LATENT SCHATTEN TT NORM FOR TENSOR COMPLETION

Andong Wang[1,2,3], *Xulin Song*[1,2], *Xiyin Wu*[1,2], *Zhihui Lai*[4] *and Zhong Jin*[1,2*]

[1]School of Computer Science and Engineering, Nanjing University of Science & Technology, China
[2]Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science & Technology, China
[3] Tensor Learning Unit, RIKEN Center for Advanced Intelligence Project (AIP), Japan
[4]School of Computer Science and Software Engineering, Shenzhen University, China

ABSTRACT

Tensor completion arouses much attention in signal processing and machine learning. The tensor train (TT) decomposition has shown better performances than the Tucker decomposition in image and video inpainting. In this paper, we propose a novel tensor completion model based on a newly defined latent Schatten TT norm. Then, the statistical performance is analyzed by establishing a non-asymptotic upper bound on the estimation error. Further, a scalable algorithm is developed to efficiently solve the model. Experimental results of color image inpainting demonstrate that the proposed norm has promising performances compared to other variants of Schatten norm.

Index Terms— tensor completion, TT decomposition, Schatten norm, statistical performance

1. INTRODUCTION

Due to the superior representation power for multi-way data, tensor based methods have been broadly studied across many fields including algebra, signal processing and machine learning [1–3]. In many applications, multi-way data arrays may be noisy and incomplete due to unavoidable reasons such as natural noise in the sensor, communication errors, and missing at random. Tensor completion aims to reconstruct a tensor from partial (and maybe noisy) observations [3, 4]. Generally speaking, it is ill-posed unless some assumptions are made on the tensor to complete. A natural assumption is that the underlying tensor is low-rank, then it has low degree of freedom and can potentially be recovered from partial observations. Due to the diversity of tensor decompositions, a tensor has multiple definitions of rank function [1]. The Tucker decomposition based rank minimization models [3] are the most influential in tensor completion. By summing Schatten 1-norms of all the mode-k unfoldings, the tensor Schatten norm (or tensor nuclear norm) models the underlying tensor as low Tucker rank and has been extensively studied [5–7]. As the low Tucker rank assumption may be too strong for some real data tensor, the latent Schatten norm [6] is proposed and attains state-of-the-art performances in many tasks [7]. Tensor completion based on other tensor decompositions (e.g. CP decomposition[8], Tensor-ring [9], t-SVD [10]) is also an active research topic [11–13].

Recently, the Tensor-train (TT) decomposition has attracted much attention. For higher-order tensors, TT decomposition provides more space-saving representation called TT format while preserving the representation power. TT decomposition based rank minimization has shown better performances than the Tucker-based algorithms in tensor completion [14–16].

In this paper, we study the tensor completion problem via a tensor norm closed related to the TT decomposition. The contributions of this paper are three-fold. First, a novel tensor completion model based on a newly defined latent Schatten TT norm is proposed. Second, statistical performance of the proposed model is analyzed by establishing a non-asymptotic upper bound on the estimation error. Third, we develop a scalable algorithm to efficiently solve the proposed model. Effectiveness of the proposed norm is demonstrated in color image inpainting experiments.

2. PRELIMINARIES AND RELATED WORK

Notations. We use lowercase bold letters for vectors (e.g. $\mathbf{v}$), uppercase bold letters for matrices (e.g. $\mathbf{M}$) and calligraphic letters for tensors (e.g. $\mathcal{T}$). If the dimension is not given explicitly, all K-th order tensors are in $\mathbb{R}^{d_1 \times \cdots \times d_K}$. For any integer n, let $[n] := \{1, \cdots, n\}$. For all $k \in [K-1]$, define $d_{\leq k} := \prod_{l \leq k} d_l$; similarly, $d_{>k}$ are defined. For simplicity, let $D = \prod_{k=1}^{K} d_k$, and $\hat{d}_k = \min\{d_{\leq k}, d_{>k}\}, \tilde{d}_k = d_{\leq k} + d_{>k}, \forall k \in [K-1]$. Given $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ with singular values σ_i, define its spectral norm and nuclear norm as $\|\mathbf{M}\| = \max_i \sigma_i$ and $\|\mathbf{M}\|_* = \sum_i \sigma_i$, respectively. The inner product between two tensors $\mathcal{T}_1$ and $\mathcal{T}_2$ is defined as $\langle \mathcal{T}_1, \mathcal{T}_2 \rangle = \mathrm{vec}(\mathcal{T}_1)^\top \mathrm{vec}(\mathcal{T}_2)$, where $\mathrm{vec}(\cdot)$ is the operation of vectorization [1]. The tensor Frobenius norm is defined as $\|\mathcal{T}\|_{\mathrm{F}} = \|\mathrm{vec}(\mathcal{T})\|_2$. The tensor

*✉ *zhongjin@njust.edu.cn*. This work is partially supported by National Natural Science Foundation of China under Grant Nos. 61872188, U1713208, 61602244, 61672287, 61702262, 61773215, 61703209, and by the Natural Science Foundation of Guangdong Province under Grant No. 2017A030313367.

l_1-norm and l_∞-norm are defined as $\|\mathcal{T}\|_1 = \|\text{vec}(\mathcal{T})\|_1$ and $\|\mathcal{T}\|_\infty = \|\text{vec}(\mathcal{T})\|_\infty$, respectively. Let $\mathbf{T}_{<k>} \in \mathbb{R}^{d_{\leq k} \times d_{>k}}$ be the mode-$(1, \cdots, k)$ unfolding matrix where the first k modes of $\mathcal{T} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ are combined into the rows and the rest $K-k$ modes are combined into the columns. Define the unfolding operator $\mathfrak{F}_k(\cdot) : \mathbb{R}^{d_1 \times \cdots \times d_K} \to \mathbb{R}^{d_{\leq k} \times d_{>k}}$ such that $\mathfrak{F}_k(\mathcal{T}) = \mathbf{T}_{<k>}$ and let $\mathfrak{F}_k^{-1}(\cdot)$ be its inverse.

Tensor-train decomposition. Given a vector of positive integers $(\mathtt{r}_0, \cdots, \mathtt{r}_K)$ and K third-order tensors $\mathcal{G}_k \in \mathbb{R}^{\mathtt{r}_{k-1} \times d_k \times \mathtt{r}_k}$ with $\mathtt{r}_0 = \mathtt{r}_K = 1$, the Tensor-train (TT) decomposition represents each elements of $\mathcal{T} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ as

$$\mathcal{T}(i_1, i_2, \cdots, i_K) = \prod\nolimits_k \mathcal{G}(:, i_k, :), \tag{1}$$

where $\mathcal{G}(:, i_k, :)$ is a $\mathtt{r}_{k-1} \times \mathtt{r}_k$ matrix. Let $\boldsymbol{\mathcal{G}} = \{\mathcal{G}_k\}_{k=1}^K$ be the set of $\mathcal{G}_k$'s and TT($\boldsymbol{\mathcal{G}}$) denotes the tensor whose elements are represented by $\boldsymbol{\mathcal{G}}$ as Eq. (1). We call $\vec{r}(\mathcal{T}) = (\mathtt{r}_0, \cdots, \mathtt{r}_K)^\top \in \mathbb{R}^{K+1}$ the TT-rank of $\mathcal{T}$. By Theorem 2.1 in [17], we have

$$\mathtt{r}_k \leq \text{rank}(\mathbf{T}_{<k>}),\ \forall k \in [K-1]. \tag{2}$$

Schatten TT norm. In [16], the Schatten TT norm defined as

$$\|\mathcal{T}\|_{\text{T}\star} := \frac{1}{K-1} \sum\nolimits_{k=1}^{K-1} \|\mathbf{T}_{<k>}\|_*, \tag{3}$$

was proposed for tensor completion. For each mode-k, the matrix nuclear norm is used to minimize $\text{rank}(\mathbf{T}_{<k>})$ which upper bounds the TT-rank $\mathtt{r}_k$ by Eq. (2). The Schatten TT norm essentially models the original tensor as simultaneously low-rank along all mode-$(1, \cdots, k)$ foldings. It can be taken as an extension of the tensor Schatten norm [6] defined as $\|\mathcal{T}\|_{\text{L}\star} := \sum_{k=1}^K \|\mathbf{T}_{(k)}\|_*$ [1], which models the underlying tensor as simultaneously low-rank along all mode-k unfoldings.

As shown in [6, 18], the "sum of Schatten norms" (e.g. the tensor Schatten norm) may lead to sub-optimal performances since it may be too strong for some real tensor data to be modeled as simultaneously low-rank. At the same times, the latent Schatten norm [6] defined as $\|\mathcal{T}\|_{\text{L}\star} := \inf_{\mathcal{T}=\sum_k \mathcal{X}^{(k)}} \sum_{k=1}^K \|\mathbf{X}_{(k)}^{(k)}\|_*$, has shown better performances than the tensor Schatten norm especially in modeling tensors only low-rank along certain modes [6, 18].

3. THE PROPOSED MODEL

3.1. The latent Schatten TT norm

Motivated by the latent Schatten norm, we define the latent Schatten TT norm as follows:

Definition 1 (Latent Schatten TT norm) *Given a tensor $\mathcal{T} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, its latent Schatten TT norm is defined as*

$$\|\mathcal{T}\|_{Tl\star} = \inf_{\mathcal{T}=\sum_{k=1}^{K-1} \mathcal{X}^{(k)}} \sum\nolimits_{k=1}^{K-1} \gamma_k \|\mathbf{X}_{<k>}^{(k)}\|_*, \tag{4}$$

where $\gamma_k > 0$'s are positive parameters.

[1] For any tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, notation $\mathbf{X}_{(k)} \in \mathbb{R}^{d_k \times (D/d_k)}$ is used to denote its mode-k unfolding [1].

By the latent Schatten TT norm, the original tensor is modeled as a mixture of $K-1$ latent tensors, each being low-rank along certain unfolding. We have the following lemma on the latent Schatten TT norm.

Lemma 1 *The dual norm of the latent Schatten TT norm $\|\cdot\|_{Tl\star}$, denoted by $\|\cdot\|_{Tl\star}^*$ can be computed as follows:*

$$\|\mathcal{T}\|_{Tl\star}^* = \max\nolimits_{k \in [K-1]} \gamma_k^{-1} \|\mathbf{T}_{<k>}\|. \tag{5}$$

The proof is a direct use of Theorem 16.4 in [19]. The dual norm of the latent Schatten TT norm plays a key role in statistical analysis and efficient computation of the proposed estimator.

3.2. Problem formulation

The observation model. Suppose we observe N scalars $\{y_i\}$ by the observation model:

$$y_i = \langle \mathcal{L}^*, \mathcal{X}_i \rangle + \sigma \xi_i, \quad \forall i \in [N], \tag{6}$$

where the standard deviation parameter $\sigma > 0$ is known, and the design tensors $\mathcal{X}_i$'s and noise variables ξ_n's satisfy the following assumptions:

A.1 *i.i.d. random design tensors.* $\mathcal{X}_i \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ are *i.i.d.* random tensors drawing from uniform distribution on the set $\{\mathbf{e}_{i_1} \circ \cdots \circ \mathbf{e}_{i_K}, \forall (i_1, \cdots, i_K) \in [d_1] \times \cdots \times [d_K]\}$.

A.2 *sub-exponential noise*[2]. ξ_n's follow independent sub-exponential distribution with moments $\mathbb{E}[\xi_n] = 0$, $\mathbb{E}[\xi_n^2] = 1$ and bounded exponential moment, i.e. $\exists$ constant $K_e > 0, \forall n \in [N]$, s.t. $\mathbb{E}[\exp(|\xi_n|/K_e)] < \infty$.

Let $\mathbf{y} = (y_1, \cdots, y_N)^\top$ and $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_N)^\top$. Define an operator $\mathfrak{X}(\mathcal{T}) := (\langle \mathcal{T}, \mathcal{X}_1 \rangle, \cdots, \langle \mathcal{T}, \mathcal{X}_N \rangle)^\top \in \mathbb{R}^N$, and its adjoint operator $\mathfrak{X}(\mathbf{z}) := \sum_{i=1}^N z_i \mathcal{X}_i$, $\forall \mathbf{z} \in \mathbb{R}^N$. Thus, the observation model (6) can be rewritten as

$$\mathbf{y} = \mathfrak{X}(\mathcal{L}^*) + \boldsymbol{\xi}. \tag{7}$$

The proposed estimator. To recover the true tensor $\mathcal{L}^*$ from noisy observations $\mathbf{y}$, we propose the following model

$$\hat{\mathcal{L}} \in \operatorname{argmin}_{\mathcal{L}} \frac{1}{2} \|\mathbf{y} - \mathfrak{X}(\mathcal{L})\|_2^2, \quad \text{s.t. } \mathcal{L} \in \underline{\mathrm{X}}(\alpha, \rho), \tag{8}$$

where $\underline{\mathrm{X}}(\alpha, \rho)$ is the set of tensors with limited l_∞-norm and latent Schatten TT norm:

$$\underline{\mathrm{X}}(\alpha, \rho) = \big\{\mathcal{L} \big| \|\mathcal{L}\|_\infty \leq \alpha, \|\mathcal{L}\|_{\text{Tl}\star} \leq \rho\big\}. \tag{9}$$

The motivation is to select the non-spiky $\mathcal{L}$ within latent Schatten TT norm of radius ρ that has least fitting error. We suppose parameters α and ρ are chosen sufficiently large such that the true tensor $\mathcal{L}^* \in \underline{\mathrm{X}}(\alpha, \rho)$.

[2] The Gaussian, symmetric Bernoulli, sub-Gaussian, Cauchy and exponential distributions are examples sub-exponential distributions.

4. STATISTICAL PERFORMANCE

Before analyzing the statistical performance on the estimation error, we first give three key lemmas. Let $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_N)^\top \in \mathbb{R}^N$ be any Rademacher sequence [13].

Lemma 2 *If the sample size $N \geq 2\max_k(\hat{d}_k \log^3 \tilde{d}_k)$, then with probability at least $1 - \sum_{k\in[K-1]} \tilde{d}_k^{-1}$, we have*

$$\|\mathfrak{X}(\boldsymbol{\xi})\|_{Tl\star}^* \leq \max_{k\in[K-1]} C_{K_e}\sqrt{\log \tilde{d}_k/(N\gamma_k \hat{d}_k)}, \tag{10}$$

where c_{K_e} is a constant depending on K_e in Assumption A.2.

The proof of Lemma 2 can be obtained by a combination of the non-commutative Bernstein inequity [20] and Lemma 1.

Lemma 3 *Define $\underline{\mathrm{Y}}(\beta, \delta) := \{\mathcal{T} \in \underline{\mathrm{X}}(1, \beta) \big| \|\mathcal{T}\|_F^2 \geq D\delta\}$. We have the following inequality for any $\Delta \in \underline{\mathrm{Y}}(\beta, \delta)$:*

$$\|\mathfrak{X}(\Delta)\|_2^2/N \;\geq\; \|\Delta\|_F^2/D - c_1\beta\mathbb{E}\big[\|\mathfrak{X}(\boldsymbol{\zeta})\|_{Tl\star}^*\big], \tag{11}$$

with probability at least $1 - \exp(-c_0 N\delta)$.

Lemma 3 shows that design operator satisfies the restricted strong convexity condition (RSC) in $\underline{\mathrm{Y}}(\beta, \delta)$ [21]. Its proof follows that of Lemma 6.1 in [22].

Lemma 4 *If $N \geq 2\max_k(\hat{d}_k \log^3 \tilde{d}_k)$, then it holds that*

$$\mathbb{E}\big[\|\mathfrak{X}(\boldsymbol{\zeta})\|_{Tl\star}^*\big] \leq \max_{k\in[K-1]} C_r\sqrt{\log \tilde{d}_k/(N\gamma_k\hat{d}_k)}, \tag{12}$$

where c_r is a constant.

The proof of Lemma 4 follows the proof of Lemma 6 in [13]. Based on Lemmas 2-4, we are able to establish an upper bound on the estimation error.

Theorem 1 *If $N \geq 2\max_k(\hat{d}_k \log^3 \tilde{d}_k)$ and Assumptions A.1 and A.2 hold, then for any solution $\hat{\mathcal{L}}$ of Model (8), the estimation error $\Delta := \hat{\mathcal{L}} - \mathcal{L}^*$ satisfies the following inequality with probability at least $1 - 2\sum_{k\in[K-1]} \tilde{d}_k^{-1}$:*

$$\frac{\|\Delta\|_F^2}{D} \leq \max\left\{C_1\rho(\alpha\vee\sigma)\max_k\sqrt{\frac{\log\tilde{d}_k}{N\gamma_k\hat{d}_k}}, C_2\frac{\alpha^2\log D}{N}\right\},$$

where C_1, C_2 are universal constants.

Following [22], the proof can be obtained by a combination of Lemmas 2-4. According to Theorem 1, when the noise level is comparable or dominated by α and the super-parameters $\gamma_k = 1/(K-1)$, $\forall k \leq K-1$, we have

$$\frac{\|\hat{\mathcal{L}} - \mathcal{L}^*\|_{\mathrm{F}}^2}{D} \leq O\bigg(\alpha\rho\max_k\Big(\frac{K\log\tilde{d}_k}{N\hat{d}_k}\Big)^{\frac{1}{2}}\bigg). \tag{13}$$

For $K = 2$, tensor completion degenerate to matrix completion and the upper bound is consistent with the bound for a nuclear norm based model in Eq. (3.22) of [22]. To the best of our knowledge, this is the first theoretical guarantee in Frobenius-norm error $\|\hat{\mathcal{L}} - \mathcal{L}^*\|_{\mathrm{F}}$ for "latent Schatten norms" based tensor completion models [7, 23].

5. OPTIMIZATION ALGORITHM

Problem (8) can be solved via the alternating direction method of multipliers (ADMM) [24]. However, in each iteration, one needs the singular value thresholding (SVT) [25] of $d_{\leq k} \times d_{>k}$ matrices with $k \in [K-1]$, which costs $O(d^{\lfloor 3K/2 \rfloor})$ for K-th order cubical tensors in $\mathbb{R}^{d\times d\times\cdots\times d}$. For large tensors, the per-iteration cost is very high. Here, we present a scalable algorithm (Algorithm 1).

Algorithm 1: APG Solver for Problem (15)

Input: $\mathcal{L}^0 = \mathcal{O}, \mathcal{V}^0 = \mathcal{O}, \mathbf{v}^0 = \mathbf{0}, \hat{\mathcal{V}}^0 = \mathcal{O}, \hat{\mathbf{v}}^0 = \mathbf{0}, z_0 = 1, \alpha, \rho, \epsilon, T_{\max}, I_{\max}$.

1: Set $t = 0$.
2: **while** $t \leq T_{\max}$ **do**
3: Compute $\mathcal{G}$ at $\boldsymbol{\Lambda}^t$ by Lemma 6;
4: **Line-search:** Set $H_{t,0} = H_t$. For $i = 0$ to $I_{\max}$ do:
4.a. Compute $\boldsymbol{\Lambda}_{i+1}^t$ by Eq. (17) at $\hat{\boldsymbol{\Lambda}}^t$ with $H_{t,i}$;
4.b. If $f(\boldsymbol{\Lambda}^{i+1}) \leq Q_{H_i}(\boldsymbol{\Lambda}^{i+1}, \hat{\boldsymbol{\Lambda}}^t) + 0.5\epsilon/z$ then $i_k = i$, and line-search loop. Otherwise, $H_{t,i+1} = 2H_{t,i}$.
End of line-search.
5: Update $\mathcal{V}^{t+1}$ and $\mathbf{v}^{t+1}$ by $\boldsymbol{\Lambda}^{t+1} = \boldsymbol{\Lambda}_{i_k+1}^t$;
6: Update $\hat{\mathcal{V}}^{t+1}$ and $\hat{\mathbf{v}}^{t+1}$ by Eq. (16);
7: Update $\mathcal{L}^{t+1}$ by $\mathcal{L}^{t+1} = (1-\gamma_t)\mathcal{L}^t + \gamma_t\mathcal{G}$;
8: $t = t + 1$.
9: **end while**

Output: $\mathcal{L}^{t+1}$.

First, Problem (8) is re-formulated as follows:

$$\begin{aligned} \min_{\mathcal{L},\mathcal{K},\mathbf{g},t} \quad & t \\ \text{s.t.} \quad & \mathcal{K} = \mathcal{L}, \mathbf{g} = \mathbf{y} - \mathfrak{X}(\mathcal{L}), \\ & \|\mathcal{L}\|_{\mathrm{Tl}\star} \leq t \leq \rho, \|\mathcal{K}\|_\infty \leq \alpha, \ \|\mathbf{g}\|_2 \leq \epsilon, \end{aligned} \tag{14}$$

where ϵ is a tolerance of noise. The Lagrangian is

$$\begin{aligned} & l(\mathcal{L}, \mathcal{K}, \mathbf{g}, t, \mathcal{V}, \mathbf{v}) = t + \langle\mathcal{V}, \mathcal{K} - \mathcal{L}\rangle + \langle\mathbf{v}, \mathfrak{X}(\mathcal{L}) + \mathbf{g} - \mathbf{y}\rangle \\ & \text{s.t. } \|\mathcal{L}\|_{\mathrm{Tl}\star} \leq t \leq \rho, \|\mathcal{K}\|_\infty \leq \alpha, \|\mathbf{g}\|_2 \leq \epsilon. \end{aligned}$$

Some algebra yields an equivalent problem as follows:

Lemma 5 *The dual problem of Problem (14) is*

$$\min_{\mathcal{V},\mathbf{v}} f(\mathcal{V}, \mathbf{v}) + r_1(\mathcal{V}) + r_2(\mathbf{v}), \tag{15}$$

where $r_1(\mathcal{V}) = \alpha\|\mathcal{V}\|_1$, $\quad r_2(\mathbf{v}) = \epsilon\|\mathbf{v}\|_2$ and

$$f(\mathcal{V}, \mathbf{v}) = \rho\max(\|\mathfrak{X}(\mathbf{v}) - \mathcal{V}\|_{Tl\star}^* - 1, 0) + \langle\mathbf{v}, \mathbf{y}\rangle.$$

Then, motivated by [26], we apply the accelerated proximal sub-gradient descent (Dual-APsG) to the dual problem (15). Specifically, we update $\boldsymbol{\Lambda} = (\mathcal{V}, \mathbf{v})$ and an auxiliary pair $\hat{\boldsymbol{\Lambda}} = (\hat{\mathcal{V}}, \hat{\mathbf{v}})$ alternatively in the t-th iteration as follows:

$$\begin{aligned} \boldsymbol{\Lambda}^{t+1} &= \mathrm{argmin}_{\boldsymbol{\Lambda}}\, Q_{H^t}(\boldsymbol{\Lambda}, \hat{\boldsymbol{\Lambda}}^t), \\ \hat{\boldsymbol{\Lambda}}^{t+1} &= \boldsymbol{\Lambda}^{t+1} + (z_t - 1)z_{t+1}^{-1}(\boldsymbol{\Lambda}^{t+1} - \boldsymbol{\Lambda}^t), \end{aligned} \tag{16}$$

where $Q_{H^t}(\mathbf{\Lambda}, \hat{\mathbf{\Lambda}}) = f(\hat{\mathbf{\Lambda}}) + \langle g(\hat{\mathbf{\Lambda}}), \mathbf{\Lambda} - \hat{\mathbf{\Lambda}}\rangle + r(\mathbf{\Lambda}) + \frac{H^t}{2}\|\mathbf{\Lambda} - \hat{\mathbf{\Lambda}}\|_{\mathrm{F}}^2$, $r(\mathbf{\Lambda}) = r_1(\mathcal{V}) + r_2(\mathbf{v})$, $g(\hat{\mathbf{\Lambda}}^t) = \big(g_1(\hat{\mathcal{V}}^t), g_2(\hat{\mathbf{v}}^t)\big)$ is the sub-gradient at $\hat{\mathbf{\Lambda}}^t$ given by Lemma 6, H_t is the reciprocal of the step size, and z_t is a scalar sequence updated iteratively by $z_{t+1} = (1 + \sqrt{1 + 4z_{t+1}^2})/2$ with initialization $z_0 = 1$.

Update $(\mathcal{V}^{t+1}, \mathbf{v}^{t+1})$**.** According to Eq. (16), we update $\Lambda^{t+1} = (\mathcal{V}^{t+1}, \mathbf{v}^{t+1})$ as follows:

$$\begin{aligned} \mathcal{V}^{t+1} &= \mathrm{prox}_{\alpha/H^{t+1}}^{\|\cdot\|_1}(\hat{\mathcal{V}}^t - g_1(\hat{\mathcal{V}}^t)/H^{t+1}), \\ \mathbf{v}^{t+1} &= \mathrm{prox}_{\epsilon/H^{t+1}}^{\|\cdot\|_2}(\hat{\mathbf{v}}^t - g_2(\hat{\mathbf{v}}^t)/H^{t+1}), \end{aligned} \tag{17}$$

with proximal operators of tensor l_1-norm and vector l_2-norm:

$$\begin{aligned} \mathrm{prox}_\tau^{\|\cdot\|_1}(\mathcal{X}) &= \mathrm{sign}(\mathcal{X}) \odot \max\{\mathrm{abs}(\mathcal{X}) - \tau, 0\}, \\ \mathrm{prox}_\tau^{\|\cdot\|_2}(\mathbf{x}) &= \max\{\|\mathbf{x}\|_2 - \tau, 0\}\|\mathbf{x}\|_2^{-1}\mathbf{x}, \end{aligned}$$

where $\odot$ denotes element-wise multiplication and we let $0/0 = 0$ for the case $\|\mathbf{x}\|_2 = 0$. Besides, the sub-gradients $g_1(\hat{\mathcal{V}}^t)$ and $g_2(\hat{\mathbf{v}}^t)$ can be computed by the following lemma.

Lemma 6 *Let $\tilde{\mathbf{\Lambda}} = \mathfrak{X}(\hat{\mathbf{v}}) - \hat{\mathcal{V}}$ and select any k^* such that $k^* \in \mathrm{argmax}_{k\in[K-1]}\gamma_k^{-1}\|\tilde{\mathbf{\Lambda}}_{<k>}\|$. Let $(\mathbf{u}^{k^*}, \mathbf{v}^{k^*})$ denote a pair of left and right singular vectors corresponding to the largest singular value of $\tilde{\mathbf{\Lambda}}_{<k^*>}$. Then one particular sub-gradient of $f(\hat{\mathcal{V}}, \hat{\mathbf{v}})$ with respect to $\hat{\mathcal{V}}$ and $\hat{\mathbf{v}}$, denoted respectively by $g_1(\hat{\mathcal{V}})$ and $g_2(\hat{\mathbf{v}})$, can be computed as*

$$g_1(\hat{\mathcal{V}}) = -\mathcal{G} \quad \text{and} \quad g_2(\hat{\mathbf{v}}) = \mathfrak{X}(\mathcal{G}) + \mathbf{y}, \tag{18}$$

where $\mathcal{G} = \mathbf{1}_{1\le\|\tilde{\mathbf{\Lambda}}\|^*_{\overline{\mathcal{T}l}_\star}} \cdot \rho\mathfrak{F}_{k^*}^{-1}(\mathbf{u}^{k^*}\mathbf{v}^{k^*\top})$.

Update H^t**.** We update H^t by linear search to reach the condition that $f(\mathbf{\Lambda})$ is smaller than $Q_{H^t}(\mathbf{\Lambda}, \hat{\mathbf{\Lambda}})$ up to some tolerance, i.e., $f(\mathbf{\Lambda}) \le Q_{H^t}(\mathbf{\Lambda}, \hat{\mathbf{\Lambda}}) + 0.5\epsilon_0/z_t$.

Update the Primal Variable. The primal variable $\mathcal{L}$ is updated by $\mathcal{L}^{t+1} = (1 - \gamma_t)\mathcal{L}^t + \gamma_t\mathcal{G}$.

Complexity and convergence analysis. The main cost lies in computing $\mathcal{G}$ in Lemma 6 and the line search step. Since only a pair of leading singular vectors are computed, per-iteration cost of Algorithm 1 is $O(I_{\max}(K-1)D)$. Thus, for K-th order cubical tensors $\mathcal{T} \in \mathbb{R}^{d\times d\times\cdots\times d}$, the per-iteration cost will be $O\big(I_{\max}(K-1)d^K\big)$, which is significantly lower than $O(d^{\lfloor 3K/2\rfloor})$ which is the per-iteration cost of SVT-based solvers. The convergence of Algorithm 1 given in Theorem 2 is a direct consequence of the analysis in [27].

Theorem 2 *The sequence of primal variables $\{(\mathcal{L}^t)\}$ generated by Algorithm 1 converges to a stationary point $\hat{\mathcal{L}}$, and the iteration number $T_{\max}$ in the worst case to achieve an ϵ-solution is $O\left(\inf_{\nu\in[0,1]}\left(\frac{M_\nu}{\epsilon}\right)^{\frac{2}{1+\nu}}\right)$, where the parameter M_ν with the Hölder smoothness order ν is defined as $M_\nu := \sup_{\mathbf{\Lambda}\neq\tilde{\mathbf{\Lambda}}}\|\nabla f(\mathbf{\Lambda}) - \nabla f(\tilde{\mathbf{\Lambda}})\|_F/\|\mathbf{\Lambda} - \tilde{\mathbf{\Lambda}}\|_F^\nu$, with $\nabla f(\cdot)$ denoting a sub-gradient of $f(\cdot)$.*

6. EXPERIMENT RESULTS

Effectiveness of the proposed norm (LatentTT) is demonstrated in color image inpainting, compared with models based on other variants of Schatten norm, i.e., tensor Schatten norm (FaLRTC) [3], scaled latent tensor Schatten norm (Latent) [28], tubal nuclear norm (ISTT) [13] and tensor Schatten TT norm (TT-ADMM) [16]. Six images of size $256 \times 256 \times 3$ are tested. The Peak Signal Noise Ratio (PSNR) is used to measure the quality of an estimation.

Given an image $\mathcal{M} \in \mathbb{R}^{d_1\times d_2\times 3}$, we randomly sample the pixels with probability $p = 0.15$ and add *i.i.d.* Gaussian noise with standard deviation $\sigma = 0.025\|\mathcal{M}\|_{\mathrm{F}}/\sqrt{D}$. We set the weight parameters $\boldsymbol{\alpha}$ for FaLRTC by $\alpha_1 : \alpha_2 : \alpha_3 = 1 : 1 : 0.01$ as suggested in [3]. For TT-ADMM, the tensor is reshaped in $\mathbb{R}^{4\times4\times4\times4\times4\times4\times4\times4\times3}$ as suggested in [14]. For the proposed LatentTT, the tensor is reshaped in $\mathbb{R}^{64\times2\times2\times768}$. Parameters γ_k $(k \le K-1)$ in Eq. (4) are set by $\gamma_k = \delta_k^{-0.3}/\sum_k \delta_k^{-0.3}$, where $\delta_k = \min\{d_{\le k}, d_{>k}\}$. The image inpainting results are shown in Fig. 1. It can be seen in Fig. 1 that the proposed norm outperform other variants of Schatten norm in most cases.

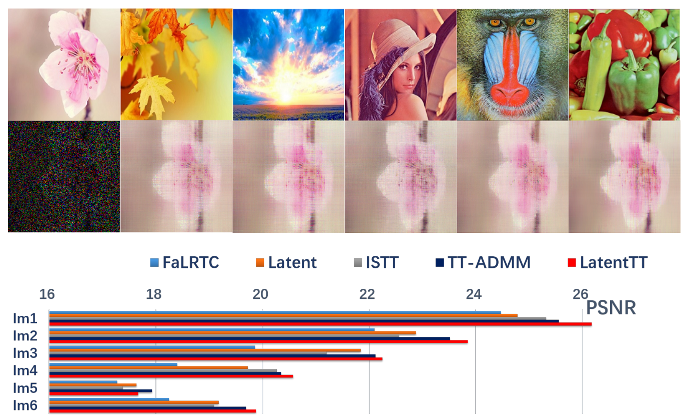

Fig. 1: Color image inpainting with 15% noisy observed entries. Top: six test images. Middle: visual comparison of Image 1; left to right: the observation, FaLRTC, Latent, ISTT, TT-ADMM and LatentTT. Bottom: PSNR values on the six images as a quantitative comparison.

7. CONCLUSION

Based on a newly defined latent Schatten TT norm, a novel tensor completion model is proposed. Then, a non-asymptotic upper bound on the estimation error is established. Further, a scalable algorithm is designed. Color image inpainting experiments show that the proposed model has promising performances compared with other Schatten norm based models. The main drawback of the non-asymptotic bound is that it is in terms of the latent Schatten TT norm of the underlying tensor rather than its TT-rank, which motivates the future work.

References

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

[2] Q. Zhao, L. Zhang, and A. Cichocki, “Bayesian cp factorization of incomplete tensors with automatic rank determination,” *IEEE TPAMI*, vol. 37, no. 9, pp. 1751–1763, 2015.

[3] J. Liu, P. Musialski, P. Wonka, and J. Ye, “Tensor completion for estimating missing values in visual data,” *IEEE TPAMI*, vol. 35, no. 1, pp. 208–220, 2013.

[4] A. Wang, Z. Lai, and Z. Jin, “Noisy low-tubal-rank tensor completion,” *Neurocomputing*, vol. 330, pp. 267–279, 2019.

[5] R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima, “Statistical performance of convex tensor decomposition,” in *NIPS*, 2011, pp. 972–980.

[6] R. Tomioka and T. Suzuki, “Convex tensor decomposition via structured schatten norm regularization,” in *NIPS*, 2013, pp. 1331–1339.

[7] M. Nimishakavi, P. Jawanpuria, and B. Mishra, “A dual framework for low-rank tensor completion,” in *NIPS*, 2018.

[8] J. D Carroll and J. Chang, “Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition,” *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[9] Q. Zhao, G. Zhou, S. Xie, et al., “Tensor ring decomposition,” *arXiv preprint arXiv:1606.05535*, 2016.

[10] M. E Kilmer, K. Braman, N. Hao, and R. C. Hoover, “Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging,” *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 1, pp. 148–172, 2013.

[11] M. Yuan and C. Zhang, “Incoherent tensor norms and their applications in higher order tensor completion,” *IEEE TIT*, vol. 63, no. 10, pp. 6753–6766, Oct 2017.

[12] W. Wang, V. Aggarwal, and S. Aeron, “Efficient low rank tensor ring completion,” in *ICCV*, Oct 2017, pp. 5698–5706.

[13] A. Wang, D. Wei, B. Wang, and Z. Jin, “Noisy low-tubal-rank tensor completion through iterative singular tube thresholding,” *IEEE Access*, vol. 6, pp. 35112–35128, 2018.

[14] J. A. Bengua, H. N. Phiem, et al., “Efficient tensor completion for color image and video recovery: Low-rank tensor train.,” *IEEE TIP*, vol. 26, no. 5, pp. 2466–2479, 2017.

[15] L. Yuan, Q. Zhao, and J. Cao, “High-order tensor completion for data recovery via sparse tensor-train optimization,” in *ICASSP*. IEEE, 2018, pp. 1258–1262.

[16] M. Imaizumi, T. Maehara, et al., “On tensor train rank minimization: Statistical efficiency and scalable algorithm,” in *NIPS*, 2017, pp. 3930–3939.

[17] I. V. Oseledets, “Tensor-train decomposition,” *SIAM Journal on Scientific Computing*, vol. 33, no. 5, pp. 2295–2317, 2011.

[18] G. Donald 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.

[19] RT Rockerfeller, *Convex analysis*, Princeton, NJ: Princeton University Press, 1970.

[20] O. Klopp, “Noisy low-rank matrix completion with general sampling distribution,” *Bernoulli*, vol. 20, no. 1, pp. 282–303, 2014.

[21] S. Negahban and M. J. Wainwright, “Restricted strong convexity and weighted matrix completion: Optimal bounds with noise,” *JMLR*, vol. 13, no. May, pp. 1665–1697, 2012.

[22] T. T. Cai and W. Zhou, “Matrix completion via max-norm constrained optimization,” *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1493–1525, 2016.

[23] K. Wimalawarne, M. Yamada, et al., “Convex coupled matrix and tensor completion,” in *NIPS*, 2018.

[24] S. Boyd, N. Parikh, et al., “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[25] J. Cai, E. J Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM J. OPTIMIZ.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[26] D. Wei, A. Wang, B. Wang, et al., “Tensor completion using spectral (k, p) -support norm,” *IEEE Access*, vol. 6, pp. 11559–11572, 2018.

[27] A. Yurtsever, Q. T Dinh, and V. Cevher, “A universal primal-dual convex optimization framework,” in *NIPS*, 2015, pp. 3150–3158.

[28] K. Wimalawarne, M. Sugiyama, and R. Tomioka, “Multitask learning meets tensor factorization: task imputation via convex optimization,” in *NIPS*, 2014, pp. 2825–2833.