The Annals of Statistics
2019, Vol. 47, No. 3, 1554–1584
https://doi.org/10.1214/18-AOS1725
© Institute of Mathematical Statistics, 2019

CONVEX REGULARIZATION FOR HIGH-DIMENSIONAL MULTIRESPONSE TENSOR REGRESSION

BY GARVESH RASKUTTI[*,1], MING YUAN[*,†,2] AND HAN CHEN[*,2]

University of Wisconsin-Madison[*] *and Columbia University*[†]

In this paper, we present a general convex optimization approach for solving high-dimensional multiple response tensor regression problems under low-dimensional structural assumptions. We consider using convex and weakly decomposable regularizers assuming that the underlying tensor lies in an unknown low-dimensional subspace. Within our framework, we derive general risk bounds of the resulting estimate under fairly general dependence structure among covariates. Our framework leads to upper bounds in terms of two very simple quantities, the *Gaussian width* of a convex set in tensor space and the *intrinsic dimension* of the low-dimensional tensor subspace. To the best of our knowledge, this is the first general framework that applies to multiple response problems. These general bounds provide useful upper bounds on rates of convergence for a number of fundamental statistical models of interest including multiresponse regression, vector autoregressive models, low-rank tensor models and pairwise interaction models. Moreover, in many of these settings we prove that the resulting estimates are minimax optimal. We also provide a numerical study that both validates our theoretical guarantees and demonstrates the breadth of our framework.

1. Introduction. Many modern scientific problems involve solving high-dimensional statistical problems where the sample size is small relative to the ambient dimension of the underlying parameter to be estimated. Over the past few decades, there has been a large amount of work on solving such problems by imposing low-dimensional structure on the parameter of interest. In particular sparsity, low-rankness and other low-dimensional subspace assumptions have been studied extensively both in terms of the development of fast algorithms and theoretical guarantees; see, for example, [4] and [9], for an overview. Most of the prior work has focused on scenarios in which the parameter of interest is a vector or matrix. Increasingly common in practice, however, the parameter or object to be estimated naturally has a higher-order tensor structure. Examples include hyperspectral image analysis [12], multienergy computed tomography [27], radar signal processing [19], audio classification [15] and text mining [6] among numerous

Received April 2017; revised May 2018.

[1] Supported in part by NSF Grant DMS-1407028.

[2] Supported in part by NSF FRG Grant DMS-1265202 and NIH Grant 1-U54AI117924-01.

MSC2010 subject classifications. 60K35.

Key words and phrases. Tensor regression, convex regularization, Gaussian width, intrinsic dimension, low-rank, sparsity.

others. It is much less clear how the low dimensional structures inherent to these problems can be effectively accounted for. The main purpose of this article is to fill in this void and provide a general and unifying framework for doing so.

Consider a general tensor regression problem where covariate tensors $X^{(i)} \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ and response tensors $Y^{(i)} \in \mathbb{R}^{d_{M+1} \times \cdots \times d_N}$ are related through

$$Y^{(i)} = \langle X^{(i)}, T \rangle + \varepsilon^{(i)}, \qquad i = 1, 2, \ldots, n. \tag{1.1}$$

Here, $T \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ is an unknown parameter of interest, and $\varepsilon^{(i)}$s are independent and identically distributed noise tensors whose entries are independent and identically distributed centred normal random variables with variance σ^2. Further, for simplicity we assume the covariates $(X^{(i)})_{i=1}^n$ are Gaussian, but with fairly general dependence assumptions. The notation $\langle \cdot, \cdot \rangle$ will refer throughout this paper to the standard inner product taken over appropriate Euclidean spaces. Hence, for $A \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ and $B \in \mathbb{R}^{d_1 \times \cdots \times d_N}$,

$$\langle A, B \rangle = \sum_{j_1=1}^{d_1} \cdots \sum_{j_M=1}^{d_M} A_{j_1,\ldots,j_M} B_{j_1,\ldots,j_M} \in \mathbb{R}$$

is the usual inner product if $M = N$; and if $M < N$, then $\langle A, B \rangle \in \mathbb{R}^{d_{M+1} \times \cdots \times d_N}$ such that its $(j_{M+1}, \ldots, j_N)$ entry is given by

$$\big(\langle A, B \rangle\big)_{j_{M+1},\ldots,j_N} = \sum_{j_1=1}^{d_1} \cdots \sum_{j_M=1}^{d_M} A_{j_1,\ldots,j_M} B_{j_1,\ldots,j_M,j_{M+1},\ldots,j_N}.$$

The goal of tensor regression is to estimate the coefficient tensor T based on observations $\{(X^{(i)}, Y^{(i)}) : 1 \leq i \leq n\}$. In addition to the canonical example of tensor regression with Y a scalar response (i.e., $M = N$), many other commonly encountered regression problems are also special cases of the general tensor regression model (1.1). Multiresponse regression (see, e.g., [1]), vector autoregressive model (see, e.g., [13]), and pairwise interaction tensor model (see, e.g., [25]) are some of the notable examples. In this article, we provide a general treatment to these seemingly different problems.

Our main focus here is on situations where the dimensionality d_k's are large when compared with the sample size n. In many practical settings, the true regression coefficient tensor T may have certain types of low-dimensional structure. Because of the high ambient dimension of a regression coefficient tensor, it is essential to account for such a low-dimensional structure when estimating it. Sparsity and low-rankness are the most common examples of such low-dimensional structures. In the case of tensors, sparsity could occur at the entry-wise level, fiber-wise level or slice-wise level, depending on the context and leading to different interpretations. There are also multiple ways in which low-rankness may be present when it comes to higher-order tensors, either at the original tensor level or at the *matricized* tensor level.

In this article, we consider a general class of convex regularization techniques to exploit either type of low-dimensional structure. In particular, we consider the standard convex regularization framework

$$\widehat{T} \in \underset{A \in \mathbb{R}^{d_1 \times \cdots \times d_N}}{\arg\min} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \| Y^{(i)} - \langle A, X^{(i)} \rangle \|_{\mathrm{F}}^2 + \lambda \mathcal{R}(A) \right\}, \tag{1.2}$$

where the regularizer $\mathcal{R}(\cdot)$ is a norm on $\mathbb{R}^{d_1 \times \cdots \times d_N}$, and $\lambda > 0$ is a tuning parameter. Hereafter, for a tensor A, $\|A\|_{\mathrm{F}} = \langle A, A \rangle^{1/2}$. We derive general risk bounds for a family of so-called *weakly decomposable* regularizers under fairly general dependence structure among the covariates. These general upper bounds apply to a number of concrete statistical inference problems including the aforementioned multiresponse regression, high-dimensional vector autoregressive models, low-rank tensor models and pairwise interaction tensors where we show that they are typically optimal in the minimax sense.

In developing these general results, we make several contributions to a fast growing literature on high-dimensional tensor estimation. First of all, we provide a unified and principled approach to exploit the low-dimensional structure in these tensor problems. In doing so, we incorporate an extension of the notion of decomposability originally introduced by [18] for vector and matrix models to *weak decomposability* previously introduced in [31] which allows us to handle more delicate tensor models such as the nuclear norm regularization for low-rank tensor models. Moreover, we provide, for the regularized least squared estimate given by (1.2), a general risk bound under an easily interpretable condition on the design tensor. The risk bound we derive is presented in terms of merely two geometric quantities, the *Gaussian width* which depends on the choice of regularization and the *intrinsic dimension* of the subspace that the tensor T lies in. We believe this is the first general framework that applies to multiple responses and general dependence structure for the covariate tensor X. Finally, our general results lead to novel upper bounds for several important regression problems involving high-dimensional tensors: multiresponse regression, multivariate autoregressive models and pairwise interaction models, for which we also prove that the resulting estimates are minimiax rate optimal with appropriate choices of regularizers.

Our framework incorporates both tensor structure and multiple responses which present a number of challenges compared to previous approaches. These challenges manifest themselves both in terms of the choice of regularizer $\mathcal{R}$ and the technical challenges in the proof of the main result. First, since the notion of low-dimensional is more generic for tensors meaning there are a number of choices of convex regularizer $\mathcal{R}$ and these must satisfy a form of weak decomposability and provide optimal rates. Multiple responses and the flexible dependence structure among the covariates also present significant technical challenges for proving restricted strong convexity, a key technical tool for establishing rates of convergence. In particular, a one-sided uniform law (Lemma 1.2 in the Supplementary Material

[23]) is required instead of classical techniques as developed in, for example, [17, 22] that only apply to univariate responses.

The remainder of the paper is organized as follows: In Section 2, we introduce the general framework of using weakly decomposable regularizers for exploiting low-dimensional structures in high-dimensional tensor regression. In Section 3, we present a general upper bound for weakly decomposable regularizers and discuss specific risk bounds for commonly used sparsity or low-rankness regularizers for tensors. In Section 4, we apply our general result to three specific statistical problems, namely, the multiresponse regression, multivariate autoregressive model and the pairwise interaction model. We show that in each of the three examples appropriately chosen weakly decomposable regularizers leads to minimax optimal estimation of the unknown parameters. Numerical experiments are presented in Section 5 to further demonstrate the merits and breadth of our approach. Proofs are deferred to the Supplementary Material [23].

2. Methodology.

Recall that the regularized least-squares estimate is given by

$$\widehat{T} = \underset{A \in \mathbb{R}^{d_1 \times \cdots \times d_N}}{\arg\min} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \| Y^{(i)} - \langle A, X^{(i)} \rangle \|_{\mathrm{F}}^2 + \lambda \mathcal{R}(A) \right\}.$$

For brevity, we assume implicitly hereafter that the minimizer on its left- hand side is uniquely defined. Our development here actually applies to the more general case where $\widehat{T}$ can be taken as an arbitrary element from the set of the minimizers. Of particular interest here is the so-called *weakly decomposable* convex regularizers, extending a similar concept introduced by [18] for vectors and matrices.

Let $\mathcal{A}$ be an arbitrary linear subspace of $\mathbb{R}^{d_1 \times \cdots \times d_N}$ and $\mathcal{A}^\perp$ its orthogonal complement:

$$\mathcal{A}^\perp := \{ A \in \mathbb{R}^{d_1 \times \cdots \times d_N} \mid \langle A, B \rangle = 0 \text{ for all } B \in \mathcal{A} \}.$$

We call a regularizer $\mathcal{R}(\cdot)$ weakly decomposable with respect to a pair $(\mathcal{A}, \mathcal{B})$ where $\mathcal{B} \subseteq \mathcal{A}$ if there exist a constant $0 < c_{\mathcal{R}} \leq 1$ such that for any $A \in \mathcal{A}^\perp$ and $B \in \mathcal{B}$,

$$\mathcal{R}(A + B) \geq \mathcal{R}(A) + c_{\mathcal{R}} \mathcal{R}(B). \tag{2.1}$$

In particular, if (2.1) holds for any $B \in \mathcal{B} = \mathcal{A}$, we say $\mathcal{R}(\cdot)$ is weakly decomposable with respect to $\mathcal{A}$. A more general version of this concept was first introduced in [31]. Because $\mathcal{R}$ is a norm, by triangular inequality, we also have

$$\mathcal{R}(A + B) \leq \mathcal{R}(A) + \mathcal{R}(B).$$

Many of the commonly used regularizers for tensors are weakly decomposable or decomposable for short. When $c_{\mathcal{R}} = 1$, our definition of decomposability naturally extends from similar notion for vectors ($N = 1$) and matrices ($N = 2$) introduced by [18]. We also allow for more general choices of $c_{\mathcal{R}}$ here to ensure a wider applicability. For example, as we shall see the popular tensor nuclear norm regularizer is decomposable with respect to appropriate linear subspaces with $c_{\mathcal{R}} = 1/2$, but not decomposable if $c_{\mathcal{R}} = 1$.

We have now described a catalogue of commonly used regularizers for tensors and argue that they are all decomposable with respect to appropriately chosen subspaces of $\mathbb{R}^{d_1 \times \cdots \times d_N}$. To fix ideas, we shall focus in what follows on estimating a third-order tensor T, that is, $N = 3$, although our discussion can be straightforwardly extended to higher-order tensors.

2.1. *Sparsity regularizers.* An obvious way to encourage entry-wise sparsity is to impose the vector ℓ_1 penalty on the entries of A:

$$\mathcal{R}(A) := \sum_{j_1=1}^{d_1} \sum_{j_2=1}^{d_2} \sum_{j_3=1}^{d_3} |A_{j_1 j_2 j_3}|, \tag{2.2}$$

following the same idea as the Lasso for linear regression (see, e.g., [29]). This is a canonical example of decomposable regularizers. For any fixed $I \subset [d_1] \times [d_2] \times [d_3]$ where $[d] = \{1, 2, \ldots, d\}$, write

$$\mathcal{A}(I) = \mathcal{B}(I) = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : A_{j_1 j_2 j_3} = 0 \text{ for all } (j_1, j_2, j_3) \notin I\}. \tag{2.3}$$

It is clear that

$$\mathcal{A}^{\perp}(I) = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : A_{j_1 j_2 j_3} = 0 \text{ for all } (j_1, j_2, j_3) \in I\},$$

and $\mathcal{R}(A)$ defined by (2.2) is decomposable with respect to $\mathcal{A}$ with $c_{\mathcal{R}} = 1$.

In many applications, sparsity arises with a more structured fashion for tensors. For example, a fiber or a slice of a tensor is likely to be zero simultaneously. Mode-1 fibers of a tensor $A \in \mathcal{R}^{d_1 \times d_2 \times d_3}$ are the collection of d_1-dimensional vectors

$$\{A_{\cdot j_2 j_3} = (A_{1 j_2 j_3}, \ldots, A_{d_1 j_2 j_3})^{\top} : 1 \le j_2 \le d_2, 1 \le j_3 \le d_3\}.$$

Mode-2 and -3 fibers can be defined in the same fashion. To fix ideas, we focus on mode-1 fibers. Sparsity among mode-1 fibers can be exploited using the group-based ℓ_1 regularizer:

$$\mathcal{R}(A) = \sum_{j_2=1}^{d_2} \sum_{j_3=1}^{d_3} \|A_{\cdot j_2 j_3}\|_{\ell_2}, \tag{2.4}$$

similar to the group Lasso (see, e.g., [32]), where $\|\cdot\|_{\ell_2}$ stands for the usual vector ℓ_2 norm. Similar to the vector ℓ_1 regularizer, the group ℓ_1-based regularizer is also decomposable. For any fixed $I \subset [d_2] \times [d_3]$, write

$$\mathcal{A}(I) = \mathcal{B}(I) = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : A_{j_1 j_2 j_3} = 0 \text{ for all } (j_2, j_3) \notin I\}. \tag{2.5}$$

It is clear that

$$\mathcal{A}^{\perp}(I) = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : A_{j_1 j_2 j_3} = 0 \text{ for all } (j_2, j_3) \in I\},$$

and $\mathcal{R}(A)$ defined by (2.4) is decomposable with respect to $\mathcal{A}$ with $c_{\mathcal{R}} = 1$. Note that in defining the regularizer in (2.4), instead of vector ℓ_2 norm, other ℓ_q $(q > 1)$ norms could also be used; see, for example, [30].

Sparsity could also occur at the slice level. The $(1, 2)$ slices of a tensor $A \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ are the collection of $d_1 \times d_2$ matrices

$$\{A_{\cdot\cdot j_3} = (A_{j_1 j_2 j_3})_{1 \le j_1 \le d_1, 1 \le j_2 \le d_2} : 1 \le j_3 \le d_3\}.$$

Let $\|\cdot\|$ be an arbitrary norm on $d_1 \times d_2$ matrices. Then the following group regularizer can be considered:

$$\mathcal{R}(A) = \sum_{j_3=1}^{d_3} \|A_{\cdot\cdot j_3}\|. \tag{2.6}$$

Typical examples of the matrix norm that can be used in (2.6) include Frobenius norm and nuclear norm among others. In the case when $\|\cdot\|_{\mathrm{F}}$ is used, $\mathcal{R}(\cdot)$ is again a decomposable regularizer with respect to

$$\mathcal{A}(I) = \mathcal{B}(I) = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : A_{j_1 j_2 j_3} = 0 \text{ for all } j_3 \notin I\}, \tag{2.7}$$

for any $I \subset [d_3]$.

Now consider the case when we use the matrix nuclear norm $\|\cdot\|_*$ in (2.6). Let P_{1j} and P_{2j}, $j = 1, \dots, d_3$ be two sequences of projection matrices on $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Let

$$\begin{aligned} &\mathcal{A}(P_{1j}, P_{2j} : 1 \le j \le d_3) \\ &\quad = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : P_{1j}^{\perp} A_{\cdot\cdot j} P_{2j}^{\perp} = 0, j = 1, \dots, d_3\} \end{aligned} \tag{2.8}$$

and

$$\begin{aligned} &\mathcal{B}(P_{1j}, P_{2j} : 1 \le j \le d_3) \\ &\quad = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : A_{\cdot\cdot j} = P_{1j} A_{\cdot\cdot j} P_{2j}, j = 1, \dots, d_3\}. \end{aligned} \tag{2.9}$$

By pinching inequality (see, e.g., [3]), it can be derived that $\mathcal{R}(\cdot)$ is decomposable with respect to $\mathcal{A}(P_{1j}, P_{2j} : 1 \le j \le d_3)$ and $\mathcal{B}(P_{1j}, P_{2j} : 1 \le j \le d_3)$.

2.2. *Low-rankness regularizers.*

In addition to sparsity, one may also consider tensors with low-rank. There are multiple notions of rank for higher-order tensors; see, for example, [11], for a recent review. In particular, the so-called CP rank is defined as the smallest number r of rank-one tensors needed to represent a tensor $A \in \mathbb{R}^{d_1 \times d_2 \times d_3}$:

$$A = \sum_{k=1}^{r} u_k \otimes v_k \otimes w_k, \tag{2.10}$$

where $u_k \in \mathbb{R}^{d_1}$, $v_k \in \mathbb{R}^{d_2}$ and $w_k \in \mathbb{R}^{d_3}$. To encourage a low rank estimate, we can consider the nuclear norm regularization. Following [33], we define the nuclear norm of A through its dual norm. More specifically, let the spectral norm of A be given by

$$\|A\|_s = \max_{\|u\|_{\ell_2}, \|v\|_{\ell_2}, \|w\|_{\ell_2} \le 1} \langle A, u \otimes v \otimes w \rangle.$$

Then its nuclear norm is defined as

$$\|A\|_* = \max_{\|B\|_s \le 1} \langle A, B \rangle.$$

We shall then consider the regularizer:

$$\mathcal{R}(A) = \|A\|_*. \tag{2.11}$$

We now show this is also a weakly decomposable regularizer.

Let P_k be a projection matrix in $\mathbb{R}^{d_k}$. Define

$$(P_1 \otimes P_2 \otimes P_3)A = \sum_{k=1}^{r} P_1 u_k \otimes P_2 v_k \otimes P_3 w_k.$$

Write

$$Q = P_1 \otimes P_2 \otimes P_3 + P_1^\perp \otimes P_2 \otimes P_3 + P_1 \otimes P_2^\perp \otimes P_3 + P_1 \otimes P_2 \otimes P_3^\perp,$$

and

$$Q^\perp = P_1^\perp \otimes P_2^\perp \otimes P_3^\perp + P_1^\perp \otimes P_2^\perp \otimes P_3 + P_1 \otimes P_2^\perp \otimes P_3^\perp + P_1^\perp \otimes P_2 \otimes P_3^\perp,$$

where $P_k^\perp = I - P_k$.

LEMMA 2.1. *For any $A \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ and projection matrices P_k in $\mathbb{R}^{d_k}$, $k = 1, 2, 3$, we have*

$$\|A\|_* \ge \|(P_1 \otimes P_2 \otimes P_3)A\|_* + \frac{1}{2}\|Q^\perp A\|_*.$$

Lemma 2.1 is a direct consequence from the characterization of sub-differential for tensor nuclear norm given by [33], and can be viewed as a tensor version of the pinching inequality for matrices.

Write

$$\mathcal{A}(P_1, P_2, P_3) = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : QA = A\} \tag{2.12}$$

and

$$\mathcal{B}(P_1, P_2, P_3) = \{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : (P_1 \otimes P_2 \otimes P_3)A = A\}. \tag{2.13}$$

By Lemma 2.1, $\mathcal{R}(\cdot)$ defined by (2.11) is weakly decomposable with respect to $\mathcal{A}(P_1, P_2, P_3)$ and $\mathcal{B}(P_1, P_2, P_3)$ with $c_{\mathcal{R}} = 1/2$. We note that a counterexample is also given by [33] which shows that, for the tensor nuclear norm, we cannot take $c_{\mathcal{R}} = 1$.

Another popular way to define tensor rank is through the so-called Tucker decomposition. Recall that the Tucker decomposition of a tensor $A \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is of the form

$$A_{j_1 j_2 j_3} = \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \sum_{k_3=1}^{r_3} S_{k_1 k_2 k_3} U_{j_1 k_1} V_{j_2 k_2} W_{j_3 k_3} \tag{2.14}$$

so that U, V and W are orthogonal matrices, and the so-called core tensor $S = (S_{k_1 k_2 k_3})_{k_1, k_2, k_3}$ is such that any two slices of S are orthogonal. The triplet (r_1, r_2, r_3) are referred to as the Tucker ranks of A. It is not hard to see that if (2.10) holds, then the Tucker ranks (r_1, r_2, r_3) can be equivalently interpreted as the dimensionality of the linear spaces spanned by $\{u_k : 1 \le k \le r\}$, $\{v_k : 1 \le k \le r\}$, and $\{w_k : 1 \le k \le r\}$, respectively. The following relationship holds between CP rank and Tucker ranks:

$$\max\{r_1, r_2, r_3\} \le r \le \min\{r_1 r_2, r_2 r_3, r_1 r_3\}.$$

A convenient way to encourage low Tucker ranks in a tensor is through matricization. Let $\mathcal{M}_1(\cdot)$ denote the mode-1 matricization of a tensor. That is $\mathcal{M}_1(A)$ is a $d_1 \times (d_2 d_3)$ matrix whose column vectors are the the mode-1 fibers of $A \in \mathbb{R}^{d_1 \times d_2 \times d_3}$. $\mathcal{M}_2(\cdot)$ and $\mathcal{M}_3(\cdot)$ can also be defined in the same fashion. It is clear

$$\operatorname{rank}(\mathcal{M}_k(A)) = r_k(A).$$

A natural way to encourage low-rankness is therefore through nuclear norm regularization:

$$\mathcal{R}(A) = \frac{1}{3} \sum_{k=1}^{3} \|\mathcal{M}_k(A)\|_*. \tag{2.15}$$

By the pinching inequality for matrices, $\mathcal{R}(\cdot)$ defined by (2.15) is also decomposable with respect to $\mathcal{A}(P_1, P_2, P_3)$ and $\mathcal{B}(P_1, P_2, P_3)$ with $c_{\mathcal{R}} = 1$.

3. Risk bounds for decomposable regularizers.

We now establish risk bounds for general decomposable regularizers. In particular, our bounds are given in terms of the *Gaussian width* of a suitable set of tensors. Recall that the Gaussian width of a set $S \subset \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ is given by

$$w_G(S) := \mathbb{E}\Big(\sup_{A \in S} \langle A, G \rangle\Big),$$

where $G \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ is a tensor whose entries are independent $\mathcal{N}(0, 1)$ random variables; see, for example, [8] for more details on Gaussian width.

Note that the Gaussian width is a geometric measure of the volume of the set S and can be related to other volumetric characterizations (see, e.g., [20]). We also define the unit ball for the norm-regularizer $\mathcal{R}(\cdot)$ as follows:

$$\mathbb{B}_{\mathcal{R}}(1) := \{A \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N} \mid \mathcal{R}(A) \leq 1\}.$$

We impose the mild assumption that $\|A\|_{\mathrm{F}} \leq \mathcal{R}(A)$ which ensures that the regularizer $\mathcal{R}(\cdot)$ encourages low-dimensional structure.

Now we define a quantity that relates the size of the norm $\mathcal{R}(A)$ to the Frobenius norm $\|A\|_{\mathrm{F}}$ over the the low-dimensional subspace $\mathcal{A}$. Following [18], for a subspace $\mathcal{A}$ of $\mathbb{R}^{d_1 \times \cdots \times d_N}$, define its compatibility constant $s(\mathcal{A})$ as

$$s(\mathcal{A}) := \sup_{A \in \mathcal{A}/\{0\}} \frac{\mathcal{R}^2(A)}{\|A\|_{\mathrm{F}}^2},$$

which can be interpreted as a notion of intrinsic dimensionality of $\mathcal{A}$.

Now we turn our attention to the covariate tensor. Denote by $X^{(i)} = \mathrm{vec}(X^{(i)})$ the vectorized covariate from the ith sample. With slight abuse of notation, write

$$X = \mathrm{vec}\big((X^{(1)})^\top, \ldots, (X^{(n)})^\top\big) \in \mathbb{R}^{n \cdot d_1 d_2 \cdots d_M}$$

the concatenated covariates from all n samples. For convenience, let $D_M = d_1 d_2 \cdots d_M$. Further for brevity, we assume a Gaussian design so that

$$X \sim \mathcal{N}(0, \Sigma),$$

where

$$\Sigma = \mathrm{cov}(X, X) \in \mathbb{R}^{nD_M \times nD_M}.$$

With more technical work, our results may be extended beyond Gaussian designs. We note that we do not require that the sample tensors $X^{(i)}$ be independent.

We shall assume that Σ has bounded eigenvalues which we later verify for a number of statistical examples. Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ represent the smallest and largest eigenvalues of a matrix, respectively. In what follows, we shall assume that

$$c_\ell^2 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq c_u^2, \tag{3.1}$$

for some constants $0 < c_\ell \leq c_u < \infty$.

Note that in particular if all covariates $\{X^{(i)} : i = 1, \ldots, n\}$ are independent and identically distributed, then Σ has a block diagonal structure, and (3.1) boils down to similar conditions on $\mathrm{cov}(X^{(i)}, X^{(i)})$. However, (3.1) is more general and applicable to settings in which the $X^{(i)}$'s may be dependent such as time-series models, which we shall discuss in further detail in Section 4.

We are now in position to state our main result on the risk bounds in terms of both Frobenius norm $\|\cdot\|_{\mathrm{F}}$ and the empirical norm $\|\cdot\|_n$ where for a tensor $A \in \mathbb{R}^{d_1 \times \cdots \times d_N}$, which we define as

$$\|A\|_n^2 := \frac{1}{n} \sum_{i=1}^n \|\langle A, X^{(i)} \rangle\|_F^2.$$

The main reason we focus on random Gaussian design is so that we can prove a one-sided uniform law that relates the empirical norm defined above to the Frobenius norm of a tensor in $\mathcal{A}$ (see Lemma 1.2 in the Supplementary Material [23]).

THEOREM 3.1. *Suppose that* (1.1) *holds for a tensor T from a linear subspace $\mathcal{A}_0 \subset \mathbb{R}^{d_1 \times \cdots \times d_N}$ where* (3.1) *holds. Let $\widehat{T}$ be defined by* (1.2) *where the regularizer $\mathcal{R}(\cdot)$ is decomposable with respect to $\mathcal{A}$ and $\mathcal{A}_0$ for some linear subspace $\mathcal{A} \supseteq \mathcal{A}_0$. If*

$$\lambda \geq \frac{2\sigma c_u (3 + c_{\mathcal{R}})}{c_{\mathcal{R}} \sqrt{n}} w_G[\mathbb{B}_{\mathcal{R}}(1)], \tag{3.2}$$

then there exists a constant $c > 0$ such that with probability at least $1 - \exp\{-c w_G^2[\mathbb{B}_{\mathcal{R}}(1)]\}$,

$$\max\{\|\widehat{T} - T\|_n^2, \|\widehat{T} - T\|_{\mathrm{F}}^2\} \leq \frac{6(1 + c_{\mathcal{R}})}{3 + c_{\mathcal{R}}} \frac{9c_u^2}{c_\ell^2} s(\mathcal{A})\lambda^2, \tag{3.3}$$

when n is sufficiently large, assuming that the right-hand side converges to zero as n increases.

As stated in Theorem 3.1, our upper bound boils down to bounding two quantities, $s(\mathcal{A})$ and $w_G[\mathbb{B}_{\mathcal{R}}(1)]$ which are both purely geometric quantities. To provide some intuition, $w_G[\mathbb{B}_{\mathcal{R}}(1)]$ captures how large the $\mathcal{R}(\cdot)$ norm is relative to the $\|\cdot\|_{\mathrm{F}}$ norm and $s(\mathcal{A})$ captures the low dimension of the subspace $\mathcal{A}$.

Several technical remarks are in order. Note that $w_G[\mathbb{B}_{\mathcal{R}}(1)]$ can be expressed as expectation of the *dual norm* of G. According to $\mathcal{R}$ (see, e.g., [26], for details), the dual norm $\mathcal{R}^*(\cdot)$ is given by

$$\mathcal{R}^*(B) := \sup_{A \in \mathbb{B}_{\mathcal{R}}(1)} \langle A, B \rangle,$$

where the supremum is taken over tensors of the same dimensions as B. It is straightforward to see that $w_G[\mathbb{B}_{\mathcal{R}}(1)] = \mathbb{E}[\mathcal{R}^*(G)]$.

To the best of our knowledge, this is the first general result that applies to multiple responses. As mentioned earlier, incorporating multiple responses presents a technical challenge (see Lemma 1.2 in the Supplementary Material [23]) which is a one-sided uniform law analogous to restricted strong convexity. While Theorem 3.1 focuses on Gaussian design, results can be extended to random sub-Gaussian design using more sophisticated techniques (see, e.g., [14, 34]) or for fixed design by assuming covariates deterministically satisfy the conditions in Lemma 1.2 in the Supplementary Material [23]. Since the focus of this paper is on general dependence structure, we assume random Gaussian design.

One important practical challenge is that σ^2, c_u and c_ℓ are typically unknown and these clearly influence the choice of λ. This is a common challenge for high-dimensional statistical inference and we do not address this issue in this paper.

In practice, λ is typically chosen through cross-validation. A more sophisticated choice of λ based on estimation of σ^2 and other constants remains an open question. Another important and open question is for what choices of $\mathcal{A}_0$ is the upper bound optimal (up to a constant). In Section 4, we provide specific examples in which we provide minimax lower bounds which match the upper bounds up to constant. However, as we see for low-rank tensor regression for low-rank tensor regression discussed in Section 3.2, we are not aware of a convex regularizer that matches the minimax lower bound.

Now we develop upper bounds on both quantities in different scenarios. As in the previous section, we shall focus on third-order tensor in the rest of the section for the ease of exposition.

3.1. *Sparsity regularizers.*

We first consider sparsity regularizers described in the previous section.

3.1.1. *Entry-wise and fiber-wise sparsity.*

Recall that vectorized ℓ_1 regularizer:

$$\mathcal{R}_1(A) = \sum_{j_1=1}^{d_1} \sum_{j_2=1}^{d_2} \sum_{j_3=1}^{d_3} |A_{j_1 j_2 j_3}|,$$

could be used to exploit entry-wise sparsity. Clearly,

$$\mathcal{R}_1^*(A) = \max_{j_1, j_2, j_3} |A_{j_1 j_2 j_3}|.$$

We can now show the following.

LEMMA 3.1. *There exists a constant* $0 < c < \infty$ *such that*

$$w_G[\mathbb{B}_{\mathcal{R}_1}(1)] \le c\sqrt{\log(d_1 d_2 d_3)}. \tag{3.4}$$

Let

$$\Theta_1(s) = \left\{ A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : \sum_{j_1=1}^{d_1} \sum_{j_2=1}^{d_2} \sum_{j_3=1}^{d_3} \mathbb{I}(A_{j_1 j_2 j_3} \neq 0) \le s \right\}.$$

For an arbitrary $A \in \Theta_1(s)$, write

$$I(A) = \{(j_1, j_2, j_3) \in [d_1] \times [d_2] \times [d_3] : A_{j_1 j_2 j_3} \neq 0\}.$$

Then $\mathcal{R}_1(\cdot)$ is decomposable with respect to $\mathcal{A}(I(A))$ as defined by (2.3). It is easy to verify that for any $A \in \Theta_1(s)$,

$$s_1(\mathcal{A}(I)) = \sup_{B \in \mathcal{A}(I(A))/\{0\}} \frac{\mathcal{R}_1^2(B)}{\|B\|_{\mathrm{F}}^2} \le s. \tag{3.5}$$

In light of (3.5) and (3.4), Theorem 3.1 implies that

$$\sup_{T\in\Theta_1(s)} \max\{\|\widehat{T}_1 - T\|_n^2, \|\widehat{T}_1 - T\|_{\mathrm{F}}^2\} \lesssim \frac{s\log(d_1d_2d_3)}{n},$$

with high probability by taking

$$\lambda \asymp \sqrt{\frac{\log(d_1d_2d_3)}{n}},$$

where $\widehat{T}_1$ is the regularized least squares estimate defined by (1.2) when using regularizer $\mathcal{R}_1(\cdot)$.

A similar argument can also be applied to fiber-wise sparsity. To fix ideas, we consider here only sparsity among mode-1 fibers. In this case, we use a group Lasso type of regularizer:

$$\mathcal{R}_2(A) = \sum_{j_2=1}^{d_2}\sum_{j_3=1}^{d_3} \|A_{\cdot j_2 j_3}\|_{\ell_2}.$$

Then

$$\mathcal{R}_2^*(A) = \max_{j_2, j_3} \|A_{\cdot j_2 j_3}\|_{\ell_2}.$$

LEMMA 3.2. *There exists a constant* $0 < c < \infty$ *such that*

$$w_G[\mathbb{B}_{\mathcal{R}_2}(1)] \le c\sqrt{\max\{d_1, \log(d_2d_3)\}}. \tag{3.6}$$

Let

$$\Theta_2(s) = \left\{A \in \mathbb{R}^{d_1\times d_2\times d_3} : \sum_{j_2=1}^{d_2}\sum_{j_3=1}^{d_3} \mathbb{I}(A_{\cdot j_2 j_3} \neq \mathbf{0}) \le s\right\}.$$

Similar to the previous case, for an arbitrary $A \in \Theta_1(s)$, write

$$I(A) = \{(j_2, j_3) \in [d_2]\times[d_3] : A_{\cdot j_2 j_3} \neq \mathbf{0}\}.$$

Then $\mathcal{R}_2(\cdot)$ is decomposable with respect to $\mathcal{A}(I(A))$ as defined by (2.5). It is easy to verify that for any $A \in \Theta_2(s)$,

$$s_2(\mathcal{A}(I)) = \sup_{B\in\mathcal{A}(I(A))/\{0\}} \frac{\mathcal{R}_2^2(B)}{\|B\|_{\mathrm{F}}^2} \le s. \tag{3.7}$$

In light of (3.7) and (3.6), Theorem 3.1 implies that

$$\sup_{T\in\Theta_2(s)} \max\{\|\widehat{T}_2 - T\|_n^2, \|\widehat{T}_2 - T\|_{\mathrm{F}}^2\} \lesssim \frac{s\max\{d_1, \log(d_2d_3)\}}{n},$$

with high probability by taking

$$\lambda \asymp \sqrt{\frac{\max\{d_1, \log(d_2 d_3)\}}{n}},$$

where $\widehat{T}_2$ is the regularized least squares estimate defined by (1.2) when using regularizer $\mathcal{R}_2(\cdot)$.

Comparing with the rates for entry-wise and fiber-wise sparsity regularization, we can see the benefit of using group Lasso type of regularizer $\mathcal{R}_2$ when sparsity is likely to occur at the fiber level. More specifically, consider the case when there are a total of s_1 nonzero entries from s_2 nonzero fibers. If an entry-wise ℓ_1 regularization is applied, we can achieve the risk bound

$$\|\widehat{T}_1 - T\|_{\mathrm{F}}^2 \lesssim \frac{s_1 \log(d_1 d_2 d_3)}{n}.$$

On the other hand, if fiber-wise group ℓ_1 regularization is applied, then the risk bound becomes

$$\|\widehat{T}_2 - T\|_{\mathrm{F}}^2 \lesssim \frac{s_2 \max\{d_1, \log(d_2 d_3)\}}{n}.$$

When nonzero entries are clustered in fibers, we may expect $s_1 \asymp s_2 d_1$. In this case, $\widehat{T}_2$ enjoys performance superior to that of $\widehat{T}_1$ since $s_2 d_1 \log(d_1 d_2 d_3)$ is larger than $s_2 \max\{d_1, \log(d_2 d_3)\}$.

3.1.2. *Slice-wise sparsity and low-rank structure.* Now we consider slice-wise sparsity and low-rank structure. Again, to fix ideas, we consider here only sparsity among $(1, 2)$ slices. As discussed in the previous section, two specific types of regularizers could be employed:

$$\mathcal{R}_3(A) = \sum_{j_3=1}^{d_3} \|A_{\cdot\cdot j_3}\|_{\mathrm{F}}$$

and

$$\mathcal{R}_4(A) = \sum_{j_3=1}^{d_3} \|A_{\cdot\cdot j_3}\|_*,$$

where recall that $\|\cdot\|_*$ denotes the nuclear norm of a matrix, that is, the sum of all singular values.

Note that

$$\mathcal{R}_3^*(A) = \max_{1 \le j_3 \le d_3} \|A_{\cdot\cdot j_3}\|_{\mathrm{F}}.$$

Then we have the following result.

LEMMA 3.3. *There exists a constant* $0 < c < \infty$ *such that*

$$w_G[\mathbb{B}_{\mathcal{R}_3}(1)] \le c\sqrt{\max\{d_1d_2, \log(d_3)\}}. \tag{3.8}$$

Let

$$\Theta_3(s) = \left\{A \in \mathbb{R}^{d_1\times d_2\times d_3} : \sum_{j_3=1}^{d_3} \mathbb{I}(A_{\cdot\cdot j_3} \ne \mathbf{0}) \le s\right\}.$$

For an arbitrary $A \in \Theta_1(s)$, write

$$I(A) = \{j_3 \in [d_3] : A_{\cdot\cdot j_3} \ne \mathbf{0}\}.$$

Then $\mathcal{R}_3(\cdot)$ is decomposable with respect to $\mathcal{A}(I(A))$ as defined by (2.7). It is easy to verify that for any $A \in \Theta_3(s)$,

$$s_3(\mathcal{A}(I(A))) = \sup_{B\in\mathcal{A}(I(A))/\{0\}} \frac{\mathcal{R}_3^2(B)}{\|B\|_{\mathrm{F}}^2} \le s. \tag{3.9}$$

Based on (3.9) and (3.8), Theorem 3.1 implies that

$$\sup_{T\in\Theta_3(s)} \max\{\|\widehat{T}_3 - T\|_n^2, \|\widehat{T}_3 - T\|_{\mathrm{F}}^2\} \lesssim \frac{s\max\{d_1d_2, \log(d_3)\}}{n},$$

with high probability by taking

$$\lambda \asymp \sqrt{\frac{\max\{d_1d_2, \log(d_3)\}}{n}},$$

where $\widehat{T}_3$ is the regularized least squares estimate defined by (1.2) when using regularizer $\mathcal{R}_3(\cdot)$.

Alternatively, for $\mathcal{R}_4(\cdot)$,

$$\mathcal{R}_4^*(A) = \max_{j_3} \|A_{\cdot\cdot j_3}\|_s,$$

we have the following.

LEMMA 3.4. *There exists a constant* $0 < c < \infty$ *such that*

$$w_G[\mathbb{B}_{\mathcal{R}_4}(1)] \le c\sqrt{\max\{d_1, d_2, \log(d_3)\}}. \tag{3.10}$$

Now consider

$$\Theta_4(r) = \left\{A \in \mathbb{R}^{d_1\times d_2\times d_3} : \sum_{j_3=1}^{d_3} \operatorname{rank}(A_{\cdot\cdot j_3}) \le r\right\}.$$

For an arbitrary $A \in \Theta_4(r)$, denote by P_{1j} and P_{2j} the projection onto the row and column space of $A_{\cdot\cdot j}$, respectively. It is clear that $A \in \mathcal{B}(P_{1j}, P_{2j} : 1 \le j \le d_3)$

as defined by (2.9). In addition, recall that $\mathcal{R}_4$ is decomposable with respect to $\mathcal{B}(P_{1j}, P_{2j} : 1 \le j \le d_3)$ and $\mathcal{A}(P_{1j}, P_{2j} : 1 \le j \le d_3)$ as defined by (2.8). It is not hard to see that for any $A \in \Theta_4(r)$, $\mathcal{A}(P_{1j}, P_{2j} : 1 \le j \le d_3) \subset \Theta_4(2r)$, from which we can derive the following.

LEMMA 3.5. *For any* $A \in \Theta_4(r)$,

$$s_4\big(\mathcal{A}(P_{1j}, P_{2j} : 1 \le j \le d_3)\big) \le \sup_{B \in \mathcal{A}/\{0\}} \frac{\mathcal{R}_4^2(B)}{\|B\|_{\mathrm{F}}^2} \le 2r. \tag{3.11}$$

In light of (3.11) and (3.10), Theorem 3.1 implies that

$$\sup_{T \in \Theta_4(r)} \max\{\|\widehat{T}_4 - T\|_n^2, \|\widehat{T}_4 - T\|_{\mathrm{F}}^2\} \lesssim \frac{r \max\{d_1, d_2, \log(d_3)\}}{n},$$

with high probability by taking

$$\lambda \asymp \sqrt{\frac{\max\{d_1, d_2, \log(d_3)\}}{n}},$$

where $\widehat{T}_4$ is the regularized least squares estimate defined by (1.2) when using regularizer $\mathcal{R}_4(\cdot)$.

Comparing with the rates for estimates with regularizers $\mathcal{R}_3$ and $\mathcal{R}_4$, we can see the benefit of using $\mathcal{R}_4$ when the nonzero slices are likely to be of low-rank. In particular, consider the case when there are s_1 nonzero slices and each nonzero slice has rank up to r. Then applying $\mathcal{R}_3$ leads to risk bound

$$\|\widehat{T}_3 - T\|_{\mathrm{F}}^2 \lesssim \frac{s_1 \max\{d_1 d_2, \log(d_3)\}}{n},$$

whereas applying $\mathcal{R}_4$ leads to

$$\|\widehat{T}_4 - T\|_{\mathrm{F}}^2 \lesssim \frac{s_1 r \max\{d_1, d_2, \log(d_3)\}}{n}.$$

It is clear that $\widehat{T}_4$ is a better estimator when $r \ll d_1 = d_2 = d_3$.

3.2. *Low-rankness regularizers.* We now consider regularizers that encourages low-rank estimates. We begin with the tensor nuclear norm regularization:

$$\mathcal{R}_5(A) = \|A\|_*.$$

Recall that $\mathcal{R}_5^*(A) = \|A\|_s$.

LEMMA 3.6. *There exists a constant* $0 < c < \infty$ *such that*

$$w_G\big[\mathbb{B}_{\mathcal{R}_5}(1)\big] \le c\sqrt{(d_1 + d_2 + d_3)}. \tag{3.12}$$

Now let

$$\Theta_5(r)=\{A\in\mathbb{R}^{d_1\times d_2\times d_3}:\max\{r_1(A),r_2(A),r_3(A)\}\le r\}.$$

For an arbitrary $A\in\Theta_5(r)$, denote by P_1, P_2, P_3 the projection onto the linear space spanned by the mode-1, -2 and -3 fibers, respectively. As we argued in the previous section, $\mathcal{R}_5(\cdot)$ is weakly decomposable with respect to $\mathcal{A}(P_1,P_2,P_3)$ and $\mathcal{B}(P_1,P_2,P_3)$, and $A\in\mathcal{B}(P_1,P_2,P_3)$ where $\mathcal{A}(P_1,P_2,P_3)$ and $\mathcal{B}(P_1,P_2,P_3)$ are defined by (2.12) and (2.13), respectively.

LEMMA 3.7. *For any* $A\in\Theta_5(r)$,

$$s_5(\mathcal{A}(P_1,P_2,P_3))=\sup_{B\in\mathcal{A}(P_1,P_2,P_3)/\{0\}}\frac{\mathcal{R}_5^2(B)}{\|B\|_{\mathrm{F}}^2}\le r^2.$$

Lemmas 3.6 and 3.7 show that

$$\sup_{T\in\Theta_5(r)}\max\{\|\widehat{T}_5-T\|_n^2,\|\widehat{T}_5-T\|_{\mathrm{F}}^2\}\lesssim\frac{r^2(d_1+d_2+d_3)}{n},$$

with high probability by taking

$$\lambda\asymp\sqrt{\frac{d_1+d_2+d_3}{n}},$$

where $\widehat{T}_5$ is the regularized least squares estimate defined by (1.2) when using regularizer $\mathcal{R}_5(\cdot)$.

Next, we consider the low-rankness regularization via matricization:

$$\mathcal{R}_6(A)=\frac{1}{3}(\|\mathcal{M}_1(A)\|_*+\|\mathcal{M}_2(A)\|_*+\|\mathcal{M}_3(A)\|_*).$$

It is not hard to see that

$$\mathcal{R}_6^*(A)=3\max\{\|\mathcal{M}_1(A)\|_s,\|\mathcal{M}_2(A)\|_s,\|\mathcal{M}_3(A)\|_s\}.$$

LEMMA 3.8. *There exists a constant* $0<c<\infty$ *such that*

$$w_G[\mathbb{B}_{\mathcal{R}_6}(1)]\le c\sqrt{\max\{d_1d_2,d_2d_3,d_1d_3\}}. \tag{3.13}$$

On the other hand, we have the following.

LEMMA 3.9. *For any* $A\in\Theta_5(r)$,

$$s_6(\mathcal{A}(P_1,P_2,P_3))=\sup_{B\in\mathcal{A}(P_1,P_2,P_3)/\{0\}}\frac{\mathcal{R}_6^2(B)}{\|B\|_{\mathrm{F}}^2}\le r.$$

Lemmas 3.8 and 3.9 suggest that

$$\sup_{T\in\Theta_5(r)} \max\{\|\widehat{T}_6 - T\|_n^2, \|\widehat{T}_6 - T\|_{\mathrm{F}}^2\} \lesssim \frac{r\max\{d_1d_2, d_2d_3, d_1d_3\}}{n},$$

with high probability by taking

$$\lambda \asymp \sqrt{\frac{\max\{d_1d_2, d_2d_3, d_1d_3\}}{n}},$$

where $\widehat{T}_6$ is the regularized least squares estimate defined by (1.2) when using regularizer $\mathcal{R}_6(\cdot)$.

Comparing with the rates for estimates with regularizers $\mathcal{R}_5$ and $\mathcal{R}_6$, we can see the benefit of using $\mathcal{R}_5$. For any $T \in \Theta_5(r)$, If we apply regularizer $\mathcal{R}_5$, then

$$\|\widehat{T}_5 - T\|_{\mathrm{F}}^2 \lesssim \frac{r^2(d_1+d_2+d_3)}{n}.$$

This is to be compared with the risk bound for matricized regularization:

$$\|\widehat{T}_6 - T\|_{\mathrm{F}}^2 \lesssim \frac{r\max\{d_1d_2, d_2d_3, d_1d_3\}}{n}.$$

Obviously, $\widehat{T}_5$ always outperform $\widehat{T}_6$ since $r \le \min\{d_1, d_2, d_3\}$. The advantage of $\widehat{T}_5$ is typically rather significant since in general $r \ll \min\{d_1, d_2, d_3\}$. On the other hand, $\widehat{T}_6$ is more amenable for computation.

Both upper bounds on Frobenius error on $\widehat{T}_5$ and $\widehat{T}_6$ are novel results and complement the existing results on tensor completion [7, 16] and [33]. Neither $\widehat{T}_5$ nor $\widehat{T}_6$ is minimax optimal and remains an interesting question as to whether there exists a convex regularization approach that is minimax optimal.

4. Specific statistical problems.

In this section, we apply our results to several concrete examples where we are attempting to estimate a tensor under certain sparse or low-rank constraints, and show that the regularized least squares estimate $\widehat{T}$ is typically minimiax rate optimal with appropriate choices of regularizers. In particular, we focus on the multiresponse aspect of the general framework to provide novel upper bounds and matching minimax lower bounds.

4.1. *Multiresponse regression with large p.*

The first example we consider is the multiresponse regression model:

$$Y_k^{(i)} = \sum_{j=1}^{p}\sum_{\ell=1}^{m} X_{j\ell}^{(i)} T_{j\ell k} + \varepsilon_k^{(i)},$$

where $1 \le i \le n$ represents the index for each sample, $1 \le k \le m$ represents the index for each response and $1 \le j \le p$ represents the index for each feature. For the multiresponse regression problem, we have $N = 3$, $M = 2$, $d_1 = d_2 = m$ which represents the total number of responses and $d_3 = p$, which represent the total number of parameters.

Since we are in the setting where p is large but only a small number s are relevant, we define the subspace

$$\mathcal{T}_1 = \left\{ A \in \mathbb{R}^{m \times m \times p} \,\middle|\, \sum_{j=1}^{p} \mathbb{I}(\|A_{\cdot\cdot j}\|_{\mathrm{F}} \neq 0) \leq s \right\}.$$

Furthermore, for each i we assume $X^{(i)} \in \mathbb{R}^{m \times p}$ where each entry of $X^{(i)}$, $[X^{(i)}]_{k,j}$, corresponds to the jth feature for the kth response. For simplicity, we assume the $X^{(i)}$'s are independent Gaussian with covariance $\widetilde{\Sigma} \in \mathbb{R}^{mp \times mp}$. The penalty function we are considering is

$$\mathcal{R}(A) = \sum_{j=1}^{p} \|A_{\cdot\cdot j}\|_{\mathrm{F}}, \tag{4.1}$$

and the corresponding dual function applied to the i.i.d. Gaussian tensor G is

$$\mathcal{R}^*(G) = \max_{1 \leq j \leq p} \|G_{\cdot\cdot j}\|_{\mathrm{F}}.$$

THEOREM 4.1. *Under the multiresponse regression model with $T \in \mathcal{T}_1$ and independent Gaussian design where $c_\ell^2 \leq \lambda_{min}(\widetilde{\Sigma}) \leq \lambda_{\max}(\widetilde{\Sigma}) \leq c_u^2$, if*

$$\lambda \geq 3\sigma c_u \sqrt{\frac{\max\{m^2, \log p\}}{n}},$$

such that $\sqrt{s}\lambda$ converges to zero as n increases, then there exist some constants $c_1, c_2 > 0$ such that with probability at least $1 - p^{-c_1}$

$$\max\{\|\widehat{T} - T\|_n^2, \|\widehat{T} - T\|_{\mathrm{F}}^2\} \leq \frac{c_2 c_u^2}{c_\ell^2} s\lambda^2,$$

when n is sufficiently large, where $\widehat{T}$ is the regularized least squares estimate defined by (1.2) with regularizer given by (4.1). In addition,

$$\min_{\widetilde{T}} \max_{T \in \mathcal{T}_1} \|\widetilde{T} - T\|_{\mathrm{F}}^2 \geq \frac{c_3 \sigma^2 s \max\{m^2, \log p/s\}}{c_u^2 n},$$

for some constant $c_3 > 0$, with probability at least $1/2$, where the minimum is taken over all estimators $\widetilde{T}$ based on data $\{(X^{(i)}, Y^{(i)}) : 1 \leq i \leq n\}$.

Theorem 4.1 shows that when taking

$$\lambda \asymp \sqrt{\frac{\max\{m^2, \log p\}}{n}},$$

the regularized least squares estimate defined by (1.2) with regularizer given by (4.1) achieves minimax optimal rate of convergence over the parameter space $\mathcal{T}_1$.

Alternatively, there are settings where the effect of covariates on the multiple tasks may be of low-rank structure. In such a situation, we may consider

$$\mathcal{T}_2 = \left\{ A \in \mathbb{R}^{m \times m \times p} \;\middle|\; \sum_{j=1}^{p} \operatorname{rank}(A_{\cdot\cdot j}) \le r \right\}.$$

An appropriate penalty function in this case is

$$\mathcal{R}(A) = \sum_{j=1}^{p} \|A_{\cdot\cdot j}\|_*, \tag{4.2}$$

and the corresponding dual function applied to G is

$$\mathcal{R}^*(G) = \max_{1 \le j \le p} \|G_{\cdot\cdot j}\|_s.$$

THEOREM 4.2. *Under the multiresponse regression model with $T \in \mathcal{T}_2$ and independent Gaussian design where $c_\ell^2 \le \lambda_{min}(\widetilde{\Sigma}) \le \lambda_{\max}(\widetilde{\Sigma}) \le c_u^2$, if*

$$\lambda \ge 3\sigma c_u \sqrt{\frac{\max\{m, \log p\}}{n}},$$

such that $\sqrt{r}\lambda$ converges to zero as n increases, then there exist some constants $c_1, c_2 > 0$ such that with probability at least $1 - p^{-c_1}$,

$$\max\{\|\widehat{T} - T\|_n^2, \|\widehat{T} - T\|_{\mathrm{F}}^2\} \le \frac{c_2 c_u^2}{c_\ell^2} r\lambda^2$$

when n is sufficiently large, where $\widehat{T}$ is the regularized least squares estimate defined by (1.2) with regularizer given by (4.2). In addition,

$$\min_{\widetilde{T}} \max_{T \in \mathcal{T}_2} \|\widetilde{T} - T\|_{\mathrm{F}}^2 \ge \frac{c_3 \sigma^2 r \max\{m, \log(p/r)\}}{c_u^2 n},$$

for some constant $c_3 > 0$, with probability at least $1/2$, where the minimum is taken over all estimators $\widetilde{T}$ based on data $\{(X^{(i)}, Y^{(i)}) : 1 \le i \le n\}$.

Again Theorem 4.2 shows that by taking

$$\lambda \asymp \sqrt{\frac{\max\{m, \log p\}}{n}},$$

the regularized least squares estimate defined by (1.2) with regularizer given by (4.2) achieves minimax optimal rate of convergence over the parameter space $\mathcal{T}_2$. Comparing with optimal rates for estimating a tensor from $\mathcal{T}_1$, one can see the benefit and importance to take advantage of the extra low rankness if the true coefficient tensor is indeed from $\mathcal{T}_2$. As far as we are aware, these are the first results that provide upper bounds and matching minimax lower bounds for high-dimensional multiresponse regression with sparse or low-rank slices.

4.2. *Multivariate sparse autoregressive models.* Now we consider the setting of vector autoregressive models. In this case, our generative model is

$$X^{(t+p)} = \sum_{j=1}^{p} A_j X^{(t+p-j)} + \varepsilon^{(t)}, \tag{4.3}$$

where $1 \le t \le n$ represents the time index, $1 \le j \le p$ represents the lag index, $\{X^{(t)}\}_{t=0}^{n+p}$ is an m-dimensional vector and $\varepsilon^{(t)} \sim \mathcal{N}(0, \sigma^2 I_{m\times m})$ represents the additive noise. Note that the parameter tensor T is an $m \times m \times p$ tensor so that $T_{\cdot\cdot j} = A_j$, and $T_{k\ell j}$ represents the co-efficient of the kth variable on the ℓ^{th} variable at lag j. This model is studied by [2] where p is relatively small (to avoid introducing long-range dependence) and m is large. Our main results allow more general structure and regularization schemes than those considered in [2].

Since we assume the number of series m is large, and there are m^2 possible interactions between the series we assume there are only $s \ll m^2$ interactions in total:

$$\mathcal{T}_3 = \left\{ A \in \mathbb{R}^{m\times m\times p} \,\middle|\, \sum_{k=1}^{m} \sum_{\ell=1}^{m} \mathbb{I}(A_{k\ell\cdot} \ne \mathbf{0}) \le s \right\}. \tag{4.4}$$

The penalty function we are considering is

$$\mathcal{R}(A) = \sum_{k=1}^{m} \sum_{\ell=1}^{m} \|A_{k\ell\cdot}\|_{\ell_2}, \tag{4.5}$$

and the corresponding dual function applied to G is

$$\mathcal{R}^*(G) = \max_{1\le k,\ell\le m} \|G_{k,\ell,\cdot}\|_{\ell_2}.$$

The challenge in this setting is that the X's are highly dependent and we use the results developed in [2] to prove that (3.1) is satisfied.

Prior to presenting the main results, we introduce concepts developed in [2] that play a role in determining the constants c_u^2 and c_ℓ^2 which relate to the stability of the autoregressive processes. A p-variate Gaussian time series is defined by its autocovariance matrix function

$$\Gamma_X(h) = \operatorname{Cov}\big(X^{(t)}, X^{(t+h)}\big),$$

for all $t, h \in \mathbb{Z}$. Further, we define the spectral density function:

$$f_X(\theta) := \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \Gamma_X(\ell) e^{-i\ell\theta}, \qquad \theta \in [-\pi, \pi].$$

To ensure the spectral density is bounded, we make the following assumption:

$$\mathcal{M}(f_X) := \operatorname*{ess\,sup}_{\theta} \Lambda_{\max}\big(f_X(\theta)\big) < \infty.$$

Further, we define the matrix polynomial

$$\mathcal{A}(z) = I_{m\times m} - \sum_{j=1}^{p} A_j z^j,$$

where $\{A_j\}_{j=1}^p$ denote the back-shift matrices, and z represents any point on the complex plane. Note that for a stable, invertible AR(p) process,

$$f_X(\theta) = \frac{1}{2\pi}\mathcal{A}^{-1}(e^{-i\theta})\overline{\mathcal{A}^{-1}(e^{-i\theta})}.$$

We also define the lower extremum of the spectral density:

$$m(f_X) := \operatorname*{ess\,inf}_{\theta} \Lambda_{\min}\big(f_X(\theta)\big).$$

Note that $m(f_X)$ and $\mathcal{M}(f_X)$ satisfy the following bounds:

$$m(f_X) \geq \frac{1}{2\pi\mu_{\max}(\mathcal{A})} \quad \text{and} \quad \mathcal{M}(f_X) \leq \frac{1}{2\pi\mu_{\min}(\mathcal{A})},$$

where

$$\mu_{\min}(\mathcal{A}) := \min_{|z|=1} \Lambda_{\min}\big(\overline{\mathcal{A}(z)}\mathcal{A}(z)\big)$$

and

$$\mu_{\max}(\mathcal{A}) := \max_{|z|=1} \Lambda_{\max}\big(\overline{\mathcal{A}(z)}\mathcal{A}(z)\big).$$

From a straightforward calculation, we have that for any fixed Δ

$$\frac{1}{\mu_{\max}}\|\Delta\|_{\mathrm{F}}^2 \leq \mathbb{E}\big[\|\Delta\|_n^2\big] \leq \frac{1}{\mu_{\min}}\|\Delta\|_{\ell_2}^2. \tag{4.6}$$

Hence $c_u^2 = 1/\mu_{\min}$ and $c_\ell^2 = 1/\mu_{\max}$. Now we state our main result for autoregressive models.

THEOREM 4.3. *Under the vector autoregressive model defined by* (4.3) *with* $T \in \mathcal{T}_3$, *if*

$$\lambda \geq 3\sigma\sqrt{\frac{\max\{p, 2\log m\}}{n\mu_{\min}}},$$

such that $\sqrt{s}\lambda$ *converges to zero as n increases, then there exist some constants* $c_1, c_2 > 0$ *such that with probability at least* $1 - m^{-c_1}$,

$$\max\big\{\|\widehat{T} - T\|_n^2, \|\widehat{T} - T\|_{\mathrm{F}}^2\big\} \leq \frac{c_2\mu_{\max}}{\mu_{\min}} s\lambda^2,$$

when n is sufficiently large, where $\widehat{T}$ is the regularized least squares estimators defined by (1.2) *with regularizer given by* (4.5). *In addition,*

$$\min_{\widetilde{T}} \max_{T \in \mathcal{T}_3} \|\widetilde{T} - T\|_{\mathrm{F}}^2 \geq c_3 \mu_{\min} \sigma^2 \frac{s \max\{p, \log(m/\sqrt{s})\}}{n},$$

for some constant $c_3 > 0$, with probability at least $1/2$, where the minimum is taken over all estimators $\widetilde{T}$ based on data $\{X^{(t)} : t = 0, \dots, n + p\}$.

Theorem 4.3 provides, to our best knowledge, the only lower bound result for multivariate time series. The upper bound is also novel and is different from Proposition 4.1 in [2] since we impose sparsity only on the large m directions and not over the p lags, whereas [2] impose sparsity through vectorization. Note that Proposition 4.1 in [2] follows directly from Lemma 3.1 with $d_1 = p$ and $d_2 = d_3 = m$. Using the sparsity regularizer, [2] vectorize the problem and prove restricted strong convexity whereas since we leave the problem as a multiresponse problem, we required the more refined technique used for proving Lemma 1.2 in the Supplementary Material [23].

4.3. *Pairwise interaction tensor models.* Finally, we consider the tensor regression (1.1) where T follows a pairwise interaction model. More specifically, $(X^{(i)}, Y^{(i)})$, $i = 1, 2, \dots, n$ are independent copies of a random couple $X \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ and $Y \in \mathbb{R}$ such that

$$Y = \langle X, T \rangle + \varepsilon$$

and

$$T_{j_1 j_2 j_3} = A^{(12)}_{j_1 j_2} + A^{(13)}_{j_1 j_3} + A^{(23)}_{j_2 j_3}.$$

Here, $A^{(k_1,k_2)} \in \mathbb{R}^{d_{k_1} \times d_{k_2}}$ such that

$$A^{(k_1,k_2)} \mathbf{1} = \mathbf{0} \quad \text{and} \quad \big(A^{(k_1,k_2)}\big)^\top \mathbf{1} = \mathbf{0}.$$

The pairwise interaction was used originally by [24, 25] for personalized tag recommendation, and later analyzed in [5]. Hoff [10] briefly introduced a single index additive model (among other tensor models) which is a subclass of the pairwise interaction model. The regularizer we consider is

$$\mathcal{R}(A) = \|A^{(12)}\|_* + \|A^{(13)}\|_* + \|A^{(23)}\|_*. \tag{4.7}$$

It is not hard to see that $\mathcal{R}$ defined above is decomposable with respect to $\mathcal{A}(P_1, P_2, P_3)$ for any projection matrices:

Let

$$\mathcal{T}_4 = \Big\{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : A_{j_1 j_2 j_3} = A^{(12)}_{j_1 j_2} + A^{(13)}_{j_1 j_3} + A^{(23)}_{j_2 j_3}, A^{(k_1,k_2)} \in \mathbb{R}^{d_{k_1} \times d_{k_2}},$$
$$A^{(k_1,k_2)} \mathbf{1} = \mathbf{0}, \text{and} \big(A^{(k_1,k_2)}\big)^\top \mathbf{1} = \mathbf{0},$$
$$\max_{k_1,k_2} \operatorname{rank}\big(A^{(k_1,k_2)}\big) \leq r\Big\}.$$

For simplicity, we assume i.i.d. Gaussian design so $c_\ell^2 = c_u^2 = 1$.

THEOREM 4.4. *Under the pairwise interaction model with $T \in \mathcal{T}_4$, if*

$$\lambda \geq 3\sigma\sqrt{\frac{\max\{d_1, d_2, d_3\}}{n}},$$

such that $\sqrt{r}\lambda$ converges to zero as n increases, then there exist constants $c_1, c_2 > 0$ such that with probability at least $1 - \min\{d_1, d_2, d_3\}^{-c_1}$,

$$\max\{\|\widehat{T} - T\|_n^2, \|\widehat{T} - T\|_{\mathrm{F}}^2\} \leq c_2 r \lambda^2,$$

when n is sufficiently large, where $\widehat{T}$ is the regularized least squares estimate defined by (1.2) with regularizer given by (4.7). In addition,

$$\min_{\widetilde{T}} \max_{T \in \mathcal{T}_4} \|\widetilde{T} - T\|_{\mathrm{F}}^2 \geq \frac{c_3 \sigma^2 r \max\{d_1, d_2, d_3\}}{n},$$

for some constant $c_3 > 0$, with probability at least $1/2$, where the minimum is taken over all estimate $\widetilde{T}$ based on data $\{(X^{(i)}, Y^{(i)}) : 1 \leq i \leq n\}$.

As in the other settings, Theorem 4.4 establishes the minimax optimality of the regularized least squares estimate (1.2) when using an appropriate convex decomposable regularizer. Since this is single response and the norm involves matricization, this result is a straightforward extension to earlier results.

5. Numerical experiments. In this section, we provide a series of numerical experiments that both support our theoretical results and display the flexibility of our general framework. In particular, we consider several different models including: third-order tensor regression with a scalar response (Section 5.1); fourth-order tensor regression (Section 5.2); matrix-response regression with both group sparsity and low-rankness regularizers (Section 5.3); multivariate sparse autoregressive models (Section 5.4) and pairwise interaction models (Section 5.5). To perform the simulations in a computationally tractable way, we adapt the block coordinate descent approaches in multiresponse case developed by [28], and those developed by [21] for univariate response settings, to capture group sparsity and low-rankness regularizers.

To fix ideas, in all numerical experiments, the covariate tensors $X^{(i)}$s were independent standard Gaussian ensembles (except for the multivariate auto-regressive models); and the noise $\varepsilon^{(i)}$s are i.i.d. random tensors with elements following $N(0, \sigma^2)$ independently. As to the choice of tuning parameter, we adopt grid search on λ to find the one with the least estimation error (in terms of mean squared error) in all our numerical examples.

5.1. *Third-order tensor regression.* First, we consider a third-order tensor regression model:

$$Y^{(i)} = \langle B, X^{(i)} \rangle + \varepsilon^{(i)},$$

where $B \in \mathbb{R}^{d\times d\times d}$, $Y^{(i)}$, $\varepsilon^{(i)} \in \mathbb{R}$, $X^{(i)} \in \mathbb{R}^{d\times d\times d}$. The regression coefficient tensor B was generated as follows: the first s slices $B_{\cdot\cdot 1}, \dots, B_{\cdot\cdot s}$ are i.i.d. standard normal ensembles; and the remaining slices $B_{\cdot\cdot s+1}, \dots, B_{\cdot\cdot d_3}$ are set to be zero. Naturally, we consider here the group-sparsity regularizer:

$$\min_{A\in\mathbb{R}^{d\times d\times d}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \|Y^{(i)} - \langle A, X^{(i)} \rangle\|_{\mathrm{F}}^2 + \lambda \sum_{j_3=1}^{d} \|A_{\cdot\cdot j_3}\|_{\mathrm{F}} \right\}.$$

Figure 1 shows the mean squared error of the estimate averaged over 50 runs (with standard deviation) versus d, n and s respectively. In the left and middle panels, we set $s = 2$, whereas in the right panel, we fixed $d = 16$. As we can observe, the mean squared error increases approximately according to d^2, s, and $1/n$ which agrees with the risk bound given in Lemma 3.3.

We also considered a setting where B is slice-wise low rank. More specifically, the s nonzero slices $B_{\cdot\cdot 1}, \dots, B_{\cdot\cdot s}$ were random rank-r matrices. In this case, the slice-wise low-rankness regularizer can be employed:

$$\min_{A\in\mathbb{R}^{d_1\times d_2\times d_3}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \|Y^{(i)} - \langle A, X^{(i)} \rangle\|_{\mathrm{F}}^2 + \lambda \sum_{j_3=1}^{d_3} \|A_{\cdot\cdot j_3}\|_{*} \right\}.$$

The performance of the estimate, averaged over 50 simulation runs, is summarized by Figure 2 where in the left and middle panels $r = 2$, and in the right panel, $d = 16$. Once again, our results are consistent with our theoretical results.

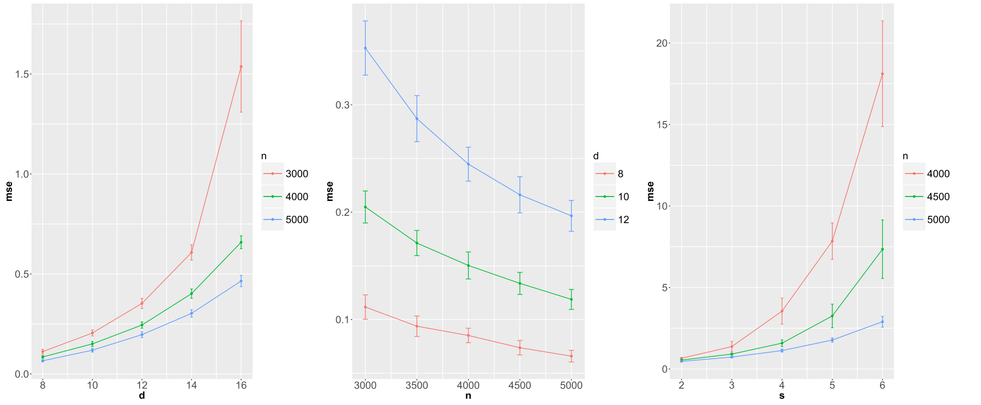

FIG. 1. *Mean squared error of the group-sparsity regularization for third-order tensor regression. The plot was based on* 50 *simulation runs and the error bars in each panel represent* ± *one standard deviation.*

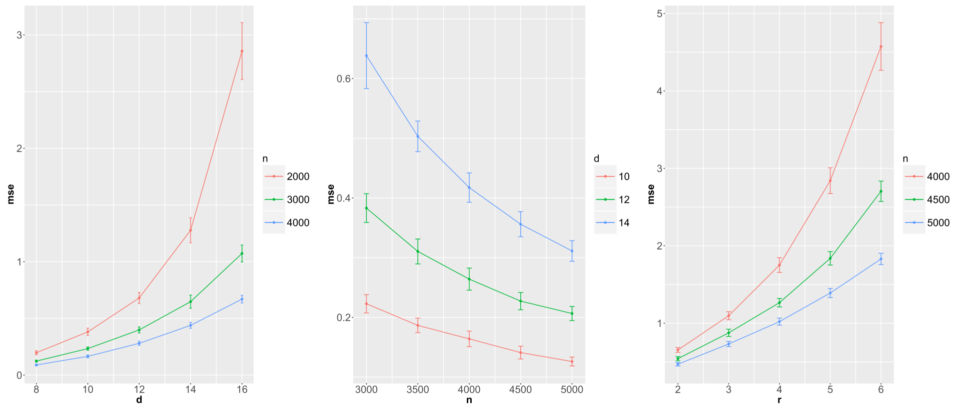

FIG. 2. *Mean squared error for third-order tensor regression with low-rank slices tensor coefficients. The plot was based on* 50 *simulation runs and the error bars in each panel represent* ± *one standard deviation.*

5.2. *Fourth-order tensor regression.* Although we have focused on third- order tensors for brevity, our treatment applies to higher-order tensors as well. As an illustration, we now consider fourth-order models where $B \in \mathbb{R}^{d\times d\times d\times d}$, $Y^{(i)}$, $\varepsilon^{(i)} \in \mathbb{R}$, $X^{(i)} \in \mathbb{R}^{d\times d\times d\times d}$.

To generate low-rank fourth-order tensors, we impose low CP rank as follows: generate four independent groups of r independent random vectors of unit length, $\{u_{k,1}\}_{k=1}^r$, $\{u_{k,2}\}_{k=1}^r$, $\{u_{k,3}\}_{k=1}^r$ and $\{u_{k,4}\}_{k=1}^r$ via performing an SVD of Gaussian random matrix two times and keeping the r pairs of leading singular vectors, and then compute the outer-product yielding a rank-r tensor $B = \sum_{k=1}^r u_{k,1} \otimes u_{k,2} \otimes u_{k,3} \otimes u_{k,4}$.

We consider two different regularization schemes. First, we impose low-rank structure through mode-1 matricization:

$$\min_{A\in\mathbb{R}^{d\times d\times d\times d}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \|Y^{(i)} - \langle A, X^{(i)}\rangle\|_{\mathrm{F}}^2 + \lambda \|\mathcal{M}_1(A)\|_* \right\}.$$

Second, we use the square matricization as follows:

$$\min_{A\in\mathbb{R}^{d\times d\times d\times d}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \|Y^{(i)} - \langle A, X^{(i)}\rangle\|_{\mathrm{F}}^2 + \lambda \|\mathcal{M}_{12}(A)\|_* \right\},$$

where $\mathcal{M}_{12}(\cdot)$ reshape a fourth-order tensor into a $d^2 \times d^2$ matrix by collapsing its first two indices, and last two indices, respectively. Table 1 shows the average root-mean-square error (RMSE, for short) for both approaches. As we can see, the 2-by-2 approach appears superior to the 1-by-3 approach which is also predicted by the theory.

TABLE 1

Tensor regression with fourth-order tensor covariates and scale response based on matricization: RMSE were computed based on 50 *simulations runs. Numbers in parentheses are standard errors*

n	d	r	σ	SNR	RMSE (Mode-1)	RMSE (Square deal)
2000	7	5	10	3.0 (0.1)	0.53 (0.01)	0.51 (0.01)
2000	7	3	10	1.5 (0.1)	0.58 (0.02)	0.49 (0.02)
4000	10	3	10	1.7 (0.1)	0.67 (0.01)	0.51 (0.02)

5.3. *Matrix-response regression.* Our general framework can handle matrix-responses in a seamless fashion. For demonstration, we consider here matrix-response regression with both group sparsity and low-rankness regularizer. More specifically, the following model was considered:

$$Y^{(i)} = \langle B, X^{(i)} \rangle + \varepsilon^{(i)},$$

where $B \in \mathbb{R}^{d\times d\times d}$, $Y^{(i)}, \varepsilon^{(i)} \in \mathbb{R}^{d\times d}$, $X^{(i)} \in \mathbb{R}^{d}$. As before, to impose group sparsity, the first s slices of B were generated as Gaussian ensembles and the remaining slices were set to zero.

For both the group sparsity and low-rankness regularizers, we used the matrix-version algorithm for group-penalized multiresponse regression in [28]. For each block of the coordinate descent, the subproblem with both ℓ_1 and nuclear norm penalty have closed-form solutions.

Figure 3 shows the average (with standard deviation) mean squared error over 50 runs versus the d, n and s parameter. (Here, $d_1 = d_2 = d_3 = d$.) As we observe, the mean-squared error increase approximately according to $\log d$, s, and $1/n$ which supports our upper bound in Theorem 4.1.

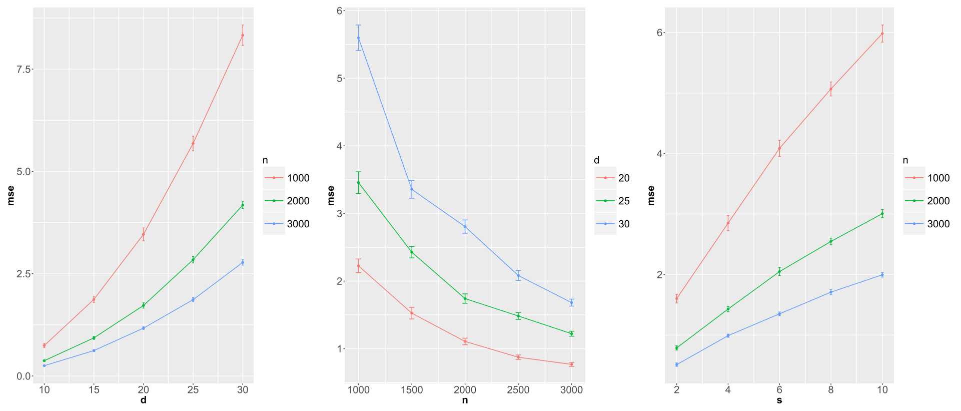

FIG. 3. *Matrix response regression with sparse slices tensor coefficients. The plot was based on* 50 *simulation runs and the error bars in each panel represent ± one standard deviation.*

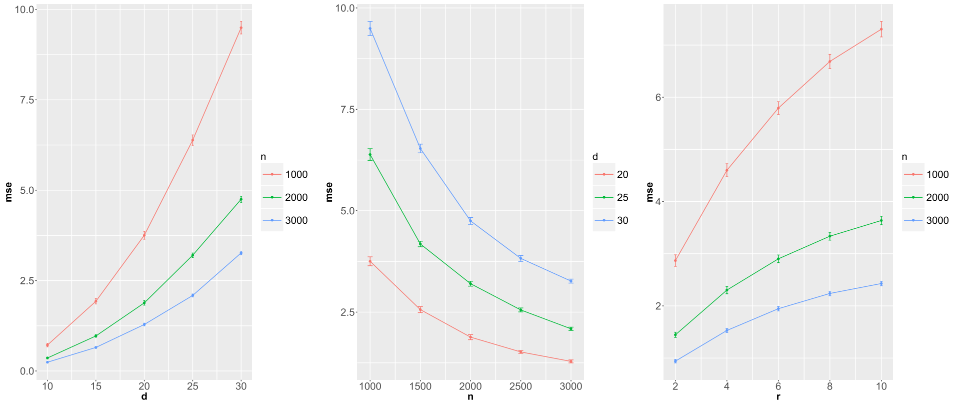

FIG. 4. *Matrix response regression with low-rank slices tensor coefficients. The plot was based on* 50 *simulation runs and the error bars in each panel represent* ± *one standard deviation.*

We also generated low-rank B in the same fashion as before. Figure 4 plots the average (with standard deviation) mean squared error against d, n and r, respectively. These results are consistent with the main result in Theorem 4.2.

5.4. *Multivariate sparse autoregressive models.* Now we consider dependent covariates and responses through the multivariate autoregressive model. Recall that the generative model is

$$X^{(t+p)} = \sum_{j=1}^{p} B_{\cdot\cdot j} X^{(t+p-j)} + \varepsilon^{(t)},$$

where $1 \le t \le n$ represents the time index, $1 \le j \le p$ represents the lag index, $\{X^{(t)}\}_{t=0}^{(n+p)}$ is an m-dimensional vector, $\varepsilon^{(t)} \sim \sigma\mathcal{N}(0, I_{m\times m})$ represents the additive noise.

We consider four different low-dimensional structures for B and we choose the entries of B to be sufficiently small to ensure the time series is stable.

- Slice-wise sparsity: $B_{\cdot\cdot 1}, \dots, B_{\cdot\cdot s}$ are s nonzero slices of diagonal matrix, where diagonal elements are constants ρ with $\rho = 2$. $B_{\cdot\cdot s+1}, \dots, B_{\cdot\cdot d_3}$ are zero slices.
- Sparse low-rank slices: $B_{\cdot\cdot 1}, \dots, B_{\cdot\cdot s}$ are s nonzero slices, which are independent random rank-r matrix [truncated matrix with i.i.d. elements from $N(0, \tau^2)$]. $B_{\cdot\cdot s+1}, \dots, B_{\cdot\cdot d_3}$ are zero slices. Here, $\tau = 1/150$ for $m = 10$ and $\tau = 1/500$ for $m = 20$.
- Group sparsity by lag (sparse normal slices):$B_{\cdot\cdot 1}, \dots, B_{\cdot\cdot s}$ are s nonzero slices, where elements follow i.i.d. $N(0, \tau^2)$ with $\tau = 0.05$. $B_{\cdot\cdot s+1}, \dots, B_{\cdot\cdot d_3}$ are zero slices.

TABLE 2

Multivariate autoregressive model with various sparsity/low-rankness: RMSE were computed based on 50 *simulations runs. Numbers in parentheses are standard errors*

Regularizer	**Coefficient tensor**	n	m	p	s/r	σ	**SNR (sd)**	**RMSE (sd)**
Vectorized	s diagonal slices	2000	10	10	5	2	5.5 (1.4)	0.43 (0.15)
sparsity	s diagonal slices	2000	20	20	5	2	2.9 (0.7)	0.25 (0.04)
Low-rank	s rank-r slices	2000	10	10	10, 5	0.05	1.0 (0.1)	0.37 (0.01)
slices	s rank-r slices	2000	20	20	10, 5	0.05	1.2 (0.1)	0.82 (0.02)
Group sparsity	s Gaussian slices	2000	10	10	5	0.2	0.42 (0.02)	0.40 (0.01)
(by lag)	s Gaussian slices	2000	20	20	5	0.2	0.42 (0.01)	0.64 (0.02)
Group sparsity	s Gaussian fibers	2000	10	10	10	0.02	0.93 (0.2)	0.35 (0.04)
(by coordinate)	s Gaussian fibers	2000	20	20	10	0.02	1.2 (0.1)	0.72 (0.07)

- Group sparsity by coordinate (sparse normal fibers): $B_{s_1 s_2 \cdot}$ is a vector of i.i.d. normal elements following $N(0, \tau^2)$ ($\tau = 0.1$) when $(s_1, s_2) \in \mathcal{S}$, which is a random sample of size s from $\{1, \dots, m\} \times \{1, \dots, m\}$, and zero otherwise.

Table 2 shows the average RMSE for 50 runs of each case as a function of m, p, s and r. In general, the smaller the n is, or the larger the m (or p) is, the harder it is to recover the coefficient B. These findings are consistent with our theoretical developments.

5.5. *Pairwise interaction tensor models.* Finally, we consider the so-called pairwise interaction tensor models as described in Section 4.3. To implement this regularization scheme, we kept iterating among the matrix slices $A_{1,2}$, $A_{1,3}$ and $A_{2,3}$ and updating one of the three at a time while assuming the other two components are fixed. For the update of A_{k_1,k_2}, we conducted an approximated projection onto the zero-row-sum/zero-column-sum subspace after each generalized gradient descent (soft thresholding) step

$$A^{(i+1)}_{k_1,k_2} = \hat{P}\big(\hat{P}_{\lambda\eta}\big(A^{(i)}_{k_1,k_2} - \eta \nabla f\big)\big),$$

where η is the step size for the gradient step, ∇f is the gradient of the least square objective function, $\hat{P}_{\lambda\eta}$ is the singular space soft-thresholding operator with threshold $\lambda\eta$ and $\hat{P}$ is the approximated projection operator that make any given matrix have zero row sums (by shifting rows) and zero column sums (by shifting columns). We simulated independent random low-rank matrix C_{k_1,k_2}s and make them have zero column sums and row sums by $B_{k_1,k_2} = \hat{P}(C_{k_1,k_2})$.

Table 3 shows the average (with standard deviation) RMSE under different r, d, n combinations under 50 runs. In general, the RMSE in estimating the tensor coefficient increases as s and d increases.

TABLE 3
Pairwise interaction model: RMSE were computed based on 50 *simulations runs. Numbers in parentheses are standard errors*

n	d_1	d_2	d_3	s	σ	**RMSE**	**SNR**
2000	40	40	40	5	10	0.54 (0.02)	2.4 (0.1)
2000	40	40	40	10	10	0.70 (0.01)	3.4 (0.1)
2000	20	20	20	5	10	0.39 (0.01)	1.7 (0.1)
2000	20	20	20	10	10	0.37 (0.01)	2.3 (0.1)
1000	20	20	20	5	10	0.58 (0.02)	1.6 (0.1)
1000	20	20	20	10	10	0.63 (0.02)	2.3 (0.1)

SUPPLEMENTARY MATERIAL

Proofs (DOI: 10.1214/18-AOS1725SUPP; .pdf). We provide all the proofs to the main theorem.

REFERENCES

[1] ANDERSON, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed. Wiley, New York. MR0771294

[2] BASU, S. and MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43** 1535–1567. MR3357870

[3] BHATIA, R. (1997). *Matrix Analysis. Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

[4] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

[5] CHEN, S., LYU, M. R., KING, I. and XU, Z. (2013). Exact and stable recovery of pairwise interaction tensors. In *Advances in Neural Information Processing Systems*.

[6] COHEN, S. B. and COLLINS, M. (2012). Tensor decomposition for fast parsing with latent-variable PCFGS. In *Advances in Neural Information Processing Systems*.

[7] GANDY, S., RECHT, B. and YAMADA, I. (2011). Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Probl.* **27** 025010, 19. MR2765628

[8] GORDON, Y. (1988). On Milman's inequality and random subspaces which escape through a mesh in $\mathbf{R}^n$. In *Geometric Aspects of Functional Analysis* (1986/87). *Lecture Notes in Math.* **1317** 84–106. Springer, Berlin. MR0950977

[9] HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations. Monographs on Statistics and Applied Probability* **143**. CRC Press, Boca Raton, FL. MR3616141

[10] HOFF, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *Ann. Appl. Stat.* **9** 1169–1193. MR3418719

[11] KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. MR2535056

[12] LI, N. and LI, B. (2010). Tensor completion for on-board compression of hyperspectral images. In 17*th IEEE International Conference on Image Processing* (*ICIP*) 517–520.

[13] LÜTKEPOHL, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin. MR2172368

[14] MENDELSON, S. (2016). Upper bounds on product and multiplier empirical processes. *Stochastic Process. Appl.* **126** 3652–3680. MR3565471

[15] MESGARANI, N., SLANEY, M. and SHAMMA, S. (2006). Content-based audio classification based on multiscale spectro-temporal features. *IEEE Trans. Speech Audio Process.* **14** 920–930.

[16] MU, C., HUANG, B., WRIGHT, J. and GOLDFARB, D. (2014). Square deal: Lower bounds and improved relaxations for tensor recovery. In *International Conference on Machine Learning*.

[17] NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **13** 1665–1697. MR2930649

[18] NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133

[19] NION, D. and SIDIROPOULOS, N. D. (2010). Tensor algebra and multidimensional harmonic retrieval in signal processing for MIMO radar. *IEEE Trans. Signal Process.* **58** 5693–5705. MR2789612

[20] PISIER, G. (1989). *The Volume of Convex Bodies and Banach Space Geometry. Cambridge Tracts in Mathematics* **94**. Cambridge Univ. Press, Cambridge. MR1036275

[21] QIN, Z., SCHEINBERG, K. and GOLDFARB, D. (2013). Efficient block-coordinate descent algorithms for the group Lasso. *Math. Program. Comput.* **5** 143–169. MR3069877

[22] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259. MR2719855

[23] RASKUTTI, G., YUAN, M. and CHEN, H. (2019). Supplement to "Convex regularization for high-dimensional multiresponse tensor regression." DOI:10.1214/18-AOS1725SUPP.

[24] RENDLE, S., MARINHO, L. B., NANOPOULOS, A. and SCHMIDT-THIEME, L. (2009). Learning optimal ranking with tensor factorization for tag recommendation. In *SIGKDD*.

[25] RENDLE, S. and SCHMIDT-THIEME, L. (2010). Pairwise interaction tensor factorization for personalized tag recommendation. In *ICDM*.

[26] ROCKAFELLAR, R. T. (1970). *Convex Analysis. Princeton Mathematical Series* **28**. Princeton Univ. Press, Princeton, NJ. MR0274683

[27] SEMERCI, O., HAO, N., KILMER, M. E. and MILLER, E. L. (2014). Tensor-based formulation and nuclear norm regularization for multienergy computed tomography. *IEEE Trans. Image Process.* **23** 1678–1693. MR3191324

[28] SIMON, N., FRIEDMAN, J. and HASTIE, T. (2013). A blockwise coordinate descent algorithm for penalized multiresponse and grouped multinomial regression. Technical report, Georgia Tech.

[29] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[30] TURLACH, B. A., VENABLES, W. N. and WRIGHT, S. J. (2005). Simultaneous variable selection. *Technometrics* **47** 349–363. MR2164706

[31] VAN DE GEER, S. (2014). Weakly decomposable regularization penalties and structured sparsity. *Scand. J. Stat.* **41** 72–86. MR3181133

[32] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574

[33] YUAN, M. and ZHANG, C.-H. (2016). On tensor completion via nuclear norm minimization. *Found. Comput. Math.* **16** 1031–1068. MR3529132

[34] ZHOU, S. (2009). Restricted eigenvalue conditions on subgaussian random matrices. Technical report, ETH, Zurich. Available at arXiv:0912.4045.

G. Raskutti
H. Chen
Department of Statistics
University of Wisconsin-Madison
330 N Orchard St.
Madison, Wisconsin 53715
USA
E-mail: raskutti@stat.wisc.edu
hanchen@stat.wisc.edu

M. Yuan
Department of Statistics
University of Wisconsin-Madison
330 N Orchard St.
Madison, Wisconsin 53715
USA
and
Statistics Department
Columbia University
1255 Amsterdam Avenue
New York, New York 10027
USA
E-mail: ming.yuan@columbia.edu