ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746705

ACCELERATING ILL-CONDITIONED ROBUST LOW-RANK TENSOR REGRESSION

Tian Tong[†], *Cong Ma*[‡], *Yuejie Chi*[†]

[†]Department of Electrical and Computer Engineering, Carnegie Mellon University, USA
[‡]Department of Statistics, University of Chicago, USA
{ttong1,yuejiec}@andrew.cmu.edu; congm@uchicago.edu

ABSTRACT

An important problem that arises across different applications in signal processing, machine learning, and data science is to reliably estimate a tensor from a small number of measurements that are possibly corrupted. Leveraging the low-rank structure under the Tucker decomposition, we propose a provably efficient algorithm that directly estimates the tensor factors by solving a nonsmooth and nonconvex composite optimization problem that minimizes the least absolute deviation loss. The proposed algorithm—built on subgradient methods—harnesses preconditioners that are designed to be equivariant w.r.t. the low-rank parameterization, and is shown to achieve local linear convergence at a constant rate under the Gaussian design. Numerical experiments are provided to corroborate the superior performance of the proposed algorithm.

Index Terms— robust low-rank tensor regression, nonconvex composite optimization, scaled subgradient method

1. INTRODUCTION

The modern data deluge has created a growing number of applications involving multi-dimensional or multi-attribute datasets, examples including video surveillance, hyperspectral imaging, neuroimaging, social network analysis, and so on. Tensors arise naturally as a suitable data structure that captures the underlying multiway interactions, offering advantages over the matrix counterpart [1,2]. An important problem, known as tensor regression, that arises frequently across different applications is to recover a tensor from a small number of its linear measurements, given by

$$\boldsymbol{y} \approx \mathcal{A}(\boldsymbol{\mathcal{X}}_\star),$$

where $\boldsymbol{\mathcal{X}}_\star \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_K}$ is a K-way tensor, $\boldsymbol{y} \in \mathbb{R}^m$ is the collected measurements, and $\mathcal{A}(\cdot)$ is a linear map that models the data collection process. For ease of presentation, we consider the case $K = 3$ throughout the paper, while our results hold for the general case without difficulty.

Practical constraints such as sensing budgets or physical limitations often lead to a highly ill-posed problem, where the number of measurements is much smaller than the ambient dimension of the tensor, i.e. $m \ll \prod_{k=1}^3 n_k$. Fortunately, many real-world datasets possess low-dimensional structures, where the corresponding tensor can be appropriately decomposed into a small number of factors using a drastically reduced number of parameters. In this paper, we focus on the Tucker decomposition, where $\boldsymbol{\mathcal{X}}_\star$ admits the following decomposition with Tucker rank or multilinear rank $\boldsymbol{r} = (r_1, r_2, r_3)$, and $r_i \ll n_i$:

$$\boldsymbol{\mathcal{X}}_\star = \boldsymbol{\mathcal{C}}_\star \times_1 \boldsymbol{U}_\star \times_2 \boldsymbol{V}_\star \times_3 \boldsymbol{W}_\star \coloneqq (\boldsymbol{U}_\star, \boldsymbol{V}_\star, \boldsymbol{W}_\star) \cdot \boldsymbol{\mathcal{C}}_\star, \quad (1)$$

where $\boldsymbol{\mathcal{C}}_\star \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the core tensor, $\boldsymbol{U}_\star \in \mathbb{R}^{n_1 \times r_1}$, $\boldsymbol{V}_\star \in \mathbb{R}^{n_2 \times r_2}$, $\boldsymbol{W}_\star \in \mathbb{R}^{n_3 \times r_3}$ are orthonormal matrices corresponding to the factors of each mode, and $\times_k$ denotes the tensor-matrix product along mode k. Many provable algorithms have been proposed to recover the low-rank tensor $\boldsymbol{\mathcal{X}}_\star$ from $\boldsymbol{y}$ both in statistically and computationally efficient manners, e.g. [3–9].

To reduce the storage and computational complexities, a popular approach is to take advantage of the low-rank factorization and optimize the factors directly by solving

$$\min_{\boldsymbol{F}=(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{\mathcal{C}})} \|\mathcal{A}(\boldsymbol{\mathcal{X}}) - \boldsymbol{y}\|_2^2, \text{ where } \boldsymbol{\mathcal{X}} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{C}}, \quad (2)$$

with the optimization variables $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r_1}$, $\boldsymbol{V} \in \mathbb{R}^{n_2 \times r_2}$, $\boldsymbol{W} \in \mathbb{R}^{n_3 \times r_3}$ and $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$.

1.1. Our contributions: robust low-rank tensor regression

In practice, due to sensor failures and malicious attacks, it is common that the collected measurements may suffer from undesirable and unknown corruptions, which are possibly adversarial. Consequently, there is an imminent need to develop low-rank tensor recovery algorithms that are provably robust and efficient, which are still lacking. To fill the gap, instead of minimizing the smooth loss function in (2), which is known to be vulnerable to outliers, we resort to the least absolute deviations (LAD) loss, which measures the residual sum of absolute errors:

$$\min_{\boldsymbol{F}=(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{\mathcal{C}})} \|\mathcal{A}(\boldsymbol{\mathcal{X}}) - \boldsymbol{y}\|_1, \text{ where } \boldsymbol{\mathcal{X}} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \cdot \boldsymbol{\mathcal{C}}. \quad (3)$$

Leveraging recent insights in preconditioning for ill-conditioned low-rank matrix and tensor estimation [9–11], this paper proposes an efficient algorithm for solving the nonconvex composite optimization problem in (3), namely the scaled subgradient method (ScaledSM), which incorporates carefully-designed preconditioners in the local updates to preserve the equivariance of the low-rank parameterization. Under the Gaussian design, the proposed method provably finds the ground truth at a constant linear rate that is *independent of the condition number* even under a constant fraction of outliers, as long as it is initialized properly. The algorithm is much more scalable than its counterpart without the preconditioners, especially when the ground truth tensor is ill-conditioned. To the best of our knowledge, our work provides the first provable algorithm that achieves robust low-rank tensor regression from corrupted measurements, together with a fast rate of convergence independent of the condition number of the ground truth tensor.

The work of T. Tong and Y. Chi is supported in part by Office of Naval Research under N00014-19-1-2404, by Air Force Research Laboratory under FA8750-20-2-0504, and by National Science Foundation under CAREER ECCS-1818571, CCF-1901199 and ECCS-2126634.

1.2. Related works

Low-rank tensor recovery has attracted significant research interest in recent years, where many algorithms have been developed with provable performance guarantees, e.g. [3–7, 9, 12–15]. Moreover, spectral methods [16–18] are often applied to provide a smart initialization from which iterative algorithms refine locally to enable global convergence despite the presence of nonconvexity. However, a majority of these algorithms are designed with respect to the smooth least-squares loss and therefore their performance is very sensitive to the existence of outliers.

Motivated by the success of robust principal component analysis for the matrix setting [19], convex relaxation approaches are proposed in [3, 20, 21] via unfolding the tensor of interest and invoking matrix-based algorithms. However, their computational complexity is often prohibitive for large-scale problems. On the other end, the LAD loss is not new to handle outliers, and has been adopted for high-dimensional signal recovery [11, 22–27], where the subgradient method has been analyzed in [23, 26–28]. Another popular strategy is to adaptively truncate or prune outliers in an iterative manner guided by quantile statistics, as done in [29–32].

The preconditioner in our approach is directly inspired by [9], which proposed a scaled gradient descent (ScaledGD) method to optimize the smooth loss function (2) for low-rank tensor regression. In particular, the proposed subgradient method can be viewed as the tensor counterpart of [11], which generalizes the preconditioner designs to the nonsmooth setting.

1.3. Paper organization and notation

The rest of this paper is organized as follows. Section 2 describes the problem formulation as well as the proposed algorithm. Section 3 provides the theoretical guarantees in terms of local linear convergence. Numerical experiments are illustrated in Section 4, and finally, we conclude in Section 5.

Notation. Throughout this paper, boldface calligraphic letters (e.g. $\boldsymbol{\mathcal{A}}$) denote tensors, and boldface capitalized letters (e.g. $\boldsymbol{A}$) denote matrices. $\sigma_i(\boldsymbol{A})$ denotes its i-th largest singular value, and $\|\boldsymbol{A}\|_{\mathsf{F}}$, $\|\boldsymbol{A}\|$, and $\|\boldsymbol{A}\|_{\infty}$ denotes the Frobenius norm, the spectral norm, and the entrywise ℓ_∞ norm of a matrix $\boldsymbol{A}$, respectively. The set of invertible matrices in $\mathbb{R}^{r\times r}$ is denoted by $\mathrm{GL}(r)$. Let $\otimes$ denote the Kronecker product, and $\mathrm{sign}(\boldsymbol{x})$ denote the vector containing the signs of the entries of $\boldsymbol{x}$.

Given a tensor $\boldsymbol{\mathcal{X}} \coloneqq [\boldsymbol{\mathcal{X}}(i_1,i_2,i_3)] \in \mathbb{R}^{n_1\times n_2\times n_3}$, its mode-1 matricization $\mathcal{M}_1(\boldsymbol{\mathcal{X}}) \in \mathbb{R}^{n_1\times(n_2n_3)}$ is defined by

$$[\mathcal{M}_1(\boldsymbol{\mathcal{X}})]\big(i_1, i_2+(i_3-1)n_2\big) = \boldsymbol{\mathcal{X}}(i_1,i_2,i_3),$$

for $1\le i_k\le n_k$, $k=1,2,3$; $\mathcal{M}_2(\boldsymbol{\mathcal{X}})$ and $\mathcal{M}_3(\boldsymbol{\mathcal{X}})$ can be defined similarly. The inner product between two tensors is defined as $\langle \boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_2\rangle = \sum_{i_1,i_2,i_3} \boldsymbol{\mathcal{X}}_1(i_1,i_2,i_3)\boldsymbol{\mathcal{X}}_2(i_1,i_2,i_3)$. The Frobenius norm of $\boldsymbol{\mathcal{X}}$ is then given by $\|\boldsymbol{\mathcal{X}}\|_{\mathsf{F}} = \sqrt{\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}\rangle}$.

2. FORMULATION AND PROPOSED ALGORITHM

Let $\boldsymbol{\mathcal{X}}_\star \coloneqq [\boldsymbol{\mathcal{X}}_\star(i_1,i_2,i_3)] \in \mathbb{R}^{n_1\times n_2\times n_3}$ be the ground truth tensor that satisfies the Tucker decomposition in (1), which equivalently means that for $k=1,2,3$, and $1\le i_k\le n_k$,

$$\boldsymbol{\mathcal{X}}_\star(i_1,i_2,i_3) = \sum_{j_1=1}^{r_1}\sum_{j_2=1}^{r_2}\sum_{j_3=1}^{r_3} \boldsymbol{U}_\star(i_1,j_1)\boldsymbol{V}_\star(i_2,j_2)\boldsymbol{W}_\star(i_3,j_3)\boldsymbol{\mathcal{C}}_\star(j_1,j_2,j_3).$$

Consider the robust low-rank tensor regression problem, in which the measurements are corrupted by sparse outliers. Specifically, assume that we have access to a set of linear observations of $\boldsymbol{\mathcal{X}}_\star$, where the measurement vector $\boldsymbol{y} = \{y_i\}_{i=1}^m$ is given as

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{\mathcal{X}}_\star) + \boldsymbol{s}, \tag{4}$$

where $\mathcal{A}(\boldsymbol{\mathcal{X}}_\star) = \{\langle \boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{X}}_\star\rangle\}_{i=1}^m$ is the measurement operator, with $\boldsymbol{\mathcal{A}}_i \in \mathbb{R}^{n_1\times n_2\times n_3}$ denoting the i-th sensing tensor, and $\boldsymbol{s} = \{s_i\}_{i=1}^m$ corresponds to the outlier vector. We assume the outlier $\boldsymbol{s}$ is a sparse vector obeying $\|\boldsymbol{s}\|_0 = p_s m$ for some $0\le p_s\le 1$, which means that $\|\boldsymbol{s}\|_0$ is much smaller than its ambient dimension m, so that only a small fraction p_s of the measurements are corrupted. However, the corrupted entries can take arbitrary or adversarial magnitudes. The goal is to recover the low-rank tensor $\boldsymbol{\mathcal{X}}_\star$ from $\boldsymbol{y}$ in a robust and scalable manner.

To cope with the outliers, it is natural to minimize the least absolute deviation (LAD) loss of the measurements, given by

$$f(\boldsymbol{\mathcal{X}}) \coloneqq \|\mathcal{A}(\boldsymbol{\mathcal{X}}) - \boldsymbol{y}\|_1 = \sum_{i=1}^m \left|\langle \boldsymbol{\mathcal{A}}_i, \boldsymbol{\mathcal{X}}\rangle - \boldsymbol{y}_i\right|. \tag{5}$$

In addition, to take advantage of the low-rank structure and minimize complexity, we factorize the tensor $\boldsymbol{\mathcal{X}} = (\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{C}}$ with $\boldsymbol{U}\in\mathbb{R}^{n_1\times r_1}$, $\boldsymbol{V}\in\mathbb{R}^{n_2\times r_2}$, $\boldsymbol{W}\in\mathbb{R}^{n_3\times r_3}$ and $\boldsymbol{\mathcal{C}}\in\mathbb{R}^{r_1\times r_2\times r_3}$, and optimize the factors directly via the following unconstrained composite optimization problem:

$$\min_{\boldsymbol{F}=(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{\mathcal{C}})} \mathcal{L}(\boldsymbol{F}) \coloneqq f\big((\boldsymbol{U},\boldsymbol{V},\boldsymbol{W})\cdot\boldsymbol{\mathcal{C}}\big), \tag{6}$$

which is nonconvex and nonsmooth.

2.1. Proposed scaled subgradient method

A natural idea to optimize (6) is via subgradient descent, which updates the factor quadruple iteratively according to

$$\begin{aligned}
\boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \eta_t \mathcal{M}_1(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{U}}_t,\\
\boldsymbol{V}_{t+1} &= \boldsymbol{V}_t - \eta_t \mathcal{M}_2(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{V}}_t,\\
\boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \eta_t \mathcal{M}_3(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{W}}_t,\\
\boldsymbol{\mathcal{C}}_{t+1} &= \boldsymbol{\mathcal{C}}_t - \eta_t \Big(\boldsymbol{U}_t^\top, \boldsymbol{V}_t^\top, \boldsymbol{W}_t^\top\Big)\cdot\boldsymbol{\mathcal{G}}_t.
\end{aligned} \tag{7}$$

where $\eta_t>0$ is the step size, $\boldsymbol{\mathcal{G}}_t = \mathcal{A}^*(\mathrm{sign}(\mathcal{A}(\boldsymbol{\mathcal{X}}_t)) - \boldsymbol{y}) \in \partial_{\boldsymbol{\mathcal{X}}} f(\boldsymbol{\mathcal{X}}_t)$ is a subgradient of $f(\boldsymbol{\mathcal{X}})$ with respect to $\boldsymbol{\mathcal{X}}$ at $\boldsymbol{\mathcal{X}}_t = (\boldsymbol{U}_t,\boldsymbol{V}_t,\boldsymbol{W}_t)\cdot\boldsymbol{\mathcal{C}}_t$, and $\mathcal{A}^*(\cdot)$ is the adjoint operator of $\mathcal{A}(\cdot)$. Furthermore, the following short-hand notation is introduced:

$$\breve{\boldsymbol{U}}_t \coloneqq (\boldsymbol{W}_t\otimes\boldsymbol{V}_t)\mathcal{M}_1(\boldsymbol{\mathcal{C}}_t)^\top, \tag{8a}$$

$$\breve{\boldsymbol{V}}_t \coloneqq (\boldsymbol{W}_t\otimes\boldsymbol{U}_t)\mathcal{M}_2(\boldsymbol{\mathcal{C}}_t)^\top, \tag{8b}$$

$$\breve{\boldsymbol{W}}_t \coloneqq (\boldsymbol{V}_t\otimes\boldsymbol{U}_t)\mathcal{M}_3(\boldsymbol{\mathcal{C}}_t)^\top. \tag{8c}$$

While simple and straightforward, this approach tends to converge very slowly when the tensor is ill-conditioned. Inspired by [9], we propose to update the iterate along a preconditioned or scaled direction of the subgradient, leading to the following scaled subgradient

method (ScaledSM):

$$\begin{aligned}
\boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \eta_t \mathcal{M}_1(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{U}}_t(\breve{\boldsymbol{U}}_t^\top \breve{\boldsymbol{U}}_t)^{-1},\\
\boldsymbol{V}_{t+1} &= \boldsymbol{V}_t - \eta_t \mathcal{M}_2(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{V}}_t(\breve{\boldsymbol{V}}_t^\top \breve{\boldsymbol{V}}_t)^{-1},\\
\boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \eta_t \mathcal{M}_3(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{W}}_t(\breve{\boldsymbol{W}}_t^\top \breve{\boldsymbol{W}}_t)^{-1},\\
\boldsymbol{\mathcal{C}}_{t+1} &= \boldsymbol{\mathcal{C}}_t - \eta_t \Big((\boldsymbol{U}_t^\top \boldsymbol{U}_t)^{-1}\boldsymbol{U}_t^\top,\\
&\qquad (\boldsymbol{V}_t^\top \boldsymbol{V}_t)^{-1}\boldsymbol{V}_t^\top, (\boldsymbol{W}_t^\top \boldsymbol{W}_t)^{-1}\boldsymbol{W}_t^\top\Big)\cdot \boldsymbol{\mathcal{G}}_t.
\end{aligned} \tag{9}$$

Step size schedules. We still need to specify the choice of the step size $\eta_t > 0$, which needs to be carefully scheduled in accordance with the scaled update. Specifically, we apply a geometrically decaying learning rate schedule [33] with proper scaling,

$$\eta_t := \frac{\lambda q^t}{N_t}, \tag{10}$$

where $q \in (0,1)$, $\lambda > 0$ and

$$\begin{aligned}
N_t^2 := &\left\|\mathcal{M}_1(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{U}}_t(\breve{\boldsymbol{U}}_t^\top \breve{\boldsymbol{U}}_t)^{-1/2}\right\|_{\mathsf{F}}^2 + \left\|\mathcal{M}_2(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{V}}_t(\breve{\boldsymbol{V}}_t^\top \breve{\boldsymbol{V}}_t)^{-1/2}\right\|_{\mathsf{F}}^2\\
&+ \left\|\mathcal{M}_3(\boldsymbol{\mathcal{G}}_t)\breve{\boldsymbol{W}}_t(\breve{\boldsymbol{W}}_t^\top \breve{\boldsymbol{W}}_t)^{-1/2}\right\|_{\mathsf{F}}^2 + \left\|\Big((\boldsymbol{U}_t^\top \boldsymbol{U}_t)^{-1/2}\boldsymbol{U}_t^\top,\right.\\
&\left.(\boldsymbol{V}_t^\top \boldsymbol{V}_t)^{-1/2}\boldsymbol{V}_t^\top, (\boldsymbol{W}_t^\top \boldsymbol{W}_t)^{-1/2}\boldsymbol{W}_t^\top\Big)\cdot \boldsymbol{\mathcal{G}}_t\right\|_{\mathsf{F}}^2.
\end{aligned} \tag{11}$$

In fact, N_t can be viewed as the norm of the subgradient under a scaled metric compatible with our preconditioners. This choice is informed by our theory.

Remark 1. Ideally, one might be tempted to apply the Polyak's step size, given by $\eta_t := \frac{f(\boldsymbol{\mathcal{X}}_t)-f(\boldsymbol{\mathcal{X}}_\star)}{N_t^2}$. However, it is impractical due to the unknown optimal function value $f(\boldsymbol{\mathcal{X}}_\star)$. As illustrated in [11], geometric step size achieves the same performance as Polyak's step size when parameters λ, q are tuned appropriately.

Equivariance to low-rank parameterization. A crucial property of ScaledSM is that the update of the low-rank tensor $\boldsymbol{\mathcal{X}}_t$ is invariant w.r.t. the low-rank parameterization. Suppose that at the t-th iteration, we reparameterize the factor $\boldsymbol{F}_t = (\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t, \boldsymbol{\mathcal{C}}_t)$ by

$$\widetilde{\boldsymbol{F}}_t = (\boldsymbol{U}_t\boldsymbol{Q}_1, \boldsymbol{V}_t\boldsymbol{Q}_2, \boldsymbol{W}_t\boldsymbol{Q}_3, (\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1})\cdot \boldsymbol{\mathcal{C}}_t)$$

via any invertible matrices $\boldsymbol{Q}_k \in \mathrm{GL}(r_k)$, $k = 1,2,3$, where both $\boldsymbol{F}_t$ and $\widetilde{\boldsymbol{F}}_t$ correspond to the same low-rank tensor $\boldsymbol{\mathcal{X}}_t = (\boldsymbol{U}_t, \boldsymbol{V}_t, \boldsymbol{W}_t)\cdot \boldsymbol{\mathcal{C}}_t$. By checking (9) and (10), it is straightforward to verify that the next iterate from $\widetilde{\boldsymbol{F}}_t$ follow the same change of parameterization, i.e.

$$\widetilde{\boldsymbol{F}}_{t+1} = (\boldsymbol{U}_{t+1}\boldsymbol{Q}_1, \boldsymbol{V}_{t+1}\boldsymbol{Q}_2, \boldsymbol{W}_{t+1}\boldsymbol{Q}_3, (\boldsymbol{Q}_1^{-1}, \boldsymbol{Q}_2^{-1}, \boldsymbol{Q}_3^{-1})\cdot \boldsymbol{\mathcal{C}}_{t+1}),$$

which ensures the update rule of ScaledSM is insensitive to the imbalance of the factors in the low-rank parameterization—a key property that is absent in the vanilla subgradient method and contributes to the performance gain.

2.2. Truncated spectral initialization

Inspired by the median-truncated spectral initialization in [29–31], we propose a tensor counterpart that is tailored to our problem to initialize ScaledSM. Denote $\boldsymbol{y}_{\mathsf{trunc}}$ as the vector after discarding p_s fraction of measurements with largest magnitudes:

$$[\boldsymbol{y}_{\mathsf{trunc}}]_i = \begin{cases} \frac{y_i}{1-p_s}, & \text{if} \quad |y_i| \le |\boldsymbol{y}|_{(\lceil p_s m\rceil)} \\ 0, & \text{otherwise} \end{cases}, \tag{12}$$

where $|\boldsymbol{y}|_{(k)}$ denotes the k-th largest amplitude of $\boldsymbol{y}$. Let $\mathcal{A}^*(\cdot)$ be the adjoint operator of $\mathcal{A}(\cdot)$. The truncated spectral initialization $\boldsymbol{F}_0 = (\boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{W}_0, \boldsymbol{\mathcal{C}}_0)$ is then given by the top-$\boldsymbol{r}$ higher-order SVD (HOSVD) of $\mathcal{A}^*(\boldsymbol{y}_{\mathsf{trunc}})$:

$$(\boldsymbol{U}_0, \boldsymbol{V}_0, \boldsymbol{W}_0)\cdot \boldsymbol{\mathcal{C}}_0 = \mathcal{H}_{\boldsymbol{r}}(\mathcal{A}^*(\boldsymbol{y}_{\mathsf{trunc}})), \tag{13}$$

where $\boldsymbol{U}_0$ is the top-r_1 left singular vectors of $\mathcal{M}_1(\mathcal{A}^*(\boldsymbol{y}_{\mathsf{trunc}}))$, analogously for $\boldsymbol{V}_0, \boldsymbol{W}_0$, and $\boldsymbol{\mathcal{C}}_0 = (\boldsymbol{U}_0^\top, \boldsymbol{V}_0^\top, \boldsymbol{W}_0^\top)\cdot \mathcal{A}^*(\boldsymbol{y}_{\mathsf{trunc}})$ is the core tensor.

3. THEORETICAL GUARANTEES

We focus on presenting the local linear convergence of the proposed scaled subgradient method while leaving a complete account of global convergence to the future work. To begin with, we introduce a key metric that defines a sort of condition number of the low-rank tensor. The condition number of $\boldsymbol{\mathcal{X}}_\star$ is defined as

$$\kappa := \frac{\sigma_{\max}(\boldsymbol{\mathcal{X}}_\star)}{\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)} = \frac{\max_{k=1,2,3}\sigma_{\max}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))}{\min_{k=1,2,3}\sigma_{\min}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star))}, \tag{14}$$

where $\sigma_{\max}(\boldsymbol{\mathcal{X}}) = \max_{k=1,2,3}\sigma_{\max}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}))$, $\sigma_{\min}(\boldsymbol{\mathcal{X}}) = \min_{k=1,2,3}\sigma_{\min}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}))$, and $\sigma_{\max}(\mathcal{M}_k(\boldsymbol{\mathcal{X}})), \sigma_{\min}(\mathcal{M}_k(\boldsymbol{\mathcal{X}}))$ are the largest and the smallest nonzero singular values of $\mathcal{M}_k(\boldsymbol{\mathcal{X}})$.

3.1. A general theory of local linear convergence

Our convergence guarantees are built on standard geometric assumptions [11, 23, 26] on the loss function $f(\cdot)$ for the analysis of subgradient-type algorithms, which are defined as follows.

Definition 1 (Restricted Lipschitz continuity)**.** A function $f : \mathbb{R}^{n_1\times n_2\times n_3} \mapsto \mathbb{R}$ is said to be rank-$\boldsymbol{r}$ restricted L-Lipschitz continuous for some quantity $L > 0$ if

$$|f(\boldsymbol{\mathcal{X}}_1) - f(\boldsymbol{\mathcal{X}}_2)| \le L\|\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2\|_{\mathsf{F}}$$

holds for any $\boldsymbol{\mathcal{X}}_1, \boldsymbol{\mathcal{X}}_2 \in \mathbb{R}^{n_1\times n_2\times n_3}$ such that $\boldsymbol{\mathcal{X}}_1 - \boldsymbol{\mathcal{X}}_2$ has multilinear rank at most $2\boldsymbol{r}$.

Definition 2 (Restricted sharpness)**.** A function $f : \mathbb{R}^{n_1\times n_2\times n_3} \mapsto \mathbb{R}$ is said to be rank-$\boldsymbol{r}$ restricted μ-sharp w.r.t. $\boldsymbol{\mathcal{X}}_\star$ for some $\mu > 0$ if

$$f(\boldsymbol{\mathcal{X}}) - f(\boldsymbol{\mathcal{X}}_\star) \ge \mu\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$$

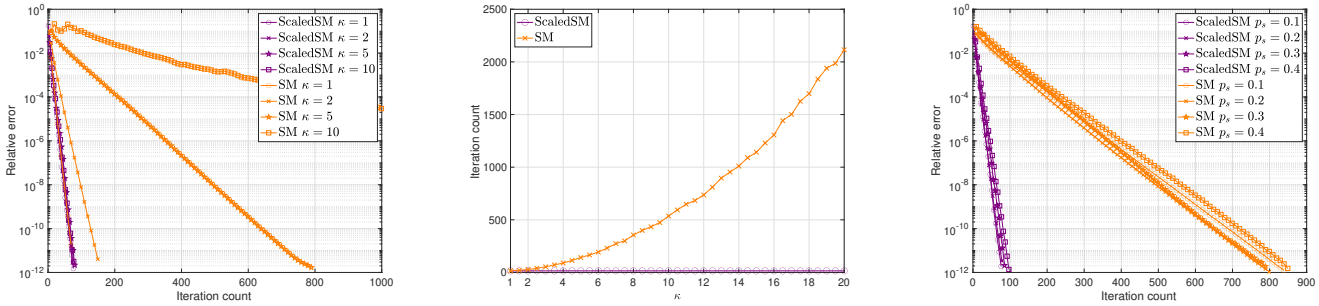
holds for any $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1\times n_2\times n_3}$ with multilinear rank at most $\boldsymbol{r}$.

The condition number of a function $f(\cdot)$ that is both restricted L-Lipschitz continuous and μ-sharp is then denoted by

$$\chi_f := L/\mu. \tag{15}$$

To fully capture the performance progress of ScaledSM, we measure the performance of factor quadruple $\boldsymbol{F} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{\mathcal{C}})$ using the following error metric [9]

$$\mathrm{dist}^2(\boldsymbol{F}, \boldsymbol{F}_\star) := \inf_{\boldsymbol{Q}_k\in\mathrm{GL}(r_k)} \|(\boldsymbol{U}\boldsymbol{Q}_1 - \boldsymbol{U}_\star)\boldsymbol{\Sigma}_{\star,1}\|_{\mathsf{F}}^2$$



(a) (b) (c)

Fig. 1. Performance comparisons of the proposed method (ScaledSM) and the vanilla subgradient method (SM). (a) The reconstruction errors $\|\boldsymbol{\mathcal{X}}_t-\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}/\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$ w.r.t. the iteration count under different condition numbers $\kappa=1,2,5,10$ with $p_s=0.2$. (b) The iteration complexities w.r.t. the condition number for achieving $\|\boldsymbol{\mathcal{X}}_t-\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}\le 10^{-3}\|\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}$ with $p_s=0.2$. (c) The reconstruction errors w.r.t. the iteration count under different amounts of outliers $p_s=0.1,0.2,0.3,0.4$ with $\kappa=5$.

$$+\|(\boldsymbol{V}\boldsymbol{Q}_2-\boldsymbol{V}_\star)\boldsymbol{\Sigma}_{\star,2}\|_{\mathsf{F}}^2+\|(\boldsymbol{W}\boldsymbol{Q}_3-\boldsymbol{W}_\star)\boldsymbol{\Sigma}_{\star,3}\|_{\mathsf{F}}^2 +\big\|(\boldsymbol{Q}_1^{-1},\boldsymbol{Q}_2^{-1},\boldsymbol{Q}_3^{-1})\cdot\boldsymbol{\mathcal{C}}-\boldsymbol{\mathcal{C}}_\star\big\|_{\mathsf{F}}^2, \quad (16)$$

which takes into consideration both the representation ambiguity of the factorization up to invertible transforms and the scaling of different factors due to the presence of preconditioners, where $\boldsymbol{\Sigma}_{\star,k}$ denotes the diagonal matrix composed of nonzero singular values of $\mathcal{M}_k(\boldsymbol{\mathcal{X}}_\star)$. With this metric in place, we state the linear convergence of the scaled subgradient method when $f(\cdot)$ satisfies both the rank-$\boldsymbol{r}$ restricted L-Lipschitz continuity and μ-sharpness, as follows.

Theorem 1 (Scaled subgradient method with exact convergence)**.** *Suppose that $f(\boldsymbol{\mathcal{X}}):\mathbb{R}^{n_1\times n_2\times n_3}\mapsto\mathbb{R}$ is convex in $\boldsymbol{\mathcal{X}}$, and satisfies rank-$\boldsymbol{r}$ restricted L-Lipschitz continuity and μ-sharpness (cf. Definitions 1 and 2). In addition, suppose that the initialization $\boldsymbol{F}_0$ satisfies*

$$\mathrm{dist}(\boldsymbol{F}_0,\boldsymbol{F}_\star)\le 10^{-3}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)/\chi_f, \quad (17)$$

and the scaled subgradient method adopts the geometrically decaying step sizes in (10) with $\lambda=\frac{(\sqrt{2}-1)^{3/2}}{2}10^{-3}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)/\chi_f^2$ and $q=(1-0.016/\chi_f^2)^{1/2}$. Then for all $t\ge 0$, the iterates satisfy

$$\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)\le(1-0.016/\chi_f^2)^{t/2}10^{-3}\sigma_{\min}(\boldsymbol{\mathcal{X}}_\star)/\chi_f,$$
$$\text{and}\quad \|\boldsymbol{\mathcal{X}}_t-\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}\le 3\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star).$$

Theorem 1 shows that the iterates of the scaled subgradient method converges at a linear rate; to reach ϵ-accuracy, i.e. $\|\boldsymbol{\mathcal{X}}_t-\boldsymbol{\mathcal{X}}_\star\|_{\mathsf{F}}\le\epsilon\sigma_r(\boldsymbol{\mathcal{X}}_\star)$, it takes at most $O(\chi_f^2\log\frac{1}{\epsilon})$ iterations, which, importantly, is *independent of the condition number* κ of $\boldsymbol{\mathcal{X}}_\star$. Finally, it is worth noting that the choices of constants in Theorem 1 are pessimistic to simplify analysis. Due to space limits, we defer the full proof to [34].

3.2. Case study: Gaussian design

It turns out that under the Gaussian design, where all the sensing tensors are composed of i.i.d. Gaussian entries, the resulting loss function obeys the rank-$\boldsymbol{r}$ restricted L-Lipschitz continuity and μ-sharpness with high probability.

Proposition 1 (Gaussian designs)**.** *Let $n:=\max\{n_1,n_2,n_3\}$ and $r:=\max\{r_1,r_2,r_3\}$. Suppose that $[\mathcal{A}(\boldsymbol{\mathcal{X}})]_i=\frac{1}{m}\langle\boldsymbol{\mathcal{A}}_i,\boldsymbol{\mathcal{X}}\rangle$ with tensors $\boldsymbol{\mathcal{A}}_1,\ldots,\boldsymbol{\mathcal{A}}_m$ composed of i.i.d. standard Gaussian entries. Then with probability exceeding $1-c_1n^{-c_2}$, the loss function $f(\boldsymbol{\mathcal{X}})=\|\mathcal{A}(\boldsymbol{\mathcal{X}})-\boldsymbol{y}\|_1$ in (5) satisfies the rank-$\boldsymbol{r}$ restricted L-Lipschitz continuity and μ-sharpness with*

$$L=0.8,\quad \mu=0.79(1-2p_s), \quad (18)$$

as long as $m\ge\frac{C(nr+r^3)}{(1-2p_s)^2}\log\left(\frac{1}{1-2p_s}\right)$. Here, C,c_1,c_2 are some universal constants.

Combining Theorem 1 and Proposition 1, ScaledSM is guaranteed to reach ϵ-accuracy in at most $O\left(\frac{1}{(1-2p_s)^2}\log\frac{1}{\epsilon}\right)$ iterations, as long as the sample size is sufficiently large. This amounts to a near-optimal sample complexity $O(nr+r^3)$ and dimension-free iteration complexity $O(\log\frac{1}{\epsilon})$ even with a constant fraction of outliers.

Beyond the Gaussian design, similar guarantees can be established when the observation operator satisfies the mixed-norm restricted isometry property [11].

4. NUMERICAL EXPERIMENTS

In this section, we provide numerical experiments to illustrate the performance of ScaledSM for robust tensor regression, and highlight its advantage compared to the vanilla subgradient method (SM). For simplicity, we set $n_1=n_2=n_3=30$, and $r_1=r_2=r_3=3$, and collect $m=5000$ measurements according to (4). The ground truth tensor $\boldsymbol{\mathcal{X}}_\star$ is generated as described in [9]. Each outlier is independently generated as $s_i=\bar{s}_i\Omega_i$, with Ω_i drawn from a Bernoulli distribution with parameter p_s, and $\bar{s}_i$ drawn from a uniform distribution in $[-10\|\mathcal{A}(\boldsymbol{\mathcal{X}}_\star)\|_\infty,10\|\mathcal{A}(\boldsymbol{\mathcal{X}}_\star)\|_\infty]$. Both ScaledSM and SM start from the same truncated spectral initialization (13), and for simplicity use the Polyak's step size (which amounts to using optimally tuned geometrically decaying step sizes).

Fig. 1 shows the detailed performance comparison of ScaledSM and SM under various settings. Thanks to the robustness of the least absolute deviation loss, both algorithms converge linearly in the presence of outliers. Noteworthily, ScaledSM converges as a fast rate that is independent with κ, while SM slows down dramatically as κ increases. Indeed, the iteration complexity of SM grows super linearly with respect to condition number κ, while ScaledSM takes a much smaller number of iterations and therefore accelerates the convergence for ill-conditioned instances.

5. CONCLUSIONS

This paper develops a scaled subgradient method for robust low-rank tensor regression from corrupted measurements, by minimizing the a natural nonsmooth and nonconvex loss function based on least absolute deviation. For future work, it is of interest to examine if it is possible to develop provably efficient algorithms for the related problem called robust low-rank tensor completion [21].

6. REFERENCES

[1] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[2] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3551–3582, 2017.

[3] B. Huang, C. Mu, D. Goldfarb, and J. Wright, "Provable models for robust low-rank tensor completion," *Pacific Journal of Optimization*, vol. 11, no. 2, pp. 339–364, 2015.

[4] H. Rauhut, R. Schneider, and Ž. Stojanac, "Low rank tensor recovery via iterative hard thresholding," *Linear Algebra and its Applications*, vol. 523, pp. 220–262, 2017.

[5] H. Chen, G. Raskutti, and M. Yuan, "Non-convex projected gradient descent for generalized low-rank tensor regression," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 172–208, 2019.

[6] R. Han, R. Willett, and A. Zhang, "An optimal statistical and computational framework for generalized tensor estimation," *arXiv preprint arXiv:2002.11255*, 2020.

[7] A. Zhang, Y. Luo, G. Raskutti, and M. Yuan, "ISLET: Fast and optimal low-rank tensor regression via importance sketching," *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 2, pp. 444–479, 2020.

[8] Y. Luo and A. R. Zhang, "Low-rank tensor estimation via Riemannian Gauss-Newton: Statistical optimality and second-order convergence," *arXiv preprint arXiv:2104.12031*, 2021.

[9] T. Tong, C. Ma, A. Prater-Bennette, E. Tripp, and Y. Chi, "Scaling and scalability: Provable nonconvex low-rank tensor estimation from incomplete measurements," *arXiv preprint arXiv:2104.14526*, 2021.

[10] T. Tong, C. Ma, and Y. Chi, "Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent," *Journal of Machine Learning Research*, pp. 1–67, 2021.

[11] ——, "Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2396–2409, 2021.

[12] G. Raskutti, M. Yuan, and H. Chen, "Convex regularization for high-dimensional multiresponse tensor regression," *The Annals of Statistics*, vol. 47, no. 3, pp. 1554–1584, 2019.

[13] B. Barak and A. Moitra, "Noisy tensor completion via the sum-of-squares hierarchy," in *Conference on Learning Theory*. PMLR, 2016, pp. 417–445.

[14] C. Cai, G. Li, H. V. Poor, and Y. Chen, "Nonconvex low-rank tensor completion from noisy data," in *Advances in Neural Information Processing Systems*, 2019, pp. 1863–1874.

[15] A. Liu and A. Moitra, "Tensor completion made practical," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[16] A. Montanari and N. Sun, "Spectral algorithms for tensor completion," *Communications on Pure and Applied Mathematics*, vol. 71, no. 11, pp. 2381–2425, 2018.

[17] C. Cai, G. Li, Y. Chi, H. V. Poor, and Y. Chen, "Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees," *The Annals of Statistics*, vol. 49, no. 2, pp. 944–967, 2021.

[18] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Spectral methods for data science: A statistical perspective," *Foundations and Trends® in Machine Learning*, vol. 14, no. 5, pp. 566–806, 2021.

[19] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, pp. 11:1–11:37, 2011.

[20] D. Goldfarb and Z. Qin, "Robust low-rank tensor recovery: Models and algorithms," *SIAM Journal on Matrix Analysis and Applications*, vol. 35, no. 1, pp. 225–253, 2014.

[21] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan, "Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5249–5257.

[22] Y. Li, Y. Sun, and Y. Chi, "Low-rank positive semidefinite matrix recovery from corrupted rank-one measurements," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 397–408, 2017.

[23] J. C. Duchi and F. Ruan, "Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval," *Information and Inference: A Journal of the IMA*, vol. 8, no. 3, pp. 471–529, 2019.

[24] V. Charisopoulos, D. Davis, M. Díaz, and D. Drusvyatskiy, "Composite optimization for robust blind deconvolution," *arXiv preprint arXiv:1901.01624*, 2019.

[25] X. Li, Z. Zhu, A. M.-C. So, and R. Vidal, "Nonconvex robust low-rank matrix recovery," *SIAM Journal on Optimization*, vol. 30, no. 1, pp. 660–686, 2020.

[26] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy, "Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence," *Foundations of Computational Mathematics*, pp. 1–89, 2021.

[27] J. Ma and S. Fattahi, "Implicit regularization of sub-gradient method in robust matrix recovery: Don't be afraid of outliers," *arXiv preprint arXiv:2102.02969*, 2021.

[28] L. Ding, L. Jiang, Y. Chen, Q. Qu, and Z. Zhu, "Rank overspecified robust matrix recovery: Subgradient method and exact recovery," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[29] H. Zhang, Y. Chi, and Y. Liang, "Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow," in *International Conference on Machine Learning*, 2016, pp. 1022–1031.

[30] Y. Li, Y. Chi, H. Zhang, and Y. Liang, "Non-convex low-rank matrix recovery with arbitrary outliers via median-truncated gradient descent," *Information and Inference: A Journal of the IMA*, vol. 9, no. 2, pp. 289–325, 2020.

[31] H. Zhang, Y. Chi, and Y. Liang, "Median-truncated nonconvex approach for phase retrieval with outliers," *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 7287–7310, 2018.

[32] X. Yi, D. Park, Y. Chen, and C. Caramanis, "Fast algorithms for robust PCA via gradient descent," in *Advances in neural information processing systems*, 2016, pp. 4152–4160.

[33] J.-L. Goffin, "On convergence rates of subgradient optimization methods," *Mathematical programming*, vol. 13, no. 1, pp. 329–347, 1977.

[34] T. Tong, "Scaled gradient methods for ill-conditioned low-rank matrix and tensor estimation," Ph.D. dissertation, Carnegie Mellon University, 2022.