Statistical Performance of Convex Tensor Decomposition

Authored by:

Ryota Tomioka
Hisashi Kashima
Taiji Suzuki
Kohei Hayashi

Abstract

We analyze the statistical performance of a recently proposed convex tensor decomposition algorithm. Conventionally tensor decomposition has been formulated as non-convex optimization problems, which hindered the analysis of their performance. We show under some conditions that the mean squared error of the convex method scales linearly with the quantity we call the normalized rank of the true tensor. The current analysis naturally extends the analysis of convex low-rank matrix estimation to tensors. Furthermore, we show through numerical experiments that our theory can precisely predict the scaling behaviour in practice.

1 Paper Body

Tensors (multi-way arrays) generalize matrices and naturally represent data having more than two modalities. For example, multi-variate time-series, for instance, electroencephalography (EEG), recorded from multiple subjects under various conditions naturally form a tensor. Moreover, in collaborative ?ltering, users? preferences on products, conventionally represented as a matrix, can be represented as a tensor when the preferences change over time or context. For the analysis of tensor data, various models and methods for the low-rank decomposition of tensors have been proposed (see Kolda & Bader [12] for a recent survey). These techniques have recently become increasingly popular in data-mining [1, 14] and computer vision [25, 26]. Besides they have proven useful in chemometrics [4], psychometrics [24], and signal processing [20, 7, 8]. Despite empirical success, the statistical performance of tensor decomposition algorithms has not been fully elucidated. The dif?culty lies in the non-convexity of the conventional tensor decomposition algorithms (e.g., alternating least squares [6]). In addition, studies have revealed many discrepancies (see [12]) between matrix rank and tensor rank, which make extension of studies on the performance of low-rank matrix models (e.g., [9]) challenging. Recently, several authors [21, 10, 13, 23] have focused on the notion of tensor mode-k rank (instead of tensor rank), which is related to the Tucker decomposition [24]. They discovered that regularized estimation based on the Schatten 1-norm, which is a popular technique for recovering low-rank matrices via convex optimization, can also be applied to tensor decomposition. In particular, the 1

Convex Tucker (exact) Optimization tolerance

0

10

?3

10

0

0.2 0.4 0.6 0.8 Fraction of observed elements

1

Figure 1: Result of estimation of rank-(7, 8, 9) tensor of dimensions 50???? 50 ? 20 from partial ??? ? ? W ? ??? is plotted against the measurements; see [23] for the details. The estimation error ???W F fraction of observed elements m = M/N . Error bars over 10 repetitions are also shown. Convex refers to the convex tensor decomposition based on the minimization problem (7). Tucker (exact) refers to the conventional (non-convex) Tucker decomposition [24] at the correct rank. Gray dashed line shows the optimization tolerance 10?3 . The question is how we can predict the point where the generalization begins (roughly m = 0.35 in this plot).

study in [23] showed that there is a clear transition at certain number of samples where the error drops dramatically from no generalization to perfect generalization (see Figure 1). In this paper, motivated by the above recent work, we mathematically analyze the performance of convex tensor decomposition. The new convex formulation for tensor decomposition allows us to generalize recent results on Schatten 1-norm-regularized estimation of matrices (see [17, 18, 5, 19]). Under a general setting we show how the estimation error scales with the mode-k ranks of the true tensor. Furthermore, we analyze the speci?c settings of (i) noisy tensor decomposition and (ii) random Gaussian design. In the ?rst setting, we assume that all the elements of a low-rank tensor is observed with noise and the goal is to recover the underlying low-rank structure. This is the most common setting a tensor decomposition algorithm is used. In the second setting, we assume that the unknown tensor is a coef?cient of a tensor-input scalar-output regression problem and the input tensors (design) are randomly given from independent Gaussian distributions. Surprisingly, it turns out that the random Gaussian setting can precisely predict the phase-transition-like behaviour in Figure 1. To the best of our knowledge, this is the ?rst paper that rigorously studies the performance of a tensor decomposition algorithm.

2

Notation

In this section, we introduce the notations we use in this paper. Moreover, we introduce a H?olderlike inequality (3) and the notion of mode-k decomposability

(5), which play central roles in our analysis. QK Let X ? Rn1 ????nK be a K-way tensor. We denote the number of elements in X by N = k=1 nk . ? The inner product between two tensors ?W, X ? is de?ned as ?W, X ? = vec(W) p), where ??? ??? vec(X vec is a vectorization. In addition, we de?ne the Frobenius norm of a tensor ???X ???F = ?X , X ?. Q The mode-k unfolding X (k) is the nk ? n ? k (? nk := k? ?=k nk?) matrix obtained by concatenating the mode-k ?bers (the vectors obtained by ?xing every index of X but the kth index) of X as column vectors. The mode-k rank of a tensor X , denoted by rankk (X), is the rank of the mode-k unfolding X (k) (as a matrix). Note that when K = 2 and X is actually a matrix, and X (2) = X (1) ? . We say a tensor X is rank (r1 , . . . , rK) when rk = rankk (X) for k = 1, . . . , K. Note that the mode-k rank can be computed in a polynomial time, because it boils down to computing a matrix rank, whereas computing tensor rank is NP complete [11]. See [12] for more details. Since for each k, the convex envelope of the mode-k rank is given as the Schatten 1-norm [18] (known as the trace norm [22] or the nuclear norm [3]), it is natural to consider the following 2

??? ??? overlapped Schatten 1-norm ???W ???S of a tensor W ? Rn1 ?????nK (see also [21]): 1

??? ??? ???W ???

S1

=

K ? 1 X? ?W (k) ? , S1 K

(1)

k=1

where W (k) is the mode-k unfolding of W. Here ? ? ?S1 is the Schatten 1-norm for a matrix Xr ?W ?S1 = ?j (W), j=1

where ?j (W) is the jth largest singular-value of W . The dual norm of the Schatten 1-norm is the Schatten ?-norm (known as the spectral norm) as follows: ?X?S? = max ?j (X). j=1,...,r

Since the two norms ? ? ?S1 and ? ? ?S? are dual to each other, we have the following inequality: —?W , X?— ? ?W ?S1 ?X?S? , (2) where ?W , X? is the inner product of W and X. The same inequality holds for the overlapped Schatten 1-norm (1) and its dual norm. The dual norm of the overlapped Schatten 1-norm can be characterized by the following lemma. ??? ??? Lemma 1. The dual norm of the overlapped Schatten 1-norm denoted as ???????S ? is de?ned as the 1 in?mum of the maximum mode-k spectral norm over the tensors whose average equals the given tensor X as follows: ??? ??? (k) ???X ??? ? = max ?Y (k) ?S? , inf S1 1 (1) +Y (2) +???+Y (K) =X k=1,...,K Y ( ) K (k)

where Y (k) is the mode-k unfolding of Y (k) . Moreover, the following upper bound on the dual norm ??? ??? ??????? ? is valid: S1

??? ??? ???X ???

S1?

??? ??? 1 XK ?X (k) ?S? . ? ???X ???mean := k=1 K

??? ??? Proof. The ?rst part can be shown by solving the dual of the maximization problem ???X ???S ? := 1 ??? ??? sup ?W, X ? s.t. ???W ???S1

? 1. The second part is obtained by setting Y (k) = PK K1/c ? X /ck , where ck = ?X (k) ?S? , and using Jensen?s inequality.

k? =1

k

According to Lemma 1, we have the ??following ? ??? ??? H? ?o??lder-like ??? inequality ??? ??? ??? —?W, X ?— ? ???W ???S1 ???X ???S ? ? ???W ???S1 ???X ???mean .

(3) ??? ??? ??? ??? Note that the above bound is tighter than the more intuitive relation — ?W, X ? — ? ???W ???S ???X ???S 1 ? ??? ??? (???X ???S? := max1,...,K ?X (k) ?S?), which one might come up as an analogy to the matrix case (2). 1

Finally, let W ? ? Rn1 ?????nK be the low-rank tensor that we wish to recover. We assume that W ? is rank (r1 , . . . , rK). Thus, for each k we have W ?(k) = U k S k V k (k = 1, . . . , K), where U k ? Rnk ?rk and V k ? Rn? k ?rk are orthogonal, and S k ? Rrk ?rk is diagonal. Let ? ? Rn1 ?????nK be an arbitrary tensor. We de?ne the mode-k orthogonal complement ???k of an unfolding ?(k) ? Rnk ??nk of ? with respect to the true low-rank tensor W ? as follows: ??k

???k = (I nk ? U k U k ?)?(k) (I n? k ? V k V k ?).

(4)

:= ?(k) ? ???k is the true tensor W ?(k) .

In addition the component having overlapped row/column space with the unfolding of Note that the decomposition ?(k) = ??k + ???k is de?ned for each mode; thus we use subscript k instead of (k). Using the decomposition de?ned above we have the following equality, which we call mode-k decomposability of the Schatten 1-norm: ?W ?(k) + ???k ?S1 = ?W ?(k) ?S1 + ????k ?S1 (k = 1, . . . , K). (5) The above decomposition is de?ned for each mode and thus it is weaker than the notion of decomposability discussed by Negahban et al. [15]. 3

3

Theory

In this section, we ?rst present a deterministic result that holds under a certain choice of regularization constant ?M and an assumption called the restricted strong convexity. Then, we focus on special cases to justify the choice of regularization constant and the restricted strong convexity assumption. We analyze the setting of (i) noisy tensor decomposition and (ii) random Gaussian design in Section 3.2 and Section 3.3, respectively. 3.1

Main result

Our goal is to estimate an unknown rank (r1 , . . . , rK) tensor W ? ? Rn1 ????nK from observations yi = ?Xi , W ? ? + ?i (i = 1, . . . , M). (6) Here the noise ?i follows the independent zero-mean Gaussian distribution with variance ? 2 . We employ the regularized empirical risk minimization problem proposed in [21, 10, 13, 23] for the estimation of W as follows: ??? ??? 1 minimize ?y ? X(W)?22 + ?M ???W ???S1 , (7) n ?????n 1 K 2M W?R where y = (y1 , . . . , yM)? is the collection of observations; X : Rn1 ?????nK ? RM is a linear operator that maps W to the M dimensional output vector X(W) = (?X1 , W? , . . . , ?XM , W?) ? ? RM . The Schatten 1-norm term penalizes every mode of W to be jointly low-rank (see Equation (1)); ?M ¿ 0 is the regularization constant. Accordingly, the solution of the minimization problem (7) is typically a low-rank tensor when ?M is suf?ciently large. In addition, we denote the adjoint operator PM of X as X? : RM ? Rn1 ?????nK ; that is X? (?) = i=1 ?i Xi ? Rn1 ?????nK . ? ? W? The ?rst step in our analysis is to characterize the particularity of the residual tensor ? := W as in the following lemma. ??? ??? ? be the solution of the minimization problem (7) with ?M ? 2???X? (?)??? Lemma 2. Let W /M , mean ? ? W ? , where W ? is the true low-rank tensor. Let ?(k) = ?? + ??? be the and let ? := W k k decomposition de?ned in Equation (4). Then we have the following inequalities: 1. rank(??k) ? 2rk for each k = 1, . . . , K. PK PK ? ?? 2. k=1 ??k ?S1 . k=1 ??k ?S1 ? 3 Proof. The proof uses the mode-k decomposability (5) and is analogous to that of Lemma 1 in [17]. The second ingredient of our analysis is the restricted strong convexity. Although, ?strong? may sound like a strong assumption, the point is that we require this assumption to hold only for the particular residual tensor we characterized in Lemma 2. The assumption can be stated as follows. Assumption 1 (Restricted strong convexity). We suppose that there is a positive constant ?(X) such that the operator X satis?es the inequality ??? ???2 1 ?X(?)?22 ??(X)???????F , (8) M PK for all ? ? Rn1 ?????nK such that for each k = 1, . . . , K, rank(??k) ? 2rk and k=1 ????k ?S1 ? PK 3 k=1 ???k ?S1 , where ??k and ???k are de?ned through the decomposition (4). Now using the above two ingredients, we are ready to prove the following deterministic guarantee on the performance of the estimation procedure (7). ??? ??? ? be the solution of the minimization problem (7) with ?M ? 2???X? (?)??? Theorem 1. Let W /M . mean Suppose that the operator X satis?es the restricted strong convexity condition. Then the following bound is true: PK ? ??? ??? ? ? W ? ??? ? 32?M k=1 rk . ???W (9) F ?(X)K 4

? ? W ? . Combining the fact that the objective value for W ? Proof. Let ? = W ??? ? ??? ??? is??smaller ? ??than ? ??? that for ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? W , the H?older-like inequality (3), the triangular inequality W S ? W S ? ?????S , and 1 1 1 ??? ??? the assumption ???X? (?)/M ??? ? ?M /2, we obtain mean

??? ??? ??? ??? ??? ??? ??? ??? 1 (10) ?X(?)?22 ? ???X? (?)/M ???mean ???????S1 + ?M ???????S1 ? 2?M ???????S1 . 2M Now the left-hand side can be lower-bounded using the restricted strong convexity (8). On the other hand, using Lemma 2, the right-hand side can be upper-bounded as follows: ??? ??? ??? ??? ??? ??? ? ??????? ? 1 PK (???k ?S1 + ????k ?S1) ? 4 PK ???k ?S1 ? 4 ? F PK 2rk , (11) k=1 k=1 k=1 K K K S1 ??? ??? where the last inequality follows because ???????F = ??(k) ?F for k = 1, . . . , K. Combining inequalities (8), (10), and (11), we obtain our claim (9). Negahban et al. [15] (see also [17]) pointed out that the key properties for establishing a sharp convergence result for a regularized M -estimator is the decomposability of the regularizer and the restricted strong convexity. What we have shown suggests that the weaker mode-k decomposability (5) suf?ce to obtain the above convergence result for the overlapped Schatten 1-norm (1) regularization. 3.2 Noisy Tensor Decomposition In this subsection, we consider the setting where all the elements are observed (with noise) and the goal is to recover the underlying low-rank tensor without noise. Since all the elements are observed only once, X is simply a vectorization (M = ??? ??? N), and the left2 ? ? W ? ??? . Therefore, the ? ? ? = W hand side of inequality (10)???gives the quantity of interest ?X(?)? 2 F ??? remaining task is to bound ???X? (?)???mean as in the following lemma. Lemma 3. Suppose ??? ?that ?? X : n1 ?? ? ??nK ? N is a vectorization of a tensor. With high probability the quantity ???X? (?)???mean is concentrated around its mean, which can be bounded as follows: K ??? ??? ? p ? X ?? E???X? (?)???mean ? nk + n ? k . K

(12)

k=1

??? ??? Setting the regularization constant as ?M = c0 E???X? (?)???mean /N , we obtain the following theorem. Theorem 2. Suppose that X : n1 ?? ? ??nK ? N is a vectorization of a tensor. There are universal constants c0 and c1 , such that, with high probability, any solution of the minimization problem (7) PK ? p with regularization constant ?M = c0 ? k=1 (nk + n ? k)/(KN) satis?es the following bound: ? !2 ? !2 K K X X ??? ???2 ?? ? p ? 1 1 ? 2 ? ? W ??? ? c1 ? ???W nk + n ? k rk . F K K k=1

k=1

Proof. Combining Equations (10)?(11) with the fact that X is simply a vectorization and M = N , we have ? 1 ? ? W ? ?F ? 16 2?M PK ?rk . ?W N

K

k=1

Substituting the choice of regularization constant ?M and squaring both sides, we obtain our claim.? We can simplify the result of Theorem 2 by noting that n ? k = N/nk ? nk , when the dimenPK ? 2 1 sions are of the same order. Introducing the notation ?r?1/2 = (K rk) and n?1 := k=1 (1/n1 , . . . , 1/nK), we have ??? ??? ? ? W ? ???2 ???W ? ? F ? Op ? 2 ?n?1 ?1/2 ?r?1/2 . (13) N We call the quantity r? = ?n?1 ?1/2 ?r?1/2 the normalized rank, because r? = r/n when the dimensions are balanced (nk = n and rk = r for all k = 1, . . . , K). 5

3.3

Random Gaussian Design

In this subsection, we consider the case the elements of the input tensors Xi (i = 1, . . . , M) in the observation model (6) are distributed according to independent identical standard Gaussian distributions. We call this setting random Gaussian design. ??? ??? First we show an upper bound on the norm ???X? (?)???mean , which we use to specify the scaling of the regularization constant ?M in Theorem 1. Lemma 4. Let X : Rn1 ?????nK ? RM be a random Gaussian design. In addition, we assume that ?i is sampled independently from N (0, ? 2). Then with high probability the quantity ? ??? ? the ??noise ???X (?)??? is concentrated around its mean, which can be bounded as follows: mean ??? ??? E???X? (?)???

mean

? K ? p ? M X ?? ? nk + n ? k . K k=1

Next the following lemma, which is a generalization of a result presented in Negahban and Wainwright [17, Proposition 1], provides a ground for the restricted strong convexity assumption (8). Lemma 5. Let X : Rn1 ?????nK ? RM be a random Gaussian design. Then it satis?es ?r ! r K n ? k ?????? ?????? 1 X ?X(?)?2 1 ?????? ?????? nk ? ? F? + ? S1 , ? 4 K M M M k=1

with probability at least 1 ? 2 exp(?N/32). Proof. The proof is analogous to that of Proposition 1 in [17] except that we use H?older-like inequality (3) for tensors instead of inequality (2) for matrices. Finally, we obtain the following convergence bound. Theorem 3. Under the random Gaussian design setup, there are universal constants c0 , c1 , and c2 PK ? PK ? 2 p 1 1 such that for a sample size M ? c1 (K n ? k))2 (K rk) , any solution of the k=1 (nk + ? Pk=1 p ? K minimization problem (7) with regularization constant ?M = c0 ? k=1 (nk + n ? k)/(K M) satis?es the following bound: PK ? PK ? 2 p 1 1 ??? ??? ?2 (K n ? k))2 (K k=1 (nk + k=1 rk) ? ? W ? ???2 ? c2 ???W , F M with high probability. Again we can simplify the result of Theorem 3 as follows: for sample size M ? c1 N r? we have ? ? ?1 ??? ??? ? ? W ? ???2 ? Op ? 2 N ?n ?1/2 ?r?1/2 , ???W (14) F M where r? = ?n?1 ?1/2 ?r?1/2 is the normalized rank. Note that the condition on the number of samples M does not depend on the noise variance ? 2 . Therefore in the limit ? 2 ? 0, the bound (14) is suf?ciently small but only valid for sample size M that exceeds c1 N r?, which implies a threshold behavior as in Figure 1. Note also that in the matrix case (K = 2), r1 = r2 = r and N ?n?1 ?1/2 = O(n1 + n2). Therefore ? ? W ? ?2 ? we can restate the above result as for sample size M ? c1 r(n1 + n2), we have ?W F Op (r(n1 + n2)/M), which is compatible with the result in [17, 18].

4

Experiments

In this section, we conduct two numerical experiments to con?rm our analysis in Section 3.2 and Section 3.3. 6

?4

3

x 10

0.03

size=[50 50 20] ?M=0.03/N size=[50 50 20] ?M=0.33/N

size=[50 50 20] ?M=2.34/N

size=[50 50 20] ? =0.54/N

0.025

M

size=[100 100 50] ?M=0.66/N

size=[100 100 50] ?M=0.69/N

Mean squared error

Mean squared error

size=[50 50 20] ? =6/N M

size=[100 100 50] ?M=0.06/N 2

size=[50 50 20] ?M=0.33/N

size=[100 100 50] ? =1.11/N M

1

0.02

size=[100 100 50] ? =4.5/N M

size=[100 100 50] ?M=12/N 0.015

0.01

0.005

0 0

0.2

0.4 0.6 Normalized rank

0.8

0 0

1

(a) Small noise (? = 0.01).

0.2

0.4 0.6 Normalized rank

0.8

1

(b) Large noise (? = 0.1).

Figure 2: Result of noisy tensor decomposition for tensors of size 50 ? 50 ? 20 and 100 ? 100 ? 50.

4.1

Noisy Tensor Decomposition

We randomly generated low-rank tensors of dimensions n(1) = (50, 50, 20) and n(2) = (100, 100, 50) for various ranks (r1 , . . . , rK). For a speci?c rank, we generated the true tensor by drawing elements of the r1 ? ? ? ? ? rK ?core tensor? from the standard normal distribution and multiplying its each mode by an orthonormal factor randomly drawn from the Haar measure. As described in Section 3.2, the observation y consists of all the elements of the original tensor once (M = N) with additive independent Gaussian noise with variance ? 2 . We used the alternating direction method of multipliers (ADMM) for ?constraint? approaches described in [23, 10] to solve the minimization problem (7). The whole experiment was repeated 10 times and averaged. ??? ??? ? ? W ? ???2 /N is plotted against The results are shown in Figure 2. The mean squared error ???W F the normalized rank r? = ?n?1 ?1/2 ?r?1/2 (of the true tensor) de?ned in Equation (13). Since the choice of the regularization constant ?M only depends on the size of the tensor and not on the ranks of the underlying tensor in Theorem 2, we ?x the regularization constant to some different values and report the dependency of the estimation error on the normalized rank r? of the true tensor. Figure 2(a) shows the result for small noise (? = 0.01) and Figure 2(b) shows the result for large ??? ??? ? ? W ? ???2 grows linearly noise (? = 0.1). As predicted by Theorem 2, the squared error ???W F against the normalized rank r?. This behaviour is consistently observed not only around the preferred regularization constant value (triangles) but also in the over-?tting case (circles) and the under?tting case (crosses). Moreover, as predicted by Theorem 2, the preferred regularization constant value scales linearly and the squared error scales quadratically to the noise standard deviation ?. As predicted by Lemma

3, the curves for the smaller 50 ? 50 ? 20 tensor and those for the larger 100 ? 100 ? 50 tensor seem to agree when the regularization constant is scaled by the factor two. p Note that the dominant term in inequality (12) is the second term n ? k , which is roughly scaled by the factor two from 50 ? 50 ? 20 to 100 ? 100 ? 50. 4.2

Tensor completion from partial observations

In this subsection, we repeat the simulation originally done by Tomioka et al. [23] and demonstrate that our results in Section 3.3 can precisely predict the empirical scaling behaviour with respect to both the size and rank of a tensor. We present results for both matrix completion (K = 2) and tensor completion (K = 3). For the matrix case, we randomly generated low-rank matrices of dimensions 50 ? 20, 100 ? 40, and 250 ? 200. For the tensor case, we randomly generated low-rank tensors of dimensions 50 ? 50 ? 20 and 100 ? 100 ? 50. We generated the matrices or tensors as in the previous subsection for various ranks. We randomly selected some elements of the true matrix/tensor for training and kept the 7

1

0.8

0.8

0.6 0.4 size=[50 20] size=[100 40] size=[250 200]

0.2 0 0

0.1 0.2 0.3 0.4 0.5 Normalized rank ——n?1—— ——r—— 1/2

Fraction at Error¡=0.01

Fraction at err¡=0.01

1

0.6 0.4 0.2 0 0

0.6

1/2

(a) Matrix completion (K = 2).

size=[50 50 20] size=[100 100 50] 0.2 0.4 0.6 Normalized rank ——n?1——1/2——r——1/2

0.8

(b) Tensor completion (K = 3).

Figure 3: Scaling behaviour of matrix/tensor completion with respect to the size n and the rank r.

remaining elements for testing. No observation noise is added. We used the ADMM for ?as a matrix? and ?constraint? approaches described in [23] to solve the minimization problem (7) for matrix completion and tensor completion, respectively. Since there is no observation noise, we chose the regularization constant ? ? 0. A single experiment for a speci?c size and rank can be visualized as in Figure 1. ?In ?? Figure ?3, ??? we plot the minimum fraction of observations m = M/N that achieved error ? ? W ??? smaller than 0.01 against the normalized rank r? = ?n?1 ?1/2 ?r?1/2 (of the true ten???W F sor) de?ned in Equation (13). The matrix case is plotted in Figure 3(a) and the tensor case is plotted in Figure 3(b). Each series (blue crosses or red circles) corresponds to different matrix/tensor size and each data-point corresponds to a different core size (rank). We can see that the fraction of observations m = M/N scales linearly against the normalized rank r?, which agrees with the condition M/N ? c1 ?n?1 ?1/2 ?r?1/2 = c1 r? in Theorem 3 (see Equation (14)). The agreement is especially good for tensor completion (Figure 3(b)), where the two series almost overlap. Interestingly, we can see that when compared at the same normalized rank, tensor completion is easier than matrix completion. For example, when nk = 50 and rk = 10 for each k = 1, . . . , K, the normalized rank is 0.2. From Figure 3, we can see that we only need to see 30% of the entries in the tensor case to achieve error smaller than 0.01, whereas we need about 60% of the entries in the matrix case.

5

Conclusion

We have analyzed the statistical performance of a tensor decomposition algorithm based on the overlapped Schatten 1-norm regularization (7). Numerical experiments show that our theory can predict the empirical scaling behaviour well. The fraction of observation m = M/N at the threshold predicted by our theory is proportional to the quantity we call the normalized rank, which re?nes conjecture (sum of the mode-k ranks) in [23]. There are numerous directions that the current study can be extended. In this paper, we have focused on the convergence of the estimation error; it would be meaningful to also analyze the condition for the consistency of the estimated rank as in [2]. Second, although we have succeeded in predicting the empirical scaling behaviour, the setting of random Gaussian design does not match the tensor completion setting in Section 4.2. In order to analyze the latter setting, the notion of incoherence in [5] or spikiness in [16] might be useful. This might also explain why tensor completion is easier than matrix completion at the same normalized rank. Moreover, when the target tensor is only low-rank in a certain mode, Schatten 1-norm regularization fails badly (as predicted by the high normalized rank). It would be desirable to analyze the ?Mixture? approach that aims at this case [23]. In a broader context, we believe that the current paper could serve as a basis for re-examining the concept of tensor rank and low-rank approximation of tensors based on convex optimization. Acknowledgments. We would like to thank Franz Kir?aly and Hiroshi Kajino for their valuable comments and discussions. This work was supported in part by MEXT KAKENHI 22700138, 23240019, 23120004, 22700289, and NTT Communication Science Laboratories. 8

2 References

[1] E. Acar and B. Yener. Unsupervised multiway data analysis: A literature survey. IEEE T. Knowl. Data. En., 21(1):6?20, 2009. [2] F.R. Bach. Consistency of trace norm minimization. J. Mach. Learn. Res., 9:1019?1048, 2008. [3] S. Boyd and L. Vandenberghe. Convex Optimization. Cambridge University Press, 2004. [4] R. Bro. PARAFAC. Tutorial and applications. Chemometr. Intell. Lab., 38(2):149?171, 1997. [5] E. J. Candes and B. Recht. Exact matrix completion via convex optimization. Found. Comput. Math., 9(6):717?772,

2009. [6] 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. [7] P. Comon. Tensor decompositions. In J. G. McWhirter and I. K. Proudler, editors, Mathematics in signal processing V. Oxford University Press, 2002. [8] L. De Lathauwer and J. Vandewalle. Dimensionality reduction in higher-order signal processing and rank-(r1 , r2 , . . . , rn) reduction in multilinear algebra. Linear Algebra Appl., 391:31?55, 2004. [9] K. Fukumizu. Generalization error of linear neural networks in unidenti?able cases. In Algorithmic Learning Theory, pages 51?62. Springer, 1999. [10] S. Gandy, B. Recht, and I. Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. Inverse Problems, 27:025010, 2011. [11] J. H?astad. Tensor rank is NP-complete. Journal of Algorithms, 11(4):644?654, 1990. [12] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. SIAM Review, 51(3):455?500, 2009. [13] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. In Prof. ICCV, 2009. [14] M. M?rup. Applications of tensor (multiway array) factorizations and decompositions in data mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1):24?40, 2011. [15] S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A uni?ed framework for high-dimensional analysis of m-estimators with decomposable regularizers. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, Advances in NIPS 22, pages 1348?1356. 2009. [16] S. Negahban and M.J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. Technical report, arXiv:1009.2118, 2010. [17] S. Negahban and M.J. Wainwright. Estimation of (near) low-rank matrices with noise and highdimensional scaling. Ann. Statist., 39(2), 2011. [18] 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. [19] A. Rohde and A.B. Tsybakov. 39(2):887?930, 2011.

Estimation of high-dimensional low-rank matrices.

Ann. Statist.,

[20] N.D. Sidiropoulos, R. Bro, and G.B. Giannakis. Parallel factor analysis in sensor array processing. IEEE T. Signal Proces., 48(8):2377?2388, 2000. [21] 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. [22] N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In Lawrence K. Saul, Yair Weiss, and L?eon Bottou, editors, Advances in NIPS 17, pages 1329?1336. MIT Press, Cambridge, MA, 2005. [23] R. Tomioka, K. Hayashi, and H. Kashima. Estimation of low-rank tensors via convex optimization. Technical report, arXiv:1010.0789, 2011. [24] L. R. Tucker. Some mathematical notes on three-mode factor analysis. Psychometrika, 31(3):279?311, 1966. [25] M. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. Computer Vision?ECCV 2002, pages 447?460, 2002. [26] H. Wang and N. Ahuja. Facial expression decomposition. In Proc. 9th ICCV, pages 958 ?

965, 2003.
9