Convex Tensor Decomposition via Structured Schatten Norm Regularization

Abstract

We discuss structured Schatten norms for tensor decomposition that includes two recently proposed norms ("overlapped" and "latent") for convex-optimization-based tensor decomposition, and connect tensor decomposition with wider literature on structured sparsity. Based on the properties of the structured Schatten norms, we mathematically analyze the performance of "latent" approach for tensor decomposition, which was empirically found to perform better than the "overlapped" approach in some settings. We show theoretically that this is indeed the case. In particular, when the unknown true tensor is low-rank in a specific mode, this approach performs as good as knowing the mode with the smallest rank. Along the way, we show a novel duality result for structures Schatten norms, establish the consistency, and discuss the identifiability of this approach. We confirm through numerical simulations that our theoretical prediction can precisely predict the scaling behaviour of the mean squared error.

1. Introduction

Decomposition of tensors (Kolda & Bader, 2009) (or multi-way arrays) into low-rank components arises naturally in many real world data analysis problems. For example, in neuroimaging, we are often interested in finding spatio-temporal patterns of neural activities that are related to certain experimental conditions or subjects; one way to do this is to compute the decomposition of the data tensor, which can be of size channels × time-points × subjects × conditions (Mørup, 2011). In computer vision, an ensemble of face images can be collected into a tensor of size pixels × subjects × illumination × viewpoints; the decomposition of this tensor yields the so called tensorfaces (Vasilescu & Terzopoulos, 2002), which can be regarded as a multi-linear generalization of eigenfaces (Sirovich & Kirby, 1987). Conventionally tensor decomposition has been tackled through non-convex optimization problems, using alternate least squares or higher order orthogonal iteration (De Lathauwer et al., 2000). Although being successful in many application areas, the statistical performance of such approaches has been widely open. Moreover, the model selection problem can be highly challenging, especially for the so called Tucker model (Tucker, 1966; De Lathauwer et al., 2000), because we need to specify the rank \( r_k \) for each mode (here a mode refers to one dimensionality of a tensor); that is, we have \( K \) hyper-parameters to choose for a \( K \)-way tensor, which is challenging even for \( K = 3 \).

Recently a convex-optimization-based approach for tensor decomposition has been proposed by several authors (Signoretto et al., 2010; Gandy et al., 2011; Liu et al., 2009; Tomioka et al., 2011a), and its performance has been analyzed in (Tomioka et al., 2011b). The basic idea behind their convex approach, which we call overlapped approach, is to unfold a tensor into matrices along different modes and penalize the un-

\(^1\)For a \( K \)-way tensor, there are \( K \) ways to unfold a tensor into a matrix. See Section 2.

Figure 1. Schematic illustrations of the overlapped approach and the latent approach for the decomposition of a three way tensor (\( K = 3 \).)
Tensor Decomposition via Structured Schatten Norm Regularization

Figure 2. Estimation of a low-rank $50 \times 50 \times 20$ tensor of rank $r \times r \times 3$ from noisy measurements. The noise standard deviation is $\sigma = 0.1$. The estimation errors of two convex optimization based methods are plotted against the rank $r$ of the first two modes. The solid lines show the error at the fixed regularization constant $\lambda$, which is 0.89 for the overlapped approach and 3.79 for the latent approach (see also Figure 3). The dashed lines show the minimum error over candidates of the regularization constant $\lambda$ from 0.1 to 100. In the inset, the errors of the two approaches are plotted against the regularization constant $\lambda$ for rank $r = 40$ (marked with gray dashed vertical line in the outset). The two values (0.89 and 3.79) are marked with vertical dashed lines. Note that both approaches need no knowledge of the true rank; the rank is automatically learned.

folded matrices to be *simultaneously low-rank* based on the Schatten 1-norm, which is also known as the trace norm and nuclear norm (Fazel et al., 2001; Srebro et al., 2005; Recht et al., 2010); see the left panel of Figure 1. The convex approach does not require the rank of the decomposition to be specified beforehand, and due to the low-rank inducing property of the Schatten 1-norm, the rank of the decomposition is automatically determined.

However, it has been noticed that the above overlapped approach has a limitation that it performs poorly for a tensor that is only low-rank in a certain mode (Tomioka et al., 2011a). They proposed an alternative approach, which we call *latent approach*, that decomposes a given tensor into a mixture of tensors that each are low-rank in a specific mode; see the right panel of Figure 1. Figure 2 demonstrates that the latent approach is preferable to the overlapped approach when the underlying tensor is almost full rank in all but one mode.

However, there are two issues that are not properly addressed so far.

The first issue is the statistical performance of the latent approach. In this paper, we show that the mean squared error of the latent approach scales no greater than the minimum mode-$k$ rank of the underlying true tensor, which clearly explains why the latent approach suffers less than the overlapped approach in Figure 2.

The second issue is the identifiability of the model underlying the latent approach, i.e., a mixture of low-rank tensors. In this paper, we show that such a mixture is identifiable only when the mixture consists of one component; in other words, when the underlying tensor is low-rank in a specific mode.

Along the way, we show a novel duality between the two types of norms employed in the above two approaches, namely the overlapped Schatten norm and the latent Schatten norm. This result is closely related and generalize the results in structured sparsity literature (Bach et al., 2011; Jenatton et al., 2011; Obozinski et al., 2011; Maurer & Pontil, 2011). In fact, the (plain) overlapped lasso constrains the weights to be simultaneously group sparse over overlapping groups. The latent group lasso predicts with a mixture of group sparse weights (see also Wright et al., 2010; Jalali et al., 2010; Agarwal et al., 2011). These approaches clearly correspond to the two variations of tensor decomposition algorithms we discussed above.

Finally we empirically compare the overlapped approach and latent approach and show that even when the unknown tensor is simultaneously low-rank, which is a favorable situation for the overlapped approach, the latent approach performs better in many cases. Thus we provide both theoretical and empirical evidence that for noisy tensor decomposition, the latent approach is preferable to the overlapped approach. Our result is complementary to the previous study (Tomioka et al., 2011a:b), which mainly focused on the noise-less tensor completion setting.

This paper is structured as follows. In Section 2, we provide basic definitions of the two variations of structured Schatten norms, namely the overlapped/latent Schatten norms, and discuss their properties, especially the duality between them. Section 3 presents our main theoretical contributions; we establish the consistency of the latent approach, we show a denoising performance bound, and discuss the identifiability of the model underlying it. In Section 4, we empirically confirm the scaling predicted by our theory. Finally, Section 5 concludes the paper.
2. Structured Schatten norms for tensors

In this section, we define the overlapped Schatten norm and the latent Schatten norm and discuss their basic properties.

First we need some basic definitions.

Let \( W \in \mathbb{R}^{n_1 \times \cdots \times n_K} \) be a \( K \)-way tensor. We denote the total number of entries in \( W \) by \( N = \prod_{k=1}^K n_k \). The dot product between two tensors \( W \) and \( X \) is defined as \( \langle W, X \rangle = \text{vec}(W)^\top \text{vec}(X) \); i.e., the dot product as vectors in \( \mathbb{R}^N \). The Frobenius norm of a tensor is defined as \( \| W \|_F = \sqrt{\langle W, W \rangle} \). Each dimensionality of a tensor is called a mode. The mode \( k \) unfolding \( \bar{W}(k) \in \mathbb{R}^{n_k \times N/n_k} \) is a matrix that is obtained by concatenating the mode-\( k \) fibers along columns; here a mode-\( k \) fiber is an \( n_k \) dimensional vector obtained by fixing all the indices but the \( k \)th index of \( W \). The mode-\( k \) rank \( \ell_k \) of \( W \) is the rank of the mode-\( k \) unfolding \( \bar{X}(k) \). We say that a tensor \( W \) has Tucker rank \( (\ell_1, \ldots, \ell_K) \) if the mode-\( k \) rank is \( \ell_k \) for \( k = 1, \ldots, K \) (Kolda & Bader, 2009). The mode \( k \) folding is the inverse of the unfolding operation.

### 2.1. Overlapped Schatten norms

The low-rank inducing norm studied in (Signoreatto et al., 2010; Gandy et al., 2011; Liu et al., 2009; Tomioka et al., 2011a), which we call overlapped Schatten 1-norm, can be written as follows:

\[
\| W \|_{S_1/1} = \sum_{k=1}^K \| W(k) \|_{S_1},
\]

(1)

In this paper, we consider the following more general overlapped \( S_p/q \)-norm, which includes the Schatten 1-norm as the special case \((p, q) = (1, 1)\). The overlapped \( S_p/q \)-norm is written as follows:

\[
\| W \|_{S_p/q} = \left( \sum_{k=1}^K \| W(k) \|_{S_p/q} \right)^{1/q},
\]

(2)

where \( 1 \leq p, q \leq \infty \); here

\[
\| W \|_{S_p} = \left( \sum_{j=1}^r \sigma_j^p(W) \right)^{1/p}
\]

is the Schatten \( p \)-norm for matrices, where \( \sigma_j(W) \) is the \( j \)th largest singular value of \( W \).

When used as a regularizer, the overlapped Schatten 1-norm penalizes all modes of \( W \) to be jointly low-rank. It is related to the overlapped group regularization (see Jenatton et al., 2011; Mairal et al., 2011) in a sense that the same object \( W \) appears repeatedly in the norm.

The following inequality relates the overlapped Schatten 1-norm with the Frobenius norm, which was a key step in the analysis of Tomioka et al. (2011b):

\[
\| W \|_{S_1/1} \leq \sum_{k=1}^K \sqrt{\ell_k} \| W \|_F,
\]

(3)

where \( \ell_k \) is the mode-\( k \) rank of \( W \).

Now we are interested in the dual norm of the overlapped \( S_p/q \)-norm, because deriving the dual norm is a key step in solving the minimization problem that involves the norm (2) (see Mairal et al., 2011), as well as computing various complexity measures, such as, Rademacher complexity (Foygel & Srebro, 2011) and Gaussian width (Chandrasekaran et al., 2010). It turns out that the dual norm of the overlapped \( S_p/q \)-norm is the latent \( S_p^*\)-\( q^* \)-norm as shown in the following lemma.

**Lemma 1.** The dual norm of the overlapped \( S_p/q \)-norm is the latent \( S_p^*\)-\( q^* \)-norm, where \( 1/p + 1/p^* = 1 \) and \( 1/q + 1/q^* = 1 \), which is defined as follows:

\[
\| X^* \|_{S_p^*/q^*} = \inf_{(X^{(1)}, \ldots, X^{(K)}) = X} \left( \sum_{k=1}^K \| X^{(k)} \|_{S_p^*} \right)^{1/q^*}.
\]

(4)

Here the infimum is taken over the \( K \)-tuple of tensors \( X^{(1)}, \ldots, X^{(K)} \) that sums to \( X \).

**Proof.** The proof is presented in Appendix A. \( \square \)

The duality in the above lemma naturally generalizes the duality between overlapped/latent group sparsity norms that have only partial overlap (in contrast to the complete overlap here). Although being recognized in special instances (Jalali et al., 2010; Obozinski et al., 2011; Maurer & Pontil, 2011; Agarwal et al., 2011), to the best of our knowledge, this duality has not been presented in the generality of Lemma 1. Note that when the groups have no overlap, the overlapped/latent group sparsity norms become identical, and the duality is the ordinary duality between the group \( S_p/q \)-norms and the group \( S_p^*/q^* \)-norms.

### 2.2. Latent Schatten norms

The latent approach for tensor decomposition proposed by Tomioka et al. (2011a) solves the following minimization problem

\[
\text{minimize}_{W^{(1)}, \ldots, W^{(K)}} L(W^{(1)} + \cdots + W^{(K)}) + \lambda \sum_{k=1}^K \| W^{(k)} \|_{S_1/1}.
\]

(5)
where $L$ is a loss function, $\lambda$ is a regularization constant, and $W^{(k)}(k)$ is the mode-$k$ unfolding of $W^{(k)}$. Intuitively speaking, the latent approach for tensor decomposition predicts with a mixture of $K$ tensors that each are regularized to be low-rank in a specific mode.

Now, since the loss term in the minimization problem (5) only depends on the sum of the tensors $W^{(1)}, \ldots, W^{(K)}$, minimization problem (5) is equivalent to the following minimization problem

$$\min_{W} L(W) + \lambda \|W\|_{S_{1}/T},$$

In other words, we have identified the structured Schatten norm employed in the latent approach as the latent $S_{1}/1$-norm (or latent Schatten 1-norm for short), which can be written as follows:

$$\|W\|_{S_{1}/T} = \inf_{(W^{(1)} + \ldots + W^{(K)}) = W} \sum_{k=1}^{K} \|W^{(k)}(k)\|_{S_{1}}.$$  (6)

According to Lemma 1, the dual norm of the latent $S_{1}/1$-norm is the overlapped $S_{\infty}/\infty$-norm

$$\|X\|_{S_{\infty}/\infty} = \max_{k} \|X^{(k)}\|_{S_{\infty}},$$  (7)

where $\|\cdot\|_{S_{\infty}}$ is the spectral norm.

The following lemma is similar to inequality (3) and is a key in our analysis.

**Lemma 2.**

$$\|W\|_{S_{1}/T} \leq \left( \min_{k} \sqrt{\lambda_{k}} \right) \|W\|_{F},$$

where $\lambda_{k}$ is the mode-$k$ rank of $W$.

**Proof.** Since we are allowed to take a singleton decomposition $W^{(k)} = W$ and $W^{(k')} = 0$ ($k' \neq k$), we have

$$\|W\|_{S_{1}/T} = \inf_{(W^{(1)} + \ldots + W^{(K)}) = W} \sum_{k=1}^{K} \|W^{(k)}(k)\|_{S_{1}}$$

$$\leq \|W^{(k)}(k)\|_{S_{1}}$$

$$\leq \sqrt{\lambda_{k}} \|W^{(k)}\|_{F} \quad (\forall k = 1, \ldots, K)$$

Choosing $k$ that minimizes the right hand side, we obtain our claim. \qed

Compared to inequality (3), the latent Schatten 1-norm is bounded by the minimal square root of the ranks instead of the sum. This is the fundamental reason why the latent approach performs better than the overlapped approach as in Figure 2.

### 3. Main theoretical results

In this section, we study the consistency, generalization performance, and identifiability of the latent approach for tensor decomposition in the context of recovering an unknown tensor $W^{*}$ from noisy measurements. This is the setting of the experiment in Figure 2.

First, we show that the latent approach is consistent. That is, the error goes to zero when the noise goes to zero, which corresponds to the situation when the entries are repeatedly observed.

Second, combining the duality we presented in the previous section with the techniques from Agarwal et al. (2011), we analyze the denoising performance of the latent approach in the context of recovering an unknown tensor $W^{*}$ from noisy measurements. This is the setting of the experiment in Figure 2. We first prove a deterministic inequality that holds under certain condition on the regularization constant. Next, we assume Gaussian noise and derive an inequality that holds with high probability under an appropriate scaling of the regularization constant.

Third, we discuss the difference between overlapped approach and latent approach and provide an explanation for the empirically observed superior performance of the latent approach in Figure 2.

Finally we discuss the condition under which the decomposition $W = \sum_{k=1}^{K} W^{(k)}$ is identifiable and show that the model is (locally) identifiable only when the mixture consists of one component.

#### 3.1. Consistency

Let $W^{*}$ be the underlying true tensor and the noisy version $Y$ is obtained as follows:

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

where $E \in \mathbb{R}^{n_{1} \times \ldots \times n_{K}}$ is the noise tensor.

First we establish the consistency of the latent approach.

**Theorem 1.** The estimator defined by

$$\hat{W} = \arg\min_{W} \left( \frac{1}{2} \|Y - W\|_{F}^{2} + \lambda \|W\|_{S_{1}/T} \right).$$  (8)

is consistent. That is, when the noise goes to zero (e.g., when the entries are repeatedly observed), $\hat{W} \to W^{*}$ for any sequence $\lambda \to 0$.

**Proof.** Due to the triangular inequality

$$\|\hat{W} - W^{*}\|_{F} \leq \|\hat{W} - Y\|_{F} + \|Y - W^{*}\|_{F}.$$
Here the second term goes to zero as the noise shrinks. Next, from the optimality of $\mathcal{W}$, the first term satisfies
\[\mathcal{Y} - \hat{\mathcal{W}} \in \lambda \partial \|\hat{\mathcal{W}}\|_{S_1/1},\]
where $\partial \|\hat{\mathcal{W}}\|_{S_1/1}$ is the subdifferential of the latent $S_1/1$ norm at $\hat{\mathcal{W}}$. Now since the dual norm of the latent $S_1/1$ norm is the overlapping $S_{\infty}/\infty$ norm, for any $\mathcal{G} \in \partial \|\hat{\mathcal{W}}\|_{S_1/1}$, we have $\|\mathcal{G}\|_{S_{\infty}/\infty} \leq 1$, and therefore
\[\|\hat{\mathcal{W}} - \mathcal{Y}\|_F \leq C\|\hat{\mathcal{W}} - \mathcal{Y}\|_{S_{\infty}/\infty} \leq C\lambda,\]
where $C$ is a constant that is independent of $\lambda$. Therefore, for any sequence $\lambda \to 0$, we have $\hat{\mathcal{W}} \to \mathcal{W}^{*}$ when $\mathcal{E} \to 0$. \hfill\(\square\)

### 3.2. Deterministic bound

The consistency statement in the previous section only deals with the sum $\hat{\mathcal{W}} = \sum_{k=1}^{K} \hat{\mathcal{W}}^{(k)}$ and its convergence to the truth $\mathcal{W}^{*}$ in the limit the noise goes to zero. In this section, we establish a stronger statement that shows the behavior of individual terms $\hat{\mathcal{W}}^{(k)}$ and also the denoising performance.

To this end we need some additional assumptions.

First, we assume that the unknown tensor $\mathcal{W}^{*}$ is a mixture of $K$ tensors that each are low-rank in a certain mode and we have a noisy observation $\mathcal{Y}$ as follows:
\[\mathcal{Y} = \mathcal{W}^{*} + \mathcal{E} = \sum_{k=1}^{K} \mathcal{W}^{*(k)} + \mathcal{E},\]
where $\bar{r}_k = \text{rank}(\mathcal{W}^{(k)})$ is the mode-$k$ rank of the $k$th component $\mathcal{W}^{*(k)}$.

Second, we assume that the spectral norm of the mode-$k$ unfolding of the $l$th component is bounded by a constant $\alpha$ for all $k \neq l$ as follows:
\[\|\mathcal{W}^{*(l)}\|_{S_{\infty}} \leq \alpha \quad (\forall l \neq k, k, l = 1, \ldots, K).\]

Note that such an additional incoherence assumption has also been used in (Candes et al., 2009; Wright et al., 2010; Agarwal et al., 2011; Hsu et al., 2011).

We employ the following optimization problem to recover the unknown tensor $\mathcal{W}^{*}$:
\[
\hat{\mathcal{W}} = \arg\min_{\mathcal{W}} \left( \frac{1}{2} \|\mathcal{Y} - \mathcal{W}\|_F^2 + \lambda \|\mathcal{W}\|_{S_1/1} \right)
\]
s.t. $\|\mathcal{W}^{(k)}\|_{S_{\infty}} \leq \alpha$, $\forall l \neq k$, \hfill (11)

where $\mathcal{W} = \sum_{k=1}^{K} \mathcal{W}^{(k)}$ denotes the optimal decomposition induced by the latent Schatten 1-norm (6); $\lambda > 0$ is a regularization constant. Notice that we have introduced additional spectral norm constraints to control the correlation between the components (see also Agarwal et al., 2011).

Our first bound can be stated as follows:

**Theorem 2.** Let $\hat{\mathcal{W}}^{(k)}$ be an optimal decomposition of $\hat{\mathcal{W}}$ induced by the latent Schatten 1-norm (6). Assume that the regularization constant $\lambda$ satisfies $\lambda \geq 2\|\mathcal{E}\|_{S_{\infty}/\infty} + \alpha(K - 1)$. Then there is a universal constant $c$ such that, any solution $\hat{\mathcal{W}}$ of the minimization problem (11) satisfies the following deterministic bound:
\[
\sum_{k=1}^{K} \|\hat{\mathcal{W}}^{(k)} - \mathcal{W}^{*(k)}\|_F^2 \leq c\lambda^2 \sum_{k=1}^{K} \bar{r}_k. \quad (12)
\]

**Proof.** The proof is presented in Appendix B. \hfill\(\square\)

We can also obtain a bound on the difference of the whole tensor $\hat{\mathcal{W}} - \mathcal{W}^{*}$ rather than the squared sum differences as in Theorem 2 as follows.

**Corollary 1.** Under the same conditions as in Theorem 2 we have
\[
\|\hat{\mathcal{W}} - \mathcal{W}^{*}\|_F^2 \leq cK\lambda^2 \sum_{k=1}^{K} \bar{r}_k. \quad (13)
\]

**Proof.** Using the triangular inequality and Cauchy-Schwarz inequality we have
\[
\|\hat{\mathcal{W}} - \mathcal{W}^{*}\|_F \leq \sum_{k=1}^{K} \|\hat{\mathcal{W}}^{(k)} - \mathcal{W}^{*(k)}\|_F \leq \sqrt{K} \sum_{k=1}^{K} \|\hat{\mathcal{W}}^{(k)} - \mathcal{W}^{*(k)}\|_F. \quad \square
\]

Since we are bounding the overall error in (13), we may exploit the arbitrariness of the decomposition $\mathcal{W}^{*} = \sum_{k=1}^{K} \mathcal{W}^{*(k)}$ to obtain a tight bound. The tightest bound is obtained when we choose the decomposition that minimizes the sum of the ranks $\sum_{k=1}^{K} \bar{r}_k$. We say $\mathcal{W}^{*}$ has the latent rank $(\bar{r}_1, \ldots, \bar{r}_K)$ for such a minimal decomposition in terms of the sum.

A simple upper bound is obtained by choosing a decomposition $\mathcal{W}^{*(k)} = \mathcal{W}^{*}$ and $\mathcal{W}^{*(k')} = 0$ for $k' \neq k$. In particular by choosing the mode with the minimum mode-$k$ rank, we obtain
\[
\|\hat{\mathcal{W}} - \mathcal{W}^{*}\|_F^2 \leq cK\lambda^2 \min_{k=1, \ldots, K} \bar{r}_k,
\]
where $\bar{r}_k$ is the mode-$k$ rank of $\mathcal{W}^{*}$. We refer to the above decomposition as the minimum rank singleton decomposition.
Note that the right-hand side of our bound (12) does not necessarily go to zero when the noise $\mathcal{E}$ goes to zero, because $\lambda \geq \alpha(K-1)$. When the noise goes to zero, $\mathcal{W} \rightarrow \mathcal{W}^*$ can be obtained by any decreasing sequence $\lambda \rightarrow 0$ as shown in the previous subsection. Therefore our bound is most useful when the noise is relatively large and the first term $2 \|\mathcal{E}\|_{\infty} \infty$ dominates the second term $\alpha(K-1)$ in the condition for the regularization constant $\lambda$.

3.3. Gaussian noise

When the elements of the noise tensor $\mathcal{E}$ are Gaussian, we obtain the following theorem.

**Theorem 3. Assume that the elements of the noise tensor $\mathcal{E}$ are independent Gaussian random variables with variance $\sigma^2$. In addition, assume without loss of generality that the dimensionalities of $\mathcal{W}^*$ are sorted in the descending order, i.e., $n_1 \geq \cdots \geq n_K$. Then there are universal constants $c_0, c_1$ such that, with high probability, any solution of the minimization problem (11) with regularization constant $\lambda = c_0 \sigma (\sqrt{N/n_K} + \sqrt{n_1} + \sqrt{\log K}) + \alpha(K-1)$ satisfies the following bound:

$$\frac{1}{N} \sum_{k=1}^{K} \|\mathcal{W}(k) - \mathcal{W}^*(k)\|_F^2 \leq c_1 F \sigma^2 \sum_{k=1}^{K} \frac{\bar{r}_k}{n_K},$$

where $F = \left(1 + \sqrt{\frac{\log K}{N}}\right) + \left(\sqrt{\log K} + \frac{\alpha(K-1)}{c_0 \sigma}\right) \sqrt{\frac{N}{n}}$ is a factor that mildly depends on the dimensionalities and the constant $\alpha$ in (10).

**Proof.** The proof is presented in Appendix C. \hfill \square

Note that the theoretically optimal choice of regularization constant $\lambda$ is independent of the Tucker/latent rank of the truth $\mathcal{W}^*$, which is unknown in practice.

Again we can obtain a bound corresponding to the minimum rank singleton decomposition as in inequality (13) as follows:

$$\frac{1}{N} \|\mathcal{W} - \mathcal{W}^*\|_F^2 \leq c_1 K F \sigma^2 \frac{\min_k \bar{r}_k}{n_K},$$

where $F$ is the same factor as in Theorem 3.

3.4. Comparison with the overlapped approach

Inequality (15) explains the superior performance of the latent approach for tensor decomposition in Figure 2. The inequality obtained in (Tomokta et al., 2011b) for the overlapped approach that uses overlapped Schatten 1-norm (1) can be stated as follows:

$$\frac{1}{N} \|\mathcal{W} - \mathcal{W}^*\|_F^2 \leq c' \sigma^2 \left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{n_k}\right)^2 \left(\frac{1}{K} \sum_{k=1}^{K} \sqrt{\bar{r}_k}\right)^2.$$

Comparing inequalities (15) and (16), we notice that the complexity of the overlapped approach depends on the average (square root) of the Tucker rank $\bar{r}_1, \ldots, \bar{r}_K$, whereas that of the latent approach only grows linearly against the minimum Tucker rank. Interestingly, the latent approach performs as if it knows the mode with the minimum rank, although such information is not available to it. However in inequality (15) we have the factor $K$. This means that if the mode with the minimum rank is known, the latent approach looses by constant factor $K$ against the simple matrix decomposition approach that unfolds the given tensor at the minimal rank mode and performs ordinary Schatten 1-norm minimization.

3.5. Discussion on the identifiability

Let $\bar{r}_k = \text{rank}(\mathcal{W}^{(k)})$ be the mode-$k$ rank of the $k$th component $\mathcal{W}^{(k)}$ in the decomposition

$$\mathcal{W} = \mathcal{W}^{(1)} + \mathcal{W}^{(2)} + \cdots + \mathcal{W}^{(K)}. $$

We say that a decomposition (17) is locally identifiable when there is no other decomposition $\sum_{k=1}^{K} \mathcal{W}^{(k)}$ having the same rank ($\bar{r}_1, \ldots, \bar{r}_K$). The following theorem fully characterizes the local identifiability of the decomposition (17).

**Theorem 4. The decomposition (17) is locally identifiable if and only if $\mathcal{W}^{(k^*)} = \mathcal{W}$ for $k = k^*$ and $\mathcal{W}^{(k)} = 0$ otherwise, for some $k^*$.**

**Proof.** The proof is given in Appendix D. \hfill \square

The above theorem partly explains the difficulty of estimating individual components $\mathcal{W}^{(k^*)}$ without additional incoherence assumption as in (10). In fact, most decompositions of the form (9) are not identifiable.

4. Numerical results

In this section, we numerically confirm the scaling behavior we have theoretically predicted in the last section.

The goal of this experiment is to recover the true low rank tensor $\mathcal{W}^*$ from a noisy observation $\mathcal{Y}$. We randomly generated the true low rank tensors $\mathcal{W}^*$ of size $50 \times 50 \times 20$ or $80 \times 80 \times 40$ with various Tucker ranks.
Tensor Decomposition via Structured Schatten Norm Regularization

A low-rank tensor is generated by first randomly drawing the \( z_1 \times z_2 \times z_3 \) core tensor from the standard normal distribution and multiplying an orthogonal factor matrix drawn from the Haar measure to each mode. The observation tensor \( Y \) is obtained by adding Gaussian noise with standard deviation \( \sigma = 0.1 \). There is no missing entries in this experiment.

For an observation \( Y \), we computed tensor decompositions using the overlapped approach and the latent approach (11) using the solver available from the webpage\(^2\) of one of the authors of Tomioka et al. (2011a).

The solver uses the alternating direction method of multipliers (Gabay & Mercier, 1976) and the algorithm is described in the above paper. We computed the solutions for 20 candidate regularization constants ranging from 0.1 to 100 and report the results for three representative values for each method.

We measured the quality of the solutions obtained by the two approaches by the mean squared error (MSE) \( ||\hat{W} - W^*||_F^2/N \). In order to make our theoretical predictions more concrete, we define the quantities in the right hand side of the bounds (16) and (14) as Tucker rank (TR) complexity and Latent rank (LR) complexity, respectively, as follows:

\[
\text{TR complexity} = \left( \frac{1}{K} \sum_{k=1}^{K} \sqrt{\frac{n_k}{N}} \right)^2 \left( \frac{1}{K} \sum_{k=1}^{K} \sqrt{n_k} \right)^2, \tag{18}
\]

\[
\text{LR complexity} = \frac{\sum_{k=1}^{K} \bar{r}_k}{n_K}, \tag{19}
\]

where without loss of generality we assume \( n_1 \geq \cdots \geq n_K \). We have ignored terms like \( \sqrt{n_K/N} \) because they are negligible for \( n_k \approx 50 \) and \( N \approx 50,000 \). The TR complexity is equivalent to the normalized rank in (Tomioka et al., 2011b). Note that the TR complexity (18) is defined in terms of the Tucker rank \((z_1, \ldots, z_K)\) of the truth \( W^* \), whereas the LR complexity (19) is defined in terms of the latent rank \((\bar{r}_1, \ldots, \bar{r}_K)\) (see Section 3.2). In order to compute the sum of latent ranks \( \sum_{k=1}^{K} \bar{r}_k \), we ran the latent approach to the true tensor \( W^* \) without noise, and took the minimum of the sums obtained from that and the minimum rank singleton decomposition. The whole procedure is repeated 10 times and averaged.

Figure 3 shows the results of the experiment. The left panel shows the MSE of the overlapped approach against the TR complexity (18). The middle panel shows the MSE of the latent approach against the LR complexity (19). The right panel shows the improvement (i.e., MSE of the overlap approach divided by that of the latent approach) against the ratio of the respective complexity measures.

First, from the left panel we can confirm that as predicted by (Tomioka et al., 2011b), the MSE of the overlapped approach scales linearly against the TR complexity (18) for each value of the regularization constant. We can also see that as predicted by Theorem 3, by scaling the regularization constant proportionally with \( \sqrt{N/n_K} \), the series corresponding to size \( 50 \times 50 \times 20 \) and those corresponding to size \( 80 \times 80 \times 40 \) almost lie on top of each others.

From the central panel, we can clearly see that the MSE of the latent approach scales linearly against the LR complexity (19) as predicted by Theorem 3. The series with \( \Delta (\lambda = 3.79 \text{ for } 50 \times 50 \times 20, \lambda = 5.46 \text{ for } 80 \times 80 \times 40) \) is mostly below other series, which means that the optimal choice of the regularization constant is independent of the rank of the true tensor and only depends on the size; this agrees with the condition on \( \lambda \) in Theorem 3. Since the blue series and red series with the same markers lie on top of each other (especially the series with \( \Delta \) for which the optimal regularization constant is chosen), we can see that our theory predicts not only the scaling against the latent ranks but also that against the size of the tensor correctly. Note that the regularization constants are scaled by roughly 1.6 to account for the difference in the dimensionality.

The right panel reveals that in many cases the latent approach performs better than the overlapped approach, i.e., MSE (overlap)/ MSE (latent) greater than one. Moreover, we can see that the success of the latent approach relative to the overlapped approach is correlated with high TR complexity to LR complexity ratio. Indeed, we found that the optimal decomposition of the true tensor \( W^* \) was typically a singleton decomposition corresponding to the smallest tucker rank (see Section 3.2).

One might think that we can fix the overlapped approach by allowing individual regularization constant for each mode. However, this would only be possible if we knew the mode with small rank.

The improvements here are milder than that in Figure 2. This is because most of the randomly generated low-rank tensors were simultaneously low-rank to some degree. It is interesting that the latent approach perform at least as good as the overlapped approach also in such situations.

\(^2\)http://www.ibis.t.u-tokyo.ac.jp/RyotaTomioka/Softwares/Tensor
5. Conclusion

In this paper, we have presented a framework for structured Schatten norms. The current framework includes both the overlapped Schatten 1-norm and latent Schatten 1-norm recently proposed in the context of convex-optimization-based tensor decomposition (Signoretto et al., 2010; Gandy et al., 2011; Liu et al., 2009; Tomioka et al., 2011a), and connects these studies to the broader studies on structured sparsity (Bach et al., 2011; Jenatton et al., 2011; Obozinski et al., 2011; Maurer & Pontil, 2011). Moreover, we have shown a duality that holds between the two types of norms.

Furthermore, we have rigorously studied the performance of the latent approach for tensor decomposition. We have shown the consistency of the latent Schatten 1-norm minimization. Next, we have analyzed the denoising performance of the latent approach and shown that the error of the latent approach is upper bounded by the minimum Tucker rank, which contrasts sharply against the average (square root) dependency of the overlapped approach analyzed in Tomioka et al. (2011b). This explains the empirically observed superior performance of the latent approach compared to the overlapped approach. The most difficult case for the overlapped approach is when the unknown tensor is only low-rank in one mode as in Figure 2.

We have also confirmed through numerical simulations that our analysis precisely predicts the scaling of the mean squared error as a function of the dimensionalities and the latent rank of the unknown tensor. Unlike Tucker rank, latent rank of a tensor is not easy to compute. However, note that the theoretically optimal scaling of the regularization constant does not depend on the latent rank.

Therefore we have theoretically and empirically shown that for noisy tensor decomposition, the latent approach is more likely to perform better than the overlapped approach. Analyzing the performance of the latent approach for tensor completion would be an important future work.

The structured Schatten norms proposed in this paper include norms for tensors that are not employed in practice yet. Therefore, we envision that this paper serve as a starting point for various extensions, e.g., using the overlapped $S_1/\infty$-norm instead of the $S_1/1$-norm or a non-sparse tensor decomposition similar to the $\ell_p$-norm MKL (Micchelli & Pontil, 2005; Kloft et al., 2011).

References

Agarwal, A., Negahban, S., and Wainwright, M.J. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. Technical report, arXiv:1102.4807v2, 2011.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. Convex optimization with sparsity-inducing norms. In Optimization for Machine Learning. MIT Press, 2011.

Candes, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? Technical report, arXiv:0912.3599, 2009.

Chandrasekaran, V., Recht, B., Parrilo, PA, and Willsky,
A. The convex geometry of linear inverse problems, preprint. Technical report, arXiv:1012.0621v2, 2010.

De Lathauwer, L., De Moor, B., and Vandewalle, J. On the best rank-1 and rank-(R1, R2, . . . , RN) approximation of higher-order tensors. SIAM J. Matrix Anal. Appl., 21(4): 1324–1342, 2000. ISSN 0895-4798.

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

Foygel, R. and Srebro, N. Concentration-based guarantees for low-rank matrix reconstruction. Technical report, arXiv:1102.3923, 2011.

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

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

Hsu, Daniel, Kakade, Sham M. and Zhang, Tong. Robust matrix decomposition with sparse corruptions. Information Theory, IEEE Transactions on, 57(11):7221–7234, 2011.

Jalali, A., Ravikumar, P., Sanghavi, S., and Ruan, C. A dirty model for multi-task learning. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R.S., and Culotta, A. (eds.), Advances in NIPS 23, pp. 964–972. 2010.

Jenatton, R., Audibert, J.Y., and Bach, F. Structured variable selection with sparsity-inducing norms. J. Mach. Learn. Res., 12:2777–2824, 2011.

Kloft, M., Brefeld, U., Sonnenburg, S., and Zien, A. Lp-norm multiple kernel learning. J. Mach. Learn. Res., 12: 953–997, 2011.

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

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

Mairal, J., Jenatton, R., Obozinski, G., and Bach, F. Convex and network flow optimization for structured sparsity. J. Mach. Learn. Res., 12:2681–2720, 2011.

Maurer, A. and Pontil, M. Structured sparsity and generalization. Technical report, arXiv:1108.3476, 2011.

Micchelli, C.A. and Pontil, M. Learning the kernel function via regularization. Journal of Machine Learning Research, 6:1099–1125, 2005.

Morup, M. Applications of tensor (multiway array) factorizations and decompositions in data mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1):24–40, 2011.

Negahban, S., Ravikumar, P., Wainwright, M., and Yu, B. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A. (eds.), Advances in NIPS 22, pp. 1348–1356. 2009.

Obozinski, G., Jacob, L., and Vert, J.P. Group lasso with overlaps: the latent group lasso approach. Technical report, arXiv:1110.0413, 2011.

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

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

Sirovich, L. and Kirby, M. Low-dimensional procedure for the characterization of human faces. J. Opt. Soc. Am. A, 4(3):519–524, 1987.

Srebro, N., Rennie, J. D. M., and Jaakkola, T. S. Maximum-margin matrix factorization. In Saul, Lawrence K., Weiss, Yair, and Bottou, Léon (eds.), Advances in NIPS 17, pp. 1329–1336. MIT Press, Cambridge, MA, 2005.

Tomliska, R., Hayashi, K., and Kashima, H. Estimation of low-rank tensors via convex optimization. Technical report, arXiv:1010.0789, 2011a.

Tomliska, R., Suzuki, T., Hayashi, K., and Kashima, H. Statistical performance of convex tensor decomposition. In Shawe-Taylor, J., Zemel, R.S., Bartlett, P., Pereira, F.C.N., and Weinberger, K.Q. (eds.), Advances in NIPS 24, pp. 972–980. 2011b.

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

Vasilescu, M. and Terzopoulos, D. Multilinear analysis of image ensembles: Tensorfaces. Computer Vision—ECCV 2002, pp. 447–460, 2002.

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

Wright, J., Ganesh, A., Rao, S., Peng, Y., and Ma, Y. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A. (eds.), Advances in NIPS 22, pp. 2080–2088. Curran Associates Inc., 2010.
Supplementary material for “Convex Tensor Decomposition via Structured Schatten Norms”

A. Proof of Lemma 1

Proof. From the definition, the dual norm \( \|X\|_{(S_p/q)}^* \) can be written as follows:

\[
\|W\|_{(S_p/q)}^* = \sup \langle W, X \rangle \quad \text{s.t.} \quad \|W\|_{S_p/q} \leq 1.
\]

The basic strategy of the proof is to rewrite the above maximization problem as a constraint optimization problem and derive the dual problem.

First, we rewrite the above maximization problem as follows:

\[
\|X\|_{(S_p/q)}^* = \frac{1}{K} \sum_{k=1}^{K} \langle Z_k, X_{(k)} \rangle
\]

s.t. \( Z_k = W_{(k)} \sum_{k=1}^{K} \|Z_k\|_{S_p}^q \leq 1, \)

where \( Z_k \in \mathbb{R}^{n_k \times N/n_k} \) \( (k = 1, \ldots, K) \) are auxiliary variables.

Next we write down the Lagrangian as follows:

\[
L = \frac{1}{K} \sum_{k=1}^{K} \langle Z_k, X_{(k)} \rangle
\] + \( \frac{1}{K} \sum_{k=1}^{K} \langle \tilde{Y}_{(k)}(k), Z_{(k)} - W \rangle \)
\] + \( \frac{\gamma}{K^q} \left( 1 - \sum_{k \neq k_0} \|Z_k\|_{S_p}^q \right), \)

where \( \tilde{Y}_{(k)} \in \mathbb{R}^{n_1 \times \cdots \times N/K} \) \( (k = 1, \ldots, K) \), and \( \gamma \geq 0 \) are Lagrangian multipliers.

Note that for \( X, Z \in \mathbb{R}^{R \times C} \), we have

\[
\sup_{Z} \left( \langle X, Z \rangle - \frac{\gamma}{q^q} \|Z\|_{S_p}^q \right)
\] \leq \( \gamma \sup_{Z} \left( \|X\|_{\gamma} \|Z\|_{S_p} - \frac{1}{q^q} \|Z\|_{S_p}^q \right) \)
\] \leq \( \gamma^{1-q} \leq \langle X \rangle_{\gamma} \|Z\|_{S_p}^{q}. \)

Here the first equality is achieved if we take \( Z = \alpha U \text{diag}(\sigma_1^{q^q}, \ldots, \sigma_p^{q^q})V^\top \), where \( U \text{diag}(\sigma_1, \ldots, \sigma_p) \) is the singular value decomposition of the matrix \( X/\gamma \), and \( \alpha \) is an arbitrary scaling constant. The second equality is achieved if we take \( \|Z\|_{S_p} = \|X\|_{\gamma} \|Z\|_{S_p}^{q}. \)

Thus, maximizing the Lagrangian with respect to \( Z_k \)

\( (k = 1, \ldots, K) \) and \( W \), we obtain the dual problem

\[
\|X\|_{(S_p/q)}^* = \inf_{\gamma \in \mathbb{R}^+} \left( \frac{\gamma^{1-q}}{K^q} \sum_{k=1}^{K} \|Y_{(k)}\|_{S_p}^q + \frac{\gamma}{K^q} \right)
\]

s.t. \( \gamma \leq 1 + \cdots + \gamma \leq K \).

where we used the change of variable \( (X + \tilde{Y}(k))/K =: \gamma \).

Furthermore, by explicitly minimizing over \( \gamma \), we have \( \gamma/K = (\sum_{k=1}^{K} \|Y_{(k)}\|_{S_p}^q)^{1/q} \) and we obtain the statement of the lemma.

B. Proof of Theorem 2

Let \( \tilde{W} = \sum_{k=1}^{K} W_{(k)} \) be the solution and its optimal decomposition of the minimization problem (11); in addition let \( \Delta_{(k)} := \tilde{W}_{(k)} - W_{(k)} \).

The proof is based on Lemmas 3 and 4, which we present below.

In order to present the first lemma, we need the following definitions. Let \( U_k S_{(k)} V_k = W_{(k)} \) be the singular value decomposition of the mode-\( k \) unfolding of the \( k \)-th component of the unknown tensor \( W \). We define the orthogonal projection of \( \Delta_{(k)} \) as follows:

\[
\Delta_{(k)} = \Delta_{(k)} + \Delta_{(k)}',
\]

where

\[
\Delta_{(k)}' = (I_{N_{n_k}} - U_k U_k^\top) \Delta_{(k)} (J_{N/n_k} - V_k V_k^\top).
\]

Intuitively speaking, \( \Delta_{(k)}' \) lies in a subspace completely orthogonal to the unfolding of the \( k \)-th component \( W_{(k)} \), whereas \( \Delta_{(k)}' \) lies in a partially correlated subspace.

The following lemma is similar to Negahban et al. (2009, Lemma 1) and Tomioka et al. (2011b, Lemma 2), and it bounds the Schatten 1-norm of the orthogonal part \( \Delta_{(k)}' \) with that of the partially correlated part \( \Delta_{(k)} \) and also bounds the rank of \( \Delta_{(k)}' \).

Lemma 3. Let \( \tilde{W} \) be the solution of the minimization problem (11) with the regularization constant \( \lambda \geq 2\|E\|_{S_{\infty}/\infty} \). Let \( \Delta_{(k)} \) and its decomposition be as defined above. Then we have

1. \( \text{rank}(\Delta_{(k)}') \leq 2r_k \).
2. \( \sum_{k=1}^{K} \|\Delta_{(k)}'\|_{S_{1}} \leq 3 \sum_{k=1}^{K} \|\Delta_{(k)}\|_{S_{1}}. \)

Note that although the proof of the above statement closely follows that of Tomioka et al. (2011b, Lemma
The following lemma relates the squared Frobenius norm of the difference of the sums $\|\sum_{k=1}^{K} \Delta^{(k)}\|^2_F$ with the sum of squared differences $\sum_{k=1}^{K} \|\Delta^{(k)}\|^2_F$.

**Lemma 4.** Let $\hat{W}$ be the solution of the minimization problem (11). Then we have,

$$\frac{1}{2} \sum_{k=1}^{K} \|\Delta^{(k)}\|^2_F \leq \frac{1}{2} \|\Delta\|^2_F + \alpha(K-1) \sum_{k=1}^{K} \|\Delta^{(k)}\|_1,$$

where $\Delta = \sum_{k=1}^{K} \Delta^{(k)}$.

**Proof of Theorem 2.** First from the optimality of $\hat{W}$, we have

$$\frac{1}{2} \|Y - \hat{W}\|_F^2 + \lambda \sum_{k=1}^{K} \|\hat{W}^{(k)}\|_1 \leq \frac{1}{2} \|Y - W^*\|_F^2 + \lambda \sum_{k=1}^{K} \|W^{(k)}\|_1,$$

which implies

$$\frac{1}{2} \|\Delta\|^2_F \leq (\Delta, \mathcal{E}) + \lambda \sum_{k=1}^{K} \|\Delta^{(k)}\|_1 \leq (\|\mathcal{E}\|_{S_w/\infty} + \lambda) \sum_{k=1}^{K} \|\Delta^{(k)}\|_1,$$

where we used the fact that $\mathcal{Y} = \mathcal{W}^* + \mathcal{E}$ and the triangular inequality in the first line, and Hölder’s inequality in the second line. Note that there is an additional looseness in the second line due to the fact that $\Delta = \sum_{k=1}^{K} \Delta^{(k)}$ is not the optimal decomposition of $\Delta$ induced by the latent Schatten 1-norm.

Next, combining inequality (20) with Lemma 4, we obtain

$$\frac{1}{2} \sum_{k=1}^{K} \|\Delta^{(k)}\|^2_F \leq 2 \lambda \sum_{k=1}^{K} \|\Delta^{(k)}\|_1,$$

where we used the fact that $\lambda \geq \|\mathcal{E}\|_{S_w/\infty} + \alpha(K-1)$.

Finally combining inequality (21) with Lemma 3, we obtain

$$\frac{1}{2} \sum_{k=1}^{K} \|\Delta^{(k)}\|^2_F \leq 2 \lambda \sum_{k=1}^{K} \|\Delta^{(k)}\|_1 + \|\Delta''^{(k)}\|_1,$$

where we used Lemma 3 in the second line, Hölder’s inequality in the third line (combined with Lemma 3), the fact that $\Delta^{(k)} = \Delta_k' + \Delta_k''$ is an orthogonal decomposition in the fourth line, and Cauchy-Schwarz inequality in the fifth line. Dividing both sides of the last inequality by $\sqrt{\sum_{k=1}^{K} \|\mathcal{W}^{(k)}\|^2_F}$, we obtain our claim.

**C. Proof of Theorem 3**

**Proof.** Since each entry of $\mathcal{E}$ is an independent zero mean Gaussian random variable with variance $\sigma^2$, for each mode $k$ we have the following tail bound (Corollary 5.35 in [Vershynin, 2010])

$$P\left(\|E^{(k)}\|_{S_w} > \sigma \left(\sqrt{N/n_k} + \sqrt{n_k}\right) + t\right) \leq \exp\left(-t^2/(2\sigma^2)\right).$$

Next, taking a union bound

$$P\left(\max_k \|E^{(k)}\|_{S_w} > \sigma \max_k \left(\sqrt{N/n_k} + \sqrt{n_k}\right) + t\right) \leq K \exp\left(-t^2/(2\sigma^2)\right).$$

Substituting $t = t + \sigma \log K$, we have

$$P\left(\max_k \|E^{(k)}\|_{S_w/\infty} > \sigma \max_k \left(\sqrt{N/n_k} + \sqrt{n_k}\right) + \sigma \log K + t\right) \leq \exp\left(-t^2/(2\sigma^2)\right) \leq \exp\left(-t^2/(2\sigma^2)\right).$$

Therefore if $c_0 > 2$,

$$\lambda = c_0 \sigma \left(\sqrt{N/nK} + \sqrt{n_1} + \sqrt{\log K}\right) + \alpha(K-1) \geq 2 \|\mathcal{E}\|_{S_w/\infty} + \alpha(K-1),$$

with probability at least $1 - \exp\left(-\frac{c_0 - 21}{2}(N/nK)\right)$, which satisfies the condition of Theorem 2. Substituting the above $\lambda$ into the right hand side of the error bound in Theorem 2 we have the statement of Theorem 3.

**D. Proof of Theorem 4**

**Proof.** We first prove the “if” direction. Suppose that there is another decomposition

$$\sum_{k=1}^{K} \mathcal{W}^{(k)} = \sum_{k=1}^{K} \mathcal{W}_h^{(k)},$$

such that $\text{rank}(\mathcal{W}^{(k)}) = \text{rank}(\mathcal{W}_h^{(k)})$. Note that $\mathcal{W} \neq \mathcal{W}_h$ can happen only when $\mathcal{W}^{(k)} \neq 0$ (otherwise
the rank would increase). Also note that $W \neq \tilde{W}$ should happen for at least two $k$'s. Combining these
we conclude that there are $k \neq \ell$ such that $W^{(k)} \neq 0$
and $W^{(\ell)} \neq 0$.

Conversely, suppose that there are $k \neq \ell$ such that
$W^{(k)} \neq 0$ and $W^{(\ell)} \neq 0$, we can write
\[ W^{(k)} = C^{(k)} \times_k U_k, \]
\[ W^{(\ell)} = C^{(\ell)} \times_\ell U_\ell, \]
where $U_k \in \mathbb{R}^{n_k \times \tilde{r}_k}$, $C^{(k)} \in \mathbb{R}^{n_1 \times \cdots \times n_{k-1} \times \tilde{r}_k \times \cdots \times n_K}$,
and $U_\ell$ and $C^{(\ell)}$ are defined similarly. Since $C^{(k)}$ and
$C^{(\ell)}$ are allowed to be full rank, we can define
\[ \tilde{C}^{(k)} = C^{(k)} + D^{(k,\ell)} \times_\ell U_\ell, \]
\[ \tilde{C}^{(\ell)} = C^{(\ell)} - D^{(k,\ell)} \times_k U_k, \]
for any $D \in \mathbb{R}^{n_1 \times \cdots \times \tilde{r}_k \times \cdots \tilde{r}_\ell \times \cdots \times n_K}$. Then we have
\[ W^{(k)} + W^{(\ell)} = C^{(k)} \times_k U_k + C^{(\ell)} \times_\ell U_\ell \]
\[ = \left( C^{(k)} + D^{(k,\ell)} \times_\ell U_\ell \right) \times_k U_k \]
\[ + \left( C^{(\ell)} - D^{(k,\ell)} \times_k U_k \right) \times_\ell U_\ell \]
\[ = \tilde{C}^{(k)} \times_k U_k + \tilde{C}^{(\ell)} \times_\ell U_\ell \]
\[ = \tilde{W}^{(k)} + \tilde{W}^{(\ell)}. \]

Note that rank($\tilde{W}^{(k')}_{(k')}$) $= \tilde{r}_{k'}$ for $k' = k, \ell$. Therefore,
there are infinitely many decompositions that have the
same rank ($\tilde{r}_1, \ldots, \tilde{r}_K$).

\[ \square \]

---

3Here the tensor mode-$k$ product $A = B \times_k C$ is defined
as $a_{i_1 \ldots i_K} = \sum_{l=1}^{d_k} b_{i_1i_2 \ldots i_{k-1}l} c_{l \ell_k}$ where $A = (a_{i_1 \ldots i_K}) \in \mathbb{R}^{n_1 \times \cdots \times n_K}$, $B = (b_{i_1 \ldots i_{k-1}l}) \in \mathbb{R}^{n_1 \times \cdots \times d_k \times \cdots \times n_K}$, and $C = (c_{l \ell_k}) \in \mathbb{R}^{d_k \times n_k}$.
Rank of the first two modes

Estimation error $||W - W^*||$

Latent Schatten $p$-norm