Convergence rate of Bayesian tensor estimator and its minimax optimality

Taiji Suzuki[†,‡] S-TAIJI@IS.TITECH.AC.JP

[†] Tokyo Institute of Technology, O-okayama 2-12-1, Meguro-ku, Tokyo 152-8552, JAPAN
[‡] PRESTO, Japan Science and Technology Agency, JAPAN

Abstract

We investigate the statistical convergence rate of a Bayesian low-rank tensor estimator, and derive the minimax optimal rate for learning a low-rank tensor. Our problem setting is the regression problem where the regression coefficient forms a tensor structure. This problem setting occurs in many practical applications, such as collaborative filtering, multi-task learning, and spatio-temporal data analysis. The convergence rate of the Bayes tensor estimator is analyzed in terms of both in-sample and out-of-sample predictive accuracies. It is shown that a fast learning rate is achieved without any strong convexity of the observation. Moreover, we show that the method has adaptivity to the unknown rank of the true tensor, that is, the near optimal rate depending on the true rank is achieved even if it is not known a priori. Finally, we show the minimax optimal learning rate for the tensor estimation problem, and thus show that the derived bound of the Bayes estimator is tight and actually near minimax optimal.

1. Introduction

Tensor modeling is a powerful tool for representing higher order relations between several data sources. The second order correlation has been a main tool in data analysis for a long time. However, because of an increase in the variety of data types, we frequently encounter a situation where higher order correlations are important for transferring information between more than two data sources. In this situation, a tensor structure is required. For example, a recommendation system with a three-mode table, such as user $\times$ movie $\times$ context, is regarded as comprising the tensor data analysis (Karatzoglou et al., 2010). The noteworthy success of tensor data analysis is based on the notion of the *low rank* property of a tensor, which is analogous to that of a matrix. The rank of a tensor is defined by a generalized version of the singular value decomposition for matrices. This enables us to decompose a tensor into a few factors and find higher order relations between several data sources.

A naive approach to computing tensor decomposition requires non-convex optimization (Kolda & Bader, 2009). Several authors have proposed convex relaxation methods to overcome the computational difficulty caused by non-convexity (Liu et al., 2009; Signoretto et al., 2010; Gandy et al., 2011; Tomioka et al., 2011; Tomioka & Suzuki, 2013). The main idea of convex relaxations is to unfold a tensor into a matrix, and apply trace norm regularization to the matrix thus obtained. This technique connects low rank tensor estimation to the well-investigated convex low rank matrix estimation. Thus, we can apply the techniques developed in low rank matrix estimation in terms of optimization and statistical theories. To address the theoretical aspects, Tomioka et al. (2011) gave the statistical convergence rate of a convex tensor estimator that utilizes the so-called *overlapped Schatten 1-norm* defined by the sum of the trace norms of all unfolded matricizations. Mu et al. (2014) showed that the bound given by Tomioka et al. (2011) is tight, but can be improved by a modified technique called *square deal*. Tomioka & Suzuki (2013) proposed another approach called *latent Schatten 1-norm* regularization that is defined by the infimum convolution of trace norms of all unfolded matricizations, and analyzed its convergence rate. These theoretical studies revealed the qualitative dependence of learning rates on the rank and size of the underlying tensor. However, one problem of convex methods is that, reducing the problem to a matrix estimation, one may lose statistical efficiency in exchange for computational efficiency.

The main question addressed in this paper is whether we can obtain a tractable method that possesses a (near) optimal learning rate. To answer this question, we consider a Bayesian learning method. Bayesian tensor learning methods have been studied extensively (Chu & Ghahramani, 2009; Xu et al., 2013; Xiong et al., 2010; Rai et al., 2014).

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

Basically, they construct a generative model of the tensor decomposition and place a prior probability on the decomposed components. As in convex methods, Bayesian methods have also been polished to efficiently process large datasets. Their statistical performances have been supported numerically, but only a few theoretical analyses have been provided.

In this paper, we present the learning rate of a Bayes estimator for low-rank tensor regression problems and give the minimax optimal rate for the tensor estimation problem. The prior probability we consider here is the most basic one, which places Gaussian priors on decomposed components and an exponentially decaying prior on the rank (see Xiong et al. (2010); Xu et al. (2013); Rai et al. (2014)). Roughly speaking, we obtain the (near) optimal convergence rate,

$$\|\hat{A}-A^*\|_n^2 = O_p\left(\frac{d^*(\sum_{k=1}^K M_k)\log(K\sqrt{n(\sum_k M_k)^K}}{n}\right),$$

where n is the sample size, $\hat{A}$ is the Bayes estimator, A^* is the true tensor, d^* is the *CP-rank* of the true tensor (its definition will be given in Section 2), and $(M_1,\dots,M_K)$ is the size ($A^* \in \mathbb{R}^{M_1\times\cdots\times M_K}$). Moreover, our analysis has the following favorable properties.

- The rate is proven *without* assuming any strong convexity on the empirical L_2 norm.
- Rank adaptivity is shown, that is, the convergence rate is automatically adjusted to the rank of the true tensor as if we knew it a priori.

In particular, the first property significantly differentiates our approach from existing approaches. A variant of strong convexity, such as restricted strong convexity (Bickel et al., 2009; Negahban et al., 2012) is usually assumed in convex sparse learning in order to derive a fast rate. However, for the analysis of the predictive accuracy of a Bayes estimator, a near optimal rate can be shown without such conditions. This is a remarkable point of the predictive accuracy analysis rather than the parameter estimation accuracy. Finally, we give the minimax optimal learning rate for the tensor estimation problem. That is roughly given by

$$\inf_{\hat{A}}\sup_{A^*\in\mathcal{T}} \mathrm{E}[\|\hat{A}-A^*\|_{L_2(P(X))}^2] \geq C\frac{d^*(\sum_{k=1}^K M_k)}{n}$$

where $\hat{A}$ is any estimator and $\mathcal{T}$ is a set of tensors. It is seen that the derived learning rate is actually near minimax optimal.

To the best of our knowledge, this is the first result that gives the minimax optimal rate of low-rank tensor estimation problem and shows (near) optimality of a computationally tractable learning method.

2. Problem Settings

In this section, the problem setting of this paper is shown. Suppose that there exists the true tensor $A^* \in \mathbb{R}^{M_1\times\cdots\times M_K}$ of order K, and we observe n samples $D_n = \{(Y_i, X_i)\}_{i=1}^n$ from the following linear model:

$$Y_i = \langle A^*, X_i\rangle + \epsilon_i.$$

Here, X_i is a tensor in $\mathbb{R}^{M_1\times\cdots\times M_K}$ and the inner product $\langle\cdot,\cdot\rangle$ between two tensors $A, X \in \mathbb{R}^{M_1\times\cdots\times M_K}$ is defined by $\langle A, X\rangle = \sum_{j_1,\dots,j_K=1}^{M_1,\dots,M_K} A_{j_1,\dots,j_K}X_{j_1,\dots,j_K}$. ϵ_i is i.i.d. noise from a normal distribution $N(0,\sigma^2)$ with mean 0 and variance σ^2.

Now, we assume the true tensor A^* is "low-rank." The notion of rank considered in this paper is *CP-rank* (Canonical Polyadic rank) (Hitchcock, 1927a;b). We say a tensor $A \in \mathbb{R}^{M_1\times\cdots\times M_K}$ has CP-rank d' if there exist matrices $U^{(k)} \in \mathbb{R}^{d'\times M_k}$ $(k = 1,\dots,K)$ such that $A_{j_1,\dots,j_K} = \sum_{r=1}^{d'} U^{(1)}_{r,j_1}U^{(2)}_{r,j_2}\dots U^{(K)}_{r,j_K}$, and d' is the minimum number to yield this decomposition (we do not require the orthogonality of $U^{(k)}$). This is called *CP-decomposition*. When A satisfies this relation for $\boldsymbol{U} = (U^{(1)}, U^{(2)},\dots,U^{(K)})$, we write

$$\begin{aligned} A = A_{\boldsymbol{U}} &= [[U^{(1)}, U^{(2)},\dots,U^{(K)}]] \\ &=: \left(\sum\nolimits_{r=1}^{d'} U^{(1)}_{r,j_1}U^{(2)}_{r,j_2}\dots U^{(K)}_{r,j_K}\right)_{j_1,\dots,j_K}. \end{aligned} \quad (1)$$

We denote by d^* the CP-rank of the true tensor A^*. Notice that, for the special case of matrices ($K = 2$), the CP-rank coincides with the usual rank of a matrix.

In this paper, we investigate the predictive accuracy of the linear model with the assumption that A^* has low CP-rank. Because of the low CP-rank assumption, the learning problem becomes more structured than an ordinary linear regression problem on a vector. This problem setting includes the well-known low rank matrix estimation as a special case $K = 2$. There are two types of predictive accuracy: *in-sample* and *out-of-sample* ones. The in-sample predictive accuracy of an estimator $\hat{A}$ is defined by

$$\|\hat{A}-A^*\|_n^2 := \frac{1}{n}\sum_{i=1}^n \langle X_i, \hat{A}-A^*\rangle^2, \quad (2)$$

where $\{X_i\}_{i=1}^n$ is the observed input samples. The out-of-sample one is defined by

$$\|\hat{A}-A^*\|_{L_2(P(X))}^2 := \mathrm{E}_{X\sim P(X)}[\langle X, \hat{A}-A^*\rangle^2], \quad (3)$$

where $P(X)$ is the distribution of X that generates the observed samples $\{X_i\}_{i=1}^n$ and the expectation is taken over independent realization X from the observed ones.

Example 1. Tensor completion under random sampling. *Suppose that we have partial observations of a tensor. A tensor completion problem consists of denoising the observational noise and completing the unobserved elements. In this problem, X_i is independently identically distributed from a set $\{e_{j_1,\dots,j_K} \mid 1 \leq j_k \leq M_k \ (k = 1,\dots,K)\}$, where $e_{j_1,\dots,j_K}$ is an indicator tensor that has 1 at its $(j_1,\dots,j_K)$-element and 0 elsewhere, and thus, Y_i is an observation of one element of A^* contaminated with noise ϵ_i.*

The out-of-sample accuracy measures how accurately we can recover the underlying tensor A^* from the partial observation. If X_i is uniformly distributed, $\|\hat{A} - A^*\|_{L_2(P(X))} = \frac{1}{\sqrt{M_1 \dots M_K}} \|\hat{A} - A^*\|_2$, where $\|\cdot\|_2$ is the ℓ_2-norm obtained by summing the squares of all the elements. If $K = 2$, this problem is reduced to the standard matrix completion problem. In that sense, our problem setting is a wide generalization of the low rank matrix completion problem.

Example 2. Multi-task learning. *Suppose that several tasks are aligned across a 2-dimensional space. For each task $(s,t) \in \{1,\dots,M_1\} \times \{1,\dots,M_2\}$ (indexed by two numbers), there is a true weight vector $a^*_{(s,t)} \in \mathbb{R}^{M_3}$. The tensor A^* is an array of the weight vectors $a^*_{(s,t)}$, that is, $A^*_{s,t,j} = a^*_{(s,t),j}$.*

The input vector X_i is a vector of predictor variables for one specific task, say (s,t), and takes a form such that $X_{i,(s',t',:)} = \begin{cases} x_i^{(s,t)} \in \mathbb{R}^{M_3}, & ((s',t') = (s,t)), \\ \mathbf{0}, & (\textit{otherwise}). \end{cases}$

By assuming A^* is low-rank in the sense of CP-rank, the problem becomes a multi-task feature learning with a two dimensional structure in the task space (Romera-Paredes et al., 2013).

As shown in the examples, the estimation problem of low-rank tensor A^* is a natural extension of low-rank matrix estimation. However, it has a much richer structure than matrix estimation. Thus far, some convex regularized learning problems have been proposed analogously to spectrum regularization on a matrix, and their theoretical analysis has also been provided. However, no method have been proved to be statistically optimal. There is a huge gap between a matrix and higher order array. One reason for this gap is the computational complexity of the convex envelope of CP-rank. It is well known that the trace norm of a matrix is a convex envelope of the matrix rank on a set of matrices with a restricted operator norm (Srebro et al., 2005). However, as for tensors, computing CP-rank and CP-decomposition themselves is NP-hard (Hillar & Lim, 2013).

In this paper, we investigate a Bayes estimator instead of the convex regularized one. It will be shown that our Bayes estimator shows a near optimal convergence rate with a much weaker assumption, while the learning procedure is computationally tractable. The rate is much improved as compared to that of the existing estimators.

3. Bayesian tensor estimator

We now provide the prior distribution of the Bayes estimator that is investigated in this paper. On a decomposition of a rank d' tensor $A = [[U^{(1)}, U^{(2)}, \dots, U^{(K)}]]$ ($U^{(k)} \in \mathbb{R}^{d' \times M_k}$), we place a Gaussian prior:

$$\pi(U^{(1)},\dots,U^{(K)}|d') \propto \exp\left\{-\frac{d'}{2\sigma_{\mathrm{p}}^2} \sum_{k=1}^{K} \mathrm{Tr}[{U^{(k)}}^\top U^{(k)}]\right\},$$

where $\sigma_{\mathrm{p}} > 0$. Moreover, we placed a prior distribution on the rank $1 \leq d' \leq d_{\max}$ as

$$\pi(d') = \tfrac{1}{N_\xi} \xi^{d'(M_1+\dots+M_K)},$$

where $0 < \xi < 1$ is some positive real number, $d_{\max}$ is a sufficiently large number that is supposed to be larger than d^*, and N_ξ is the normalizing constant, $N_\xi = \frac{1-\xi^{M_1+\dots+M_K}}{\xi - \xi^{d_{\max}(M_1+\dots+M_K)}}$.

Now, since the noise is Gaussian, the likelihood of a tensor A is given by

$$p(D_n|A) =: p_{n,A} \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \langle A, X_i\rangle)^2\right\}.$$

The posterior distribution is given by

$$\Pi(A \in \mathcal{C}|D_n) = \frac{\sum_{d=1}^{d_{\max}} \int_{A_U \in \mathcal{C}} p_{n,A_U} \pi((U^{(k)})_{k=1}^K|d)\pi(d)\mathrm{d}U^{(1)}\dots\mathrm{d}U^{(K)}}{\sum_{d=1}^{d_{\max}} \int p_{n,A_U} \pi((U^{(k)})_{k=1}^K|d)\pi(d)\mathrm{d}U^{(1)}\dots\mathrm{d}U^{(K)}},$$

where $\mathcal{C} \subseteq \mathbb{R}^{M_1 \times \dots M_K}$. It is noteworthy that the posterior distribution of $U^{(k)}$ conditioned by d and $U^{(k')}$ ($k' \neq k$) is a Gaussian distribution, because the prior is conjugate to Gaussian distributions. Therefore, the posterior mean $\int_A f(A)\Pi(\mathrm{d}A|D_n)$ of a function $f : \mathbb{R}^{M_1 \times \dots M_K} \to \mathbb{R}$ can be computed by an MCMC method, such as Gibbs sampling (as for the Bayes tensor estimator, see Xiong et al. (2010); Xu et al. (2013); Rai et al. (2014)). In this paper, we consider the posterior mean estimator $\hat{A} = \int A\Pi(\mathrm{d}A|D_n)$ which is the Bayes estimator corresponding to the square loss.

4. Convergence rate analysis

In this section, we present the statistical convergence rate of the Bayes estimator. Before we give the convergence rate,

we define some quantities and provide the assumptions. We define the *max-norm* of A^* as

$$\|A^*\|_{\max,2} := \min_{\{U^{(k)}\}} \{\max_{i,k} \|U^{(k)}_{:,i}\|_2 \mid A^* = [[U^{(1)}, \dots, U^{(K)}]],\ U^{(k)} \in \mathbb{R}^{d^* \times M_k}\}.$$

The ℓ_p-norm of a tensor A is given by $\|A\|_p := (\sum_{j_1,\dots,j_K} |A_{j_1,\dots,j_K}|^p)^{\frac{1}{p}}$. The prior mass around the true tensor A^* is a key quantity for characterizing the convergence rate, which is denoted by Ξ:

$$\Xi(\delta) := -\log(\Pi(A : \|A - A^*\|_n < \delta)),$$

where $\delta > 0$. Small $\Xi(\delta)$ means that the prior is well concentrated around the truth. Thus, it is natural to consider that, if Ξ is large, the Bayes estimator could be close to the truth. However, clearly, we do not know beforehand the location of the truth. Thus, it is not beneficial to place too much prior mass around one specific point. Instead, the prior mass should cover a wide range of possibilities of A^*. The balance between concentration and dispersion has a similar meaning to that of the bias-variance trade-off.

To normalize the scale, we assume that the ℓ_1-norm of X_i is bounded[1].

Assumption 1. *We assume that the ℓ_1-norm of X_i is bounded by 1: $\|X_i\|_1 \le 1$, a.s...*

For theoretical simplicity, we utilize the unnormalized prior on the rank[2], that is $\pi(d) = \xi^{d(\sum_k M_k)}$. It should be noted that removing the normalization constant does not affect the posterior. Under these assumptions, $\Xi(\delta)$ can be bounded as follows.

Lemma 1. *Under Assumption 1, the prior mass Ξ has the bound*

$$\Xi(\tfrac{r}{\sqrt{n}}) \le \frac{d^* \sum_{k=1}^K \|U^{*(k)}\|_F^2}{2\sigma_{\mathrm{p}}^2} + d^* \left(\sum_{k=1}^K M_k\right) \log\left[\frac{6}{\xi}\left(\frac{\sqrt{n}\sigma_{\mathrm{p}}^K K \left(\frac{\|A^*\|_{\max,2}}{\sigma_{\mathrm{p}}} + 1\right)^{K-1}}{r} \vee 1\right)\right].$$

for all $r > 0$, where $\{U^{*(k)}\}_k$ are any tensors satisfying $A^* = [[U^{*(1)}, \dots, U^{*(K)}]]$.

It should be noted that $\Xi(r/\sqrt{n}) \le \Xi(1/\sqrt{n})$ for all $r > 1$. Finally, we define the technical quantities $C_{n,K} := 3K\sqrt{n}\left(\frac{4\sigma_{\mathrm{p}}^2 \Xi(\frac{1}{\sqrt{n}})}{d^*}\right)^{\frac{K}{2}}$ and $c_\xi := \min\{|\log(\xi)|/\log(C_{n,K}), 1\}/4$.

[1] ℓ_1-norm could be replaced by another norm such as ℓ_2. This difference affects the analysis of out-of-sample accuracies, but rejecting samples with $\max_X |\langle X, A\rangle| \le R$ gives an analogous result for other norms.

[2] This is not essential, but for just making the expression of Ξ simple.

4.1. In-sample predictive accuracy

We now give the convergence rate of the in-sample predictive accuracy. We suppose the inputs $\{X_i\}_{i=1}^n$ are fixed (not random). The in-sample predictive accuracy conditioned by $\{X_i\}_{i=1}^n$ is given as follows.

Theorem 1. *Under Assumption 1, there exists a universal constant C such that the posterior mean of the in-sample accuracy is upper bounded by*

$$\mathrm{E}\left[\int \|A - A^*\|_n^2 \mathrm{d}\Pi(A|Y_{1:n})\right] \le \frac{C}{n}\left[d^*\left(\sum\nolimits_k M_k + \frac{1}{|\log(\xi)|}\right)\log(C_{n,K}) + \frac{\Xi(\sqrt{\frac{c_\xi}{n}})}{c_\xi} + \log(d_{\max}) + K + 8^K(K+1)!\right]. \quad (4)$$

The proof is given in Appendix A.2 in the supplementary material. This theorem provides the speed at which the posterior mass concentrates around the true A^*. It should be noted that the integral in the LHS of Eq. (4) is taken *outside* $\|A - A^*\|_n^2$. This gives not only information about the posterior concentration but also the convergence rate of the posterior mean estimator. This can be shown as follows. By Jensen's inequality, we have $\mathrm{E}\left[\|\int A \mathrm{d}\Pi(A|Y_{1:n}) - A^*\|_n^2\right] \le \mathrm{E}\left[\int \|A - A^*\|_n^2 \mathrm{d}\Pi(A|Y_{1:n})\right]$. Therefore, Theorem 1 gives a much stronger claim on the posterior than just stating the convergence rate of the posterior mean estimator.

Since the rate (4) is rather complicated, we give a simplified bound. By assuming $\log(d_{\max})$ and $K!$ are smaller than $d^*(\sum_k M_k)$, we rearrange it as

$$\mathrm{E}\left[\int \|A - A^*\|_n^2 \mathrm{d}\Pi(A|Y_{1:n})\right] = O\left(\frac{d^*(M_1 + \cdots + M_K)}{n} \log\left(K\sqrt{n(\textstyle\sum_{k=1}^K M_k)^K}\frac{\sigma_{\mathrm{p}}^K}{\xi}\right)\right).$$

Inside the $O(\cdot)$ symbol, a constant factor depending on $K, \|A^*\|_{\max,2}, \sigma_{\mathrm{p}}, \xi$ is hidden. This bound means that the convergence rate is characterized by the actual degree of freedom up to a log term. That is, since the true tensor has rank d^* and thus has a decomposition (1), the number of unknown parameters is bounded by $d^*(M_1 + \cdots + M_K)$. Thus, the rate is basically $O(\frac{\text{degree of freedom}}{n})$ (up to $\log$ order), which is optimal (see Section 5 for more precise argument). Here, we would like to emphasize that the true rank d^* is unknown, but by placing a prior distribution on a rank the Bayes estimator can appropriately estimate the rank and gives an almost optimal rate depending on the true rank. In this sense, the Bayes estimator has adaptivity to the true rank.

More importantly, we do not assume *any strong convexity* on the design. Usually, to derive a fast convergence rate of sparse estimators, such as Lasso and the trace norm regularization estimator, we assume a variant of strong convexity, such as a restricted eigenvalue condition (Bickel et al., 2009) and restricted strong convexity (Negahban et al., 2012). It is difficult to check the strong convexity condition in practice. However, our convergence rate does not require such conditions. One reason why this is possible is that we are interested in the predictive accuracies rather than the actual distance between the tensors $\|A - A^*\|_2^2$ (parameter estimation accuracy) . It is known that this phenomenon occurs also in high dimensional regression of vectors, see Dalalyan & Tsybakov (2008) for example.

4.2. Out-of-sample predictive accuracy

Next, we turn to the convergence rate of the out-of-sample predictive accuracy. In this setting, the input sequence $\{X_i\}_{i=1}^n$ is not fixed, but an i.i.d. random variable generated by a distribution $P(X)$.

To obtain fast convergence of the out-of-sample accuracy, we need to bound the difference between the empirical and population L_2-errors: $\|A - A^*\|_n^2 - \|A - A^*\|_{L_2(P(X))}^2$. To ensure that this quantity is small using Bernstein's inequality, $\max_X |\langle X, A\rangle|$ should be bounded. However, the infinity norm of the posterior mean could be large in tensor estimation. This difficulty can be avoided by rejecting posterior sample A with a large infinity norm.

4.2.1. Infinity norm thresholding

Now, define $\|A\|_\infty = \max_{j_1,\dots,j_K} |A_{j_1,\dots,j_K}|$. Then, under Assumption 1, we have $\langle X, A\rangle \leq \|A\|_\infty$. Here, we assume that the infinity norm $\|A^*\|_\infty$ of the true tensor is approximately known, that is, we know $R > 0$ such that $2\|A^*\|_\infty < R$. This is usually true. For example, we know the upper bound in tensor completion for a recommendation system. Otherwise, we may apply cross validation. Our strategy is to put the infinity norm restriction on the prior, that is, we utilize the "truncated" prior the support of which is restricted to $\|A\|_\infty \leq R$. The estimation with this prior can be implemented merely by rejecting the posterior samples with an infinity norm larger than a threshold R during the sampling scheme.

The resultant posterior distribution is expressed as the conditional posterior distribution $\Pi(\cdot|\|A\|_\infty \leq R, D_n)$. Accordingly, we investigate the out-of-sample accuracy of the conditional posterior:

$$\mathrm{E}_{D_n}\left[\int \|A - A^*\|_{L_2(P(X))}^2 \mathrm{d}\Pi(A|\|A\|_\infty \leq R, D_n)\right].$$

Theorem 2. *Suppose Assumption 1 and $\|A^*\|_\infty < \frac{1}{2}R$ are satisfied, then the out-of-sample accuracy is bounded as*

$$\begin{aligned}
&\mathrm{E}_{D_n}\left[\int \|A - A^*\|_{L_2(P(X))}^2 \mathrm{d}\Pi(A|\|A\|_\infty \leq R, D_n)\right] \\
\leq &\frac{C(R^2 \vee 1)}{n}\left[d^*\left(\sum\nolimits_k M_k + \frac{3}{|\log(\xi)|}\right)\log(C_{n,K})\right. \\
&\left. + \frac{\Xi(\sqrt{\frac{c_\xi}{n}})}{c_\xi} + \log(d_{\max}) + 8^K(K+1)!\right],
\end{aligned}$$

where C is a universal constant.

The proof is given in Appendix B in the supplementary material.. The only difference between this and Theorem 1 is that R^2 appears in front of the bound. The source of this factor is the gap between the empirical and population L_2-norms. Here again, the convergence rate can be simplified as

$$\begin{aligned}
&\mathrm{E}_{D_n}\left[\int \|A - A^*\|_{L_2(P(X))}^2 \mathrm{d}\Pi(A|\|A\|_\infty \leq R, D_n)\right] \\
\leq &O\left(\frac{d^*(M_1 + \cdots + M_K)}{n}(R^2 \vee 1)\times\right. \\
&\left.\log\left(K\sqrt{n(\textstyle\sum_{k=1}^K M_k)^K}\frac{\sigma_{\mathrm{p}}^K}{\xi}\right)\right), \qquad (5)
\end{aligned}$$

if $K!$ and $|\log(\xi)|$ are smaller than $d^*(\sum_k M_k)$. Here, we observe that the convergence rate achieved is optimal up to the log-term. We would like to emphasize again that the optimal rate is achieved, although we do not assume any strong convexity on the distribution $L_2(\Pi)$. This can be so because we are not analyzing the actual L_2-norm $\|A - A^*\|_2$. If we do not assume strong convexity like $\|A - A^*\|_2 \leq C\|A - A^*\|_{L_2(P(X))}$, it is impossible to derive fast convergence of $\|A - A^*\|_2$. The trick is that we focus on the "weighted" L_2-norm $\|A - A^*\|_{L_2(P(X))}$ instead of $\|A - A^*\|_2$.

Finally, it is remarked that, if X_i is the uniform at random observation in the tensor completion problem, then $\|A - A^*\|_{L_2(P(X))}^2 = \frac{1}{\prod_k M_k}\|A - A^*\|_2^2$ (note that in this setting $\|X_i\|_1 = 1$). Thus, our analysis yields fast convergence of the tensor recovery. If $K = 2$, the analysis recovers the well known rate of matrix completion problems up to a $\log(nM_1M_2)$ term (Negahban & Wainwright, 2012; Negahban et al., 2012; Rohde & Tsybakov, 2011):

$$\frac{1}{M_1M_2}\|\hat{A} - A^*\|_2^2 = O_p\left(\frac{d^*(M_1 + M_2)}{n}\log(nM_1M_2)\right).$$

4.2.2. Max-norm thresholding

Finally, we briefly describe the convergence rate of the Bayes estimator based on the rejection sampling with respect to restricted *max-norm*. We reject the posterior sample with a max-norm larger than R; that is, we accept

only a sample $A_{\boldsymbol{U}}$ that satisfies $\boldsymbol{U} \in \{(U^{(1)}, \dots, U^{(K)}) \mid \|U^{(k)}_{:,j}\| \leq R \, (1 \leq k \leq K, 1 \leq j \leq M_k)\} =: \mathcal{U}_R$. Then, we have the following bound.

Theorem 3. *Under Assumption 1 and $\|A^*\|_{\max,2} + \sigma_{\mathrm{p}} < R$, we have*

$$\mathrm{E}_{D_n}\left[\int \|A_{\boldsymbol{U}} - A^*\|^2_{L_2(P(X))} \mathrm{d}\Pi(A_{\boldsymbol{U}} | \boldsymbol{U} \in \mathcal{U}_R, D_n)\right] \leq C\left(\frac{d^*(\sum_{k=1}^K M_k)}{n}(1 \vee R^{2K}) \log\left(K\sqrt{n} R^{\frac{K}{2}} \frac{\sigma_{\mathrm{p}}^K}{\xi}\right)\right),$$

where C is a constant depending on $K, \log(d_{\max}), \sigma_{\mathrm{p}}$.

The proof is given in Appendix C in the supplementary material. In a setting where M_k is much larger than R, this bound gives a much better rate than the previous ones, because the term inside $\log$ is improved from $\prod M_k^{\frac{1}{2}}$ to $R^{\frac{K}{2}}$. On the other hand, the rejection rate during the sampling would be increased.

5. Minimax optimality

In this section, the learning rates derived above are actually minimax optimal up to $\log$ terms. To prove this, we specify the $L_2(P(X))$ norm. We take $L_2(P(X))$ as a uniform observation of the entries. That is, $\langle X, A \rangle = A_{i_1, i_2, \dots, i_K}$ for some $i_k \in \{1, \dots, M_k\}$ $(k = 1, \dots, K)$, and the choice of $(i_1, i_2, \dots, i_K)$ is uniform. The hypothesis space is given by

$$\mathcal{T}_R = \{[[U^{(1)}, \dots, U^{(K)}]] \mid U^{(k)} \in \mathbb{R}^{d^* \times M_k}, \|U^{(k)}_{:,j}\| \leq R \, (1 \leq k \leq K, 1 \leq j \leq M_k)\},$$

the set of tensors with rank d^* and the max norm not more than R. The minimax optimal rate is the convergence rate that can not be improved by *any* estimator: for any estimator, there is a tensor $A^* \in \mathcal{T}_R$ such that the predictive accuracy corresponding to the true tensor A^* is lower bounded by the minimax optimal risk.

Theorem 4. *The minimax learning rate of the tensor estimation is lower bounded as follows. Suppose that $R \geq 1$, $M_k > 4$ $(\forall k < K)$, $M_K / d^* > 4$ and M_K / d^* is integer. Then there exists a constant C such that*

$$\inf_{\hat{A}} \sup_{A^* \in \mathcal{T}_R} \mathrm{E}[\|\hat{A} - A^*\|^2_{L_2(P(X))}] \geq C \min\left\{\sigma^2\left(\frac{d^*(\sum_{k=1}^K M_k)}{n}\right), (R^2/d^*)^K\right\}$$

where $\inf_{\hat{A}}$ is taken over all estimator and the expectation is taken for the training samples.

The proof is given in Appendix E in the supplementary material. The theorem is proven by using the information theoretic argument developed by Yang & Barron (1999) to derive the minimax optimal rate. The assumption, M_K / d^* is integer, is just a technical assumption and is not essential. We can see that the convergence rates given in Theorems 2 and 3 are minimax optimal upto $\log$ and R^{2K} terms. Note that R in Theorem 2 is about the infinity norm, not max-norm. The infinity norm $\|A\|_\infty$ is roughly bounded by $\|A\|_\infty \leq \|A\|^K_{\max,2}$. Thus there is no contradiction. This result supports the use of the Bayes estimators.

6. Related works

In this section, we describe the existing works and clarify their relation to our work. Recently, theoretical analyses of convex regularized low-rank tensor estimators have been developed. The pioneering work (Tomioka et al., 2011) analyzed a method that utilizes unfolded matricization of a tensor. Let $M = \prod_{k=1}^K M_k$ (the number of whole elements). The authors used so-called *overlapped Schatten 1-norm* regularization $\sum_{k=1}^K \|A^{(k)}\|_{\mathrm{Tr}}$ where $\|\cdot\|_{\mathrm{Tr}}$ is the trace norm and $A^{(k)} \in \mathbb{R}^{M_k \times M/M_k}$ is the *mode-k unfolding* of a tensor A that is a matrix obtained by unfolding the tensor with the k-th index fixed. Their analysis assumes the true tensor A^* has a low *Tucker-rank* (Tucker, 1966). Tucker-rank is a general notion of CP-rank. In this sense, their analysis is more general than ours. However, strong convexity of the empirical and population L_2-norm is assumed. Under this setting and $n = M$, the following bound is obtained:

$$\frac{1}{M}\|\hat{A} - A^*\|_2^2 \leq C \frac{d^*}{n}\left(\frac{1}{K}\sum_{k=1}^K \sqrt{\frac{M}{M_k}}\right)^2. \quad (6)$$

It can be seen that our bound $\frac{d^*(\sum_{k=1}^K M_k)}{n}\log(nM)$ is smaller than this bound; in particular, if M_k is large, the difference is significantly large. In Mu et al. (2014), it was shown that the bound (6) is tight and cannot be improved if the overlapped Schatten 1-norm is used. In Mu et al. (2014), a novel method called *square deal* was proposed and was shown to achieve the following rate. For $n = M$,

$$\frac{1}{M}\|\hat{A} - A^*\|_2^2 \leq C \frac{d^*}{n}\left(\prod_{k \in I_1} M_k + \prod_{k \in I_2} M_k\right),$$

where I_1 and I_2 are any disjoint decomposition of index set $\{1, \dots, K\}$. This improves the rate (6), but is still larger than the rate of the Bayes estimator, because the product of M_k appears instead of the sum.

Another study on the regularization approach was presented in Tomioka & Suzuki (2013). The authors proposed using the *latent Schatten 1-norm*: $\inf_{\{A_k\}: A = \sum_{k=1}^K A_k} \sum_{k=1}^K \|A_k^{(k)}\|_{\mathrm{Tr}}$, where $A_k^{(k)} \in \mathbb{R}^{M_k \times M/M_k}$ is the mode-k unfolding of the tensor A_k. A nice point of this method is that it automatically finds the minimum rank direction, that is, the mode-k unfolding $A^{*(k)}$ with the minimum rank. It was shown that

the rate is

$$\frac{1}{M}\|\hat{A} - A^*\|_2^2 \leq C\frac{d^* \max_{k=1}^K \{M_k + M/M_k\}}{n}.$$

This rate is also larger than that of the Bayes estimator. We would like to remark that this rate is obtained for the low "Tucker-rank" situation, which is more general than our low CP-rank setting, and thus, it is not best suited to our situation. However, it is not apparent that the latent Schatten 1-norm achieves the same rate as that of the Bayes estimator in low CP-rank settings. Recently, an extention of the overlapped and latent Schatten 1-norms, called *scaled trace norm*, has been proposed by Wimalawarne et al. (2014).

As for Bayesian counter part, a Bayesian low rank matrix estimator is analyzed in Alquier (2013). The prior in this study is similar to ours with $K = 2$, but, instead of placing prior on the rank, the authors placed a Gamma prior on the variance of the Gaussian prior. They utilized the novel PAC-Bayes technique (McAllester, 1998; Catoni, 2004) to show

$$\|\hat{A} - A^*\|_n^2 \leq C\frac{d^*(M_1 + M_2)\log(nM_1M_2)}{n}.$$

This work has a similar flavor to ours in the sense that no strong convexity is required to obtain the convergence rate (see also Dalalyan & Tsybakov (2008) for the use of Bayes estimator in high dimensional regression). However, our analysis deals with general tensor estimation ($K \geq 3$) and the posterior concentration is also given ($\int \|A - A^*\|_n^2 \mathrm{d}\Pi(A|D_n)$ instead of $\|\hat{A} - A^*\|_n^2$ where $\hat{A}$ is the posterior mean). More recently, Mai & Alquier (2015) established a PAC-Bayes bound of the out-of-sample accuracy of the low rank matrix estimation. However, the analysis for $K \geq 3$ is not covered by there analysis. As for the Bayesian tensor estimator, in Zhou et al. (2013) a Bayes estimator of probabilistic tensors was investigated. The model applies fully observed multinomial random variables and the rank is determined beforehand. Therefore, the setting is different from ours. Yang & Dunson (2013) investigated classification of high-dimensional tensors where the covariate is categorical and its distribution is characterized by a low rank probabilistic tensor. However, all distributions are multinomial, and thus the setting is different from our regression problem.

7. Numerical experiments

We now present numerical experiments to justify our theoretical results. Two experiments are executed: the first one is a comparison between the convex optimiation method and the Bayes method, and the second one is verifying the convergence rate. The problem is the tensor completion problem where each observation is a random selection of one element of A^* with observational noise $N(0, 1)$ (see Example 1). The true tensor A^* was randomly generated such that each element of $U^{(k)}$ $(k = 1, \ldots, K)$ was uniformly distributed on $[-1, 1]$. σ_{p} was set at 5, and the true tensor was estimated by the posterior mean obtained by the rejection sampling scheme with $R = 10$. $d_{\max}$ and ξ were set at 10 and 0.5. The posterior sampling was terminated after 500 iterations. The experiments were executed in five different settings, called settings 1 to 5: $\{(M_1, \ldots, M_K), d^*\} = \{(10, 10, 10), 4\}$, $\{(10, 10, 40), 5\}, \{(20, 20, 30), 8\}$, $\{(20, 30, 40), 5\}$, $\{(30, 30, 40), 6\}$. For each setting, we repeated the experiments five times and computed the average of the in-sample predictive accuracy and out-of-sample accuracy over all five repetitions. The number of samples was chosen as $n = n_{\mathrm{s}} \prod_k M_k$, where n_{s} varied from 0.3 to 0.9.

7.1. Comparison with a convex approach

We compare the Bayes estimator with the overlapped Schatten 1-norm regularization approach (Tomioka et al., 2011). The comparison is executed in the settings 2 and 5. As for the regularization parameter of the convex regularized approach, we have chosen the best parameter at each sample size and each problem setting. The dashed lines correspond to the convex approach, and the solid line correspond to the Bayes approach. The accuracies of both methods are improved as the sample size increases. It can be seen that the Bayes approach much outperforms the convex approach in terms of both in-sample and out-of-sample accuracies.

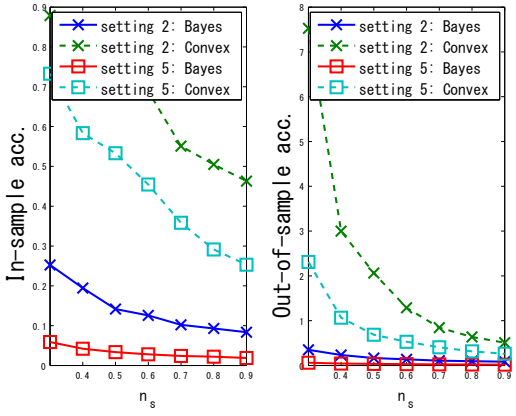

Figure 1. In-sample and out-of-sample accuracy comparison between the convex regularization approach and the Bayes approach, averaged over five repetitions.

7.2. Verification of the convergence rate

To verify our convergence analysis of the Bayes estimator, we consider the "scaled" accuracy in addition to the actual

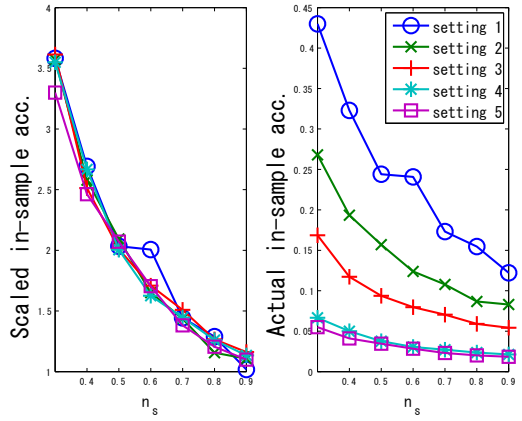

Figure 2. Scaled in-sample accuracy (left) and actual in-sample accuracy (right) versus n_s, averaged over five repetitions.

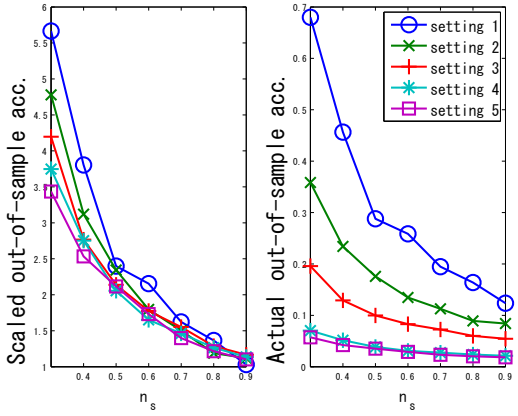

Figure 3. Scaled out-of-sample accuracy (left) and actual out-of-sample accuracy (right) versus n_s, averaged over five repetitions.

accuracy. The scaled accuracy is defined by $\|\hat{A} - A^*\|_n^2 \times \left(\frac{\prod_k M_k}{d^*(\sum_k M_k)}\right)$; the scaled out-of-sample accuracy is also defined in the same manner. Figure 2 shows the in-sample accuracies and the scaled in-sample accuracies against the sample ratio $n_{\rm s}$. The same plot for the out-of-sample accuracy is shown in Figure 3. It can be seen that the curves of the scaled accuracies in all settings are satisfactorily overlapped. This means that our bound accurately describes the sample complexity of the Bayesian tensor estimator, because according to our bounds the scaled accuracies should behave as $1/n_s$ up to a constant factor (and a $\log$ term). The figures show that the scaling factor given by our theories is well matched to the actual predictive accuracy.

8. Conclusion and discussions

In this paper, we investigated the statistical convergence rate of a Bayesian low rank tensor estimator. The notion of a tensor's rank in this paper was based on CP-rank. It is noteworthy that the predictive accuracy was derived *without* any strong convexity assumption. Moreover, we showed the minimax optimal rate of the out-of-sample predictive accuracy. The minimax rate confirms that the obtained bound is (near) optimal. It was also shown that the Bayes estimator has adaptivity to the unknown rank. Numerical experiments showed that our theories indeed describe the actual behavior of the Bayes estimator.

Our bound includes the $\log$ term, which is not negligible when K is large. However, numerical experiments showed that the scaling factor without the $\log$ term explains well the actual behavior. The $\log$ term could be removed by assuming a specific condition on the distribution of X. Clarification of this issue is an important future work.

Acknowledgment

We would like to thank Ryota Tomioka, Pierre Alquier and anonymous referees for suggestive comments. This work was partially supported by MEXT Kakenhi (25730013, 25120012, and 26280009), JST-PRESTO and JST-CREST.

References

Alquier, P. Bayesian methods for low-rank matrix estimation: Short survey and theoretical study. In *Algorithmic Learning Theory*, volume 8139 of *Lecture Notes in Artificial Intelligence*, pp. 309–323. Springer-Verlag, 2013.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Catoni, O. *Statistical Learning Theory and Stochastic Optimization*. Lecture Notes in Mathematics. Springer, 2004. Saint-Flour Summer School on Probability Theory 2001.

Chu, W. and Ghahramani, Z. Probabilistic models for incomplete multi-dimensional arrays. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5 of *JMLR Workshop and Conference Proceedings*, 2009.

Dalalyan, A. S. and Tsybakov, A. B. Aggregation by exponential weighting sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72:39–61, 2008.

Dasgupta, S. and Gupta, A. An elementary proof of the Johnson-Lindenstrauss lemma. *Random Structures and Algorithms*, 22:60–65, 2002.

Gandy, S., Recht, B., and Yamada, I. Tensor completion

and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27:025010, 2011.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. Convergence rates of posterior distributions. *The Annals of Statistics*, 2000(2):500–531, 2000.

Hillar, C. J. and Lim, L.-H. Most tensor problems are np-hard. *Journal of the ACM*, 60(6):45:1–45:39, 2013.

Hitchcock, F. L. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6:164–189, 1927a.

Hitchcock, F. L. Multilple invariants and generalized rank of a p-way matrix or tensor. *Journal of Mathematics and Physics*, 7:39–79, 1927b.

Karatzoglou, A., Amatriain, X., Baltrunas, L., and Oliver, N. Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the 4th ACM Conference on Recommender Systems 2010*, pp. 79–86, 2010.

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

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

Li, W. V. and Linde, W. Approximation, metric entropy and small ball estimates for gaussian measures. *The Annals of Probability*, 27(3):1556–1578, 1999.

Liu, J., Musialski, P., Wonka, P., and Ye, J. Tensor completion for estimating missing values in visual data. In *Proceedings of the 12th International Conference on Computer Vision (ICCV)*, pp. 2114–2121, 2009.

Mai, T. T. and Alquier, P. A bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electronic Journal of Statistics*, 9:823–841, 2015.

McAllester, D. Some PAC-Bayesian theorems. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pp. 230–234, 1998.

Mu, C., Huang, B., Wright, J., and Goldfarb, D. Square deal: Lower bounds and improved relaxations for tensor recovery. In *Proceedings of the 31th International Conference on Machine Learning*, pp. 73–81, 2014.

Negahban, S. and Wainwright, M. J. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13: 1665–1697, 2012.

Negahban, S., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Rai, P., Wang, Y., Guo, S., Chen, G., Dunson, D., and Carin, L. Scalable Bayesian low-rank decomposition of incomplete multiway tensors. In *Proceedings of the 31th International Conference on Machine Learning*, volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 1800–1808, 2014.

Rohde, A. and Tsybakov, A. B. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

Romera-Paredes, B., Aung, H., Bianchi-Berthouze, N., and Pontil, M. Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 1444–1452, 2013.

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

Srebro, N., Rennie, J., and Jaakkola, T. Maximum margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, pp. 1329–1336. MIT Press, 2005.

Tomioka, R. and Suzuki, T. Convex tensor decomposition via structured schatten norm regularization. In *Advances in Neural Information Processing Systems 26*, pp. 1331–1339, 2013. NIPS2013.

Tomioka, R., Suzuki, T., Hayashi, K., and Kashima, H. Statistical performance of convex tensor decomposition. In *Advances in Neural Information Processing Systems 24*, pp. 972–980, 2011. NIPS2011.

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

van der Vaart, A. W. and van Zanten, J. H. Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12:2095–2119, 2011.

Wimalawarne, K., Sugiyama, M., and Tomioka, R. Multitask learning meets tensor factorization: task imputation via convex optimization. In *Advances in Neural Information Processing Systems 27*, pp. 2825–2833. 2014.

Xiong, L., Chen, X., Huang, T.-K., Schneider, J., and Carbonell, J. G. Temporal collaborative filtering with bayesian probabilistic tensor factorization. In *Proceedings of SIAM Data Mining*, pp. 211–222, 2010.

Xu, Z., Yan, F., and Qi, Y. Bayesian nonparametric models for multiway data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99:1, 2013. PrePrints.

Yang, Y. and Barron, A. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

Yang, Y. and Dunson, D. B. Bayesian conditional tensor factorizations for high-dimensional classification, 2013. arXiv:1301.4950.

Zhou, J., Bhattacharya, A., Herring, A., and Dunson, D. Bayesian factorizations of big sparse tensors, 2013. arXiv:1306.1598.