Convex Regularization for High-Dimensional Tensor Regression

Garvesh Raskutti* and Ming Yuan†

University of Wisconsin-Madison

Abstract

In this paper we present a general convex optimization approach for solving high-dimensional 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. These general bounds provide useful upper bounds on rates of convergence for a number of fundamental statistical models of interest including multi-response regression, vector auto-regressive models, low-rank tensor models and pairwise interaction models. Moreover, in many of these settings we prove that the resulting estimates are minimax optimal.

*Departments of Statistics and Computer Science, and Optimization Group at Wisconsin Institute for Discovery, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706. The research of Garvesh Raskutti is supported in part by NSF Grant DMS-1407028

†Department of Statistics and Morgridge Institute for Research, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706. The research of Ming Yuan was supported in part by NSF FRG Grant DMS-1265202, and NIH Grant 1-U54AI117924-01.

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, e.g., Buhlmann and van de Geer (2011) and Hastie et al. (2015), for an overview. Most of the prior work has focussed 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 hyper-spectral image analysis (Li and Li, 2010), multi-energy computed tomography (Semerci et al., 2014), radar signal processing (Sidiropoulos and Nion, 2010), audio classification (Mesgarani et al., 2006) and text mining (Cohen and Collins, 2012) among numerous 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:

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

Here $T \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ is an unknown parameter of interest, and $\epsilon^{(i)}$s are independent and identically distributed noise tensors whose entries are independent and identically distributed centered normal random variables. 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

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

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). Multi-response regression (see, e.g., Anderson, 1984), vector autoregressive model (see, e.g., Lütkepolhl, 2006), and pairwise interaction tensor model (see, e.g., Rendle and Schmidt-Thieme, 2010) 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_ks 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:

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

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 multi-response regression, high-dimensional vector auto-regressive 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 principled approach to exploit the low dimensional structure in these problems. In doing so, we extend the notion of decomposability originally introduced by Negahban et al. (2012) for vector and matrix models to *weak decomposability* 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 (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. Finally, our general results lead to novel upper bounds for several important regression problems involving high-dimensional tensors: multi-response regression, multi-variate auto-regressive models and pairwise interaction models, for which we also prove that the resulting estimates are minimiax rate optimal with appropriate choices of regularizers.

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, multi-response 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. The proofs are presented in Section 5.

2 Methodology

Recall the regularized least-squares objective:

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

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 particularly interest here is the so-called *weakly decomposable* convex regularizers, extending a similar concept introduced by Negahban et al. (2012) 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 $B \in \mathcal{B}$,

$$\mathcal{R}(A + B) \geq \mathcal{R}(A) + c_{\mathcal{R}} \mathcal{R}(B). \tag{3}$$

In particular, if (3) holds for any $B \in \mathcal{B} = \mathcal{A}$, we say $\mathcal{R}(\cdot)$ is weakly decomposable with respect to $\mathcal{A}$. 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 Negahban et al. (2012). 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 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:

$$\mathcal{R}(A) := \sum_{j_1=1}^{d_1} \sum_{j_2=1}^{d_2} \sum_{j_3=1}^{d_3} |A_{j_1 j_2 j_3}|, \tag{4}$$

following the same idea as the Lasso for linear regression (see, e.g., Tibshirani, 1996). 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

$$\mathcal{A}(I) = \mathcal{B}(I) = \left\{ A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : A_{j_1 j_2 j_3} = 0 \text{ for all } (j_1, j_2, j_3) \notin I \right\}. \tag{5}$$

It is clear that

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

and $\mathcal{R}(A)$ defined by (4) 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

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

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:

$$\mathcal{R}(A) = \sum_{j_2=1}^{d_2} \sum_{j_3=1}^{d_3} \|A_{\cdot j_2 j_3}\|_{\ell_2}, \tag{6}$$

similar to the group Lasso (see, e.g., Yuan and Lin, 2006), 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

$$\mathcal{A}(I) = \mathcal{B}(I) = \left\{ A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : A_{j_1 j_2 j_3} = 0 \text{ for all } (j_2, j_3) \notin I \right\}. \tag{7}$$

It is clear that

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

and $\mathcal{R}(A)$ defined by (6) is decomposable with respect to $\mathcal{A}$ with $c_{\mathcal{R}} = 1$. Note that in defining the regularizer in (6), instead of vector ℓ_2 norm, other ℓ_q $(q > 1)$ norms could also be used. See, e.g., Turlach et al. (2005).

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:

$$\mathcal{R}(A) = \sum_{j_3=1}^{d_3} \|A_{\cdot\cdot j_3}\|. \tag{8}$$

Typical examples of the matrix norm that can be used in (8) 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

$$\mathcal{A}(I) = \mathcal{B}(I) = \left\{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : A_{j_1 j_2 j_3} = 0 \text{ for all } j_3 \notin I\right\}. \tag{9}$$

for any $I \subset [d_3]$.

Now consider the case when we use the matrix nuclear norm $\|\cdot\|_*$ in (8). Let P_{1j} and P_{2j}, $j = 1, \ldots, d_3$ be two sequences of projection matrices on $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$ respectively. Let

$$\mathcal{A}(P_{1j}, P_{2j} : 1 \le j \le d_3) = \left\{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : P_{1j}^{\perp} A_{\cdot\cdot j} P_{2j}^{\perp} = 0, j = 1, \ldots, d_3\right\}, \tag{10}$$

and

$$\mathcal{B}(P_{1j}, P_{2j} : 1 \le j \le d_3) = \left\{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : A_{\cdot\cdot j} = P_{1j} A_{\cdot\cdot j} P_{2j}, j = 1, \ldots, d_3\right\}. \tag{11}$$

By pinching inequality (see, e.g., Bhatia, 1997), 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, e.g., Koldar and Bader (2009), 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 $A \in \mathbb{R}^{d_1 \times d_2 \times d_3}$:

$$A = \sum_{k=1}^{r} u_k \otimes v_k \otimes w_k \tag{12}$$

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 Yuan and Zhang (2014), we define the nuclear norm of A through its dual norm. More specifically, let the spectral norm of A be given by

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

Then its nuclear norm is defined as

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

We shall then consider the regularizer:

$$\mathcal{R}(A) = \|A\|_*. \tag{13}$$

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 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*

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

Lemma 1 is a direct consequence from the characterization of sub-differential for tensor nuclear norm given by Yuan and Zhang (2014), and can be viewed as a tensor version of the pinching inequality for matrices.

Write

$$\mathcal{A}(P_1, P_2, P_3) = \left\{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : QA = A\right\}, \tag{14}$$

and

$$\mathcal{B}(P_1, P_2, P_3) = \left\{A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : (P_1 \otimes P_2 \otimes P_3)A = A\right\}. \tag{15}$$

By Lemma 1, $\mathcal{R}(\cdot)$ defined by (13) 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 Yuan and Zhang (2014) 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:

$$A_{j_1 j_2 j_3} = \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \sum_{k_3=1}^{r_3} S_{k_1 k_2 k_3} U_{j_1 k_1} V_{j_2 k_2} W_{j_3 k_3} \tag{16}$$

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 (12) 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:

$$\mathcal{R}(A) = \frac{1}{3} \sum_{k=1}^{3} \|\mathcal{M}_k(A)\|_*. \tag{17}$$

By the pinching inequality for matrices, $\mathcal{R}(\cdot)$ defined by (17) 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 $S \subset \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ is given by

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

where $G \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ is a tensor whose entries are independent $\mathcal{N}(0,1)$ random variables. See, e.g., Gordon (1988).

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., Pisier, 1989).We also define the unit ball for the norm-regularizer $\mathcal{R}(.)$ as follows:

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

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 Negahban et al. (2012), 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

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

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

$$c_\ell^2 \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le c_u^2, \tag{18}$$

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

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 (18) boils down to similar conditions on $\mathrm{cov}(X^{(i)}, X^{(i)})$. However (18) 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:

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

Theorem 1. *Suppose that (1) holds for a tensor T from a linear subspace $\mathcal{A}_0 \subset \mathbb{R}^{d_1 \times \cdots \times d_N}$ where (18) holds. Let $\widehat{T}$ be defined by (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*

$$\lambda \ge \frac{2c_u(3 + c_{\mathcal{R}})}{c_{\mathcal{R}}\sqrt{n}} w_G[\mathbb{B}_{\mathcal{R}}(1)], \tag{19}$$

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

$$\max\left\{\|\widehat{T} - T\|_n^2, \|\widehat{T} - T\|_{\mathrm{F}}^2\right\} \le \frac{6(1 + c_{\mathcal{R}})}{3 + c_{\mathcal{R}}} \frac{9c_u^2}{c_\ell^2} s(\mathcal{A})\lambda^2, \tag{20}$$

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

As stated in Theorem 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 $\mathcal{A}$.

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., Rockafellar, 1970, 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 $w_G[\mathbb{B}_{\mathcal{R}}(1)] = \mathbb{E}[\mathcal{R}^*(G)]$.

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,

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

It can then be shown that:

Lemma 2. *There exists a constant $0 < c < \infty$ such that*

$$w_G[\mathbb{B}_{\mathcal{R}_1}(1)] \leq c\sqrt{\log(d_1 d_2 d_3)}. \tag{21}$$

Let

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

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 (5). It is easy to verify that for any $A \in \Theta_1(s)$,

$$s_1(\mathcal{A}(I)) = \sup_{B \in \mathcal{A}(I(A))/\{0\}} \frac{\mathcal{R}_1^2(B)}{\|B\|_{\mathrm{F}}^2} \leq s. \tag{22}$$

In light of (22) and (21), Theorem 1 implies that

$$\sup_{T \in \Theta_1(s)} \max\left\{\|\widehat{T}_1 - T\|_n^2, \|\widehat{T}_1 - T\|_{\mathrm{F}}^2\right\} \lesssim \frac{s \log(d_1 d_2 d_3)}{n},$$

with high probability by taking

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

where $\widehat{T}_1$ is the regularized least squares estimate defined by (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. *There exists a constant $0 < c < \infty$ such that*

$$w_G[\mathbb{B}_{\mathcal{R}_2}(1)] \leq c\sqrt{\max\{d_1, \log(d_2 d_3)\}}. \tag{23}$$

Let

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

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 (7). It is easy to verify that for any $A \in \Theta_2(s)$,

$$s_2(\mathcal{A}(I)) = \sup_{B \in \mathcal{A}(I(A))/\{0\}} \frac{\mathcal{R}_2^2(B)}{\|B\|_{\mathrm{F}}^2} \leq s. \tag{24}$$

In light of (24) and (30), Theorem 1 implies that

$$\sup_{T \in \Theta_2(s)} \max\left\{\|\widehat{T}_2 - T\|_n^2, \|\widehat{T}_2 - T\|_{\mathrm{F}}^2\right\} \lesssim \frac{s \max\{d_1, \log(d_2 d_3)\}}{n},$$

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 (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 $s_1 \approx s_2 d_1$. In this case, $\widehat{T}_2$ enjoys performance superior to that of $\widehat{T}_1$.

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 $\|.\|_*$ 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 4. *There exists a constant $0 < c < \infty$ such that*

$$w_G[\mathbb{B}_{\mathcal{R}_3}(1)] \le c\sqrt{\max\{d_1 d_2, \log(d_3)\}}. \tag{25}$$

Let

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

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

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

Then $\mathcal{R}_3(\cdot)$ is decomposable with respect to $\mathcal{A}(I(A))$ as defined by (9). It is easy to verify that for any $A \in \Theta_3(s)$,

$$s_3(\mathcal{A}(I(A))) = \sup_{B\in\mathcal{A}(I(A))/\{0\}} \frac{\mathcal{R}_3^2(B)}{\|B\|_{\mathrm{F}}^2} \le s. \tag{26}$$

Based on (26) and (25), Theorem 1 implies that

$$\sup_{T\in\Theta_3(s)} \max\left\{ \|\widehat{T}_3 - T\|_n^2, \|\widehat{T}_3 - T\|_{\mathrm{F}}^2 \right\} \lesssim \frac{s\max\{d_1 d_2, \log(d_3)\}}{n},$$

with high probability by taking

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

where $\widehat{T}_3$ is the regularized least squares estimate defined by (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 5. *There exists a constant $0 < c < \infty$ such that*

$$w_G[\mathbb{B}_{\mathcal{R}_4}(1)] \leq c\sqrt{\max\{d_1, d_2, \log(d_3)\}}. \tag{27}$$

Now consider

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

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 $A \in \mathcal{B}(P_{1j}, P_{2j} : 1 \leq j \leq d_3)$ as defined by (11). In addition, recall that $\mathcal{R}_4$ is decomposable with respect to $\mathcal{B}(P_{1j}, P_{2j} : 1 \leq j \leq d_3)$ and $\mathcal{A}(P_{1j}, P_{2j} : 1 \leq j \leq d_3)$ as defined by (10). It is not hard to see that for any $A \in \Theta_4(r)$, $\mathcal{A}(P_{1j}, P_{2j} : 1 \leq j \leq d_3) \subset \Theta_4(2r)$, from which we can derive that:

Lemma 6. *For any $A \in \Theta_4(r)$,*

$$s_4(\mathcal{A}(P_{1j}, P_{2j} : 1 \leq j \leq d_3)) \leq \sup_{B \in \mathcal{A}/\{0\}} \frac{\mathcal{R}_4^2(B)}{\|B\|_{\mathrm{F}}^2} \leq 2r. \tag{28}$$

In light of (28) and (27), Theorem 1 implies that

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

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 (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 7. *There exists a constant $0 < c < \infty$ such that*

$$w_G[\mathbb{B}_{\mathcal{R}_5}(1)] \leq c\sqrt{(d_1 + d_2 + d_3)}. \tag{29}$$

Now let

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

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 (14) and (15) respectively. It can also be shown that

Lemma 8. *For any $A \in \Theta_5(r)$,*

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

Lemmas 7 and 8 show that

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

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 (2) when using regularizer $\mathcal{R}_5(\cdot)$.

Next we consider the low-rankness regularization via matricization:

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

It is not hard to see that

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

Lemma 9. *There exists a constant* $0 < c < \infty$ *such that*

$$w_G[\mathbb{B}_{\mathcal{R}_6}(1)] \leq c\sqrt{\max\{d_1d_2, d_2d_3, d_1d_3\}}. \tag{30}$$

On the other hand,

Lemma 10. *For any* $A \in \Theta_5(r)$,

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

Lemmas 9 and 10 suggest that

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

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 (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 $r \leq \min\{d_1, d_2, d_3\}$. 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 Gandy et al. (2011); Mu et al. (2014) and Yuan and Zhang (2014).

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.

4.1 Multi-Response regression with large p

The first example we consider is the multi-response regression model:

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

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 multi-response 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:

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

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 j^{th} feature for the k^{th} 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:

$$\mathcal{R}(A) = \sum_{j=1}^{p} \|A_{\cdot\cdot j}\|_{\mathrm{F}}, \tag{31}$$

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

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

Theorem 2. *Under the multi-response regression model with $T \in \mathcal{T}_1$ and independent Gaussian design where $c_\ell^2 \le \lambda_{min}(\widetilde{\Sigma}) \le \lambda_{max}(\widetilde{\Sigma}) \le c_u^2$, if*

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

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 $1 - p^{-c_1}$

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

when n is sufficiently large, where $\widehat{T}$ is the regularized least squares estimate defined by (2) with regularizer given by (31). In addition,

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

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 2 shows that when taking

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

the regularized least squares estimate defined by (2) with regularizer given by (31) 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

$$\mathcal{T}_2 = \left\{ A \in \mathbb{R}^{m \times m \times p} \mid \sum_{j=1}^{p} \mathrm{rank}(A_{\cdot\cdot j}) \leq r \right\}.$$

An appropriate penalty function in this case is:

$$\mathcal{R}(A) = \sum_{j=1}^{p} \|A_{\cdot\cdot j}\|_*, \tag{32}$$

and the corresponding dual function applied to G is:

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

Theorem 3. *Under the multi-response regression model with $T \in \mathcal{T}_2$ and independent Gaussian design where $c_\ell^2 \leq \lambda_{min}(\widetilde{\Sigma}) \leq \lambda_{max}(\widetilde{\Sigma}) \leq c_u^2$, if*

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

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* $1 - p^{-c_1}$,

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

when n *is sufficiently large, where* $\widehat{T}$ *is the regularized least squares estimate defined by (2) with regularizer given by (32). In addition,*

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

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* $\{(X^{(i)}, Y^{(i)}) : 1 \leq i \leq n\}$.

Again Theorem 3 shows that by taking

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

the regularized least squares estimate defined by (2) with regularizer given by (32) 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 $\mathcal{T}_2$.

4.2 Multivariate Sparse Auto-regressive Models

Now we consider the setting of vector auto-regressive models. In this case, our generative model is:

$$X^{(t+p)} = \sum_{j=1}^{p} A_j X^{(t+p-j)} + \epsilon^{(t)}, \tag{33}$$

where $1 \leq t \leq n$ represents the time index, $1 \leq j \leq p$ represents the lag index, $\{X^{(t)}\}_{t=0}^{n+p}$ is an m-dimensional vector, $\epsilon^{(t)} \sim \mathcal{N}(0, I_{m\times m})$ 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 k^{th} variable on the ℓ^{th} variable at lag j. This model is studied by Basu and Michailidis (2015) 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 Basu and Michailidis (2015).

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.

$$\mathcal{T}_3 = \left\{ A \in \mathbb{R}^{m \times m \times p} \mid \sum_{k=1}^{m} \sum_{\ell=1}^{m} \mathbb{I}(A_{k\ell\cdot} \neq \mathbf{0}) \leq s \right\}. \tag{34}$$

The penalty function we are considering is:

$$\mathcal{R}(A) = \sum_{k=1}^{m} \sum_{\ell=1}^{m} \|A_{k\ell\cdot}\|_{\ell_2}, \tag{35}$$

and the corresponding dual function applied to G is:

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

The challenge in this setting is that the X's are highly dependent and we use the results developed in Basu and Michailidis (2015) to prove that (18) is satisfied.

Prior to presenting the main results, we introduce concepts developed in Basu and Michailidis (2015) that play a role in determining the constants c_u^2 and c_ℓ^2 which relate to the stability of the auto-regressive processes. A p-variate Gaussian time series is defined by its auto-covariance matrix function

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

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

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

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

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

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:

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

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

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

where

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

and

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

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

$$\frac{1}{\mu_{\max}}\|\Delta\|_{\mathrm{F}}^2 \leq \mathbb{E}\left[\|\Delta\|_n^2\right] \leq \frac{1}{\mu_{\min}}\|\Delta\|_2^2. \tag{36}$$

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

Theorem 4. *Under the vector auto-regressive model defined by* (33) *with* $T \in \mathcal{T}_3$*, if*

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

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* $1 - m^{-c_1}$*,*

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

when n *is sufficiently large, where* $\widehat{T}$ *is the regularized least squares estimators defined by (2) with regularizer given by (35). In addition,*

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

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* $\{X^{(t)} : t = 0, \ldots, n + p\}$*.*

Theorem 4 provides, to our best knowledge, the only lower bound result for multivariate time series, and the upper bound is different from Proposition 4.1 in Basu and Michailidis (2015) since we impose sparsity only on the large m directions and not over the p lags. Our framework also extends to any low-dimensional structure on the tensor A defined by matricization, whereas Basu and Michailidis (2015) impose sparsity through vectorization. Note that Proposition 4.1 in Basu and Michailidis (2015) follows directly from Lemma 2 with $d_1 = p$ and $d_2 = d_3 = m$.

4.3 Pairwise interaction tensor models

Finally, we consider the tensor regression (1) where T follows a pairwise interaction model. More specifically, $(X^{(i)}, Y^{(i)})$, $i = 1, 2, \ldots, n$ 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

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

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

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

The pairwise interaction was used originally by Rendle et al. (2009); Rendle and Schmidt-Thieme (2010) for personalized tag recommendation, and later analyzed in Chen et al. (2013). Hoff (2003) briefly introduced a single index additive model (amongst other tensor models) which is a sub-class of the pairwise interaction model. The regularizer we consider is:

$$\mathcal{R}(A) = \|A^{(12)}\|_* + \|A^{(13)}\|_* + \|A^{(23)}\|_*. \tag{37}$$

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

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

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

Theorem 5. *Under the pairwise interaction model with $T \in \mathcal{T}_4$, if*

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

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 $1 - \min\{d_1, d_2, d_3\}^{-c_1}$,

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

when n is sufficiently large, where $\widehat{T}$ is the regularized least squares estimate defined by (2) with regularizer given by (37). In addition,

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

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 5 establishes the minimax optimality of the regularized least squares estimate (2) when using an appropriate convex decomposable regularizer.

5 Proofs

In this section, we present the proofs to our main results.

5.1 Proof of Theorem 1

The initial steps exploit weak decomposability and are similar to those from Negahban et al. (2012). After the initial steps, we use properties of Gaussian random variables and suprema of Gaussian processes to derive our general upper bound. Throughout $\mathcal{R}(A)$ refers to the weakly decomposable regularizer over the tensor A. For a tensor A, we shall write A_0 and $A^\perp$ as its projections onto $\mathcal{A}_0$ and $\mathcal{A}^\perp$ with respect to the Frobenius norm, respectively.

Since $\widehat{T}$ is the empirical minimizer,

$$\frac{1}{2n}\sum_{i=1}^n \|Y^{(i)} - \langle X^{(i)}, \widehat{T}\rangle\|_{\mathrm{F}}^2 + \lambda\mathcal{R}(\widehat{T}) \;\leq\; \frac{1}{2n}\sum_{i=1}^n \|Y^{(i)} - \langle X^{(i)}, T\rangle\|_{\mathrm{F}}^2 + \lambda\mathcal{R}(T).$$

Substituting $Y^{(i)} = \langle X^{(i)}, T\rangle + \epsilon^{(i)}$ and $\Delta = \widehat{T} - T$,

$$\begin{aligned}
\frac{1}{2n}\sum_{i=1}^{n} \|\langle X^{(i)}, \Delta\rangle\|_{\mathrm{F}}^2 &\leq \frac{1}{n}\left|\sum_{i=1}^{n}\langle \epsilon^{(i)} \otimes X^{(i)}, \Delta\rangle\right| + \lambda(\mathcal{R}(T) - \mathcal{R}(\widehat{T})) \\
&\leq \mathcal{R}^*\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon^{(i)} \otimes X^{(i)}\right)\mathcal{R}(\Delta) + \lambda(\mathcal{R}(T) - \mathcal{R}(\widehat{T}_0) - c_{\mathcal{R}}\mathcal{R}(\widehat{T}^{\perp})) \\
&\leq \mathcal{R}^*\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon^{(i)} \otimes X^{(i)}\right)\mathcal{R}(\Delta) + \lambda(\mathcal{R}(\Delta_0) - c_{\mathcal{R}}\mathcal{R}(\Delta^{\perp})),
\end{aligned}$$

where the second inequality follows from the decomposability and the last one follows from triangular inequality.

Let $G \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ be an tensor where each entry is i.i.d. $\mathcal{N}(0,1)$. Recall the definition of Gaussian width:

$$w_G[\mathbb{B}_{\mathcal{R}}(1)] = \mathbb{E}[\mathcal{R}^*(G)].$$

For simplicity let

$$\eta_{\mathcal{R}} = \frac{3 + c_{\mathcal{R}}}{2c_{\mathcal{R}}}$$

and recall that $\lambda \geq 2c_u\eta_{\mathcal{R}}n^{-1/2}\mathbb{E}[\mathcal{R}^*(G)]$. We have the following Lemma:

Lemma 11. *If* $\lambda \geq 2c_u\eta_{\mathcal{R}}n^{-1/2}\mathbb{E}[\mathcal{R}^*(G)]$, *then*

$$\lambda \geq \eta_{\mathcal{R}}\mathcal{R}^*\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon^{(i)} \otimes X^{(i)}\right),$$

with probability at least $1 - \exp\{-\eta_{\mathcal{R}}^2\mathbb{E}[\mathcal{R}^*(G)]^2/4\}$

The proof relies on Gaussian comparison inequalities and concentration inequalities.

Proof of Lemma 11. Recall that we have set:

$$\lambda \geq \frac{2\eta_{\mathcal{R}}c_u}{\sqrt{n}}\mathbb{E}[\mathcal{R}^*(G)].$$

First we show that $\lambda \geq 2c_u\eta_{\mathcal{R}}n^{-1/2}\mathcal{R}^*(G)$ with high probability using concentration of Lipschitz functions for Gaussian random variables (see Theorem 8 in Appendix A). First we prove that $f(G) = \mathcal{R}^*(G) = \sup_{A\in\mathbb{B}_{\mathcal{R}}(1)}\langle G, A\rangle$ is a 1-Lipschitz function in terms of G. In particular note that:

$$f(G) - f(G') = \sup_{A:\mathcal{R}(A)\leq 1}\langle G, A\rangle - \sup_{A:\mathcal{R}(A)\leq 1}\langle G', A\rangle.$$

Let $\widetilde{A} := \arg\max_{A:\mathcal{R}(A)\leq 1}\langle G, A\rangle$. Then

$$\begin{aligned}
\sup_{A:\mathcal{R}(A)\leq 1}\langle G, A\rangle - \sup_{A:\mathcal{R}(A)\leq 1}\langle G', A\rangle &= \langle G, \widetilde{A}\rangle - \sup_{\mathcal{R}(A)\leq 1}\langle G', A\rangle \\
&\leq \langle G, \widetilde{A}\rangle - \langle G', \widetilde{A}\rangle \\
&\leq \langle G - G', \widetilde{A}\rangle \\
&\leq \sup_{A:\mathcal{R}(A)\leq 1}\langle G - G', A\rangle \\
&\leq \sup_{A:\|A\|_{\mathrm{F}}\leq 1}\langle G - G', A\rangle \\
&\leq \|G - G'\|_{\mathrm{F}},
\end{aligned}$$

where recall that $\|A\|_F \leq \mathcal{R}(A)$ which implies the second last inequality. Therefore $f(G)$ is a 1-Lipschitz function with respect to the Frobenius norm. Therefore, by applying Theorem 8 in Appendix A,

$$\mathbb{P}\left\{\left|\sup_{A\in\mathbb{B}_{\mathcal{R}}(1)}\langle G, A\rangle - \mathbb{E}[\sup_{A\in\mathbb{B}_{\mathcal{R}}(1)}\langle G, A\rangle]\right| > w_G(\mathbb{B}_{\mathcal{R}}(1))\right\} \leq 2\exp\left(-\frac{1}{2}w_G^2[\mathbb{B}_{\mathcal{R}}(1)]\right).$$

Therefore

$$\lambda \geq \frac{\eta_{\mathcal{R}} c_u}{\sqrt{n}}\mathcal{R}^*(G)$$

with probability at least $1 - 2\exp\{-w_G^2[\mathbb{B}_{\mathcal{R}}(1)]/2\}$.

To complete the proof, we use a Gaussian comparison inequality between the supremum of the process $c_u n^{-1/2}\langle G, A\rangle$ and $n^{-1}\sum_{i=1}^n\langle \epsilon^{(i)}\otimes X^{(i)}, A\rangle$ over the set $\mathbb{B}_{\mathcal{R}}(1)$. Recall that:

$$\mathcal{R}^*\left(\frac{1}{n}\sum_{i=1}^n \epsilon^{(i)}\otimes X^{(i)}\right) = \sup_{A\in\mathbb{B}_{\mathcal{R}}(1)}\left\langle A, \frac{1}{n}\sum_{i=1}^n \epsilon^{(i)}\otimes X^{(i)}\right\rangle.$$

Recall that each $\epsilon^{(i)} \in \mathbb{R}^{d_{M+1}\times d_{M+2}\times\ldots\times d_N}$ is an i.i.d. standard Gaussian tensor and $\mathrm{vec}(X) \in \mathbb{R}^{nd_1d_2\cdots d_M}$ is a Gaussian vector covariance $\Sigma \in \mathbb{R}^{(nD_M)\times(nD_M)}$. Further let $\{w^{(i)} : i = 1, \ldots, n\}$ be i.i.d. standard normal Gaussian tensors where $w^{(i)} \in \mathbb{R}^{d_1\times d_2\times\ldots\times d_M}$. Assuming (18) and using a standard Gaussian comparison inequality due to Lemma 13 in Appendix A proven earlier in Anderson (1955), we get

$$\mathbb{P}\left\{\sup_{A:\mathcal{R}(A)\leq 1}\frac{1}{n}\sum_{i=1}^n\langle\epsilon^{(i)}\otimes X^{(i)}, A\rangle > x\right\} \leq \mathbb{P}\left\{\sup_{A:\mathcal{R}(A)\leq 1}\frac{1}{n}\sum_{i=1}^n\langle\epsilon^{(i)}\otimes w^{(i)}, A\rangle > \frac{x}{c_u}\right\},$$

since

$$\mathrm{Cov}(\mathrm{vec}(X)) = \Sigma \preceq c_u^2 I_{(nD_M)\times(nD_M)}.$$

Now we apply Slepian's lemma (Slepian, 1962) to complete the proof. For completeness, Slepian's lemma is included in Appendix A. Clearly for any A,

$$\frac{1}{n}\sum_{i=1}^{n} \langle \mathbb{E}[\epsilon^{(i)} \otimes w^{(i)}], A\rangle = 0.$$

Further a simple calculation shows that for any A,

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} \langle [\epsilon^{(i)} \otimes w^{(i)}], A\rangle\right) = \frac{\|A\|_{\mathrm{F}}^2}{n},$$

where we have exploited independence between across all samples and fibers, and the fact that ϵ and w are independent. Further, for all A, A',

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} \langle [\epsilon^{(i)} \otimes w^{(i)}], A - A'\rangle\right) = \frac{\|A - A'\|_{\mathrm{F}}^2}{n}.$$

Now let $G \in \mathbb{R}^{d_1\times d_2\times\ldots\times d_N}$ be an i.i.d. standard normal tensor and define the zero-mean Gaussian process,

$$\frac{1}{\sqrt{n}}\langle G, A\rangle,$$

for any $A \in \mathbb{B}_{\mathcal{R}}(1)$. It is straightforward to show that,

$$\mathrm{Var}\left(\frac{1}{\sqrt{n}}\langle G, A\rangle\right) = \frac{\|A\|_{\mathrm{F}}^2}{n},$$

and

$$\mathrm{Var}\left(\frac{1}{\sqrt{n}}\langle G, A - A'\rangle\right) = \frac{\|A - A'\|_{\mathrm{F}}^2}{n},$$

for all A, A'. Therefore, directly applying Slepian's lemma (Lemma 14 in Appendix A),

$$\mathbb{P}\left\{\sup_{\mathcal{R}(A)\le 1} \frac{1}{n}\sum_{i=1}^{n} \langle \epsilon^{(i)} \otimes w^{(i)}, A\rangle > x\right\} \le \mathbb{P}\left\{\sup_{\mathcal{R}(A)\le 1} \frac{1}{\sqrt{n}}\langle G, A\rangle > x\right\},$$

for all $x > 0$. Substituting x by x/c_u means that

$$\mathbb{P}\left\{\mathcal{R}^*\left(\frac{1}{n}\sum_{i=1}^{n} \langle \epsilon^{(i)} \otimes w^{(i)}, A\rangle\right) > x\right\} \le \mathbb{P}\left\{\frac{c_u}{\sqrt{n}}\mathcal{R}^*(G) > x\right\},$$

for any $x > 0$. This completes the proof. □

In light of Lemma 11, for the remainder of the proof, we can condition on the event that

$$\lambda \geq \eta_{\mathcal{R}} \mathcal{R}^* \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon^{(i)} \otimes X^{(i)} \right).$$

Under this event,

$$\begin{aligned} \frac{1}{2n} \sum_{i=1}^{n} \|\langle X^{(i)}, \Delta \rangle\|_{\mathrm{F}}^2 &\leq \frac{1}{\eta_{\mathcal{R}}} \lambda \mathcal{R}(\Delta) + \lambda(\mathcal{R}(\Delta_0) - c_{\mathcal{R}} \mathcal{R}(\Delta^{\perp})) \\ &\leq \left(1 + \frac{1}{\eta_{\mathcal{R}}}\right) \lambda \mathcal{R}(\Delta_0) - \left(c_{\mathcal{R}} - \frac{1}{\eta_{\mathcal{R}}}\right) \lambda \mathcal{R}(\Delta^{\perp}). \end{aligned}$$

Since

$$\frac{1}{2n} \sum_{i=1}^{n} \|\langle \Delta, X^{(i)} \rangle\|_{\mathrm{F}}^2 \geq 0,$$

we get

$$\mathcal{R}(\Delta^{\perp}) \leq \frac{3}{c_{\mathcal{R}}} \mathcal{R}(\Delta_0).$$

Hence we define the cone

$$\mathcal{C} = \left\{ \Delta \mid \mathcal{R}(\Delta^{\perp}) \leq 3 c_{\mathcal{R}}^{-1} \mathcal{R}(\Delta_0) \right\},$$

and know that $\Delta \in \mathcal{C}$. Hence

$$\frac{1}{2n} \sum_{i=1}^{n} \|\langle X^{(i)}, \Delta \rangle\|_{\mathrm{F}}^2 \leq \frac{3(1 + c_{\mathcal{R}})}{3 + c_{\mathcal{R}}} \lambda \mathcal{R}(\Delta_0) \leq \frac{3(1 + c_{\mathcal{R}})}{3 + c_{\mathcal{R}}} \sqrt{s(\mathcal{A})} \lambda \|\Delta\|_{\mathrm{F}}.$$

Recall that

$$\frac{1}{n} \sum_{i=1}^{n} \|\langle X^{(i)}, \Delta \rangle\|_{\mathrm{F}}^2 = \|\Delta\|_n^2.$$

Thus,

$$\|\Delta\|_n^2 \leq \frac{6(1 + c_{\mathcal{R}})}{3 + c_{\mathcal{R}}} \sqrt{s(\mathcal{A})} \lambda \|\Delta\|_{\mathrm{F}}.$$

For convenience, in the remainder of this proof let

$$\delta_n := \frac{6(1 + c_{\mathcal{R}})}{3 + c_{\mathcal{R}}} \sqrt{s(\mathcal{A})} \lambda.$$

Now we split into three cases. (i) If $\|\Delta\|_n \geq \|\Delta\|_{\mathrm{F}}$, then

$$\max\{\|\Delta\|_n, \|\Delta\|_{\mathrm{F}}\} \leq \delta_n.$$

On the other hand if (ii) $\|\Delta\|_n \leq \|\Delta\|_{\mathrm{F}}$ and $\|\Delta\|_{\mathrm{F}} \leq \frac{c_u}{c_\ell}\delta_n$, then

$$\max\{\|\Delta\|_n, \|\Delta\|_{\mathrm{F}}\} \leq \frac{c_u}{c_\ell}\delta_n.$$

Hence the only case we need to consider is (iii) $\|\Delta\|_n \leq \|\Delta\|_{\mathrm{F}}$ and $\|\Delta\|_{\mathrm{F}} \geq c_u c_\ell^{-1}\delta_n$. Now we follow a similar proof technique to the proof for Theorem 1 in Raskutti et al. (2012).

Let us define the following set:

$$\mathcal{C}(\delta_n) := \left\{\Delta \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N} \mid \mathcal{R}(\Delta^{\perp}) \leq 3c_{\mathcal{R}}^{-1}\mathcal{R}(\Delta_0),\ \|\Delta\|_n \leq \|\Delta\|_{\mathrm{F}}\right\}.$$

Further, let us define the event:

$$\mathcal{E}(\delta_n) := \left\{\|\Delta\|_n^2 \geq \frac{1}{4}\|\Delta\|_{\mathrm{F}}^2 \mid \Delta \in \mathcal{C}(\delta_n),\ \|\Delta\|_{\mathrm{F}} \geq \frac{c_u}{c_\ell}\delta_n\right\}.$$

Let us define the alternative event:

$$\mathcal{E}'(\delta_n) := \{\|\Delta\|_n^2 \geq \frac{1}{4}\|\Delta\|_{\mathrm{F}}^2 \mid \Delta \in \mathcal{C}(\delta_n),\ \|\Delta\|_{\mathrm{F}} = \frac{c_u}{c_\ell}\delta_n\}.$$

We claim that it suffices to show that $\mathcal{E}'(\delta_n)$ holds with probability at least $1 - \exp(-cn)$ for some constant $c > 0$. In particular, given an arbitrary non-zero $\Delta \in \mathcal{C}(\delta_n)$, consider the re-scaled tensor

$$\widetilde{\Delta} = \frac{c_u\delta_n}{c_\ell}\frac{\Delta}{\|\Delta\|_{\mathrm{F}}}.$$

Since $\Delta \in \mathcal{C}(\delta_n)$ and $\mathcal{C}(\delta_n)$ is star-shaped, we have $\widetilde{\Delta} \in \mathcal{C}(\delta_n)$ and $\|\widetilde{\Delta}\|_{\mathrm{F}} = c_u c_\ell^{-1}\delta_n$ by construction. Consequently, it is sufficient to prove that $\mathcal{E}'(\delta_n)$ holds with high probability.

Lemma 12. *Under the assumption that for any $c' > 0$, there exists an n such that $\sqrt{s}\lambda \leq c'$, there exists a $\tilde{c} > 0$ such that*

$$\mathbb{P}\big(\mathcal{E}'(\delta_n)\big) \geq 1 - \exp(-\tilde{c}n).$$

Proof of Lemma 12. Denote by $D_N = d_1 d_2 \cdots d_N$ and $D_M = d_1 d_2 \cdots d_M$. Now we define the random variable

$$Z_n(\mathcal{C}(\delta_n)) = \sup_{\Delta \in \mathcal{C}(\delta_n)} \left\{\frac{c_u^2}{c_\ell^2}\delta_n^2 - \frac{1}{n}\sum_{i=1}^{n}\|\langle \Delta, X^{(i)}\rangle\|_F^2\right\},$$

then it suffices to show that

$$Z_n(\mathcal{C}(\delta_n)) \leq \frac{c_u^2\delta_n^2}{2c_\ell^2}.$$

Recall that the norm

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

Let $N_{\rm pr}(\epsilon; \mathcal{C}(\delta_n); \|\cdot\|_n)$ denote the proper covering number of $\mathcal{C}(\delta_n)$ in $\|\cdot\|_n$ norm. Now let $\Delta^1, \Delta^2, \ldots, \Delta^{\mathcal{N}}$, be a minimal $c_u\delta_n/(8c_\ell)$-proper covering of $\mathcal{C}(\delta_n)$, so that for all $\Delta \in \mathcal{C}(\delta_n)$, there exists a k such that

$$\|\Delta^k - \Delta\|_n \leq \frac{c_u\delta_n}{8c_\ell},$$

and

$$\mathcal{N} = N_{\rm pr}\left(\frac{c_u\delta_n}{8c_\ell}; \mathcal{C}(\delta_n); \|\cdot\|_n\right).$$

Note that

$$\begin{aligned}\frac{c_u^2\delta_n^2}{c_\ell^2} - \frac{1}{n}\sum_{i=1}^n \|\langle \Delta, X^{(i)}\rangle\|_F^2 &= \left(\frac{c_u^2\delta_n^2}{c_\ell^2} - \frac{1}{n}\sum_{i=1}^n \|\langle \Delta^k, X^{(i)}\rangle\|_F^2\right) \\ &\quad + \left(\frac{1}{n}\sum_{i=1}^n \|\langle \Delta^k, X^{(i)}\rangle\|_F^2 - \frac{1}{n}\sum_{i=1}^n \|\langle \Delta, X^{(i)}\rangle\|_F^2\right).\end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \|\langle \Delta^k, X^{(i)}\rangle\|_{\rm F}^2 - \|\langle \Delta, X^{(i)}\rangle\|_{\rm F}^2 \\ =\ & \frac{1}{n}\sum_{i=1}^n \left\langle \langle \Delta^k - \Delta, X^{(i)}\rangle, \langle \Delta^{(k)} + \Delta, X^{(i)}\rangle \right\rangle \\ \leq\ & \left(\frac{1}{n}\sum_{i=1}^n \|\langle \Delta^k - \Delta, X^{(i)}\rangle\|_{\rm F}^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n \|\langle \Delta^k + \Delta, X^{(i)}\rangle\|_{\rm F}^2\right)^{1/2} \\ =\ & \|\Delta^k - \Delta\|_n \left(\frac{1}{n}\sum_{i=1}^n \|\langle \Delta^k + \Delta, X^{(i)}\rangle\|_{\rm F}^2\right)^{1/2}.\end{aligned}$$

By our choice of covering, $\|\Delta^k - \Delta\|_n \leq c_u\delta_n/8c_\ell$. On the other hand, we have

$$\left(\frac{1}{n}\sum_{i=1}^n \|\langle \Delta^k + \Delta, X^{(i)}\rangle\|_{\rm F}^2\right) \leq \left(2\|\Delta^k\|_n^2 + 2\|\Delta\|_n^2\right)^{1/2} \leq \sqrt{4c_u^2\delta_n^2/c_\ell^2} = 2\frac{c_u\delta_n}{c_\ell}.$$

Overall, we have established the upper bound

$$\frac{1}{n}\|\langle \Delta^k, X^{(i)}\rangle\|_{\rm F}^2 - \|\langle \Delta, X^{(i)}\rangle\|_{\rm F}^2 \leq \frac{c_u^2\delta_n^2}{4c_\ell^2}.$$

Hence we have:

$$Z_n(\mathcal{C}(\delta_n)) \leq \max_{1\leq k\leq \mathcal{N}} \left\{ \frac{c_u^2\delta_n^2}{c_\ell^2} - \frac{1}{n}\sum_{i=1}^{n} \|\langle \Delta^k, X^{(i)}\rangle\|_{\mathrm{F}}^2 \right\} + \frac{c_u^2\delta_n^2}{4c_\ell^2}.$$

Now we use (18) combined with the Hanson-Wright inequality (Hanson and Wright, 1971) to prove that for any $\Delta^{(k)}$ in our covering set,

$$\mathbb{P}\left\{ \frac{c_u^2}{c_\ell^2}\delta_n^2 - \|\Delta^k\|_n^2 > \frac{c_u^2\delta_n^2}{4c_\ell^2} \right\} \leq \exp(-cn),$$

for some constant $c > 0$. Recall that

$$\Sigma = \mathbb{E}[\mathrm{vec}(X)\mathrm{vec}(X)^\top] \in \mathbb{R}^{(nD_M)\times(nD_M)}.$$

Further, recall $[M] = \{1, 2, ..., M\}$ and define an extension of the standard matricization

$$\widetilde{\Delta} := \mathcal{M}_{[M]}(\Delta) \in \mathbb{R}^{D_M \times D_N/D_M}$$

which groups together the first M modes. Further we define the matrix $Q \in \mathbb{R}^{(nD_M)\times(nD_M)}$ such that

$$Q_{r\ell,sm} = \mathbb{I}(r = s)\langle \widetilde{\Delta}_\ell, \widetilde{\Delta}_m \rangle$$

where $1 \leq r, s \leq n$ and $1 \leq \ell, m \leq D_M$ and $\widetilde{\Delta}_\ell, \widetilde{\Delta}_m \in \mathbb{R}^n$. Simple algebra shows that

$$\|\Delta\|_n^2 = \frac{1}{n} Z^\top Q^{1/2}\Sigma Q^{1/2} Z.$$

for some $Z \in \mathbb{R}^{nD_M}$ such that

$$Z \sim \mathcal{N}(0, I_{(nD_M)\times(nD_M)}).$$

Note that

$$\mathbb{E}[\|\Delta\|_n^2] = \frac{1}{n}\mathbb{E}[Z^\top Q^{1/2}\Sigma Q^{1/2} Z] \geq \frac{c_\ell^2}{n}\mathbb{E}[Z^\top Q Z],$$

using (18). Furthermore,

$$\frac{c_\ell^2}{n}\mathbb{E}[Z^\top Q Z] = c_\ell^2 \|\widetilde{\Delta}\|_{\mathrm{F}}^2 = c_u^2\delta_n^2.$$

Now we apply the Hanson-Wright inequality (see, e.g., Hanson and Wright, 1971) to get

$$\mathbb{P}\left\{ \frac{c_u^2}{c_\ell^2}\delta_n^2 - \|\Delta\|_n^2 > \frac{c_u^2}{c_\ell^2}\delta_n^2\zeta \right\} \leq 2\exp\left(-c\min\left\{ \frac{n^2\zeta^2\delta_n^4}{\|Q^{1/2}\Sigma Q^{1/2}\|_{\mathrm{F}}^2}, \frac{n\zeta\delta_n^2}{\|Q^{1/2}\Sigma Q^{1/2}\|_s} \right\} \right).$$

First we upper bound $\|Q^{1/2}\Sigma Q^{1/2}\|_{\mathrm{F}}^2$. If (18) holds, then

$$\|Q^{1/2}\Sigma Q^{1/2}\|_{\mathrm{F}}^2 \leq c_u^2 \|Q\|_{\mathrm{F}}^2.$$

Furthermore,

$$\begin{aligned}
\|Q\|_{\mathrm{F}}^2 &= \sum_{s=1}^{n}\sum_{r=1}^{n}\mathbb{I}(r=s)\sum_{\ell=1}^{D_M}\sum_{m=1}^{D_M}\langle\widetilde{\Delta}_\ell,\widetilde{\Delta}_m\rangle^2 \\
&= \sum_{r=1}^{n}\sum_{\ell=1}^{D_M}\sum_{m=1}^{D_M}\langle\widetilde{\Delta}_\ell,\widetilde{\Delta}_m\rangle^2 \\
&= n\sum_{\ell=1}^{D_M}\sum_{m=1}^{D_M}\langle\widetilde{\Delta}_\ell,\widetilde{\Delta}_m\rangle^2 \\
&\leq n\sum_{\ell=1}^{D_M}\|\widetilde{\Delta}_\ell\|_{\ell_2}^2\sum_{m=1}^{D_M}\|\widetilde{\Delta}_m\|_{\ell_2}^2 \\
&= \frac{c_u^2}{c_\ell^2}n\delta_n^4.
\end{aligned}$$

Thus,

$$\|Q^{1/2}\Sigma Q^{1/2}\|_{\mathrm{F}}^2 \leq \frac{c_u^4}{c_\ell^2}n\delta_n^4.$$

Next we upper bound $\|Q^{1/2}\Sigma Q^{1/2}\|_s$. If (18) holds, then

$$\|Q^{1/2}\Sigma Q^{1/2}\|_s \leq c_u\|Q\|_s.$$

Let $v \in \mathbb{R}^{nD_M}$ such that $\|v\|_{\ell_2}^2 = 1$. Then

$$\begin{aligned}
v^\top Q v &= \sum_{s=1}^{n}\sum_{r=1}^{n}\sum_{\ell=1}^{D_M}\sum_{m=1}^{D_M}v_{r\ell}\langle\widetilde{\Delta}_\ell,\widetilde{\Delta}_m\rangle v_{sm}\mathbb{I}(r=s) \\
&= \sum_{r=1}^{n}\sum_{\ell=1}^{D_M}\sum_{m=1}^{D_M}\langle\widetilde{\Delta}_\ell,\widetilde{\Delta}_m\rangle v_{r\ell}v_{rm} \\
&= \sum_{\ell=1}^{D_M}\sum_{m=1}^{D_M}\langle\widetilde{\Delta}_\ell,\widetilde{\Delta}_m\rangle\sum_{r=1}^{n}v_{r\ell}v_{rm} \\
&\leq \|v\|_{\ell_2}^2\|\Delta\|_{\mathrm{F}}^2 \\
&= \frac{c_u^2}{c_\ell^2}\delta_n^2.
\end{aligned}$$

This implies that

$$\|Q^{1/2}\Sigma Q^{1/2}\|_s \leq \frac{c_u^3}{c_\ell^2}\delta_n^2.$$

Hence, applying the Hanson-Wright inequality yields:

$$\mathbb{P}\left\{\frac{c_u^2}{c_\ell^2}\delta_n^2 - \|\Delta\|_n^2 > \frac{c_u^2}{c_\ell^2}\delta_n^2\zeta\right\} \leq 2\exp\left(-\frac{cc_\ell^2}{c_u^2}\min\{n\zeta^2, n\zeta\}\right).$$

Setting $\zeta = 1/4$ yields

$$\mathbb{P}\left\{\frac{c_u^2}{c_\ell^2}\delta_n^2 - \|\Delta\|_n^2 > \frac{c_u^2}{4c_\ell^2}\delta_n^2\right\} \leq 2\exp\left(-\frac{cc_\ell^2 n}{16c_u^2}\right).$$

Next using the union bound, we have

$$\mathbb{P}\left\{\max_{s=1,2,\ldots,\mathcal{N}}\left\{\frac{c_u^2}{c_\ell^2}\delta_n^2 - \|\Delta^{(s)}\|_n^2\right\} > \frac{\delta_n^2}{4}\right\} \leq \exp\left(\log N_{\mathrm{pr}}\left(\frac{c_u\delta_n}{8c_\ell}, \mathcal{C}(\delta_n), \|\cdot\|_n\right) - cn\right).$$

It remains to bound $\log N_{\mathrm{pr}}(c_u\delta_n/(8c_\ell), \mathcal{C}(\delta_n), \|\cdot\|_n)$. Since the proper covering entropy is upper bounded by the standard covering entropy so that

$$\log N_{\mathrm{pr}}\left(\frac{c_u\delta_n}{8c_\ell}, \mathcal{C}(\delta_n), \|\cdot\|_n\right) \leq \log N\left(\frac{c_u\delta_n}{16c_\ell}, \mathcal{C}(\delta_n), \|\cdot\|_n\right),$$

it suffices to upper bound $\log N(c_u\delta_n/(16c_\ell), \mathcal{C}(\delta_n), \|\cdot\|_n)$. Viewing the samples X as fixed, let us define the zero-mean Gaussian process $\{W_\Delta\}_{\Delta\in\mathcal{B}}$ via

$$W_\Delta = \frac{1}{\sqrt{n}}\sum_{i=1}^n \langle \epsilon^{(i)} \otimes X^{(i)}, \Delta\rangle$$

where $\{\epsilon^{(i)} : i = 1, \ldots, n\}$ are i.i.d. standard Gaussian random variables. By construction, we have

$$\mathrm{var}[(W_\Delta - W_{\Delta'})] = \|\Delta - \Delta'\|_n^2.$$

By the Sudakov minoration (see, e.g., Ledoux and Talagrand, 1991), for all $\eta > 0$ we have

$$\eta\sqrt{\log N(\eta, \mathcal{C}(\delta_n), \|\cdot\|_n)} \leq 4\mathbb{E}_\epsilon\left(\sup_{\Delta\in\mathcal{C}(\delta_n)} W_\Delta\right).$$

Setting $\eta = c_u\delta_n/(16c_\ell)$, we obtain the upper bound:

$$\sqrt{\frac{1}{n}\log N\left(\frac{c_u\delta_n}{16c_\ell}, \mathcal{C}(\delta_n), \|.\|_n\right)} \leq \frac{64c_\ell}{c_u\delta_n}\mathbb{E}_\epsilon\left(\sup_{\Delta\in\mathcal{C}(\delta_n)}\frac{1}{n}\sum_{i=1}^n\langle\epsilon^{(i)}\otimes X^{(i)}, \Delta\rangle\right).$$

The final step is to upper bound the Gaussian complexity

$$\mathbb{E}_\epsilon \left(\sup_{\Delta \in \mathcal{C}(\delta_n)} \frac{1}{n} \sum_{i=1}^n \langle \epsilon^{(i)} \otimes X^{(i)}, \Delta \rangle \right).$$

Clearly,

$$\frac{1}{n} \sum_{i=1}^n \langle \epsilon^{(i)} \otimes X^{(i)}, \Delta \rangle \leq \mathcal{R}^* \left(\frac{1}{n} \sum_{i=1}^n \epsilon^{(i)} \otimes X^{(i)} \right) \mathcal{R}(\Delta) \leq \frac{\lambda}{\eta_{\mathcal{R}}} \mathcal{R}(\Delta).$$

by the definition of λ and our earlier argument. Since $\Delta \in \mathcal{C}(\delta_n)$,

$$\frac{\lambda}{\eta_{\mathcal{R}}} \mathcal{R}(\Delta) \leq \frac{\lambda(1 + 3c_{\mathcal{R}}^{-1})}{\eta_{\mathcal{R}}} \mathcal{R}(\Delta_0) \leq \frac{\lambda(1 + 3c_{\mathcal{R}}^{-1})}{\eta_{\mathcal{R}}} \sqrt{s(\mathcal{A})} \|\Delta_0\|_F \leq \frac{c_u(1 + 3c_{\mathcal{R}}^{-1})}{c_\ell \eta_{\mathcal{R}}} \delta_n \sqrt{s(\mathcal{A})} \lambda.$$

Therefore,

$$\mathbb{E}_\epsilon \left(\sup_{\Delta \in \mathcal{C}(\delta_n)} \frac{1}{n} \sum_{i=1}^n \langle \epsilon^{(i)} \otimes X^{(i)}, \Delta \rangle \right) \leq \frac{c_u(1 + 3c_{\mathcal{R}}^{-1})}{c_\ell \eta_{\mathcal{R}}} \delta_n \sqrt{s(\mathcal{A})} \lambda,$$

and

$$\sqrt{\frac{1}{n} \log N(\frac{c_u \delta_n}{16 c_\ell}, \mathcal{C}(\delta_n), \|.\|_n)} \leq 64 \frac{(1 + 3c_{\mathcal{R}}^{-1})}{\eta_{\mathcal{R}}} \sqrt{s(\mathcal{A})} \lambda.$$

Hence

$$\sqrt{\log N(\frac{c_u \delta_n}{16 c_\ell}, \mathcal{C}(\delta_n), \|.\|_n)} \leq 64 \frac{(1 + 3c_{\mathcal{R}}^{-1})}{\eta_{\mathcal{R}}} \sqrt{n} \sqrt{s(\mathcal{A})} \lambda$$

and

$$\mathbb{P} \left\{ \max_{s=1,2,\ldots,N} \left\{ \frac{c_u^2}{c_\ell^2} \delta_n^2 - \|\Delta^{(s)}\|_n^2 \right\} > \frac{c_u^2 \delta_n^2}{4 c_\ell^2} \right\} \leq \exp(64^2 cns(\mathcal{A})\lambda^2 - cn) \leq \exp(-\tilde{c}n)$$

where the finally inequality holds since $s(\mathcal{A})\lambda^2$ converges to 0 so if we choose n to be sufficiently large. □

Finally we return to the main proof. On the event $\mathcal{E}(\delta_n)$, it now follows easily that,

$$\max\{\|\Delta\|_2^2, \|\Delta\|_n^2\} \leq \frac{\eta_{\mathcal{R}} c_u^2}{c_\ell^2} s(\mathcal{A}) \lambda^2.$$

This completes the proof for Theorem 1.

5.2 Proof of other results in Section 3

In this section we present proofs for the other main results from Section 3, deferring the more technical parts to the appendix.

Proof of Lemmas 2, 3 and 4. We prove these three lemmas together since the proofs follow a very similar argument. First let $S \subset \{1,2,3\}$ denote the directions in which sparsity is applied and $D_S = \prod_{k \in S} d_k$ denote the total dimension in all these directions. For example, in Lemma 2 $S = \{1,2,3\}$ and $D_S = d_1 d_2 d_3$, for Lemma 3, $S = \{2,3\}$ and $D_S = d_2 d_3$ and for Lemma 4, $S = \{1\}$ and $D_S = d_1$. Recall $N = \{1,2,3\}$ and $D_N = d_1 d_2 d_3$.

Note that $\mathcal{R}^*(G)$ can be represented by the variational form:

$$\mathcal{R}^*(G) = \sup_{\|\mathrm{vec}(u)\|_{\ell_1} \le 1, \|v\|_{\mathrm{F}} \le 1} \langle G, u \otimes v \rangle,$$

where $u \in \mathbb{R}^{d_{S_1} \times \ldots \times d_{S_{|S|}}}$ and $v \in \mathbb{R}^{d_{S_1^c} \times \ldots \times d_{S_{N-|S|}^c}}$. Now we express the supremum of this Gaussian process as:

$$\sup_{(u,v) \in V} \mathrm{vec}(u)^\top \mathcal{M}_S(G) \mathrm{vec}(v),$$

where recall $\mathcal{M}_S$ is the matricization involving either slice or fiber S. The remainder of the proof follows from Lemma 15 in Appendix B. □

Proof of Lemma 5. Recall that

$$\mathcal{R}^*(G) := \max_{1 \le j_3 \le d_3} \|G_{..j_3}\|_s .$$

For each $1 \le j_3 \le d_3$, Lemma 16 in Appendix B with $N = 2$ satisfies the concentration inequality

$$\mathbb{E}[\|G_{..j_3}\|_s] \le \sqrt{6(d_1 + d_2)}.$$

Applying standard bounds on the maximum of functions of independent Gaussian random variables,

$$\mathbb{E}[\max_{1 \le j_3 \le d_3} \|G..j_3\|_s] \le \sqrt{6(d_1 + d_2 + \log d_3)}.$$

This completes the proof. □

Proof of Lemma 6. Using the standard nuclear norm upper bound for a matrix in terms of rank and Frobenius norm:

$$\begin{aligned}\mathcal{R}_4^2(A) &= \left(\sum_{j_3=1}^{d_3} \|A_{\cdot\cdot j_3}\|_*\right)^2 \\ &\leq \left(\sum_{j_3=1}^{d_3} \sqrt{\operatorname{rank}(A_{\cdot\cdot j_3})}\|A_{\cdot\cdot j_3}\|_{\mathrm{F}}\right)^2 \\ &\leq \sum_{j_3=1}^{d_3} \operatorname{rank}(A_{\cdot\cdot j_3}) \sum_{j_3=1}^{d_3} \|A_{\cdot\cdot j_3}\|_{\mathrm{F}}^2 = \sum_{j_3=1}^{d_3} \operatorname{rank}(A_{\cdot\cdot j_3})\|A\|_{\mathrm{F}}^2,\end{aligned}$$

where the final inequality follows from the Cauchy-Schwarz inequality. Finally, note that for any $A \in \Theta_4(r)/\{0\}$,

$$\sum_{j_3=1}^{d_3} \operatorname{rank}(A_{\cdot\cdot j_3}) \leq r,$$

which completes the proof. □

Proof of Lemma 7. Note that $\mathcal{R}^*(G) = \|G\|_s$, we can directly apply Lemma 16 with $N = 3$ from Appendix B. □

Proof of Lemma 8. From Tucker decomposition (16), it is clear that for any $A \in \Theta_5(r)$, we can find sets of vectors $\{u_k : k = 1, \ldots, r^2\}$, $\{v_k : k = 1, \ldots, r^2\}$ and $\{w_k : k = 1, \ldots, r^2\}$ such that

$$A = \sum_{k=1}^{r^2} u_k \otimes v_k \otimes w_k,$$

and in addition,

$$u_k^\top u_{k'} = (v_k^\top v_{k'})(w_k^\top w_{k'}) = 0$$

for any $k \neq k'$. It is not hard to see that

$$\|A\|_{\mathrm{F}}^2 = \sum_{k=1}^{r^2} \left(\|u_k\|_{\ell_2}^2 \|v_k\|_{\ell_2}^2 \|w_k\|_{\ell_2}^2\right).$$

On the other hand, as shown by Yuan and Zhang (2014),

$$\|A\|_* = \sum_{k=1}^{r^2} \left(\|u_k\|_{\ell_2}\|v_k\|_{\ell_2}\|w_k\|_{\ell_2}\right).$$

The claim then follows from an application of Cauchy-Schwartz inequality. □

Proof of Lemma 9. Recall that we are considering the regularizer

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

and our goal is to upper bound

$$\mathcal{R}_6^*(G) = 3\max_{1\le k\le 3}\|\mathcal{M}_k(G)\|_s.$$

Once again apply Lemma 16 in Appendix B with $N = 2$ for each matricization implies

$$\mathbb{E}[\mathcal{R}_6^*(G)] \le 4\max(\sqrt{d_1}, \sqrt{d_2}, \sqrt{d_3}).$$

□

Proof of Lemma 10. It is not hard to see that

$$\begin{aligned}
\mathcal{R}_6(A)^2 &= \frac{1}{9}\left(\|\mathcal{M}_1(A)\|_* + \|\mathcal{M}_2(A)\|_* + \|\mathcal{M}_3(A)\|_*\right)^2 \\
&\le \frac{1}{9}(\sqrt{r_1} + \sqrt{r_2} + \sqrt{r_3})^2\|A\|_{\mathrm{F}}^2 \\
&\le \max\{r_1(A), r_2(A), r_3(A)\}\|A\|_{\mathrm{F}}^2,
\end{aligned}$$

which completes the proof. □

5.3 Proof of results in Section 4

In this section we prove the results in Section 4. First we provide a general minimax lower result that we apply to our main results. Let $\mathcal{T} \subset \mathbb{R}^{d_1\times d_2\times\cdots\times d_N}$ be an arbitrary subspace of order-N tensors.

Theorem 6. *Assume that (18) holds and there exists a finite set* $\{A^1, A^2, \ldots, A^m\} \in \mathcal{T}$ *of tensors such that* $\log m \geq 128 n\delta^2$*, such that*

$$nc_u^{-2}\delta^2 \leq \|A^{\ell_1} - A^{\ell_2}\|_{\mathrm{F}}^2 \leq 8nc_u^{-2}\delta^2,$$

for all $\ell_1 \neq \ell_2 \in [m]$ *and all* $\delta > 0$*. Then*

$$\min_{\widetilde{T}} \max_{T \in \mathcal{T}} \|\widetilde{T} - T\|_{\mathrm{F}}^2 \geq cc_u^{-2}\delta^2,$$

with probability at least $1/2$ *for some* $c > 0$*.*

Proof. We use standard information-theoretic techniques developed in Ibragimov and Has'minskii (1981) and extended in Yang and Barron (1999). Let $\{A^1, A^2, \ldots, A^m\}$ be a set such that

$$\|A^{\ell_1} - A^{\ell_2}\|_{\mathrm{F}}^2 \geq nc_u^{-2}\delta^2$$

for all $\ell_1 \neq \ell_2$, and let $\widetilde{m}$ be a random variable uniformly distributed over the index set $[m] = \{1, 2, \ldots, m\}$.

Now we use a standard argument which allows us to provide a minimax lower bound in terms of the probability of error in a multiple hypothesis testing problem (see, e.g., Yang and Barron, 1999; Yu, 1996) then yields the lower bound (write out steps here).

$$\inf_{\widetilde{T}} \sup_{T \in \mathcal{T}} \mathbb{P}\left\{\|\widetilde{T} - T\|_{\mathrm{F}}^2 \geq \frac{c_u^{-2}\delta^2}{2}\right\} \geq \inf_{\widetilde{T}} \mathbb{P}(\widetilde{T} \neq A^{\widetilde{m}})$$

where the infimum is taken over all estimators $\widetilde{T}$ that are measurable functions of X and Y.

Let $X = \{X^{(i)} : i = 1, \ldots, n\}$, $Y = \{Y^{(i)} : i = 1, \ldots, n\}$ and $E = \{\epsilon^{(i)} : i = 1, \ldots, n\}$. Using Fano's inequality (see, e.g., Cover and Thomas, 1991), for any estimator $\widetilde{T}$, we have:

$$\mathbb{P}[\widetilde{T} \neq A^{\widetilde{m}}|X] \geq 1 - \frac{I_X(A^{\widetilde{m}}; Y) + \log 2}{\log m}.$$

Taking expectations over X on both sides, we have

$$\mathbb{P}[\widetilde{T} \neq A^{\widetilde{m}}] \geq 1 - \frac{\mathbb{E}_X[I_X(A^{\widetilde{m}}; Y)] + \log 2}{\log m}.$$

For $\ell = 1, 2, \ldots, m$, let $\mathbb{Q}^\ell$ denote the condition distribution of Y conditioned on X and the event $\{T = A^\ell\}$, and $D_{\mathrm{KL}}(\mathbb{Q}^{\ell_1}||\mathbb{Q}^{\ell_2})$ denote the Kullback-Leibler divergence between $\mathbb{Q}^{\ell_1}$

and $\mathbb{Q}^{\ell_2}$. From the convexity of mutual information (see, e.g., Cover and Thomas, 1991), we have the upper bound

$$I_X(T;Y) \leq \frac{1}{\binom{m}{2}} \sum_{\ell_1,\ell_2=1}^{m} D_{\mathrm{KL}}(\mathbb{Q}^{\ell_1}||\mathbb{Q}^{\ell_2}).$$

Given our linear Gaussian observation model (1),

$$D_{\mathrm{KL}}(\mathbb{Q}^{\ell_1}||\mathbb{Q}^{\ell_2}) = \frac{1}{2}\sum_{i=1}^{n} \left(\langle A^{\ell_1}, X^{(i)}\rangle - \langle A^{\ell_2}, X^{(i)}\rangle\right)^2 = \frac{n\|A^{\ell_1} - A^{\ell_2}\|_n^2}{2}.$$

Further if (18) holds, then

$$\mathbb{E}_X[I_X(T;Y)] \leq \frac{n}{2\binom{m}{2}} \sum_{\ell_1\neq\ell_2} \mathbb{E}_X[\|A^{\ell_1} - A^{\ell_2}\|_n^2] \leq c_u^2 \frac{n}{2\binom{m}{2}} \sum_{\ell_1\neq\ell_2} \|A^{\ell_1} - A^{\ell_2}\|_{\mathrm{F}}^2.$$

Based on our construction, there exists a set $\{A^1, A^2, \ldots, A^m\}$ where each $A^\ell \in \mathcal{T}$ such that $\log m \geq Cn\delta^2$ and

$$c_u^{-1}\delta \leq \|A^{\ell_1} - A^{\ell_2}\|_{\mathrm{F}} \leq 8c_u^{-1}\delta$$

for all $\ell_1 \neq \ell_2 \in \{1, 2, \ldots, m\}$. If (18) holds, then

$$\mathbb{E}_X\left(\|A^{\ell_1} - A^{\ell_2}\|_n^2\right) \leq c_u^2\|A^{\ell_1} - A^{\ell_2}\|_{\mathrm{F}}^2$$

and we can conclude that

$$\mathbb{E}_X[I_X(T;Y)] \leq 32c_u^2 n\delta^2,$$

and from the earlier bound due to Fano's inequality, for and $\delta > 0$ such that

$$\frac{32c_u^2 n\delta^2 + \log 2}{\log m} \leq \frac{1}{2},$$

we are guaranteed that

$$\mathbb{P}\left\{\widetilde{T} \neq A^{\tilde{m}}\right\} \geq \frac{1}{2}.$$

The proof is now completed because $\log m \geq 128n\delta^2$ and $32n\delta^2 \geq \log 2$. □

Proof of Theorem 2. The proof for the upper bound follows directly from Lemma 4 with $d_1 = d_2 = m$ and $d_3 = p$ and noting that the overall covariance $\Sigma \in \mathbb{R}^{(nD_M)\times(nD_M)}$ is block-structured with blocks $\widetilde{\Sigma}$ since each of the samples is independent. Hence

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

To prove the lower bound, we use Theorem 6 and construct a suitable packing set for $\mathcal{T}_1$. The way we construct this packing is to construct two separate packing sets and select the set with the higher packing number using a similar argument to that used in Raskutti et al. (2012) which also uses two separate packing sets. The first packing set we consider involves selecting the s-dimensional slice $A_{..S}$ where $A \subset [j_3]$ and $S = \{1, 2, ..., s\}$. Consider vectorizing each slice so $v = \text{vec}(A_{..S}) \in \mathbb{R}^{sm^2}$. Hence in order to apply Theorem 6, we define the set $\mathcal{T}$ to be slices which is isomorphic to the vector space $\mathbb{R}^{sm^2}$. Using Lemma 17 in Appendix C, there exists a packing set $\{v^1, v^2, ..., v^N\} \in \mathbb{R}^{sm^2}$ such that $\log N \geq csm^2$ and for all v^{ℓ_1}, v^{ℓ_2} where $\ell_1 \neq \ell_2$,

$$\frac{\delta^2}{4} \leq \|v^{\ell_1} - v^{\ell_2}\|_{\text{F}}^2 \leq \delta^2$$

for any $\delta > 0$. If we choose $\delta = c\sqrt{s}m/\sqrt{n}$, then Theorem 6 implies the lower bound

$$\min_{\widetilde{T}} \max_{T \in \mathcal{T}_1} \|\widetilde{T} - T\|_{\text{F}}^2 \geq cc_u^{-2} \frac{sm^2}{n},$$

with probability greater than 1/2.

The second packing set we construct is for the slice $A_{11\cdot} \in \mathbb{R}^p$. Since in the third direction only s of the p co-ordinates are non-zero, the packing number for any slice is analogous to the packing number for s-sparse vectors with ambient dimension p. Letting $v = A_{11\cdot}$, we need to construct a packing set for

$$\{v \in \mathbb{R}^p \mid \|v\|_{\ell_0} \leq s\}.$$

Using Lemma 18 in Appendix C, there exists a discrete set $\{v^1, v^2, ..., v^N\}$ such that $\log N \geq cs \log(p/s)$ for some $c > 0$ and

$$\frac{\delta^2}{8} \leq \|v^k - v^\ell\|_2^2 \leq \delta^2$$

for $k \neq \ell$ for any $\delta > 0$. Setting $\delta^2 = sn^{-1} \log(p/s)$,

$$\min_{\widetilde{T}} \max_{T \in \mathcal{T}_1} \|\widetilde{T} - T\|_{\text{F}}^2 \geq cc_u^{-2} \frac{s \log(p/s)}{n},$$

with probability greater than 1/2.

Taking a maximum over lower bounds involving both packing sets completes the proof of the lower bound in Theorem 2. □

Proof of Theorem 3. The upper bound follows directly from Lemma 5 with $d_1 = d_2 = m$ and $d_3 = p$ and noting that the overall covariance $\Sigma \in \mathbb{R}^{(nD_M)\times(nD_M)}$ is block-structured with blocks $\widetilde{\Sigma}$ since each of the samples is independent.

To prove the lower bound, we use Theorem 6 and construct a suitable packing set for $\mathcal{T}_2$. Once again we construct two separate packings and choose the set that leads to the larger minimax lower bound. For our first packing set, we construct a packing a long once slice. Let us assume $A = (A_{\cdot\cdot 1}, ..., A_{\cdot\cdot p})$, where $\text{rank}(A_{\cdot\cdot 1}) \leq r$ and

$$A_{\cdot\cdot 2} = \cdots = A_{\cdot\cdot p} = 0.$$

If we let $A_{\cdot\cdot 1} = M$ where $M \in \mathbb{R}^{m\times m}$ then $A = (M, 0, .., 0) \in \mathbb{R}^{m\times m\times p}$. Using Lemma 19 in Appendix C, there exists a set $\{A^1, A^2, ..., A^N\}$ such that $\log N \geq crm$ and

$$\frac{\delta^2}{4} \leq \|A^{\ell_1} - A^{\ell_2}\|_{\text{F}}^2 \leq \delta^2$$

for all $\ell_1 \neq \ell_2$ and any $\delta > 0$. Here we set $\delta = \sqrt{rm/n}$. Therefore using Theorem 6

$$\min_{\widetilde{T}} \max_{T\in\mathcal{T}_2} \|\widetilde{T} - T\|_{\text{F}}^2 \geq cc_u^{-2}\frac{rm}{n},$$

with probability greater than 1/2.

The second packing set for $\mathcal{T}_2$ involves a packing in the space of singular values since

$$\sum_{j=1}^{p} \text{rank}(A_{\cdot\cdot j}) \leq r.$$

Let $\{\sigma_{jk} : k = 1, \ldots, m\}$ be the singular values of the matrix $A_{\cdot\cdot j}$. Under our rank constraint, we have

$$\sum_{j=1}^{p}\sum_{k=1}^{m} \mathbb{I}(\sigma_{jk} \neq 0) \leq s.$$

Let $v \in \mathbb{R}^{mp}$ where

$$v = \text{vec}((\sigma_{jk})_{1\leq j\leq p, 1\leq k\leq m}).$$

Note that

$$\sum_{j=1}^{p}\sum_{k=1}^{m} \mathbb{I}(\sigma_{jk} \neq 0) \leq r$$

implies $\|v\|_{\ell_0} \leq r$. Using Lemma 18, there exists a set $\{v^1, v^2, ..., v^N\}$, such that $\log N \geq cr\log(mp/r)$ and for all $\ell_1 \neq \ell_2$,

$$\frac{\delta^2}{4} \leq \|v^{\ell_1} - v^{\ell_2}\|_2^2 \leq \delta^2$$

for any $\delta > 0$. If we set $\delta^2 = rn^{-1}\log(mp/r)$. Therefore using Theorem 6,

$$\min_{\widetilde{T}} \max_{T \in \mathcal{T}_2} \|\widetilde{T} - T\|_{\mathrm{F}}^2 \geq cc_u^{-2} \frac{r\log(mp/r)}{n},$$

with probability greater than 1/2. Hence taking a maximum over both bounds,

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

with probability greater than 1/2. □

Proof of Theorem 4. The upper bound with

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

follows directly from Lemma 3 with $d_1 = p$ and $d_2 = d_3 = m$ and (18) is satisfied with $c_u^2 = 1/\mu_{\min}$ and $c_\ell^2 = 1/\mu_{\max}$ according to (36).

To prove the lower bound is similar to the proof for the lower bound in Theorem 2. Once again we use Theorem 6 and construct a two suitable packing sets for $\mathcal{T}_3$. The first packing set we consider involves selecting an arbitrary subspace

$$\widetilde{\mathcal{T}} := \{A = (A_{j_1,j_2,j_3})_{j_1,j_2,j_3} \mid 1 \leq j_1 \leq \sqrt{s},\ 1 \leq j_2 \leq \sqrt{s},\ 1 \leq j_3 \leq p\}.$$

Now if we let $v = \mathrm{vec}(A)$, then v comes from an sp-dimensional vector space for any $A \in \widetilde{\mathcal{T}}$. Using Lemma 17 in Appendix C, there exists a packing set $\{v^1, v^2, ..., v^N\} \in \mathbb{R}^{sp}$ such that $\log N \geq csp$ and for all v^{ℓ_1}, v^{ℓ_2} where $\ell_1 \neq \ell_2$,

$$\frac{\delta^2}{4} \leq \|v^{\ell_1} - v^{\ell_2}\|_{\mathrm{F}}^2 \leq \delta^2$$

for any $\delta > 0$. If we choose $\delta = \sqrt{sp/n}$, then Theorem 6 implies the lower bound

$$\min_{\widetilde{T}} \max_{T \in \mathcal{T}_3} \|\widetilde{T} - T\|_{\mathrm{F}}^2 \geq cc_u^{-2} \frac{sp}{n},$$

with probability greater than 1/2. Further $c_u^2 = 1/\mu_{\min}$.

For the second packing set we construct is for the slice A_{1,j_2,j_3} for any $1 \leq j_2, j_3 \leq m$. Since in the second and third direction only s of the co-ordinates are non-zero, we consider the vector space

$$\{v \in \mathbb{R}^{m^2} \mid \|v\|_{\ell_0} \leq s\}.$$

Once again using the standard standard hypercube construction in Lemma 18 in Appendix C, there exists a discrete set $\{v^1, v^2, ..., v^N\}$ such that $\log N \geq cs\log(m^2/s)$ for some $c > 0$ and

$$\frac{\delta^2}{8} \leq \|v^{\ell_1} - v^{\ell_2}\|_2^2 \leq \delta^2$$

for $\ell_1 \neq \ell_2$ for any $\delta > 0$. Setting $\delta = sn^{-1}\log(m^2/s)$ yields

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

with probability greater than 1/2. Taking a maximum over lower bounds involving both packing sets completes the proof of of our lower bound. □

Proof of Theorem 5. The upper bound follows from a slight modification of the statement in Lemma 5. In particular since $\mathcal{R}(A) = \|A^{(12)}\|_* + \|A^{(13)}\|_* + \|A^{(23)}\|_*$, the dual norm is

$$\mathcal{R}^*(A) = \max_{1 \leq k_1 < k_2 \leq 3} \|A^{(k_1 k_2)}\|_s. \tag{38}$$

Hence, following the same technique as used in Lemma 5

$$\mathbb{E}[\mathcal{R}^*(G)] \leq c \max_{1 \leq k_1 < k_2 \leq 3} \sqrt{\frac{\max\{d_{k_1}, d_{k_2}\}}{n}} = c\sqrt{\frac{\max\{d_1, d_2, d_3\}}{n}}. \tag{39}$$

It is also straightforward to see that $s(\mathcal{T}_4) \leq r$.

To prove the lower bound, we construct three packing sets and select the one with the largest packing number. Recall that

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

Therefore our three packings are for $A^{(12)} \in \mathbb{R}^{d_1 \times d_2}$, $A^{(13)} \in \mathbb{R}^{d_1 \times d_3}$, and $A^{(23)} \in \mathbb{R}^{d_2 \times d_3}$ assuming each has rank r. We focus on packing in $A^{(12)} \in \mathbb{R}^{d_1 \times d_2}$ since the approach is similar in the other two cases. Using Lemma 16 from Appendix B in combination with Theorem 6,

$$\min_{\widetilde{T}} \max_{T \in \mathcal{T}_4} \|\widetilde{T} - T\|_{\mathrm{F}}^2 \geq c c_u^{-2} \frac{r \min\{d_1, d_2\}}{n},$$

with probability greater than $1/2$. Repeating this process for packings in $A^{(13)} \in \mathbb{R}^{d_1 \times d_3}$, and $A^{(23)} \in \mathbb{R}^{d_2 \times d_3}$ assuming each has rank r and taking a maximum over all three bounds yields the overall minimax lower bound

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

with probability greater than $1/2$. □

References

A. Agarwal, S. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 27(2):1171–1197, 2012.

T. W. Anderson. The integral of a symmetric convex set and some probability inequalities. *Proc. of American Mathematical Society*, 6:170–176, 1955.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1984.

S. Basu and G. Michailidis. Regularized estimation in sparse high-dimensional time series models. *Annals of Statistics*, 43(4):1535–1567, 2015.

R. Bhatia. *Matrix Analysis*. Springer, New York, 1997.

P. Buhlmann and S. van de Geer. *Statistical for High-Dimensional Data*. Springer Series in Statistics. Springer, New York, 2011.

S. Chen, M. R. Lyu, I. King, and Z. Xu. Exact and stable recovery of pairwise interaction tensors. In *Advances in Neural Information Processing Systems*, 2013.

S. Cohen and M. Collins. Tensor decomposition for fast parsing with latent-variable pcfgs. In *Advances in Neural Information Processing Systems*, 2012.

T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

S. Gandy, B. Recht, and I. Yamada. Tensor completion and low-n rank tensor recovery via convex optimization. *Inverse Problems*, 27, 2011.

Y. Gordon. On milmans inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. *Geometric aspects of functional analysis, Israel Seminar 1986-87, Lecture Notes*, 1317:84–106, 1988.

D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Monographs on Statistics and Applied Probability 143. CRC Press, New York, 2015.

P. Hoff. Multilinear tensor regression for longitudinal relational data. Technical Report TR-2003-08, Department of Statistics, University of Washington, 2003.

I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York, 1981.

T. G. Koldar and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51: 455–500, 2009.

M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

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

H. Lütkepolhl. *New introduction to multiple time series analysis*. Springer, New York, 2006.

P. Massart. *Concentration Inequalties and Model Selection*. Ecole d'Eté de Probabilités, Saint-Flour. Springer, New York, 2003.

N. Mesgarani, M. Slaney, and S. Shamma. Content-based audio classification based on multiscale spectro-temporal features. *IEEE Transactions on Speech and Audio Processing*, 14:920–930, 2006.

C. Mu, B. Huang, J. Wright, and D. Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *International Conference on Machine Learning*, 2014.

S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39:1069–1097, 2011.

S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, UK, 1989.

G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue conditions for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions of Information Theory*, 57(10):6976–6994, 2011.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13: 398–427, 2012.

S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *ICDM*, 2010.

S. Rendle, L. B. Marinho, A. Nanopoulos, and L. Schmidt-Thieme. Learning optimal ranking with tensor factorization for tag recommendation. In *SIGKDD*, 2009.

G. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

O. Semerci, N. Hao, M. Kilmer, and E. Miller. Tensor based formulation and nuclear norm regularizatin for multienergy computed tomography. *IEEE Transactions on Image Processing*, 23:1678–1693, 2014.

N.D. Sidiropoulos and N. Nion. Tensor algebra and multi-dimensional harmonic retrieval in signal processing for mimo radar. *IEEE Transactions on Signal Processing*, 58:5693–5705, 2010.

D. Slepian. The one-sided barrier problem for gaussian noise. *Bell System Tech. J*, 41: 463–501, 1962.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

B. Turlach, W.N. Venables, and S.J. Wright. Simultaneous variable selection. *Technometrics*, 27:349–363, 2005.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

B. Yu. Assouad, Fano and Le Cam. *Research Papers in Probability and Statistics: Festschrift in Honor of Lucien Le Cam*, pages 423–435, 1996.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 68(1):49–67, 2006.

M. Yuan and C-H. Zhang. On tensor completion via nuclear norm minimization. *Foundation of Computational Mathematics*, to appear, 2014.

A Results for Gaussian random variables

In this section we provide some standard concentration bounds that we use throughout this paper. First, we provide the definition for sub-Gaussian random variables. A zero-mean random variable X is sub-Gaussian if there is a positive number σ such that,

$$\mathbb{E}[e^{\gamma X}] \leq e^{\sigma^2\gamma^2/2}.$$

For example, if $X \sim \mathcal{N}(0, \sigma^2)$ is a sub-Gaussian random variable with parameter σ.

For quadratic forms involving independent sub-Gaussian random variables, we have the useful Hanson-Wright inequality.

Theorem 7 (Hanson-Wright inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be independent zero-mean sub-Gaussian random variables with sub-Gaussian parameter upper bounded σ. Further let A be an $n \times n$ matrix. Then for every $t \geq 0$ there exists a constant $c > 0$,*

$$\mathbb{P}\left\{|X^\top AX - \mathbb{E}(X^\top AX)| > t\right\} \leq 2\exp\left(-c\min\left\{\frac{t^2}{\sigma^4\|A\|_{\mathrm{F}}^2}, \frac{t}{\sigma^2\|A\|_s}\right\}\right),$$

where $X = (X_1, \ldots, X_n)^\top$.

A.1 Gaussian comparison inequalities

The first result is a classical result from Anderson (1955).

Lemma 13 (Anderson's comparison inequality)**.** *Let X and Y be zero-mean Gaussian random vectors with covariance Σ_X and Σ_Y respectively. If $\Sigma_X - \Sigma_Y$ is positive semi-definite then for any convex symmetric set C,*

$$\mathbb{P}(X \in C) \leq \mathbb{P}(Y \in C).$$

The following Lemma is Slepian's inequality Slepian (1962) which allows to upper bound the supremum of one Gaussian process by the supremum of another Gaussian process.

Lemma 14 (Slepian's Lemma)**.** *Let $\{G_s, s \in S\}$ and $\{H_s, s \in S\}$ be two centered Gaussian processes defined over the same index set S. Suppose that both processes are almost surely*

bounded. For each $s, t \in S$, if $\mathbb{E}(G_s - G_t)^2 \leq \mathbb{E}(H_s - H_t)^2$, then $\mathbb{E}[\sup_{s\in S} G_s] \leq \mathbb{E}[\sup_{s\in S} H_s]$. Further if $\mathbb{E}(G_s^2) = \mathbb{E}(H_s^2)$ for all $s \in S$, then

$$\mathbb{P}\left\{\sup_{s\in S} G_s > x\right\} \leq \mathbb{P}\left\{\sup_{s\in S} H_s > x\right\},$$

for all $x > 0$.

Finally, we require a standard result on the concentration of Lipschitz functions over Gaussian random variables.

Theorem 8 (Theorem 3.8 from Massart (2003))**.** *Let $g \sim \mathcal{N}(0, I_{d\times d})$ be a d-dimensional Gaussian random variable. Then for any function $F : \mathbb{R}^d \to \mathbb{R}$ such that $|F(x) - F(y)| \leq L\|x - y\|_{\ell_2}$ for all $x, y \in \mathbb{R}^d$, we have*

$$\mathbb{P}\big[|F(g) - \mathbb{E}[F(g)]| \geq t\big] \leq 2\exp\left(-\frac{t^2}{2L^2}\right),$$

for all $t > 0$.

B Suprema for i.i.d. Gaussian tensors

In this section we provide important results on suprema of i.i.d. Gaussian tensors over different sets.

B.1 The group ℓ_2-ℓ_∞ norm

Let $G \in \mathbb{R}^{d_1\times d_2}$ be an i.i.d. Gaussian matrix and define the set

$$V := \{(u, v) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \mid \|u\|_{\ell_2} \leq 1\ , \|v\|_{\ell_1} \leq 1\}.$$

Using this notation, let us define the define the random quantity:

$$M(G, V) := \sup_{(u,v)\in V} u^\top G v.$$

Then we have the following overall bound.

Lemma 15.

$$\mathbb{E}[M(G, V)] \leq 3(\sqrt{d_1} + \sqrt{\log d_2}).$$

Proof. Our proof user similar ideas to the proof of Theorem 1 in Raskutti et al. (2010). We need to upper bound $\mathbb{E}[M(G,V)]$. We are taking the supremum of the Gaussian process

$$\sup_{\|u\|_{\ell_2}\leq 1,\, \|v\|_{\ell_2}\leq 1} u^\top G v.$$

We now construct a second Gaussian process $\widetilde{G}_{u,v}$ over the set V and apply Slepian's inequality (see Lemma 14 in Appendix A.1) to upper bound

$$\sup_{\|u\|_{\ell_2}\leq 1,\, \|v\|_{\ell_2}\leq 1} u^\top G v$$

by the supremum over our second Gaussian process. $\widetilde{G}_{u,v}$. In particular, let us define the process as:

$$\widetilde{G}_{u,v} = g^\top u + h^\top v,$$

where the vectors $(g,h) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ are i.i.d. standard normals (also independent of each other). It is straightforward to show that both $u^\top G v$ and $g^\top u + h^\top v$ are zero-mean. Further it is straightforward to show that

$$\mathrm{Var}(\widetilde{G}_{u,v} - \widetilde{G}_{u',v'}) = \|u-u'\|_{\ell_2}^2 + \|v-v'\|_{\ell_2}^2.$$

Now we show that

$$\mathrm{Var}(u^\top G v - u'^\top G v') \leq \|u-u'\|_{\ell_2}^2 + \|v-v'\|_{\ell_2}^2.$$

To this end, observe that

$$\begin{aligned}
\mathrm{Var}(u^\top G v - u'^\top G v') &= \|uv^\top - u'v'^\top\|_{\mathrm{F}}^2 \\
&= \|(u-u')v^\top + u'(v-v')^\top\|_{\mathrm{F}}^2 \\
&= \|v\|_{\ell_2}^2\|u-u'\|_{\ell_2^2} + \|u'\|_{\ell_2}^2\|v-v'\|_{\ell_2}^2 \\
&\quad +2(u^\top u' - \|u'\|_{\ell_2}\|u\|_{\ell_2})(v^\top v' - \|v'\|_{\ell_2}\|v\|_{\ell_2}).
\end{aligned}$$

First note that $\|v\|_{\ell_2}^2 \leq \|v\|_{\ell_1}^2 \leq 1$ for all $v \in V$ and $\|u'\|_{\ell_2}^2 \leq 1$. By the Cauchy-Schwarz inequality, $v^\top v' - \|v'\|_{\ell_2}\|v\|_{\ell_2} \leq 0$ and $u^\top u' - \|u'\|_{\ell_2}\|u\|_{\ell_2} \leq 0$. Therefore

$$\mathrm{Var}(u^\top G v - u'^\top G v') \leq \|u-u'\|_{\ell_2^2} + \|v-v'\|_{\ell_2}^2.$$

Consequently using Lemma 14

$$\mathbb{E}[M(G,V)] \leq \mathbb{E}[\sup_{\|u\|_{\ell_2}\leq 1} g^\top u + \sup_{\|v\|_{\ell_1}\leq 1} h^\top v].$$

Therefore:

$$\begin{aligned}\mathbb{E}[M(G,V)] &\leq \mathbb{E}[\sup_{\|u\|_{\ell_2}\leq 1} g^\top u + \sup_{\|v\|_{\ell_1}\leq 1} h^\top v] \\ &= \mathbb{E}[\sup_{\|u\|_{\ell_2}\leq 1} g^\top u] + \mathbb{E}[\sup_{\|v\|_{\ell_1}\leq 1} h^\top v] \\ &= \mathbb{E}[\|g\|_{\ell_2}] + \mathbb{E}[\|h\|_{\ell_\infty}].\end{aligned}$$

By known results on Gaussian maxima (see e.g. Ledoux and Talagrand, 1991),

$$\mathbb{E}[\|h\|_{\ell_\infty}] \leq 3\sqrt{\log d_2}$$

and

$$\mathbb{E}[\|g\|_{\ell_2}] \leq \sqrt{d_1} + o(\sqrt{d_1}) \leq \frac{3}{2}\sqrt{D_j}.$$

Therefore

$$\mathbb{E}[M(G,V)] \leq \frac{3}{2}\sqrt{d_1} + 3\sqrt{\log d_2}.$$

□

B.2 Spectral norm of tensors

Our proof is based on an extension of the proof techniques used for the proof of Proposition 1 in Negahban and Wainwright (2011).

Lemma 16. *Let $G \in \mathbb{R}^{d_1\times d_2\times\cdots\times d_N}$ be a random sample from an i.i.d. Gaussian tensor ensemble. Then we have*

$$\mathbb{E}[\|G\|_s] \leq 4\log(4N)\sum_{k=1}^{N}\sqrt{d_k}.$$

Proof. Recall the definition of $\|G\|_s$:

$$\|G\|_s = \sup_{(u_1,u_2,\ldots,u_N)\in S^{d_1-1}\times S^{d_2-1}\times\cdots\times S^{d_N-1}} \langle u_1\otimes u_2\otimes\cdots\otimes u_N, G\rangle.$$

Since each entry $\langle u_1 \otimes u_2 \otimes \cdots \otimes u_N, G\rangle$ is a zero-mean Gaussian random variable, $\|G\|_s$ is the supremum of a Gaussian process and therefore the concentration bound follows from Theorem 7.1 in Ledoux Ledoux (2001).

We use a standard covering argument to upper bound $\mathbb{E}[\|G\|_s]$. Let $\{u_1^1, u_1^2, \ldots, u_1^{M_1}\}$ be a $1/2N$ covering number of the sphere S^{d_1-1} in terms of vector ℓ_2-norm. Similarly for all $2 \leq k \leq N$, let $\{u_k^1, u_k^2, \ldots, u_k^{M_k}\}$ be a $1/2N$ covering number of the sphere S^{d_k-1}. Therefore

$$\begin{aligned} & \langle u_1 \otimes u_2 \otimes \cdots \otimes u_{N-1} \otimes u_N, G\rangle \\ \leq \; & \langle u_1 \otimes u_2 \otimes \cdots \otimes u_{N-1} \otimes u_N^j, G\rangle + \langle u_1 \otimes u_2 \otimes \cdots \otimes u_{N-1} \otimes (u_N - u_N^j), G\rangle. \end{aligned}$$

Taking a supremum over both sides,

$$\|G\|_s \leq \max_{j=1,\ldots,M_N} \langle u_1 \otimes u_2 \otimes \cdots \otimes u_{N-1} \otimes u_N^j, G\rangle + \frac{1}{2N}\|G\|_s.$$

Repeating this argument over all N directions,

$$\|G\|_s \leq 2 \max_{j_1=1,2,\ldots,M_1,\ldots,j_N=1,2,\ldots,M_N} \langle u_1^{j_1} \otimes u_2^{j_2} \otimes \cdots \otimes u_N^{j_N}, G\rangle.$$

By construction, each variable $\langle u_1^{j_1} \otimes u_2^{j_2} \otimes \cdots \otimes u_N^{j_N}, G\rangle$ is a zero-mean Gaussian with variance at most 1, so by standard bounds on Gaussian maxima,

$$\mathbb{E}[\|G\|_s] \leq 4\sqrt{\log(M_1 \times M_2 \times \cdots \times M_N)} \leq 4[\sqrt{\log M_1} + \cdots + \sqrt{\log M_N}].$$

There exist a $1/2N$-coverings of S^{d_k-1} with $\log M_k \leq d_k \log(4N)$ which completes the proof. □

C Hypercube packing sets

In this section, we provide important results for the lower bound results. One key concept is the so-called *Hamming distance.*The Hamming distance is between two vectors $v \in \mathbb{R}^d$ and $v' \in \mathbb{R}^d$ is defined by:

$$d_H(v, v') = \sum_{j=1}^{d} \mathbb{I}(v_j \neq v'_j).$$

Lemma 17. *Let $\mathcal{C} = [-1, +1]^d$ where $d \geq 6$. Then there exists a discrete subset $\{v^1, v^2, ..., v^m\} \subset \mathcal{C}$, such that $\log m \geq cd$ for some constant $c > 0$, and for all $\ell_1 \neq \ell_2$,*

$$\frac{\delta^2}{4} \leq \|v^{\ell_1} - v^{\ell_2}\|_{\ell_2}^2 \leq \delta^2,$$

for any $\delta > 0$.

Proof. Let

$$v^\ell \in \left\{-\frac{\delta}{\sqrt{d}}, \frac{\delta}{\sqrt{d}}\right\}^d,$$

i.e. a member of the d-dimensional hypercube re-scaled by $\sqrt{3}\delta/(2\sqrt{d})$. Recall the definition of Hamming distance provided above. In this case amounts to the places either v_j or v'_j is negative, but both or not negative. Then according to Lemma 4 in Yu (1996), there exists a subset re-scaled of this hypercube $v^1, v^2, ..., v^m$, such that

$$d_H(v^{\ell_1}, v^{\ell_2}) \geq \frac{d}{3}$$

and $\log m \geq cd$. Clearly,

$$\|v^{\ell_1} - v^{\ell_2}\|_{\ell_2}^2 = \frac{3\delta^2}{4d} d_H(v^{\ell_1}, v^{\ell_2}) \geq \frac{\delta^2}{4}.$$

Further,

$$\|v^{\ell_1} - v^{\ell_2}\|_{\ell_2}^2 \leq \frac{3\delta^2}{4d} \times d \leq \frac{3\delta^2}{4} \leq \delta^2.$$

This completes the proof. □

Next we provide a hupercube packing set for the sparse subset of vectors. That is the set

$$V := \{v \in \mathbb{R}^d \mid \|v\|_{\ell_0} \leq s\}.$$

This follows from Lemma 4 in Raskutti et al. (2011) which we state here for completeness.

Lemma 18. *Let $\mathcal{C} = [-1, +1]^d$ where $d \geq 6$. Then there exists a discrete subset $\{v^1, v^2, ..., v^m\} \subset V \cap \mathcal{C}$, such that $\log m \geq cs \log(d/s)$ for some $c > 0$, and for all $\ell_1 \neq \ell_2$,*

$$\frac{\delta^2}{8} \leq \|v^{\ell_1} - v^{\ell_2}\|_{\ell_2}^2 \leq \delta^2,$$

for any $\delta > 0$.

Finally we present a packing set result from Lemma 6 in Agarwal et al. (2012) that packs into the set of rank-r $d_1 \times d_2$ matrices.

Lemma 19. *Let* $\min\{d_1, d_2\} \geq 10$*, and let* $\delta > 0$*. Then for each* $1 \leq r \leq \min\{d_1, d_2\}$*, there exists a set of* $d_1 \times d_2$ *matrices* $\{A^1, A^2, ..., A^m\}$ *with rank-*r *with cardinality* $\log m \geq cr \min\{d_1, d_2\}$ *for some constant* $c > 0$ *such that*

$$\frac{\delta^2}{4} \leq \|A^{\ell_1} - A^{\ell_2}\|_{\mathrm{F}}^2 \leq \delta^2,$$

for all $\ell_1 \neq \ell_2$*.*