Square Deal: Lower Bounds and Improved Convex 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,*
[2]*Department of Electrical Engineering,*
Columbia University, New York, NY 10027, USA

Editor:

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.

Keywords: Tensor Recovery, Convex Relaxation, Low Rank, Lower Bound, Sample Complexity

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 and 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 la-

tent variable graphical models (Anandkumar et al., 2014), classifying audio (Mesgarani et al., 2006), mining text (Cohen and Collins, 2012), processing radar signals (Nion and Sidiropoulos, 2010), multilinear multitask learning (Romera-Paredes et al., 2013) , to name a few.

We consider the problem of recovering a K-way tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_K}$ from linear measurements $z = \mathcal{G}[\mathcal{X}] \in \mathbb{R}^m$. Typically, $m \ll N = \prod_{i=1}^{K} n_i$, and so the problem of recovering $\mathcal{X}$ from 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., 2014). 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 and Lim, 2013). For example, even computing a tensor's (CP) rank,

$$\text{rank}_{\text{cp}}(\mathcal{X}) := \min\left\{ r \mid \mathcal{X} = \sum_{i=1}^{r} a_1^{(i)} \circ \cdots \circ a_K^{(i)} \right\}, \tag{1}$$

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 recover tensors of small *Tucker rank* (Tucker, 1966). 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 $\mathcal{X}$:

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

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 (2) suggests a 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 $\mathcal{G}[\mathcal{X}] = z$. 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 and Sacchi, 2013; Li and 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 $\text{rank}_{\text{tc}}(\mathcal{X}) \leq (r, r, \ldots, r)$. Let Υ_r denote the set

of all such tensors,[1] namely

$$\Upsilon_r := \{\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n \times n \times \cdots \times n} \mid \text{rank}_{\text{tc}}(\boldsymbol{\mathcal{X}}) \preceq (r, r, \ldots, r)\}. \tag{3}$$

We will consider the problem of estimating an element $\boldsymbol{\mathcal{X}}$ of Υ_r from Gaussian measurements $\mathcal{G}$ (i.e., $z_i = \langle \boldsymbol{\mathcal{G}}_i, \boldsymbol{\mathcal{X}} \rangle$, where $\boldsymbol{\mathcal{G}}_i$ has i.i.d. standard normal entries). To describe a generic tensor in Υ_r, we need at most $r^K + rnK$ parameters. In Section 2, we show that a certain nonconvex strategy can recover all $\boldsymbol{\mathcal{X}} \in \Upsilon_r$ exactly when $m > (2r)^K + 2nrK$. In contrast, the best known theoretical guarantee for SNN minimization, due to Tomioka et al. (2011), shows that $\boldsymbol{\mathcal{X}} \in \Upsilon_r$ 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.

Our theoretical results pertain to Gaussian operators $\mathcal{G}$. The motivation for studying Gaussian measurements is threefold. 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). Moreover, 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. Furthermore, the results with respect to Gaussian measurements have direct implications to the minimax risk for denoising (Oymak and Hassibi, 2013; Amelunxen et al., 2014). In Section 4, 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 and 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., 2014). 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.

---

1. To keep the presentation in this paper compact, we state most of our results regarding tensors in Υ_r, although it is not difficult to modify them for general tensors.

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 $\mathcal{G}[\mathcal{X}_0] \in \mathbb{R}^m$. We would like to attempt to recover $\mathcal{X}_0$ by minimizing some combination of the ranks of the unfoldings, over all tensors $\mathcal{X}$ that are consistent with our observations. This suggests a *vector optimization* problem (Boyd and Vandenberghe, 2004, Chap. 4.7):

$$\text{minimize}_{(\text{w.r.t. } \mathbb{R}_+^K)} \quad \text{rank}_{\text{tc}}(\mathcal{X}) \quad \text{subject to} \quad \mathcal{G}[\mathcal{X}] = \mathcal{G}[\mathcal{X}_0]. \tag{4}$$

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 (4) recovers $\mathcal{X}_0$ if there does not exist any other tensor $\mathcal{X}$ that is consistent with the observations and has no larger rank along each mode:

Definition 1 *We call $\mathcal{X}_0$ recoverable by (4) if the set*

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

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

$$\text{minimize}_{\mathcal{X}} \quad \max_i \left\{ \frac{\text{rank}(\mathcal{X}_{(i)})}{\text{rank}(\mathcal{X}_{0(i)})} \right\} \quad \text{subject to} \quad \mathcal{G}[\mathcal{X}] = \mathcal{G}[\mathcal{X}_0]. \tag{5}$$

The problems (4)-(5) are not tractable. However, they do serve as a baseline for understanding how many generic measurements are required to recover $\mathcal{X}_0$ from an information theoretic perspective.

The recovery performance of program (4) depends heavily on the properties of $\mathcal{G}$. Suppose (4) fails to recover $\mathcal{X}_0 \in \Upsilon_r$. Then there exists another $\mathcal{X}' \in \Upsilon_r$ such that $\mathcal{G}[\mathcal{X}'] = \mathcal{G}[\mathcal{X}_0]$. So, to guarantee that (4) recovers *any* $\mathcal{X}_0 \in \Upsilon_r$, a necessary and sufficient condition is that $\mathcal{G}$ is injective on Υ_r, which can be implied by the condition $\text{null}(\mathcal{G}) \cap \Upsilon_{2r} = \{0\}$. Consequently, if $\text{null}(\mathcal{G}) \cap \Upsilon_{2r} = \{0\}$, (4) will recover any $\mathcal{X}_0 \in \Upsilon_r$. We expect this to occur when the number of measurements significantly exceeds the number of intrinsic degrees of freedom of a generic element of Υ_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 Υ_r:

Theorem 2 *Whenever $m \geq (2r)^K + 2nrK + 1$, with probability one, $\text{null}(\mathcal{G}) \cap \Upsilon_{2r} = \{0\}$, and hence (4) recovers every $\mathcal{X}_0 \in \Upsilon_r$.*

The proof of Theorem 2 follows from a covering argument, which we establish in several steps. Let

$$\mathfrak{S}_{2r} = \{\mathcal{D} \mid \mathcal{D} \in \Upsilon_{2r}, \|\mathcal{D}\|_F = 1\}. \tag{6}$$

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:

4

Lemma 3 *Suppose that the covering number for $\mathfrak{S}_{2r}$ with respect to Frobenius norm, satisfies*

$$N(\mathfrak{S}_{2r}, \|\cdot\|_F, \varepsilon) \leq (\beta/\varepsilon)^d, \tag{7}$$

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* $\mathrm{null}(\mathcal{G}) \cap \Upsilon_{2r} = \{\mathbf{0}\}$.

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

$$[[\mathcal{C}; \mathbf{U}_1, \mathbf{U}_2, \cdots, \mathbf{U}_K]] := \mathcal{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \cdots \times_K \mathbf{U}_K, \tag{8}$$

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

Lemma 4 *Let $\mathcal{C}, \mathcal{C}' \in \mathbb{R}^{r_1, \dots, r_K}$, and $\mathbf{U}_1, \mathbf{U}'_1 \in \mathbb{R}^{n_1 \times r_1}, \dots, \mathbf{U}_K, \mathbf{U}'_K \in \mathbb{R}^{n_K \times r_K}$ with $\mathbf{U}_i^* \mathbf{U}_i = \mathbf{U}_i'^* \mathbf{U}_i' = \mathbf{I}$, and $\|\mathcal{C}\|_F = \|\mathcal{C}'\|_F = 1$. Then*

$$\left\| [[\mathcal{C}; \mathbf{U}_1, \dots, \mathbf{U}_K]] - [[\mathcal{C}'; \mathbf{U}'_1, \dots, \mathbf{U}'_K]] \right\|_F \leq \left\| \mathcal{C} - \mathcal{C}' \right\|_F + \sum_{i=1}^{K} \left\| \mathbf{U}_i - \mathbf{U}'_i \right\|. \tag{9}$$

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 $\mathcal{C}$ and $\{\mathbf{U}_i\}$. The total size of the resulting ε net is thus bounded by the following lemma:

Lemma 5 $N(\mathfrak{S}_{2r}, \|\cdot\|_F, \varepsilon) \leq (3(K+1)/\varepsilon)^{(2r)^K + 2nrK}$

With these observations in hand, Theorem 2 follows immediately.

3. Convexification: Sum of Nuclear Norms?

Since the nonconvex problem (4) 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 (4) with nuclear norms. Due to convexity, the resulting vector optimization problem can be solved by the following scalar optimization:

$$\min_{\mathcal{X}} \quad \sum_{i=1}^{K} \lambda_i \|\mathcal{X}_{(i)}\|_* \quad \text{s.t.} \quad \mathcal{G}[\mathcal{X}] = \mathcal{G}[\mathcal{X}_0], \tag{10}$$

where $\lambda \geq \mathbf{0}$. The optimization (10) was first introduced by (Liu et al., 2009) and has been used successfully in applications in imaging (Semerci et al., 2013; Kreimer and Sacchi, 2013; Li and 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 (10) provably recovers the underlying tensor $\mathcal{X}_0$, when the number of measurements m is sufficiently large. The following is a (simplified) corollary of results of Tomioka et. al. (2011) [2]:

---

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

Corollary 6 (of (Tomioka et al., 2011), Theorem 3) *Suppose that $\mathcal{X}_0$ has Tucker rank $(r, \ldots, r)$, and $m \geq Crn^{K-1}$, where C is a constant. Then with high probability, $\mathcal{X}_0$ is the optimal solution to (10), with each $\lambda_i = 1$.*

This result shows that there *is* a range in which (10) 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 $\mathcal{X}_0$, 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 (10):

Theorem 7 *Let $\mathcal{X}_0 \in \Upsilon_r$ be nonzero. Set $\kappa = \min_i \left\{ \left\| (\mathcal{X}_0)_{(i)} \right\|_*^2 / \|\mathcal{X}_0\|_F^2 \right\} \times n^{K-1}$. Then if the number of measurements $m \leq \kappa - 2$, $\mathcal{X}_0$ is not the unique solution to (10), with probability at least $1 - 4\exp(-\frac{(\kappa-m-2)^2}{16(\kappa-2)})$. Moreover, there exists $\mathcal{X}_0 \in \Upsilon_r$ for which $\kappa = rn^{K-1}$.*

This implies that Corollary 6 (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 (10): although a generic element $\mathcal{X}_0$ of Υ_r can be described using at most $r^K + nrK$ real numbers, we require $\Omega(rn^{K-1})$ observations to recover it using (10). Theorem 7 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.

3.1 General lower bound for multiple structures

The poor behavior of (10) is an instance of a much more general phenomenon, first discovered by Oymak et. al. (2012). Our target tensor $\mathcal{X}_0$ 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 combining the convex relaxations for each of the individual structures. In the tensor case, this yields (10). Surprisingly, this combination is often not significantly more powerful than the best single regularizer (Oymak et al., 2012). We obtain Theorem 7 as a consquence of a new, general result of this nature, using a geometric framework introduced in (Amelunxen et al., 2014). 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.

Setup.   In general, we are interested in recovering a signal x_0 with several low-dimensional structures simultaneously, based on generic measurements with respect to x_0. Here the target signal x_0 could lie in any finite dimensional Hilbert space (e.g. a vector in $\mathbb{R}^n$, a matrix

---

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

in $\mathbb{R}^{n_1 \times n_2}$, a tensor in $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_K}$), but without loss of generality, we will consider $x_0 \in \mathbb{R}^n$. Let $\|\cdot\|_{(i)}$ be the penalty norm corresponding to the i-th structure (e.g. ℓ_1, nuclear norm). Consider the following *sum-of-norms (SON)* model,

$$\min_{x \in \mathbb{R}^n} \quad f(x) := \lambda_1 \|x\|_{(1)} + \lambda_2 \|x\|_{(2)} + \cdots + \lambda_K \|x\|_{(K)} \qquad \text{subject to} \quad \mathcal{G}[x] = \mathcal{G}[x_0], \qquad (11)$$

where $\mathcal{G}[\cdot]$ is a Gaussian measurement operator, and $\lambda > 0$. In the subsequent analysis, we will evaluate the performance of (11) in terms of recovering x_0, where the only assumption we require is:

Assumption 8 *The target signal x_0 is nonzero.*

Optimality condition. Is x_0 the unique optimal solution to (11)? Recall that the descent cone of a function f at a point x_0 is defined as

$$\mathcal{C}(f, x_0) := \text{cone}\{v \mid f(x_0 + v) \leq f(x_0)\}, \qquad (12)$$

which, in short, will be denoted as $\mathcal{C}$. Then x_0 is the unique optimal solution if and only if $\text{null}(\mathcal{G}) \cap \mathcal{C} = \{0\}$. Conversely, recovery fails if $\text{null}(\mathcal{G})$ has nontrivial intersection with $\mathcal{C}$.

Since $\mathcal{G}$ 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.

Denote the polar cone of $\mathcal{C}$ as $\mathcal{C}^\circ$, i.e.

$$\mathcal{C}^\circ := \left\{ u \in \mathbb{R}^n \mid \sup_{x \in \mathcal{C}} \langle u, x \rangle \leq 0 \right\}. \qquad (13)$$

Because polarity reverses inclusion, we expect that $\mathcal{C}$ will be *large* whenever $\mathcal{C}^\circ$ is *small*, which leads us to control the size of $\mathcal{C}^\circ$.

As $f(x_0) \neq 0 = \min_{x \in \mathbb{R}^n} f(x)$, it can be verified that (Rockafellar, 1997, Thm. 23.7)

$$\mathcal{C}^\circ = \text{cone}\left(\partial f(x_0)\right) = \text{cone}\left(\sum_{i \in [K]} \lambda_i \partial f_i(x_0) \right), \qquad (14)$$

where the sum is made in Minkowski sense.

In order to control the size of C^o based on (14), we will next establish some basic geometric properties for each single norm.

Properties for each single norm. Consider a general single norm $\|\cdot\|_\diamond$ and denote its dual norm (a.k.a. polar function) as $\|\cdot\|_\diamond^\circ$, i.e. for any $u \in \mathbb{R}^n$,

$$\|u\|_\diamond^\circ := \sup_{\|x\|_\diamond \leq 1} \langle x, u \rangle. \qquad (15)$$

Define $L := \sup_{x \neq 0} \|x\|_\diamond / \|x\|$, which implies that $\|\cdot\|_\diamond$ is L-Lipschitz: $\|x\|_\diamond \leq L\|x\|$ for all x. Then we also have $\|u\| \leq L\|u\|_\diamond^\circ$ for all u as

$$\|u\|_\diamond^\circ = \sup_{\|x\|_\diamond \leq 1} \langle x, u \rangle \geq \sup_{L\|x\| \leq 1} \langle x, u \rangle = \sup_{\|x\| \leq 1/L} \langle x, u \rangle = \frac{1}{L} \|u\|. \qquad (16)$$
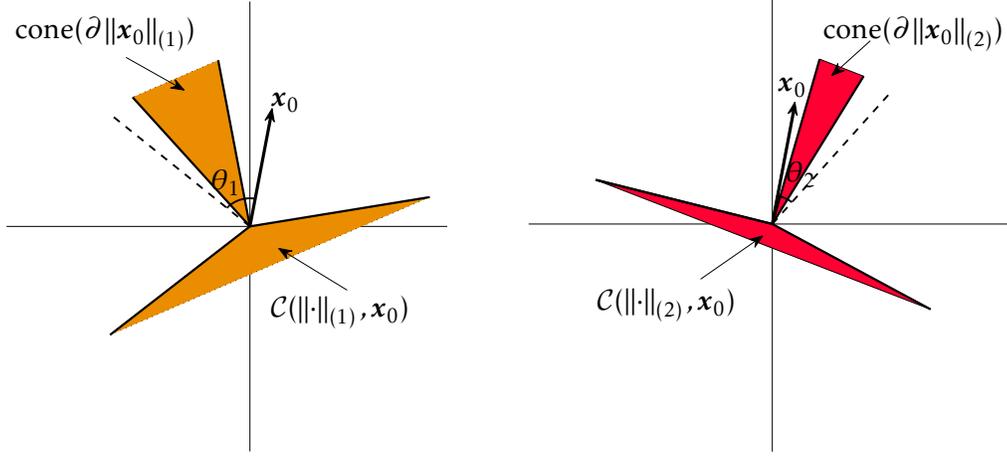
Figure 1: **Cones and their polars for convex regularizers $\|\cdot\|_{(1)}$ and $\|\cdot\|_{(2)}$ respectively.**
Suppose our x_0 has two sparse structures simultaneously. Regularizer $\|\cdot\|_{(1)}$ has a larger conic hull of subdifferential at x_0, i.e. $\mathrm{cone}(\partial\|x_0\|_{(1)})$, which results in a smaller descent cone. Thus minimizing $\|\cdot\|_{(1)}$ is more likely to recover x_0 than minimizing $\|\cdot\|_{(2)}$. Consider convex regularizer $f(x) = \|x_0\|_{(1)} + \|x_0\|_{(2)}$. Suppose as depicted, $\theta_1 \geq \theta_2$. Then both $\mathrm{cone}(\partial\|x_0\|_{(1)})$ and $\mathrm{cone}(\partial\|x_0\|_{(2)})$ are in the circular cone $\mathrm{circ}(x_0, \theta_1)$. Thus we have: $\mathrm{cone}\big(\partial f(x_0)\big) = \mathrm{cone}(\partial\|x_0\|_{(1)} + \partial\|x_0\|_{(2)}) \subseteq \mathrm{conv}\big\{\mathrm{circ}(x_0, \theta_1), \mathrm{circ}(x_0, \theta_2)\big\} = \mathrm{circ}(x_0, \theta_1)$.

In addition, noting that

$$\partial\|\cdot\|_\diamond(x) = \{u \mid \langle u, x\rangle = \|x\|_\diamond,\ \|u\|_\diamond^\circ \leq 1\}, \tag{17}$$

for any $u \in \partial\|\cdot\|_\diamond(x_0)$, we have

$$\cos\left(\angle(u, x_0)\right) := \frac{\langle u, x_0\rangle}{\|u\|\,\|x_0\|} \ \geq\ \frac{\|x_0\|_\diamond}{L\|u\|_\diamond^\circ\|x_0\|} \ \geq\ \frac{\|x_0\|_\diamond}{L\|x_0\|}. \tag{18}$$

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

$$\mathrm{circ}(x, \theta) = \{z \mid \angle(z, x) \leq \theta\}, \tag{19}$$

denote the *circular cone* with axis x and angle θ. Then with $\theta := \cos^{-1}(\|x_0\|_\diamond/L\|x_0\|)$,

$$\partial\|\cdot\|_\diamond(x_0) \subseteq \mathrm{circ}(x_0, \theta). \tag{20}$$

Table 1 describes the angle parameters θ for various structure inducing norms. Notice that in general, more complicated x_0 leads to smaller angles θ. For example, if 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 x_0 becomes more dense, $\partial\|\cdot\|_\diamond$ is contained in smaller and smaller circular cones.

Polar cone $\subseteq$ circular cone. 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 $\boldsymbol{x}_0$, $\partial f(\boldsymbol{x}_0)$ itself is also contained in a circular cone, and thus based on (14), we have

Lemma 9 *For $\boldsymbol{x}_0 \neq \boldsymbol{0}$, set $\theta_i = \cos^{-1}\left(\|\boldsymbol{x}_0\|_{(i)}/L_i\|\boldsymbol{x}_0\|\right)$, where $L_i = \sup_{\boldsymbol{x} \neq \boldsymbol{0}} \|\boldsymbol{x}\|_{(i)}/\|\boldsymbol{x}\|$. Then*

$$\mathcal{C}^\circ = \mathrm{cone}\left(\partial f(\boldsymbol{x}_0)\right) \subseteq \mathrm{circ}\left(\boldsymbol{x}_0, \max_{i \in [K]} \theta_i\right). \tag{21}$$

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. Figure 1 visualizes this geometry.

Statistical Dimension. How does this behavior affect the recoverability of $\boldsymbol{x}_0$ via (11)? 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 the elegant framework introduced by Amelunxen et al. (2014). They define the *statistical dimension* of the convex cone $\mathcal{C}$ to be the expected norm square of the projection of a standard Gaussian vector onto C:

$$\delta(\mathcal{C}) := \mathbb{E}_{\mathbf{g} \sim \text{i.i.d.} \mathcal{N}(0,1)}\left[\|\mathcal{P}_{\mathcal{C}}(\mathbf{g})\|^2\right]. \tag{22}$$

Using tools from spherical integral geometry, Amelunxen et al. (2014) 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 (11), we need only one side of their result with slight modifications:

Corollary 10 *Let $\mathcal{G} : \mathbb{R}^n \to \mathbb{R}^m$ be a Gaussian operator, and $\mathcal{C}$ a convex cone. Then if $m \leq \delta(\mathcal{C})$,*

$$\mathbb{P}\left[\mathcal{C} \cap \mathrm{null}(\mathcal{G}) = \{\boldsymbol{0}\}\right] \leq 4\exp\left(-\frac{(\delta(\mathcal{C}) - m)^2}{16\delta(\mathcal{C})}\right). \tag{23}$$

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 9, we calculate

$$\delta(\mathcal{C}) = n - \delta(\mathcal{C}^\circ) = n - \delta(\mathrm{cone}(\partial f(\boldsymbol{x}_0))) \geq n - \delta(\mathrm{circ}(\boldsymbol{x}_0, \max_i \theta_i)). \tag{24}$$

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 $\boldsymbol{x} \in \mathbb{R}^{n_1 \times n_2}$	$c = \#\{j \mid \boldsymbol{x}\mathbf{e}_j \neq \boldsymbol{0}\}$	$\sum_j \|\boldsymbol{x}\mathbf{e}_j\|$	$[\frac{1}{n_2}, \frac{c}{n_2}]$	$[n_1, cn_1]$
Low-rank $\boldsymbol{x} \in \mathbb{R}^{n_1 \times n_2}$ $(n_1 \geq n_2)$	$r = \mathrm{rank}(\boldsymbol{x})$	$\|\boldsymbol{x}\|_*$	$[\frac{1}{n_2}, \frac{r}{n_2}]$	$[n_1, rn_1]$

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., 2014; McCoy, 2013).

Lemma 11 $\delta(\text{circ}(\boldsymbol{x}_0, \theta)) \leq n \sin^2 \theta + 2$.

Finally, by combining (24) and Lemma 11, we have $\delta(\mathcal{C}) \geq n \min_i \cos^2 \theta_i - 2$. Using Corollary 10, we obtain:

Theorem 12 (SON model.) *Suppose the target signal $\boldsymbol{x}_0 \neq \boldsymbol{0}$. For each i-th norm ($i \in [K]$), define $L_i := \sup_{\boldsymbol{x} \neq \boldsymbol{0}} \|\boldsymbol{x}\|_{(i)} / \|\boldsymbol{x}\|$. Set*

$$\kappa_i \;=\; \frac{n \|\boldsymbol{x}_0\|_{(i)}^2}{L_i^2 \|\boldsymbol{x}_0\|^2} \;=\; n \cos^2(\theta_i), \quad and \quad \kappa = \min_i \kappa_i.$$

Then the statistical dimension of the descent cone of f at the point $\boldsymbol{x}_0$: $\delta(\mathcal{C}(f, \boldsymbol{x}_0)) \geq \kappa - 2$, and thus if the number of generic measurements $m \leq \kappa - 2$,

$$\mathbb{P}\left[\, \boldsymbol{x}_0 \text{ is the unique optimal solution to (11)} \,\right] \leq 4 \exp\left(-\frac{(\kappa - m - 2)^2}{16(\kappa - 2)}\right). \qquad (25)$$

Consequently, for reliable recovery, the number of measurements needs to be at least proportional to κ.[4] 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 12 together with Table 1 leads us to the phenomenon that 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 that the subdifferential of a norm at $\boldsymbol{x}_0$ is contained in a relatively small circular cone with central axis $\boldsymbol{x}_0$.

Extension. Here we consider a slightly more general setup: a signal $\boldsymbol{x}_0 \in \mathbb{R}^n$, *after appropriate linear transforms*, has K low-dimensional structures simultaneously. These linear transforms can be quite general, and could be either prescribed by experts or adaptively learned from training data.

In specific, for any i in $[K]$, there exists an appropriate linear transform $\mathcal{A}_i : \mathbb{R}^n \to \mathbb{R}^{m_i}$ such that $\mathcal{A}_i[\boldsymbol{x}_0]$ follows a parsimonious model in $\mathbb{R}^{m_i}$ (e.g. sparsity, low-rank). Let $\|\cdot\|_{(i)}$ be the penalty norms corresponding to the i-th structure (e.g. ℓ_1, nuclear norm). Based on generic measurements collected, it is natural to recover $\boldsymbol{x}_0$ using the following *sum-of-composite-norms (SOCN)* formulation

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(x) := \lambda_1 \|\mathcal{A}_1[\boldsymbol{x}]\|_{(1)} + \lambda_2 \|\mathcal{A}_2[\boldsymbol{x}]\|_{(2)} + \cdots + \lambda_K \|\mathcal{A}_K[\boldsymbol{x}]\|_{(K)} \qquad \text{s.t.} \quad \mathcal{G}[\boldsymbol{x}] = \mathcal{G}[\boldsymbol{x}_0], \quad (26)$$

where $\mathcal{G}[\cdot]$ is a Gaussian measurement operator, and $\lambda > 0$. Essentially following the same reasoning as above, a result similar to Theorem 12, stating a lower bound on the number of generic measurements required, can be achieved:

---

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

Theorem 13 (SOCN model) *Suppose the target signal $x_0 \notin \cap_{i\in[K]} null(\mathcal{A}_i)$. For each $i \in [K]$, define $L_i = \sup_{x \in \mathbb{R}^{m_i}\setminus\{0\}} \|x\|_{(i)} / \|x\|$. Set*

$$\kappa_i = \frac{n\|\mathcal{A}_i x_0\|_{(i)}^2}{L_i^2 \|\mathcal{A}_i\|^2 \|x_0\|^2}, \quad and \quad \kappa = \min_i \kappa_i.$$

Then if $m \leq \kappa - 2$,

$$\mathbb{P}\left[\, x_0 \text{ is the unique optimal solution to (26)} \right] \leq 4\exp\left(-\frac{(\kappa - m - 2)^2}{16(\kappa - 2)}\right). \tag{27}$$

Remark 14 *Clearly, Theorem 12 can be regarded as a special case of Theorem 13, where $\mathcal{A}_i's$ are all identity operators.*

3.2 Low-rank tensors

We can specialize Theorem 12 to low-rank tensors as follows: if the target signal $\mathcal{X}_0 \in \Upsilon_r$, i.e. a K-mode $n \times n \times \cdots \times n$ tensor of Tucker rank $(r, r, \ldots, r)$, then for each $i \in [K]$, $\|\cdot\|_{(i)} := \|(\cdot)_{(i)}\|_*$ is $L_i = \sqrt{n}$-Lipschitz. Hence

$$\kappa = \min_i \left\{ \|(\mathcal{X}_0)_{(i)}\|_*^2 / \|\mathcal{X}_0\|_F^2 \right\} n^{K-1}. \tag{28}$$

The term $\min_i \left\{ \|(\mathcal{X}_0)_{(i)}\|_*^2 / \|\mathcal{X}_0\|_F^2 \right\}$ lies between 1 and r, inclusively. For example, if $\mathcal{X}_0 \in \Upsilon_1$, then that term is equal to 1; if $\mathcal{X}_0 = [[\mathcal{C}, U_1, \ldots, U_K]]$ with $U_i^* U_i = I$ and $\mathcal{C}$ (super)diagonal ($\mathcal{C}_{i_1 \ldots i_r} = \mathbf{1}_{\{i_1 = i_2 = \cdots = i_r\}}$), then that term is equal to r. That exactly yields Theorem 7.

Empirical estimates of the statistical dimension. As noted in Theorem 12, the statistical dimension of the descent cone $\delta(\mathcal{C})$ plays a crucial role in deriving our lower bound for the number of generic measurements. In the following, we will empirically justify our theoretical result for $\delta(\mathcal{C})$ under the setting of our interest, low-rank tensors.

Consider $\mathcal{X}_0$ as a K-mode $n \times n \times \cdots \times n$ (super)diagonal tensor with only the first r diagonal entries as 1 and 0 elsewhere. Clearly, $\mathcal{X}_0 \in \Upsilon_r$, and Corollary 6, Theorem 12 and expression (28) yield

$$\delta(\mathcal{C}) := \delta\left(\mathcal{C}\left(\sum_{i=1}^K \|\mathcal{X}_{(i)}\|_*, \, \mathcal{X}_0\right)\right) \geq rn^{K-1} - 2, \quad and \quad \delta(\mathcal{C}) = \Theta(rn^{K-1}). \tag{29}$$

In the following, we will empirically corroborate (29) based on recent results developed in statistical decision theory.

In order to estimate $\delta(\mathcal{C})$, we construct a perturbed observation $\mathcal{Z}_0 = \mathcal{X}_0 + \sigma\mathcal{G}$, where vec($\mathcal{G}$) is a standard normal vector and σ is the standard deviation parameter. Then

$$\hat{\mathcal{X}} := \arg\min_{\mathcal{X}} \|\mathcal{Z}_0 - \mathcal{X}\|_F \quad \text{s.t.} \quad \sum_{i=1}^K \|\mathcal{X}_{(i)}\|_* \leq Kr = \sum_{i=1}^K \|(\mathcal{X}_0)_{(i)}\|_*, \tag{30}$$

is computed as an estimate of $\mathcal{X}_0$. Due to the recent results from Oymak and Hassibi (2013), the normalized mean-squared error (NMSE), defined as

$$\text{NMSE}(\sigma) := \frac{\mathbb{E}\left[\left\|\hat{\mathcal{X}} - \mathcal{X}_0\right\|_F^2\right]}{\sigma^2}, \tag{31}$$

is a decreasing function over $\sigma > 0$ and

$$\delta(\mathcal{C}) := \lim_{\sigma \to 0^+} \text{NMSE}(\sigma). \tag{32}$$

Therefore, for small σ, NMSE serves a good estimator for $\delta(\mathcal{C})$. For more discussions on related tensor denoising problems, see Appendix D.

In our experiment, we take $\sigma = 10^{-8}$ and for different triples of (K, r, n), we measure the empirical NMSE averaged over 10 repeats. Dykstra's Algorithm (see Appendix E.1) is exploited to solve problem (30). The numerical outputs are presented in Figure 2, which firmly conforms to the theoretical results displayed in (29).

4. A Better Convexification: Square Deal

The number of measurements promised by Corollary 6 and Theorem 7 is actually the same (up to constants) as the number of measurements required to recover a tensor $\mathcal{X}_0$ 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

$$\text{minimize } \|\mathcal{X}_{(1)}\|_* \quad \text{subject to} \quad \mathcal{G}[\mathcal{X}] = \mathcal{G}[\mathcal{X}_0] \tag{33}$$

also recovers $\mathcal{X}_0$ w.h.p. when $m \geq Crn^{K-1}$.

This suggests a more mundane explanation for the difficulty with (10): the term rn^{K-1} comes from the need to reconstruct the right singular vectors of the $n \times n^{K-1}$ matrix $\mathcal{X}_{(1)}$. 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 $\mathcal{X}_0$ is four or larger.

Square reshaping. 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 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 $\mathcal{X}_{(1)}$ to a more square matrix while preserving the low-rank property. Let $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_K}$. Select some $j \in [K]$. Then we
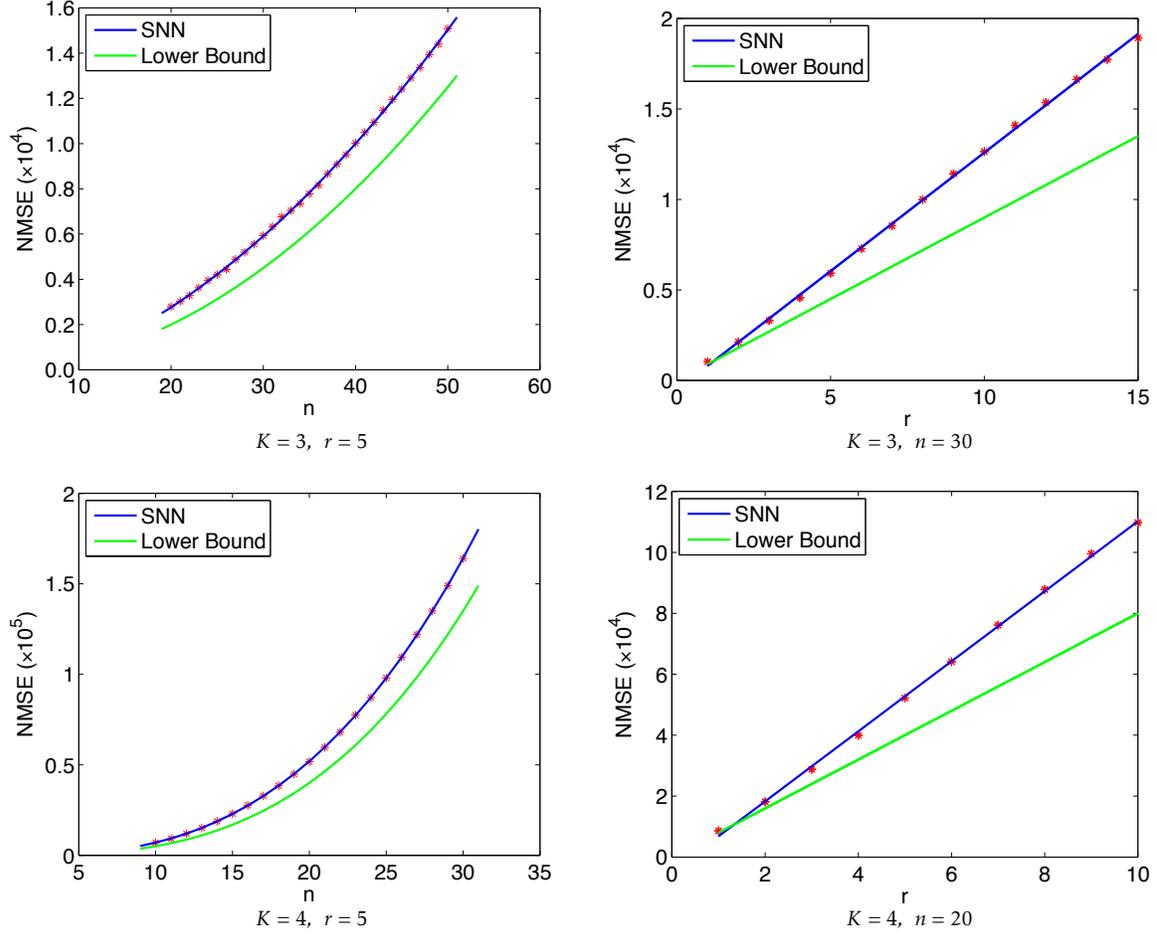
Figure 2: **Lower bound for statistical dimension.** Each red cross represents the empirical estimate of $\delta(\mathcal{C})$ for one particular triple (K, r, n). The blue curves fit the red dots based on the relationship $\delta(\mathcal{C}) = \Theta(rn^{K-1})$. In specific, in the left top (resp. left bottom) figure, we fit the red crosses with a quadratic (resp. cubic) curve; and in the right figures, we fit the red crosses with linear curves. Note that the red crosses fit pretty well with the blue curves, which is consistent with our result that $\delta(\mathcal{C}) = \Theta(rn^{K-1})$. The blue curves correspond to our lower bound $rn^{K-1} - 2$, which tightly lie below the red crosses. This empirically corroborates the lower bound result $\delta(\mathcal{C}) \geq rn^{K-1} - 2$.

define matrix $\boldsymbol{\mathcal{X}}_{[j]}$ as[5]

$$\boldsymbol{\mathcal{X}}_{[j]} = \mathsf{reshape}\left(\boldsymbol{\mathcal{X}}_{(1)}, \prod_{i=1}^{j} n_i, \prod_{i=j+1}^{K} n_i\right). \tag{34}$$

---

5. One can also think of (34) as embedding the tensor $\boldsymbol{\mathcal{X}}$ into the matrix $\boldsymbol{\mathcal{X}}_{[j]}$ as follows: $\boldsymbol{\mathcal{X}}_{i_1, i_2, \cdots, i_K} = \left(\boldsymbol{\mathcal{X}}_{[j]}\right)_{a,b}$, where

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

13

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

Lemma 15 *(1) If $\mathcal{X}$ has CP decomposition $\mathcal{X} = \sum_{i=1}^{r} \lambda_i a_i^{(1)} \circ a_i^{(2)} \circ \cdots \circ a_i^{(K)}$, then*

$$\mathcal{X}_{[j]} = \sum_{i=1}^{r} \lambda_i (a_i^{(j)} \otimes \cdots \otimes a_i^{(1)}) \circ (a_i^{(K)} \otimes \cdots \otimes a_i^{(j+1)}). \tag{35}$$

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

$$\mathcal{X}_{[j]} = (\mathbf{U}_j \otimes \cdots \otimes \mathbf{U}_1) \mathcal{C}_{[j]} (\mathbf{U}_K \otimes \cdots \otimes \mathbf{U}_{j+1})^*. \tag{36}$$

Using Lemma 15 and the fact that $\text{rank}(\mathbf{A} \otimes \mathbf{B}) = \text{rank}(\mathbf{A})\text{rank}(\mathbf{B})$, we obtain:

Lemma 16 *Let $\text{rank}_{tc}(\mathcal{X}) = (r_1, r_2, \cdots, r_K)$, and $\text{rank}_{cp}(\mathcal{X}) = r_{cp}$. Then $\text{rank}(\mathcal{X}_{[j]}) \le r_{cp}$, and $\text{rank}(\mathcal{X}_{[j]}) \le \min\left\{ \prod_{i=1}^{j} r_i, \prod_{i=j+1}^{K} r_i \right\}$.*

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

$$\text{minimize} \quad \left\| \mathcal{X}_{[j]} \right\|_* \quad \text{subject to} \quad \mathcal{G}[\mathcal{X}] = \mathcal{G}[\mathcal{X}_0]. \tag{37}$$

Using Lemma 16 and (Chandrasekaran et al., 2012), we can prove that this relaxation exactly recovers $\mathcal{X}_0$, when the number of measurements is sufficiently large:

Theorem 17 *Consider a K-way tensor with the same length (say n) along each mode. (1) If $\mathcal{X}_0$ has CP rank r, using (37) with $j = \lceil \frac{K}{2} \rceil$, $m \ge Crn^{\lceil \frac{K}{2} \rceil}$ is sufficient to recover $\mathcal{X}_0$ with high probability. (2) If $\mathcal{X}_0$ has Tucker rank $(r, r, \cdots, r)$, using (37) 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 $\mathcal{X}_0$ with high probability.*

The number of measurements $O(r^{\lfloor \frac{K}{2} \rfloor} n^{\lceil \frac{K}{2} \rceil})$ required to recover $\mathcal{X}$ with square reshaping (37), 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.

Low-rank tensor completion. We corroborate the improvement of square reshaping with numerical experiments on *low-rank tensor completion (LRTC)*. LRTC attempts to reconstruct the 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 recovery guarantees for both our square model (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, it is more difficult to establish a negative result for the SNN model (like Theorem 7). Nonetheless, numerical results below clearly indicate the advantage of our square approach, complementing our theoretical results established in previous sections.

We generate our four-way tensors $\boldsymbol{\mathcal{X}}_0 \in \mathbb{R}^{n \times n \times n \times n}$ as $\boldsymbol{\mathcal{X}}_0 = \boldsymbol{\mathcal{C}}_0 \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 \times_4 \mathbf{U}_4$, where $\boldsymbol{\mathcal{C}}_0 \in \mathbb{R}^{r_1 \times r_2 \times r_3 \times r_4}$ and $\mathbf{U}_i \in \mathbb{R}^{n_i \times r_i}$ for each $i \in [4]$ are constructed under the random Gaussian models (by MATLAB command): each entry of $\mathbf{C}_0, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3$ and $\mathbf{U}_4$ is generated using randn(). The observed entries are chosen uniformly with ratio ρ. We compare the recovery performances between

$$\text{minimize}_{\boldsymbol{\mathcal{X}}} \qquad \sum_{i=1}^{K} \left\| \boldsymbol{\mathcal{X}}_{(i)} \right\|_* \qquad \text{subject to} \qquad \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0], \quad \text{and} \qquad (38)$$

$$\text{minimize}_{\boldsymbol{\mathcal{X}}} \qquad \left\| \boldsymbol{\mathcal{X}}_{\{1,2\}} \right\|_* \qquad \text{subject to} \qquad \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}] = \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_0]. \qquad (39)$$

We fix (r_1, r_2, r_3, r_4) as $(1,1,1,1)$ and $(1,1,2,2)$ respectively. For each choice of (r_1, r_2, r_3, r_4), 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 $\left\| \boldsymbol{\mathcal{X}}^\star - \boldsymbol{\mathcal{X}}_0 \right\|_F / \|\boldsymbol{\mathcal{X}}_0\|_F \leq 10^{-2}$.

The optimization problems are solved using efficient first-order methods. Since (39) is equivalent to standard matrix completion, we use the existing solver ALM (Lin et al., 2010). For the sum of nuclear norms minimization (38), we implement the Douglas-Rachford algorithm (see Appendix E.2 for details).

Figure 3 plots the fraction of correct recovery for each pair. Clearly, the square approach succeeds in a much larger region.

General reshaping. 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 (34) but with a relabeling preprocessing. In specific, for $1 \leq k \leq K$, we relabel the k-th mode as the original i_k-th mode. Regarding this relabeled tensor $\overline{\boldsymbol{\mathcal{X}}}$, we can define

$$\boldsymbol{\mathcal{X}}_{\mathcal{I}} := \overline{\boldsymbol{\mathcal{X}}}_{[j]} = \text{reshape}\left(\overline{\boldsymbol{\mathcal{X}}}_{(1)}, \prod_{k=1}^{j} n_{i_k}, \prod_{k=j+1}^{K} n_{i_k} \right). \qquad (40)$$

Lemma 16 and Theorem 17 can then also be easily extended. As suggested by Theorem 17 (after modification), to maximize the effect of our square model, we would like to choose $\mathcal{I}$ to minimize the quantity,

$$\text{rank}(\boldsymbol{\mathcal{X}}_{\mathcal{I}}) \cdot \max\left\{ \prod_{k=1}^{j} n_{i_k}, \prod_{k=j+1}^{K} n_{i_k} \right\}. \qquad (41)$$

15
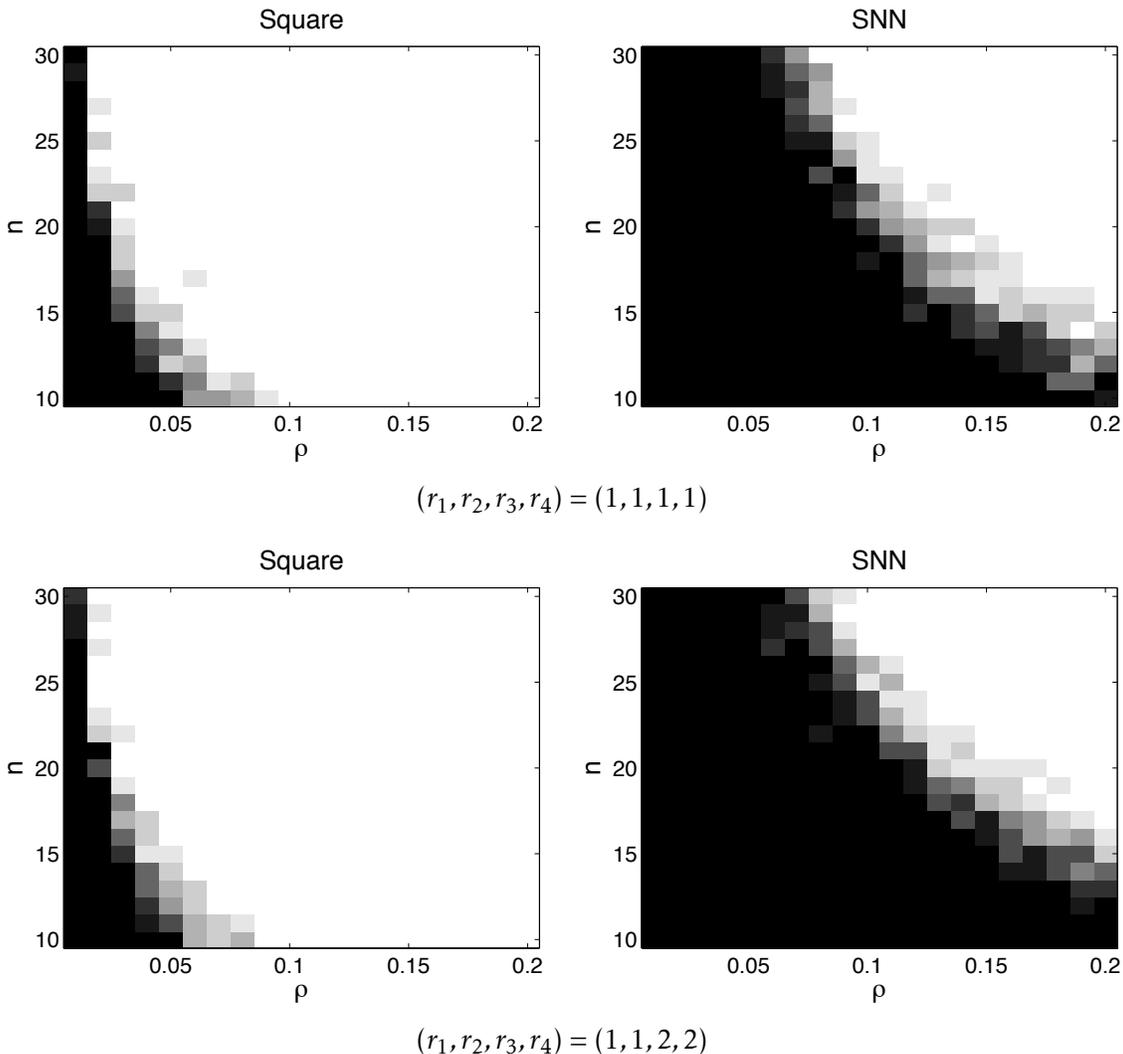
Figure 3: **Tensor completion with Gaussian random data.** The colormap indicates the fraction of instances that are correctly recovered for each (ρ, n)-pair, which increases with brightness from 100% failure (black) to 100% success (white).

In practice, normally we do not know the exact rank of each mode, and hence (41) cannot be computed directly. However, prior knowledge of the physical properties of the underlying tensor can provide some guidance – e.g., in multi-spectral video data, the video tensor tends to be low rank in both the wavelength and the temporal modes, so grouping these two modes would lead to a natural low-rank matrix. Hence, practically, we should set $\mathcal{I}$ by taking both the size and the physical characteristics of the true tensor into consideration.

Remark 18 *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 the SNN model.*

5. 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 (4), we show that $(2r)^K + 2nrK + 1$ measurements are sufficient to recover any $\mathcal{X}_0 \in \Upsilon_r$ almost surely. For the conventional convex surrogate sum-of-nuclear-norms (SNN) model (10), 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 (37). We then prove that $O(r^{\lfloor \frac{K}{2} \rfloor} n^{\lceil \frac{K}{2} \rceil})$ measurements are sufficient to recover a tensor $\mathcal{X}_0 \in \Upsilon_r$ with high probability. Though the theoretical results only pertain to Gaussian measurements, our numerical experiments still suggest the square reshaping model outperforms the SNN model in other settings. 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 (37), is always within a constant of it, and is much better for small r and $K \geq 4$. Although this is a significant improvement, in contrast with the nonconvex model (4), 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 $K > 2$.

Since the release of our work (Mu et al., 2013) online, we note that several interesting models and algorithms have been proposed and analyzed, focusing on the low-rank tensor completion (LRTC) problem. Oymak et al. (2014) extended the negative result for the SNN model to more general sampling schemes. Yuan and Zhang (2014) analyzed the tensor nuclear norm model (though not computationally tractable) and established better sampling complexity result. Several other works – e.g. (Jain and Oh, 2014; Aswani, 2014), achieved better sample complexity using tractable methods by considering special subclasses of low-rank tensors. In addition, many works in the field of numerical optimization have designed efficient methods to solve LRTC related non-convex models, e.g. alternating minimization (Romera-Paredes et al., 2013; Xu et al., 2013), Riemannian optimization (Kressner et al., 2014), where empirical successes have been greatly witnessed. 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 12 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, 2014) 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.

Acknowledgments

References

D. Amelunxen. *Geometric analysis of the condition of the convex feasibility problem*. PhD thesis, PhD Thesis, Univ. Paderborn, 2011.

D. Amelunxen, M. Lotz, M. McCoy, and J. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Inform. Inference*, 3(3):224–294, 2014.

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

A. Aswani. Positive low-rank tensor completion. *arXiv preprint arXiv:1412.0620*, 2014.

B. W. Bader and T. G. Kolda. Matlab tensor toolbox version 2.6. *http://www.sandia.gov/ tgkolda/TensorToolbox/index-2.6.html*, 2015.

H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

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

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

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

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

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

P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

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

Y. C. Eldar, D. Needell, and Y. Plan. Uniqueness conditions for low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 33(2):309–314, 2012.

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

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

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

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

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

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

B. Huang, C. Mu, D. Goldfarb, and J. Wright. Provable low-rank tensor recovery. To appear in *Pacific Journal of Optimization*, preprint available at `http://www.optimization-online.org/DB_HTML/2014/02/4252.html`, 2014.

P. Jain and S. Oh. Provable tensor factorization with missing data. In *NIPS*, 2014.

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

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

D. Kressner, M. Steinlechner, and B. Vandereycken. Low-rank tensor completion by riemannian optimization. *BIT Numerical Mathematics*, 54(2):447–468, 2014.

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

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

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

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

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

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

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

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

C. Mu, B. Huang, J. Wright, and D. Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *ICML'14*, preprint available at *arXiv:1307.5870.*, 2013.

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.

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

S. Oymak and B. Hassibi. Sharp MSE bounds for proximal denoising. *arXiv preprint arXiv:1305.2714*, 2013.

S. Oymak, A. Jalali, M. Fazel, Y. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *arXiv preprint arXiv:1212.3753v1*, 2012.

S. Oymak, A. Jalali, M. Fazel, Y. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *arXiv preprint arXiv:1212.3753v3*, 2014.

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

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

E. Richard, G. R Obozinski, and J.P. Vert. Tight convex relaxations for sparse matrix factorization. In *NIPS*, 2014.

R. Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

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

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

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

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

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

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

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing*, *Theory and Applications*, pages 210–268. Cambridge University Press.

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

Y. Xu, R. Hao, W. Yin, and Z. Su. Parallel matrix factorization for low-rank tensor completion. To appear in *Inverse Problems and Imaging*, preprint available at *arXiv:1312.1254.*, 2013.

M. Yuan and C.-H. Zhang. On tensor completion via nuclear norm minimization. *arXiv preprint arXiv:1405.1773*, 2014.

Appendix A. Proofs for Section 2

Proof of Lemma 3. The arguments we used below are primarily adapted from (Eldar et al., 2012), where their interest is to establish the number of Gaussian measurements required to recover a low rank matrix by rank minimization.

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

$$\forall\, t > 0, \quad \mathbb{P}\left[\left|\langle \mathcal{G}_i, \mathcal{D} \rangle\right| < t\right] < 2t \cdot \frac{1}{\sqrt{2\pi}} = t\sqrt{\frac{2}{\pi}}. \tag{42}$$

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

$$\mathbb{P}\left[\left\|\mathcal{G}[\bar{\mathcal{D}}]\right\|_{\infty} < t\right] < \left(t\sqrt{2/\pi}\right)^m. \tag{43}$$

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

$$\left\|\mathcal{G}[\mathcal{D}]\right\|_{\infty} \geq \max_{\bar{\mathcal{D}} \in \Omega}\left\{\left\|\mathcal{G}[\bar{\mathcal{D}}]\right\|_{\infty} - \|\mathcal{G}\|_{F \to \infty}\left\|\bar{\mathcal{D}} - \mathcal{D}\right\|_F\right\} \tag{44}$$

$$\geq \min_{\bar{\mathcal{D}} \in \Omega}\left\{\left\|\mathcal{G}[\bar{\mathcal{D}}]\right\|_{\infty}\right\} - \varepsilon\|\mathcal{G}\|_{F \to \infty}. \tag{45}$$

Hence,

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

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

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

$$\leq \beta^d(2\sqrt{2/\pi})^m \varepsilon^{m-d}\log(1/\varepsilon)^m + \mathbb{P}\left[\|\mathcal{G}\|_{F \to \infty} > \log(1/\varepsilon)\right]. \tag{46}$$

Since $m \geq d + 1$, (46) goes to zero as $\varepsilon \searrow 0$. Hence, taking a sequence of decreasing ε, we can show that $\mathbb{P}\left[\inf_{\mathcal{D} \in \mathfrak{S}_{2r}} \|\mathcal{G}[\mathcal{D}]\|_{\infty} = 0\right] \leq t$ for every positive t, establishing the result.

Proof of Lemma 4. This follows from the basic fact that for any tensor $\mathcal{X}$ and matrix $\mathbf{U}$ of compatible size,

$$\|\mathcal{X} \times_k \mathbf{U}\|_F = \left\|\mathbf{U}\mathcal{X}_{(k)}\right\|_F \leq \|\mathbf{U}\|\left\|\mathcal{X}_{(k)}\right\|_F = \|\mathbf{U}\|\|\mathcal{X}\|_F. \tag{47}$$

Write

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

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

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

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

where the first inequality follows from triangle inequality and the second inequality follows from the fact that $\|\mathcal{C}\|_F = 1$, $\left\|\mathbf{U}_j\right\| = 1$, $\mathbf{U}_i^* \mathbf{U}_i = \mathbf{I}$ and $\mathbf{U}_i'^* \mathbf{U}_i' = \mathbf{I}$.

Proof of Lemma 5. 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 $\mathcal{O} = \{\mathbf{U} \in \mathbb{R}^{n \times r} \mid \mathbf{U}^*\mathbf{U} = \mathbf{I}\}$. Clearly, for any $\mathcal{C} \in \mathfrak{C}$, $\|\mathcal{C}\|_F = 1$, and for any $\mathbf{U} \in \mathcal{O}$, $\|\mathbf{U}\| = 1$. Thus by (Vershynin, 2007, Prop. 4) and (Vershynin, Lemma 5.2), there exists an $\frac{\varepsilon}{K+1}$-net $\mathfrak{C}'$ covering $\mathfrak{C}$ with respect to the Frobenius norm such that $\#\mathfrak{C}' \le (\frac{3(K+1)}{\varepsilon})^{(2r)^K}$, and there exists an $\frac{\varepsilon}{K+1}$-net $\mathcal{O}'$ covering $\mathcal{O}$ with respect to the operator norm such that $\#\mathcal{O}' \le (\frac{3(K+1)}{\varepsilon})^{2nr}$. Construct

$$\mathfrak{S}'_{2r} = \{[[\mathcal{C}'; \mathbf{U}'_1, \ldots, \mathbf{U}'_K]] \mid \mathcal{C}' \in \mathfrak{C}', \ \mathbf{U}'_i \in \mathcal{O}'\}. \tag{48}$$

Clearly $\#\mathfrak{S}'_{2r} \le \left(\frac{3(K+1)}{\varepsilon}\right)^{(2r)^K + 2nrK}$. 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 $\mathcal{D} = [[\mathcal{C}; \mathbf{U}_1, \cdots, \mathbf{U}_K]] \in \mathfrak{S}_{2r}$ where $\mathcal{C} \in \mathfrak{C}$ and $\mathbf{U}_i \in \mathcal{O}$, by our constructions above, there exist $\mathcal{C}' \in \mathfrak{C}'$ and $\mathbf{U}'_i \in \mathcal{O}'$ such that $\left\|\mathcal{C} - \mathcal{C}'\right\|_F \le \frac{3(K+1)}{\varepsilon}$ and $\left\|\mathbf{U}_i - \mathbf{U}'_i\right\| \le \frac{3(K+1)}{\varepsilon}$. Then $\mathcal{D}' = [[\mathcal{C}'; \mathbf{U}'_1, \cdots, \mathbf{U}'_K]] \in \mathfrak{S}'_{2r}$ is within ε-distance from $\mathcal{D}$, since by the triangle inequality derived in Lemma 2, we have

$$\left\|\mathcal{D} - \mathcal{D}'\right\|_F = \left\|[[\mathcal{C}; \mathbf{U}_1, \ldots, \mathbf{U}_K]] - [[\mathcal{C}'; \mathbf{U}'_1, \ldots, \mathbf{U}'_K]]\right\|_F \le \left\|\mathcal{C} - \mathcal{C}'\right\|_F + \sum_{i=1}^{K} \left\|\mathbf{U}_i - \mathbf{U}'_i\right\| \le \varepsilon. \tag{49}$$

This completes the proof.

Appendix B. Proofs for Section 3

Proof of Corollary 10. Denote $\lambda = \delta(\mathcal{C}) - m$. Then following the result derived by Amelunxen et al. (2014, Theorem 7.2), we have

$$\mathbb{P}\left[\mathcal{C} \cap \mathrm{null}(\mathcal{G}) = \{\mathbf{0}\}\right] \le 4\exp\left(-\frac{\lambda^2/8}{\min\{\delta(\mathcal{C}), \delta(\mathcal{C}^\circ)\} + \lambda}\right) \tag{50}$$

$$\le 4\exp\left(-\frac{\lambda^2/8}{\delta(\mathcal{C}) + \lambda}\right)$$

$$\le 4\exp\left(-\frac{(\delta(\mathcal{C}) - m)^2}{16\delta(\mathcal{C})}\right).$$

Proof of Lemma 11. Denote $\mathrm{circ}(\mathbf{e}_n, \theta)$ as $\mathrm{circ}_n(\theta)$, where $\mathbf{e}_n$ is the nth standard basis for $\mathbb{R}^n$. Since $\delta\big(\mathrm{circ}(\mathbf{x}_0, \theta)\big) = \delta\big(\mathrm{circ}(\mathbf{e}_n, \theta)\big)$, it is sufficient to prove $\delta\big(\mathrm{circ}_n(\theta)\big) \le n\sin^2\theta + 2$.

Let us first consider the case where n is *even*. Define a discrete random variable V supported on $\{0, 1, 2, \cdots, n\}$ with probability mass function $\mathbb{P}[V = k] = v_k$. Here v_k denotes the k-th intrinsic volumes of $\mathrm{circ}_n(\theta)$. Then it can be verified (see Amelunxen, 2011, Ex. 4.4.8)

$$v_k = \frac{1}{2}\binom{\frac{1}{2}(n-2)}{\frac{1}{2}(k-1)}\sin^{k-1}(\theta)\cos^{n-k-1}(\theta) \qquad \text{for } k = 1, 2, \cdots, n-1. \tag{51}$$

From Prop. 5.11 of (Amelunxen et al., 2014), we know that

$$\delta\big(\text{circ}_n(\theta)\big) = \mathbb{E}[V] = \sum_{k=1}^{n} \mathbb{P}[V \geq k]. \tag{52}$$

Moreover, by the interlacing result (Amelunxen et al., 2014, Prop. 5.6) and the fact that $\mathbb{P}[V \geq 2k] = \mathbb{P}[V \geq 2k-1] - \mathbb{P}[V = 2k-1]$, we have

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

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

$$
\begin{aligned}
&\vdots \qquad\qquad \vdots \quad \vdots
\end{aligned}
$$

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

Summing up the above inequalities, we have

$$
\begin{aligned}
\mathbb{E}[V] &= \sum_{k=1}^{n} \mathbb{P}[V \geq k] \tag{53}\\
&\leq \sum_{k=1,3,\cdots,n-1} 2(k-1)v_k + \sum_{k=1,3,\cdots,n-1} 3v_k \\
&\leq (n-2)\sin^2\theta + \frac{3}{2}\sum_{k=0}^{n} v_k \\
&\leq (n-2)\sin^2\theta + \frac{3}{2} = n\sin^2\theta + 2\cos^2\theta - \frac{1}{2},
\end{aligned}
$$

where the second last inequality follows from 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 (Amelunxen et al., 2014, Prop. 5.6).

Suppose n is *odd*. Since the intersection of $\text{circ}_{n+1}(\theta)$ with any n-dimensional linear subspace containing $\mathbf{e}_{n+1}$ is an isometric image of $\text{circ}_n(\theta)$, by Prop. 4.1 of (Amelunxen et al., 2014), we have

$$\delta(\text{circ}_n(\theta)) = \delta(\text{circ}_n(\theta) \times \{\mathbf{0}\}) \leq \delta(\text{circ}_{n+1}(\theta)) \leq (n+1)\sin^2\theta + 2\cos^2\theta - \frac{1}{2} \leq n\sin^2\theta + \cos^2\theta + \frac{1}{2}. \tag{54}$$

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

$$\delta\big(\text{circ}_n(\theta)\big) \leq n\sin^2\theta + \cos^2\theta + \frac{1}{2} < n\sin^2\theta + 2. \tag{55}$$

24

Proof of Theorem 12.   Notice that for any fixed $m > 0$, the function $f : t \to 4\exp\left(-\frac{(t-m)^2}{16t}\right)$ is decreasing for $t \geq m$. Then due to Corollary 10 and the fact that $\delta(\mathcal{C}) \geq \kappa - 2 \geq m$, we have

$$
\begin{aligned}
\mathbb{P}\left[\,x_0 \text{ is the unique optimal solution to (11)}\,\right] &= \mathbb{P}\left[\mathcal{C} \cap \text{null}(\mathcal{G}) = \{\mathbf{0}\}\right] &\qquad(56)\\
&\leq 4\exp\left(-\frac{(\delta(\mathcal{C}) - m)^2}{16\delta(\mathcal{C})}\right)\\
&\leq 4\exp\left(-\frac{(\kappa - m - 2)^2}{16(\kappa - 2)}\right).
\end{aligned}
$$

Proof of Theorem 13.   The argument for Theorem 12 can be easily adapted to prove Theorem 13, with the following additional observation regarding the function $\|\mathcal{A}[\cdot]\|_\diamond$, where $\|\cdot\|_\diamond$ is a norm in $\mathbb{R}^m$ with dual norm $\|\cdot\|_\diamond^\circ$, and $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^m$ is a linear operator satisfying $\mathcal{A}[x_0] \neq \mathbf{0}$. Essentially, we will next prove that $\partial\|\mathcal{A}[\cdot]\|_\diamond(x_0)$ is contained in a circular cone, which is analogous to (20).

For any $u \in \partial\|\mathcal{A}[\cdot]\|_\diamond(x_0)$, there exists a $v \in \partial\|\cdot\|_\diamond(\mathcal{A}[x_0])$ such that $u = \mathcal{A}^\star v$. Thus we have

$$
\cos(\angle(u, x_0)) = \frac{\langle u, x_0 \rangle}{\|u\|\,\|x_0\|} = \frac{\langle \mathcal{A}^* v, x_0 \rangle}{\|\mathcal{A}^* v\|\,\|x_0\|} \geq \frac{\langle v, \mathcal{A}x_0 \rangle}{\|\mathcal{A}\|\,\|v\|\,\|x_0\|}. \qquad(57)
$$

Define $L := \sup_{x \neq \mathbf{0}} \|x\|_\diamond/\|x\|_2$, which implies that $\|\cdot\|_\diamond$ is L-Lipschitz: $\|x\|_\diamond \leq L\|x\|$ for all x. Then $\|v\| \leq L\|v\|_\diamond^\circ$ for all v as well. Thus, we have

$$
\cos(\angle(u, x_0)) \geq \frac{\langle v, \mathcal{A}x_0 \rangle}{L\|\mathcal{A}\|\,\|v\|_\diamond^\circ\,\|x_0\|}. \qquad(58)
$$

Recall that

$$
\partial\|\cdot\|_\diamond(x) = \{v \mid \langle v, x \rangle = \|x\|_\diamond,\ \|v\|_\diamond^\circ \leq 1\}. \qquad(59)
$$

We can therefore further simplify

$$
\cos(\angle(u, x_0)) \geq \frac{\|\mathcal{A}x_0\|_\diamond}{L\|\mathcal{A}\|\,\|x_0\|}, \qquad(60)
$$

which is equivalent to saying

$$
\partial\|\mathcal{A}[\cdot]\|_\diamond(x_0) \subseteq \text{circ}(x_0, \theta), \qquad(61)
$$

with $\theta := \cos^{-1}\left(\frac{\|\mathcal{A}x_0\|_\diamond}{L\|\mathcal{A}\|\|x_0\|}\right)$.

Appendix C. Proofs for Section 4

C.1 Proof of Lemma 15.

Proof (1) By the definition of $\mathcal{X}_{[j]}$, it is sufficient to prove that the vectorization of the right hand side of (4.3) equals $\text{vec}(\mathcal{X}_{(1)})$.

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

$$
\begin{aligned}
\text{vec}(\mathcal{X}_{(1)}) &= \text{vec}\Big(\sum_{i=1}^{r} \lambda_i a_i^{(1)} \circ (a_i^{(K)} \otimes a_i^{(K-1)} \otimes \cdots \otimes a_i^{(2)}) \Big) \qquad (62) \\
&= \sum_{i=1}^{r} \lambda_i \text{vec}\Big(a_i^{(1)} \circ (a_i^{(K)} \otimes a_i^{(K-1)} \otimes \cdots \otimes a_i^{(2)}) \Big) \\
&= \sum_{i=1}^{r} \lambda_i (a_i^{(K)} \otimes a_i^{(K-1)} \otimes \cdots \otimes a_i^{(2)} \otimes a_i^{(1)}),
\end{aligned}
$$

where the last equality follows from the fact that $\text{vec}(a \circ b) = b \otimes a$. Similarly, we can derive that the vectorization of the right hand side of (4.3),

$$
\begin{aligned}
&\text{vec}\Big(\sum_{i=1}^{r} \lambda_i (a_i^{(j)} \otimes a_i^{(j-1)} \otimes \cdots \otimes a_i^{(1)}) \circ (a_i^{(K)} \otimes a_i^{(K-1)} \cdots \otimes a_i^{(j+1)}) \Big) \qquad (63) \\
&= \sum_{i=1}^{r} \lambda_i \text{vec}\Big((a_i^{(j)} \otimes a_i^{(j-1)} \otimes \cdots \otimes a_i^{(1)}) \circ (a_i^{(K)} \otimes a_i^{(K-1)} \cdots \otimes a_i^{(j+1)}) \Big) \\
&= \sum_{i=1}^{r} \lambda_i (a_i^{(K)} \otimes a_i^{(K-1)} \otimes \cdots \otimes a_i^{(2)} \otimes a_i^{(1)}) \\
&= \text{vec}(\mathcal{X}_{(1)}).
\end{aligned}
$$

Thus, equation (4.3) is valid.

(2) The above argument can be easily adapted to prove the second claim. Since $\mathcal{X} = \mathcal{C} \times_1 U_1 \times_2 U_2 \times_3 \cdots \times_K U_K$, we have

$$
\begin{aligned}
\text{vec}(\mathcal{X}_{(1)}) &= \text{vec}\Big(U_1 \mathcal{C}_{(1)} (U_K \otimes U_{K-1} \otimes \cdots \otimes U_2)^* \Big) \qquad (64) \\
&= (U_K \otimes U_{K-1} \otimes \cdots \otimes U_1) \text{vec}(\mathcal{C}_{(1)}),
\end{aligned}
$$

where the last equality follows from the fact that $\text{vec}(ABC) = (C^* \otimes A)\text{vec}(B)$. Similarly, we can derive that the vectorization of the right hand side of (4.4),

$$
\begin{aligned}
&\text{vec}\Big((U_j \otimes U_{j-1} \otimes \cdots \otimes U_1) \mathcal{C}_{[j]} (U_K \otimes U_{K-1} \otimes \cdots \otimes U_{j+1})^* \Big) \qquad (65) \\
&= (U_K \otimes U_{K-1} \otimes \cdots \otimes U_1) \text{vec}(\mathcal{C}_{[j]}) \\
&= (U_K \otimes U_{K-1} \otimes \cdots \otimes U_1) \text{vec}(\mathcal{C}_{(1)}) \\
&= \text{vec}(\mathcal{X}_{(1)}).
\end{aligned}
$$

Thus, equation (4.4) is valid. ∎

Appendix D. Tensor denoising

A classical problem in statistical inference is to estimate the target signal with Gaussian perturbed observations. Here, we briefly discuss this denoising problem under the context of low-rank tensors.

In specific, the target signal is a low-rank tensor (in terms of Tucker rank), say $\mathcal{X}_0 \in \Upsilon_r$, and we observe $\mathcal{Z}_0 = \mathcal{X}_0 + \sigma \mathcal{G}$, where $\text{vec}(\mathcal{G})$ is a standard norm vector and σ is an unknown standard deviation parameter. To estimate $\mathcal{X}_0$, a natural way is to solve the following convex optimization problem[6]

$$\widehat{\mathcal{X}}_\tau \quad := \quad \arg\min_{\mathcal{X}} \|\mathcal{Z}_0 - \mathcal{X}\|_F \quad \text{s.t.} \quad f(\mathcal{X}_0) \le \tau, \tag{67}$$

where f is a convex function promoting the low-rank tensor structure, and $\tau > 0$ balances the structural penalty and the date fidelity term.

One way to evaluate the denoising performance of this convex regularizer f is to measure the minimax normalized mean-squared-error (NMSE) risks, defined as

$$R_{mm}(f, f(\mathcal{X}_0)) \quad := \quad \sup_{\mathcal{X}_0 \in \Upsilon_r, \sigma > 0} \frac{1}{\sigma^2} \mathbb{E}\left[\left\|\widehat{\mathcal{X}}_{f(\mathcal{X}_0)} - \mathcal{X}_0\right\|_F^2\right], \tag{68}$$

$$R_{mm}(f) \quad := \quad \sup_{\mathcal{X}_0 \in \Upsilon_r, \sigma > 0} \inf_{\tau > 0} \frac{1}{\sigma^2} \mathbb{E}\left[\left\|\widehat{\mathcal{X}}_\tau - \mathcal{X}_0\right\|_F^2\right], \tag{69}$$

i.e. the risk corresponds to the normalized mean-squared error (NMSE) for either the fixed oracle value $\tau = f(\mathcal{X}_0)$ or the best tuned τ, at worst choices of the underlying signal $\mathcal{X}_0$ and the noise level σ. Due to the general result proved by Oymak and Hassibi (2013, Theorem 3.1), quantities in (68) and (69) are closely related with statistical dimension,

$$R_{mm}(f, f(\mathcal{X}_0)) = \sup_{\mathcal{X}_0 \in \Upsilon_r} \delta(\mathcal{C}(f, \mathcal{X}_0)) \quad \text{and} \quad R_{mm}(f) = \sup_{\mathcal{X}_0 \in \Upsilon_r} \delta(\mathcal{C}(f, \mathcal{X}_0)) - O(n^{K/2}), \tag{70}$$

where we recall that $\mathcal{C}(f, \mathcal{X}_0)$ denotes the descent cone of f at the point $\mathcal{X}_0$. Based on this result, we can easily characterize the Minimax MSE risks of several convex functions f discussed in the paper (see Table 2). [7]

To empirically verify the results in Table 2, we construct $\mathcal{X}_0$ as a 4-mode $n \times n \times n \times n$ (super)diagonal tensor with only the first r diagonal entries as 1 and 0 elsewhere, and choose

---

6. Problem (67) is equivalent to its Lagrangian formulation

$$\min_{\mathcal{X}} \quad \frac{1}{2} \|\mathcal{Z}_0 - \mathcal{X}\|_F^2 + \lambda f(\mathcal{X}_0) \tag{66}$$

with a proper choice of $\lambda \ge 0$.

7. For the SNN model ($f = \sum_{i \in [K]} \lambda_i \|\mathcal{X}_{(i)}\|_*$), we also have λ_i's to choose, and so $R_{mm}(f, f(\mathcal{X}_0))$ and $R_{mm}(f)$ are instead naturally defined as

$$R_{mm}(f, f(\mathcal{X}_0)) \quad := \quad \sup_{\mathcal{X}_0 \in \Upsilon_r, \sigma > 0} \inf_{\lambda > 0} \frac{1}{\sigma^2} \mathbb{E}\left[\left\|\widehat{\mathcal{X}}_{f(\mathcal{X}_0)} - \mathcal{X}_0\right\|_F^2\right], \tag{71}$$

$$R_{mm}(f) \quad := \quad \sup_{\mathcal{X}_0 \in \Upsilon_r, \sigma > 0} \inf_{\tau > 0, \lambda > 0} \frac{1}{\sigma^2} \mathbb{E}\left[\left\|\widehat{\mathcal{X}}_\tau - \mathcal{X}_0\right\|_F^2\right]. \tag{72}$$

Table 2: **Minimax NMSE risks of different convex regularizers for the low-rank tensor estimation.** Note that the risks for the Single Norm model and the SNN model are essentially on the same order, which is substantially higher than the one for the Square model. This can be viewed as a dual phenomenon of our results (Theorem 7 and Theorem 17) regarding the exact low-rank tensor recovery using generic measurements. Both of these two results arise from the study on the statistical dimension of the descent cone of certain convex function f at the target signal x_0.

Model	Convex regularizer $f(\cdot)$	$R_{mm,f(\mathcal{X}_0)}(f)$	$R_{mm}(f)$
Single Norm	$\left\|\mathcal{X}_{(1)}\right\|_*$	$\Theta(rn^{K-1})$	$\Theta(rn^{K-1})$
SNN	$\sum_{i\in[K]} \lambda_i \left\|\mathcal{X}_{(i)}\right\|_*$	$\Theta(rn^{K-1})$	$\Theta(rn^{K-1})$
Square	$\|\mathcal{X}_\square\|_*$	$\Theta(r^{\lfloor\frac{K}{2}\rfloor}n^{\lceil\frac{K}{2}\rceil})$	$\Theta(r^{\lfloor\frac{K}{2}\rfloor}n^{\lceil\frac{K}{2}\rceil})$

$\sigma = 10^{-8}$. Convex regularizers $f(\cdot)$ listed in Table 2: $\left\|\mathcal{X}_{(1)}\right\|_*$, $\sum_{i\in[K]}\left\|\mathcal{X}_{(i)}\right\|_*$, and $\|\mathcal{X}_\square\|_*$, are respectively tested. For different pairs (r,n), we compute the empirical NMSE by averaging $\frac{1}{\sigma^2}\left\|\widehat{\mathcal{X}}_{f(\mathcal{X}_0)} - \mathcal{X}_0\right\|_F^2$ over 10 repeats. Curves are fitted based on the complexities displayed in Table 2. It can be clearly observed that the obtained curves fit the empirical NMSE quite tightly.
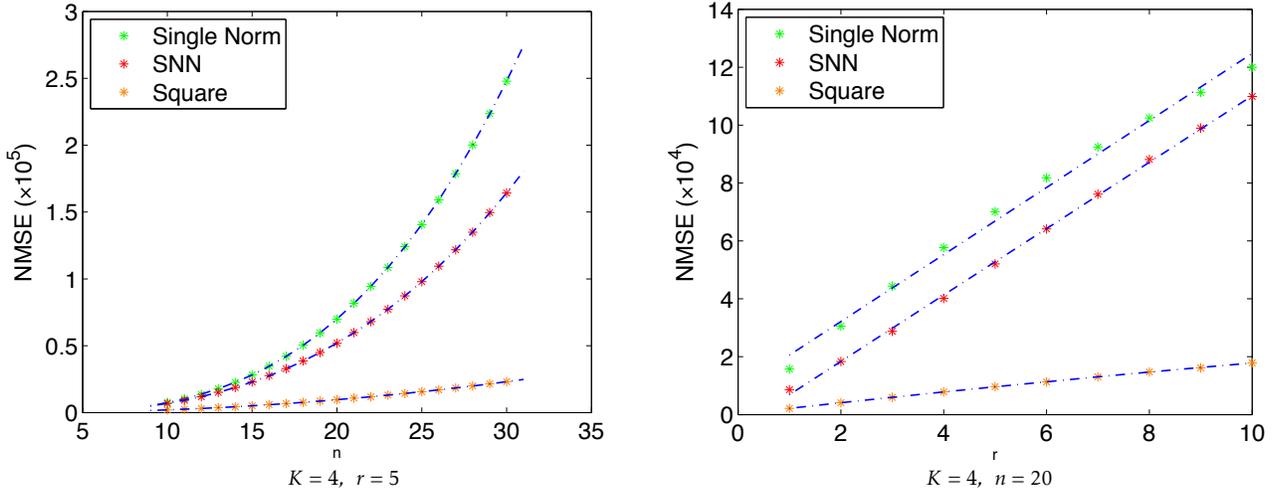
Figure 4: **Tensor denoising.** Each cross corresponds to the empirical estimate of NMSE for a (r,n)-pair. Different colors are used to indicate different convex models. The blue dashed lines are fitted using polynomials consistent with the complexities displayed in Table 2.

Appendix E. Algorithms

Here, we briefly describe the first-order methods we designed for tensor optimization problems involved in the paper. [8]

E.1 Dykstra's Algorithm for (30)

By splitting $\mathcal{X}$ into $\{\mathcal{X}_i\}_{i \in [K]}$, problem (30) can be reformulated as

$$
\min_{\{\mathcal{X}_i\}_{i \in [K]}} \quad \sum_{i \in [K]} \left\| \mathcal{Z}_0 - \mathcal{X}_{(i)} \right\|_F^2
$$

$$
\text{s.t.} \quad \sum_{i \in [K]} \left\| (\mathcal{X}_i)_{(i)} \right\|_* \leq \tau := Kr \tag{73}
$$

$$
\mathcal{X}_1 = \mathcal{X}_2 = \cdots = \mathcal{X}_K \in \mathbb{R}^{n \times n \times \cdots \times n}.
$$

This is essentially to compute $\mathcal{P}_{\mathcal{C}_1 \cap \mathcal{C}_2}[z_0]$: the projection of $z_0 := \underbrace{(\mathcal{Z}_0, \mathcal{Z}_0, \cdots, \mathcal{Z}_0)}_{K}$ onto the intersection of two closed convex sets $\mathcal{C}_1$ and $\mathcal{C}_2$ in the Hilbert space $\mathcal{H}$, where $\mathcal{H} := \bigtimes_{i \in [K]} \mathbb{R}^{n \times n \times \cdots \times n}$ and

$$
\mathcal{C}_1 := \left\{ (\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_K) \in \mathcal{H} \;\middle|\; \sum_{i \in [K]} \left\| (\mathcal{X}_i)_{(i)} \right\|_* \leq \tau \right\}, \tag{74}
$$

$$
\mathcal{C}_2 := \{ (\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_K) \in \mathcal{H} \mid \mathcal{X}_1 = \mathcal{X}_2 = \cdots = \mathcal{X}_K \}. \tag{75}
$$

As both $\mathcal{P}_{\mathcal{C}_1}[\cdot]$ and $\mathcal{P}_{\mathcal{C}_2}[\cdot]$ have closed form solutions that can be easily computed, we apply *Dykstra's algorithm* (see Bauschke and Combettes, 2011, Chap. 29.1) to tackle problem (30).

Algorithm 1 Dykstra's algorithm for problem (30)

1: **Initialization:** $z^{(0)} \leftarrow (\mathcal{Z}_0, \mathcal{Z}_0, \cdots, \mathcal{Z}_0) \in \mathcal{H}$, $q^{(-1)} \leftarrow 0 \in \mathcal{H}$, $q^{(0)} \leftarrow 0 \in \mathcal{H}$;
2: **for** $n \leftarrow 1, 2, \ldots$ **do**
3:     **if** $2 \mid n$ **then**
4:         $z^{(n)} \leftarrow \mathcal{P}_{\mathcal{C}_2}[z^{(n-1)} + q^{(n-2)}]$;
5:         $q^{(n)} \leftarrow z^{(n-1)} + q^{(n-2)} - z^{(n)}$;
6:     **else**
7:         $z^{(n)} \leftarrow \mathcal{P}_{\mathcal{C}_1}[z^{(n-1)} + q^{(n-2)}]$;
8:         $q^{(n)} \leftarrow z^{(n-1)} + q^{(n-2)} - z^{(n)}$;
9:     **end if**
10: **end for**

For the sequence $\{z^{(n)}\}$ generated by Algorithm 1, its convergence to the optimal solution of problem (30) follows directly from Theorem 29.2 of the book by Bauschke and Combettes (2011).

---

8. MATLAB codes are available on CM's personal website: `https://sites.google.com/site/mucun1988/`. MATLAB Tensor Toolbox (Bader and Kolda, 2015) has been utilized in our implementation.

E.2 Douglas-Rachford Algorithm for problem (38)

By splitting $\boldsymbol{\mathcal{X}}$ into $\{\boldsymbol{\mathcal{X}}_i\}_{i\in[K+1]}$, problem (38) can be reformulated as

$$
\begin{aligned}
\min_{\{\boldsymbol{\mathcal{X}}_i\}_{i\in[K+1]}} \quad & \sum_{i\in[K]} \left\| (\boldsymbol{\mathcal{X}}_i)_{(i)} \right\|_* \\
\text{s.t.} \quad & \mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_{K+1}] = \boldsymbol{\mathcal{M}} \\
& \boldsymbol{\mathcal{X}}_1 = \boldsymbol{\mathcal{X}}_2 = \cdots = \boldsymbol{\mathcal{X}}_{K+1} \in \mathbb{R}^{n\times n\times\cdots\times n}.
\end{aligned}
\tag{76}
$$

If we denote $\boldsymbol{x} := (\boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_2, \cdots, \boldsymbol{\mathcal{X}}_{K+1}) \in \mathcal{H} := \bigtimes_{i\in[K+1]} \mathbb{R}^{n\times n\times\cdots\times n}$, then problem (76) can be compactly expressed as

$$
\min_{\boldsymbol{x}\in\mathcal{H}} \quad F(\boldsymbol{x}) + G(\boldsymbol{x}),
\tag{77}
$$

where

$$
F(\boldsymbol{x}) := \sum_{i\in[K]} \left\| (\boldsymbol{\mathcal{X}}_i)_{(i)} \right\|_* + \mathbf{1}_{\{\mathcal{P}_\Omega[\boldsymbol{\mathcal{X}}_{K+1}]=\boldsymbol{\mathcal{M}}\}},
\tag{78}
$$

$$
G(\boldsymbol{x}) := \mathbf{1}_{\{\boldsymbol{\mathcal{X}}_1=\boldsymbol{\mathcal{X}}_2=\cdots=\boldsymbol{\mathcal{X}}_{K+1}\}}, \quad \text{and}
\tag{79}
$$

here the indicator function for a set C, $\mathbf{1}_C(\boldsymbol{x})$, equals 0 if $\boldsymbol{x} \in C$ and $+\infty$ otherwise. Note that the proximity operators of F and G, i.e.

$$
\text{prox}_F(\boldsymbol{x}) := \arg\min_{\boldsymbol{y}\in\mathcal{H}} \ F(\boldsymbol{y}) + \frac{1}{2}\left\| \boldsymbol{x}-\boldsymbol{y} \right\|^2 \quad \text{and}
\tag{80}
$$

$$
\text{prox}_G(\boldsymbol{x}) := \arg\min_{\boldsymbol{y}\in\mathcal{H}} \ G(\boldsymbol{y}) + \frac{1}{2}\left\| \boldsymbol{x}-\boldsymbol{y} \right\|^2
\tag{81}
$$

can be both easily computed. Therefore, it is quite suitable to apply the *Douglas-Rachford algorithm* here (see Combettes and Pesquet, 2011, for more details).

Algorithm 2 Douglas-Rachford algorithm for problem (38)

1: **Initialization:** $\boldsymbol{x}^{(0)} \leftarrow \boldsymbol{0} \in \mathcal{H}$;
2: **for** $n \leftarrow 0, 1, 2, \ldots$ **do**
3:    $\boldsymbol{y}^{(n)} \leftarrow \text{prox}_G(\boldsymbol{x}^{(n)})$;
4:    $\boldsymbol{x}^{(n+1)} \leftarrow \text{prox}_F(2\boldsymbol{y}^{(n)} - \boldsymbol{x}^{(n)}) + \boldsymbol{x}^{(n)} - \boldsymbol{y}^{(n)}$;
5: **end for**
6: **Output** $\text{prox}_G(\boldsymbol{x}^{(n+1)})$;