first simplify $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xy}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}\widetilde{\boldsymbol{\Theta}}_0$ to

$$\underbrace{\frac{1}{n}\sum_{k=1}^n \dot{\boldsymbol{x}}_k\big(\operatorname{vec}(\boldsymbol{E}_k) + \operatorname{vec}(\boldsymbol{\xi}_{k2})\big)^\top}_{\mathcal{T}_1}$$

$$\underbrace{- \frac{1}{n}\sum_{k=1}^n \Big(\dot{\boldsymbol{x}}_k\boldsymbol{\xi}_{k1}^\top - \frac{\delta_1^2}{4}\boldsymbol{I}_s\Big)\widetilde{\boldsymbol{\Theta}}_0}_{\mathcal{T}_2}.$$

Thus, we have $|\mathcal{T}| \le |\langle \mathcal{T}_1, \widetilde{\boldsymbol{\Delta}}\rangle| + |\langle \mathcal{T}_2, \widetilde{\boldsymbol{\Delta}}\rangle|$, and it amounts to estimating $|\langle \mathcal{T}_1, \widetilde{\boldsymbol{\Delta}}\rangle|$ and $|\langle \mathcal{T}_2, \widetilde{\boldsymbol{\Delta}}\rangle|$. For the first term, by turning back to the $\mathbb{R}^{p \times q}$ we have

$$|\langle \mathcal{T}_1, \widetilde{\boldsymbol{\Delta}}\rangle| = \Big|\Big\langle \frac{1}{n}\sum_{k=1}^n \dot{\boldsymbol{x}}_k\big(\operatorname{vec}(\boldsymbol{E}_k) + \operatorname{vec}(\boldsymbol{\xi}_{k2})\big)^\top, \widetilde{\boldsymbol{\Delta}}\Big\rangle\Big|$$

$$= \Big|\sum_{i=1}^s \Big\langle \frac{1}{n}\sum_{k=1}^n \dot{x}_{ki}(\boldsymbol{E}_k + \boldsymbol{\xi}_{k2}), \boldsymbol{\Delta}^{(i)}\Big\rangle\Big|$$

$$\le \Big(\sum_{i=1}^s \|\boldsymbol{\Delta}^{(i)}\|_{nu}\Big)\Big(\max_{i \in [s]}\Big\|\frac{1}{n}\sum_{k=1}^n \dot{x}_{ki}(\boldsymbol{E}_k + \boldsymbol{\xi}_{k2})\Big\|_{op}\Big)$$

$$\le C(K + \delta_1)(E + \delta_2)\sqrt{\frac{p+q}{n}} \cdot \Big(\sum_{i=1}^s \|\boldsymbol{\Delta}^{(i)}\|_{nu}\Big), \quad (32)$$

where in the last inequality we invoke Lemma 5 and a union bound over $i \in [s]$; it holds with probability at least $1 - 2\exp(-c(p+q))$ because $\log s = O(p+q)$. Note that the second term $|\langle \mathcal{T}_2, \widetilde{\boldsymbol{\Delta}}\rangle|$ vanishes in partial quantization ($\delta_1 = 0$), thus we estimate it on the complete quantization case ($\delta_1 > 0$) where we further assume $\sum_i \|\boldsymbol{\Theta}_0^{(i)}\|_{op}^2 \le R^2$ and $s = O(p + q)$. In particular, we define

$$\boldsymbol{\Psi} = [\psi_{ij}] = \frac{1}{n}\sum_{k=1}^n \Big(\dot{\boldsymbol{x}}_k\boldsymbol{\xi}_{k1}^\top - \frac{\delta_1^2}{4}\boldsymbol{I}_s\Big) \in \mathbb{R}^{s \times s},$$

and note that we have $\mathbb{E}\boldsymbol{\Psi} = 0$. Moreover Lemma 5 provides that, $\|\boldsymbol{\Psi}\|_{op} \lesssim (K + \delta_1)\delta_1\sqrt{\frac{s}{n}}$ holds with probability at least $1 - \exp(-s)$. On this event, we estimate that

$$|\langle \mathcal{T}_2, \widetilde{\boldsymbol{\Delta}}\rangle| = |\langle \boldsymbol{\Psi}\widetilde{\boldsymbol{\Theta}}_0, \widetilde{\boldsymbol{\Delta}}\rangle|$$

$$= \Big|\sum_{i=1}^s \Big\langle \sum_{j=1}^s \psi_{ij}\boldsymbol{\Theta}_0^{(j)}, \boldsymbol{\Delta}^{(i)}\Big\rangle\Big|$$

$$\le \Big(\max_{i \in [s]}\Big\|\sum_{j=1}^s \psi_{ij}\boldsymbol{\Theta}_0^{(j)}\Big\|_{op}\Big)\Big(\sum_{i=1}^s \|\boldsymbol{\Delta}^{(i)}\|_{nu}\Big)$$

$$\le C(K + \delta_1)R\delta_1\sqrt{\frac{p+q}{n}}\Big(\sum_{i=1}^s \|\boldsymbol{\Delta}^{(i)}\|_{nu}\Big), \quad (33)$$

where the last inequality is because for $i \in [s]$,

$$\Big\|\sum_j \psi_{ij}\boldsymbol{\Theta}_0^{(j)}\Big\|_{op} \le \sum_j |\psi_{ij}|\|\boldsymbol{\Theta}_0^{(j)}\|_{op}$$

$$\le \Big(\sum_j \psi_{ij}^2\Big)^{1/2}\Big(\sum_j \|\boldsymbol{\Theta}_0^{(j)}\|_{op}^2\Big)^{1/2}$$

$$\le R\|\boldsymbol{\Psi}\|_{op} = O\Big((K + \delta_1)R\delta_1\sqrt{\frac{s}{n}}\Big) \quad (34)$$

also recall that $s = O(p + q)$.

To conclude, in "partial quantization" we have shown $\mathcal{T} = O\big(A_6\sqrt{\frac{p+q}{n}}\big)$, and in "complete quantization" $\mathcal{T} = O\big(A_7\sqrt{\frac{p+q}{n}}\big)$. Compared to our choices of λ, we can

assume $2\mathscr{T} \leq \frac{1}{2}\lambda\left(\sum_i \|\boldsymbol{\Delta}^{(i)}\|_{nu}\right)$ with the promised probability. Because the left-hand side of (31) is non-negative (deterministically if $\delta_1 = 0$, with the promised probability if $\delta_1 > 0$), and $\lambda > 0$, we arrive at

$$\sum_{i=1}^{s}\left(\|\widehat{\boldsymbol{\Theta}}^{(i)}\|_{nu} - \|\boldsymbol{\Theta}_0^{(i)}\|_{nu}\right) \leq \frac{1}{2}\sum_{i=1}^{s}\|\boldsymbol{\Delta}^{(i)}\|_{nu}. \quad (35)$$

Step 3. Conclude the proof.

We use a decomposability argument. In particular, we let $r_i = \mathrm{rank}(\boldsymbol{\Theta}_0^{(i)})$, and exactly the same as the definition of $(\mathcal{M}, \overline{\mathcal{M}}, \overline{\mathcal{M}}^\perp)$ at the beginning of *Step 2* in the proof of Theorem 1 (regarding $\boldsymbol{\Theta}_0$ thereof), we now define $(\mathcal{M}_i, \overline{\mathcal{M}}_i, \overline{\mathcal{M}}_i^\perp)$ regarding $\boldsymbol{\Theta}_0^{(i)}$. Similarly, we have the decomposability

$$\|\mathcal{P}_{\mathcal{M}_i}\boldsymbol{A} + \mathcal{P}_{\overline{\mathcal{M}}_i^\perp}\boldsymbol{B}\|_{nu} = \|\mathcal{P}_{\mathcal{M}_i}\boldsymbol{A}\|_{nu} + \|\mathcal{P}_{\overline{\mathcal{M}}_i^\perp}\boldsymbol{B}\|_{nu}$$

holds for all $i \in [s]$ and $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{p\times q}$. Thus, we can use (17) to obtain

$$\|\widehat{\boldsymbol{\Theta}}^{(i)}\|_{nu} - \|\boldsymbol{\Theta}_0^{(i)}\|_{nu} \geq \|\mathcal{P}_{\overline{\mathcal{M}}_i^\perp}\boldsymbol{\Delta}^{(i)}\|_{nu} - \|\mathcal{P}_{\overline{\mathcal{M}}_i}\boldsymbol{\Delta}^{(i)}\|_{nu}.$$

Putting this into the left-hand side of (35), and also apply

$$\|\boldsymbol{\Delta}^{(i)}\|_{nu} \leq \|\mathcal{P}_{\overline{\mathcal{M}}_i^\perp}\boldsymbol{\Delta}^{(i)}\|_{nu} + \|\mathcal{P}_{\overline{\mathcal{M}}_i}\boldsymbol{\Delta}^{(i)}\|_{nu}$$

to the right-hand side, it provides $\sum_i\|\mathcal{P}_{\overline{\mathcal{M}}_i^\perp}\boldsymbol{\Delta}^{(i)}\|_{nu} \leq 3\sum_i\|\mathcal{P}_{\overline{\mathcal{M}}_i}\boldsymbol{\Delta}^{(i)}\|_{nu}$, which leads to

$$\sum_i \|\boldsymbol{\Delta}^{(i)}\|_{nu} \leq 4\sum_i\|\mathcal{P}_{\overline{\mathcal{M}}_i}\boldsymbol{\Delta}^{(i)}\|_{nu} \leq 4\sum_i\sqrt{2r_i}\|\boldsymbol{\Delta}^{(i)}\|_F$$

$$\leq 4\sqrt{2}\left(\sum_i r_i\right)^{1/2}\left(\sum_i\|\boldsymbol{\Delta}^{(i)}\|_F^2\right)^{1/2} \leq 4\sqrt{2r}\cdot\|\boldsymbol{\Delta}\|_F.$$

Now we are ready to put pieces together. Because $\|\boldsymbol{\Theta}_0^{(i)}\|_{nu} - \|\widehat{\boldsymbol{\Theta}}^{(i)}\|_{nu} \leq \|\boldsymbol{\Delta}^{(i)}\|_{nu}$, overall, the right-hand side of (31) has the bound $O\left(\lambda\sum_i\|\boldsymbol{\Delta}^{(i)}\|_{nu}\right) = O\left(\sqrt{r}\lambda\|\boldsymbol{\Delta}\|_F\right)$, while the left-hand side is lower bounded by $\frac{\kappa_0}{2}\|\boldsymbol{\Delta}\|_F^2$, so it holds with the promised probability that, $\|\boldsymbol{\Delta}\|_F = O\left(\frac{\sqrt{r}\lambda}{\kappa_0}\right)$. The proof can be concluded by using the chosen value of λ. $\quad\square$

D. Auxiliary Facts

1) The Proof of Lemma 3:

Proof: The proof is a standard covering argument for controlling the matrix operator norm. We construct $\mathcal{N}_1 \subset \mathbb{S}^{d_1-1}$ as a $\frac{1}{4}$-net of $\mathbb{S}^{d_1-1}$, meaning that for any $\boldsymbol{v} \in \mathbb{S}^{d_1-1}$ there exists $\boldsymbol{x} \in \mathcal{N}_1$ such that $\|\boldsymbol{x} - \boldsymbol{v}\|_2 \leq \frac{1}{4}$. Similarly, let $\mathcal{N}_2$ be a $\frac{1}{4}$-net of $\mathbb{S}^{d_2-1}$. By [62, Corollary 4.2.13] we can assume $|\mathcal{N}_1| \leq 9^{d_1}$, $|\mathcal{N}_2| \leq 9^{d_2}$. Note that for any $\boldsymbol{u} \in \mathcal{N}_1$, $\boldsymbol{v} \in \mathcal{N}_2$, we have

$$\left\|\boldsymbol{u}^\top\boldsymbol{a}_k\boldsymbol{b}_k^\top\boldsymbol{v} - \mathbb{E}(\boldsymbol{u}^\top\boldsymbol{a}_k\boldsymbol{b}_k^\top\boldsymbol{v})\right\|_{\psi_1}$$

$$\overset{(i)}{\leq} C_1\left\|(\boldsymbol{u}^\top\boldsymbol{a}_k)(\boldsymbol{v}^\top\boldsymbol{b}_k)\right\|_{\psi_1}$$

$$\overset{(ii)}{\leq} C_1\|\boldsymbol{u}^\top\boldsymbol{a}_k\|_{\psi_2}\|\boldsymbol{v}^\top\boldsymbol{b}_k\|_{\psi_2} \leq C_1 E_1 E_2 \quad (36)$$

Note that (i) is due to centering [62, Exercise 2.7.10], and we use (4) in (ii). Thus, we can use Bernstein's inequality (see [62, Theorem 2.8.1]) to obtain the concentration of $\frac{1}{n}\sum_k\left(\boldsymbol{u}^\top\boldsymbol{a}_k\boldsymbol{b}_k^\top\boldsymbol{v} - \mathbb{E}(\boldsymbol{u}^\top\boldsymbol{a}_k\boldsymbol{b}_k^\top\boldsymbol{v})\right)$; Followed by a union

bound over $(\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{N}_1 \times \mathcal{N}_2$, then for any $t > 0$

$$\mathbb{P}\left(\sup_{\boldsymbol{u}\in\mathcal{N}_1}\sup_{\boldsymbol{v}\in\mathcal{N}_2}\left|\frac{1}{n}\sum_k\left(\boldsymbol{u}^\top\boldsymbol{a}_k\boldsymbol{b}_k^\top\boldsymbol{v} - \mathbb{E}(\boldsymbol{u}^\top\boldsymbol{a}_k\boldsymbol{b}_k^\top\boldsymbol{v})\right)\right| \geq t\right)$$

$$\leq 2\exp\left((d_1 + d_2)\log 9 - c_2 n\min\left\{\left(\frac{t}{E_1 E_2}\right)^2, \frac{t}{E_1 E_2}\right\}\right). \quad (37)$$

We take $t = C_3 E_1 E_2\sqrt{\frac{d_1+d_2}{n}}$ with sufficiently large C_3, recall that $n \geq d_1 + d_2$, then the event

$$\sup_{\boldsymbol{u}\in\mathcal{N}_1}\sup_{\boldsymbol{v}\in\mathcal{N}_2}\left|\frac{1}{n}\sum_k\left(\boldsymbol{u}^\top\boldsymbol{a}_k\boldsymbol{b}_k^\top\boldsymbol{v} - \mathbb{E}(\boldsymbol{u}^\top\boldsymbol{a}_k\boldsymbol{b}_k^\top\boldsymbol{v})\right)\right|$$

$$\leq C_3 E_1 E_2\sqrt{\frac{d_1 + d_2}{n}} \quad (38)$$

holds with probability at least $1 - \exp(-c_4(d_1 + d_2))$. Note that [62, Exercise 4.4.3] gives $\|\frac{1}{n}\sum_k\{\boldsymbol{a}_k\boldsymbol{b}_k^\top - \mathbb{E}(\boldsymbol{a}_k\boldsymbol{b}_k^\top)\}\|_{op} \leq 2\cdot$ (the left hand side of (38)), the proof is complete. $\quad\square$

2) A Lemma for the Proof of Theorem 3:

Lemma 5: Assume $a_1, ..., a_n \in \mathbb{R}$ are independent and satisfy $\max_k\|a_k\|_{\psi_2} \leq K$; $\boldsymbol{B}_1, ..., \boldsymbol{B}_n \in \mathbb{R}^{p\times q}$ are independent and satisfy $\sup_{\boldsymbol{u}\in\mathbb{R}^{p-1}}\sup_{\boldsymbol{v}\in\mathbb{R}^{q-1}}\|\boldsymbol{u}^\top\boldsymbol{B}_k\boldsymbol{v}\|_{\psi_2} \leq E$ for each k. Assume $n \gtrsim p + q$, then it holds with probability at least $1 - 2\exp(-c(p+q))$ that,

$$\left\|\frac{1}{n}\sum_{k=1}^{n}\left\{a_k\boldsymbol{B}_k - \mathbb{E}(a_k\boldsymbol{B}_k)\right\}\right\|_{op} \leq CKE\sqrt{\frac{p+q}{n}}.$$

Proof: Similarly to that of Lemma 3, the proof is essentially a standard covering argument for controlling operator norm of random matrix. For simplicity we assume $\mathbb{E}(a_k\boldsymbol{B}_k) = 0$; the proof extends to $\mathbb{E}(a_k\boldsymbol{B}_k) \neq 0$ by simple centering technique [62, Exercise 2.7.10]. We invoke a covering argument: let $\mathcal{N}_1, \mathcal{N}_2$ be the $\frac{1}{4}$-net of $\mathbb{S}^{p-1}$, $\mathbb{S}^{q-1}$, respectively; we can assume $|\mathcal{N}_1| \leq 9^p, |\mathcal{N}_2| \leq 9^q$. By [62, Exercise 4.4.3] we have

$$\left\|\frac{1}{n}\sum_{k=1}^{n}a_k\boldsymbol{B}_k\right\|_{op} \leq 2\sup_{\boldsymbol{u}\in\mathcal{N}_1}\sup_{\boldsymbol{v}\in\mathcal{N}_2}\frac{1}{n}\sum_{k=1}^{n}a_k\boldsymbol{u}^\top\boldsymbol{B}_k\boldsymbol{v}. \quad (39)$$

For fixed $\boldsymbol{u}, \boldsymbol{v}$,

$$\|a_k\boldsymbol{u}^\top\boldsymbol{B}_k\boldsymbol{v}\|_{\psi_1} \leq \|a_k\|_{\psi_2}\|\boldsymbol{u}^\top\boldsymbol{B}_k\boldsymbol{v}\|_{\psi_2} \leq KE.$$

Thus, we can apply Bernstein's inequality [62, Theorem 2.8.1], together with a union bound on $\mathcal{N}_1 \times \mathcal{N}_2$, to obtain that for any $t > 0$,

$$\mathbb{P}\left(\sup_{\boldsymbol{u}\in\mathcal{N}_1}\sup_{\boldsymbol{v}\in\mathcal{N}_2}\frac{1}{n}\sum_{k=1}^{n}a_k\boldsymbol{u}^\top\boldsymbol{B}_k\boldsymbol{v} \geq t\right)$$

$$\leq 2\exp\left((p+q)\log 9 - cn\cdot\min\left\{\frac{t^2}{K^2 E^2}, \frac{t}{KE}\right\}\right).$$

We set $t = CKE\sqrt{\frac{p+q}{n}}$ with sufficiently large C, recall that we assume $n \gtrsim p + q$, we obtain that with probability at least $1 - 2\exp(-c_1(p+q))$,

$$\sup_{\boldsymbol{u}\in\mathcal{N}_1}\sup_{\boldsymbol{v}\in\mathcal{N}_2}\frac{1}{n}\sum_{k=1}^{n}a_k\boldsymbol{u}^\top\boldsymbol{B}_k\boldsymbol{v} \leq CKE\sqrt{\frac{p+q}{n}}.$$

Combined with (39), the result follows. $\quad\square$

References

[1] C. W. Anderson, E. A. Stolz, and S. Shamsunder, "Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 3, pp. 277–286, Mar. 1998.

[2] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learn.*, vol. 73, pp. 243–272, 2008.

[3] T. C. Aysal, M. J. Coates, and M. G. Rabbat, "Distributed average consensus with dithered quantization," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4905–4918, Oct. 2008.

[4] S. A. Bhaskar, "Probabilistic low-rank matrix completion from quantized measurements," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2131–2164, 2016.

[5] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *Proc. 42nd Annu. Conf. Inf. Sci. Syst.* Piscataway, NJ, USA: IEEE, 2008, pp. 16–21.

[6] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends® Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[7] E. N. Brown, R. E. Kass, and P. P. Mitra, "Multiple neural spike train data analysis: State-of-the-art and future challenges," *Nature Neurosci.*, vol. 7, no. 5, pp. 456–461, 2004.

[8] F. Bunea, Y. She, and M. Wegkamp, "Adaptive rank penalized estimators in multivariate regression," FL, USA, Tech. Rep., 2010. [Online]. Available: https://stat.ufl.edu/wp-content/uploads/sites/120/WegkampABSTRACT2011.pdf

[9] F. Bunea, Y. She, and M. H. Wegkamp, "Optimal selection of reduced rank estimators of high-dimensional matrices," *Ann. Statist.*, vol. 39, no. 2, pp. 1282–1309, 2011. [Online]. Available: http://www.jstor.org/stable/29783674

[10] T. Cai and W.-X. Zhou, "A max-norm constrained minimization approach to 1-bit matrix completion," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 3619–3647, 2013.

[11] E. J. Candes and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2342–2359, Apr. 2011.

[12] R. Caruana, "Multitask Learning," *Machine Learning*, vol. 28, pp. 41–75, 1997.

[13] J. Chen and M. K. Ng, "Color image inpainting via robust pure quaternion matrix completion: Error bound and weighted loss," *SIAM J. Imag. Sci.*, vol. 15, no. 3, pp. 1469–1498, 2022.

[14] J. Chen, M. K. Ng, and D. Wang, "Quantizing heavy-tailed data in statistical estimation: (Near) Minimax rates, covariate quantization, and uniform recovery," 2022, *arXiv:2212.14562*.

[15] J. Chen, C.-L. Wang, M. K. Ng, and D. Wang, "High dimensional statistical estimation under uniformly dithered one-bit quantization," *IEEE Trans. Inf. Theory*, vol. 69, no. 8, pp. 5151–5187, Aug. 2023.

[16] K. Chen, H. Dong, and K.-S. Chan, "Reduced rank regression via adaptive nuclear norm penalization," *Biometrika*, vol. 100, no. 4, pp. 901–920, 2013.

[17] Y. Chen, X. Xiao, and Y. Zhou, "Low-rank quaternion approximation for color image processing," *IEEE Trans. Image Process.*, vol. 29, pp. 1426–1439, 2020.

[18] A. Danaee, R. C. de Lamare, and V. H. Nascimento, "Distributed quantization-aware RLS learning with bias compensation and coarsely quantized signals," *IEEE Trans. Signal Process.*, vol. 70, pp. 3441–3455, 2022.

[19] M. A. Davenport, Y. Plan, E. Van Den Berg, and M. Wootters, "1-bit matrix completion," *Inf. Inference: J. IMA*, vol. 3, no. 3, pp. 189–223, 2014.

[20] S. Dirksen, "Quantized compressed sensing: A survey," in *Compressed Sensing and Its Applications*. Springer, 2019, pp. 67–95. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-73074-5_2#citeas

[21] S. Dirksen, J. Maly, and H. Rauhut, "Covariance estimation under one-bit quantization," *Ann. Stat.*, vol. 50, no. 6, pp. 3538–3562, 2022.

[22] S. Dirksen and S. Mendelson, "Non-Gaussian hyperplane tessellations and robust one-bit compressed sensing," *J. Eur. Math. Soc.*, vol. 23, no. 9, pp. 2913–2947, 2021.

[23] J. Fan, W. Wang, and Z. Zhu, "A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery," *Ann. Stat.*, vol. 49, no. 3, p. 1239, 2021.

[24] C. Giraud, "Low rank multivariate regression," *Electron. J. Stat.*, vol. 5, pp. 775–799, 2011.

[25] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[26] R. M. Gray and T. G. Stockham, "Dithered quantizers," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 805–812, May 1993.

[27] N. Hamdi and M. Bayati, "On low-rank trace regression under general sampling distribution," *J. Mach. Learn. Res.*, vol. 23, no. 1, pp. 14424–14472, 2022.

[28] O. A. Hanna, Y. H. Ezzeldin, C. Fragouli, and S. Diggavi, "Quantization of distributed data for learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 3, pp. 987–1001, Sep. 2021.

[29] M. Hao, L. Qu, D. Kong, L. Sun, and H. Zhu, "Optimal minimax variable selection for large-scale matrix linear regression model," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 6499–6537, 2021.

[30] L. Harrison, W. D. Penny, and K. Friston, "Multivariate autoregressive modeling of fMRI time series," *Neuroimage*, vol. 19, no. 4, pp. 1477–1491, 2003.

[31] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6869–6898, 2017.

[32] B. Jacob et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2704–2713.

[33] N. Jayant and L. Rabiner, "The application of dither to the quantization of speech signals," *Bell Syst. Tech. J.*, vol. 51, no. 6, pp. 1293–1304, 1972.

[34] A. Kipnis, Y. C. Eldar, and A. J. Goldsmith, "Fundamental distortion limits of analog-to-digital compression," *IEEE Trans. Inf. Theory*, vol. 64, no. 9, pp. 6013–6033, Sep. 2018.

[35] O. Klopp, J. Lafond, É. Moulines, and J. Salmon, "Adaptive multinomial matrix completion," *Electron. J. Stat.*, vol. 9, no. 2, pp. 2950–2975, 2015.

[36] A. Koloskova, S. Stich, and M. Jaggi, "Decentralized stochastic optimization and gossip algorithms with compressed communication," in *Proc. Int. Conf. Mach. Learn.* Long Beach, CA, USA: PMLR, 2019, pp. 3478–3487.

[37] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *Ann. Stat.*, vol. 39, no. 5, pp. 2302–2329, 2011.

[38] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2016, *arXiv:1610.05492*.

[39] D. Kong, B. An, J. Zhang, and H. Zhu, "L2RM: Low-rank linear regression models for high-dimensional matrix responses," *J. Am. Stat Assoc.*, vol. 115, no. 529, pp. 403–424, 2019, doi: 10.1080/01621459.2018.1555092.

[40] J. S. Lim, *Two-Dimensional Signal and Image Processing*. Englewood Cliffs, NJ, USA: Prentice Hall, 1990.

[41] J. Limb, "Design of dither waveforms for quantized visual signals," *Bell Syst. Tech. J.*, vol. 48, no. 7, pp. 2555–2582, 1969.

[42] P.-L. Loh and M. J. Wainwright, "Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima," *J. Mach. Learn. Res.*, vol. 16, no. 19, pp. 559–616, 2015.

[43] S. E. Medland, N. Jahanshad, B. M. Neale, and P. M. Thompson, "Whole-genome analyses of whole-brain data: Working within an expanded search space," *Nat. Neurosci.*, vol. 17, no. 6, pp. 791–800, 2014.

[44] S. Negahban and M. J. Wainwright, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *Ann. Stat.*, vol. 39, no. 2, pp. 1069–1097, 2011.

[45] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 1665–1697, 2012.

[46] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers," *Stat. Sci.*, vol. 27, no. 4, pp. 538–557, 2012.

[47] Y. Plan and R. Vershynin, "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 482–494, Jan. 2013.

[48] Y. Plan and R. Vershynin, "The generalized lasso with non-linear observations," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1528–1537, Mar. 2016.

[49] G. Raskutti, M. Yuan, and H. Chen, "Convex regularization for high-dimensional multiresponse tensor regression," *Ann. Stat.*, vol. 47, no. 3, pp. 1554–1584, 2019.

[50] G. C. Reinsel, R. P. Velu, and K. Chen, *Multivariate Reduced-Rank Regression: Theory, Methods and Applications*. Englewood Cliffs, NY, USA: Springer Nature, 2023, vol. 225.

[51] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization," in *Proc. Int. Conf. Artif. Intell. Stat.* PMLR, 2020, pp. 2021–2031.

[52] L. Roberts, "Picture coding using pseudo-random noise," *IRE Trans. Inf. Theory*, vol. 8, no. 2, pp. 145–154, 1962.

[53] A. Rohde and A. B. Tsybakov, "Estimation of high-dimensional low-rank matrices," *Ann. Stat.*, vol. 39, no. 2, pp. 887–930, 2011.

[54] L. Schuchman, "Dither signals and their effect on quantization noise," *IEEE Trans. Commun. Technol.*, vol. 12, no. 4, pp. 162–165, Dec. 1964.

[55] Y. She and K. Chen, "Robust reduced-rank regression," *Biometrika*, vol. 104, no. 3, pp. 633–647, 2017.

[56] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "UVeQFed: Universal vector quantization for federated learning," *IEEE Trans. Signal Process.*, vol. 69, pp. 500–514, 2020.

[57] Z. Sun, W. Cui, and Y. Liu, "Quantized corrupted sensing with random dithering," *IEEE Trans. Signal Process.*, vol. 70, pp. 600–615, 2022.

[58] P. M. Thompson, T. Ge, D. C. Glahn, N. Jahanshad, and T. E. Nichols, "Genetics of the connectome," *Neuroimage*, vol. 80, pp. 475–488, 2013.

[59] C. Thrampoulidis and A. S. Rawat, "The Generalized Lasso for sub-Gaussian measurements with dithered quantization," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2487–2500, Apr. 2020.

[60] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat Soc.: Ser. B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[61] N. Vaswani, "Nonconvex structured phase retrieval: A focus on provably correct approaches," *IEEE Signal Process. Mag.*, vol. 37, no. 5, pp. 67–77, Sep. 2020.

[62] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*. Cambridge, U.K.: Cambridge Univ. Press, 2018, vol. 47.

[63] C. Wen, H. Ba, W. Pan, M. Huang, and A. D. N. Initiative, "Co-sparse reduced-rank regression for association analysis between imaging phenotypes and genetic variants," *Bioinformatics*, vol. 36, no. 21, pp. 5214–5222, 2020.

[64] B. Widrow and I. Kollár, *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[65] A. Wille et al., "Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana," *Genome Biol.*, vol. 5, no. 11, pp. 1–13, 2004.

[66] C. Xu and L. Jacques, "Quantized compressive sensing with RIP matrices: The benefit of dithering," *Inf. Inference: J. IMA*, vol. 9, no. 3, pp. 543–586, 2020.

[67] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro, "Dimension reduction and coefficient estimation in multivariate linear regression," *J. R. Stat Soc.: Ser. B (Stat. Methodol.)*, vol. 69, no. 3, pp. 329–346, 2007.

[68] H. Zhang, J. Li, K. Kara, D. Alistarh, J. Liu, and C. Zhang, "ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning," in *Proc. Int. Conf. Mach. Learn*. Sydney, Australia: PMLR, 2017, pp. 4035–4043.

[69] Y. Zhang, W. Shen, and D. Kong, "Covariance estimation for matrix-valued data," *J. Am. Stat. Assoc.*, pp. 1–12, 2022, doi: 10.1080/01621459.2022.2068419.

[70] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, "DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients," 2016, *arXiv:1606.06160*.

Junren Chen received the B.Sc. degree in mathematics and applied mathematics from Sun Yat-sen University. He is currently pursuing the Ph.D. degree with the Department of Mathematics, The University of Hong Kong. He received a Hong Kong Ph.D. Fellowship from Hong Kong Research Grants Council for supporting the Ph.D. study. His research interests include compressed sensing, high-dimensional statistics, signal and image processing, quantization, and optimization.

Yueqi Wang received the B.S. degree from Zhejiang University, Zhejiang, China, in 2021. She is currently pursuing the Ph.D. degree with the University of Hong Kong, Hong Kong, China. Her major research interests include photonic dispersion relation reconstruction, topological optimization, and machine learning.

Michael K. Ng (Senior Member, IEEE) received the B.Sc. and M.Phil. degrees from The University of Hong Kong, Hong Kong, in 1990 and 1992, respectively, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 1995. From 1995 to 1997, he was a Research Fellow with the Computer Sciences Laboratory, The Australian National University, Canberra, ACT, Australia. He was an Assistant Professor/Associate Professor with The University of Hong Kong, from 1997 to 2005. He was a Professor/Chair Professor, from 2005 to 2019, with the Department of Mathematics, Hong Kong Baptist University, Hong Kong, and a Chair Professor, from 2019 to 2023, with the Department of Mathematics, The University of Hong. He is currently a Chair Professor in mathematics and a Chair Professor in data science with Hong Kong Baptist University. His research interests include applied and computational mathematics, machine learning and artificial intelligence, and data science. He serves as an editorial board member of several international journals. He was selected for the 2017 Class of Fellows of the Society for Industrial and Applied Mathematics. He received the Feng Kang Prize for his significant contributions to scientific computing.