Journal of Machine Learning Research 20 (2019) 1-37 Submitted 12/16; Revised 1/18; Published 2/19

Non-Convex Projected Gradient Descent for Generalized Low-Rank Tensor Regression

Han Chen HAN.CHEN@WISC.EDU

Garvesh Raskutti RASKUTTI@STAT.WISC.EDU
Department of Statistics
University of Wisconsin–Madison
Madison, WI 53706, USA

Ming Yuan MING.YUAN@COLUMBIA.EDU
Department of Statistics
Columbia University
New York, NY 10027, USA

Editor: Sujay Sanghavi

Abstract

In this paper, we consider the problem of learning high-dimensional tensor regression problems with low-rank structure. One of the core challenges associated with learning high-dimensional models is computation since the underlying optimization problems are often non-convex. While convex relaxations could lead to polynomial-time algorithms they are often slow in practice. On the other hand, limited theoretical guarantees exist for non-convex methods. In this paper we provide a general framework that provides theoretical guarantees for learning high-dimensional tensor regression models under different low-rank structural assumptions using the projected gradient descent algorithm applied to a potentially non-convex constraint set Θ in terms of its *localized Gaussian width* (due to Gaussian design). We juxtapose our theoretical results for non-convex projected gradient descent algorithms with previous results on regularized convex approaches. The two main differences between the convex and non-convex approach are: (i) from a computational perspective whether the non-convex projection operator is computable and whether the projection has desirable contraction properties and (ii) from a statistical error bound perspective, the non-convex approach has a superior rate for a number of examples. We provide three concrete examples of low-dimensional structure which address these issues and explain the pros and cons for the non-convex and convex approaches. We supplement our theoretical results with simulations which show that, under several common settings of generalized low rank tensor regression, the projected gradient descent approach is superior both in terms of statistical error and run-time provided the step-sizes of the projected descent algorithm are suitably chosen.

Keywords: tensors, non-convex optimization, high-dimensional regression, low-rank;

1. Introduction

Parameter estimation in high-dimensional regression has received substantial interest over the past couple of decades. See, e.g., Buhlmann and van de Geer (2011); Hastie et al. (2015). One of the more recent advances in this field is the study of problems where the parameters and/or data take the form of a multi-way array or *tensor*. Such problems arise in many practical settings (see, e.g., Cohen and Collins, 2012; Li and Li, 2010; Semerci et al., 2014; Sidiropoulos and Nion, 2010) and present a number of additional challenges that do not arise in the vector or matrix setting. In particular, one of the challenges associated with high-dimensional tensor regression models is how to define low-dimensional structure since the notion of rank is ambiguous for tensors (see, e.g., Koldar and Bader, 2009). Different approaches on how to impose low-rank and sparsity structure that lead to implementable algorithms have been considered. See, e.g., Gandy et al. (2011); Mu et al. (2014); Raskutti and Yuan (2015); Tomioka et al. (2013); Yuan and Zhang (2014), and references therein. All of the previously mentioned approaches have relied on penalized convex relaxation schemes and in particular, many of these different approaches have been encompassed by Raskutti and Yuan (2015). The current work complements these earlier developments by studying the non-convex projected gradient descent (PGD) approaches to generalized low-rank tensor regression.

While convex approaches are popular since greater theoretical guarantees have been provided for them, non-convex approaches have gained popularity as recently more theoretical guarantees have been provided for specific high-dimensional settings. See, e.g., Fan and Li (2001); Jain et al. (2014, 2016); Loh and Wainwright (2015). Furthermore, even though non-convex problems do not in general lead to polynomial-time computable methods, they often work well in practice. In particular, inspired by the recent work of Jain et al. (2014, 2016) who demonstrated the effectiveness of non-convex projected gradient descent approaches for high-dimensional linear regression and matrix regression, we consider applying similar techniques to high-dimensional low-rank tensor regression problems with a generalized linear model loss function.

Low-rankness in higher order tensors may occur in a variety of ways (see e.g. Koldar and Bader (2009) for examples). To accommodate these different notions of low-rankness, we develop a general framework which provides theoretical guarantees for projected gradient descent algorithms applied to tensors residing in general low-dimensional subspaces. Our framework relies on two properties ubiquitous in low-rank tensor regression problems: (i) that the parameter space is a member of a class of subspaces super-additive when indexed over a partially ordered set; and (ii) there exists an approximate projection onto each subspace satisfying a contractive property. Assuming that the coefficient tensor lies in a low-dimensional subspace Θ satisfying these properties, we establish general risk bounds for non-convex projected gradient descent based methods applied to a generalized tensor regression model with covariates with Gaussian design. By developing this general framework in terms of these properties, we only need to verify that our set satisfies these properties and also demonstrate that these properties are fundamental to providing theoretical results for projected gradient descent.

Our main theoretical result shows that the Frobenius norm scales as $n^{-1/2}w_G[\Theta \cap \mathbb{B}_{\mathrm{F}}(1)]$, where n is the sample size, $\mathbb{B}_{\mathrm{F}}(1) := \{A : \|A\|_F \leq 1\}$ refers to the Frobenius-norm ball with radius 1 and $w_G[\Theta \cap \mathbb{B}_{\mathrm{F}}(1)]$ refers to the *localized Gaussian width* of Θ. While statistical rates in terms of Gaussian widths are already established for convex regularization approaches (see, e.g., Chandrasekaran et al., 2012; Raskutti and Yuan, 2015), this to the best of our knowledge is the first general error bound for non-convex projected gradient descent in terms of a localized Gaussian width.

Another major contribution we make is to provide a comparison both in terms of statistical error rate and computation to existing convex approaches to low rank tensor regression. Using our statistical error bound for non-convex projected gradient descent which is stated in terms of the localized Gaussian width of Θ, we show explicitly that our error bound for the non-convex approach is no larger (up to a constant) than those for convex regularization schemes (see, e.g., Theorem 1 of Raskutti and Yuan, 2015), and in some cases is smaller. To make this comparison more concrete, we focus on three particular examples of low-rank tensor structure: (i) sum of ranks of each slice of a tensor being small; (ii) sparsity and low-rank structure for slices; and (iii) low Tucker rank. In case (i), both approaches are applicable and achieve the same rate of convergence. For case (ii), the non-convex approach is still applicable whereas a convex regularization approach is not naturally applicable. In case (iii) again both approaches are applicable but a superior statistical performance can be achieved via the non-convex method. We supplement our theoretical comparison with a simulation comparison. Our simulation results show that our non-convex projected gradient descent based approach compares favorably to the convex regularization approach using a generic `cvx` solver in terms of both run-time and statistical performance provided optimal step-size choices in the projected gradient descent and regularization parameters in the convex regularization approach are used. Furthermore the projected gradient descent scales to much larger-scale data than generic convex solvers.

1.1. Our contributions

To summarize, we make the following three contributions in our paper:

- Firstly, we provide general error bounds for projected gradient descent applied to generalized tensor regression problems in terms of the localized Gaussian width of the constraint set Θ. In particular, we provide three novel results. Theorem 1 provides an upper bound for projected gradient descent with tensor parameters and applies to any Θ satisfying the super-additivity and contractive properties described above and explained in greater detail in Section 3. Theorem 1 substantially generalizes prior results in Jain et al. (2014, 2016) which focus on sparse vectors and low-rank matrices. Theorem 2 and 3 apply the general result in Theorem 1 to generalized linear models and Gaussian linear models respectively. Significantly, Theorems 2 and 3 show that the localized Gaussian width for the constraint set for Θ play a crucial role in the mean-squared error bound. This is the first analysis we are aware of that expresses the statistical error of PGD in terms of localized Gaussian width which allows us to deal with PGD in a more unified manner.

- Using Theorem 3, our second major contribution is to provide a comparison in terms of mean-squared error to standard convex regularization schemes studied in Raskutti and Yuan (2015). We show using the comparison of Gaussian widths in the convex and non-convex cases that unlike for vector and matrix problems where convex regularization schemes provably achieve the same statistical error bounds as non-convex approaches, the more complex structure of tensors means that our non-convex approach could yield a superior statistical error bound in some examples compared to convex regularization schemes. We also prove that our non-convex error bound is no larger than the convex regularization bound developed in Raskutti and Yuan (2015).

- Lastly, we demonstrate in Section 5 the benefits of the non-convex approach compared to existing convex regularization schemes for various low-rank tensor regression problems. We also show through simulations the benefit of using low-rank tensor regularization schemes compared to using a low-rank matrix scheme.

The remainder of the paper is organized as follows: Section 2 introduces the basics of the low-rank tensor regression models we consider and introduces the projected gradient descent algorithm. Section 3 presents the general theoretical results for non-convex projected gradient descent and specific examples are discussed in Section 4. A simulation comparison between the convex and non-convex approach is provided in Section 5 and proofs are provided in Section 7.

2. Methodology

Consider a generalized tensor regression framework where the conditional distribution of a scalar response Y given a covariate tensor $X \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ is given by

$$p(Y|X,T) = h(Y)\exp\left\{Y\langle X,T\rangle - a(\langle X,T\rangle)\right\}, \tag{1}$$

where $a(\cdot)$ is a strictly convex and differentiable log-partition function, $h(\cdot)$ is a nuisance parameter, and $T \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ is the parameter tensor of interest. Typical examples of $a(\cdot)$ include $a(\theta) = \frac{1}{2}\theta^2$ leading to the usual normal linear regression, $a(\theta) = \log(1+e^\theta)$ corresponding to logistic regression, and $a(\theta) = e^\theta$ which can be identified with Poisson regression where θ is a scalar. The goal is to estimate tensor T based on the training data $\{(X^{(i)}, Y^{(i)}) : 1 \le i \le n\}$. For convenience we assume $(X^{(i)}, Y^{(i)})$'s are independent copies of (X, Y). Hence the negative log-likelihood risk objective, for any $A \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$, is:

$$\mathcal{L}(A) = \frac{1}{n}\sum_{i=1}^{n}\left[a(\langle X^{(i)}, A\rangle) - Y^{(i)}\langle X^{(i)}, A\rangle - \log h(Y^{(i)})\right]. \tag{2}$$

The notation $\langle \cdot, \cdot \rangle$ will refer throughout this paper to the standard inner product taken over appropriate Euclidean spaces. Hence, for $A \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ and $B \in \mathbb{R}^{d_1 \times \cdots \times d_N}$:

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

Using the standard notion of inner product, for a tensor A, $\|A\|_{\mathrm{F}} = \langle A, A\rangle^{1/2}$. And the empirical norm $\|\cdot\|_n$ for a tensor $A \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ is define as:

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

Also, for any linear subspace $\mathcal{A} \subset \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$, $A_{\mathcal{A}}$ denotes the projection of a tensor A onto $\mathcal{A}$. More precisely

$$A_{\mathcal{A}} := \arg\min_{M \in \mathcal{A}} \|A - M\|_{\mathrm{F}}.$$

2.1. Background on tensor algebra

One of the major challenges associated with low-rank tensors is that the notion of higher-order tensor decomposition and rank is ambiguous. See, e.g., Koldar and Bader (2009) for a review. There are two standard decompositions, the so-called canonical polyadic (CP) decomposition and the Tucker decomposition. In addition to that, there are various notions of tensor low-rankness considered in the application. In order to simplify notation, we focus our discussion on third-order tensors ($N = 3$) but point out that many of the notions generalize to $N > 3$. Our general theorem in Section 3 works for general $N \geq 3$.

The CP decomposition of a third-order tensor is defined as the smallest number r of rank-one tensors needed to represent a tensor $A \in \mathbb{R}^{d_1 \times d_2 \times d_3}$:

$$A = \sum_{k=1}^{r} u_{k,1} \otimes u_{k,2} \otimes u_{k,3} \tag{3}$$

where $u_{k,m} \in \mathbb{R}^{d_m}$, for $1 \leq k \leq r$ and $1 \leq m \leq 3$.

A second popular decomposition is the so-called Tucker decomposition. The Tucker decomposition of a tensor $A \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is of the form:

$$A_{j_1 j_2 j_3} = \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \sum_{k_3=1}^{r_3} S_{k_1 k_2 k_3} U_{j_1 k_1, 1} U_{j_2 k_2, 2} U_{j_3 k_3, 3}$$

so that $U_m \in \mathbb{R}^{d_m \times r_m}$ for $1 \leq m \leq 3$ are factors matrices (which are usually orthogonal) and $S \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the so-called core tensor (see, e.g., Koldar and Bader, 2009). The vector (r_1, r_2, r_3) is referred to as the Tucker ranks of A. It is not hard to see that if (3) holds, then the Tucker ranks (r_1, r_2, r_3) can be equivalently interpreted as the dimensionality of the linear spaces spanned by $\{u_{k,1} : 1 \leq k \leq r\}$, $\{u_{k,2} : 1 \leq k \leq r\}$, and $\{u_{k,3} : 1 \leq k \leq r\}$ respectively.

A convenient way to represent low Tucker ranks of a tensor is through *matricization.* Denote by $\mathcal{M}_1(\cdot)$ the mode-1 matricization of a tensor, that is $\mathcal{M}_1(A)$ is the $d_1 \times (d_2 d_3)$ matrix whose column vectors are the mode-1 fibers of $A \in \mathbb{R}^{d_1 \times d_2 \times d_3}$. A fiber is defined by fixing every index but one, which is the first index in the case of mode-1 fiber. Here A has $d_2 d_3$ mode-1 fibers. $\mathcal{M}_2(\cdot)$, $\mathcal{M}_3(\cdot)$ are defined in the same fashion. By defining

$$\text{rank}(\mathcal{M}_m(A)) = r_m(A),$$

it follows that $(r_1(A), r_2(A), r_3(A))$ represent the Tucker ranks of A. For later discussion, define $\mathcal{M}_i^{-1}(\cdot)$ to be the inverse of mode-i matricization, so

$$\mathcal{M}_1^{-1} : \mathbb{R}^{d_1 \times (d_2 \cdot d_3)} \to \mathbb{R}^{d_1 \times d_2 \times d_3},$$

$$\mathcal{M}_2^{-1} : \mathbb{R}^{d_2 \times (d_1 \cdot d_3)} \to \mathbb{R}^{d_1 \times d_2 \times d_3},$$

$$\mathcal{M}_3^{-1} : \mathbb{R}^{d_3 \times (d_1 \cdot d_2)} \to \mathbb{R}^{d_1 \times d_2 \times d_3},$$

such that $\mathcal{M}_i^{-1}(\mathcal{M}_i(A)) = A$. Also, a tensor can be vectorized via collapsing all its dimensions sequentially, i.e. $\text{vec}(A) \in \mathbb{R}^{d_1 \cdot d_2 \cdot d_3}$ for $A \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, and A's $(j_1, j_2, j_3)^{th}$ element is placed in the position of $\sum_{i=1}^{3}[(j_i - 1)\Pi_{l=i+1}^{3} d_l] + 1$ in $\text{vec}(A)$.

Further, we define *slices* of a tensor as follows. For an order-3 tensor A, the $(1,2)$ slices of A are the collection of d_3 matrices of $d_1 \times d_2$

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

2.2. Low-dimensional structural assumptions

As mentioned earlier, there is not a unique way of defining tensor and there are multiple ways to impose low-rankness/low-dimensionality on a tensor space. Given the ambiguity of tensor low-rankness, one of the goals of the paper is to develop a general framework that applies to many different notions of low-rankness. To be more concrete, we focus on three specific examples of low-rank structure. These three examples fall in the general framework of our analysis of the PGD algorithm applied to low-dimensional tensor regression. Among the three, the first two are closely related to viewing the tensor as slices of matrix or lower order tensor (e.g. 4th order tensor has 'slices' of 3rd order tensor), and hence measuring hybrid of group sparsity and lower order low-rankness. The third is the maximum of the standard Tucker ranks. For simplicity, we focus on the case when $N = 3$ and then discuss potential generalization to higher order.

Firstly we place low-rank structure on the matrix slices. In particular first define:

$$\Theta_1(r) = \left\{ A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : \sum_{j_3=1}^{d_3} \text{rank}(A_{\cdot\cdot j_3}) \le r \right\}.$$

We can view $\Theta_1(r)$ as a specific rank-r or less subset of 3rd order tensor space with the rank defined to be the sum of the rank of the matrix slices. To connect to matrices, an alternative way to parameterize $\Theta_1(r)$ is as the set of $\mathbb{R}^{d_1 . d_3 \times d_2 . d_3}$ block diagonal matrices, where each block corresponds to the matrix slice $A_{\cdot\cdot j_3} \in \mathbb{R}^{d_1 \times d_2}$ and $1 \le j_3 \le d_3$. The rank constraint in $\Theta_1(r)$, corresponds exactly to placing a rank constraint on the corresponding $\mathbb{R}^{d_1 . d_3 \times d_2 . d_3}$ matrix.

Secondly we can impose a related notion where we bound the maximum of the rank of each slice and sparsity along the matrix slices.

$$\Theta_2(r, s) = \left\{ A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : \max_{j_3} \text{rank}(A_{\cdot\cdot j_3}) \le r, \sum_{j_3=1}^{d_3} \mathbb{I}(A_{\cdot\cdot j_3} \ne 0) \le s \right\}.$$

We can view $\Theta_2(r, s)$ as a specific rank-(r,s) or less subset of the third-order tensor. One natural example where imposing this combination of low-rankness and sparsity for is vector auto-regressive (VAR) models (see e.g. Basu and Michailidis (2015)). For example, if we have an M-variate time series and consider a VAR(p) model, $d_1 = d_2 = M$ and $d_3 = p$. We want each auto-regressive matrix slice $A_{\cdot\cdot j_3}$ to have low-rank and the total number of lags involved in the problem to be sparse (e.g. to account for immediate effects and seasonal effects).

Finally, we impose the assumption that Tucker ranks are upper bounded:

$$\Theta_3(r_1, r_2, r_3) = \left\{ A \in \mathbb{R}^{d_1 \times d_2 \times d_3} : r_i(A) \le r_i \text{ for } i = 1, 2, 3 \right\}.$$

Note that all these low-dimensional structural assumptions $\Theta_1(r)$, $\Theta_2(r,s)$ and $\Theta_3(r_1,r_2,r_3)$ are non-convex sets. Note that $\Theta_2(r,s)$ and $\Theta_3(r_1,r_2,r_3)$ have two parameters (r,s) and three parameters (r_1,r_2,r_3) respectively and our framework deals with this. In the next subsection we introduce a general projected gradient descent (PGD) algorithm for minimizing the generalized linear model objective (2) subject to the parameter tensor A belonging to a potentially non-convex constraint set Θ. Also note that it is straightforward to extend $\Theta_3(r_1,r_2,r_3)$ to $N > 3$.

2.3. Projected Gradient Descent (PGD)

In this section we introduce the non-convex projected gradient descent (PGD) approaches developed in Jain et al. (2014, 2016) adapted to a general tensor space Θ. The problem we are interested in is minimizing the generalized linear model objective (2) subject to A belonging to a potentially non-convex set. The PGD algorithm for minimizing a general loss function $f(A)$ subject to the constraint $A \in \Theta$ is as follows:

Algorithm 1 Projected Gradient Descent

```
Input : data $\mathbf{Y}, \mathbf{X}$, parameter space $\Theta$, iterations $K$, step size $\eta$
Initialize : $k = 0$, $\widehat{T}_0 \in \Theta$
for $k = 1, 2, \ldots, K$ do
    $g_k = \widehat{T}_k - \eta \nabla f(\widehat{T}_k)$ (gradient step)
    $\widehat{T}_{k+1} = P_\Theta(g_k)$ or $\widehat{T}_{k+1} = \widehat{P}_\Theta(g_k)$ ((approximate) projection step)
end for
Output : $\widehat{T}_K$
```

The notation $\widehat{P}_\Theta(\cdot)$ refers to an approximate projection on to Θ if an exact projection is not implementable. The PGD algorithm has been widely used for both convex and non-convex objectives and constraint sets. In our setting, we choose the negative log-likelihood for the generalized linear model as the functions $f(A)$ to minimize.

An important question associated with projected gradients is whether the projection on to Θ is implementable. The projections we consider for $\Theta_1(r)$, $\Theta_2(r,s)$ and $\Theta_3(r_1,r_2,r_3)$ can all be implemented approximately as combinations of projections on to matrix or vector subspaces defined in Jain et al. (2014, 2016). By using sparsity and low-rankness projections in the vector and matrix respectively, the projections in Θ_1, Θ_2 and Θ_3 are implementable.

In particular, for a vector $v \in \mathbb{R}^d$, we define the projection operator $\tilde{P}_s(v)$ as the projection on to the set of s-sparse vectors by selecting the s largest elements of v in ℓ_2-norm. That is:

$$\tilde{P}_s(v) := \operatorname*{arg\,min}_{\|z\|_{\ell_0} \le s} \|z - v\|_{\ell_2}.$$

For a matrix $M \in \mathbb{R}^{d_1 \times d_2}$, let $\bar{P}_r(M)$ denote the rank-r projection:

$$\bar{P}_r(M) := \operatorname*{arg\,min}_{\mathrm{rank}(Z) \le r} \|Z - M\|_{\mathrm{F}}.$$

As mentioned in Jain et al. (2014, 2016), both projections are computable. $\tilde{P}_s(v)$ is the thresholding operator which selects the s largest elements of v in ℓ_2-norm and $\bar{P}_r(M)$ can be selecting the top r singular vectors of M. For the remainder of this paper we use both of these projection operators for vectors and matrices respectively.

3. Main Results

In this section we present our general theoretical results where we provide a statistical guarantee for the PGD algorithm applied to a low-dimensional space Θ.

3.1. Properties for Θ and its projection

To ensure the PGD algorithm converges for a given subspace Θ, we view it as a member of a collection of subspaces $\{\Theta(t) : t \in \Xi\}$ for some $\Xi \subset \mathbb{Z}_+^k$ and require some general properties for this collection. The index t typically represents a sparsity and/or low-rank index and may be multi-dimensional. For example, $\Theta_1(r)$ is indexed by rank r where

$$\Xi = \{0, \ldots, d_3 \cdot \min\{d_1, d_2\}\}.$$

Similarly, $\Theta_2(r, s)$ is indexed by $t = (r, s)$ so that

$$\Xi = \{(0, 0), \ldots, (\min\{d_1, d_2\}, d_3)\},$$

and $\Theta_3(r_1, r_2, r_3)$ is indexed by rank (r_1, r_2, r_3) so that

$$\Xi = \{(0, 0, 0), \ldots, (\min\{d_1, d_2d_3\}, \min\{d_2, d_1d_3\}, \min\{d_3, d_1d_2\})\}.$$

Note that the Ξ is partially ordered where $a \geq (\leq, <, >)b$ for two vectors a and b of conformable dimension means the inequality holds in an element-wise fashion, i.e. the inequality holds for each element of a and its counterparty in b.

Definition 1. A set $\{\Theta(t) : t \in \Xi\}$ is a *superadditive and partially ordered collection of symmetric cones* if

(1) each member $\Theta(t)$ is a *symmetric cone* in that if $z \in \Theta(t)$, then $cz \in \Theta(t)$ for any $c \in \mathbb{R}$;

(2) the set is *partially ordered* in that for any $t_1 \leq t_2$, $\Theta(t_1) \subset \Theta(t_2)$;

(3) the set is *superadditive* in that $\Theta(t_1) + \Theta(t_2) \subset \Theta(t_1 + t_2)$.

The first two properties basically state that we have a set of symmetric cones in the tensor space with a partial ordering indexed by t. The last property requires that the collection of subspaces be superadditive in that the Minkowski sum of any two subspaces is contained in the subspace of dimension that is the sum of the two lower dimensions. All three properties are essential for deriving theoretical guarantees for the PGD algorithm. By relying on these properties alone, we obtain a general result that provides a unified way of dealing with $\Theta_1(r), \Theta_2(r, s)$ and $\Theta_3(r_1, r_2, r_3)$ as well as many other collections of subspaces.

Furthermore, we introduce the following property of contractive projection, for P_Θ or $\widehat{P}_\Theta$ in Algorithm 1, that is essential for the theoretical performance of the PGD algorithm. Again, we shall view these operators as members of a collection of operators $Q_{\Theta(t)}$:

$\cup_t\Theta(t) \mapsto \Theta(t)$. The contractive projection property says that, when $Q_{\Theta(t)} : \cup_t\Theta(t) \mapsto \Theta(t)$ are viewed as projections, projection onto a larger "dimension" incurs less approximation error *per dimension* compared to projection onto a smaller dimension, up to a constant factor. This property is adopted from earlier study on vector/matrix case. (see, e.g., Jain et al., 2014, 2016). The difference here is that in the tensor case, the constant factor δ is no longer one as in vector/matrix case. And later theorems will show that any finite δ can guarantee linear convergence with error bound expressed in Gaussian width for PGD in tensor case. This means that certain low-cost approximate low-rank projections in the tensor case are sufficient for PGD and later we will see Tucker rank is the case.

Definition 2. We say that a set $\{\Theta(t) : t \geq 0\}$ and corresponding operators $Q_{\Theta(t)} : \cup_t\Theta(t) \mapsto \Theta(t)$ satisfy the *contractive projection property* for some $\delta > 0$, denoted by CPP(δ), if for any $t_1 < t_2 < t_0$, $Y \in \Theta(t_1)$, and $Z \in \Theta(t_0)$:

$$\|Q_{\Theta(t_2)}(Z) - Z\|_{\mathrm{F}} \leq \delta \left\| \frac{t_0 - t_2}{t_0 - t_1} \right\|_{\ell_\infty}^{1/2} \cdot \|Y - Z\|_{\mathrm{F}}.$$

Here, when $\Theta(t)$ is indexed by multi-dimensional t, the division $\frac{t_0-t_2}{t_0-t_1}$ refers to a vector with the jth element to be $\frac{(t_0)_j-(t_2)_j}{(t_0)_j-(t_1)_j}$.

It is clear that $\Theta_1(r)$ is isomorphic to rank-r block diagonal matrices with diagonal blocks $A_{\cdot\cdot1}$, $A_{\cdot\cdot2}$,..., $A_{\cdot\cdot d_3}$ so that $\{\Theta_1(r)\}$ satisfies Definition 1. It is also easy to verify that $\{\Theta_1(r)\}$ and its projections $\{P_{\Theta_1(r)}\}$ which applies $\bar{P}_r(.)$ to the $d_1.d_3 \times d_2.d_3$ block-diagonal matrix described earlier obeys CPP(1). Later, we will see in Lemmas 2 and 3 that these two properties are also satisfied by $\{\Theta_2(r,s)\}$ and $\{\Theta_3(r_1,r_2,r_3)\}$, and their appropriate (approximate) projections.

3.2. Restricted strong convexity

Now we state some general requirements on the loss function, namely the restricted strong convexity and smoothness conditions (RSCS), that is another essential part for the guarantee of PGD performance (see, e.g., Jain et al., 2014, 2016). Recall that for $f(A)$, a function of tensor A, we can abuse the notation to view $f(A)$ as a function of vectorized tensor, i.e. $f(A) = f(\mathrm{vec}(A))$and use $\nabla^2 f(A)$ to denote the Hessian of function f on vectorized tensor. Please refer to Section 2.1 for the formal definition of vectorization of a tensor.

Definition 3. We say that a function f satisfies *restricted strong convexity and smoothness conditions $RSCS(\Theta, C_l, C_U)$* for a set Θ, and $0 < C_l < C_u < \infty$ if for any $A \in \Theta$, $\nabla^2 f(A)$ is positive semidefinite such that for any $B \in \Theta$

$$C_l \cdot \|B\|_{\mathrm{F}} \leq \|\nabla^2 f(A) \cdot \mathrm{vec}(B)\|_{\ell_2} \leq C_u \cdot \|B\|_{\mathrm{F}},$$

for some constants $C_l < C_u$. Note that $RSCS(\Theta, C_l, C_U)$ reduces to restricted strong convexity and restricted smoothness assumptios for vectors and matrices (see e.g. Jain et al. (2014, 2016)) when A and B are vectors and matrices. We first state the following Theorem about the PGD performance under general loss function which is a tensor version of the results in Jain et al. (2014, 2016).

Theorem 1 ***(PGD Error Bound for General Loss Function)*** *Suppose that $\{\Theta(t) : t \geq 0\}$ is a superadditive and partially ordered collection of symmetric cones, together with operators $\{P_{\Theta(t)} : t \geq 0\}$ which obey CPP(δ) for some constant $\delta > 0$, and f satisfies $RSCS(\Theta(t_0), C_l, C_u)$ for some constants C_l and C_u. Let $\widehat{T}_K$ be the output from the Kth iteration of applying PGD algorithm with step size $\eta = 1/C_u$, and projection $P_{\Theta(t_1)}$ where*

$$t_1 = \left\lceil \frac{4\delta^2 C_u^2 C_l^{-2}}{1 + 4\delta^2 C_u^2 C_l^{-2}} \cdot t_0 \right\rceil .$$

Then

$$\sup_{T \in \Theta(t_0 - t_1)} \|\widehat{T}_K - T\|_{\mathrm{F}} \leq 4\eta C_u C_l^{-1} \cdot \sup_{A \in \Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)} \langle \nabla f(T), A \rangle + \epsilon,$$

for any

$$K \geq 2 C_u C_l^{-1} \log \left(\frac{\|T\|_{\mathrm{F}}}{\epsilon} \right).$$

Basically, Theorem 1 guarantees that the PGD applied to general loss minimization problem with general low-rank constraint enjoys linear convergence rate with statistical error bounded by the restricted norm of the gradient of loss function evaluation at the true parameter. Note that we are taking a supremum over the set $\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)$ which relates to the local Gaussian width which we define shortly. More insight arises when we specialize f to the generalized linear model.

3.3. Generalized linear models

To use Theorem 1 for a specific f and collection of $\{\Theta(t) : t \geq 0\}$, we need to verify the conditions on $\{\Theta(t) : t \geq 0\}$, $\{P_{\Theta(t)} : t \geq 0\}$ and f satisfying $RSCS(\Theta, C_l, C_U)$, and choose the step-size in the PGD accordingly.

First we turn our attention to the covariate tensor $(X^{(i)})_{i=1}^n$ where $X^{(i)} \in \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ and how it relates to the $RSCS(\Theta, C_l, C_U)$. With slight abuse of notation, write

$$\mathrm{vec}(X^{(i)}) \in \mathbb{R}^{d_1 d_2 \cdots d_N}$$

for $1 \leq i \leq n$ which is the vectorization of each tensor covariate $X^{(i)}$. For convenience let $D_N = d_1 d_2 \cdots d_N$. Further as mentioned for technical convenience we assume a Gaussian design of independent sample tensors $X^{(i)}$ s.t.

$$\mathrm{vec}(X^{(i)}) \sim \mathcal{N}(0, \Sigma) \text{ where } \Sigma \in \mathbb{R}^{D_N \times D_N}. \tag{4}$$

With more technical work our results may be extended beyond random Gaussian designs. We shall assume that Σ has bounded eigenvalues. Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ represent the smallest and largest eigenvalues of a matrix, respectively. In what follows, we shall assume that

$$c_\ell^2 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq c_u^2, \tag{5}$$

for some constants $0 < c_\ell \leq c_u < \infty$. For our analysis of the non-convex projected gradient descent algorithm, we define the condition number $\kappa = c_u / c_l$.

A quantity that emerges from our analysis is the *Gaussian width* (see, e.g., Gordon, 1988) of a set $S \subset \mathbb{R}^{d_1 \times d_2 \times \ldots \times d_N}$ which is defined to be:

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

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. The Gaussian width is a standard notion of size or complexity of a subset of tensors S.

In addition to the positive semi-definiteness of the Hessian in the GLM, the following Lemma extends a restricted upper and lower eigenvalue condition to the sample version of Σ and hence guarantees the RSCS for the GLM with Gaussian covariates with quite general covariance structure.

Lemma 1 *Assume that* (4) *and* (5) *hold. For any $\tau > 1$, there exist constants $c_1, c_2, c_3 > 0$ such that if $n \geq c_1 w_G^2[\Theta \cap \mathbb{B}_{\mathrm{F}}(1)]$, then with probability at least $1 - c_2 \exp(-c_3 w_G^2[\Theta \cap \mathbb{B}_{\mathrm{F}}(1)])$,*

$$\left(\tau^{-1} c_l\right)^2 \|A\|_{\mathrm{F}}^2 \leq \frac{1}{n}\sum_{i=1}^{n} \langle X^{(i)}, A \rangle^2 \leq (\tau c_u)^2 \|A\|_{\mathrm{F}}^2, \qquad \forall A \in \Theta.$$

Lemma 1 is a direct consequence of Theorem 6 in Banerjee et al. (2015). Using these definitions, we are in a position to state the main result for generalized linear models.

Theorem 2 ***(PGD Error Bound for Generalized Linear Model)*** *Suppose that $\{\Theta(t) : t \geq 0\}$ is a superadditive and partially ordered collection of symmetric cones, and together with operators $\{P_{\Theta(t)} : t \geq 0\}$ which obey CPP(δ) for some constant $\delta > 0$. Assume that $\{(X^{(i)}, Y^{(i)}) : i = 1, \ldots, n\}$ follow the generalized linear model* (1) *and $X^{(i)}$'s satisfy* (4) *and* (5)*, $\mathbb{E}|Y^{(i)}|^q \leq M_Y$ for some $q > 2$ and $M_Y > 0$, $1/\tau_0^2 \leq Var(Y^{(i)}) \leq \tau_0^2$ for $i = 1, \ldots, n$ and some $\tau_0 > 0$, and $n > c_1 w_G^2[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)]$ for some t_0 and $c_1 > 0$. Let $\widehat{T}_K$ be the output from the K^{th} iteration of applying PGD algorithm to* (2) *with step size $\eta = (\tau c_u)^{-2}$ and projection $P_{\Theta(t_1)}$ where*

$$t_1 = \left\lceil \frac{4\delta^2 \tau^8 \kappa^4}{1 + 4\delta^2 \tau^8 \kappa^4} \cdot t_0 \right\rceil,$$

for any given $\tau > \tau_0$. Then there exist constants $c_2, c_3, c_4, c_5 > 0$ such that

$$\sup_{T \in \Theta(t_0 - t_1)} \|\widehat{T}_K - T\|_{\mathrm{F}} \leq \frac{c_5 \eta \tau^4 \kappa^2 c_u M_Y^{1/q}}{\sqrt{n}} \cdot w_G[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)] + \epsilon,$$

with probability at least

$$1 - K c_2 \exp\left\{-c_3 w_G^2[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)]\right\} - K c_4 n^{-(q/2-1)} \log^q n,$$

for any

$$K \geq 2\tau^4 \kappa^2 \log\left(\frac{\|T\|_{\mathrm{F}}}{\epsilon}\right).$$

Notice that the statistical error we have is related to the Gaussian width of the intersection of a unit Frobenius ball and an (often non-convex) subset of low-dimensional structure $w_G[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)]$. The intersection of $\Theta(t_0)$ with $\mathbb{B}_{\mathrm{F}}(1)$ means we are *localizing* the Gaussian width to a unit Frobenius norm ball around T. *Localization* of the Gaussian width means a sharper statsitical error bound can be proven and the benefits of localization in empirical risk minimization have been previously discussed in Bartlett et al. (2005). Later we will see how the main result leads to sample complexity bounds applied to $\Theta_2(r, s)$ and $\Theta_3(r_1, r_2, r_3)$. To the best of our knowledge this is the first general result that provides statistical guarantees in terms of the local Gaussian width of $\Theta(t_0)$ for the projected gradient descent algorithm. Expressing the error bound in terms of the Gaussian width allows an easy comparison to already established error bounds for convex regularization schemes which we discuss in Section 3.4.

The moment conditions on the response in Theorem 2 are in place to ensure that the restricted strong convexity and restricted smoothness conditions are satisfied for a non-quadratic loss. When specialized to normal linear regression, these conditions could be further removed.

3.4. Gaussian model and comparison to convex regularization

Consider the Gaussian linear regression setting which corresponds to the GLM in Equation (1) with $a(\theta) = \frac{\theta^2}{2}$. In particular

$$Y^{(i)} = \langle X^{(i)}, T \rangle + \zeta^{(i)}, \tag{6}$$

where $\zeta^{(i)}$'s are independent $\mathcal{N}(0, \sigma^2)$ random variables. Furthermore, substituting $a(\theta) = \frac{\theta^2}{2}$ into the GLM objective (2), we have the least-squares objective:

$$f(A) = \frac{1}{2n} \sum_{i=1}^{n} (Y^{(i)} - \langle X^{(i)}, A \rangle)^2. \tag{7}$$

Now we state our main result for the normal linear regression.

Theorem 3 ***(PGD Error Bound for Normal Linear Regression)*** *Suppose that $\{\Theta(t) : t \geq 0\}$ is a superadditive and partially ordered collection of symmetric cones, and together with operators $\{P_{\Theta(t)} : t \geq 0\}$ which obey CPP(δ) for some constant $\delta > 0$. Assume that $\{(X^{(i)}, Y^{(i)}) : i = 1, \ldots, n\}$ follow the Gaussian linear model (6) where $n > c_1 w_G^2[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)]$ for some t_0 and $c_1 > 0$. Let $\widehat{T}_K$ be the output from the K^{th} iteration of applying PGD algorithm to (2) with step size $\eta = (\tau c_u)^{-2}$ and projection $P_{\Theta(t_1)}$ where*

$$t_1 = \left\lceil \frac{4\delta^2 \tau^8 \kappa^4}{1 + 4\delta^2 \tau^8 \kappa^4} \cdot t_0 \right\rceil,$$

for any given $\tau > 1$. Then there exist constants $c_2, c_3 > 0$ such that

$$\sup_{T \in \Theta(t_0 - t_1)} \|\widehat{T}_K - T\|_{\mathrm{F}} \leq \frac{8\eta \tau^4 \kappa^2 c_u \sigma}{\sqrt{n}} w_G[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)] + \epsilon,$$

with probability at least

$$1 - Kc_2 \exp\left\{-c_3 w_G^2[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)]\right\},$$

for any

$$K \geq 2\tau^4\kappa^2 \log\left(\frac{\|T\|_{\mathrm{F}}}{\epsilon}\right).$$

One of the contributions of this paper outlined in the introduction is to compare the non-convex PGD approach in tensor regression to the existing convex regularization approach analyzed in Raskutti and Yuan (2015) applied to the Gaussian linear model (6). In this section we first summarize the general result from Raskutti and Yuan (2015) and then provide a comparison to the error bound for the non-convex PGD approach. In particular, the following estimator for T is considered in Raskutti and Yuan (2015):

$$\widehat{T} \in \operatorname*{arg\,min}_{A \in \mathbb{R}^{d_1 \times \cdots \times d_N}} \left\{\frac{1}{2n}\sum_{i=1}^{n} \|Y^{(i)} - \langle A, X^{(i)}\rangle\|_{\mathrm{F}}^2 + \lambda \mathcal{R}(A)\right\}, \tag{8}$$

where the convex regularizer $\mathcal{R}(\cdot)$ is a norm on $\mathbb{R}^{d_1 \times \cdots \times d_N}$, and $\lambda > 0$ is a tuning parameter. The *convex conjugate* for $\mathcal{R}$ (see e.g. Rockafellar (1970) for details) is given by:

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

For example if $\mathcal{R}(A) = \|A\|_*$, then $\mathcal{R}^*(B) = \|B\|_s$. Following Negahban et al. (2012), for a subspace Θ of $\mathbb{R}^{d_1 \times \cdots \times d_N}$, define its compatibility constant $s(\Theta)$ as

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

which can be interpreted as a notion of low-dimensionality of Θ.

Raskutti and Yuan (2015) show that if $\widehat{T}$ is defined by (8) and the regularizer $\mathcal{R}(\cdot)$ is *decomposable* with respect to Θ, then if

$$\lambda \geq 2w_G(\mathbb{B}_{\mathcal{R}}(1)), \tag{9}$$

where recall that $w_G(\mathbb{B}_{\mathcal{R}}(1)) = \mathbb{E}\big(\sup_{A \in \mathbb{B}_{\mathcal{R}}(1)} \langle A, G\rangle\big)$. Then according to Theorem 1 in Raskutti and Yuan (2015),

$$\max\left\{\|\widehat{T} - T\|_n, \|\widehat{T} - T\|_{\mathrm{F}}\right\} \lesssim \frac{\sqrt{s(\Theta)}\lambda}{\sqrt{n}}. \tag{10}$$

with probability at least $1 - \exp(-cn)$ for some constant $c > 0$. In particular setting $\lambda = 2w_G(\mathbb{B}_{\mathcal{R}}(1))$,

$$\max\left\{\|\widehat{T} - T\|_n, \|\widehat{T} - T\|_{\mathrm{F}}\right\} \lesssim \frac{\sqrt{s(\Theta)}w_G(\mathbb{B}_{\mathcal{R}}(1))}{\sqrt{n}}.$$

The error bound boils down to bounding two quantities, $s(\Theta)$ and $w_G(\mathbb{B}_{\mathcal{R}}(1))$, noting that for comparison pursposes the subpace Θ in the convex case refers to $\Theta(t_0)$ in the non-convex case. In the next section we provide a comparison between the error bound for the non-convex PGD approach and the convex regularization approach. To be clear, Raskutti and Yuan (2015) consider multi-response models where the response $Y^{(i)}$ can be a tensor which are not considered in this paper.

The error bound for the convex regularization scheme scales as $\sqrt{s(\Theta(t_0))}w_G[\mathbb{B}_{\mathcal{R}}(1)]/\sqrt{n}$ while we recall that the error bound we prove in this paper for the non-convex PGD approach scales as $w_G[\Theta(t_0) \cap \mathbb{B}_{\rm F}(1)]/\sqrt{n}$. Hence how the Frobenius error for the non-convex and convex approach scales depends on which of the quantities $\sqrt{s(\Theta(t_0))}w_G[\mathbb{B}_{\mathcal{R}}(1)]/\sqrt{n}$ or $w_G[\Theta(t_0) \cap \mathbb{B}_{\rm F}(1)]/\sqrt{n}$ is larger. It follows easily that $w_G[\Theta(t_0) \cap \mathbb{B}_{\rm F}(1)] \le \sqrt{s(\Theta(t_0))}w_G[\mathbb{B}_{\mathcal{R}}(1)]$ since

$$\begin{aligned} w_G[\Theta(t_0) \cap \mathbb{B}_{\rm F}(1)] &= \mathbb{E}\big[\sup_{A \in \Theta(t_0), \|A\|_{\rm F} \le 1} \langle A, G \rangle\big] \\ &\le \mathbb{E}\big[\sup_{\mathcal{R}(A) \le \sqrt{s(\Theta(t_0))}} \langle A, G \rangle\big] \\ &= \sqrt{s(\Theta(t_0))}\mathbb{E}\big[\sup_{\mathcal{R}(A) \le 1} \langle A, G \rangle\big] = \sqrt{s(\Theta(t_0))}w_G[\mathbb{B}_{\mathcal{R}}(1)]. \end{aligned}$$

The first inequality follows from the subspace compatibility constant since for all $A \in \Theta(t_0) \cap \mathbb{B}_{\rm F}(1)$, $\mathcal{R}(A) \le \sqrt{s(\Theta(t_0))}\|A\|_{\rm F} \le \sqrt{s(\Theta(t_0))}$ and the final equality follows since $\mathcal{R}(\cdot)$ is a convex function. Therefore the non-convex error bound is always no larger than the convex error bound and the important question is whether there is a gap between the convex and non-convex bounds which implies a superior bound in the non-convex case. For examples involving sparse vectors and low-rank matrices as studied in e.g., Buhlmann and van de Geer (2011); Jain et al. (2014, 2016), these two quantities end up being identical up to a constant. On the other hand for tensors, as we see in this paper for $\Theta_3(r_1, r_2, r_3)$, the Gaussian width using the non-convex approach is smaller which presents an additional benefit for the non-convex projection approach.

In terms of implementation, the regularizer $\mathcal{R}(\cdot)$ needs to be defined in the convex approach and the important question is whether the convex regularizer is implementable for the low-dimensional structure of interest. For the non-convex approach, the important implementation issue is whether exact or approximate projection that satisfies the contractive projection property is implementable. These implementation issues have been resolved in the vector and matrix cases (see, e.g., Jain et al., 2014, 2016). In Section 5 in this paper, we focus on whether they apply in the low-rank tensor case under the low-dimensional structure Θ_1, Θ_2 and Θ_3.

4. Specific low rank structure

In this section, we apply Theorem 3 (and by extension Theorem 2) to $\Theta_1(r)$, $\Theta_2(r, s)$ and $\Theta_3(r_1, r_2, r_3)$ and compare our theoretical result to the theoretical bound achieved by the convex regularization approach. Recall that $\Theta_1(r)$ is isomorphic to rank-r block diagonal matrices with diagonal blocks $A_{\cdot\cdot 1}$, $A_{\cdot\cdot 2}$,..., $A_{\cdot\cdot d_3}$ so that its treatment is identical to the case of low rank matrix estimation. See Jain et al. (2016) for further discussions. Hence we will focus on $\Theta_2(r,s)$ and $\Theta_3(r_1,r_2,r_3)$. To prove error bounds using Theorem 3 we find an exact or approximate projection $P_{\Theta(t)}$, prove the contractive projection property and then find an error bound on the Gaussian width $w_G[\Theta(t) \cap \mathbb{B}_{\mathrm{F}}(1)]$.

4.1. Low-rank structure for matrix slices

Recall that

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

For the projection, we define the two-step projection $P_{\Theta_2(r,s)}$:

(1) for each matrix slice $A_{\cdot\cdot j_3}$ where $1 \le j_3 \le d_3$, let $\tilde{A}_{\cdot\cdot j_3} := \bar{P}_r(A_{\cdot\cdot j_3})$ be the best rank r approximation of $A_{\cdot\cdot j_3}$;

(2) to impose sparsity, retain s out of d_3 slices with the largest magnitude $\|\tilde{A}_{\cdot\cdot j_3}\|_{\mathrm{F}}$, and zero out all other slices.

As discussed earlier both steps are easily computable using thresholding and SVD operators as discussed in Jain et al. (2014, 2016).The following lemma proves that the contractive property of projection onto $\Theta_2(r,s)$ holds for our $P_{\Theta_2(r,s)}$.

Lemma 2 *Let the projection operator $P_{\Theta_2(r,s)}$ be defined above. Suppose $Z \in \Theta_2(r_0,s_0)$, and $r_1 < r_2 < r_0, s_1 < s_2 < s_0$. Then for any $Y \in \Theta_2(r_1,s_1)$, we have*

$$\|P_{\Theta_2(r_2,s_2)}(Z) - Z\|_{\mathrm{F}} \le (\alpha + \beta + \alpha\beta) \cdot \|Y - Z\|_{\mathrm{F}}.$$

where $\alpha = \sqrt{(s_0 - s_2)/(s_0 - s_1)}$, $\beta = \sqrt{(r_0 - r_2)/(r_0 - r_1)}$.

Consequently we have the following Theorem:

Theorem 4 *Let $\{X^{(i)}, Y^{(i)}\}_{i=1}^n$ follow a Gaussian linear model as defined by (6) with $T \in \Theta_2(r,s)$ and*

$$n \ge c_1 \cdot sr(d_1 + d_2 + \log d_3)$$

for some constant $c_1 > 0$. Then, applying the PGD algorithm with step size $\eta = (\tau c_u)^{-2}$ and projection $P_{\Theta(r',s')}$ where

$$s' = \lceil 36\tau^8\kappa^4 s \rceil, \qquad \text{and} \qquad r' = \lceil 36\tau^8\kappa^4 r \rceil,$$

guarantees that, with probability at least $1 - Kc_2 \exp\{-c_3 \max(d_1, d_2, \log d_3)\}$, after $K \ge 2\tau^4\kappa^2 \log(\|T\|_{\mathrm{F}}/\epsilon)$ iterations,

$$\|\widehat{T}_K - T\|_{\mathrm{F}} \le c_4 \sigma \sqrt{\frac{sr \max\{d_1, d_2, \log(d_3)\}}{n}} + \epsilon$$

for any $\tau > 1$, and some constants $c_2, c_3, c_4 > 0$.

The convex regularization approach defined by Raskutti and Yuan (2015) is not directly applicable for $\Theta_2(r,s)$ since there is no suitable choice of regularizer that imposes both low-rankness of each slice and sparsity. Therefore we discuss the convex regularization approach applied to the parameter space $\Theta_1(r)$ for which a natural choice of regularizer is:

$$\mathcal{R}_1(A) = \sum_{j_3=1}^{d_3} \|A_{\cdot\cdot j_3}\|_*,$$

where $\|\cdot\|_*$ refers to the standard nuclear norm of a matrix. Let $\widehat{T}$ be an estimator corresponding to the minimizer of the regularized least-squares estimator defined by (8) with regularizer $\mathcal{R}_1(A)$. Lemma 6 in Raskutti and Yuan (2015) proves that

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

Notice that both $\Theta_1(r)$ and $\Theta_2(r,s)$ focus on the low-rankness of matrix slices of a tensor, and actually $\Theta_1(\cdot)$ can be seen as relaxation of $\Theta_2(\cdot,\cdot)$ since $\Theta_2(s,r) \subset \Theta_1(sr)$. Theorem 4 guarantees that, under the restriction of sparse slices of low-rank matrices, PGD achieves the linear convergence rate with the statistical error of order

$$\sqrt{\frac{sr\max\{d_1, d_2, \log(d_3)\}}{n}}.$$

If we compare this result with the risk bound of the convex regularization approach where the true tensor parameter lies in $\Theta_1(r)$ we see that replacing r by sr yields the same rate which makes intuitive sense in light of the observation that $\Theta_2(s,r) \subset \Theta_1(sr)$.

4.2. Low Tucker rank

We now consider the general set of tensors with low Tucker rank:

$$\Theta_3(r_1, r_2, r_3) = \left\{ A \in \mathbb{R}^{d_1\times d_2\times d_3} : r_i(A) \le r_i \text{ for } i = 1,2,3 \right\}.$$

Although we focus on $N = 3$, note that $\Theta_3(r_1,r_2,r_3)$ can be easily extended to general N and we also consider $N = 4$ in the simulations.

To define the projection $P_{\Theta_3(r_1,r_2,r_3)}$ on to $\Theta_3(r_1,r_2,r_3)$, we exploit the connection between Tucker ranks and ranks of different matricizations mentioned earlier. Recall that the matricization operator $\mathcal{M}_j$ maps a tensor to a matrix and the inverse operator $\mathcal{M}_j^{-1}$ maps a matrix back to a tensor. Let $\bar{P}_r(M)$ be the low-rank projection operator that maps a matrix M to its best rank r approximation. Then we can define the approximate projection $\widehat{P}_{\Theta_3(r_1,r_2,r_3)}$ as follows:

$$\widehat{P}_{\Theta_3(r_1,r_2,r_3)}(A) := (\mathcal{M}_3^{-1} \circ \bar{P}_{r_3} \circ \mathcal{M}_3) \circ (\mathcal{M}_2^{-1} \circ \bar{P}_{r_2} \circ \mathcal{M}_2) \circ (\mathcal{M}_1^{-1} \circ \bar{P}_{r_1} \circ \mathcal{M}_1)(A). \quad (11)$$

The order of which matricization is performed is nonessential. Similar to before, we have the following projection lemma to be essential in the analysis of PGD applied to the restricted parameter space Θ_3.

Lemma 3 *Suppose $Z \in \Theta_3(r_1^{(0)}, r_2^{(0)}, r_3^{(0)})$, and $r_i^{(1)} < r_i^{(2)} < r_i^{(0)}$ for $i = 1, 2, 3$. Then for any $Y \in \Theta_3(r_1^{(1)}, r_2^{(1)}, r_3^{(1)})$, we have*

$$\|\widehat{P}_{\Theta_3(r_1,r_2,r_3)}(Z) - Z\|_{\mathrm{F}} \leq [(\beta_1 + 1)(\beta_2 + 1)(\beta_3 + 1) - 1]\|Y - Z\|_{\mathrm{F}}$$

where $\beta_i = \sqrt{(r_i^{(0)} - r_i^{(2)})/(r_i^{(0)} - r_i^{(1)})}$.

This allows us to derive the following result for the PGD algorithm applied with projection operator $\widehat{P}_{\Theta_3(r_1',r_2',r_3')}(\cdot)$. Basically, the sequential matrix low-rank projection, as an approximate projection onto low Tucker rank subset, could achieve the same order error rate as the exact low Tucker rank projection which might involve expensive iterative computation.

Theorem 5 *Let $\{X^{(i)}, Y^{(i)}\}_{i=1}^n$ follow a Gaussian linear model as defined by (6) with $T \in \Theta_3(r_1, r_2, r_3)$ and*

$$n \geq c_1 \cdot \min\{r_1(d_1 + d_2 d_3), r_2(d_2 + d_1 d_3), r_3(d_3 + d_1 d_2)\},$$

for some constant $c_1 > 0$. Then, applying the PGD algorithm with step size $\eta = (\tau c_u)^{-2}$ and projection $\widehat{P}_{\Theta_3(r_1',r_2',r_3')}$ where

$$r_i' = \lceil 196\tau^8\kappa^4 r_i\rceil \text{ for } i = 1, 2, 3$$

guarantees that, with probability at least $1 - Kc_2 \exp\{-c_3 \min(d_1 + d_2 d_3, d_2 + d_1 d_3, d_3 + d_1 d_2)\}$, after $K \geq 2\tau^4\kappa^2 \log(\|T\|_{\mathrm{F}}/\epsilon)$ iterations,

$$\|\widehat{T}_K - T\|_{\mathrm{F}} \leq c_4\sigma\sqrt{\frac{\min\{r_1(d_1 + d_2 d_3), r_2(d_2 + d_1 d_3), r_3(d_3 + d_1 d_2)\}}{n}} + \epsilon$$

for any $\tau > 1$, and some constants $c_2, c_3, c_4 > 0$.

In Raskutti and Yuan (2015), the following convex low-rankness regularizer is considered for the space $\Theta_3(r)$:

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

Let $\widehat{T}$ be an estimator corresponding to the minimizer of the regularized least-squares estimator defined by (8) with regularizer $\mathcal{R}_2(A)$. An adaptation of the proof of Lemma 10 in Raskutti and Yuan (2015) to $\Theta_3(r_1, r_2, r_3)$ proves that:

$$\|\widehat{T} - T\|_{\mathrm{F}} \lesssim \sqrt{\frac{\max(r_1, r_2, r_3) \cdot \max(d_1 + d_2 d_3, d_2 + d_1 d_3, d_3 + d_1 d_2)}{n}}.$$

This shows that convex relaxation in this particular case has greater mean-squared error since the minimum is replaced by the maximum. The underlying reason is that the non-convex PGD approach selects the optimal choice of matricization whereas the convex regularization approach takes an average of the three matricizations which is sub-optimal. While it may be argued that one could use the regularizer corresponding with only the optimal $\mathcal{M}_1(.)$, $\mathcal{M}_2(.)$, or $\mathcal{M}_3(.)$, since the Tucker ranks (r_1, r_2, r_3) are unknown, it is impossible to know which matricization to use in the convex case.

5. Simulations

In this section, we provide a simulation study that firstly verifies that the non-convex PGD algorithm performs well in solving least-squares, logistic and Poisson regression problems, compares the non-convex PGD approach with the convex regularization approach we discussed earlier, and also compares using tensor regularization and naive matricization schemes. Our simulation study includes both third and fourth order tensors. For the purpose of illustration, we consider the balanced-dimension situation where $d = d_1 = d_2 = d_3 (= d_4)$, and hence the number of elements is $p = d^3$ for a third order tensor and $p = d^4$ for a fourth order tensor.

5.1. Data generation

We first describe three different ways of generating random tensor coefficient T with different types of low tensor rank structure.

1. (Low CP rank) Generate three independent groups of r independent random vectors of unit length, $\{u_{k,1}\}_{k=1}^r$, $\{u_{k,2}\}_{k=1}^r$ and $\{u_{k,3}\}_{k=1}^r$. To do this we perform the SVD of a Gaussian random matrix three times and keep the r leading singular vectors, and then compute the outer-product

$$T = \sum_{k=1}^r u_{k,1} \otimes u_{k,2} \otimes u_{k,3}.$$

   The T produced in this way is guaranteed to have CP rank at most r. This can easily be extended to $N = 4$.

2. (Low Tucker rank) Generate $M_{d\times d\times d}$ with i.i.d. $\mathcal{N}(0,1)$ elements and then do approximate Tucker rank-r projection (successive low rank approximation of mode-1, mode-2 and mode-3 matricization) to get $T = \widehat{P}_{\Theta_3(r,r,r)}(M)$. The T produced in this way is guaranteed to have largest element of Tucker rank at most r. Once again this is easily extended to the $N = 4$ case.

3. (Sparse slices of low-rank matrices) In this case $N = 3$. Generate s slices of random rank-r matrices, (with eigenvalues all equal to one and random eigenvectors), and fill up the remaining $d - s$ slices with zero matrices to get $d \times d \times d$ tensor T. The T produced in this way is guaranteed to fall in $\Theta_2(r, s)$.

Then we generate covariates $\{X^{(i)}\}_{i=1}^n$ to be i.i.d random matrices filled with i.i.d $\mathcal{N}(0,1)$ entries. Finally, we simulate three GLM model, the Gaussian linear model, logistic regression and Poisson regression as follows.

1. (Gaussian linear model) We simulated noise $\{\epsilon^{(i)}\}_{i=1}^n$ independently from $\mathcal{N}(0, \sigma^2)$ and we vary σ^2. The noisy observation is then

$$Y^{(i)} = \langle X^{(i)}, T\rangle + \epsilon^{(i)}.$$

2. (Logistic regression) We simulated Binomial random variables:

$$Y^{(i)} \sim \text{Binomial}(m, p_i), \text{ where } p_i = \text{logit}(\alpha \cdot \langle X^{(i)}, T\rangle).$$

3. (Poisson regression) We simulated

$$Y^{(i)} \sim \text{Poisson}(\lambda_i), \text{ where } \lambda_i = m \exp(\alpha \cdot \langle X^{(i)}, T \rangle).$$

5.2. Convergence of PGD under restricted tensor regression

The first set of simulations investigates the convergence performance of PGD under various constraints and step-size choices for three different types of low-rankness. One of the important challenges when using the projected gradient descent algorithm is choosing the step-sizes (just like selecting the regularization parameter for convex regularization schemes) and the step-size choices stated in Theorem 2 depend on non-computable parameters (e.g. $c_u, c_\ell, ...$). In practice, the step is very important in that large step size would lead to divergence and small step size could cause slow convergence. We suggest either start with relatively large step size and continuously decrease the step size if divergence behavior is observed until tolerable step size is found, or start with relatively small step size that can guarantee convergence and gradually increase step size along the way while going back to small step size when divergence behavior is observed, which is similar to warm start strategy usually used to speed up convergence. In all our simualtions, the step-size η is set as a constant specified in each plot.

5.2.1. Third-order tensors

In the first two cases (see cases below), PGD with approximate projection $\widehat{P}_{\Theta_3(r',r',r')}$ were applied with different choices of (r', η) while in the third case the PGD with exact projection $P_{\Theta_2(r',s')}$ were adopted with different choices of (r', s', η).

Case 1a: (Gaussian) Low CP Rank with $p = 50^3$, $n = 4000$, $r = 5$, $\sigma = 0.5$ (SNR ≈ 4.5);

Case 2a: (Gaussian) Low Tucker Rank with $p = 50^3$, $n = 4000$, $r = 5$, $\sigma = 5$ (SNR ≈ 7.2);

Case 3a: (Gaussian) Slices of Low-rank Matrices with $p = 50^3$, $n = 4000$, $r = 5$, $s = 5$, $\sigma = 1$ (SNR ≈ 5.2).

Figures 1, 2 and 3 plot normalized rooted mean squared error (rmse) $\|\widehat{T} - T\|_{\text{F}} / \|T\|_{\text{F}}$ versus number of iterations, showing how fast rmse decreases as the number of iterations increases, under different (r', η) or (r', s', η). Notice that here we plot the average rmse with error bar to be the standard deviation for ten runs for Cases 1a, 2a and 3a.

Overall, the plots show the convergence of rmse's, and that the larger the r' or s' is, the greater the converged rmse will be, meaning that misspecification of rank/sparsity will do harm to the performance of PGD. In terms of the choice of step size, the experiments inform us that if η is too large, the algorithm may not converge and the range of tolerable step-size choices varies in different cases. In general, the more misspecified the constraint parameter(s) is(are), the lower the tolerance for step size will be. On the other hand, as we can see in all cases, given η under a certain tolerance level, the larger the η is, the faster the convergence will be.

5.2.2. Fourth-order tensors

Although we have focused on third order tensor for brevity, our method applies straightforwardly to higher order tensors. For illustration, we considered the following two examples which focus on estimating fourth order low rank tensors.

Case 4a: (Gaussian) Low CP Rank with $p = 20^4$, $n = 4000$, $r = 5$, $\sigma = 0.5$ (SNR $\approx$ 4.4);

Case 5a: (Gaussian) Low Tucker Rank with $p = 20^4$, $n = 4000$, $r = 5$, $\sigma = 5$ (SNR $\approx$ 7.4).

Figure 4 plots average rmse (with s.d. as error bar) for ten runs vs number of iterations for Case 4a and Case 5a using $\eta = 0.2$ under various choices of low-rankness constraint parameter r'. In general the convergence behavior for Case 4a and Case 5a are similar to those for Case 1a and Case 2a.

5.2.3. Logistic and Poisson regression

In the next set of simulations, we study the convergence behavior of the PGD applied to logistic and Poisson regression situation.

Case 1b: (Logistic) Low CP Rank with $p = 50^3$, $n = 4000$, $r = 5$, $m = 10$, $\alpha = 0.1$ (SNR $\approx$ 3.8);

Case 1c: (Poisson) Low CP Rank with $p = 30^3$, $n = 4000$, $r = 5$, $m = 5$, $\alpha = 0.5$ (SNR $\approx$ 4.8).

The results presented in Figures 5 and 6 exhibit similar pattern of convergence as in Figure 1. We observe also that in the case of low Tucker rank and sparse slices of low-rank matrices, logistic and Poisson regression have similar convergence behavior to least-squares regression. In general, a relaxed projection step is inferior to using the true rank parameter for projection. Once again as the step-size increases, the convergence will speed up until the step size becomes too large to guarantee convergence.

5.3. Comparison of non-convex PGD to convex regularization

Next, we compare the PGD method with convex regularization methods (implemented via `cvx`). In general, the `cvx` based regularization algorithm is significantly slower than the PGD method. This is partly due to the fact that the infrastructure of generic `cvx` is not tailored to solve the specific convex optimization problems. On the other hand, the PGD is much easier to implement and enjoys fast rates of convergence, which may also contribute to its improved performance in terms of run-time. Besides, `cvx` cannot handle p as large as those in Cases 1a, 2a and 3a. Hence, in order to do comparison in terms of the estimation error, we resort to moderate p so that `cvx` runs to completion. The simulation setup is as follows:

Case 6a: (Gaussian) Low CP Rank with $p = 10^3$, $n = 1000$, $r = 5$, $\sigma = 0.5, 1,$ or 2 (SNR $\approx$ 4.8, 2.4 or 1.2);

Case 7a: (Gaussian) Low Tucker Rank with $p = 10^3$, $n = 1000$, $r = 5$, $\sigma = 2.5, 5,$ or 10 (SNR $\approx$ 7.2, 3.6, or 1.8);

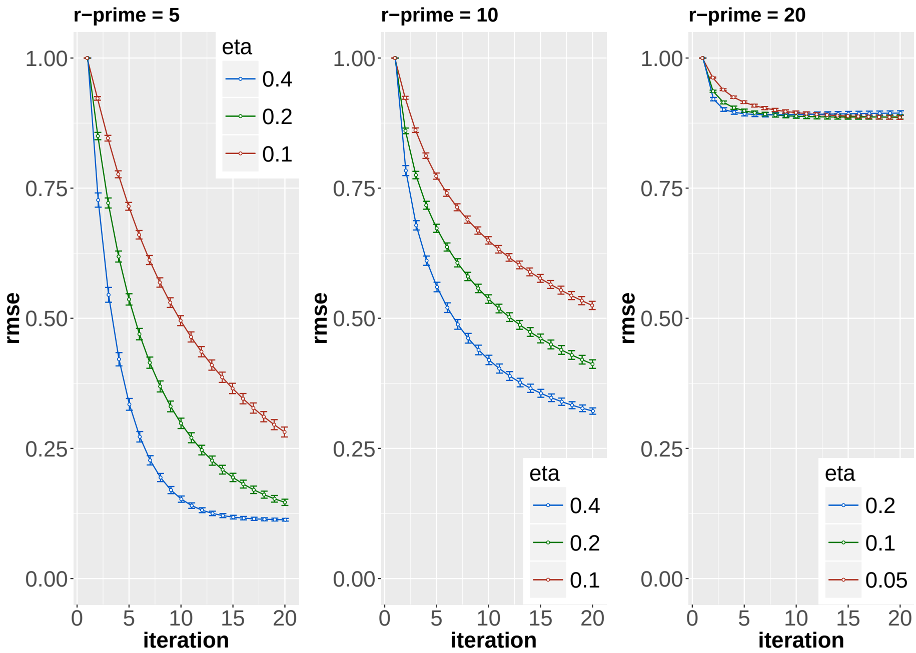

Figure 1: Case 1a: Low CP rank

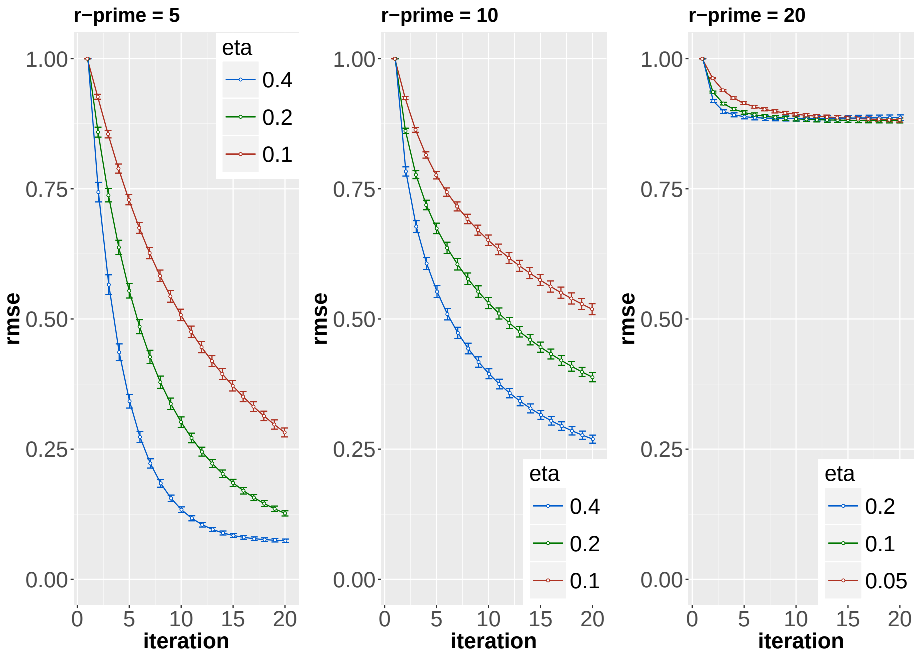

Figure 2: Case 2a: Low Tucker rank

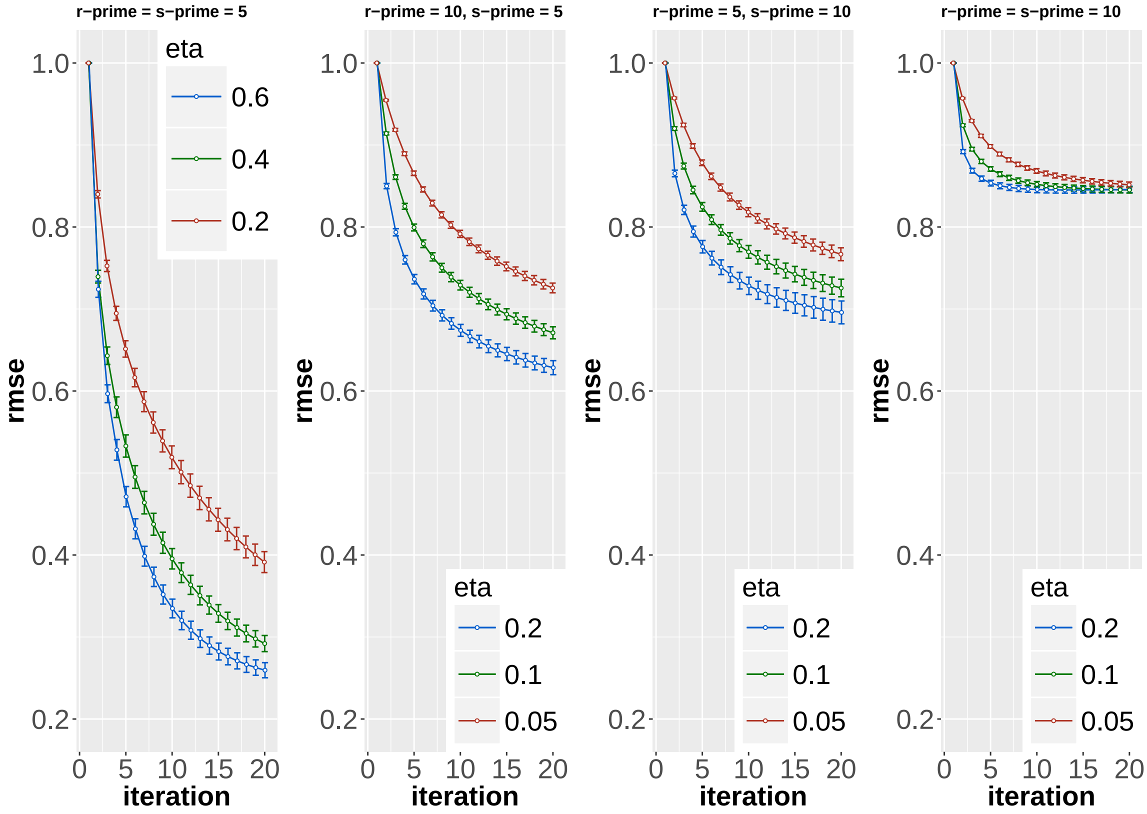

Figure 3: Case 3a: Sparse slices of low-rank matrices

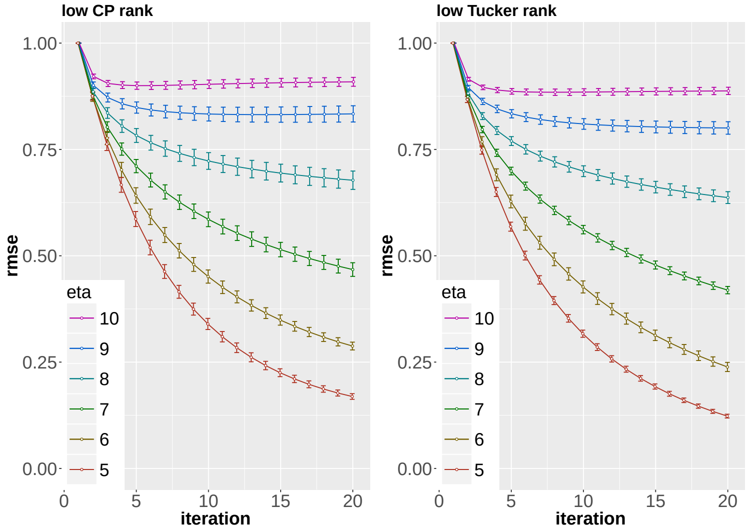

Figure 4: Case 4a, 5a: 4th order tensor

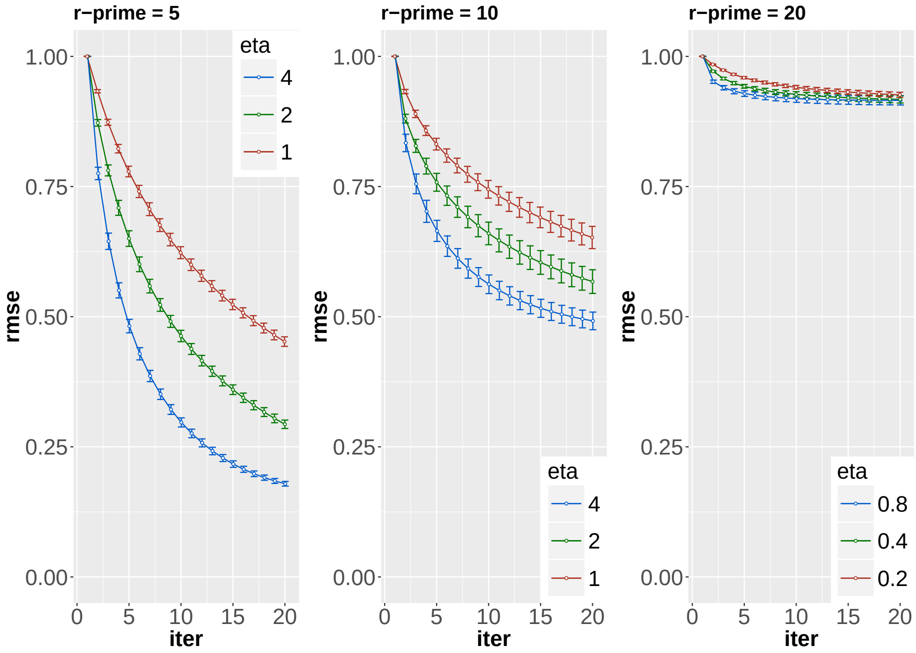

Figure 5: Case 1b: (Logistic) Low CP rank

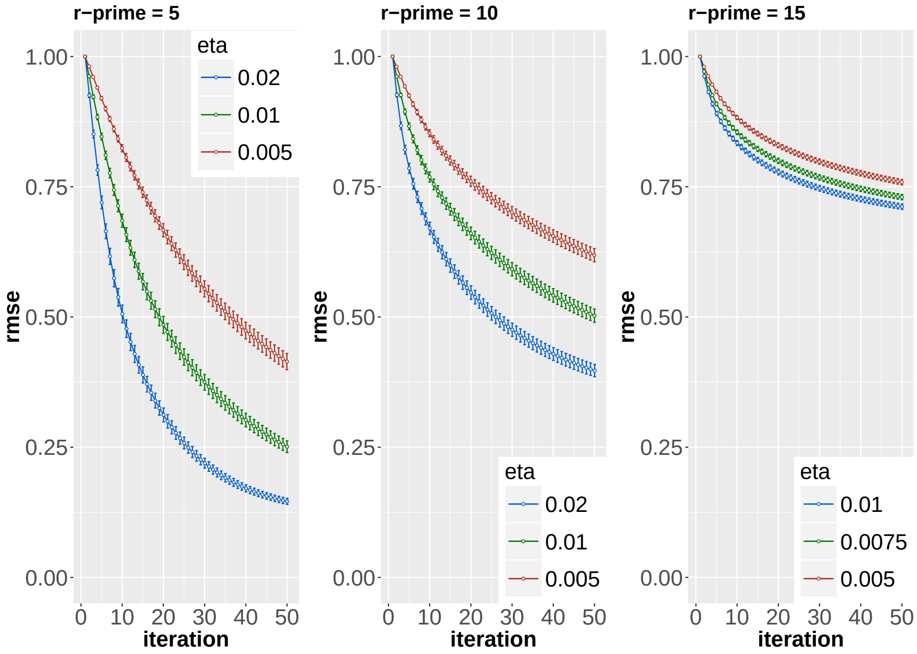

Figure 6: Case 1c: (Poisson) Low CP rank

Case 8a: (Gaussian) Slices of Low-rank Matrices with $p = 10^3$, $n = 1000$, $r = 5$, $s = 5$, $\sigma = 0.5, 1,$ or 2 (SNR $\approx$ 9.6, 4.8, or 2.4);

Case 6b: (Logistic) Low CP Rank with $p = 10^3$, $n = 1000$, $\alpha = 3.5$, $r = 5$, $m = 20, 5,$ or 1 (SNR $\approx$ 9.0, 4.5 or 2.0);

Case 7b: (Logistic) Low Tucker Rank with $p = 10^3$, $n = 1000$, $\alpha = 0.5$, $r = 5$, $m = 20, 5,$ or 1 (SNR $\approx$ 9.6 , 4.9 or 2.2);

Case 8b: (Logistic) Slices of Low-rank Matrices with $p = 10^3$, $n = 1000$, $\alpha = 1.2$, $r = 5$, $s = 5$, $m = 20, 5,$ or 1 (SNR $\approx$ 7.7 , 3.8 , or 1.7);

Case 6c: (Poisson) Low CP Rank with $p = 10^3$, $n = 1000$, $\alpha = 0.5$, $r = 5$, $m = 20, 5,$ or 1 (SNR $\approx$ 9.6 , 4.7, or 2.1);

Case 7c: (Poisson) Low Tucker Rank with $p = 10^3$, $n = 1000$, $\alpha = 0.06$, $r = 5$, $m = 20, 5,$ or 1 (SNR $\approx$ 9.0, 4.5 or 2.0);

Case 8c: (Poisson) Slices of Low-rank Matrices with $p = 10^3$, $n = 1000$, $\alpha = 0.25$, $r = 5$, $s = 5$, $m = 30, 10,$ or 5 (SNR $\approx$ 15.4, 8.8 or 6.2).

Cases 6(a,b,c), 7(a,b,c) and 8(a,b,c) were constructed to represent different types of tensor low-rankness structure in least-square, logistic and Poisson regression. In each case, three levels of SNR(high, moderate and low) are considered. For each setting, we simulated 50 groups of (T, ϵ, X) and run PGD and convex-regularization methods for the recovery of T to get average rmse with standard deviation for the two approaches respectively. Here we are comparing the best performance achieved by the PGD and convex regularization method respectively: for the PGD we use true parameter as the constraint parameter $r' = r$ (and $s' = s$); for convex regularization method, we do a grid search to choose the tuning parameter that yields the smallest rmse.

The results are summarized in Table 1. These results show that in general, the PGD method produces smaller rmse's than convex regularization methods regardless of the noise level of the data.

5.4. Comparison of tensor low-rankness approach and simple matricization approach

Finally, we display the value of considering low-rank tensors rather than a naive low-rank matricazation scheme. To begin, we consider the following low tensor ranks schemes.

Case 9a: (Gaussian) Low Tucker Rank with $p = 30^3$, $n = 4000$, $r = 5$, $\sigma = 5, 10,$ or 20 (SNR $\approx$ 5.9, 2.9 or 1.5);

Case 10a: (Gaussian) Low CP Rank with $p = 30^3$, $n = 4000$, $r = 5$, $\sigma = 0.5, 1, 2$ or (SNR $\approx$ 4.5, 2.2, or 1.1);

Case 11a: (Gaussian) Slices of Low-rank Matrices with $p = 30^3$, $n = 4000$, $r = 5$, $s = 5$, $\sigma = 1, 2,$ or 4 (SNR $\approx$ 5.0, 2.5 or 1.3).

rmse (sd)	SNR	PGD	Convex Regularization
Case 6a	High	**0.11 (0.01)**	0.28 (0.02)
	Moderate	**0.22 (0.01)**	0.47 (0.02)
	Low	**0.46 (0.03)**	0.69 (0.02)
Case 7a	High	**0.07 (0.01)**	0.18 (0.01)
	Moderate	**0.14 (0.01)**	0.32 (0.02)
	Low	**0.28 (0.02)**	0.51 (0.02)
Case 8a	High	**0.08 (0.01)**	0.12 (0.01)
	Moderate	**0.16 (0.01)**	0.23 (0.01)
	Low	**0.30 (0.01)**	0.41 (0.02)
Case 6b	High	**0.16 (0.01)**	0.44 (0.02)
	Moderate	**0.20 (0.01)**	0.54 (0.02)
	Low	**0.35 (0.02)**	0.66 (0.02)
Case 7b	High	**0.17 (0.01)**	0.46 (0.02)
	Moderate	**0.22 (0.01)**	0.55 (0.02)
	Low	**0.35 (0.01)**	0.67 (0.01)
Case 8b	High	**0.26 (0.01)**	0.37 (0.02)
	Moderate	**0.34 (0.02)**	0.50 (0.01)
	Low	**0.56 (0.04)**	0.68 (0.02)
Case 6c	High	**0.09 (0.01)**	0.57 (0.03)
	Moderate	**0.17 (0.01)**	0.61 (0.04)
	Low	**0.39 (0.04)**	0.71 (0.03)
Case 7c	High	**0.12 (0.01)**	0.74 (0.02)
	Moderate	**0.21 (0.02)**	0.75 (0.02)
	Low	**0.43 (0.06)**	0.80 (0.02)
Case 8c	High	**0.13 (0.01)**	0.79 (0.03)
	Moderate	**0.22 (0.03)**	0.81 (0.03)
	Low	**0.32 (0.03)**	0.83 (0.02)

Table 1: rmse of nonconvex PGD vs convex regularization

rmse (sd)	SNR	sequential low-rank projection	naive matricization
Case 9a	High	**0.22 (0.01)**	0.57 (0.01)
	Moderate	**0.24 (0.01)**	0.65 (0.01)
	Low	**0.31 (0.01)**	0.80 (0.01)
Case 10a	High	**0.10 (0.01)**	0.58 (0.01)
	Moderate	**0.35 (0.01)**	0.69 (0.01)
	Low	**0.35 (0.01)**	0.85 (0.01)
Case 11a	High	**0.15 (0.01)**	0.56 (0.01)
	Moderate	**0.29 (0.01)**	0.67 (0.01)
	Low	**0.57 (0.02)**	0.83 (0.01)

Table 2: rmse of approximate Tucker projection vs naive matricization

Case 9a and Case 10a is when the true tensor parameter has low Tucker rank or low CP rank. Case 11a is when the true tensor parameter is sparse slices of low-rank matrices. Three levels of SNR(high, moderate and low) are considered. For each setting, we simulated 50 groups of (T, ϵ, X) and run two approaches to recover T. The first approach we use is the PGD with our tensor low-rank (approximate) projection. In the Case 9a and 10a, tensor Tucker low-rankness is used via sequential matrix low rank projection, i.e. the approximate low Tucker rank projection $\hat{\Theta}_3(r', r', r')$ with $r' = 5$. While in Case 11a we exploit combination of group-sparsity and matrix low-rankness via $\Theta_2(r', s')$ with $r' = s' = 5$. The second approach is to naively use the matricization where we view the $p \times p \times p$ tensor simply as a $p \times p^2$ matrix and then impose low-rank structure. For naive matricization approach, we report the results of matrix rank-r' projection with $r' = 5$ which yields the best performance of its kind. Table 2 summarizes the average rmse with standard deviation of the two approaches for 50 random runs respectively. Hence we can see that if we partially discard the tensor structure by using only the low-rankness after matricization, the performance is greatly inferior to the low Tucker approximate projection approach. This applies to all three cases considered here, i.e. Tucker, CP and sparse slices of low-rank matrices.

6. Conclusion

In this paper, we provide a general framework that offers theoretical guarantees for learning high-dimensional tensor regression models under different low-rank structural assumptions using the PGD algorithm applied to a potentially non-convex constraint set Θ in terms of its *localized Gaussian width*. Our framework is the first general theory for PGD applied to tensor problems and given that the notion of low-rank structure is ambiguous for tensors, our general framework applies treats them in a unified way.

By providing statistical guarantees in terms of localized Gaussian width, we prove that the PGD approach has mean-squared error that is no worse than the convex regularization counter-part studied in Raskutti and Yuan (2015). We also provided three concrete examples Θ_1, Θ_2 and Θ_3 of low-dimensional tensor structure and provide implementable (approximate) projections and provide mean-squared error guarantees. For Θ_1 and Θ_2 we provide a convex regularization scheme that achieves the same rate, while for Θ_3 we show that the non-convex PGD approach achieves superior mean-squared error.

We supplement our theoretical results with simulations which show that, under several common settings of generalized low rank tensor regression, the projected gradient descent approach is superior both in terms of statistical error and run-time compared to convex approaches provided the step-sizes of the projected descent algorithm are suitably chosen. Additional simulation results show that PGD with tensor low-rankness constraint also outperforms naive matricization approaches.

7. Proofs

7.1. Proof of general results

We first prove the results of Section 3: Theorems 1, 2 and 3. In particular, we first provide a proof for Theorem 1. For convenience we first state the proof for the Gaussian case (Theorem 3) and then describe the necessary changes needed for the more general GLM case (Theorem 2).

7.1.1. Proof of Theorem 1

The proof follows very similar steps to those developed in Jain et al. (2014, 2016). Recall that $\widehat{T}_{k+1} = P_{\Theta(t_1)}(g_k)$ where $g_k = \widehat{T}_k - \eta\nabla f(\widehat{T}_k)$. For $\widehat{T}_{k+1} \in \Theta(t_1)$ and any $T \in \Theta(t_0 - t_1)$, the superadditivity condition guarantees that there exists a linear subspace $\mathcal{A} = \{\alpha_1 \widehat{T}_{k+1} + \alpha_2 T \ |\alpha_1, \alpha_2 \in \mathbb{R}\}$ such that $\widehat{T}_{k+1} \in \mathcal{A}$, $T \in \mathcal{A}$ and $\mathcal{A} \subset \Theta(t_0)$.

The contractive projection property CPP(δ) implies that for any $T \in \Theta(t_0 - t_1)$,

$$\|(\widehat{T}_{k+1} - g_k)_{\mathcal{A}}\|_{\mathrm{F}} \leq \delta \left\| \frac{t_0 - t_1}{t_1} \right\|_{\ell_\infty}^{1/2} \cdot \|(T - g_k)_{\mathcal{A}}\|_{\mathrm{F}}.$$

Since $t_1 = \left\lceil \frac{4\delta^2 C_u^2 C_l^{-2}}{1+4\delta^2 C_u^2 C_l^{-2}} \cdot t_0 \right\rceil$,

$$\delta \left\| \frac{t_0 - t_1}{t_1} \right\|_{\ell_\infty}^{1/2} \leq (2C_u C_l^{-1})^{-1}.$$

Hence,

$$\begin{aligned}
\|\widehat{T}_{k+1} - T\|_{\mathrm{F}} &\leq \|\widehat{T}_{k+1} - g_k\|_{\mathrm{F}} + \|T - g_k\|_{\mathrm{F}} \\
&\leq \left(1 + \delta \left\| \frac{t_0 - t_1}{t_1} \right\|_{\ell_\infty}^{1/2}\right) \|(T - g_k)_{\mathcal{A}}\|_{\mathrm{F}} \\
&\leq (1 + (2C_u C_l^{-1})^{-1}) \|(T - g_k)_{\mathcal{A}}\|_{\mathrm{F}} \\
&\leq \left(1 + \frac{C_l}{2C_u}\right) \|[T - \widehat{T}_k - \eta(\nabla f(T) - \nabla f(\widehat{T}_k))]_{\mathcal{A}}\|_{\mathrm{F}} + 2\eta\|[\nabla f(T)]_{\mathcal{A}}\|_{\mathrm{F}},
\end{aligned}$$

where the final inequality follows from the triangle inequality. If we define the Hessian matrix of the function f of a vectorized tensor as

$$H(A) = \nabla^2 f(A),$$

the Mean Value Theorem implies that

$$\mathrm{vec}(\nabla f(T) - \nabla f(\widehat{T}_k)) = H(\widehat{T}_k + \alpha(T - \widehat{T}_k)) \cdot (T - \widehat{T}_k),$$

for some $0 < \alpha < 1$, and

$$\|\widehat{T}_{k+1} - T\|_{\mathrm{F}} \leq (1 + (2C_u C_l^{-1})^{-1}) \|[(I - \eta H(\widehat{T}_k + \alpha(T - \widehat{T}_k)))\mathrm{vec}(\widehat{T}_k - T)]_{\mathrm{vec}(\mathcal{A})}\|_{\ell_2} \\ +2\eta \|[\nabla f(T)]_{\mathcal{A}}\|_{\mathrm{F}}.$$

We now appeal to the following lemma:

Lemma 4 *Suppose $\mathcal{S}$ is a linear subspace of $\mathbb{R}^d$, and H is an $d \times d$ positive semidefinite matrix. For any given $0 < c < 1$, if for any $x \in \mathcal{S}$,*

$$c x^\top x \leq x^\top H x \leq (2 - c) x^\top x, \tag{12}$$

then for any $z \in \mathcal{S}$, we have

$$\|[(I - H)z]_{\mathcal{S}}\|_{\ell_2} \leq (1 - c)\|z\|_{\ell_2},$$

$(\cdot)_{\mathcal{S}}$ stands for the projection onto the subspace $\mathcal{S}$.

Proof [Proof of Lemma 4] Suppose the orthonomal basis of $\mathcal{S}$ is $e_1 \ldots, e_q$, and then

$$\mathbb{R}^d = \{ce_1 | c \in \mathbb{R}\} \oplus \ldots \oplus \{ce_q | c \in \mathbb{R}\} \oplus \mathcal{S}^\perp$$

For positive semidefinite H, it can be decomposed as follows

$$H = D^\top D.$$

Hence we can decompose the rows of D to get

$$D = \sum_{i=1}^q \lambda_i e_i^\top + (y_1, \ldots, y_n)^\top$$

where $y_1, \ldots, y_n \in \mathcal{S}^\perp$, and $\lambda_i \in \mathbb{R}^n$ for $i = 1, \ldots, q$. Therefore,

$$D^\top D = \sum_{i=1}^q \sum_{j=1}^q (\lambda_i^\top \lambda_j) e_i e_j^\top + \sum_{k=1}^n y_k y_k^\top + (y_1, \ldots, y_n) \sum_{i=1}^q \lambda_i e_i^\top + \sum_{i=1}^q e_i \lambda_i^\top \cdot (y_1, \ldots, y_n)^\top.$$

Now for any $(\alpha_1, \ldots, \alpha_q)^\top \in \mathbb{R}^q$, we have $x = \sum_{i=1}^q \alpha_i e_i \in \mathcal{S}$, and hence

$$x^\top D^\top D x = (\sum_{i=1}^q \alpha_i e_i)^\top D^\top D (\sum_{i=1}^q \alpha_i e_i) = \sum_{i=1}^q \sum_{j=1}^q (\lambda_i^\top \lambda_j) \alpha_i \alpha_j.$$

The equation 12 then implies that the matrix

$$\Lambda = \{\Lambda_{i,j}\}_{i,j=1}^q \text{ where } \Lambda_{i,j} = \lambda_i^\top \lambda_j$$

has eigenvalues bounded by c from below and $2-c$ from above. Next, notice that for any $z \in \mathcal{S}$, we have $z = \sum_{i=1}^{q} \beta_i e_i$ for some $(\beta_1, \dots, \beta_q)^\top \in \mathbb{R}^q$, and hence due to the fact that $y_1, \dots y_n \in \mathcal{S}^\perp$

$$(I - D^\top D)z = \left(I - \sum_{i=1}^{q} \sum_{j=1}^{q} (\lambda_i^\top \lambda_j) e_i e_j^\top + (y_1, \dots, y_n) \sum_{i=1}^{q} \lambda_i e_i^\top \right) \sum_{i=1}^{q} \beta_i e_i,$$

and furthermore

$$\begin{aligned} [(I - D^\top D)z]_{\mathcal{S}} &= \left(I - \sum_{i=1}^{q} \sum_{j=1}^{q} (\lambda_i^\top \lambda_j) e_i e_j^\top \right) \sum_{i=1}^{q} \beta_i e_i \\ &= (e_1, \dots, e_q)(I_{q\times q} - \Lambda)(\beta_1, \dots \beta_q)^\top, \end{aligned}$$

and

$$\|[(I - D^\top D)z]_{\mathcal{S}}\|_{\ell_2}^2 = (\beta_1, \dots \beta_q)(I_{q\times q} - \Lambda)(\beta_1, \dots \beta_q)^\top,$$

which completes the proof. ■

Condition $RSCS(\Theta(t_0), C_l, C_u)$ guarantees the condition of Lemma 4 is satisfied with $H = \eta H(\widehat{T}_k + \alpha(T - \widehat{T}_k))$, $c = (C_u C_l^{-1})^{-1}$ and $\mathcal{S} = \mathcal{A}$. Hence Lemma 4 implies that

$$\|\widehat{T}_{k+1} - T\|_{\mathrm{F}} \quad \leq \quad (1 + (2C_u C_l^{-1})^{-1}) \left(1 - (C_u C_l^{-1})^{-1}\right) \|\widehat{T}_k - T\|_{\mathrm{F}} + 2\eta \|[\nabla f(T)]_{\mathcal{A}}\|_{\mathrm{F}}.$$

Therefore for any k,

$$\|\widehat{T}_{k+1} - T\|_{\mathrm{F}} \leq (1 - (2C_u C_l^{-1})^{-1}) \|\widehat{T}_k - T\|_{\mathrm{F}} + 2\eta Q,$$

where

$$Q \quad = \quad \sup_{\mathcal{A}_0 \subset \Theta(t_0)} \|(\nabla f(T))_{\mathcal{A}_0}\|_{\mathrm{F}},$$

where $\mathcal{A}_0$ is any linear subspace of $\Theta(t_0)$.We then appeal to the following result.

Lemma 5 *Suppose $\mathcal{A}$ is a linear subspace of tensor space Ω. For any $L \in \Omega$,*

$$\|(L)_{\mathcal{A}}\|_{\mathrm{F}} = \sup_{A \in \mathcal{A} \cap \mathbb{B}_{\mathrm{F}}(1)} \langle A, L \rangle$$

Proof [Proof of Lemma 5] First, we are going to show that

$$\|(L)_{\mathcal{A}}\|_{\mathrm{F}} \leq \sup_{A \in \mathcal{A} \cap \mathbb{B}_{\mathrm{F}}(1)} \langle A, L \rangle$$

Suppose we have $(L)_{\mathcal{A}} = P \in \mathcal{A}$. Since for any $\alpha > -1$, $P + \alpha P \in \mathcal{A}$, and hence

$$\|P + \alpha P - L\|_{\mathrm{F}} = \|P - L\|_{\mathrm{F}} + \alpha^2 \|P\|_{\mathrm{F}} + \alpha \langle P, P - L \rangle \leq \|P - L\|_{\mathrm{F}}$$

we must have $\langle P, P - L\rangle = 0$, i.e. $\langle P, L\rangle = \langle P, P\rangle$. (otherwise α of small magnitude with the same sign of $\langle P, P - L\rangle$ will violate the inequality). Therefore,

$$\sup_{A\in\mathcal{A}\cap\mathbb{B}_{\mathrm{F}}(1)} \langle A, L\rangle \geq \left\langle \frac{P}{\|P\|_{\mathrm{F}}}, L\right\rangle = \|P\|_{\mathrm{F}} = \|(L)_{\mathcal{A}}\|_{\mathrm{F}}$$

What remains is to show

$$\|(L)_{\mathcal{A}}\|_{\mathrm{F}} \geq \sup_{A\in\mathcal{A}\cap\mathbb{B}_{\mathrm{F}}(1)} \langle A, L\rangle$$

For any $D \in \mathcal{A}\cap\mathbb{B}_{\mathrm{F}}(1)$, let D_α be the projection of L onto $\{\alpha D|\alpha \geq 0\}$, and hence

$$\langle D_\alpha, L\rangle = \langle D_\alpha, D_\alpha\rangle \leq \langle P, P\rangle.$$

Therefore, we have

$$\langle D, L\rangle \leq \left\langle \frac{D}{\|D\|_{\mathrm{F}}}, L\right\rangle = \left\langle \frac{D_\alpha}{\|D_\alpha\|_{\mathrm{F}}}, L\right\rangle \leq \|P\|_{\mathrm{F}}$$

which completes the proof. ■

Lemma 5 then implies

$$Q = \sup_{\mathcal{A}_0\subset\Theta(t_0)} \|(\nabla f(T))_{\mathcal{A}_0}\|_{\mathrm{F}} = \sup_{\mathcal{A}_0\subset\Theta(t_0)} \sup_{A\in\mathcal{A}_0\cap\mathbb{B}_{\mathrm{F}}(1)} \langle \nabla f(T), A\rangle \leq \sup_{A\in\Theta(t_0)\cap\mathbb{B}_{\mathrm{F}}(1)} \langle \nabla f(T), A\rangle\,.$$

Therefore, after

$$K = \lceil 2C_u C_l^{-1} \log\frac{\|T\|_{\mathrm{F}}}{\epsilon}\rceil$$

iterations,

$$\|\widehat{T}_k - T\|_{\mathrm{F}} \leq \epsilon + 4\eta C_u C_l^{-1} \sup_{A\in\Theta(t_0)\cap\mathbb{B}_{\mathrm{F}}(1)} \langle \nabla f(T), A\rangle\,,$$

which completes the proof.

7.1.2. Proof of Theorem 3

Recall that the original least-squares objective is

$$f(A) = \frac{1}{2n}\sum_{i=1}^{n}(Y^{(i)} - \langle X^{(i)}, A\rangle)^2.$$

Hence, the gradient at true tensor coefficient T is :

$$\nabla f(T) = \frac{1}{n}\sum_{i=1}^{n} X^{(i)} \otimes [\langle T, X^{(i)}\rangle - Y^{(i)}] = -\frac{1}{n}\sum_{i=1}^{n} X^{(i)}\zeta^{(i)}$$

for the least-squares objective we consider. Further

$$\nabla^2 f(T) = \frac{1}{n}\sum_{i=1}^{n} X^{(i)} \otimes X^{(i)}.$$

Through vectorization, the Hessian matrix H is

$$H = \sum_{i=1}^{n} \text{vec}(X^{(i)})\text{vec}(X^{(i)})^\top.$$

Lemma 1 then implies that given $n \geq c_1 w_G^2[\Theta(t_0) \cap \mathbb{B}_\text{F}(1)]$, with probability at least

$$1 - c_2/2 \exp(-c_3 w_G^2[\Theta(t_0) \cap \mathbb{B}_\text{F}(1)]),$$

we have, for any $A \in \Theta(t_0)$,

$$\left(\tau^{-1}c_l\right)^2 \langle A, A\rangle \leq \frac{1}{n}\sum_{i=1}^{n} \langle X^{(i)}, A\rangle^2 \leq (\tau c_u)^2 \langle A, A\rangle,$$

which guarantees the $RSCS(\Theta(t_0), C_l, C_u)$ condition with $C_u = \tau c_u$ and $C_l = \tau^{-1}c_l$. Thus Theorem 1 implies that

$$\|\widehat{T}_k - T\|_\text{F} \leq \epsilon + 4\eta\tau^2\kappa \sup_{A\in\Theta(t_0)\cap\mathbb{B}_\text{F}(1)} \left\langle \frac{1}{n}\sum_{i=1}^{n} \zeta^{(i)} X^{(i)}, A \right\rangle.$$

The last step is to show that

$$\sup_{A\in\Theta(t_0)\cap\mathbb{B}_\text{F}(1)} \left\langle \frac{1}{n}\sum_{i=1}^{n} \zeta^{(i)} X^{(i)}, A \right\rangle \leq 2c_u\sigma n^{-1/2} w_G[\Theta(t_0) \cap \mathbb{B}_\text{F}(1)],$$

with probability at least

$$1 - c_2/2 \exp\left\{-c_3 w_G^2[\Theta(t_0) \cap \mathbb{B}_\text{F}(1)]\right\}.$$

This can be shown by simply applying Lemma 11 in Raskutti and Yuan (2015) and replacing $\{A|\mathcal{R}(A) \leq 1\}$ with $\Theta(t_0) \cap \mathbb{B}_\text{F}(1)$. Note that all the proof steps for Lemma 11 in Raskutti and Yuan (2015) are identical for $\Theta(t_0) \cap \mathbb{B}_\text{F}(1)$ since the sets $\Theta(t)$'s are symmetric.

7.1.3. Proof of Theorem 2

The proof follows the same flow as Theorem 3 but we requires an important concentration result from Mendelson (2014). Recall that in the GLM setting, according to (2),

$$f(A) = \frac{1}{n}\sum_{i=1}^{n} (a(\langle X^{(i)}, A\rangle) - Y^{(i)}\langle X^{(i)}, A\rangle).$$

Hence the gradient at true coefficient T is

$$\nabla f(T) = \frac{1}{n}\sum_{i=1}^{n} (\mu_i - Y^{(i)})X^{(i)},$$

where $\mu_i = a'(\langle X^{(i)}, T\rangle)$, and the Hessian matrix at vectorized tensor T is

$$\nabla^2 f(T) = \sum_{i=1}^{n} W_{ii}\text{vec}(X^{(i)})\text{vec}(X^{(i)})^\top.$$

where $W_{ii} = a''(\langle X^{(i)}, T\rangle)$.

Since $\mathrm{Var}(Y^{(i)}) = a''(\langle X^{(i)}, T\rangle) = W_{ii}$, the moment assumption $1/\tau_0^2 \leq \mathrm{Var}(Y^{(i)}) \leq \tau_0^2$ guarantees that

$$\frac{1}{\tau_0^2} \leq W_{ii} \leq \tau_0^2.$$

Plus, for any $\tau > \tau_0$, Lemma 1 guarantees that when $n > c_1 w_G^2[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)]$,

$$\left((\tau/\tau_0)^{-1} c_l\right)^2 \langle A, A\rangle \leq \frac{1}{n}\sum_{i=1}^n \langle X^{(i)}, A\rangle^2 \leq ((\tau/\tau_0)c_u)^2 \langle A, A\rangle.$$

Therefore $RSCS(\Theta(t_0), C_l, C_u)$ condition holds with $C_l = \tau^{-1}c_l$ and $C_u = \tau c_u$. Thus Theorem 1 implies that

$$\|\widehat{T}_k - T\|_{\mathrm{F}} \leq \epsilon + 4\eta\tau^2\kappa \sup_{A\in\Theta(t_0)\cap\mathbb{B}_{\mathrm{F}}(1)} \left\langle \frac{1}{n}\sum_{i=1}^n (Y^{(i)} - \mu_i)X^{(i)}, A \right\rangle.$$

For the last step, by applying a concentration result on the following multiplier empirical process

$$\sup_{A\in\Theta(t_0)\cap\mathbb{B}_{\mathrm{F}}(1)} \left\langle \frac{1}{n}\sum_{i=1}^n (Y^{(i)} - \mu_i)X^{(i)}, A \right\rangle = \sup_{A\in\Theta(t_0)\cap\mathbb{B}_{\mathrm{F}}(1)} \frac{1}{n}\sum_{i=1}^n (Y^{(i)} - \mu_i)\left\langle X^{(i)}, A\right\rangle,$$

we can bound the quantity by the Gaussian width with large probability, up to some constant.

More specifically, denote

$$\omega^{(i)} = \Sigma^{-1/2}\mathrm{vec}(X^{(i)})$$

then $\{\omega^{(i)}\}_{i=1}^n$ are i.i.d Gaussian random vectors and hence

$$\begin{aligned}
& \sup_{A\in\Theta(t_0)\cap\mathbb{B}_{\mathrm{F}}(1)} \frac{1}{n}\sum_{i=1}^n (Y^{(i)} - \mu_i)\left\langle X^{(i)}, A\right\rangle \\
= & \sup_{F\in\mathrm{vec}(\Theta(t_0)\cap\mathbb{B}_{\mathrm{F}}(1))} \frac{1}{n}\sum_{i=1}^n (Y^{(i)} - \mu_i)\mathrm{vec}(X^{(i)})^\top F \\
= & \sup_{F\in\Sigma^{1/2}\cdot\mathrm{vec}(\Theta(t_0)\cap\mathbb{B}_{\mathrm{F}}(1))} \frac{1}{n}\sum_{i=1}^n (Y^{(i)} - \mu_i)(\omega^{(i)})^\top F \\
\leq & \; c_5 M_Y^{1/q} \sup_{F\in\Sigma^{1/2}\cdot\mathrm{vec}(\Theta(t_0)\cap\mathbb{B}_{\mathrm{F}}(1))} \frac{1}{n}\sum_{i=1}^n (\omega^{(i)})^\top F \\
= & \; c_5 M_Y^{1/q} \sup_{A\in\Theta(t_0)\cap\mathbb{B}_{\mathrm{F}}(1)} \frac{1}{n}\sum_{i=1}^n \left\langle X^{(i)}, A\right\rangle \\
\leq & \; \frac{c_5 M_Y^{1/q} c_u}{\sqrt{n}} w_G^2[\Theta(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)],
\end{aligned}$$

where the first inequality follows from Theorem 1.9 of Mendelson (2014), and the second inequality holds in view of Lemma 11 of Raskutti and Yuan (2015), and both inequalities hold with probability greater than

$$1 - c_2\exp\left\{-c_3 w_G^2[\Theta(\theta' + \theta) \cap \mathbb{B}_{\mathrm{F}}(1)]\right\} - c_4 n^{-(q/2-1)}\log^q n.$$

7.2. Proofs of results in Section 4

We now present the proofs for the two main examples $\Theta_2(r, s)$ and $\Theta_3(r)$. Our proofs involve: (i) proving that the projection properties hold for both sets of cones and (ii) finding an error bound for the Gaussian width $w_G[\Theta(t) \cap \mathbb{B}_{\mathrm{F}}(1)]$.

7.2.1. Proof of Theorem 4

First, it is straightforward to verify that $\{\Theta_2(r, s)\}$ is a superadditive family of symmetric cones. We then verify the contraction properties as stated by Lemma 2.
Proof [Proof of Lemma 2] We need to develop an error bound for $\|P_{\Theta_2(r_2,s_2)}(Z) - Z\|_{\mathrm{F}}$ for a general tensor $Z \in \Theta_2(r_0, s_0)$. Let $\tilde{Z} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ denote the tensor whose slices $\tilde{Z}_{\cdot\cdot j_3}$, ($j_3 = 1, \dots, d_3$) are the rank-$r_2$ approximation of the corresponding slices of Z. First, it follows from the contraction property of low rank matrix projector (see, e.g., Jain et al., 2016) that for all $1 \le j_3 \le d_3$, for any $Y_{\cdot\cdot j_3}$ such that $\mathrm{rank}(Y_{\cdot\cdot j_3}) \le r_1$

$$\|\tilde{Z}_{\cdot\cdot j_3} - Z_{\cdot\cdot j_3}\|_{\mathrm{F}} \le \beta \|Y_{\cdot\cdot j_3} - Z_{\cdot\cdot j_3}\|_{\mathrm{F}}.$$

By summing over j_3 it follows that for any $Y \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ where $\mathrm{rank}(Y_{\cdot\cdot j_3}) \le r_1$ for all j_3

$$\|\tilde{Z} - Z\|_{\mathrm{F}} \le \beta \|Y - Z\|_{\mathrm{F}}.$$

The projection $P_{\Theta_2(r_2,s_2)}(Z)$ involves zeroing out the slices of $\tilde{Z}$ with the smallest magnitude. Let $v_{\tilde{Z}} := \mathrm{vec}(\|\tilde{Z}_{\cdot\cdot 1}\|_{\mathrm{F}}, \|\tilde{Z}_{\cdot\cdot 2}\|_{\mathrm{F}}, \dots, \|\tilde{Z}_{\cdot\cdot d_3}\|_{\mathrm{F}})$. As shown by Jain et al. (2014), for all Y where $v_Y = \mathrm{vec}(\|Y_{\cdot\cdot 1}\|_{\mathrm{F}}, \|Y_{\cdot\cdot 2}\|_{\mathrm{F}}, \dots, \|Y_{\cdot\cdot d_3}\|_{\mathrm{F}})$ and $\|v_Y\|_{\ell_0} \le s_1$,

$$\|\tilde{P}_s(v_{\tilde{Z}}) - v_{\tilde{Z}}\|_{\ell_2} \le \alpha \|v_Y - v_{\tilde{Z}}\|_{\ell_2}.$$

Therefore

$$\begin{aligned} \|P_{\Theta_2(r_2,s_2)}(Z) - \tilde{Z}\|_{\mathrm{F}} &\le \alpha \|Y - \tilde{Z}\|_{\mathrm{F}} \\ &\le \alpha(\|Y - Z\|_{\mathrm{F}} + \|\tilde{Z} - Z\|_{\mathrm{F}}) \\ &\le (\alpha + \alpha\beta)\|Y - Z\|_{\mathrm{F}}. \end{aligned}$$

Hence using the triangle inequality:

$$\|P_{\Theta_2(r_2,s_2)}(Z) - Z\|_{\mathrm{F}} \le \|\tilde{Z} - Z\|_{\mathrm{F}} + \|P_{\Theta_2(r_2,s_2)}(Z) - \tilde{Z}\|_{\mathrm{F}} \le (\alpha + \beta + \alpha\beta) \cdot \|Y - Z\|_{\mathrm{F}},$$

which completes the proof. ■

Lemma 2 guarantees that $P_{\Theta_2(r,s)}$ satisfies the contractive projection property CPP(δ) with $\delta = 3$. Hence, by setting $t_1 = (r', s')$ and $t_0 = (r' + r, s' + s)$, Theorem 3 directly implies the linear convergence rate result with statistical error of order

$$n^{-1/2} w_G[\Theta_2(r' + r, s' + s) \cap \mathbb{B}_{\mathrm{F}}(1)].$$

It remains to calibrate the Gaussian width. Recall the definition of the convex regularizer:

$$\mathcal{R}_1(A) = \sum_{j_3=1}^{d_3} \|A_{\cdot\cdot j_3}\|_*.$$

It is straightforward to show that

$$\Theta_2(r'+r, s'+s) \cap \mathbb{B}_{\mathrm{F}}(1) \subset \mathbb{B}_{\mathcal{R}_1}(\sqrt{(r'+r)(s'+s)}).$$

Then Lemma 5 of Raskutti and Yuan (2015) implies that

$$\begin{aligned} w_G[\Theta_2(r'+r, s'+s) \cap \mathbb{B}_{\mathrm{F}}(1)] &\leq w_G[\mathbb{B}_{\mathcal{R}_1}(\sqrt{(r'+r)(s'+s)})] \\ &\leq \sqrt{(s'+s)(r'+r)}w_G[\mathbb{B}_{\mathcal{R}_1}(1)] \\ &\leq \sqrt{(s'+s)(r'+r)}\sqrt{6(d_1+d_2+\log d_3)} \end{aligned}$$

which completes the proof.

7.2.2. Proof of Theorem 5

Once again, it is straightforward to verify that $\{\Theta_3(r_1, r_2, r_3)\}$ is a superadditive family of symmetric cones. We now verify the contraction properties
Proof [Proof of Lemma 3] To develop an error bound for $\|\widehat{P}_{\Theta_3(r_1^{(2)}, r_2^{(2)}, r_3^{(2)})}(Z) - Z\|_{\mathrm{F}}$ for a general tensor $Z \in \Theta_3(r_1^{(0)}, r_2^{(0)}, r_3^{(0)})$, we introduce the following three tensors (recursively):

$$\begin{aligned} Z_{(1)} &:= (\mathcal{M}_1^{-1} \circ \bar{P}_{r_1} \circ \mathcal{M}_1)(Z) \\ Z_{(2)} &:= (\mathcal{M}_2^{-1} \circ \bar{P}_{r_2} \circ \mathcal{M}_2)(Z_{(1)}) \\ Z_{(3)} &:= (\mathcal{M}_3^{-1} \circ \bar{P}_{r_3} \circ \mathcal{M}_3)(Z_{(2)}), \end{aligned}$$

where we recall that $\mathcal{M}_1(\cdot)$, $\mathcal{M}_2(\cdot)$ and $\mathcal{M}_3(\cdot)$ are the mode-1, mode-2 and mode-3 matricization operators. Therefore $\widehat{P}_{\Theta_3(r_1^{(2)}, r_2^{(2)}, r_3^{(2)})}(Z) = Z_{(3)}$ and:

$$\|\widehat{P}_{\Theta_3(r_1^{(2)}, r_2^{(2)}, r_3^{(2)})}(Z) - Z\|_{\mathrm{F}} \leq \|\widehat{P}_{\Theta_3(r_1^{(2)}, r_2^{(2)}, r_3^{(2)})}(Z) - Z_{(2)}\|_{\mathrm{F}} + \|Z_{(2)} - Z_{(1)}\|_{\mathrm{F}} + \|Z_{(1)} - Z\|_{\mathrm{F}}.$$

Next note that

$$\|Z_{(1)} - Z\|_{\mathrm{F}} = \|(\mathcal{M}_1^{-1} \circ \bar{P}_{r_1} \circ \mathcal{M}_1)(Z) - Z\|_{\mathrm{F}} = \|\bar{P}_{r_1}(\mathcal{M}_1(Z)) - \mathcal{M}_1(Z)\|_{\mathrm{F}}.$$

As shown by Jain et al. (2016), for any Y such that $r_1(Y) \leq r_1^{(1)}$,

$$\|Z_{(1)} - Z\|_{\mathrm{F}} \leq \beta_1 \|Y - Z\|_{\mathrm{F}}.$$

Using a similar argument and the triangle inequality

$$\|Z_{(2)} - Z_{(1)}\|_{\mathrm{F}} \leq \beta_2 \|Y - Z_{(1)}\|_{\mathrm{F}} \leq \beta_2(\|Y - Z\|_{\mathrm{F}} + \|Z_{(1)} - Z\|_{\mathrm{F}}) \leq (\beta_2 + \beta_1\beta_2)\|Y - Z\|_{\mathrm{F}}.$$

Furthermore,

$$\begin{aligned} \|Z_{(3)} - Z_{(2)}\|_{\mathrm{F}} &\leq \beta_3 \|Y - Z_{(2)}\|_{\mathrm{F}} \\ &\leq \beta_3(\|Y - Z\|_{\mathrm{F}} + \|Z_{(2)} - Z\|_{\mathrm{F}}) \\ &\leq \beta_3(1 + \beta_1 + \beta_2 + \beta_1\beta_2)\|Y - Z\|_{\mathrm{F}}. \end{aligned}$$

Therefore for all $Y \in \Theta_3(r_1^{(1)}, r_2^{(1)}, r_3^{(1)})$

$$\begin{aligned}\|\widehat{P}_{\Theta_3(r_1^{(2)},r_2^{(2)},r_3^{(2)})}(Z) - Z\|_{\mathrm{F}} &= \|\widehat{P}_{\Theta_3(r_1^{(2)},r_2^{(2)},r_3^{(2)})}(Z) - Z_{(2)}\|_{\mathrm{F}} + \|Z_{(2)} - Z_{(1)}\|_{\mathrm{F}} + \|Z_{(1)} - Z\|_{\mathrm{F}} \\ &\leq [(\beta_1+1)(\beta_2+1)(\beta_3+1) - 1]\|Y - Z\|_{\mathrm{F}}.\end{aligned}$$

■

Lemma 2 guarantees the approximate projection $\widehat{P}_{\Theta_3(r_1,r_2,r_3)}$ fulfills the contractive projection property CPP(δ) with $\delta = 7$. And hence via setting $t_1 = (r_1', r_2', r_3')$ and $t_0 = (r_1' + r_1, r_2' + r_2, r_3' + r_3)$, Theorem 3 directly implies the linear convergence rate result with statistical error of order $n^{-1/2} w_G[\Theta_3(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)]$. To upper bound the Gaussian width, we define the following nuclear norms:

$$\mathcal{R}_{(i)}(A) = \|\mathcal{M}_i(A)\|_*,$$

where $1 \leq i \leq 3$ and $\|.\|_*$ is the standard nuclear norm. Then it clearly follows that

$$\Theta_3(t_0) \cap \mathbb{B}_{\mathrm{F}}(1) \subset \cap_{i=1}^3 \mathbb{B}_{\mathcal{R}_{(i)}}(\sqrt{r_i' + r_i}).$$

Lemma 5 in Raskutti and Yuan (2015) then implies that

$$\begin{aligned} w_G[\Theta_3(t_0) \cap \mathbb{B}_{\mathrm{F}}(1)] &\leq w_G[\cap_i \mathbb{B}_{\mathcal{R}_{(i)}}(\sqrt{r_i' + r_i})] \\ &\leq \min_i w_G[\mathbb{B}_{\mathcal{R}_{(i)}}(\sqrt{r_i' + r_i})] \\ &\leq \min_i \sqrt{r_i' + r_i} w_G[\mathbb{B}_{\mathcal{R}_{(i)}}(1)] \\ &\leq \sqrt{6 \min\{(r_1' + r_1)(d_1 + d_2 d_3), (r_2' + r_2)(d_2 + d_1 d_3), (r_3' + r_3)(d_3 + d_1 d_2)\}} \end{aligned}$$

which completes the proof.

Acknowledgement. Han Chen and Ming Yuan was supported in part by NSF FRG Grant DMS-1265202, and NIH Grant 1-U54AI117924-01. Ming Yuan was also supported by NSF Grant DMS-1803450 Garvesh Raskutti was supported in part by NSF Grants DMS-1407028, DMS-1811767, NIH Grant R01 GM131381-01, ARO Grant W911NF-17-1-0357 and NGA Grant HM0476-17-1-2003.

References

Arindam Banerjee, Sheng Chen, Farideh Fazayeli, and Vidyashankar Sivakumar. Estimation with norm regularization. Technical Report arXiv:1505.02294, November 2015.

P. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

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

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

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex algebraic geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12: 805–849, 2012.

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

J. Fan and R.Z. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, (32):407–499, 2001.

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.

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. Jain, A. Tewari, A. Nanopoulos, and P. Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Proceedings of NIPS*, 2014.

P. Jain, N. Rao, and I. Dhillon. Structured sparse regression via greedy hard-thresholding. Technical Report arXiv:1602.06042, February 2016.

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

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.

P. Loh and M. J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 15:559-616, 2015.

Shahar Mendelson. Upper bounds on product and multiplier empirical processes. Technical Report arXiv:1410.8003, Technion, I.I.T, 2014.

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, 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. Raskutti and M. Yuan. Convex regularization for high-dimensional tensor regression. Technical Report arXiv:1512.01215, University of Wisconsin-Madison, December 2015.

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.

R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima. Statistical performance of convex tensor decomposition. In *Advances in Neural Information Processing Systems*, pages 972–980, 2013.

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