Square Deal: Lower Bounds and Improved Relaxations for Tensor Recovery

Cun Mu[1] CM3052@COLUMBIA.EDU
Bo Huang[1] BH2359@COLUMBIA.EDU
John Wright[2] JOHNWRIGHT@EE.COLUMBIA.EDU
Donald Goldfarb[1] GOLDFARB@COLUMBIA.EDU
[1]Department of Industrial Engineering and Operations Research, Columbia University
[2]Department of Electrical Engineering, Columbia University

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 (SNN) of the unfolding matrices of the tensor. We show that this approach can be substantially suboptimal: reliably recovering a K-way $n \times n \times \cdots \times n$ tensor of Tucker rank $(r, r, \ldots, r)$ 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 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. The lower bound for the SNN model follows from our new result on recovering signals with multiple structures (e.g. sparse, low rank), which indicates the significant suboptimality of the common approach of minimizing the sum of individual sparsity inducing norms (e.g. ℓ_1, nuclear norm). Our new tractable formulation for low-rank tensor recovery shows how the sample complexity can be reduced by designing convex regularizers that exploit 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 (Kolda & Bader, 2009), 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 (Anandkumar et al., 2012), classifying audio (Mesgarani et al., 2006), mining text (Cohen & Collins, 2012), processing radar signals (Nion & Sidiropoulos, 2010), multilinear multitask learning (Romera-Paredes et al., 2013) , to name a few.

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 recovering $\boldsymbol{\mathcal{X}}$ from $\boldsymbol{z}$ 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 (Candès et al., 2006) or low-rankness for matrices (Recht et al., 2010) 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 (Negahban et al., 2012; Chandrasekaran et al., 2012; Amelunxen et al., 2013). 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 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 (Hillar & Lim, 2013). For example, even computing a tensor's (CP) rank,

$$\operatorname{rank}_{\mathrm{cp}}(\boldsymbol{\mathcal{X}}) := \min\Big\{r \mid \boldsymbol{\mathcal{X}} = \sum\nolimits_{i=1}^{r} \boldsymbol{a}_1^{(i)} \circ \cdots \circ \boldsymbol{a}_K^{(i)}\Big\},$$

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 re-

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

searchers have studied how to recover tensors of small *Tucker rank* (Tucker, 1966). The Tucker rank of a K-way tensor $\boldsymbol{\mathcal{X}}$ is a K-dimensional vector whose i-th entry is the (matrix) rank of the mode-i unfolding $\boldsymbol{\mathcal{X}}_{(i)}$ of $\boldsymbol{\mathcal{X}}$:

$$\operatorname{rank}_{\text{tc}}(\boldsymbol{\mathcal{X}}) := \big(\operatorname{rank}(\boldsymbol{\mathcal{X}}_{(1)}), \cdots, \operatorname{rank}(\boldsymbol{\mathcal{X}}_{(K)})\big). \quad (1.1)$$

Here, the matrix $\boldsymbol{\mathcal{X}}_{(i)} \in \mathbb{R}^{n_i \times \prod_{j \neq i} n_j}$ is obtained by concatenating all the mode-i fibers of $\boldsymbol{\mathcal{X}}$ as column vectors. Each *mode-i fiber* is an n_i-dimensional vector obtained by fixing every index of $\boldsymbol{\mathcal{X}}$ but the i-th one. The Tucker rank of $\boldsymbol{\mathcal{X}}$ 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 (1.1) suggests a natural, tractable convex approach to recovering low-rank tensors: seek the $\boldsymbol{\mathcal{X}}$ that minimizes $\sum_i \lambda_i \left\|\boldsymbol{\mathcal{X}}_{(i)}\right\|_*$ out of all $\boldsymbol{\mathcal{X}}$ satisfying $\mathcal{G}[\boldsymbol{\mathcal{X}}] = \boldsymbol{z}$. We will refer to this as the *sum-of-nuclear-norms (SNN)* model. Originally proposed in (Liu et al., 2009), this approach has been widely studied (Gandy et al., 2011; Signoretto et al., 2010; 2013; Tomioka et al., 2011) and applied to various datasets in imaging (Semerci et al., 2013; Kreimer & Sacchi, 2013; Li & Li, 2010; Li et al., 2010).

Perhaps surprisingly, we show that this natural approach can be substantially suboptimal. Moreover, we will suggest a simple new convex regularizer with provably better performance. Suppose $n_1 = \cdots = n_K = n$, and $\operatorname{rank}_{\text{tc}}(\boldsymbol{\mathcal{X}}) \preceq (r, r, \ldots, r)$. Let $\mathfrak{T}_r$ denote the set of all such tensors.[1] We will consider the problem of estimating an element $\boldsymbol{\mathcal{X}}_0$ of $\mathfrak{T}_r$ from Gaussian measurements $\mathcal{G}$ (i.e., $\boldsymbol{z}_i = \langle \boldsymbol{\mathcal{G}}_i, \boldsymbol{\mathcal{X}} \rangle$, where $\boldsymbol{\mathcal{G}}_i$ has i.i.d. standard normal entries). To describe a generic tensor in $\mathfrak{T}_r$, we need at most $r^K + rnK$ parameters. In Section 2, we show that a certain nonconvex strategy can recover all $\boldsymbol{\mathcal{X}} \in \mathfrak{T}_r$ exactly when $m > (2r)^K + 2nrK$. In contrast, the best known theoretical guarantee for SNN minimization, due to Tomioka et al. (2011), shows that $\boldsymbol{\mathcal{X}}_0 \in \mathfrak{T}_r$ 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 reshaping* model, which reduces the required number of measurements to $O(r^{\lfloor K/2 \rfloor} n^{\lceil K/2 \rceil})$. For $K > 3$, we obtain an improvement of a multiplicative factor polynomial in n.

[1] To keep the presentation in this paper compact, we state most of our results regarding tensors in $\mathfrak{T}_r$, although it is not difficult to modify them for general tensors.

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 (Donoho, 2006), either directly as a measurement strategy, or indirectly due to universality phenomena (Bayati et al., 2012). 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 bounds on the number of observations required. In Section 5, we demonstrate that our qualitative conclusions carry over to more realistic measurement models, such as random subsampling (Liu et al., 2009). We expect our results to be of great interest for a wide range of problems in tensor completion (Liu et al., 2009), robust tensor recovery/decomposition (Li et al., 2010; Goldfarb & Qin, 2014) and sensing.

Our technical approach 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. (2012): 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 novel geometric framework of (Amelunxen et al., 2013). Compared to (Oymak et al., 2012), our result pertains to a more general family of regularizers, and gives sharper constants. In addition, for low-rank tensor recovery problem, we demonstrate the possibility to reduce the number of generic measurements through a new convex regularizer that exploits 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 number of measurements.

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:

$$\min_{(\text{w.r.t. } \mathbb{R}^K_+)} \operatorname{rank}_{\text{tc}}(\boldsymbol{\mathcal{X}}) \quad \text{s.t.} \quad \mathcal{G}[\boldsymbol{\mathcal{X}}] = \mathcal{G}[\boldsymbol{\mathcal{X}}_0]. \quad (2.1)$$

In vector optimization, a feasible point is called *Pareto optimal* if no other feasible point dominates it in every criterion. Similarly, 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, i.e.

the set $\{\boldsymbol{\mathcal{X}}' \neq \boldsymbol{\mathcal{X}}_0 \mid \mathcal{G}[\boldsymbol{\mathcal{X}}'] = \mathcal{G}[\boldsymbol{\mathcal{X}}_0],\ \mathrm{rank_{tc}}(\boldsymbol{\mathcal{X}}') \preceq_{\mathbb{R}^K_+} \mathrm{rank_{tc}}(\boldsymbol{\mathcal{X}}_0)\}$ is empty.[2]

The recovery performance of (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 $\mathcal{G}[\boldsymbol{\mathcal{X}}'] = \mathcal{G}[\boldsymbol{\mathcal{X}}_0]$. 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 is implied by the condition $\mathrm{null}(\mathcal{G}) \cap \mathfrak{T}_{2r} = \{\mathbf{0}\}$. So, if $\mathrm{null}(\mathcal{G}) \cap \mathfrak{T}_{2r} = \{\mathbf{0}\}$, (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,* $\mathrm{null}(\mathcal{G}) \cap \mathfrak{T}_{2r} = \{\mathbf{0}\}$*, 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 present in the appendix.

Although problem (2.1) is not tractable, it can serve as a baseline for understanding how many generic measurements are required to recover $\boldsymbol{\mathcal{X}}_0$, and thus provide a benchmark for tractable formulations.

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 (Fazel, 2002; Recht et al., 2010; Gross, 2011). It seems natural, then, to replace the ranks in (2.1) with nuclear norms. Due to convexity, the resulting vector optimization problem can be solved by the following scalar optimization:

$$\min_{\boldsymbol{\mathcal{X}}} \quad \sum_{i=1}^{K} \lambda_i \|\boldsymbol{\mathcal{X}}_{(i)}\|_* \quad \text{s.t.} \quad \mathcal{G}[\boldsymbol{\mathcal{X}}] = \mathcal{G}[\boldsymbol{\mathcal{X}}_0], \qquad (3.1)$$

where $\boldsymbol{\lambda} \geq \mathbf{0}$. The optimization (3.1) was first introduced by (Liu et al., 2009) and has been used successfully in applications in imaging (Semerci et al., 2013; Kreimer & Sacchi, 2013; Li & Li, 2010; Ely et al., 2013; Li et al., 2010). Similar convex relaxations have been considered in a number of theoretical and algorithmic works (Gandy et al., 2011; Signoretto et al., 2010; Tomioka et al., 2011; Signoretto et al., 2013). It is not too surprising, then, that (3.1) provably recovers the underlying tensor $\boldsymbol{\mathcal{X}}_0$, when the number of measurements m is sufficiently large. The following is a (simplified) corollary of results of Tomioka et. al. (2011)

Corollary 2 (of (Tomioka et al., 2011), Theorem 3)**.** *Suppose that* $\boldsymbol{\mathcal{X}}_0$ *has Tucker rank* $(r, \ldots, r)$*, and* $m \geq Crn^{K-1}$*, where* C *is a constant. Then with high probability,* $\boldsymbol{\mathcal{X}}_0$ *is the optimal solution to* (3.1)*, with each* $\lambda_i = 1$*.*

This result shows that there *is* a range in which (3.1) 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 (3.1):

Theorem 3. *Let* $\boldsymbol{\mathcal{X}}_0 \in \mathfrak{T}_r$ *be nonzero. Set* $\kappa = \min_i \left\{ \left\|(\boldsymbol{\mathcal{X}}_0)_{(i)}\right\|_*^2 / \left\|\boldsymbol{\mathcal{X}}_0\right\|_F^2 \right\} \times n^{K-1}$*. Then if the number of measurements* $m \leq \kappa - 2$*,* $\boldsymbol{\mathcal{X}}_0$ *is not the unique solution to* (3.1)*, 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 (as well as some other results of (Tomioka et al., 2011)) is essentially tight. Unfortunately, it has negative implications for the efficacy of the SNN model in (3.1): 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 (3.1). Theorem 3 is a direct consequence of a much more general principle underlying multi-structured recovery, which is elaborated next. After that, in Section 4, we show that for low-rank tensor recovery, better convexifying schemes are available.

General lower bound for multiple structures

The poor behavior of (3.1) is an instance of a much more general phenomenon, first discovered by Oymak et. al. (2012). 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 could also be of interest. For sparse phase retrieval problems in signal processing (Oymak et al., 2012), the task can be rephrased to infer a block sparse matrix, which implies both sparse and low-rank structures. In robust metric learning (Lim et al., 2013), the goal is to estimate a matrix that is column sparse and low rank concurrently. In computer vision, many signals of interest are both low-rank and sparse in an appropriate basis (Liang et al., 2012). To recover such simultaneously structured objects, it is tempting to build a convex relaxation by

[2] Equivalently, it means $\boldsymbol{\mathcal{X}}_0$ is the unique optimal solution to:

$$\min_{\boldsymbol{\mathcal{X}}} \ \max_i \left\{ \frac{\mathrm{rank}(\boldsymbol{\mathcal{X}}_{(i)})}{\mathrm{rank}((\boldsymbol{\mathcal{X}}_0)_{(i)})} \right\} \quad \text{s.t.} \ \ \mathcal{G}[\boldsymbol{\mathcal{X}}] = \mathcal{G}[\boldsymbol{\mathcal{X}}_0].$$

combining the convex relaxations for each of the individual structures. In the tensor case, this yields (3.1). Surprisingly, this combination is often not significantly more powerful than the best single regularizer (Oymak et al., 2012). We obtain Theorem 3 as a consquence of a new, general result of this nature, using a geometric framework introduced in (Amelunxen et al., 2013). Compared to (Oymak et al., 2012), this approach has a clearer geometric intuition, covers a more general class of regularizers[3] 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.)[4]. Let $\|\cdot\|_{(i)}$ be the penalty norms corresponding to the i-th structure (e.g. ℓ_1, nuclear norm). Consider the composite norm optimization

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; f(\boldsymbol{x}) := \lambda_1 \|\boldsymbol{x}\|_{(1)} + \lambda_2 \|\boldsymbol{x}\|_{(2)} + \cdots + \lambda_K \|\boldsymbol{x}\|_{(K)}$$
$$\text{s.t. } \mathcal{G}[\boldsymbol{x}] = \mathcal{G}[\boldsymbol{x}_0], \tag{3.2}$$

where $\mathcal{G}[\cdot]$ is a Gaussian measurement operator, and $\boldsymbol{\lambda} > \mathbf{0}$. Is $\boldsymbol{x}_0$ the unique optimal solution to (3.2)? Recall that the descent cone of a function f at a point $\boldsymbol{x}_0$ is defined as

$$\mathcal{C}(f, \boldsymbol{x}_0) = \operatorname{cone}\{\boldsymbol{v} \mid f(\boldsymbol{x}_0 + \boldsymbol{v}) \le f(\boldsymbol{x}_0)\}, \tag{3.3}$$

which, in short, will be denoted as $\mathcal{C}$. Then $\boldsymbol{x}_0$ is the unique optimal solution if and only if $\operatorname{null}(\mathcal{G}) \cap \mathcal{C} = \{\mathbf{0}\}$. Thus, recovery fails if $\operatorname{null}(\mathcal{G})$ has nontrivial intersection with $\mathcal{C}$. Since $\mathcal{G}$ is a Gaussian operator, $\operatorname{null}(\mathcal{G})$ 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 $\mathcal{C}^o = \operatorname{cone}\big(\partial f(\boldsymbol{x}_0)\big)$. Because polarity reverses inclusion, $\mathcal{C}$ will be *large* whenever $\mathcal{C}^o$ is *small*.

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

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

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

$$\frac{\langle \boldsymbol{v}, \boldsymbol{x}_0\rangle}{\|\boldsymbol{v}\|_2\|\boldsymbol{x}_0\|_2} \ge \frac{\|\boldsymbol{x}_0\|_\diamond}{L\|\boldsymbol{v}\|_\diamond^*\|\boldsymbol{x}_0\|_2} \ge \frac{\|\boldsymbol{x}_0\|_\diamond}{L\|\boldsymbol{x}_0\|_2}. \tag{3.4}$$

A more geometric way of summarizing the above fact is as follows: for $\boldsymbol{x} \neq \mathbf{0}$, let us denote the *circular cone* with axis $\boldsymbol{x}$ and angle θ as

$$\operatorname{circ}(\boldsymbol{x}, \theta) := \{\boldsymbol{z} \mid \angle(\boldsymbol{z}, \boldsymbol{x}) \le \theta\}. \tag{3.5}$$

[3] (Oymak et al., 2012) studies decomposable norms, with some additional assumptions. Our result holds for arbitrary norms.

[4] $\boldsymbol{x}_0 \in \mathbb{R}^n$ is the underlying signal of interest, perhaps after vectorization. For example, suppose we want to recover the matrix $\mathbf{X}_0 \in \mathbb{R}^{n_1\times n_2}$. Then $\boldsymbol{x}_0 = \operatorname{vec}(\mathbf{X}_0) \in \mathbb{R}^{n_1 n_2}$.

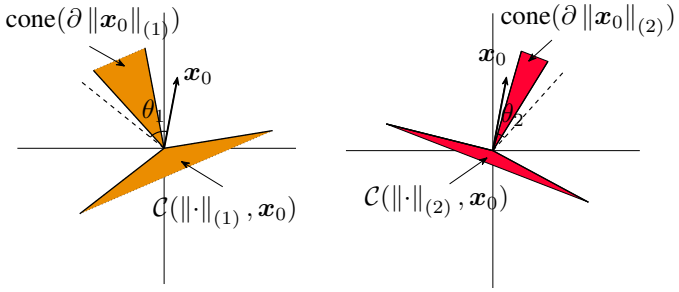

Figure 1. **Cones and their polars for convex regularizers $\|\cdot\|_{(1)}$ and $\|\cdot\|_{(2)}$ respectively.** Suppose that $\boldsymbol{x}_0$ has two sparse structures simultaneously. Consider convex regularizer $f(\boldsymbol{x}) = \|\boldsymbol{x}_0\|_{(1)} + \|\boldsymbol{x}_0\|_{(2)}$. Suppose, as depicted, $\theta_1 \ge \theta_2$. Then both $\operatorname{cone}(\partial\|\boldsymbol{x}_0\|_{(1)})$ and $\operatorname{cone}(\partial\|\boldsymbol{x}_0\|_{(2)})$ are in the circular cone $\operatorname{circ}(\boldsymbol{x}_0, \theta_1)$. Thus we have: $\operatorname{cone}\big(\partial f(\boldsymbol{x}_0)\big) = \operatorname{cone}(\partial\|\boldsymbol{x}_0\|_{(1)} + \partial\|\boldsymbol{x}_0\|_{(2)}) \subseteq \operatorname{conv}\big\{\operatorname{circ}(\boldsymbol{x}_0, \theta_1), \operatorname{circ}(\boldsymbol{x}_0, \theta_2)\big\} = \operatorname{circ}(\boldsymbol{x}_0, \theta_1)$.

Then if $\boldsymbol{x}_0 \neq \mathbf{0}$, and $\theta = \cos^{-1}(\|\boldsymbol{x}_0\|_\diamond / L\|\boldsymbol{x}_0\|_2)$,

$$\partial\|\cdot\|_\diamond(\boldsymbol{x}_0) \subseteq \operatorname{circ}(\boldsymbol{x}_0, \theta). \tag{3.6}$$

Table 1 describes the angle parameters θ for various popular 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 vector 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 circular cones.

For $f = \sum_i \lambda_i \|\cdot\|_{(i)}$, 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 $\boldsymbol{x}_0$, $\operatorname{cone}(\partial f(\boldsymbol{x}_0))$ is also contained in a circular cone:

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

$$\operatorname{cone}(\partial f(\boldsymbol{x}_0)) \subseteq \operatorname{circ}\left(\boldsymbol{x}_0, \max_{i=1\ldots K} \theta_i\right). \tag{3.7}$$

So, the subdifferential of our combined regularizer f is contained in a circular cone whose angle is given by the largest of $\{\theta_i\}_{i=1}^K$. Figure 1 visualizes this geometry.

How does this behavior affect the recoverability of $\boldsymbol{x}_0$ via (3.2)? 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. (2013). 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:

$$\delta(\mathcal{C}) \doteq \mathbb{E}_{\mathbf{g}\sim\text{i.i.d.}\mathcal{N}(0,1)}\left[\|\mathcal{P}_{\mathcal{C}}(\mathbf{g})\|_2^2\right]. \tag{3.8}$$

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.

Object	Complexity Measure	Relaxation	$\cos^2\theta$	$\kappa = n\cos^2\theta$
Sparse $\boldsymbol{x} \in \mathbb{R}^n$	$k = \|\boldsymbol{x}\|_0$	$\|\boldsymbol{x}\|_1$	$[\frac{1}{n}, \frac{k}{n}]$	$[1, k]$
Column-sparse $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$	$c = \#\{j \mid \mathbf{X}\mathbf{e}_j \neq \mathbf{0}\}$	$\sum_j \|\mathbf{X}\mathbf{e}_j\|_2$	$[\frac{1}{n_2}, \frac{c}{n_2}]$	$[n_1, cn_1]$
Low-rank $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ $(n_1 \geq n_2)$	$r = \operatorname{rank}(\mathbf{X})$	$\|\mathbf{X}\|_*$	$[\frac{1}{n_2}, \frac{r}{n_2}]$	$[n_1, rn_1]$

Using tools from spherical integral geometry, (Amelunxen et al., 2013) shows that for linear inverse problems with Gaussian measurements, a sharp phase transition in recoverability occurs around $m = \delta(\mathcal{C})$. Since we attempt to derive a necessary condition for the success of (3.2), we need only one side of their result, with slight modifications:

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

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

To apply this result to our problem, we need to have a lower bound on 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 1, we calculate

$$\begin{aligned}\delta(\mathcal{C}) \;&=\; n - \delta(\mathcal{C}^o) \;=\; n - \delta\left(\operatorname{cone}\left(\partial f(\boldsymbol{x}_0)\right)\right) \\ &\geq\; n - \delta(\operatorname{circ}(\boldsymbol{x}_0, \max_i \theta_i)).\end{aligned} \tag{3.9}$$

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 (Amelunxen et al., 2013; McCoy, 2013).

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

By combining (3.9) and Lemma 2, we have $\delta(C) \geq n\min_i \cos^2\theta_i - 2$. Using Corollary 4, we finally obtain:

Theorem 5. *Let $\boldsymbol{x}_0 \neq \mathbf{0}$. Suppose 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 $m \leq \kappa - 2$,

$$\begin{aligned}&\mathbb{P}\left[\,\boldsymbol{x}_0 \text{ is the unique optimal solution to (3.2)}\,\right] \\ &\leq 4\exp\left(-\frac{(\kappa - m - 2)^2}{16\,(\kappa - 2)}\right).\end{aligned} \tag{3.10}$$

Thus, for reliable recovery, the number of measurements needs to be at least proportional to κ.[5] 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 recently discovered by Oymak et. al. (2012): 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. The subdifferential of a norm at $\boldsymbol{x}_0$ is contained in a relatively small circular cone with central axis $\boldsymbol{x}_0$.

Theorem 3 can then be easily deduced by specializing 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, $\|\boldsymbol{\mathcal{X}}\|_{(i)} \doteq \left\|\boldsymbol{\mathcal{X}}_{(i)}\right\|_*$ is $L = \sqrt{n}$-Lipschitz. Hence,

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

The term $\min_i \left\{\left\|\boldsymbol{\mathcal{X}}_{(i)}\right\|_*^2 / \|\boldsymbol{\mathcal{X}}\|_F^2\right\}$ lies between 1 and r, inclusively. For example, if $\boldsymbol{\mathcal{X}}$ has Tucker decomposition[6] $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{C}} \times_1 \mathbf{U}_1 \times_2 \cdots \times_K \mathbf{U}_K$, with $\mathbf{U}_i^T\mathbf{U}_i = \mathbf{I}$ and $\boldsymbol{\mathcal{C}}$ supersymmetric ($\boldsymbol{\mathcal{C}}_{i_1 \ldots i_K} = \mathbf{1}_{\{i_1 = i_2 = \cdots = i_K\}}$), then that term is equal to r.

4. A Better Convexification: Square Deal

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 $m \geq Cr(n_1 + n_2)$ (Chandrasekaran et al., 2012), 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 recovers $\boldsymbol{\mathcal{X}}_0$ w.h.p. when $m \geq Crn^{K-1}$.

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

[6] The *mode-i (matrix) product* $\boldsymbol{\mathcal{A}} \times_i \mathbf{B}$ of size-compatible tensor $\boldsymbol{\mathcal{A}}$ and matrix $\mathbf{B}$ outputs the tensor $\boldsymbol{\mathcal{C}}$ such that $\boldsymbol{\mathcal{C}}_{(i)} = \mathbf{B}\boldsymbol{\mathcal{A}}_{(i)}$.

This suggests a more mundane explanation for the difficulty with (3.1): 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 remedy 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 $\mathbf{A} \in \mathbb{R}^{m_1 \times n_1}$, and integers m_2 and n_2 satisfying $m_1 n_1 = m_2 n_2$, the reshaping operator $\text{reshape}(\mathbf{A}, m_2, n_2)$ returns an $m_2 \times n_2$ matrix whose elements are taken columnwise from $\mathbf{A}$. This operator rearranges elements in $\mathbf{A}$ 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[7]

$$\boldsymbol{\mathcal{X}}_{[j]} = \text{reshape}\Big(\boldsymbol{\mathcal{X}}_{(1)},\ \prod_{i=1}^{j} n_i,\ \prod_{i=j+1}^{K} n_i\Big). \tag{4.2}$$

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, $\boldsymbol{\mathcal{X}}_{[j]}$ could become 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.

Lemma 3. *(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*

$$\boldsymbol{\mathcal{X}}_{[j]} = \sum_{i=1}^{r} \lambda_i (\boldsymbol{a}_i^{(j)} \otimes \cdots \otimes \boldsymbol{a}_i^{(1)}) \circ (\boldsymbol{a}_i^{(K)} \otimes \cdots \otimes \boldsymbol{a}_i^{(j+1)}). \tag{4.3}$$

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

$$\boldsymbol{\mathcal{X}}_{[j]} = (\mathbf{U}_j \otimes \cdots \otimes \mathbf{U}_1)\, \boldsymbol{\mathcal{C}}_{[j]}\, (\mathbf{U}_K \otimes \cdots \otimes \mathbf{U}_{j+1})^*. \tag{4.4}$$

Using Lemma 3 and the fact that $\text{rank}(\mathbf{A} \otimes \mathbf{B}) = \text{rank}(\mathbf{A})\,\text{rank}(\mathbf{B})$, we obtain:

Lemma 4. *Let $\text{rank}_{\text{tc}}(\boldsymbol{\mathcal{X}}) = (r_1, r_2, \cdots, r_K)$, and $\text{rank}_{\text{cp}}(\boldsymbol{\mathcal{X}}) = r_{\text{cp}}$. Then $\text{rank}(\boldsymbol{\mathcal{X}}_{[j]}) \le r_{\text{cp}}$, and $\text{rank}(\boldsymbol{\mathcal{X}}_{[j]}) \le \min\Big\{ \prod_{i=1}^{j} r_i,\ \prod_{i=j+1}^{K} r_i \Big\}$.*

[7] One can also think of (4.2) as embedding the tensor $\boldsymbol{\mathcal{X}}$ into the matrix $\boldsymbol{\mathcal{X}}_{[j]}$ as follows: $\boldsymbol{\mathcal{X}}_{i_1,i_2,\cdots,i_K} = \big(\boldsymbol{\mathcal{X}}_{[j]}\big)_{a,b}$, where

$$a = 1 + \sum_{m=1}^{j} \left((i_m - 1) \prod_{l=1}^{m-1} n_l \right)$$
$$b = 1 + \sum_{m=j+1}^{K} \left((i_m - 1) \prod_{l=j+1}^{m-1} n_l \right).$$

Thus, $\boldsymbol{\mathcal{X}}_{[j]}$ is not only more balanced but also maintains the low-rank property of the tensor $\boldsymbol{\mathcal{X}}$, which motivates us to recover $\boldsymbol{\mathcal{X}}_0$ by solving

$$\text{minimize} \quad \big\|\boldsymbol{\mathcal{X}}_{[j]}\big\|_* \quad \text{subject to} \quad \mathcal{G}[\boldsymbol{\mathcal{X}}] = \mathcal{G}[\boldsymbol{\mathcal{X}}_0]. \tag{4.5}$$

Using Lemma 4 and (Chandrasekaran et al., 2012), we can prove that this relaxation exactly recovers $\boldsymbol{\mathcal{X}}_0$, when the number of measurements is sufficiently large:

Theorem 6. *Consider a K-way tensor with the same length (say n) along each mode. (1) If $\boldsymbol{\mathcal{X}}_0$ has CP rank r, using (4.5) with $j = \lceil \frac{K}{2} \rceil$, $m \ge Crn^{\lceil \frac{K}{2} \rceil}$ 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 (4.5) with $j = \lceil \frac{K}{2} \rceil$, $m \ge Cr^{\lfloor \frac{K}{2} \rfloor} n^{\lceil \frac{K}{2} \rceil}$ is sufficient to recover $\boldsymbol{\mathcal{X}}_0$ with high probability.*

The number of measurements $O(r^{\lfloor \frac{K}{2} \rfloor} n^{\lceil \frac{K}{2} \rceil})$ required to recover $\boldsymbol{\mathcal{X}}$ with square reshaping (4.5), is always within a constant of the number $O(rn^{K-1})$ with the sum-of-nuclear-norms model, and is significantly smaller when r is small and $K \ge 4$. E.g., we obtain an improvement of a multiplicative factor of $n^{\lfloor K/2 \rfloor - 1}$ when r is a constant. This is a significant improvement.

Our square reshaping can be generalized to group together any j modes (say modes $i_1, i_2, \ldots, i_j$) rather than the first j modes. Denote $\mathcal{I} = \{i_1, i_2, \ldots, i_j\} \subseteq [K]$ and $\mathcal{J} = [K] \backslash \mathcal{I} = \{i_{j+1}, i_{j+2}, \ldots, i_K\}$. Then the embedded matrix $\boldsymbol{\mathcal{X}}_{\mathcal{I}} \in \mathbb{R}^{\prod_{k=1}^{j} n_{i_k} \times \prod_{k=j+1}^{K} n_{i_k}}$ can be defined similarly as in (4.2) but with a relabeling preprocessing. In specific, for $1 \le k \le K$, we relabel the k-th mode as the original i_k-th mode. Regarding this relabeled tensor $\widehat{\boldsymbol{\mathcal{X}}}$, we can define

$$\boldsymbol{\mathcal{X}}_{\mathcal{I}} := \widehat{\boldsymbol{\mathcal{X}}}_{[j]} = \text{reshape}\Big(\widehat{\boldsymbol{\mathcal{X}}}_{(1)},\ \prod_{k=1}^{j} n_{i_k},\ \prod_{k=j+1}^{K} n_{i_k}\Big).$$

Lemma 4 and Theorem 6 can then also be easily extended.

Note that for tensors with different lengths or ranks, the comparison between SNN and our square reshaping becomes more subtle. It is possible to construct examples for which the square reshaping model does not have an advantage over the SNN model, even for $K > 3$. Nevertheless, for a large class of tensors, our square reshaping is capable of reducing the number of generic measurements required by SNN model, both in theory and in numerical experiments.

5. Numerical Experiments

We corroborate the improvement of square reshaping with numerical experiments on *low-rank tensor completion (LRTC)* for both noise-free (synthetic data) and noisy (real data) cases. LRTC attempts to reconstruct the (approximately) low-rank tensor $\boldsymbol{\mathcal{X}}_0$ from a subset Ω of its entries. By imposing appropriate incoherence conditions, it is possible to prove exact/stable recovery guarantees for both our

square deal formulation (Gross, 2011) and the SNN model (Huang et al., 2014) for LRTC. However, unlike the recovery problem under Gaussian random measurements, due to the lack of sharp bounds, we have no proof that the square model is guaranteed to outperform the SNN model here. Nonetheless, numerical results below clearly indicate the advantage of our square approach, complementing our theoretical results established in previous sections.

5.1. Simulation

We generate a four-way tensor $\boldsymbol{\mathcal{X}}_0 \in \mathbb{R}^{n\times n\times n\times n}$ as $\boldsymbol{\mathcal{X}}_0 = c_0 \times_1 \boldsymbol{u}_1 \times_2 \boldsymbol{u}_2 \times_3 \boldsymbol{u}_3 \times_4 \boldsymbol{u}_4$, where $c_0 \sim \mathcal{N}(0,1)$, and $\boldsymbol{u}_i$'s are generated uniformly over the unit sphere $\mathcal{S}^{n-1} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\|_2 = 1\}$. The observed entries are chosen uniformly with ratio ρ. Since the unfolding matrix of $\boldsymbol{\mathcal{X}}_0$ along each mode has the same distribution, we set each $\lambda_i = 1$. Therefore, we compare the recovery performances between

$$\min_{\boldsymbol{\mathcal{X}}} \sum_{i=1}^{K} \|\boldsymbol{\mathcal{X}}_{(i)}\|_* \quad \text{s.t.} \quad \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0], \tag{5.1}$$

$$\min_{\boldsymbol{\mathcal{X}}} \left\|\boldsymbol{\mathcal{X}}_{\{1,2\}}\right\|_* \quad \text{s.t.} \quad \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0]. \tag{5.2}$$

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 10 test instances and declare a trial to be successful if the recovered $\boldsymbol{\mathcal{X}}^\star$ satisfies $\|\boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{X}}_0\|_F / \|\boldsymbol{\mathcal{X}}_0\|_F \le 10^{-2}$.

The optimization problems are solved using efficient first-order methods. Since (5.2) is equivalent to standard matrix completion, we use the existing solver ALM (Lin et al., 2010). For the sum of nuclear norms minimization (5.1), we implement the accelerated linearized Bregman algorithm (Huang et al., 2013) (see appendix for details).

Figure 2 plots the fraction of correct recovery for each pair. Clearly, the square approach succeeds in a much larger region.

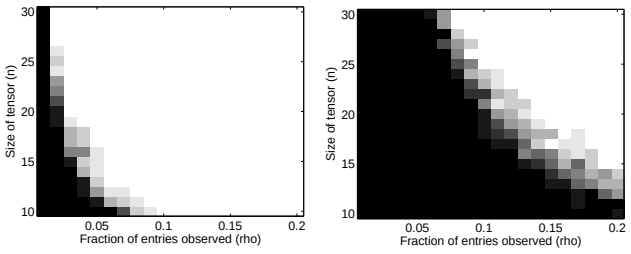

Figure 2. **Tensor completion.** The colormap indicates the fraction of instances that are correctly recovered for each (ρ, n)-pair, which increases with brightness from 100% failure (black) to 100% success (white).

5.2. Video Completion

Color videos can be naturally represented as four-mode tensors (length×width×channels×frames). We compare the performances of our square model and the SNN model on video completion from randomly missing pixels. We consider three video datasets: Ocean video ($112 \times 160 \times 3 \times 32$), Campus video ($128 \times 160 \times 3 \times 199$) and Face video ($96 \times 65 \times 3 \times 994$).[8]

For our square model, we set $\mathcal{I} = \{1, 4\}$ for the Ocean video, and set $\mathcal{I} = \{1, 2\}$ for the Campus and the Face videos, to construct more balanced embedded matrices $\boldsymbol{\mathcal{X}}_\mathcal{I}$. Due to the existence of noise in real data, we would solve the regularized least square problems

$$\min_{\boldsymbol{\mathcal{X}}} \quad \frac{1}{2}\|\mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] - \boldsymbol{\mathcal{D}}\|_F^2 + \sum_{i=1}^{4} \lambda_i \left\|\boldsymbol{\mathcal{X}}_{(i)}\right\|_* \tag{5.3}$$

$$\min_{\boldsymbol{\mathcal{X}}} \quad \frac{1}{2}\|\mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] - \boldsymbol{\mathcal{D}}\|_F^2 + \lambda \|\boldsymbol{\mathcal{X}}_\mathcal{I}\|_*, \tag{5.4}$$

where $\boldsymbol{\mathcal{D}} = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0]$ is the observed tensor, $\lambda_i \ge 0$ and $\lambda \ge 0$ are tuning parameters. Since the purpose of our experiment is to compare SNN and square models, to make the comparison fair and meaningful, we should tune those parameters as optimally as possible. That is not an easy task, especially for the SNN model (5.3), which involves four tuning parameters. As a remedy, we solve equivalent nuclear norm constrained programs,[9]

$$\min_{\boldsymbol{\mathcal{X}}} \quad \tfrac{1}{2}\|\mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] - \boldsymbol{\mathcal{D}}\|_F^2 \quad \text{s.t.} \left\|\boldsymbol{\mathcal{X}}_{(i)}\right\|_* \le \beta_i, \forall i \in [4], \tag{5.5}$$

$$\min_{\boldsymbol{\mathcal{X}}} \quad \tfrac{1}{2}\|\mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] - \boldsymbol{\mathcal{D}}\|_F^2 \quad \text{s.t.} \|\boldsymbol{\mathcal{X}}_\mathcal{I}\|_* \le \beta, \tag{5.6}$$

where we set β_i and β to their oracle values, i.e. $\beta_i = \left\|(\boldsymbol{\mathcal{X}}_0)_{(i)}\right\|_*$ and $\beta = \|(\boldsymbol{\mathcal{X}}_0)_\mathcal{I}\|_*$, which can be reasonably considered as (nearly) optimal settings. Therefore, our comparisons here are quite fair to both models.

To solve (5.5) and (5.6), we exploit the Frank-Wolfe algorithm, interests in which have recently resurged (Jaggi, 2013), due to its good scalability for dealing with nuclear norm constrained problems (see appendix for details). Figures 3 and 4 display the results obtained using (5.5) and (5.6). Clearly, our square approach outperforms the SNN model.

We expect the benefits of using our square formulation will be magnified in multi-spectral video data, where the number of channels could be much larger than three. In such data, the video tensor tends to be low rank in both the wavelength and the temporal modes. Thus we can group these two modes to form our low-rank matrix $\boldsymbol{\mathcal{X}}_\mathcal{I}$. When the

[8] A detailed description of these data is included in the appendix.

[9] The equivalence between the penalized problem and the norm constrained problem is shown in the appendix.

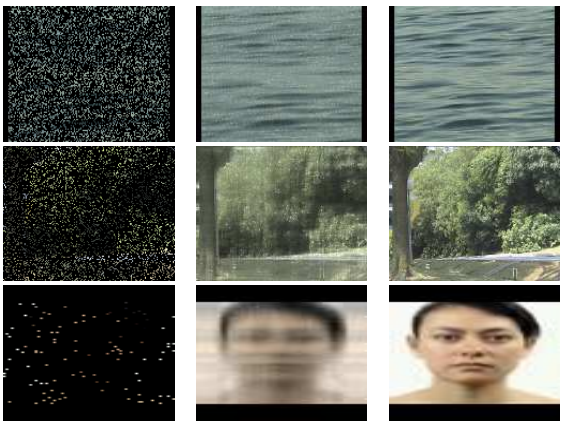

Figure 3. Sample snapshots from our datasets: Ocean, Campus, Face. **Left**: sampled video (20 % for the Ocean video, 20% for the Campus video and 2.5% for the Face video). **Middle**: video recovered by SNN model. **Right**: video recovered by square model.

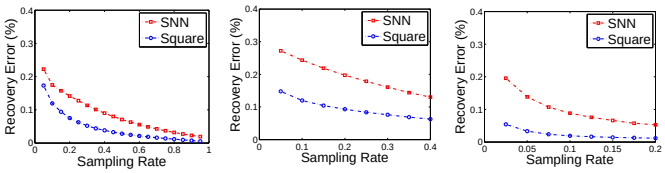

Figure 4. **Video Completion.** (relative) recovery error vs. sampling rate, for the three videos in Figure 3.

technique of taking multi-spectral data becomes mature in the future, we believe our square reshaping model will be more useful and more significant for the completion task.

6. Discussions

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 $(2r)^K+2nrK+1$ measurements are sufficient to recover any $\boldsymbol{\mathcal{X}}_0 \in \mathfrak{T}_r$ almost surely. For the conventional convex surrogate sum-of-nuclear-norms (SNN) model (3.1), 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 of multi-structured object recovery in a very general setting, which can be applied to many other scenarios (e.g. signal processing, metric learning, computer vision). 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 reshaping model (4.5). We then 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 numerical experiments for tensor completion still suggest the square reshaping model outperforms the SNN model generally. Compared with $\Omega(rn^{K-1})$ measurements required by the SNN model, the sample complexity, $O(r^{\lfloor \frac{K}{2} \rfloor} n^{\lceil \frac{K}{2} \rceil})$, required by the square reshaping (4.5), is always within a constant of it, and is 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 the square model is still suboptimal. It remains an open and intriguing problem to obtain near-optimal tractable convex relaxations for all $K > 2$.

Very recently, other interesting models and algorithms have been proposed for low-rank tensor recovery. In (Romera-Paredes & Pontil, 2013), the nuclear norm in the SNN model is replaced by a new convex regularizer along each mode. Since this new convex penalty function is not a norm, Theorem 3 does not imply its suboptimality. In (Romera-Paredes et al., 2013; Xu et al., 2013), efficient methods based on alternating minimization are designed to solve related non-convex models. Empirical improvements over the SNN model are shown in all the above convex and non-convex approaches. However, to the best of our knowledge, no theoretical guarantees have been obtained yet. Further analyzing these methods is an interesting problem for future research.

Putting our work in a broader setting, 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 (Oymak et al., 2012)). The resulting sample requirements tend to be much larger than the intrinsic degrees of freedom of the low-dimensional manifold in which the structured signal lies. Our square model for low-rank tensor recovery demonstrates the possibility that a better exploitation of those structures can significantly reduce this sample complexity (see also (Richard et al., 2013) for ideas in this direction). However, there are still no clear clues on how to intelligently utilize several simultaneous structures generally, and moreover how to design tractable methods to recover multi-structured objects with near minimal numbers of measurements. These problems are definitely worth future study.

Acknowledgements

We thank the reviewers for helpful comments. CM was supported by the Class of 1988 Doctoral Fellowship and NSF Grant DMS-1016571. BH and DG were supported by NSF Grant DMS-1016571. JW was supported by Columbia University startup funding and Office of Naval Research award N00014-13-1-0492.

References

Amelunxen, D., Lotz, M., McCoy, M., and Tropp, J. Living on the edge: A geometric theory of phase transitions in convex

optimization. *arXiv preprint arXiv:1303.6672*, 2013.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S., and Telgarsky, M. Tensor decompositions for learning latent variable models. *arXiv preprint arXiv:1210.7559*, 2012.

Bayati, Mohsen, Lelarge, Marc, and Montanari, Andrea. Universality in polytope phase transitions and message passing algorithms. *arXiv preprint arXiv:1207.7321*, 2012.

Candès, E., Romberg, J., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8), 2006.

Chandrasekaran, V., Recht, B., Parrilo, P., and Willsky, A. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

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

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

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

Fazel, M. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.

Gandy, S., Recht, B., and Yamada, I. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

Goldfarb, D. and Qin, Z. Robust low-rank tensor recovery: Models and algorithms. *SIAM Journal on Matrix Analysis and Applications*, 35(1):225–253, 2014.

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

Hillar, C. and Lim, L. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.

Huang, B., Ma, S., and Goldfarb, D. Accelerated linearized bregman method. *Journal of Scientific Computing*, 54(2-3):428–453, 2013.

Huang, B., Mu, C., Goldfarb, D., and Wright, J. Provable low-rank tensor recovery. *preprint*, 2014.

Jaggi, M. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML*, 2013.

Kolda, T. and Bader, B. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

Kreimer, N., Stanton A. and Sacchi, M. D. Nuclear norm minimization and tensor completion in exploration seismology. In *ICASSP*, 2013.

Li, N. and Li, B. Tensor completion for on-board compression of hyperspectral images. In *ICIP*, 2010.

Li, Y., Yan, J., Zhou, Y., and Yang, J. Optimum subspace learning and error correction for tensors. In *ECCV*, 2010.

Liang, X., Ren, X., Zhang, Z., and Ma, Y. Repairing sparse low-rank texture. In *ECCV 2012*, 2012.

Lim, D., McFee, B., and Lanckriet, G. Robust structural metric learning. In *ICML*, 2013.

Lin, Z., Chen, M., and Ma, Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

Liu, J., Musialski, P., Wonka, P., and Ye, J. Tensor completion for estimating missing values in visual data. In *ICCV*, 2009.

McCoy, M. *A geometric analysis of convex demixing*. PhD thesis, California Institute of Technology, 2013.

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

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

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

Oymak, S., Jalali, A., Fazel, M., Eldar, Y., and Hassibi, B. Simultaneously structured models with application to sparse and low-rank matrices. *arXiv preprint arXiv:1212.3753*, 2012.

Recht, B., Fazel, M., and Parrilo, P. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Richard, E., Bach, F., and Vert, J. Intersecting singularities for multi-structured estimation. In *ICML*, 2013.

Romera-Paredes, B. and Pontil, M. A new convex relaxation for tensor completion. In *NIPS*, 2013.

Romera-Paredes, B., Aung, H., Bianchi-Berthouze, N., and Pontil, M. Multilinear multitask learning. In *ICML*, 2013.

Semerci, O., Hao, N., Kilmer, M., and Miller, E. Tensor-based formulation and nuclear norm regularization for multi-energy computed tomography. *arXiv preprint arXiv:1307.5348*, 2013.

Signoretto, M., De Lathauwer, L., and Suykens, J. Nuclear norms for tensors and their use for convex multilinear estimation. *Submitted to Linear Algebra and Its Applications*, 43, 2010.

Signoretto, M., Tran Dinh, Q., Lathauwer, L., and Suykens, J. Learning with tensors: a framework based on convex optimization and spectral regularization. *Machine Learning*, pp. 1–49, 2013.

Tomioka, R., Suzuki, T., Hayashi, K., and Kashima, H. Statistical performance of convex tensor decomposition. In *NIPS*, 2011.

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

Xu, Y., Hao, R., Yin, W., and Su, Z. Parallel matrix factorization for low-rank tensor completion. *arXiv preprint arXiv:1312.1254*, 2013.