Yours Truly

Sunil Template

Contents

Chapter 1

Low-rank tensor denoising and recovery via convex optimization

Ryota Tomioka

Toyota Technological Institute at Chicago, USA

Taiji Suzuki

Tokyo Institute of Technology, Japan

Kohei Hayashi

National Institute of Informatics, Japan

Hisashi Kashima

University of Tokyo, Japan

1.1 Introduction

Low-rank decomposition of tensors (multi-way arrays) naturally arises in many application areas, including signal processing, neuroimaging, bioinformatics, recommender systems and other relational data analysis [29, 35, 20].

This chapter reviews convex-optimization-based algorithms for . There are several reasons to look into convex optimization for tensor decomposition. First, it allows us to prove worst-case-performance guarantees. Although we might be able to give a performance guarantee for an estimator based on a non-convex optimization (see e.g., [43]), the practical relevance of the bound would be limited if we cannot obtain the optimum efficiently. Second, the convex methods allow us to side-step the tensor rank selection problem; in practice misspecification of tensor rank can significantly deteriorate the performance, whereas choosing a continuous regularization parameter can be considered an easier task. Third, it allows us to use various efficient techniques developed in the mathematical programming communities, such as proximity operation, alternating direction method of multipliers (ADMM), and duality gap monitoring that enable us to apply these algorithms to a variety of settings reliably. The norms we propose can be used for both denoising of a fully observed noisy tensor and reconstruction of a low-rank tensor from incomplete measurements. Of course there are limitations to what we can achieve with convex optimization, which we will discuss in Section 1.6. Nevertheless we hope that the methods we discuss here serve to connect tensor decomposition with statistics and (convex) optimization, which have been largely disconnected until recently, and contribute to the better understanding of the hardness and challenges of this area.

This chapter is structured as follows: in the next section, we introduce different notions of tensor ranks and present two norms that induce low-rank tensors, namely the overlapped Schatten 1-norm and latent Schatten 1-norm. In Section 1.3, we present denoising and recovery bounds for the two norms. The proofs of the theorems can be found in original papers [54, 53]. In Section 1.4, we propose optimization algorithms for the two norms based on primal and dual ADMM, respectively. Although ADMM has become a standard practice these days, our implementation allows us to deal with the noisy case and the exact case in the same framework (no need for continuation). We also discuss the choice of the penalty parameter η. Section 1.5 consists of some simple demonstrations of the implication of the theorems. Full quantitative evaluation of the bounds can be found in original papers [54, 53]. We discuss various extensions and related work in Section 1.6. We conclude this chapter with possible future directions.

1.2 Ranks and norms

Let $\mathcal{W} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_K}$ be a K-way tensor. We denote the total number of entries in $\mathcal{W}$ by $N = \prod_{k=1}^{K} n_k$.

1.2.1 Rank and multilinear rank

A tensor $\mathcal{W}$ is *rank one* if it can be expressed as an outer product of K vectors as

$$\mathcal{W} = \boldsymbol{u}^{(1)} \circ \boldsymbol{u}^{(2)} \circ \cdots \circ \boldsymbol{u}^{(K)},$$

which can be written element-wise as follows:

$$W_{i_1 i_2 \cdots i_K} = u^{(1)}_{i_1} u^{(2)}_{i_2} \cdots u^{(K)}_{i_K}, \quad (1 \leq i_k \leq n_k, k = 1, \ldots, K).$$

It is easy to verify that a tensor is rank one.

The $\mathcal{W}$ is the smallest number r such that $\mathcal{W}$ can be expressed as the sum of r rank-one tensors as follows:

$$\mathcal{W} = \sum_{j=1}^{r} \boldsymbol{u}_j^{(1)} \circ \boldsymbol{u}_j^{(2)} \circ \cdots \circ \boldsymbol{u}_j^{(K)}. \tag{1.1}$$

The above decomposition is known as the [22]. It is known that finding the rank r or computing the best rank r approximation (even for $r = 1$) is an NP hard problem [23, 21].

The multilinear rank of $\mathcal{W}$ is the K tuple $(r_1, \ldots, r_K)$ such that r_k is the dimension of the space spanned by the mode-k fibers [13, 29]; here mode-k fibers are the n_k dimensional vectors obtained by fixing all but the kth index. If $\mathcal{W}$ admits decomposition (1.1), r_k is at most r, in which case the multilinear rank of $\mathcal{W}$ is at most $(r, \ldots, r)$.

In contrast to the rank, the $(r_1, \ldots, r_K)$ can be computed efficiently. To this end, it is convenient to define the mode-k unfolding operation. The mode-k $\boldsymbol{W}_{(k)}$ is a $n_k \times N/n_k$ matrix obtained by concatenating the mode-k fibers along columns. Then r_k is the matrix rank of the mode-k unfolding $\boldsymbol{W}_{(k)}$.

The notion of multilinear rank is connected to another decomposition known as the [56] or the higher-order SVD [13, 14]

$$\mathcal{W} = \mathcal{C} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \cdots \times_K \boldsymbol{U}_K, \tag{1.2}$$

where $\times_k$ denotes the mode-k product [29].

Computation of decompositions (1.1) and (1.2) from large noisy tensor with possibly missing entries is a challenging task. Alternating least squares (ALS) [11] and higher-order orthogonal iteration (HOOI) [14] are well known and many extensions of them are proposed [29]. However, they typically come with no theoretical guarantee either about global optimality of the obtained solution or the statistical performance of the estimator. Kannan and Vempala (see Chapter 8) [27] proposed a sampling based algorithm with a performance bound, which requires knowledge of the Frobenius norms of the slices.

1.2.2 Convex relaxations

Recently, motivated by the success of the (also known as the and) for the recovery of low-rank matrices [16, 50, 42, 10, 43, 38], several authors have proposed norms that induce low-rank tensors.

These approaches solve convex problems of the following form:

$$\underset{\mathcal{W}}{\text{minimize}} \quad L(\mathcal{W}) + \lambda \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \mathcal{W} \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_{\star}, \tag{1.3}$$

where $L : \mathbb{R}^{n_1 \times \cdots \times n_K} \rightarrow \mathbb{R}$ is a convex loss function that measures how well $\mathcal{W}$ fits the data, $\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \mathcal{W} \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert_{\star}$ is a norm (we discuss in detail below), and $\lambda > 0$ is a regularization parameter.

For example, let's assume that the measurements $\boldsymbol{y} = (y_i)_{i=1}^{M}$ are generated as

$$y_i = \langle \mathcal{X}_i, \mathcal{W}^* \rangle + \epsilon_i, \tag{1.4}$$

where $\langle \mathcal{X}, \mathcal{W} \rangle$ denotes the inner product between two tensors viewed as vectors in $\mathbb{R}^N$; more precisely, $\langle \mathcal{X}, \mathcal{W} \rangle = \sum_{i_1,\ldots,i_K} X_{i_1 \ldots i_K} W_{i_1 \ldots i_K}$. Then the loss function can be defined as the sum of squared residuals

$$L(\mathcal{W}) = \frac{1}{2} \|\boldsymbol{y} - \mathfrak{X}(\mathcal{W})\|_2^2,$$

where $\mathfrak{X}(\mathcal{W}) := (\langle \mathcal{X}_i, \mathcal{W} \rangle)_{i=1}^{M}$.

The minimization problem (1.3) minimizes the loss function keeping also the norm small. Difference from the conventional optimization based approaches for tensor decomposition is that instead of constraining the (multilinear)rank of the decomposition, it only constrains the complexity of the solution measured by a particular norm.

In the case of matrices, it is well known that the Schatten 1-norm

$$\|\boldsymbol{W}\|_{S_1} = \sum_{j=1}^{r} \sigma_j(\boldsymbol{W}),$$

where $\sigma_j(\boldsymbol{W})$ is the jth singular value of $\boldsymbol{W}$ and r is the rank of $\boldsymbol{W}$, promotes the solution of (1.3) to be low-rank; see e.g., [15]. Intuitively, this can be understood analogous to the sparsity inducing property of the ℓ_1 norm; it promotes the spectrum of $\boldsymbol{W}$ to be sparse, i.e., a spectral version of lasso [51].

It is known that the Schatten 1-norm of a rank r matrix $\boldsymbol{W}$ can be related to its Frobenius norm as follows [50]:

$$\|\boldsymbol{W}\|_{S_1} \leq \sqrt{r}\|\boldsymbol{W}\|_F.$$

Thus a low-rank matrix has a small Schatten 1-norm relative to its Frobenius norm.

The following norm has been proposed by several authors [47, 18, 33, 52]:

$$\left|\!\left|\!\left|\mathcal{W}\right|\!\right|\!\right|_{\underline{S_1/1}} = \sum_{k=1}^{K} \|\boldsymbol{W}_{(k)}\|_{S_1}. \tag{1.5}$$

We call the norm (1.5) . Intuitively, it penalizes the Schatten 1-norms of the K unfoldings, and minimizing the norm promotes $\mathcal{W}$ to have low-multilinear rank. In fact, it is easy to show (see [54]) the inequality

$$\left|\!\left|\!\left|\mathcal{W}\right|\!\right|\!\right|_{\underline{S_1/1}} \leq \sum_{k=1}^{K} \sqrt{r_k} \left|\!\left|\!\left|\mathcal{W}\right|\!\right|\!\right|_F, \tag{1.6}$$

where $\left|\!\left|\!\left|\mathcal{W}\right|\!\right|\!\right|_F$ is the Frobenius norm $\left|\!\left|\!\left|\mathcal{W}\right|\!\right|\!\right|_F = \sqrt{\langle \mathcal{W}, \mathcal{W} \rangle}$. Thus, tensors that have low multilinear rank (in average) have low overlapped Schatten 1-norm relative to the Frobenius norm.

Another norm proposed in [52, 53] is the

$$\left|\!\left|\!\left|\mathcal{W}\right|\!\right|\!\right|_{\overline{S_1/1}} = \inf_{(\mathcal{W}^{(1)}+\cdots+\mathcal{W}^{(K)})=\mathcal{W}} \sum_{k=1}^{K} \|\boldsymbol{W}_{(k)}^{(k)}\|_{S_1}. \tag{1.7}$$

Here the norm is defined as the infimum over all tuple of K tensors that sums to the original tensor $\mathcal{W}$. It is also easy to relate the latent Schatten 1-norm to the multilinear rank of $\mathcal{W}$. In [53], it was shown that

$$\left|\!\left|\!\left|\mathcal{W}\right|\!\right|\!\right|_{\overline{S_1/1}} \leq \min_k \sqrt{r_k} \left|\!\left|\!\left|\mathcal{W}\right|\!\right|\!\right|_F. \tag{1.8}$$

Note that the sum in inequality (1.6) is replaced by the minimum in inequality (1.8). Therefore, the latent Schatten 1-norm is small when the minimum mode-k rank of $\mathcal{W}$ is small.

1.3 Statistical Guarantees

In this section we present statistical performance guarantee for the estimators defined by the overlapped and latent Schatten 1-norms.

1.3.1 Denoising bounds

The first two theorems concern the denoising performance of the two norms.

Suppose that the observation $\mathcal{Y} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ is obtained as follows:

$$\mathcal{Y} = \mathcal{W}^* + \mathcal{E},$$

where $\mathcal{W}^*$ is the true low-rank tensor with multilinear rank $(r_1, \ldots, r_K)$ and $\mathcal{E} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ is the noise tensor whose entries are independently identically distributed zero-mean Gaussian random variables with variance σ^2.

Define the estimator $\hat{\mathcal{W}}$ by

$$\hat{\mathcal{W}} = \operatorname*{argmin}_{\mathcal{W}} \left(\frac{1}{2} \left|\!\left|\!\left| \mathcal{Y} - \mathcal{W} \right|\!\right|\!\right|_F^2 + \lambda \left|\!\left|\!\left| \mathcal{W} \right|\!\right|\!\right|_{\underline{S_1/1}} \right), \tag{1.9}$$

where $\lambda > 0$ is a regularization parameter.

Then we have the following denoising performance guarantee.

Theorem 1 (Denoising via the overlapped Schatten 1-norm [54]) *There are universal constants $c_i > 0$ $(i = 0, 1)$ such that any minimizer of* (1.9) *with $\lambda = c_0 \frac{\sigma}{K} \sum_{k=1}^{K} (\sqrt{N/n_k} + \sqrt{n_k})$ satisfies the following bound*

$$\frac{1}{N} \left|\!\left|\!\left| \hat{\mathcal{W}} - \mathcal{W}^* \right|\!\right|\!\right|_F^2 \leq c_1 \sigma^2 \left(\frac{1}{K} \sum_{k=1}^{K} (\sqrt{1/n_k} + \sqrt{n_k/N}) \right)^2 \left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{r_k} \right)^2 .$$

with probability at least $1 - \exp(-(\frac{1}{K} \sum_{k=1}^{K} (\sqrt{N/n_k} + \sqrt{n_k}))^2)$.

In particular, if $n_k = n$, the above bound implies the following:

$$\frac{1}{N} \left|\!\left|\!\left| \hat{\mathcal{W}} - \mathcal{W}^* \right|\!\right|\!\right|_F^2 \leq O_p \left(\sigma^2 \frac{\|\boldsymbol{r}\|_{1/2}}{n} \right), \tag{1.10}$$

where $\|\boldsymbol{r}\|_{1/2} := (\frac{1}{K} \sum_{k=1}^{K} \sqrt{r_k})^2$.

In order to state a bound for the latent Schatten 1-norm, we need additional assumptions. Suppose the following observation model

$$\mathcal{Y} = \mathcal{W}^* + \mathcal{E} = \sum_{k=1}^{K} \mathcal{W}^{*(k)} + \mathcal{E},$$

where $\mathcal{W}^* = \sum_{k=1}^{K} \mathcal{W}^{*(k)}$ is the true tensor composed of factors $\mathcal{W}^{*(k)}$ that each are low-rank in the corresponding mode, i.e., $\mathrm{rank}(\boldsymbol{W}_{(k)}^{*(k)}) = \bar{r}_k$. Note that generally $\bar{r}_k$ is different from the mode-k rank of $\mathcal{W}^*$ denoted by r_k. The entries of the noise tensor $\mathcal{E}$ are distributed according to the Gaussian distribution $\mathcal{N}(0, \sigma^2)$ as above. In addition, we assume that the spectral norm of a factor $\mathcal{W}^{*(k)}$ is bounded when unfolded at a different mode as follows:

$$\|\mathcal{W}_{(k')}^{*(k)}\|_{S_\infty} \leq \frac{\alpha}{K} \sqrt{N/n_{k'}} \quad (k \neq k'). \tag{1.11}$$

In other words, we assume that the spectral norm of the kth factor unfolded at the k'th mode is comparable to that of a random matrix for $k' \neq k$; note that the spectral norm of a random $m \times n$ matrix whose entries are independently distributed centered random variables with finite fourth moment scales as $O(\sqrt{m} + \sqrt{n})$ [57]. This means that we want the kth factor $\mathcal{W}^{(k)}$ to look only low-rank in the kth mode as the spectral norm of a low-rank matrix would be larger than a random full rank matrix.

Now let's consider the estimator

$$\hat{\mathcal{W}} = \operatorname*{argmin}_{\mathcal{W}} \left(\frac{1}{2} |\!|\!| \mathcal{Y} - \mathcal{W} |\!|\!|_F^2 + \lambda |\!|\!| \mathcal{W} |\!|\!|_{\overline{S_1/1}} \right.$$
$$\left. \text{s.t. } \mathcal{W} = \sum_{k=1}^{K} \mathcal{W}^{(k)}, \ \| \boldsymbol{W}^{(k)}_{(k')} \|_{S_\infty} \leq \frac{\alpha}{K} \sqrt{N/n_{k'}}, \quad \forall k \neq k' \right). \tag{1.12}$$

The following theorem states the denoising performance of the latent Schatten 1-norm.

Theorem 2 (Denoising via the latent Schatten 1-norm [53]) *There are universal constants $c_i > 0$ $(i = 0, 1)$ such that, any solution of the minimization problem* (1.12) *with regularization constant $\lambda = c_0 \sigma \max_k (\sqrt{N/n_k} + \sqrt{n_k})$ satisfies*

$$\frac{1}{N} \sum_{k=1}^{K} |\!|\!| \hat{\mathcal{W}}^{(k)} - \mathcal{W}^{*(k)} |\!|\!|_F^2 \leq c_1 \sigma^2 \left(\max_k (1/\sqrt{n_k} + \sqrt{n_k/N}) \right)^2 \sum_{k=1}^{K} \bar{r}_k, \tag{1.13}$$

with probability at least $1 - K \exp(-(\max_k(\sqrt{N/n_k} + \sqrt{n_k}))^2)$. Moreover, the total error $\hat{\mathcal{W}} - \mathcal{W}^*$ can be bounded as follows:

$$\frac{1}{N} |\!|\!| \hat{\mathcal{W}} - \mathcal{W}^* |\!|\!|_F^2 \leq c_1 \sigma^2 \left(\max_k (1/\sqrt{n_k} + \sqrt{n_k/N}) \right)^2 \min_k r_k, \tag{1.14}$$

with the same probability as above.

In particular, if $n_k = n$, the above bound implies the following:

$$\frac{1}{N} |\!|\!| \hat{\mathcal{W}} - \mathcal{W}^* |\!|\!|_F^2 \leq O_p \left(\sigma^2 \frac{\min_k r_k}{n} \right) \tag{1.15}$$

Comparing inequalities (1.10) and (1.15), we can see that the bound for the latent approach scales by the minimum mode-k rank, whereas that for the overlap approach scales by the average (square-root) of the mode-k ranks; see [53] for more details.

1.3.2 Tensor recovery guarantee

The next theorem concerns the problem of recovering a low-rank tensor from a small number of linear measurements. Suppose that the observations $\boldsymbol{y} = (y_i)_{i=1}^{M}$ are obtained as in (1.4) with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. In addition, we assume that the entries of the observation operator $\mathfrak{X}$ are drawn independently and identically from standard Gaussian distribution.

Now consider the estimator

$$\hat{\mathcal{W}} = \underset{\mathcal{W}}{\operatorname{argmin}} \left(\frac{1}{2M} \|\boldsymbol{y} - \mathfrak{X}(\mathcal{W})\|_2^2 + \lambda_M \left\|\!\left|\mathcal{W}\right|\!\right\|_{\underline{S_1/1}} \right). \qquad (1.16)$$

The following theorem gives a bound for tensor reconstruction from a small number of noisy measurements.

Theorem 3 (Tensor recovery with the overlapped Schatten 1-norm [54]) *There are universal constants $c_i > 0$ ($i = 0, 1, 2, 3, 4$) such that for a sample size $M \geq c_1(\frac{1}{K}\sum_{k=1}^{K}(\sqrt{N/n_k} + \sqrt{n_k}))^2(\frac{1}{K}\sum_{k=1}^{K}\sqrt{r_k})^2$, any solution $\hat{\mathcal{W}}$ of the minimization problem (1.16) with the regularization constant $\lambda_M = c_0\sigma\big(\frac{1}{K}\sum_{k=1}^{K}(\sqrt{N/n_k} + \sqrt{n_k})\big)/\sqrt{M}$ satisfies the following bound:*

$$\left\|\!\left|\hat{\mathcal{W}} - \mathcal{W}^*\right|\!\right\|_F^2 \leq c_2 \frac{\sigma^2 \left(\frac{1}{K}\sum_{k=1}^{K}(\sqrt{n_k} + \sqrt{N/n_k})\right)^2 (\frac{1}{K}\sum_{k=1}^{K}\sqrt{r_k})^2}{M},$$

with probability at least $1 - c_3 e^{-c_4 M} - \exp(-(\frac{1}{K}\sum_{k=1}^{K}(\sqrt{n_k} + \sqrt{N/n_k}))^2)$.

In particular, if $n_k = n$ the above bound implies the following:

$$\left\|\!\left|\hat{\mathcal{W}} - \mathcal{W}^*\right|\!\right\|_F^2 \leq O_p\left(\frac{\sigma^2 \|\boldsymbol{r}\|_{1/2} n^{K-1}}{M}\right), \qquad (1.17)$$

where $\|\boldsymbol{r}\|_{1/2} := (\frac{1}{K}\sum_{k=1}^{K}\sqrt{r_k})^2$.

The above theorem tells us that the number of samples that we need scales as $O(\|\boldsymbol{r}\|_{1/2} n^{K-1})$. This is rather disappointing because it is only better by a factor $\|\boldsymbol{r}\|_{1/2}/n$ compared to not assuming any low-rank-ness of the underlying truth. This motivates some of the extensions we discuss in Section 1.6.

1.4 Optimization

In this section, we discuss optimization algorithms for overlapped Schatten 1-norm (1.5) and latent Schatten 1-norm (1.7) based on [17].

ADMM is a general technique that can be used whenever splitting makes the problem easier to solve; see [8, 55].

1.4.1 ADMM for the overlapped Schatten 1-norm regularization

We reformulate the overlapped Schatten 1-norm based tensor recovery problem as follows:

$$\underset{\mathcal{W},\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_K}{\text{minimize}} \quad \frac{1}{2\lambda}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2+\sum_{k=1}^{K}\|\boldsymbol{Z}_k\|_{S_1}, \tag{1.18}$$

$$\text{subject to} \quad \boldsymbol{P}_k\boldsymbol{w}=\boldsymbol{z}_k \quad (k=1,\ldots,K). \tag{1.19}$$

Here $\boldsymbol{Z}_k \in \mathbb{R}^{n_k \times N/n_k}$ $(k=1,\ldots,K)$ are auxiliary variables and $\boldsymbol{z}_k$ is the vectorization of $\boldsymbol{Z}_k$. We also denote the vectorization of $\mathcal{W}$ by $\boldsymbol{w}$ and $\boldsymbol{X}\boldsymbol{w} = \mathfrak{X}(\mathcal{W})$. $\boldsymbol{P}_k$ denotes the mode-k unfolding operation; i.e., $\mathrm{vec}(\boldsymbol{W}_{(k)}) = \boldsymbol{P}_k\boldsymbol{w}$. Note that the regularization parameter λ is in the denominator of the loss term. Although dividing the objective by λ does not change the minimizer, it keeps the regularization term from becoming negligible in the limit $\lambda \to 0$; this is useful for dealing with the noiseless case as we explain below.

The augmented Lagrangian function for optimization problem (1.18) can be defined as

$$\begin{aligned}\mathcal{L}(\boldsymbol{w},(\boldsymbol{z}_k)_{k=1}^K,(\boldsymbol{\alpha}_k)_{k=1}^K) &= \frac{1}{2\lambda}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2+\sum_{k=1}^{K}\|\boldsymbol{Z}_k\|_{S_1}\\ &\quad + \eta\sum_{k=1}^{K}\left(\boldsymbol{\alpha}_k{}^\top(\boldsymbol{z}_k-\boldsymbol{P}_k\boldsymbol{w})+\frac{1}{2}\|\boldsymbol{z}_k-\boldsymbol{P}_k\boldsymbol{w}\|_2^2\right),\end{aligned}$$

where $\boldsymbol{\alpha}_k$ is the Lagrange multiplier vector corresponding to the equality constraint $\boldsymbol{z}_k = \boldsymbol{P}_k\boldsymbol{w}$.

The basic idea of ADMM is to minimize the augmented Lagrangian function with respect to $\boldsymbol{w}$ and $(\boldsymbol{z}_k)$ while maximizing it with respect to $(\boldsymbol{\alpha}_k)$. Following a standard derivation (see [8, 55]), we obtain the following iterations (see [52] for the derivation):

$$\begin{cases} \boldsymbol{w}^{t+1} = \left(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\eta K\boldsymbol{I}\right)^{-1}\left(\boldsymbol{X}^\top\boldsymbol{y}+\lambda\eta\sum_{k=1}^{K}\boldsymbol{P}_k{}^\top(\boldsymbol{z}_k^t+\boldsymbol{\alpha}_k^t)\right), \\ \boldsymbol{z}_k^{t+1} = \mathrm{prox}_{1/\eta}\left(\boldsymbol{P}_k\boldsymbol{w}^{t+1}-\boldsymbol{\alpha}_k^t\right) \quad (k=1,\ldots,K), \\ \boldsymbol{\alpha}_k^{t+1} = \boldsymbol{\alpha}_k^t + (\boldsymbol{z}_k^{t+1}-\boldsymbol{P}_k\boldsymbol{w}^{t+1}) \quad (k=1,\ldots,K). \end{cases}$$

Here $\mathrm{prox}_{1/\eta}$ is the proximity operator with respect to Schatten 1-norm and is defined as follows:

$$\mathrm{prox}_\theta(\boldsymbol{z}) = \mathrm{vec}\left(\boldsymbol{U}\max(\boldsymbol{S}-\theta,0)\boldsymbol{V}^\top\right), \tag{1.20}$$

where $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ is the singular-value decomposition (SVD) of $\boldsymbol{Z}$, $\boldsymbol{z}$ is the vectorization of $\boldsymbol{Z}$, and $\theta \geq 0$ is a nonnegative parameter.

The first step can be carried out efficiently, for example, by precomputing the Cholesky factorization of $(\boldsymbol{X}^\top \boldsymbol{X} + \lambda\eta K \boldsymbol{I})$ or linearization (see [60]). Note that assuming $M \leq N$ and $\mathrm{rank}(\boldsymbol{X}) = M$, we can express the limit of the first step as $\lambda \to 0$ as follows:

$$\boldsymbol{w}^{t+1} = \boldsymbol{X}^{+}\boldsymbol{y} + (\boldsymbol{I} - \boldsymbol{X}^{+}\boldsymbol{X})\frac{1}{K}\sum_{k=1}^{K} {\boldsymbol{P}_k}^\top(\boldsymbol{z}_k^t + \boldsymbol{\alpha}_k^t),$$

where $\boldsymbol{X}^{+} := \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}$ is the pseudo inverse of $\boldsymbol{X}$. Taking the limit $\lambda \to 0$ corresponds to solving the noise-free problem

$$\underset{\mathcal{W}}{\text{minimize}} \quad \sum_{k=1}^{K} \|\boldsymbol{W}_{(k)}\|_{S_1} \quad \text{subject to} \quad \boldsymbol{y} = \mathfrak{X}(\mathcal{W}).$$

Putting $1/\lambda$ in front of the loss term allows us to deal with the two problems in the same framework.

In particular, in the case of tensor completion, $\boldsymbol{X}$ is a zero-or-one matrix that has one non-zero entry in every row corresponding to the observed position. In this case, the update can be further simplified as follows:

$$w_i^{t+1} = \begin{cases} (\boldsymbol{X}^\top \boldsymbol{y})_i & \text{(if position } i \text{ is observed)}, \\ (\frac{1}{K}\sum_{k=1}^{K} {\boldsymbol{P}_k}^\top(\boldsymbol{z}_k^t + \boldsymbol{\alpha}_k^t))_i & \text{(otherwise)}. \end{cases}$$

Although careful tuning of the parameter η is not essential for the convergence of the above algorithm, in practice the speed of convergence can be quite different. Here we suggest the following heuristic choice. Consider scaling the truth $\mathcal{W}^*$ and the noise $\boldsymbol{\epsilon}$ by a constant c as $\mathcal{W}'^* = c\mathcal{W}^*$ and $\boldsymbol{\epsilon}' = c\boldsymbol{\epsilon}$. Using $\lambda' = c\lambda$, we get the original solution multiplied by the same constant. Now we require that the process of optimization should also be essentially the same. To this end, we need to scale η inversely as $1/c$ so that all the terms appearing in the augmented Lagrangian function scales linearly against c. Therefore, we choose η as $\eta = \eta_0/\mathrm{std}(\boldsymbol{y})$ where η_0 is a constant and $\mathrm{std}(\boldsymbol{y})$ is the standard deviation of $\boldsymbol{y}$.

As a stopping criterion we use the primal-dual gap; see [52] for details.

1.4.2 ADMM for latent Schatten 1-norm regularization

In this section, we present the ADMM for solving the dual of the latent Schatten 1-norm regularized least squares regression problem:

$$\underset{\mathcal{W}}{\text{minimize}} \quad \frac{1}{2\lambda}\|\boldsymbol{y} - \mathfrak{X}(\textstyle\sum_{k=1}^{K} \mathcal{W}^{(k)})\|_2^2 + \displaystyle\sum_{k=1}^{K} \|\boldsymbol{W}_{(k)}^{(k)}\|_{S_1}. \tag{1.21}$$

The dual problem can be written as follows:

$$\begin{aligned} &\underset{\boldsymbol{\alpha}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_K}{\text{minimize}} \quad \frac{\lambda}{2}\|\boldsymbol{\alpha}\|_2^2 - \boldsymbol{\alpha}^\top \boldsymbol{y} + \sum_{k=1}^{K} \delta_{S_\infty}(\boldsymbol{Z}_k), \\ &\text{subject to} \quad \boldsymbol{z}_k = \boldsymbol{P}_k \boldsymbol{X}^\top \boldsymbol{\alpha} \qquad (k = 1, \ldots, K), \end{aligned} \tag{1.22}$$

where δ_{S_∞} is the indicator function of the unit spectral norm ball, i.e.,

$$\delta_{S_\infty}(\boldsymbol{Z}) = \begin{cases} 0 & (\text{if } \|\boldsymbol{Z}\|_{S_\infty} \leq 1), \\ +\infty & (\text{otherwise}). \end{cases}$$

The augmented Lagrangian function can be written as follows:

$$\begin{aligned}\mathcal{L}_\eta\left(\boldsymbol{\alpha}, (\boldsymbol{Z}_k), (\boldsymbol{W}_k)\right) =& \frac{\lambda}{2}\|\boldsymbol{\alpha}\|_2^2 - \boldsymbol{\alpha}^\top \boldsymbol{y} + \sum_{k=1}^{K} \delta_{S_\infty}(\boldsymbol{Z}_k) \\ &+ \sum_{k=1}^{K} \left(\boldsymbol{w}_k{}^\top (\boldsymbol{P}_k \boldsymbol{X}^\top \boldsymbol{\alpha} - \boldsymbol{z}_k) + \frac{\eta}{2} \|\boldsymbol{P}_k \boldsymbol{X}^\top \boldsymbol{\alpha} - \boldsymbol{z}_k\|_2^2 \right),\end{aligned}$$

where $\boldsymbol{W}_k$ $(k = 1, \ldots, K)$ is the Lagrange multiplier vector corresponding to the equality constraint (1.22) and equals the mode-k unfolding of primal variable $\mathcal{W}^{(k)}$ at the optimality.

The iterations can be derived as follows (see [52] for details):

$$\begin{cases} \boldsymbol{w}_k^{t+1} = \mathrm{prox}_\eta \left(\boldsymbol{w}_k^t + \eta \boldsymbol{P}_k \boldsymbol{X}^\top \boldsymbol{\alpha}^t \right), \\ \boldsymbol{z}_k^{t+1} = (\boldsymbol{w}_k^t + \eta \boldsymbol{P}_k \boldsymbol{X}^\top \boldsymbol{\alpha}^t - \boldsymbol{w}_k^{t+1})/\eta, \\ \boldsymbol{\alpha}^{t+1} = \frac{1}{\lambda + \eta K} \left(\boldsymbol{y} + \eta \boldsymbol{X} \sum_{k=1}^K \boldsymbol{P}_k{}^\top (\boldsymbol{z}_k^{t+1} - \boldsymbol{w}_k^{t+1}/\eta) \right), \end{cases}$$

where prox_η is the proximity operator (1.20).

We can see that the algorithm updates the dual variables ($(\boldsymbol{z}_k)$ and $\boldsymbol{\alpha}$) and the primal variables $(\boldsymbol{w}_k)$ alternately. In particular, the update equation for the primal variables $(\boldsymbol{w}_k)$ is a popular proximal-gradient-type update. In fact, $\boldsymbol{P}_k \boldsymbol{X}^\top \boldsymbol{\alpha}^t$ converges to the gradient of the loss term at the optimality.

Note that setting $\lambda = 0$ gives the correct update equations for the noiseless case $\lambda \to 0$ in (1.21).

Consideration on the scale invariance of the algorithm similar to that in the previous subsection suggests that we should scale η linearly as the scale of $\boldsymbol{y}$; thus we set $\eta = \eta_0 \mathrm{std}(\boldsymbol{y})$.

1.5 Experiments

1.5.1 Tensor denoising

We generated synthetic problems as follows. First each entry of the core tensor $\mathcal{C}$ was sampled independently from standard normal distribution. Then orthogonal factors drawn from the uniform (Haar) measure were multiplied to each of its modes to obtain the true tensor $\mathcal{W}^*$. Then the observed tensor $\mathcal{Y}$ was obtained by adding zero-mean Gaussian noise with standard deviation $\sigma = 0.1$ to each entry.

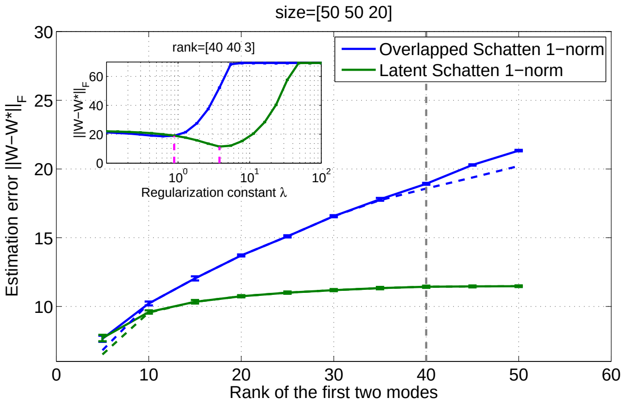

FIGURE 1.1: Estimation of a low-rank 50×50×20 tensor of rank $r \times r \times 3$ from noisy measurements. The noise standard deviation was $\sigma = 0.1$. The estimation errors of overlapped and latent approaches are plotted against the rank r of the first two modes. The solid lines show the error at the fixed regularization constant λ, which was 0.89 for the overlapped approach and 3.79 for the latent approach. The dashed lines show the minimum error over candidates of the regularization constant λ from 0.1 to 100. In the inset, the errors of the two approaches are plotted against the regularization constant λ for rank $r = 40$ (marked with vertical gray dashed line in the outset). The two values (0.89 and 3.79) are marked with vertical dashed lines.

The two approaches (overlap and latent Schatten 1-norms) were applied with different values of the regularization parameter λ ranging from 0.01 to 100. The incoherence parameter α for the latent Schatten 1-norm was set to a sufficiently large constant value so that it had no effect on the solution.

Figure 1.1 shows the result of applying the two approaches to tensors of multilinear rank $(r, r, 3)$ for different r. This experiment was specifically designed to highlight the dependency of the denoising performance of the two methods. The error of the overlapped Schatten 1-norm increases as r increases although the rank of the third mode is constant; this is because the right-hand side of (1.10) depends on the average (square-root) of multilinear ranks. On the other hand, the error of the latent Schatten 1-norm stays almost constant; this is because the minimum multilinear rank 3 is constant; see Theorem 2. Of course, this is just one well constructed example, and we refer the readers to [53] for more results that quantitatively validate Theorem 2.

1.5.2 Tensor completion

A synthetic problem was generated as follows. The true tensor $\mathcal{W}^*$ was generated the same way as in the previous subsection. Then we randomly split the entries into training and testing. No observational noise was added.

We trained overlapped and latent Schatten 1-norms using the optimization algorithms discussed in the previous section. The operator $\mathfrak{X}$ was defined as

$$\mathfrak{X}(\mathcal{W}) = (\mathcal{W}_{i_s j_s k_s})_{s=1}^{M},$$

where $(i_s, j_s, k_s)_{s=1}^{M}$ is the set of indices corresponding to the observed positions. Since there is no observational noise, we took the limit $\lambda \to 0$ in the update equations.

The result for $50 \times 50 \times 20$ tensor of multilinear rank (7,8,9) is shown in Figure 1.2. As baselines we included an expectation-maximization-based Tucker decomposition algorithm in [4] with the correct rank (exact) and 20% higher rank (large). We also included matrix completion algorithm that treated a tensor as a matrix by unfolding the tensor at a prespecified mode. This method was implemented by instantiating only one of the auxiliary variables $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_K$ in the ADMM for overlapped Schatten 1-norm presented in Section 1.4.1.

The result shows that first, treating tensor as a matrix yields a rather disappointing result, especially when we choose mode 3. This is because the dimensions are not balanced, which is often the case in practice, and unluckily the mode with the smallest dimension (mode 3) has the highest rank. On the other hand, the overlapped Schatten 1-norm can recover this tensor reliably from about 35% of the entries *without any assumption about the low-rank-ness of the modes.*

Second, the reconstruction is exact (up to optimization tolerance) above the sufficient sampling density (35%). This can be predicted from Theorem 3 in the following way: first note that the condition for the sample size M does not depend on the noise variance σ^2; second, the right-hand side of the bound is proportional to the noise variance σ^2. Therefore, if we take the limit $\sigma^2 \to 0$, the theorem predicts zero error whenever the condition for the sample size M is satisfied. We would need a lower bound to make this claim more precise, which may be obtained by following the work of [2].

Compared to the overlapped approach, the latent approach recovers the true tensor exactly only around 70% observation. Although we don't have a theory for tensor recovery via the latent approach, it seems to suggest that the number of samples that we need scales faster than the minimum multilinear rank, which appeared in the right-hand side of the denoising bound (1.14).

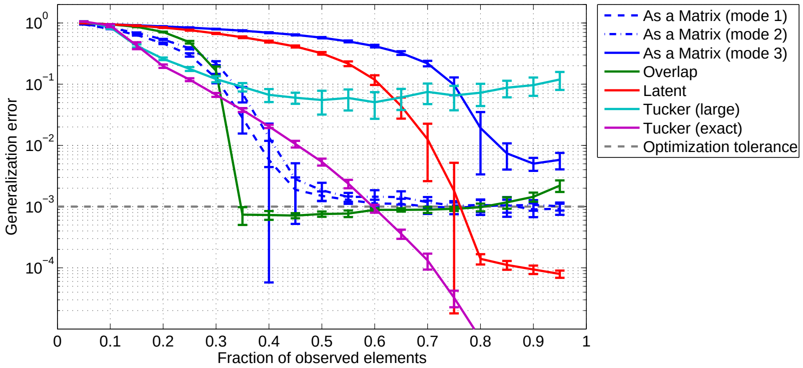

FIGURE 1.2: Comparison of tensor completion performance of overlapped and latent Schatten 1-norm regularization. As baselines, Tucker decomposition with the correct rank (exact) and 20% higher rank (large), and convex optimization based matrix completion (as a matrix) that focuses on a pre-specified mode are included. The size of the tensor is $50 \times 50 \times 20$ and the true multilinear rank is $(7, 8, 9)$. The generalization error is plotted against the fraction of observed elements (M/N) of the underlying low-rank tensor. Also the tolerance of optimization (10^{-3}) is shown.

1.6 Extensions and related work

1.6.1 Balanced unfolding

For a balanced-sized K-way tensor (i.e., $n_k = n$), CP decomposition (1.1) or Tucker decomposition (1.2) has only linearly many parameters in n. Thus we would expect that a reasonable estimator would decrease the error as $O(n/M)$. However, the scaling we see in inequality (1.17) is $O(n^{K-1}/M)$, which is far larger than what we expect.

Looking at the way the bound is derived, we notice (we thank Nam H. Nguyen for pointing this out) that the unbalancedness of the unfolding is the cause. More specifically, the term $\sqrt{n_k} + \sqrt{N/n_k}$ is the spectral norm of a random $n_k \times N/n_k$ matrix with independent centered entries with bounded fourth moment [57]. Thus, we can ask what happens if we unfold the tensor evenly.

Let $\boldsymbol{W}_{(i_1,i_2,\ldots,i_k;j_1,j_2,\ldots,j_l)}$ denote the $\prod_{a=1}^{k} n_{i_a} \times \prod_{b=1}^{l} n_{j_b}$ matrix obtained by concatenating the $\prod_{a=1}^{k} n_{i_a}$ dimensional slices of $\mathcal{W}$ specified by indices in $[n_{j_1}] \times \cdots \times [n_{j_l}]$ along columns. For example, $\boldsymbol{W}_{(1;2,3,4)}$ is the same as $\boldsymbol{W}_{(1)}$ in the original notation defined in Section 1.2. We say that an unfolding

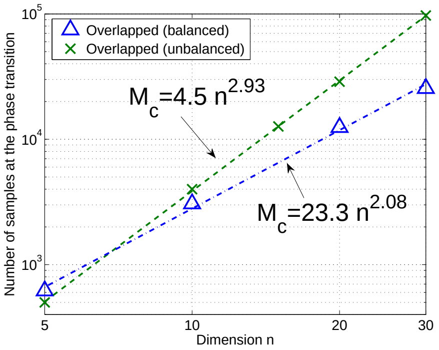

FIGURE 1.3: The number of samples necessary to recover a $n \times n \times n \times n$ tensor of multilinear rank $(2, 2, 2, 2)$. The number of samples at the phase transition M_c was defined as the number of samples at which the empirical probability of obtaining error smaller than 0.01 exceeded 1/2.

is balanced if the number of rows and columns are the same, e.g., $\boldsymbol{W}_{(1,2;3,4)}$ when $n_k = n$.

Figure 1.3 shows the number of samples at the phase transition M_c against n for the completion of 4th order balanced-sized tensors. We compared the original overlapped Schatten 1-norm (1.5) against the following norm based on three balanced unfoldings

$$|||\mathcal{W}|||_{\text{balanced}} = \|\boldsymbol{W}_{(1,2;3,4)}\|_{S_1} + \|\boldsymbol{W}_{(1,3;2,4)}\|_{S_1} + \|\boldsymbol{W}_{(1,4;2,3)}\|_{S_1}.$$

See Mu et al. [37] for a related approach, though they only considered one of the three possible balanced unfoldings.

The threshold M_c was defined as the number of samples at which the probability that the reconstruction error $|||\hat{\mathcal{W}} - \mathcal{W}^*|||_F$ was smaller than 0.01 exceeded 1/2. The dashed line corresponds to the original overlapped Schatten 1-norm (one mode against the rest). The dash-dotted line corresponds to the overlapped Schatten 1-norm based on balanced unfoldings.

We can see that the empirical scaling of the balanced version is $n^{2.08}$, whereas that of the ordinary version is $n^{2.93}$. Both of them were close to the theoretically predicted scaling n^2 and n^3, respectively.

However, computationally this approach is more challenging. The major computational cost for optimization is that of SVD. Since SVD scales as $O(m^2n + m^3)$ for an $m \times n$ matrix with $m \leq n$, the more balanced, the more challenging the computation becomes. Note that the comparison here is made assuming that both approaches use the same ADMM-based optimization algorithm (see Section 1.4.1). Thus there might be another optimization algorithm (see e.g., Jaggi [25]) that work better in the balanced case.

Recently Mu et al. [37] derived a lower-bound for the overlapped Schatten 1-norm based on the framework developed by Amelunxen et al. [2]. The lower-bound indeed shows that rn^{K-1} samples is unavoidable for the vanilla version of the overlapped Schatten 1-norm. Motivated by the lower bound, they proposed a balanced version (without overlap), which they call the .

1.6.2 Tensor nuclear norm

Chandrasekaran et al. [12] discuss a norm for tensors within the framework of . Let $\mathcal{A}$ be an atomic set that consists of rank one tensors of unit Frobenius norm:

$$\mathcal{A} = \{\boldsymbol{u}_1 \circ \boldsymbol{u}_2 \circ \cdots \circ \boldsymbol{u}_K : \|\boldsymbol{u}_k\| = 1 \quad (k = 1, \ldots, K)\}.$$

The is defined as follows:

$$|\!|\!|\mathcal{W}|\!|\!|_{\mathrm{nuc}} = \inf \sum_{a \in \mathcal{A}} c_a \quad \text{s.t.} \quad \mathcal{W} = \sum_{a \in \mathcal{A}} c_a \boldsymbol{u}_1^{(a)} \circ \cdots \circ \boldsymbol{u}_K^{(a)},$$

where with a slight abuse of notation, we use $a \in \mathcal{A}$ as an index for an element in the atomic set.

It can be shown that for a tensor that admits an [28] with R terms (decomposition (1.1) with orthogonality constraints between the components), the nuclear norm can be related to the Frobenius norm as follows:

$$|\!|\!|\mathcal{W}|\!|\!|_{\mathrm{nuc}} \leq \sqrt{R} |\!|\!|\mathcal{W}|\!|\!|_F.$$

Moreover, the tensor spectral norm

$$|\!|\!|\mathcal{X}|\!|\!|_{\mathrm{op}} = \max_{a \in \mathcal{A}} \mathcal{X} \times_1 \boldsymbol{u}_1^{(a)} \times_2 \boldsymbol{u}_2^{(a)} \cdots \times_K \boldsymbol{u}_K^{(a)},$$

which is dual to the nuclear norm, is known to be of order $O(\sqrt{n})$ for a random Gaussian tensor; see [40]. Thus it is natural to hope that we can prove that the nuclear norm would achieve an optimal $O(Rn)$ convex relaxation for tensors. However, computationally, the tensor nuclear norm seems to be intractable for $K \geq 3$. Although it is convex, it involves infinitely many variables. There is no analogue of linear matrix inequality or semidefinite programming for matrices that can be used here to the best of our knowledge.

1.6.3 Interpretation of the result

That we can bound the error in Frobenius norm as we have presented in Section 1.3 does not mean that our method is useful in practice. In fact, tensor decomposition methods are often used to uncover latent factors and gain insight about the data.

Here we present how such an insight can be gained from the solutions of the two algorithms we presented in Section 1.4.

For the overlapped approach, the factor matrices $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_K$ corresponding to the Tucker decomposition (1.2) can be obtained by computing the left singular vectors of the auxiliary matrices $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_K$. The mode-$k$ rank r_k is determined *automatically* by the proximity operator (1.20); importantly, the rank at an optimum does not depend on the choice of η, though the rank during optimization may depend on η. Once the factors are obtained, the core can be obtained as follows:

$$\mathcal{C} = \mathcal{W} \times_1 \boldsymbol{U}_1{}^\top \times_2 \boldsymbol{U}_2{}^\top \cdots \times_K \boldsymbol{U}_K{}^\top.$$

To get the stronger CP decomposition (1.1), one can perform any off-the-shelf CP decomposition algorithm on the *core* $\mathcal{C}$. This post-processing step is easier than applying CP decomposition directly to the original large tensor with noise and missing entries. In other words, this two-step approach allows us to separate the tasks of generalization and interpretation; see [52] for details.

The latent approach is less easier to interpret the solution because in general the sum $\mathcal{W} = \sum_{k=1}^{K} \mathcal{W}^{(k)}$ is not low-rank even when each $\mathcal{W}^{(k)}$ is. However in practice we found that the solution is often singleton, i.e., only one non-zero component $\mathcal{W}^{(k)}$. This corresponds to the intuition that the latent Schatten 1-norm focuses on the low-rank-ness of the mode with the minimum mode-k rank and do not care about the other modes. The fact that the solution is only low-rank in one mode is still disappointing. This could be solved by including more terms in the latent approach, which can be partially low-rank (there are 2^K possibilities to penalize the sum of the Schatten 1-norms of some of the modes) or balanced unfolding. If the resulting solution is a singleton, then the model automatically chose which mode should be low-rank.

1.6.4 Related work

Liu et al. [33, 34] proposed the overlapped approach in the context of image and video imputation. They used a penalty method to deal with the equality constraints in (1.19). Li et al. [32] extended Liu et al.'s work to sparse+low-rank decomposition of tensors (also known as sparse PCA) and applied to background/shadow removal and face recognition. Li et al. also used a penalty method for the optimization.

Signoretto et al. [47, 49, 48, 46] proposed and extended the overlapped Schatten 1-norm in the context of kernel-based learning, i.e., learning higher-order operators over Hilbert spaces. The use of kernel allows us to incorporate smoothness (or side-information) when we see an entry in a K-way tensor as a representation of a relation among objects from K different domains; see also [1]. They also proposed an optimization algorithm that supports general differentiable loss function L based on an (accelerated) proximal gradient method [39, 6]; the algorithm employs ADMM to compute the proximal operator corresponding to the overlapped Schatten 1-norm.

Gandy et al. [18] proposed ADMM and Douglas-Rachford splitting algorithm for the overlapped approach.

Yang et al. [58] proposed a fast optimization algorithm for the overlapped approach based on a fixed-point iteration combined with continuation.

Goldfarb and Qin [19] studied low-rank+sparse tensor decomposition based on the overlapped and latent Schatten 1-norms. They also proposed adaptive weighting of the terms appearing in the overlapped Schatten 1-norm (1.5) and reported that the adaptive version outperformed other methods in many cases. They also studied the relationship between the normalized rank $\|\boldsymbol{n}^{-1}\|_{1/2}\|\boldsymbol{r}\|_{1/2}$ (the quantity that appears in the condition for M in Theorem 3), the necessary sampling density, and the allowable fraction of corrupted entries.

Zhang et al. [61] extended Li et al.'s work [32] on sparse+low-rank decomposition in several interesting ways. They have incorporated transformations that align each image in order to make the spatial low-rank assumption (on the first two modes) as valid as possible (see also [36] for related work), while keeping the sequence of images smooth by also penalizing the Schatten 1-norm for the mode corresponding to the temporal dimension.

On the theoretical side, Nickel and Tresp [41] presented a generalization bound for low multilinear rank tensor in the context of relational data analysis by counting the number of possible sign patterns that low multilinear rank tensors can attain. Although the theory does not lead to a model selection criterion as we cannot provably compute the low-multilinear-rank decomposition at a given rank, yet it would be fruitful to study a convex relaxation for the set of low-rank sign tensors; see e.g., [50].

Romera-Paredes and Pontil [45] proposed a convex relaxation of mode-k rank with respect to the Frobenius norm ball and showed that it is tighter than the overlapped Schatten 1-norm at some points. Although the resulting regularizer is not a continuous function and thus challenging to compute, they proposed a subgradient-based optimization algorithm.

Jiang et al. [26] studied the best rank-one approximation of super symmetric even order tensors. They noticed that for super symmetric (meaning that the tensor is invariant to arbitrary permutation of indices) even order tensors, being rank-one is equivalent to a balanced unfolding (see Section 1.6.1) being rank-one (as a matrix). Then they solved the best rank-one approximation problem with the Schatten 1-norm regularization (which promotes rank-one solution) under linear equality constraints that ensured that the solution was super-symmetric. Empirically the solution was found to be rank one in most of the cases. They also showed many non-symmetric problems can be reduced to the symmetric case.

Krishnamurthy and Singh [30] proposed an adaptive sampling algorithm for tensor completion and showed that it succeeds with high probability with $O(n)$ samples. The number of samples required in their adaptive setting also depends on the true rank r and the coherence parameter μ_0. They also showed a lower bound that scales as $O(r^{K-1}n)$ under the incoherence assumption.

Application of the overlapped approach includes language models [24], hyper-spectral imaging [49], and multi-task learning [46, 44], besides image reconstruction discussed in [33, 34, 32, 61].

1.7 Future directions

Compared to the overlapped Schatten 1-norm, the behavior of the latent Schatten 1-norm is still unclear in some parts. First, although we have argued that empirically the solution of the optimization problem is *often* a singleton (only one non-zero component), this needs a better explanation. Second, although we believe that the incoherence assumption is necessary to prove the stronger inequality (1.13), it may not be necessary to obtain the weaker one (1.14).

Given that the sample complexities of both the overlapped and latent Schatten 1-norms are far from optimal, it would be extremely interesting to explore the statistics-computation trade-off between what we can provably achieve and how much it would be computationally expensive. Balanced unfolding [37], tensor nuclear norm [12], and the new convex relaxation [45] discussed in the previous section are candidates to be evaluated and analyzed. It would also be interesting to study recent work on decomposition of tensors arising from higher order moments of latent variable models [3] in this context.

Finally, nonnegativity [9, 31] and positive semidefiniteness [59] are constraints that are useful to impose on the factors in practice. Generalization of the results for separable nonnegative matrix factorization [5, 7] to tensors would be an interesting direction.

Acknowledgements

The authors would like to thank Franz J. Király, Nati Srebro, Nam H. Nguyen, and Gilles Blanchard for discussions and the editors of this book for comments. This work was partially supported by MEXT Kakenhi 25730013, and JST, CREST.

Bibliography

[1] J. Abernethy, F. Bach, T. Evgeniou, and J.P. Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *J. Mach. Learn. Res.*, 10:803–826, 2009.

[2] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. Technical report, arXiv:1303.6672, 2013.

[3] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. Technical report, arXiv:1210.7559, 2012.

[4] C. A. Andersson and R. Bro. The n-way toolbox for matlab. *Chemometr. Intell. Lab.*, 52(1):1–4, 2000. http://www.models.life.ku.dk/source/nwaytoolbox/.

[5] S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization–provably. In *Proceedings of the 44th symposium on Theory of Computing*, pages 145–162, 2012.

[6] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[7] V. Bittorf, B. Recht, C. Ré, and J. A. Tropp. Factoring nonnegative matrices with linear programs. In *Adv. Neural. Inf. Process. Syst. 25*, pages 1223–1231. 2012.

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[9] R. Bro and S. De Jong. A fast non-negativity-constrained least squares algorithm. *J. Chemometr.*, 11(5):393–401, 1997.

[10] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE T. Inform. Theory*, 56(5):2053–2080, 2010.

[11] J.D. Carroll and J.J. Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319, 1970.

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

[13] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278, 2000.

[14] L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1 and rank-$(R_1, R_2, \ldots, R_N)$ approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.*, 21(4):1324–1342, 2000.

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

[16] M. Fazel, H. Hindi, and S. P. Boyd. A Rank Minimization Heuristic with Application to Minimum Order System Approximation. In *Proc. of the American Control Conference*, 2001.

[17] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.*, 2(1):17–40, 1976.

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

[19] D. Goldfarb and Z. Qin. Robust low-rank tensor recovery: Models and algorithms. Technical report, arXiv:1311.6182, 2013.

[20] R. A. Harshman. Models for analysis of asymmetrical relationships among n objects or stimuli. In *First Joint Meeting of the Psychometric Society and the Society for Mathematical Psychology, McMaster University, Hamilton, Ontario*, 1978.

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

[22] F. L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *J. Math. Phys.*, 6(1):164–189, 1927.

[23] J. Håstad. Tensor rank is NP-complete. *Journal of Algorithms*, 11(4):644–654, 1990.

[24] B. Hutchinson, M. Ostendorf, and M. Fazel. Low rank language models for small training sets. *IEEE Signal Proc. Let.*, 18(9):489–492, 2011.

[25] M. Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, pages 427–435, 2013.

[26] B. Jiang, S. Ma, and S. Zhang. Tensor principal component analysis via convex optimization. Technical report, arXiv:1212.2702, 2012.

[27] R. Kannan and S. Vempala. Spectral algorithms. *Foundations and Trends in Theoretical Computer Science*, 4(3–4):157–288, 2008.

[28] T. G. Kolda. Orthogonal tensor decompositions. *SIAM J. Matrix Anal. Appl.*, 23(1):243–255, 2001.

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

[30] A. Krishnamurthy and A. Singh. Low-rank matrix and tensor completion via adaptive sampling. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Adv. Neural. Inf. Process. Syst. 26*, pages 836–844. 2013.

[31] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[32] Y. Li, J. Yan, Y. Zhou, and J. Yang. Optimum subspace learning and error correction for tensors. In *Computer Vision–ECCV 2010*, pages 790–803. Springer, 2010.

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

[34] J. Liu, J. Ye, P. Musialski, and P. Wonka. Tensor completion for estimating missing values in visual data. *IEEE T. Pattern. Anal.*, 35(1):208–220, 2013.

[35] M. Mørup. Applications of tensor (multiway array) factorizations and decompositions in data mining. *Wiley Interdisciplinary Rev.: Data Min. Knowl. Dicov.*, 1(1):24–40, 2011.

[36] M. Mørup, L. K. Hansen, S. M. Arnfred, L.-H. Lim, and K. H. Madsen. Shift-invariant multilinear decomposition of neuroimaging data. *Neuroimage*, 42(4):1439–1450, 2008.

[37] C. Mu, B. Huang, J. Wright, and D. Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. *arXiv preprint arXiv:1307.5870*, 2013.

[38] S. Negahban and M.J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.*, 39(2):673–1333, 2011.

[39] Y. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 2007/76, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, 2007.

[40] N. H. Nguyen, P. Drineas, and T. D. Tran. Tensor sparsification via a bound on the spectral norm of random tensors. Technical report, arXiv:1005.4732, 2010.

[41] M. Nickel and V. Tresp. An analysis of tensor models for learning on structured data. In *Machine Learning and Knowledge Discovery in Databases*, pages 272–287. Springer, 2013.

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

[43] A. Rohde and A. B. Tsybakov. Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, 39(2):887–930, 2011.

[44] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil. Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1444–1452, 2013.

[45] B. Romera-Paredes and M. Pontil. A new convex relaxation for tensor completion. In *Adv. Neural. Inf. Process. Syst. 26*, pages 2967–2975, 2013.

[46] M. Signoretto, L. De Lathauwer, and J. A. K. Suykens. Learning tensors in reproducing kernel hilbert spaces with multilinear spectral penalties. Technical report, arXiv:1310.4977, 2013.

[47] M. Signoretto, L. De Lathauwer, and J.A.K. Suykens. Nuclear norms for tensors and their use for convex multilinear estimation. Technical Report 10-186, ESAT-SISTA, K.U.Leuven, 2010.

[48] M. Signoretto, Q. T. Dinh, L. De Lathauwer, and J. A. K. Suykens. Learning with tensors: a framework based on convex optimization and spectral regularization. *Mach. Learn.*, 94(3):303–351, 2013.

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

[50] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *Proc. of the 18th Annual Conference on Learning Theory (COLT)*, pages 545–560. Springer, 2005.

[51] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B*, 58(1):267–288, 1996.

[52] R. Tomioka, K. Hayashi, and H. Kashima. Estimation of low-rank tensors via convex optimization. Technical report, arXiv:1010.0789, 2011.

[53] R. Tomioka and T. Suzuki. Convex tensor decomposition via structured schatten norm regularization. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Adv. Neural. Inf. Process. Syst. 26*, pages 1331–1339. 2013.

[54] R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima. Statistical performance of convex tensor decomposition. In *Adv. Neural. Inf. Process. Syst. 24*, pages 972–980. 2011.

[55] R. Tomioka, T. Suzuki, and M. Sugiyama. Augmented lagrangian methods for learning, selecting, and combining features. In Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright, editors, *Optimization for Machine Learning*. MIT Press, 2011.

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

[57] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. Technical report, arXiv:1011.3027, 2010.

[58] L. Yang, Z Huang, and X. Shi. A fixed point iterative method for low n-rank tensor pursuit. *IEEE T. Signal Proces.*, 61(11):2952–2962, 2013.

[59] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto. Infinite positive semidefinite tensor factorization for source separation of mixture signals. In *Proceedings of the 30th International Conference on Machine Learning*, pages 576–584, 2013.

[60] X. Zhang, M. Burger, and S. Osher. A unified primal-dual algorithm framework based on bregman iteration. *J. Sci. Comput.*, 46(1):20–46, 2010.

[61] X. Zhang, D. Wang, Z. Zhou, and Y. Ma. Simultaneous rectification and alignment via robust recovery of low-rank tensors. In *Adv. Neural. Inf. Process. Syst. 26*, pages 1637–1645, 2013.

Index