Square Deal: Lower Bounds and Improved Relaxations for Tensor Recovery

Cun Mu[1], Bo Huang[1], John Wright[2], Donald Goldfarb[1]

[1]Department of Industrial Engineering and Operations Research, Columbia University

[2]Department of Electrical Engineering, Columbia University

August 16, 2013

Abstract

Recovering a low-rank tensor from incomplete information is a recurring problem in signal processing and machine learning. The most popular convex relaxation of this problem minimizes the sum of the nuclear norms of the unfoldings of the tensor. We show that this approach can be substantially suboptimal: reliably recovering a K-way tensor of length n and Tucker rank r from Gaussian measurements requires $\Omega(rn^{K-1})$ observations. In contrast, a certain (intractable) nonconvex formulation needs only $O(r^K + nrK)$ observations. We introduce a very simple, new convex relaxation, which partially bridges this gap. Our new formulation succeeds with $O(r^{\lfloor K/2 \rfloor} n^{\lceil K/2 \rceil})$ observations. While these results pertain to Gaussian measurements, simulations strongly suggest that the new norm also outperforms the sum of nuclear norms for tensor completion from a random subset of entries.

Our lower bound for the sum-of-nuclear-norms model follows from a new result on recovering signals with multiple sparse structures (e.g. sparse, low rank), which perhaps surprisingly demonstrates the significant suboptimality of the commonly used recovery approach via minimizing the sum of individual sparsity inducing norms (e.g. l_1, nuclear norm). Our new formulation for low-rank tensor recovery however opens the possibility in reducing the sample complexity by exploiting several structures jointly.

1 Introduction

Tensors arise naturally in problems where the goal is to estimate a multi-dimensional object whose entries are indexed by several continuous or discrete variables. For example, a video is indexed by two spatial variables and one temporal variable; a hyperspectral datacube is indexed by two spatial variables and a frequency/wavelength variable. While tensors often reside in extremely high-dimensional data spaces, in many applications, the tensor of interest is *low-rank*, or approximately so [KB09], and hence has much lower-dimensional structure. The general problem of estimating a low-rank tensor has applications in many different areas, both theoretical and applied: e.g., estimating latent variable graphical models [AGH+12], classifying audio [MSS06], mining text [CC12], processing radar signals [DN10], to name a few.

In most part of the paper, we consider the problem of recovering a K-way tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_K}$ from linear measurements $\boldsymbol{z} = \mathcal{G}[\boldsymbol{\mathcal{X}}] \in \mathbb{R}^m$. Typically, $m \ll N = \prod_{i=1}^{K} n_i$, and so the problem of

1. REPORT DATE **16 AUG 2013**	2. REPORT TYPE	3. DATES COVERED **00-00-2013 to 00-00-2013**
4. TITLE AND SUBTITLE **Square Deal: Lower Bounds and Improved Relaxations for Tensor Recovery**		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER
		5e. TASK NUMBER
		5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Columbia University,Department of Industrial Engineering and Operations Research,New York City,NY,10027**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)
		11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Recovering a low-rank tensor from incomplete information is a recurring problem in signal processing and machine learning. The most popular convex relaxation of this problem minimizes the sum of the nuclear norms of the unfoldings of the tensor. We show that this approach can be substantially suboptimal: reliably recovering a K-way tensor of length n and Tucker rank r from Gaussian measurements requires (rn^{K-1}) observations. In contrast, a certain (intractable) nonconvex formulation needs only O(rK+nrK) observations. We introduce a very simple, new convex relaxation, which partially bridges this gap. Our new formulation succeeds with $O(rb^{K/2cnd^{K/2}e})$ observations. While these results pertain to Gaussian measurements simulations strongly suggest that the new norm also outperforms the sum of nuclear norms for tensor completion from a random subset of entries. Our lower bound for the sum-of-nuclear-norms model follows from a new result on recover- ing signals with multiple sparse structures (e.g. sparse, low rank), which perhaps surprisingly demonstrates the signi cant suboptimality of the commonly used recovery approach via minimiz- ing the sum of individual sparsity inducing norms (e.g. l1, nuclear norm). Our new formulation for low-rank tensor recovery however opens the possibility in reducing the sample complexity by exploiting several structures jointly.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**22**	

recovering $\mathcal{X}$ from $\mathbf{z}$ is *ill-posed*. In the past few years, tremendous progress has been made in understanding how to exploit structural assumptions such as sparsity for vectors [CRT] or low-rankness for matrices [RFP10] to develop computationally tractable methods for tackling ill-posed inverse problems. In many situations, convex optimization can estimate a structured object from near-minimal sets of observations [NRWY12, CRPW12, ALMT13]. For example, an $n \times n$ matrix of rank r can, with high probability, be exactly recovered from Cnr generic linear measurements, by minimizing the nuclear norm $\|\mathbf{X}\|_* = \sum_i \sigma_i(\mathbf{X})$. Since a generic rank r matrix has $r(2n - r)$ degrees of freedom, this is nearly optimal.

In contrast, the correct generalization of these results to low-rank tensors is not obvious. The numerical algebra of tensors is fraught with hardness results [HL09]. For example, even computing a tensor's (CP) rank,

$$\text{rank}_{\text{cp}}(\mathcal{X}) = \min \left\{ r \mid \mathcal{X} = \sum_{i=1}^{r} \mathbf{a}_1^{(i)} \circ \mathbf{a}_2^{(i)} \circ \cdots \circ \mathbf{a}_K^{(i)} \right\}, \tag{1.1}$$

is NP-hard in general. The nuclear norm of a tensor is also intractable, and so we cannot simply follow the formula that has worked for vectors and matrices.

With an eye towards numerical computation, many researchers have studied how to estimate or recover tensors of small *Tucker rank* [Tuc66]. The Tucker rank of a K-way tensor $\mathcal{X}$ is a K-dimensional vector whose i-th entry is the (matrix) rank of the mode-i unfolding $\mathcal{X}_{(i)}$ of $\mathcal{X}$:

$$\text{rank}_{\text{tc}}(\mathcal{X}) := \big(\text{rank}(\mathcal{X}_{(1)}), \text{rank}(\mathcal{X}_{(2)}), \cdots, \text{rank}(\mathcal{X}_{(K)})\big). \tag{1.2}$$

Here, the matrix $\mathcal{X}_{(i)} \in \mathbb{R}^{n_i \times \prod_{j \neq i} n_j}$ is obtained by concatenating all the mode-i fibers of $\mathcal{X}$ as column vectors. Each *mode-i fiber* is an n_i-dimensional vector obtained by fixing every index of $\mathcal{X}$ but the i-th one. The Tucker rank of $\mathcal{X}$ can be computed efficiently using the (matrix) singular value decomposition. For this reason, we focus on tensors of low Tucker rank. However, we will see that our proposed regularization strategy also automatically adapts to recover tensors of low CP rank, with some reduction in the required number of measurements.

The definition (1.2) suggests a very natural, tractable convex approach to recovering low-rank tensors: seek the $\mathcal{X}$ that minimizes $\sum_i \lambda_i \|\mathcal{X}_{(i)}\|_*$ out of all $\mathcal{X}$ satisfying $\mathcal{G}[\mathcal{X}] = \mathbf{z}$. We will refer to this as the *sum-of-nuclear-norms (SNN)* model. Originally, proposed in [LMWY09], this approach has been widely studied [GRY11, SDS10, THK10, TSHK11, STDLS13] and applied to various datasets in imaging [SVdPDMS11, SHKM13, KS13, LL10, LYZY10].

Perhaps surprisingly, we show that this natural approach can be *substantially suboptimal*, and introduce a simple new convex regularizer with provably better performance. For ease of stating results, suppose that $n_1 = \cdots = n_K = n$, and $\text{rank}_{\text{tc}}(\mathcal{X}) \preceq (r, r, \cdots, r)$. Let $\mathfrak{T}_r$ denote the set of all such tensors. We will consider the problem of estimating an element $\mathcal{X}_0$ of $\mathfrak{T}_r$ from Gaussian measurements $\mathcal{G}$ (i.e., $\mathbf{z}_i = \langle \mathcal{G}_i, \mathcal{X} \rangle$, where $\mathcal{G}_i$ has i.i.d. standard normal entries). To describe a generic tensor in $\mathfrak{T}_r$, we need at most $r^K + rnK$ parameters. Section 2 shows that a certain nonconvex strategy can recover all $\mathcal{X} \in \mathfrak{T}_r$ exactly when $m > (2r)^K + 2nrK$. In contrast, the best known theoretical guarantee for SNN minimization, due to Tomioka et. al. [TSHK11], shows that $\mathcal{X}_0 \in \mathfrak{T}_r$ can be recovered (or accurately estimated) from Gaussian measurements $\mathcal{G}$, provided $m = \Omega(rn^{K-1})$. In Section 3, we prove that this number of measurements is also *necessary*: accurate recovery is unlikely unless $m = \Omega(rn^{K-1})$. Thus, there is a substantial gap between an ideal nonconvex approach and the best known tractable surrogate. In Section 4, we introduce a

simple alternative, which we call the *square norm* model, which reduces the required number of measurements to $O(r^{\lfloor K/2 \rfloor} n^{\lceil K/2 \rceil})$. For $K > 3$, this improves by a multiplicative factor polynomial in n.

Our theoretical results pertain to Gaussian operators $\mathcal{G}$. The motivation for studying Gaussian measurements is twofold. First, Gaussian measurements may be of interest for compressed sensing recovery [Don06], either directly as a measurement strategy, or indirectly due to universality phenomena [DT09, BLM12]. Second, the available theoretical tools for Gaussian measurements are very sharp, allowing us to rigorously investigate the efficacy of various regularization schemes, and prove both upper and lower bonds on the number of observations required. In simulation, our qualitative conclusions carry over to more realistic measurement models, such as random subsampling [LMWY09] (see Section 5). We expect our results to be of interest for a wide range of problems in tensor completion [LMWY09], robust tensor recovery / decomposition [LYZY10, GQ12] and sensing.

Our technical methodology draws on, and enriches, the literature on general structured model recovery. The surprisingly poor behavior of the SNN model is an example of a phenomenon first discovered by Oymak et. al. [OJF$^+$12]: for recovering objects with multiple structures, a combination of structure-inducing norms is often not significantly more powerful than the best individual structure-inducing norm. Our lower bound for the SNN model follows from a general result of this nature, which we prove using the geometric framework of [ALMT13]. Compared to [OJF$^+$12], our result pertains to a more general family of regularizers, and gives sharper constants. In addition, we demonstrate the possibility to reduce the number of generic measurements through a new convex regularizer that exploit several sparse structures jointly.

2 Bounds for Non-Convex Recovery

In this section, we introduce a non-convex model for tensor recovery, and show that it recovers low-rank tensors from near-minimal numbers of measurements. While our nonconvex formulation is computationally intractable, it gives a baseline for evaluating tractable (convex) approaches.

For a tensor of low Tucker rank, the matrix unfolding along each mode is low-rank. Suppose we observe $\mathcal{G}[\boldsymbol{\mathcal{X}}_0] \in \mathbb{R}^m$. We would like to attempt to recover $\boldsymbol{\mathcal{X}}_0$ by minimizing some combination of the ranks of the unfoldings, over all tensors $\boldsymbol{\mathcal{X}}$ that are consistent with our observations. This suggests a *vector optimization* problem [BV04, Chap. 4.7]:

$$\text{minimize}_{(\text{w.r.t. } \mathbb{R}_+^K)} \quad \text{rank}_{\text{tc}}(\boldsymbol{\mathcal{X}}) \quad \text{subject to} \quad \mathcal{G}[\boldsymbol{\mathcal{X}}] = \mathcal{G}[\boldsymbol{\mathcal{X}}_0]. \tag{2.1}$$

In vector optimization, a feasible point is called *Pareto optimal* if no other feasible point dominates it in every criterion. In a similar vein, we say that (2.1) recovers $\boldsymbol{\mathcal{X}}_0$ if there does not exist any other tensor $\boldsymbol{\mathcal{X}}$ that is consistent with the observations and has no larger rank along each mode:

Definition 1. *We call $\boldsymbol{\mathcal{X}}_0$ recoverable by (2.1) if the set*

$$\{\boldsymbol{\mathcal{X}}' \neq \boldsymbol{\mathcal{X}}_0 \mid \mathcal{G}[\boldsymbol{\mathcal{X}}'] = \mathcal{G}[\boldsymbol{\mathcal{X}}_0], \ \text{rank}_{\text{tc}}(\boldsymbol{\mathcal{X}}') \preceq_{\mathbb{R}_+^K} \text{rank}_{\text{tc}}(\boldsymbol{\mathcal{X}}_0)\} = \emptyset.$$

This is equivalent to saying that $\boldsymbol{\mathcal{X}}_0$ is the unique optimal solution to the *scalar* optimization:

$$\text{minimize}_{\boldsymbol{\mathcal{X}}} \quad \max_i \left\{ \frac{\text{rank}(\boldsymbol{\mathcal{X}}_{(i)})}{\text{rank}(\boldsymbol{\mathcal{X}}_{0(i)})} \right\} \quad \text{subject to} \quad \mathcal{G}[\boldsymbol{\mathcal{X}}] = \mathcal{G}[\boldsymbol{\mathcal{X}}_0]. \tag{2.2}$$

3

The problems (2.1)-(2.2) are not tractable. However, they do serve as a baseline for understanding how many generic measurements are required to recover $\boldsymbol{\mathcal{X}}_0$. The recovery performance of program (2.1) depends heavily on the properties of $\mathcal{G}$. Suppose (2.1) fails to recover $\boldsymbol{\mathcal{X}}_0 \in \mathfrak{T}_r$. Then there exists another $\boldsymbol{\mathcal{X}}' \in \mathfrak{T}_r$ such that $\mathcal{G}[\boldsymbol{\mathcal{X}}'] = \mathcal{G}[\boldsymbol{\mathcal{X}}_0]$. So, to guarantee that (2.1) recovers *any* $\boldsymbol{\mathcal{X}}_0 \in \mathfrak{T}_r$, a necessary and sufficient condition is that $\mathcal{G}$ is injective on $\mathfrak{T}_r$, which can be implied by the condition $\mathrm{null}(\mathcal{G}) \cap \mathfrak{T}_{2r} = \{\boldsymbol{0}\}$. Consequently, if $\mathrm{null}(\mathcal{G}) \cap \mathfrak{T}_{2r} = \{\boldsymbol{0}\}$, (2.1) will recover any $\boldsymbol{\mathcal{X}}_0 \in \mathfrak{T}_r$. We expect this to occur when the number of measurements significantly exceeds the number of intrinsic degrees of freedom of a generic element of $\mathfrak{T}_r$, which is $O(r^K + nrK)$. The following theorem shows that when m is approximately twice this number, with probability one, $\mathcal{G}$ is injective on $\mathfrak{T}_r$:

Theorem 1. *Whenever* $m \geq (2r)^K + 2nrK + 1$, *with probability one,* $\mathrm{null}(\mathcal{G}) \cap \mathfrak{T}_{2\mathrm{r}} = \{\boldsymbol{0}\}$, *and hence (2.1) recovers every* $\boldsymbol{\mathcal{X}}_0 \in \mathfrak{T}_r$.

The proof of Theorem 1 follows from a covering argument, which we establish in several steps. Let

$$\mathfrak{S}_{2r} = \{\boldsymbol{\mathcal{D}} \mid \boldsymbol{\mathcal{D}} \in \mathfrak{T}_{2r}, \|\boldsymbol{\mathcal{D}}\|_F = 1\}. \tag{2.3}$$

The following lemma shows that the required number of measurements can be bounded in terms of the exponent of the covering number for $\mathfrak{S}_{2r}$, which can be considered as a proxy for dimensionality:

Lemma 1. *Suppose that the covering number for* $\mathfrak{S}_{2r}$ *with respect to Frobenius norm, satisfies*

$$N(\mathfrak{S}_{2r}, \|\cdot\|_F, \varepsilon) \leq (\beta/\varepsilon)^d, \tag{2.4}$$

for some integer d and scalar β that does not depend on ε. Then if $m \geq d+1$, with probability one $\mathrm{null}(\mathcal{G}) \cap \mathfrak{S}_{2r} = \emptyset$, *which implies that* $\mathrm{null}(\mathcal{G}) \cap \mathfrak{T}_{2r} = \{\boldsymbol{0}\}$.

It just remains to find the covering number of $\mathfrak{S}_{2r}$. We use the following lemma, which uses the triangle inequality to control the effect of perturbations in the factors of the Tucker decomposition

$$[[\boldsymbol{\mathcal{C}}; \boldsymbol{U}_1, \boldsymbol{U}_2, \cdots, \boldsymbol{U}_K]] := \boldsymbol{\mathcal{C}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \cdots \times_K \boldsymbol{U}_K, \tag{2.5}$$

where the *mode-i (matrix) product* of tensor $\boldsymbol{\mathcal{A}}$ with matrix $\boldsymbol{B}$ of compatible size, denoted as $\boldsymbol{\mathcal{A}} \times_i \boldsymbol{B}$, outputs a tensor $\boldsymbol{\mathcal{C}}$ such that $\boldsymbol{\mathcal{C}}_{(i)} = \boldsymbol{B}\boldsymbol{\mathcal{A}}_{(i)}$.

Lemma 2. *Let* $\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{C}}' \in \mathbb{R}^{r_1, \ldots, r_K}$, *and* $\boldsymbol{U}_1, \boldsymbol{U}_1' \in \mathbb{R}^{n_1 \times r_1}, \ldots, \boldsymbol{U}_K, \boldsymbol{U}_K' \in \mathbb{R}^{n_K \times r_K}$ *with* $\boldsymbol{U}_i^* \boldsymbol{U}_i = \boldsymbol{U}_i'^* \boldsymbol{U}_i' = \boldsymbol{I}$, *and* $\|\boldsymbol{\mathcal{C}}\|_F = \|\boldsymbol{\mathcal{C}}'\|_F = 1$. *Then*

$$\left\| [[\boldsymbol{\mathcal{C}}; \boldsymbol{U}_1, \ldots, \boldsymbol{U}_K]] - [[\boldsymbol{\mathcal{C}}'; \boldsymbol{U}_1', \ldots, \boldsymbol{U}_K']] \right\|_F \leq \left\| \boldsymbol{\mathcal{C}} - \boldsymbol{\mathcal{C}}' \right\|_F + \sum_{i=1}^{K} \left\| \boldsymbol{U}_i - \boldsymbol{U}_i' \right\|. \tag{2.6}$$

Using this result, we construct an ε-net for $\mathfrak{S}_{2r}$ by building $\varepsilon/(K+1)$-nets for each of the $K+1$ factors $\boldsymbol{\mathcal{C}}$ and $\{\boldsymbol{U}_i\}$. The total size of the resulting ε net is thus bounded by the following lemma:

Lemma 3. $N(\mathfrak{S}_{2r}, \|\cdot\|_F, \varepsilon) \leq (3(K+1)/\varepsilon)^{(2r)^K + 2nrK}$

With these observations in hand, Theorem 1 follows immediately.

3 Convexification: Sum of Nuclear Norms?

Since the nonconvex problem (2.1) is NP-hard for general $\mathcal{G}$, it is tempting to seek a convex surrogate. In matrix recovery problems, the nuclear norm is often an excellent convex surrogate for the rank [Faz02, RFP10, Gro11]. It seems natural, then, to replace the ranks in (2.1) with nuclear norms, and solve

$$\text{minimize } h(\boldsymbol{\mathcal{X}}) := \left(\left\|\boldsymbol{\mathcal{X}}_{(1)}\right\|_*, \left\|\boldsymbol{\mathcal{X}}_{(2)}\right\|_*, \cdots, \left\|\boldsymbol{\mathcal{X}}_{(K)}\right\|_*\right) \quad \text{subject to} \quad \mathcal{G}[\boldsymbol{\mathcal{X}}] = \mathcal{G}[\boldsymbol{\mathcal{X}}_0]. \qquad (3.1)$$

Since $\left\|\boldsymbol{\mathcal{X}}_{(i)}\right\|_*$ is a convex function, the set $\mathcal{H} := \bigcup_{\boldsymbol{\mathcal{X}}:\mathcal{G}[\boldsymbol{\mathcal{X}}]=\mathcal{G}[\boldsymbol{\mathcal{X}}_0]} \{\boldsymbol{\mathcal{Y}} \mid h(\boldsymbol{\mathcal{Y}}) \succeq h(\boldsymbol{\mathcal{X}})\}$ is convex. For any pareto optimal point $\widehat{\boldsymbol{\mathcal{X}}}$, there is a hyperplane supporting $\mathcal{H}$ passing through $h(\widehat{\boldsymbol{\mathcal{X}}})$, with normal vector $\boldsymbol{\lambda} \geq \mathbf{0}$. Therefore, $\widehat{\boldsymbol{\mathcal{X}}}$ is an optimal solution to the following scalar optimization:

$$\text{minimize} \quad \sum_{i=1}^K \lambda_i \|\boldsymbol{\mathcal{X}}_{(i)}\|_* \quad \text{subject to} \quad \mathcal{G}[\boldsymbol{\mathcal{X}}] = \mathcal{G}[\boldsymbol{\mathcal{X}}_0]. \qquad (3.2)$$

The optimization (3.2) was first introduced by [LMWY09] and has been used successfully in applications in imaging [SVdPDMS11, SHKM13, KS13, LL10, GEK13, LYZY10]. Similar convex relaxations have been considered in a number of theoretical and algorithmic works [GRY11, SDS10, THK10, TSHK11, STDLS13]. It is not too surprising, then, that (3.2) provably recovers the underlying tensor $\boldsymbol{\mathcal{X}}_0$, when the number of measurements m is sufficiently large. For example, the following is a (simplified) corollary of results of Tomioka et. al. [THK10]:[1]

Corollary 2 (of [THK10], Theorem 3). *Suppose that $\boldsymbol{\mathcal{X}}_0$ has Tucker rank $(r, \dots, r)$, and $m \geq Crn^{K-1}$. With high probability, $\boldsymbol{\mathcal{X}}_0$ is an optimal solution to (3.2), with each $\lambda_i = 1$. Here, C is numerical.*

This result shows that there *is* a range in which (3.2) succeeds: loosely, when we undersample by at most a factor of $m/N \sim r/n$. However, the number of observations $m \sim rn^{K-1}$ is significantly larger than the number of degrees of freedom in $\boldsymbol{\mathcal{X}}_0$, which is on the order of $r^K + nrK$. Is it possible to prove a better bound for this model? Unfortunately, we show that in general $O(rn^{K-1})$ measurements are also *necessary* for reliable recovery using (3.2):

Theorem 3. *Let $\boldsymbol{\mathcal{X}}_0 \in \mathfrak{T}_r$ be nonzero. Set $\kappa = \min_i \left\{ \left\|(\boldsymbol{\mathcal{X}}_0)_{(i)}\right\|_*^2 / \|\boldsymbol{\mathcal{X}}_0\|_F^2 \right\} \times n^{K-1}$. Then if the number of measurements $m \leq \kappa - 2$, $\boldsymbol{\mathcal{X}}_0$ is not the unique solution to (3.2), with probability at least $1 - 4\exp(-\frac{(\kappa-m-2)^2}{16(\kappa-2)})$. Moreover, there exists $\boldsymbol{\mathcal{X}}_0 \in \mathfrak{T}_r$ for which $\kappa = rn^{K-1}$.*

This implies that Corollary 2 (and other results of [THK10]) is essentially tight. Unfortunately, it has negative implications for the efficacy of the sum of nuclear norms in (3.2): although a generic element $\boldsymbol{\mathcal{X}}_0$ of $\mathfrak{T}_r$ can be described using at most $r^K + nrK$ real numbers, we require $\Omega(rn^{K-1})$ observations to recover it using (3.2). Theorem 3 is a direct consequence of a much more general principle underlying multi-structured recovery, which is elaborated next.

---

[1]Tomioka et. al. also show noise stability when $m = \Omega(rn^{K-1})$ and give extensions to the case where $\text{rank}_{tc}(\boldsymbol{\mathcal{X}}_0) = (r_1, \dots, r_K)$ differs from mode to mode.

Recovering objects with multiple structures

The poor behavior of (3.2) is actually an instance of a much more general phenomenon, first discovered by Oymak et. al. [OJF$^+$12]. Our target tensor $\boldsymbol{\mathcal{X}}_0$ has *multiple* low-dimensional structures simultaneously: it is low-rank along *each* of the K modes. In practical applications, many other such *simultaneously structured* objects may be of interest – for example, matrices that are simultaneously sparse and low-rank [RSV12, OJF$^+$12]. To recover such a simultaneously structured object, it is tempting to build a convex relaxation by combining the convex relaxations for each of the individual structures. In the tensor case, this yields (3.2). Surprisingly, this combination is often not significantly more powerful than the best single regularizer [OJF$^+$12]. We obtain Theorem 3 as a consquence of a new, general result of this nature, using a geometric framework introduced in [ALMT13]. Compared to the proof strategy in [OJF$^+$12], this approach has a clearer geometric intuition, covers a more general class of regularizers and yields sharper bounds.

Consider a signal $\boldsymbol{x}_0 \in \mathbb{R}^n$ having K low-dimensional structures simultaneously (e.g. sparsity, low-rank, etc.)[2]. Let $\|\cdot\|_{(i)}$ be the penalty norms corresponding to the i-th structure (e.g. ℓ_1, nuclear norm). Consider the composite norm optimization

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(x) := \lambda_1 \|\boldsymbol{x}\|_{(1)} + \lambda_2 \|\boldsymbol{x}\|_{(2)} + \cdots + \lambda_K \|\boldsymbol{x}\|_{(K)} \qquad \text{subject to} \quad \mathcal{G}[\boldsymbol{x}] = \mathcal{G}[\boldsymbol{x}_0], \qquad (3.3)$$

where $\mathcal{G}[\cdot]$ is a Gaussian measurement operator, and $\boldsymbol{\lambda} > \boldsymbol{0}$. Is $\boldsymbol{x}_0$ the unique optimal solution to (3.3)? Recall that the descent cone of a function f at a point $\boldsymbol{x}_0$ is defined as

$$\mathcal{C}(f, \boldsymbol{x}_0) = \text{cone} \left\{ \boldsymbol{v} \mid f(\boldsymbol{x}_0 + \boldsymbol{v}) \leq f(\boldsymbol{x}_0) \right\}, \qquad (3.4)$$

which, in short, will be denoted as $\mathcal{C}$. Then $\boldsymbol{x}_0$ is the unique optimal solution if and only if $\text{null}(\mathcal{G}) \cap \mathcal{C} = \{\boldsymbol{0}\}$. Conversely, recovery fails if $\text{null}(\mathcal{G})$ has nontrivial intersection with $\mathcal{C}$. If $\mathcal{G}$ is a Gaussian operator, $\text{null}(\mathcal{G})$ is a uniformly oriented random subspace of dimension $n-m$. This random subspace is more likely to have nontrivial intersection with $\mathcal{C}$ if $\mathcal{C}$ is "large," in a sense we will make precise. The polar of $\mathcal{C}$ is $\mathcal{C}^o = \text{cone}(\partial f(\boldsymbol{x}_0))$. Because polarity reverses inclusion, we expect that $\mathcal{C}$ will be "large" whenever $\mathcal{C}^o$ is "small". Figure 1 visualizes this geometry.

To control the size of C^o, first consider a single norm $\|\cdot\|_\diamond$, with dual norm $\|\cdot\|_\diamond^*$. Suppose that $\|\cdot\|_\diamond$ is L-Lipschitz: $\|\boldsymbol{x}\|_\diamond \leq L \|\boldsymbol{x}\|_2$ for all $\boldsymbol{x}$. Then $\|\boldsymbol{x}\|_2 \leq L \|\boldsymbol{x}\|_\diamond^*$ for all $\boldsymbol{x}$ as well. Noting that

$$\partial \|\cdot\|_\diamond (\boldsymbol{x}) = \left\{ \boldsymbol{v} \mid \langle \boldsymbol{v}, \boldsymbol{x} \rangle = \|\boldsymbol{x}\|_\diamond, \ \|\boldsymbol{v}\|_\diamond^* \leq 1 \right\},$$

for any $\boldsymbol{v} \in \partial \|\cdot\|_\diamond (\boldsymbol{x}_0)$, we have

$$\frac{\langle \boldsymbol{v}, \boldsymbol{x}_0 \rangle}{\|\boldsymbol{v}\|_2 \|\boldsymbol{x}_0\|_2} \geq \frac{\|\boldsymbol{x}_0\|_\diamond}{L \|\boldsymbol{v}\|_\diamond^* \|\boldsymbol{x}_0\|_2} \geq \frac{\|\boldsymbol{x}_0\|_\diamond}{L \|\boldsymbol{x}_0\|_2}. \qquad (3.5)$$

A more geometric way of summarizing this is as follows: for $\boldsymbol{x} \neq \boldsymbol{0}$, let

$$\text{circ}(\boldsymbol{x}, \theta) = \left\{ \boldsymbol{z} \mid \angle(\boldsymbol{z}, \boldsymbol{x}) \leq \theta \right\}, \qquad (3.6)$$

and denote the *circular cone* with axis $\boldsymbol{x}$ and angle θ. Then if $\boldsymbol{x}_0 \neq \boldsymbol{0}$, and $\theta = \cos^{-1}(\|\boldsymbol{x}_0\|_\diamond / L \|\boldsymbol{x}_0\|_2)$,

$$\partial \|\cdot\|_\diamond (\boldsymbol{x}_0) \subseteq \text{circ}(\boldsymbol{x}_0, \theta). \qquad (3.7)$$

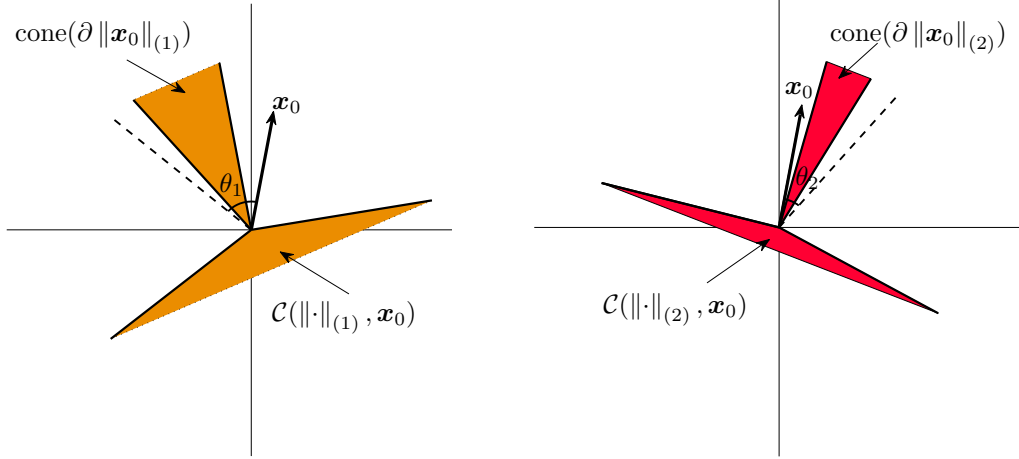[2] $\boldsymbol{x}_0$ is the underlying signal of our interest (perhaps after vectorization).

Figure 1: **Cones and their polars for convex regularizers** $\|\cdot\|_{(1)}$ **and** $\|\cdot\|_{(2)}$ **respectively.** Suppose our $\boldsymbol{x}_0$ has two sparse structures simultaneously. Regularizer $\|\cdot\|_{(1)}$ has a larger conic hull of subdifferential at $\boldsymbol{x}_0$, i.e. cone$(\partial \|\boldsymbol{x}_0\|_{(1)})$, which results in a smaller descent cone. Thus minimizing $\|\cdot\|_{(1)}$ is more likely to recover $\boldsymbol{x}_0$ than minimizing $\|\cdot\|_{(2)}$. Consider convex regularizer $f(\boldsymbol{x}) = \|\boldsymbol{x}_0\|_{(1)} + \|\boldsymbol{x}_0\|_{(2)}$. Suppose as depicted, $\theta_1 \geq \theta_2$. Then both cone$(\partial \|\boldsymbol{x}_0\|_{(1)})$ and cone$(\partial \|\boldsymbol{x}_0\|_{(2)})$ are in the circular cone circ$(\boldsymbol{x}_0, \theta_1)$. Thus we have: cone$\big(\partial f(\boldsymbol{x}_0)\big) = $ cone$(\partial \|\boldsymbol{x}_0\|_{(1)} + \partial \|\boldsymbol{x}_0\|_{(2)}) \subseteq$ conv$\big\{$circ$(\boldsymbol{x}_0, \theta_1),$ circ$(\boldsymbol{x}_0, \theta_2)\big\} = $ circ$(\boldsymbol{x}_0, \theta_1)$.

Table 1 describes the angle parameters θ for various structure inducing norms. Notice that in general, more complicated $\boldsymbol{x}_0$ leads to smaller angles θ. For example, if $\boldsymbol{x}_0$ is a k-sparse vectors with entries all of the same magnitude, and $\|\cdot\|_\diamond$ the ℓ^1 norm, $\cos^2 \theta = k/n$. As $\boldsymbol{x}_0$ becomes more dense, $\partial \|\cdot\|_\diamond$ is contained in smaller and smaller circular cones.

For $f = \sum_i \lambda_i \|\cdot\|_{(i)}$, notice that every element of $\partial f(\boldsymbol{x}_0)$ is a conic combination of elements of the $\partial \|\cdot\|_{(i)}(\boldsymbol{x}_0)$. Since each of the $\partial \|\cdot\|_{(i)}(\boldsymbol{x}_0)$ is contained in a circular cone with axis $\boldsymbol{x}_0$, $\partial f(\boldsymbol{x}_0)$ is also contained in a circular cone:

Lemma 4. *Suppose that* $\|\cdot\|_{(i)}$ *is* L_i-*Lipschitz. For* $\boldsymbol{x}_0 \neq \boldsymbol{0}$, *set* $\theta_i = \cos^{-1}\left(\|\boldsymbol{x}_0\|_{(i)} / L \|\boldsymbol{x}_0\|_2\right)$. *Then*

$$\partial f(\boldsymbol{x}_0) \subseteq \text{circ}\left(\boldsymbol{x}_0, \max_{i=1...K} \theta_i\right). \tag{3.8}$$

So, the subdifferential of our combined regularizer f is contained in a circular cone whose angle is given by the largest of the θ_i.

How does this behavior affect the recoverability of $\boldsymbol{x}_0$ via (3.3)? The informal reasoning above suggests that as θ becomes smaller, the descent cone $\mathcal{C}$ becomes larger, and we require more measurements to recover $\boldsymbol{x}_0$. This can be made precise using an elegant framework introduced by Amelunxen et. al. [ALMT13]. They define the *statistical dimension* of the convex cone $\mathcal{C}$ to be the expected norm of the projection of a standard Gaussian vector onto C:

$$\delta(\mathcal{C}) \doteq \mathbb{E}_{\boldsymbol{g} \sim \text{i.i.d.} \mathcal{N}(0,1)} \left[\|\mathcal{P}_\mathcal{C}(\boldsymbol{g})\|_2^2 \right]. \tag{3.9}$$

Object	Complexity Measure	Relaxation	$\cos^2\theta$	κ
Sparse $\boldsymbol{x} \in \mathbb{R}^n$	$k = \|\boldsymbol{x}\|_0$	$\|\boldsymbol{x}\|_1$	$[\frac{1}{n}, \frac{k}{n}]$	$[1, k]$
Column-sparse $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$	$c = \#\{j \mid \boldsymbol{X}e_j \neq \boldsymbol{0}\}$	$\sum_j \|\boldsymbol{X}e_j\|_2$	$[\frac{1}{n_2}, \frac{c}{n_2}]$	$[n_1, cn_1]$
Low-rank $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ $(n_1 \geq n_2)$	$r = \text{rank}(\boldsymbol{X})$	$\|\boldsymbol{X}\|_*$	$[\frac{1}{n_2}, \frac{r}{n_2}]$	$[n_1, rn_1]$
Low-rank $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n \times n \times \cdots \times n}$	$\text{rank}_{\text{tc}}(\boldsymbol{\mathcal{X}})$	$\sum_i \|\boldsymbol{\mathcal{X}}_{(i)}\|_*$	$[\frac{1}{n}, \frac{r}{n}]$	$[n^{K-1}, rn^{K-1}]$
	$\text{rank}_{\text{tc}}(\boldsymbol{\mathcal{X}})$	$\|\boldsymbol{\mathcal{X}}\|_\square$	$[(\frac{1}{n})^{\lfloor\frac{K}{2}\rfloor}, (\frac{r}{n})^{\lfloor\frac{K}{2}\rfloor}]$	$[n^{\lceil\frac{K}{2}\rceil}, r^{\lfloor\frac{K}{2}\rfloor}n^{\lceil\frac{K}{2}\rceil}]$

Table 1: **Concise models and their surrogates.** For each norm $\|\cdot\|_\diamond$, the third column describes the range of achievable angles θ. Larger $\cos\theta$ corresponds to a smaller C^o, a larger C, and hence a larger number of measurements required for reliable recovery. The fourth line is the sum of nuclear norms; the last line is the square norm introduced in Section 4.

Using tools from spherical integral geometry, [ALMT13] shows that for linear inverse problems with Gaussian measurements, a sharp phase transition in recoverability occurs around $m = \delta(\mathcal{C})$. We will need only one side of their result; for more details see [ALMT13]. We state a slight variant here:

Corollary 4. *Let* $\mathcal{G} : \mathbb{R}^n \to \mathbb{R}^m$ *be a Gaussian operator, and* $\mathcal{C}$ *a convex cone. Then if* $m \leq \delta(\mathcal{C})$,

$$\mathbb{P}\left[\mathcal{C} \cap \text{null}(\mathcal{G}) = \{\boldsymbol{0}\}\right] \leq 4\exp\left(-\frac{(\delta(\mathcal{C}) - m)^2}{16\delta(\mathcal{C})}\right). \tag{3.10}$$

To apply this result to our problem, we lower bound the statistical dimension $\delta(\mathcal{C})$, of the descent cone $\mathcal{C}$ of f at $\boldsymbol{x}_0$. Using the Pythagorean theorem, monotonicity of $\delta(\cdot)$, and Lemma 4, we calculate

$$\delta(\mathcal{C}) = n - \delta(\mathcal{C}^o) = n - \delta\left(\text{cone}(\partial f(\boldsymbol{x}_0))\right) \geq n - \delta(\text{circ}(\boldsymbol{x}_0, \max_i \theta_i)). \tag{3.11}$$

Moreover, using the properties of statistical dimension, we are able to prove an upper bound for the statistical dimension of circular cone, which improves the constant in existing results [ALMT13, McC13].

Lemma 5. $\delta(\text{circ}(\boldsymbol{x}_0, \theta)) \leq n\sin^2\theta + 2$.

Finally, by combining (3.11) and Lemma 5, we have $\delta(C) \geq n\min_i \cos^2\theta_i - 2$. Using Corollary 4, we obtain:

Theorem 5. *Let* $\boldsymbol{x}_0 \neq \boldsymbol{0}$. *Suppose that for each* i, $\|\cdot\|_{(i)}$ *is* L_i-*Lipschitz. Set*

$$\kappa_i = \frac{n\|\boldsymbol{x}_0\|_{(i)}^2}{L_i^2 \|\boldsymbol{x}_0\|_2^2} = n\cos^2(\theta_i),$$

and $\kappa = \min_i \kappa_i$. *Then if* $m \leq \kappa - 2$,

$$\mathbb{P}\left[\boldsymbol{x}_0 \text{ is the unique optimal solution to (3.3)}\right] \leq 4\exp\left(-\frac{(\kappa - m - 2)^2}{16(\kappa - 2)}\right). \tag{3.12}$$

8

Thus, for reliable recovery, the number of measurements needs to be at least proportional to κ.[3] Notice that $\kappa = \min_i \kappa_i$ is determined by only the best of the structures. Per Table 1, κ_i is often on the order of the number of degrees of freedom in a generic object of the i-th structure. For example, for a k-sparse vector whose nonzeros are all of the same magnitude, $\kappa = k$.

Theorem 5 together with Table 1 leads us to the phenomenon that recently discovered by Oymak et. al. [OJF+12]: for recovering objects with multiple structures, a combination of structure-inducing norms tends to be not significantly more powerful than the best individual structure-inducing norm. As we demonstrate, this general behavior follows a clear geometric interpretation that the subdifferential of a norm at $\boldsymbol{x}_0$ is contained in a relatively small circular cone with central axis $\boldsymbol{x}_0$.

We can specialize Theorem 5 to low-rank tensors as follows: if $\boldsymbol{\mathcal{X}}$ is a K-mode $n \times n \times \cdots \times n$ tensor of Tucker rank $(r, r, \ldots, r)$, then for each i, $\|\boldsymbol{\mathcal{X}}\|_{(i)} \doteq \|\boldsymbol{\mathcal{X}}_{(i)}\|_*$ is $L = \sqrt{n}$-Lipschitz. Hence,

$$\kappa = \min_i \left\{ \|\boldsymbol{\mathcal{X}}_{(i)}\|_*^2 / \|\boldsymbol{\mathcal{X}}\|_F^2 \right\} n^{K-1}. \tag{3.13}$$

The term in brackets lies between 1 and r, inclusive. For example, if $\boldsymbol{\mathcal{X}} = [[\boldsymbol{\mathcal{C}}, \boldsymbol{U}_1, \ldots, \boldsymbol{U}_K]]$, with $\boldsymbol{U}_i^T \boldsymbol{U}_i = \boldsymbol{I}$ and $\boldsymbol{\mathcal{C}}$ supersymmetric ($\boldsymbol{\mathcal{C}}_{i_1 \ldots i_K} = \mathbb{1}_{i_1 = i_2 = \cdots = i_K}$), then this term is equal to r.

4   A Better Convexification: Square Norm

The number of measurements promised by Corollary 2 and Theorem 3 is actually the same (up to constants) as the number of measurements required to recover a tensor $\boldsymbol{\mathcal{X}}_0$ which is low-rank along just one mode. Since matrix nuclear norm minimization correctly recovers a $n_1 \times n_2$ matrix of rank r when $m \geq Cr(n_1 + n_2)$ [CRPW12], solving

$$\text{minimize } \|\boldsymbol{\mathcal{X}}_{(1)}\|_* \quad \text{subject to} \quad \mathcal{G}[\boldsymbol{\mathcal{X}}] = \mathcal{G}[\boldsymbol{\mathcal{X}}_0] \tag{4.1}$$

also exactly recovers $\boldsymbol{\mathcal{X}}_0$ with high probability when $m \geq Crn^{K-1}$.

This suggests a more mundane explanation for the difficulty with (3.2): the term rn^{K-1} comes from the need to reconstruct the right singular vectors of the $n \times n^{K-1}$ matrix $\boldsymbol{\mathcal{X}}_{(1)}$. If we had some way of matricizing a tensor that *produced a more balanced (square) matrix* and also *preserved the low-rank property*, we could substantially reduce this effect, and reduce the overall sampling requirement. In fact, this is possible when the order K of $\boldsymbol{\mathcal{X}}_0$ is four or larger.

For $\boldsymbol{A} \in \mathbb{R}^{m_1 \times n_1}$, and integers m_2 and n_2 satisfying $m_1 n_1 = m_2 n_2$, the reshaping operator reshape($\boldsymbol{A}, m_2, n_2$) returns a $m_2 \times n_2$ matrix whose elements are taken columnwise from $\boldsymbol{A}$. This operator rearranges elements in $\boldsymbol{A}$ and leads to a matrix of different shape. In the following, we reshape matrix $\boldsymbol{\mathcal{X}}_{(1)}$ to a more square matrix while preserving the low-rank property. Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_K}$. Select $j \in [K] := \{1, 2, \cdots, K\}$. Then we define matrix $\boldsymbol{\mathcal{X}}_{[j]}$ as

$$\boldsymbol{\mathcal{X}}_{[j]} = \text{reshape}\left(\boldsymbol{\mathcal{X}}_{(1)}, \prod_{i=1}^{j} n_i, \prod_{i=j+1}^{K} n_i\right).$$

We can view $\boldsymbol{\mathcal{X}}_{[j]}$ as a natural generalization of the standard tensor matricization. When $j = 1$, $\boldsymbol{\mathcal{X}}_{[j]}$ is nothing but $\boldsymbol{\mathcal{X}}_{(1)}$. However, when some $j > 1$ is selected, $\boldsymbol{\mathcal{X}}_{[j]}$ becomes a more balanced matrix. This reshaping also preserves some of the algebraic structures of $\boldsymbol{\mathcal{X}}$. In particular, we will see that if $\boldsymbol{\mathcal{X}}$ is a low-rank tensor (in either the CP or Tucker sense), $\boldsymbol{\mathcal{X}}_{[j]}$ will be a low-rank matrix.

---

[3]E.g., if $m = (\kappa - 2)/2$, the probability of success is at most $4\exp(-(\kappa - 2)/64)$.

Lemma 6. *(1) If $\boldsymbol{\mathcal{X}}$ has CP decomposition $\boldsymbol{\mathcal{X}} = \sum_{i=1}^{r} \lambda_i \boldsymbol{a}_i^{(1)} \circ \boldsymbol{a}_i^{(2)} \circ \cdots \circ \boldsymbol{a}_i^{(K)}$, then*

$$\boldsymbol{\mathcal{X}}_{[j]} = \sum_{i=1}^{r} \lambda_i (\boldsymbol{a}_i^{(j)} \otimes \boldsymbol{a}_i^{(j-1)} \otimes \cdots \otimes \boldsymbol{a}_i^{(1)}) \circ (\boldsymbol{a}_i^{(K)} \otimes \boldsymbol{a}_i^{(K-1)} \cdots \otimes \boldsymbol{a}_i^{(j+1)}). \tag{4.2}$$

(2) If $\boldsymbol{\mathcal{X}}$ has Tucker decomposition $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{C}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \cdots \times_K \boldsymbol{U}_K$, then

$$\boldsymbol{\mathcal{X}}_{[j]} = (\boldsymbol{U}_j \otimes \boldsymbol{U}_{j-1} \otimes \cdots \otimes \boldsymbol{U}_1) \, \boldsymbol{\mathcal{C}}_{[j]} \, (\boldsymbol{U}_K \otimes \boldsymbol{U}_{K-1} \otimes \cdots \otimes \boldsymbol{U}_{j+1})^*. \tag{4.3}$$

Using Lemma 6 and the fact that $\mathrm{rank}(\boldsymbol{A} \otimes \boldsymbol{B}) = \mathrm{rank}(\boldsymbol{A})\,\mathrm{rank}(\boldsymbol{B})$, we obtain:

Lemma 7. *Let $\mathrm{rank}_{\mathrm{tc}}(\boldsymbol{\mathcal{X}}) = (r_1, r_2, \cdots, r_K)$, and $\mathrm{rank}_{\mathrm{cp}}(\boldsymbol{\mathcal{X}}) = r_{\mathrm{cp}}$. Then $\mathrm{rank}(\boldsymbol{\mathcal{X}}_{[j]}) \leq r_{\mathrm{cp}}$, and $\mathrm{rank}(\boldsymbol{\mathcal{X}}_{[j]}) \leq \min\left\{ \prod_{i=1}^{j} r_i, \prod_{i=j+1}^{K} r_i \right\}$.*

Thus, $\boldsymbol{\mathcal{X}}_{[j]}$ is not only more balanced but also maintains the low-rank property of tensor $\boldsymbol{\mathcal{X}}$. In the following, we show how this new matricization can lead to better relaxations for tensor recovery. For ease of discussion, we assume $\boldsymbol{\mathcal{X}}$ has the same length (say n) along each mode and has Tucker rank $(r, r, \cdots, r)$. We write $\boldsymbol{\mathcal{X}}_{\square} \doteq \boldsymbol{\mathcal{X}}_{[\lfloor \frac{K}{2} \rfloor]}$ and call $\|\boldsymbol{\mathcal{X}}\|_{\square} := \|\boldsymbol{\mathcal{X}}_{\square}\|_{\star}$ the *square norm* of tensor $\boldsymbol{\mathcal{X}}$. Since $\boldsymbol{\mathcal{X}}_{\square}$ is low-rank, we can attempt to recover $\boldsymbol{\mathcal{X}}$ by solving

$$\text{minimize} \quad \|\boldsymbol{\mathcal{X}}\|_{\square} \quad \text{subject to} \quad \mathcal{G}[\boldsymbol{\mathcal{X}}] = \mathcal{G}[\boldsymbol{\mathcal{X}}_0]. \tag{4.4}$$

Using Lemma 7 and Proposition 3.11 of [CRPW12], we can prove that this relaxation exactly recovers $\boldsymbol{\mathcal{X}}_0$, when the number of measurements is sufficiently large:

Theorem 6. *(1) If $\boldsymbol{\mathcal{X}}_0$ has CP rank r, using (4.4), $m \geq Crn^{\lceil \frac{K}{2} \rceil}$ is sufficient to recover $\boldsymbol{\mathcal{X}}_0$ with high probability. (2) If $\boldsymbol{\mathcal{X}}_0$ has Tucker rank $(r, r, \cdots, r)$, using (4.4), $m \geq Cr^{\lfloor \frac{K}{2} \rfloor} n^{\lceil \frac{K}{2} \rceil}$ is sufficient to recover $\boldsymbol{\mathcal{X}}_0$ with high probability.*

Compared with $\Omega(rn^{K-1})$ measurements required by the sum-of-nuclear-norms model, the sample complexity, $O(r^{\lfloor \frac{K}{2} \rfloor} n^{\lceil \frac{K}{2} \rceil})$, required by the square reshaping (4.4), is always within a constant of it, much better for small r and $K \geq 4$ – e.g., by a multiplicative factor of $n^{\lfloor K/2 \rfloor - 1}$ when r is a constant. This is a significant improvement. However, there are also two clear limitations. First, no improvement is obtained for the case $K = 3$. Second, the improved sample complexity in Theorem 6 is still suboptimal compared to the nonconvex model (2.1).

It is also worth noting that for tensors with different lengths or ranks, Theorem 3 and Theorem 6 can be easily modified. It remains true that for a large class of tensors, our square reshaping is capable of reducing the number generic measurements required by SNN model. However, the comparison between sum-of-nuclear-norms and square norm becomes quite subtle then. Concrete instances can be definitely constructed so that square norm model does not have any advantage over the SNN model even for $K > 3$ (e.g. a tensor of size $1000 \times 10 \times 10 \times 10$ with Tucker rank $(1,1,1,1)$). On the other hand, our square norm model can sometimes be blessed by unbalanced tensors. For example, consider a tensor of size $4 \times 9 \times 12 \times 3$ with Tucker rank $(2,2,1,1)$. Then our reshaping matrix is a 36×36 square matrix with rank 1, which is a matrix with very good (perfect) conditions.

5   Simulation Results for Tensor Completion

Tensor completion attempts to reconstruct the low-rank tensor $\boldsymbol{\mathcal{X}}_0$ based on observations over a subset of its entries Ω. By imposing appropriate incoherence conditions (and modifying slightly arguments in [Gro11]), it is possible to prove recovery guarantees for each of the following programs:

$$\text{minimize} \quad \sum_{i=1}^{K} \lambda_i \|\boldsymbol{\mathcal{X}}_{(i)}\|_* \quad \text{subject to} \quad \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0]; \tag{5.1}$$

$$\text{minimize} \quad \|\boldsymbol{\mathcal{X}}\|_\square \quad \text{subject to} \quad \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0]. \tag{5.2}$$

Unlike the recovery problem under Gaussian random measurements, due to the lack of sharp upper bounds, we have no proof that our square norm formulation outperforms the SNN model here. However, our simulation results below strongly suggest that (5.2) also performs much better than (5.1) for tensor completion case.
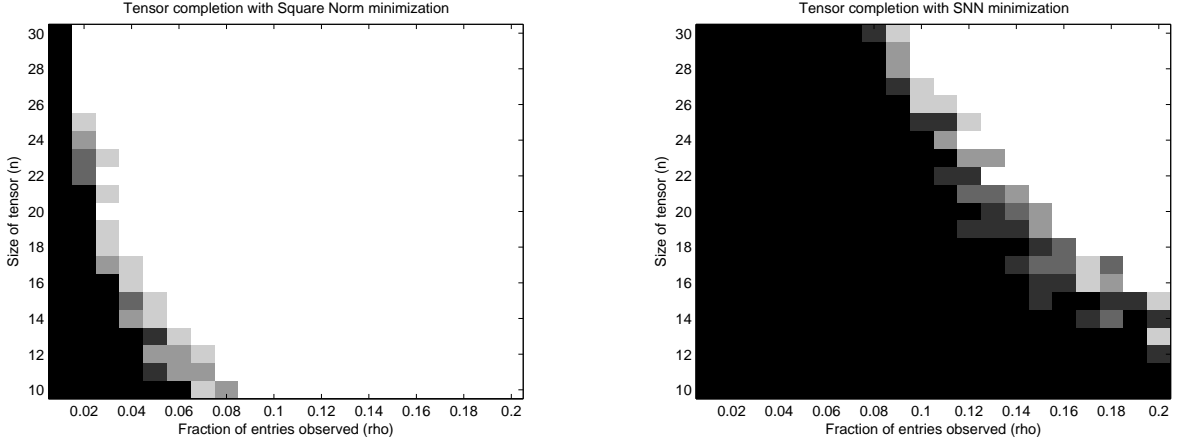

Figure 2: **Tensor completion.** The colormap indicates the fraction of correct recovery, which increases with brightness from certain failure (black) to certian success (white).

Our experiment is set up as follows. We generate a 4-way tensor, $\boldsymbol{\mathcal{X}}_0 \in \mathbb{R}^{n \times n \times n \times n}$, as

$$\boldsymbol{\mathcal{X}}_0 = [[\boldsymbol{\mathcal{C}}_0; \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3, \boldsymbol{U}_4]] = \boldsymbol{\mathcal{C}}_0 \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3 \times_4 \boldsymbol{U}_4,$$

where the core tensor $\boldsymbol{\mathcal{C}}_0 \in \mathbb{R}^{1 \times 1 \times 2 \times 2}$ has i.i.d. standard Gaussian entries, and matrices $\boldsymbol{U}_1, \boldsymbol{U}_2 \in \mathbb{R}^{n \times 1}$ and matrices $\boldsymbol{U}_3, \boldsymbol{U}_4 \in \mathbb{R}^{n \times 2}$, satisfying $\boldsymbol{U}_i^* \boldsymbol{U}_i = \boldsymbol{I}$, are drawn uniformly at random (by the command orth(randn($\cdot, \cdot$)) in Matlab). The observed entries are chosen uniformly with ratio ρ. We increase the problem size n from 10 to 30 with increment 1, and the observation ratio ρ from 0.01 to 0.2 with increment 0.01. For each (ρ, n)-pair, we simulate 5 test instances and declare a trial to be successful if the recovered $\boldsymbol{\mathcal{X}}^\star$ satisfies $\frac{\|\boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{X}}_0\|_F}{\|\boldsymbol{\mathcal{X}}_0\|_F} \leq 10^{-2}$. The optimization problems are solved using efficient first-order methods. Since (5.2) is in the form of standard matrix completion, we use the Augmented Lagrangian Method (ALM) proposed in [LCM10] to solve it. For the sum of nuclear norms minimization (5.1) with $\lambda_i = 1$, we implement the accelerated linearized Bregman

11

algorithm [HMG13], of which we include a detailed discussion in the appendix. Figure 2 plots the fraction of correct recovery for each pair (black = 0% and white = 100%). Clearly much larger white region is produced by square norm, which empirically suggests that (5.2) outperforms (5.1) for tensor completion problem.

6   Conclusion

In this paper, we establish several theoretical bounds for the problem of low-rank tensor recovery using random Gaussian measurements. For the nonconvex model (2.1), we show that $\big((2r)^K + 2nrK + 1\big)$ measurements are sufficient to recover any $\boldsymbol{\mathcal{X}}_0 \in \mathfrak{T}_r$ almost surely. We highlight that though the nonconvex recovery program is NP-hard in general, it does serve a baseline for evaluating tractable (convex) approaches. For the conventional convex surrogate sum-of-nuclear-norms (SNN) model (3.2), we prove a necessary condition that $\Omega(rn^{K-1})$ Gaussian measurements are required for reliable recovery. This lower bound is derived from our study on multi-structured object recovery under a very general setting, which can be applied to many scenarios. To narrow the apparent gap between the non-convex model and the SNN model, we unfold the tensor into a more balanced matrix while preserving its low-rank property, leading to our square-norm model (4.4). We prove that $O(r^{\lfloor \frac{K}{2} \rfloor} n^{\lceil \frac{K}{2} \rceil})$ measurements are sufficient to recover a tensor $\boldsymbol{\mathcal{X}}_0 \in \mathfrak{T}_r$ with high probability. Though the theoretical results only pertain to Gaussian measurements, our simulation result for tensor completion also suggests that square-norm model outperforms the SNN model.

Model	sample complexity
non-convex	$(2r)^K + 2nrK + 1$
SNN	$\Theta(rn^{K-1})$
square-norm	$O(r^{\lfloor \frac{K}{2} \rfloor} n^{\lceil \frac{K}{2} \rceil})$

Table 2: **Summary of results derived in our paper.**

Compared with $\Omega(rn^{K-1})$ measurements required by the sum-of-nuclear-norms model, the sample complexity, $O(r^{\lfloor \frac{K}{2} \rfloor} n^{\lceil \frac{K}{2} \rceil})$, required by the square reshaping (4.4), is always within a constant of it, much better for small r and $K \geq 4$. Although this is a significant improvement, compared with the nonconvex model (2.1), the improved sample complexity achieved by square norm model is however still suboptimal. It remails an open problem to obtain near-optimal convex relaxations for all $K > 2$.

More broadly speaking, to recover objects with multiple structures, regularizing with a combination of individual structure-inducing norms is proven to be substantially suboptimal (Theorem 5 and also [OJF$^+$12]). The resulting sample requirements tend to be much larger than the intrinsic degrees of freedom of the low-dimensional manifold that the structured signal lies in. Our square-norm model for the low-rank tensor recovery demonstrates the possibility that a better exploitation in those structures can significantly reduce this sample complexity. However, there are still no clear clues on how to intelligently utilize several simultaneous structures generally, and moreover how to design tractable method to recover multi-structured objects with near minimal number of measurements. These problems are definitely worth pursuing in future study and we hope that our work may also inspire researchers working in many other multi-structured recovery problems.

Acknowledgment

References

[AGH⁺12]   Anima Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *CoRR*, abs/1210.7559, 2012.

[ALMT13]   Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. *arXiv preprint arXiv:1303.6672*, 2013.

[Ame11]   Dennis Amelunxen. *Geometric analysis of the condition of the convex feasibility problem*. PhD thesis, PhD Thesis, Univ. Paderborn, 2011.

[BLM12]   M. Bayati, M. Lelarge, and A. Montanari. Universality in Polytope Phase Transitions and Message Passing Algorithms. *ArXiv e-prints*, 2012.

[BV04]   Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[CC12]   S. B. Cohen and M. Collins. Tensor decomposition for fast latent-variable PCFG parsing. In *Proceedings of NIPS*, 2012.

[CRPW12]   Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[CRT]   Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8).

[DN10]   N.D. Sidiropoulos D. Nion. Tensor algebra and multi-dimensional harmonic retrieval in signal processing for mimo radar. *IEEE Trans. on Signal Processing*, 58(11):5693–5705, 2010.

[Don06]   D. Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, 2006.

[DT09]   D. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Trans. Royal Society A*, 367(1906):4273–4293, 2009.

[ENP11]   Yonina C Eldar, Deanna Needell, and Yaniv Plan. Unicity conditions for low-rank matrix recovery. *arXiv preprint arXiv:1103.5479*, 2011.

[Faz02]   Maryam Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.

[GEK13]   N. Hao G. Ely, S. Aeron and M. E. Kilmer. 5d and 4d pre-stack seismic data completion using tensor nuclear norm (tnn). *preprint*, 2013.

[GQ12]   D. Goldfarb and Z. Qin. Robust low-rank tensor recovery: Models and algorithms. *preprint*, 2012.

[Gro11]   D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Info. Theory*, 57(3):1548–1566, 2011.

[GRY11]   Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

[HL09]   Christopher J. Hillar and Lek-Heng Lim. Most tensor problems are np hard. *CoRR*, abs/0911.1393, 2009.

[HMG13]   Bo Huang, Shiqian Ma, and Donald Goldfarb. Accelerated linearized bregman method. *Journal of Scientific Computing*, 54(2-3):428–453, 2013.

[KB09]   Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM Rev.*, 51(3):455–500, 2009.

[KS13]   Stanton A. Kreimer, N. and M. D. Sacchi. Nuclear norm minimization and tensor completion in exploration seismology. In *International Conference on Acoustics, Speech and Signal Processing*, 2013.

[LCM10]   Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[LL10]   Nan Li and Baoxin Li. Tensor completion for on-board compression of hyperspectral images. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 517–520, 2010.

[LMWY09]   Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. In *ICCV*, pages 2114–2121, 2009.

[LY12]   Ming-Jun Lai and Wotao Yin. Augmented l1 and nuclear-norm models with a globally linearly convergent algorithm. *arXiv preprint arXiv:1201.4615*, 2012.

[LYZY10]   Yin Li, Junchi Yan, Yue Zhou, and Jie Yang. Optimum subspace learning and error correction for tensors. In *ECCV*, pages 790–803, 2010.

[McC13]   Michael B McCoy. *A geometric analysis of convex demixing*. PhD thesis, California Institute of Technology, 2013.

[MSS06]   N. Mesgarani, M. Slaney, and S.A. Shamma. Discrimination of speech from non-speech based on multiscale spectro-temporal modulations. *IEEE Trans. Audio, Speech, and Language Processing*, 14(3):920–930, 2006.

[Nes83]   Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/k2). In *Doklady AN SSSR*, volume 269, pages 543–547, 1983.

[NRWY12]   S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Stat. Sci.*, 27(4):528–557, 2012.

[OJF+12]   Samet Oymak, Amin Jalali, Maryam Fazel, Yonina C Eldar, and Babak Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *arXiv preprint arXiv:1212.3753*, 2012.

[RFP10]   Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[RSV12]   Emile Richard, Pierre-André Savalle, and Nicolas Vayatis. Estimation of simultaneously sparse and low rank matrices. *arXiv preprint arXiv:1206.6474*, 2012.

[SDS10]   Marco Signoretto, Lieven De Lathauwer, and Johan AK Suykens. Nuclear norms for tensors and their use for convex multilinear estimation. *Submitted to Linear Algebra and Its Applications*, 43, 2010.

[SHKM13]   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*, 2013.

[STDLS13]   Marco Signoretto, Quoc Tran Dinh, Lieven Lathauwer, and JohanA.K. Suykens. Learning with tensors: a framework based on convex optimization and spectral regularization. *Machine Learning*, pages 1–49, 2013.

[SVdPDMS11] M. Signoretto, R. Van de Plas, B. De Moor, and J. A K Suykens. Tensor versus matrix completion: A comparison with application to spectral data. *IEEE SPL*, 18(7):403–406, 2011.

[THK10]   Ryota Tomioka, Kohei Hayashi, and Hisashi Kashima. Estimation of low-rank tensors via convex optimization. *arXiv preprint arXiv:1010.0789*, 2010.

[TSHK11]   Ryota Tomioka, Taiji Suzuki, Kohei Hayashi, and Hisashi Kashima. Statistical performance of convex tensor decomposition. *Advances in Neural Information Processing Systems (NIPS)*, page 137, 2011.

[Tuc66]   L. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[Ver07]   Roman Vershynin. Math 280 lectue notes. `http://www-personal.umich.edu/~romanv/teaching/2006-07/280/lec6.pdf`, 2007.

[Ver10]   Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[ZCCZ11]   Hui Zhang, Jian-Feng Cai, Lizhi Cheng, and Jubo Zhu. Strongly convex programming for exact matrix completion and robust principal component analysis. *arXiv preprint arXiv:1112.3946*, 2011.

A Proofs for Section 2

Proof of Lemma 1.

Proof. The arguments we used below are primarily adapted from [ENP11], where their interest is to establish the number of Gaussian measurements required to recover a low rank matrix by rank minimization.

Notice that every $\boldsymbol{D} \in \mathfrak{S}_{2r}$, and every i, $\langle \boldsymbol{\mathcal{G}}_i, \boldsymbol{D} \rangle$ is a standard Gaussian random variable, and so

$$\forall t > 0, \quad \mathbb{P}\left[\,|\langle \boldsymbol{\mathcal{G}}_i, \boldsymbol{D} \rangle| < t\,\right] < 2t \cdot \frac{1}{\sqrt{2\pi}} = t\sqrt{\frac{2}{\pi}}. \tag{A.1}$$

Let $\mathfrak{N}$ be an ε-net for $\mathfrak{S}_{2r}$ in terms of $\|\cdot\|_F$. Because the measurements are independent, for any fixed $\bar{\boldsymbol{D}} \in \mathfrak{S}_{2r}$,

$$\mathbb{P}\left[\,\left\|\mathcal{G}[\bar{\boldsymbol{D}}]\right\|_\infty < t\,\right] \;<\; \left(t\sqrt{2/\pi}\right)^m. \tag{A.2}$$

Moreover, for any $\boldsymbol{D} \in \mathfrak{S}_{2r}$, we have

$$\left\|\mathcal{G}[\boldsymbol{D}]\right\|_\infty \;\geq\; \max_{\bar{\boldsymbol{D}} \in \mathfrak{N}} \left\{\, \left\|\mathcal{G}[\bar{\boldsymbol{D}}]\right\|_\infty - \left\|\mathcal{G}\right\|_{F \to \infty} \left\|\bar{\boldsymbol{D}} - \boldsymbol{D}\right\|_F \,\right\} \tag{A.3}$$

$$\geq\; \min_{\bar{\boldsymbol{D}} \in \mathfrak{N}} \left\{\left\|\mathcal{G}[\bar{\boldsymbol{D}}]\right\|_\infty\right\} - \varepsilon \left\|\mathcal{G}\right\|_{F \to \infty}. \tag{A.4}$$

Hence,

$$\mathbb{P}\left[\inf_{\boldsymbol{D} \in \mathfrak{S}_{2r}} \left\|\mathcal{G}[\boldsymbol{D}]\right\|_\infty < \varepsilon \log(1/\varepsilon)\right]$$

$$\leq\; \mathbb{P}\left[\min_{\boldsymbol{D} \in \mathfrak{N}} \left\|\mathcal{G}[\boldsymbol{D}]\right\|_\infty < 2\varepsilon \log(1/\varepsilon)\right] + \mathbb{P}\left[\left\|\mathcal{G}\right\|_{F \to \infty} > \log(1/\varepsilon)\right]$$

$$\leq\; \#\mathfrak{N} \times \left(2\sqrt{2/\pi} \times \varepsilon \log(1/\varepsilon)\right)^m + \mathbb{P}\left[\left\|\mathcal{G}\right\|_{F \to \infty} > \log(1/\varepsilon)\right]$$

$$\leq\; \beta^d (2\sqrt{2/\pi})^m \varepsilon^{m-d} \log(1/\varepsilon)^m + \mathbb{P}\left[\left\|\mathcal{G}\right\|_{F \to \infty} > \log(1/\varepsilon)\right]. \tag{A.5}$$

Since $m \geq d+1$, (A.5) goes to zero as $\varepsilon \searrow 0$. Hence, taking a sequence of decreasing ε, we can show that $\mathbb{P}\left[\inf_{\boldsymbol{D} \in \mathfrak{S}_{2r}} \left\|\mathcal{G}[\boldsymbol{D}]\right\|_\infty = 0\right] \leq t$ for every positive t, establishing the result. $\qquad\square$

Proof of Lemma 2.

Proof. This follows from the basic fact that for any tensor $\boldsymbol{\mathcal{X}}$ and matrix $\boldsymbol{U}$ of compatible size,

$$\left\|\boldsymbol{\mathcal{X}} \times_k \boldsymbol{U}\right\|_F \;\leq\; \left\|\boldsymbol{U}\right\| \left\|\boldsymbol{\mathcal{X}}\right\|_F, \tag{A.6}$$

which can be established by direct calculation. Write

$$\left\|[[\boldsymbol{\mathcal{C}}; \boldsymbol{U}_1, \ldots, \boldsymbol{U}_K]] - [[\boldsymbol{\mathcal{C}}'; \boldsymbol{U}_1', \ldots, \boldsymbol{U}_K']]\right\|_F$$

$$\leq\; \left\|[[\boldsymbol{\mathcal{C}}; \boldsymbol{U}_1, \ldots, \boldsymbol{U}_K]] - [[\boldsymbol{\mathcal{C}}'; \boldsymbol{U}_1, \ldots, \boldsymbol{U}_K]]\right\|_F$$

$$+\; \left\|\sum_{i=1}^{K} [[\boldsymbol{\mathcal{C}}'; \boldsymbol{U}_1', \ldots, \boldsymbol{U}_i', \boldsymbol{U}_{i+1}, \ldots \boldsymbol{U}_k]] - [[\boldsymbol{\mathcal{C}}'; \boldsymbol{U}_1', \ldots, \boldsymbol{U}_{i-1}', \boldsymbol{U}_i, \ldots \boldsymbol{U}_K]]\right\|_F$$

$$\leq\; \left\|\boldsymbol{\mathcal{C}} - \boldsymbol{\mathcal{C}}'\right\|_F + \sum_{i=1}^{K} \left\|\boldsymbol{U}_i - \boldsymbol{U}_i'\right\|,$$

17

where the first inequality follows from triangle inequality and the second inequality follows from the fact that $\|\mathcal{C}\|_F = 1$, $\|U_j\| = 1$, $U_i^* U_i = I$ and $U_i'^* U_i' = I$. $\qquad\square$

Proof of Lemma 3.

Proof. The idea of this proof is to construct a net for each component of the Tucker decomposition and then combine those nets to form a *compound* net with the desired cardinality.

Denote $\mathfrak{C} = \{\mathcal{C} \in \mathbb{R}^{2r \times 2r \times \cdots \times 2r} \mid \|\mathcal{C}\|_F = 1\}$ and $\mathcal{O} = \{U \in \mathbb{R}^{n \times r} \mid U^* U = I\}$. Clearly, for any $\mathcal{C} \in \mathfrak{C}$, $\|\mathcal{C}\|_F = 1$, and for any $U \in \mathcal{O}$, $\|U\| = 1$. Thus by [Ver07, Prop. 4] and [Ver10, Lemma 5.2], there exists an $\frac{\varepsilon}{K+1}$-net $\mathfrak{C}'$ covering $\mathfrak{C}$ with respect to the Frobenius norm such that $\#\mathfrak{C}' \leq (\frac{3(K+1)}{\varepsilon})^{(2r)^K}$, and there exists an $\frac{\varepsilon}{K+1}$-net $\mathcal{O}'$ covering $\mathcal{O}$ with respect to the operator norm such that $\#\mathcal{O}' \leq (\frac{3(K+1)}{\varepsilon})^{2nr}$. Construct

$$\mathfrak{S}_{2r}' = \{[[\mathcal{C}'; U_1', \ldots, U_K']] \mid \mathcal{C}' \in \mathfrak{C}', \ U_i' \in \mathcal{O}'\}.$$

Clearly $\#\mathfrak{S}_{2r}' \leq \left(\frac{3(K+1)}{\varepsilon}\right)^{(2r)^K + 2nrK}$. The rest is to show that $\mathfrak{S}_{2r}'$ is indeed an ε-net covering $\mathfrak{S}_{2r}$ with respect to the Frobenius norm.

For any fixed $\mathcal{D} = [[\mathcal{C}; U_1, \cdots, U_K]] \in \mathfrak{S}_{2r}$ where $\mathcal{C} \in \mathfrak{C}$ and $U_i \in \mathcal{O}$, by our constructions above, there exist $\mathcal{C}' \in \mathfrak{C}'$ and $U_i' \in \mathcal{O}'$ such that $\|\mathcal{C} - \mathcal{C}'\|_F \leq \frac{3(K+1)}{\varepsilon}$ and $\|U_i - U_i'\| \leq \frac{3(K+1)}{\varepsilon}$. Then $\mathcal{D}' = [[\mathcal{C}'; U_1', \cdots, U_K']] \in \mathfrak{S}_{2r}'$ is within ε-distance from $\mathcal{D}$, since by the triangle inequality derived in Lemma 2, we have

$$\left\|\mathcal{D} - \mathcal{D}'\right\|_F = \left\|[[\mathcal{C}; U_1, \ldots, U_K]] - [[\mathcal{C}'; U_1', \ldots, U_K']]\right\|_F \leq \left\|\mathcal{C} - \mathcal{C}'\right\|_F + \sum_{i=1}^K \left\|U_i - U_i'\right\| \leq \varepsilon.$$

This completes the proof. $\qquad\square$

B   Proofs for Section 3

Proof of Corollary 4.

Proof. Denote $\lambda = \delta(\mathcal{C}) - m$. Then following [ALMT13, Thm. 7.1], we have

$$
\begin{aligned}
\mathbb{P}\left[C \cap \mathrm{null}(\mathcal{G}) = \{\mathbf{0}\}\right] &\leq 4\exp\left(-\frac{\lambda^2/8}{\min\{\delta(\mathcal{C}), \delta(\mathcal{C}^\circ)\} + \lambda}\right) \\
&\leq 4\exp\left(-\frac{\lambda^2/8}{\delta(\mathcal{C}) + \lambda}\right) \\
&\leq 4\exp\left(-\frac{(\delta(C) - m)^2}{16\delta(C)}\right).
\end{aligned}
$$

$\qquad\square$

Proof of Lemma 5.

Proof. Denote $\mathrm{circ}(\boldsymbol{e}_n, \theta)$ as $\mathrm{circ}_n(\theta)$, where $\boldsymbol{e}_n$ is the nth standard basis for $\mathbb{R}^n$. Since $\delta\big(\mathrm{circ}(\boldsymbol{x}_0, \theta)\big) = \delta\big(\mathrm{circ}(\boldsymbol{e}_n, \theta)\big)$, it is sufficient to prove $\delta\big(\mathrm{circ}_n(\theta)\big) \leq n \sin^2\theta + 2$.

Let us first consider the case where n is *even*. Define a discrete random variable V supporting on $\{0, 1, 2, \cdots, n\}$ with probability mass function $\mathbb{P}\left[V = k\right] = v_k$. Here v_k denotes the k-th intrinsic volumes of $\mathrm{circ}_n(\theta)$. As specified in [Ame11, Ex. 4.4.8], we have

$$v_k = \frac{1}{2}\binom{\frac{1}{2}(n-2)}{\frac{1}{2}(k-1)}\sin^{k-1}(\theta)\cos^{n-k-1}(\theta) \qquad \text{for } k = 1, 2, \cdots, n-1.$$

From [ALMT13, Prop. 5.11], we know that

$$\delta\big(\mathrm{circ}_n(\theta)\big) = \mathbb{E}\left[V\right] = \sum_{k=1}^{n}\mathbb{P}\left[V \geq k\right].$$

Moreover, by the interlacing result from [ALMT13, Prop. 5.6] and the fact that $\mathbb{P}\left[V \geq 2k\right] = \mathbb{P}\left[V \geq 2k-1\right] - \mathbb{P}\left[V = 2k-1\right]$, we have

$$
\begin{aligned}
\mathbb{P}\left[V \geq 1\right] &\leq 2\mathbb{P}\left[V = 1\right] + 2\mathbb{P}\left[V = 3\right] + \cdots + 2\mathbb{P}\left[V = n-1\right], \\
\mathbb{P}\left[V \geq 2\right] &\leq \mathbb{P}\left[V = 1\right] + 2\mathbb{P}\left[V = 3\right] + \cdots + 2\mathbb{P}\left[V = n-1\right];
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{P}\left[V \geq 3\right] &\leq 2\mathbb{P}\left[V = 3\right] + 2\mathbb{P}\left[V = 5\right] + \cdots + 2\mathbb{P}\left[V = n-1\right], \\
\mathbb{P}\left[V \geq 4\right] &\leq \mathbb{P}\left[V = 3\right] + 2\mathbb{P}\left[V = 5\right] + \cdots + 2\mathbb{P}\left[V = n-1\right];
\end{aligned}
$$

$$\vdots \qquad\qquad \vdots \quad \vdots$$

$$
\begin{aligned}
\mathbb{P}\left[V \geq n-1\right] &\leq 2\mathbb{P}\left[V = n-1\right], \\
\mathbb{P}\left[V \geq n\right] &\leq \mathbb{P}\left[V = n-1\right].
\end{aligned}
$$

Summing up the above inequalities, we have

$$
\begin{aligned}
\mathbb{E}\left[V\right] &= \sum_{k=1}^{n}\mathbb{P}\left[V \geq k\right] \\
&\leq \sum_{k=1,3,\cdots,n-1} 2(k-1)v_k + \sum_{k=1,3,\cdots,n-1} 3v_k \\
&\leq (n-2)\sin^2\theta + \frac{3}{2}\sum_{k=0}^{n}v_k \\
&\leq (n-2)\sin^2\theta + \frac{3}{2} = n\sin^2\theta + 2\cos^2\theta - \frac{1}{2},
\end{aligned}
$$

where the second last inequality follows the observations that $\sum_{k=1,3,\cdots,n-1}\frac{k-1}{2}\cdot(2v_k) = \mathbb{E}\left[Bin(\frac{n-2}{2}, \sin^2\theta)\right]$ and $\sum_{k=0}^{n}v_k \geq \sum_{k=1,3,\cdots,n-1}2v_k$ again by the interlacing result [ALMT13, Prop. 5.6].

Suppose n is *odd*. Since the intersection of $\mathrm{circ}_{n+1}(\theta)$ with any n-dimensional linear subspace containing $\boldsymbol{e}_{n+1}$ is an isometric image of $\mathrm{circ}_n(\theta)$, by [ALMT13, Prop. 4.1], we have

$$\delta(\mathrm{circ}_n(\theta)) = \delta(\mathrm{circ}_n(\theta) \times \{\boldsymbol{0}\}) \leq \delta(\mathrm{circ}_{n+1}(\theta)) \leq (n+1)\sin^2\theta + 2\cos^2\theta - \frac{1}{2} \leq n\sin^2\theta + \cos^2\theta + \frac{1}{2}.$$

Thus, taking both cases (n is even and n is odd) into consideration, we have

$$\delta\big(\mathrm{circ}_n(\theta)\big) \le n\sin^2\theta + \cos^2\theta + \frac{1}{2} \le n\sin^2\theta + 2.$$

$\square$

Proof of Theorem 5.

Proof. Notice that for any fixed $m > 0$, the function $f : t \to 4\exp\left(-\frac{(t-m)^2}{16t}\right)$ is decreasing for $t \ge m$. Then due to Corollary 4 and the fact that $\delta(\mathcal{C}) \ge \kappa - 2 \ge m$, we have

$$
\begin{aligned}
\mathbb{P}\left[\,\boldsymbol{x}_0 \text{ is the unique optimal solution to (3.3)}\,\right] &= \mathbb{P}\left[\,C \cap \mathrm{null}(\mathcal{G}) = \{\boldsymbol{0}\}\,\right] \\
&\le 4\exp\left(-\frac{(\delta(C) - m)^2}{16\delta(C)}\right) \\
&\le 4\exp\left(-\frac{(\kappa - m - 2)^2}{16\,(\kappa - 2)}\right).
\end{aligned}
$$

$\square$

C   Proofs for Section 4

Proof of Lemma 6.

Proof. (1) By the definition of $\boldsymbol{\mathcal{X}}_{[j]}$, it is sufficient to prove that the vectorization of the right hand side of (4.2) equals $\mathrm{vec}(\boldsymbol{\mathcal{X}}_{(1)})$.

Since $\boldsymbol{\mathcal{X}} = \sum_{i=1}^{r} \lambda_i \boldsymbol{a}_i^{(1)} \circ \boldsymbol{a}_i^{(2)} \circ \cdots \circ \boldsymbol{a}_i^{(K)}$, we have

$$
\begin{aligned}
\mathrm{vec}(\boldsymbol{\mathcal{X}}_{(1)}) &= \mathrm{vec}\Big(\sum_{i=1}^{r} \lambda_i \boldsymbol{a}_i^{(1)} \circ (\boldsymbol{a}_i^{(K)} \otimes \boldsymbol{a}_i^{(K-1)} \otimes \cdots \otimes \boldsymbol{a}_i^{(2)})\Big) \\
&= \sum_{i=1}^{r} \lambda_i \mathrm{vec}\big(\boldsymbol{a}_i^{(1)} \circ (\boldsymbol{a}_i^{(K)} \otimes \boldsymbol{a}_i^{(K-1)} \otimes \cdots \otimes \boldsymbol{a}_i^{(2)})\big) \\
&= \sum_{i=1}^{r} \lambda_i (\boldsymbol{a}_i^{(K)} \otimes \boldsymbol{a}_i^{(K-1)} \otimes \cdots \otimes \boldsymbol{a}_i^{(2)} \otimes \boldsymbol{a}_i^{(1)}),
\end{aligned}
$$

where the last equality follows from the fact that $\mathrm{vec}(\boldsymbol{a} \circ \boldsymbol{b}) = \boldsymbol{b} \otimes \boldsymbol{a}$. Similarly, we can derive that the vectorization of the right hand side of (4.2),

$$
\begin{aligned}
&\mathrm{vec}\Big(\sum_{i=1}^{r} \lambda_i (\boldsymbol{a}_i^{(j)} \otimes \boldsymbol{a}_i^{(j-1)} \otimes \cdots \otimes \boldsymbol{a}_i^{(1)}) \circ (\boldsymbol{a}_i^{(K)} \otimes \boldsymbol{a}_i^{(K-1)} \cdots \otimes \boldsymbol{a}_i^{(j+1)})\Big) \\
&= \sum_{i=1}^{r} \lambda_i \mathrm{vec}\big((\boldsymbol{a}_i^{(j)} \otimes \boldsymbol{a}_i^{(j-1)} \otimes \cdots \otimes \boldsymbol{a}_i^{(1)}) \circ (\boldsymbol{a}_i^{(K)} \otimes \boldsymbol{a}_i^{(K-1)} \cdots \otimes \boldsymbol{a}_i^{(j+1)})\big) \\
&= \sum_{i=1}^{r} \lambda_i (\boldsymbol{a}_i^{(K)} \otimes \boldsymbol{a}_i^{(K-1)} \otimes \cdots \otimes \boldsymbol{a}_i^{(2)} \otimes \boldsymbol{a}_i^{(1)}) \\
&= \mathrm{vec}(\boldsymbol{\mathcal{X}}_{(1)}).
\end{aligned}
$$

Thus, equation (4.2) is valid.

(2) The above argument can be easily adapted to prove the second claim. Since $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{C}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \cdots \times_K \boldsymbol{U}_K$, we have

$$
\begin{aligned}
\mathrm{vec}(\boldsymbol{\mathcal{X}}_{(1)}) &= \mathrm{vec}\left(\boldsymbol{U}_1 \, \boldsymbol{\mathcal{C}}_{(1)} \, (\boldsymbol{U}_K \otimes \boldsymbol{U}_{K-1} \otimes \cdots \otimes \boldsymbol{U}_2)^* \right) \\
&= (\boldsymbol{U}_K \otimes \boldsymbol{U}_{K-1} \otimes \cdots \otimes \boldsymbol{U}_1) \, \mathrm{vec}(\boldsymbol{\mathcal{C}}_{(1)}),
\end{aligned}
$$

where the last equality follows from the fact that $\mathrm{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^* \otimes \boldsymbol{A})\mathrm{vec}(\boldsymbol{B})$. Similarly, we can derive that the vectorization of the right hand side of (4.3),

$$
\begin{aligned}
&\mathrm{vec}\left((\boldsymbol{U}_j \otimes \boldsymbol{U}_{j-1} \otimes \cdots \otimes \boldsymbol{U}_1) \, \boldsymbol{\mathcal{C}}_{[j]} \, (\boldsymbol{U}_K \otimes \boldsymbol{U}_{K-1} \otimes \cdots \otimes \boldsymbol{U}_{j+1})^* \right) \\
&= (\boldsymbol{U}_K \otimes \boldsymbol{U}_{K-1} \otimes \cdots \otimes \boldsymbol{U}_1) \, \mathrm{vec}(\boldsymbol{\mathcal{C}}_{[j]}) \\
&= (\boldsymbol{U}_K \otimes \boldsymbol{U}_{K-1} \otimes \cdots \otimes \boldsymbol{U}_1) \, \mathrm{vec}(\boldsymbol{\mathcal{C}}_{(1)}) \\
&= \mathrm{vec}(\boldsymbol{\mathcal{X}}_{(1)}).
\end{aligned}
$$

Thus, equation (4.3) is valid. $\qquad\square$

D   Algorithms for Section 5

In this section, we will discuss in detail our implementation of accelerated linearized Bregman algorithm for the following problem:

$$
\mathrm{minimize}_{\boldsymbol{\mathcal{X}}} \quad \sum_{i=1}^{K} \|\boldsymbol{\mathcal{X}}_{(i)}\|_* \quad \text{subject to} \quad \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0]. \tag{D.1}
$$

By introducing auxiliary variable $\boldsymbol{\mathcal{W}}$ and splitting $\boldsymbol{\mathcal{X}}$ into $\boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_2, \cdots, \boldsymbol{\mathcal{X}}_K$, it can be easily verified that problem (D.1) is equivalent to

$$
\begin{aligned}
\min_{(\{\boldsymbol{\mathcal{X}}_i\}, \boldsymbol{\mathcal{W}})} \quad & \sum_{i=1}^{K} \|(\boldsymbol{\mathcal{X}}_i)_{(i)}\|_* \\
\text{s.t.} \quad & \boldsymbol{\mathcal{X}}_i = \boldsymbol{\mathcal{W}}, \quad i = 1, 2, \cdots, K, \\
& \mathcal{P}_\Omega[\boldsymbol{\mathcal{W}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0],
\end{aligned} \tag{D.2}
$$

whose objective function is now separable.

The accelerated linearized Bregman (ALB) algorithm, proposed in [HMG13], is an efficient first-order method designed for solving convex optimization problems with nonsmooth objective functions and linear constraints. It has been successfully applied to solve ℓ^1 and nuclear norm minimization problems [HMG13]. The ALB algorithm solves nonsmooth problem by firstly smoothing the objective function (e.g. adding a small l_2 perturbation), and then exploiting Nesterov's accelerated scheme [Nes83] to the dual problem, which can be verified to be unconstrained and Lipschitz differentiable. In Algorithm 1, we describe our ALM algorithm adapted to problem (D.2). Algorithm 1

solves exactly the smoothed version of problem (D.2):

$$
\min_{(\{\boldsymbol{\mathcal{X}}_i\},\boldsymbol{\mathcal{W}})} \quad \sum_{i=1}^{K} \left(\|(\boldsymbol{\mathcal{X}}_i)_{(i)}\|_* + \frac{1}{2\mu}\|(\boldsymbol{\mathcal{X}}_i)_{(i)}\|_F^2 \right) + \frac{1}{2\mu}\|\boldsymbol{\mathcal{W}}\|_F^2
$$

$$
\begin{aligned}
\text{s.t.} \quad & \boldsymbol{\mathcal{X}}_i = \boldsymbol{\mathcal{W}}, \quad i = 1, 2, \cdots, K, \\
& \mathcal{P}_\Omega[\boldsymbol{\mathcal{W}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0],
\end{aligned} \tag{D.3}
$$

where we denote $\boldsymbol{\mathcal{Y}}_i$ as the dual variable for the constraint $\boldsymbol{\mathcal{X}}_i = \boldsymbol{\mathcal{W}}$ and denote $\boldsymbol{\mathcal{Z}}$ as the dual variable for the last constraint $\mathcal{P}_\Omega[\boldsymbol{\mathcal{W}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0]$. Since the objective function in (D.3) is separable, each setup of the ALB algorithm is easy to solve as we can see from Algorithm 1 [4].

Algorithm 1: accelerated linearized Bregman algorithm for SNN model (D.1)

---

1   **Initialization:** $\boldsymbol{\mathcal{Y}}_i^0 = \tilde{\boldsymbol{\mathcal{Y}}}_i^0 = \mathbf{0}$ for each $i \in [K]$, $\boldsymbol{\mathcal{Z}}^0 = \tilde{\boldsymbol{\mathcal{Z}}}^0 = \mathbf{0}$, $\mu > 0$, $\tau > 0$, $t_0 = 1$;

2   **for** $k = 0, 1, 2, \cdots$ **do**

3   **for** $i = 1, 2, \cdots, K$ **do**

4         $\boldsymbol{\mathcal{X}}_i^{k+1} = \mu \cdot \text{Shrinkage}(\boldsymbol{\mathcal{Y}}_i^k, 1)$;

5      $\boldsymbol{\mathcal{W}}^{k+1} = \mu \cdot \left(\mathcal{P}_\Omega\left[\boldsymbol{\mathcal{Z}}^k \right] - \sum_i \boldsymbol{\mathcal{Y}}_i^k \right)$;

6   **for** $i = 1, 2, \cdots, K$ **do**

7         $\tilde{\boldsymbol{\mathcal{Y}}}_i^k = \boldsymbol{\mathcal{Y}}_i^k - \tau \cdot \left(\boldsymbol{\mathcal{X}}_i^{k+1} - \boldsymbol{\mathcal{W}}^{k+1} \right)$;

8      $\tilde{\boldsymbol{\mathcal{Z}}}^k = \boldsymbol{\mathcal{Z}}^k - \tau \cdot \mathcal{P}_\Omega\left[\boldsymbol{\mathcal{W}}^{k+1} - \boldsymbol{\mathcal{X}}_0 \right]$;

9      $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

10   **for** $i = 1, 2, \cdots, K$ **do**

11        $\boldsymbol{\mathcal{Y}}_i^{k+1} = \tilde{\boldsymbol{\mathcal{Y}}}_i^k + \frac{t_k-1}{t_{k+1}}\left(\tilde{\boldsymbol{\mathcal{Y}}}_i^k - \tilde{\boldsymbol{\mathcal{Y}}}_i^{k-1} \right)$;

12     $\boldsymbol{\mathcal{Z}}^{k+1} = \tilde{\boldsymbol{\mathcal{Z}}}^k + \frac{t_k-1}{t_{k+1}}\left(\tilde{\boldsymbol{\mathcal{Z}}}^k - \tilde{\boldsymbol{\mathcal{Z}}}^{k-1} \right)$;

For our numerical experiment ($K = 4$), we choose smoothing parameter $\mu = 50\|\boldsymbol{\mathcal{X}}_0\|_F$ and step size $\tau = \frac{1}{5\mu}$. Empirically, we observe that larger values of μ do not result in a better recovery performance. This is consistent with the theoretical results established in [LY12, ZCCZ11].

---

[4] The Shrinkage operator in line 4 of Algorithm 1 performs the regular shrinkage on the singular values of the ith unfolding matrix of $\boldsymbol{\mathcal{Y}}_i^k$, i.e. $(\boldsymbol{\mathcal{Y}}_i^k)_{(i)}$, and then folds the resulting matrix back into tensor.