ノンパラメトリックテンソルの推定理論と計算理論

Taiji Suzuki
Tokyo Institute of Technology; JST, PRESTO

Abstract

We investigate the statistical efficiency and computational complexity of some nonparametric estimators for a nonlinear tensor estimation problem. Low-rank tensor estimation has been used as a method to learn higher order relations among several data sources in a wide range of applications, such as multi-task learning, recommendation systems, and spatiotemporal analysis. We consider a general setting where a common linear tensor learning is extended to a nonlinear learning problem in reproducing kernel Hilbert space and propose two nonparametric estimators such as a Bayes estimator [16] and an alternating minimization procedure [30]. It is shown that the Bayes estimator achieves a near minimax optimal convergence rate without any strong convexity assumption, such as restricted strong convexity. We also show that the alternating minimization method achieves linear convergence as an optimization algorithm and that the generalization error of the resultant estimator yields the minimax optimality.

1 Problem formulation

Suppose that we are given n input-output samples $\{(x_i, y_i)\}_{i=1}^n$. The input x_i is a concatenation of K variables, i.e., $x_i = (x_i^{(1)}, \cdots, x_i^{(K)}) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_K = \mathcal{X}$, where each $x_i^{(k)}$ is an element of a set $\mathcal{X}_k$. We consider the regression problem where these samples are generated according to the non-parametric model [25]:

$$y_i = \sum_{r=1}^{d^*} \prod_{k=1}^{K} f^*_{(r,k)}(x_i^{(k)}) + \epsilon_i, \tag{1}$$

where $\{\epsilon_i\}_{i=1}^n$ represents an i.i.d. zero-mean noise. In this regression problem, our objective is to estimate the true function $f^*(x^{(1)}, \ldots, x^{(K)}) = \sum_{r=1}^{d^*} \prod_{k=1}^{K} f^*_{(r,k)}(x^{(k)})$.

This model captures the effect of non-linear higher order interactions among the input components $\{x^{(k)}\}_{k=1}^K$ to the output y, and thus, is useful for a regression problem where the output is determined by complex relations between the input components. This type of regression problem appears in several applications, such as multi-task learning, recommendation systems and spatiotemporal data analysis [17, 23, 3].

To understand the model in Eq. (1), it is helpful to consider a linear case as a special case [10, 31]. In general, the linear tensor model is formulated as

$$Y_i = \langle A^*, X_i \rangle + \epsilon_i. \tag{2}$$

Here, X_i, A^* are tensors in $\mathbb{R}^{M_1 \times \cdots \times M_K}$ and the inner product $\langle \cdot, \cdot \rangle$ is defined by $\langle A, X \rangle = \sum_{i_1,\ldots,i_K=1}^{M_1,\ldots,M_K} A_{i_1 \ldots i_K} X_{i_1 \ldots i_K}$. A^* is assumed to be low rank in the sense of CP-rank [13, 14], i.e., A^* is decomposed as $\sum_{r=1}^{d^*} u_r^{*(1)} \circ \cdots \circ u_r^{*(K)}$, where the vector $u_r^{*(k)} \in \mathbb{R}^{M_k}$ and the symbol $\circ$ represents the vector outer product. If we also assume X_i is rank-1, i.e., $X_i = x_i^{(1)} \circ \cdots \circ x_i^{(K)}$, then the inner product in Eq.(2) is written as: $\langle A^*, X_i \rangle = \left\langle \sum_{r=1}^{d} u_r^{*(1)} \circ \cdots \circ u_r^{*(K)}, x_i^{(1)} \circ \cdots \circ x_i^{(K)} \right\rangle = \sum_{r=1}^{d^*} \prod_{k=1}^{K} \left\langle u_r^{*(k)}, x_i^{(k)} \right\rangle$. This is equivalent to the case where we limit $f^*_{(r,k)}$ in Eq. (1) to the linear function $\left\langle u_r^{*(k)}, x^{(k)} \right\rangle$. Hence, the linear model based on CP-decomposition can be understood as a special case of our proposed model.

We propose two estimators for the nonlinear tensor model: a Bayes estimator [16] and an alternating least squares estimator [30].

2 Bayes estimator based on Gaussian process priors

Here we describe the Bayes estimator [16] for the nonparametric tensor estimation problem.

2.1 Gaussian process prior and corresponding reproducing kernel Hilbert space

We place a zero-mean Gaussian process prior $\mathrm{GP}_{r,k}$ with a kernel $\mathrm{k}_{(r,k)}$ to estimate the function $f^*_{(r,k)}$ on $\mathcal{X}_k$. A zero-mean Gaussian process $f = (f(x) : x \in \mathcal{X})$ on some input space $\mathcal{X}$ is a set of random variables $(f(x))_{x\in\mathcal{X}}$ indexed by $\mathcal{X}$ such that each finite subset $(f(x_1), \dots, f(x_j))$ $(j = 1, 2, \dots)$ obeys a zero-mean multivariate normal distribution, where $(x_1, \dots, x_j) \subseteq \mathcal{X}$ is an arbitrary finite subset of $\mathcal{X}$. The kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ corresponding to the Gaussian process is the covariance function defined by $k(x, x') = E[f(x)f(x')]$. Since the kernel function is symmetric and positive definite, we can define its corresponding RKHS in the usual manner [1].

We denote by $\mathcal{H}_{r,k}$ the RKHS corresponding to the kernel $\mathrm{k}_{(r,k)}$. It is known that the RKHS is usually much "smaller" than the support of the Gaussian process in an infinite dimensional setting. In fact, typically the prior has probability mass 0 on the infinite dimensional RKHS [33]. This leads to the fact that, under the assumption $f^*_{(r,k)} \in \mathcal{H}_{r,k}$, estimating the function $f^*_{(r,k)}$ through the standard Bayesian procedure with a Gaussian process prior never achieves the optimal rate in some important examples [33]. To overcome this issue, we scale the process by the factor of $\lambda_{(r,k)}$ and make the estimator close to the small space $\mathcal{H}_{r,k}$.

2.2 The posterior distribution and the corresponding estimator

Given a rank d, let $\mathcal{F} = (f_{(r,k)})_{r=1,\dots,d,\ k=1,\dots,K}$ be a concatenation of functions $\{f_{(r,k)}\}_{r=1,\dots,d,\ k=1,\dots,K}$. Let the Gaussian process prior $\mathrm{GP}_{r,k}(\cdot|\lambda_{(r,k)})$ with a parameter $\lambda_{(r,k)} > 0$ be the process associated with a "scaled" kernel function $\mathrm{k}_{(r,k)}/\lambda_{(r,k)}$. We consider the following prior distribution on the product space $\mathrm{d}\mathcal{F} = (\mathrm{d}f_{(r,k)})_{r=1,\dots,d,\ k=1,\dots,K}$:

$$\Pi(\mathrm{d}\mathcal{F}|d) = \prod_{r=1}^{d}\prod_{k=1}^{K}\int_{\lambda_{(r,k)}>0} \mathrm{GP}_{r,k}(\mathrm{d}f_{(r,k)}|\lambda_{(r,k)})\mathcal{G}(\mathrm{d}\lambda_{(r,k)}),$$

where $\mathcal{G}$ denotes the exponential distribution, $\mathcal{G}(\mathrm{d}\lambda_{(r,k)}) = \exp(-\lambda_{(r,k)})\mathrm{d}\lambda_{(r,k)}$, which is a conjugate prior for the scale of the Gaussian process priors. It will be shown that, by involving the scaling parameter $\lambda_{(r,k)}$, the estimator is able to achieve the optimal convergence rate while it can not without scaling as described above. Putting a prior distribution on $\lambda_{(r,k)}$ rather than fixing it to some optimally chosen value is rather for theoretical purpose, but by doing so, the estimator possesses an adaptivity against a property of f^*. Finally, we place a prior distribution on the rank $1 \le d \le d_{\max}$ as

$$\pi(d) = \frac{\xi^d}{\sum_{d'=1}^{d_{\max}} \xi^{d'}}, \tag{3}$$

where $0 < \xi < 1$ is some positive real number and $d_{\max}$ is a sufficiently larger number than the supposed true rank d.

We now provide the posterior distribution and the corresponding Bayesian estimator. For some $\beta > 0$, the posterior measure is constructed as

$$\Pi(\mathrm{d}\mathcal{F}|D_n) = \frac{\sum_{d=1}^{d_{\max}} \Pi(D_n|\mathcal{F})}{\sum_{d=1}^{d_{\max}} \int \Pi(D_n|\tilde{\mathcal{F}})\Pi(\mathrm{d}\tilde{\mathcal{F}}|d)\pi(d)}\Pi(\mathrm{d}\mathcal{F}|d)\pi(d),$$

where $\Pi(D_n|\mathcal{F})$ is a *quasi likelihood* defined by

$$\Pi(D_n|\mathcal{F}) = \exp\left\{-\frac{1}{\beta}\sum_{i=1}^{n}\left(y_i - \sum_{r=1}^{d}\prod_{k=1}^{K} f_{(r,k)}(x_i^{(k)})\right)^2\right\}$$

with a temperature parameter $\beta > 0$. Although the noise ϵ_i is not necessarily Gaussian, we suggest using the Gaussian likelihood as above. It will be shown that even with this quasi likelihood, we obtain a nice convergence property. Corresponding to the posterior, we have the postierior mean estimator $\widehat{f}$:

$$\widehat{f} = \int f\Pi(\mathrm{d}\mathcal{F}|D_n).$$

Algorithm 1 Alternating minimization procedure for nonlinear tensor estimation

Require: Training data $D_n = \{(x_i, y_i)\}_{i=1}^n$, the regularization parameter $\lambda^{(n)}$, iteration number T.
Ensure: $\widehat{f} = \sum_{r=1}^d \hat{v}_r^{(T)} \prod_{k=1}^K \widehat{f}_{(r,k)}^{(T)}$ as the estimator
  for $t = 1, \dots, T$ **do**
    Set $\tilde{f}_{(r,k)} = \widehat{f}_{(r,k)}^{(t-1)}$ $(\forall (r,k))$, and $\tilde{v}_r = \hat{v}_r^{(t-1)}$ $(\forall r)$.
    for $(r,k) \in \{1, \dots, d\} \times \{1, \dots, K\}$ **do**
      The (r,k)-element of $\tilde{f}$ is updated as

$$\tilde{f}'_{(r,k)} = \operatorname*{argmin}_{f_{(r,k)} \in \mathcal{H}_{r,k}} \left\{ \frac{1}{n} \sum_{i=1}^n \left[y_i - \left(f_{(r,k)} \prod_{k' \neq k} \tilde{f}_{(r,k')} + \sum_{r' \neq r} \tilde{v}_{r'} \prod_{k'=1}^K \tilde{f}_{(r',k')} \right)(x_i) \right]^2 + C_n \|f\|_{\mathcal{H}_{r,k}}^2 \right\}. \tag{5}$$

    $\tilde{v}_r \leftarrow \|\tilde{f}'_{(r,k)}\|_n$, $\tilde{f}_{(r,k)} \leftarrow \tilde{f}'_{(r,k)} / \tilde{v}_r$.
    end for
    Set $\widehat{f}_{(r,k)}^{(t)} = \tilde{f}_{(r,k)}$ $(\forall (r,k))$ and $\hat{v}_r^{(t)} = \tilde{v}_r$ $(\forall r)$.
  end for

The posterior sampling can be easily executed by the Gibbs sampling procedure. See [16] for more details.

3 Alternating regularized least squares procedure

Here, we present the alternating minimization procedure that optimizes the regularized empirical risk in an alternating way [30]. In that procedure, we optimize each component $f_{(r,k)}$ with the other fixed components $\{f_{(r',k')}\}_{(r',k') \neq (r,k)}$. Basically, we want to execute the following optimization problem:

$$\min_{\{f_{(r,k)}\}_{(r,k)} : f_{(r,k)} \in \mathcal{H}_{r,k}} \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{r=1}^d \prod_{k=1}^K f_{(r,k)}(x_i^{(k)}) \right)^2 + C_n \sum_{r=1}^d \sum_{k=1}^d \|f_{(r,k)}\|_{\mathcal{H}_{r,k}}^2, \tag{4}$$

where the first term is the loss function for measuring how our guess $\sum_{r=1}^d \prod_{k=1}^K f_{(r,k)}$ fits the data and the second term is a regularization term for controlling the complexity of the learning function. However, this optimization problem is not convex. Thus, it is difficult to exactly compute the optimal. In fact, it is known that this optimization problem includes an NP-hard problem even for the linear model.

We found that this computational difficulty could be overcome if we assume some additional assumptions and aim to achieve a better *generalization error* instead of exactly minimizing the *training error*. The optimization procedure we discuss to obtain such an estimator is the *alternating minimization procedure*, which minimizes the objective function (4) alternately with respect to each component $f_{(r,k)}$. For each component $f_{(r,k)}$, the objective function (4) is a convex function, and thus, it is easy to obtain the optimal solution. Actually, the subproblem is reduced to a variant of the kernel ridge regression, and the solution can be analytically obtained.

The algorithm we call alternating minimization procedure (AMP) is summarized in Algorithm 1. After minimizing the objective (Eq. (5)), the obtained solution is normalized so that its empirical L_2-norm becomes 1 to adjust the scaling factor freedom. The parameter C_n in Eq. (5) is a regularization parameter that is appropriately chosen.

For theoretical simplicity, we consider the following equivalent constraint formula instead of the penalization one (5):

$$\tilde{f}'_{(r,k)} \in \operatorname*{argmin}_{\substack{f_{(r,k)} \in \mathcal{H}_{r,k} \\ \|f_{(r,k)}\|_{\mathcal{H}_{r,k}} \leq \tilde{R}}} \left\{ \frac{1}{n} \sum_{i=1}^n \left(y_i - f_{(r,k)}(x_i^{(k)}) \prod_{k' \neq k} \tilde{f}_{(r,k')}(x_i^{(k')}) - \sum_{r' \neq r} \tilde{v}_{r'} \prod_{k'=1}^K \tilde{f}_{(r',k')}(x_i^{(k')}) \right)^2 \right\} \tag{6}$$

where the parameter $\tilde{R}$ is a regularization parameter for controlling the complexity of the estimated function.

4 Convergence rate analysis of the Bayes estimator

In this section, we provide the statistical convergence rate of the Bayes estimator, and show that the derived convergence rate is actually minimax optimal (up to constants).

First, we assume a condition on the noise ϵ_i as follows.

Assumption 1 $\mathrm{E}[\epsilon_1^2] < \infty$ and $\mathrm{E}[\epsilon_1] = 0$. Let $m_\epsilon(z) := \int_z^\infty y \mathrm{d}F_\epsilon(y)$ where $F_\epsilon(z) = P(\epsilon_1 \leq z)$ is the cumulative distribution function of the noise ϵ_i. The measure $m_\epsilon(z)\mathrm{d}z$ is absolutely continuous with respect to the distribution function $F_\epsilon(z)$ with a bounded Radon-Nikodym derivative, i.e., there exists a bounded function $g_\epsilon : \mathbb{R} \to \mathbb{R}_+$ such that

$$\int_a^b m_\epsilon(z)\mathrm{d}z = \int_a^b g_\epsilon(z)\mathrm{d}F_\epsilon(z), \quad \forall a, b \in \mathbb{R}.$$

Roughly speaking, this assumption indicates the noise has a light tail probability. In fact, the Gaussian noise $N(0,1)$ satisfies this assumption with $g_\epsilon(z) = \sigma^2$. See [11] for more details.

Next, we introduce a quantity that measures the complexity of the RKHSs. More specifically, we assume that the RKHSs defined by the kernels have a polynomial order complexity of the *metric entropy* of their unit balls. Let $N(B, \epsilon, d)$ denote the ϵ-covering number of the space B with respect to the metric d [34], that is, the smallest number of ϵ-balls that are required to cover B, where the radius ϵ of the ϵ-balls is measured by the metric d. The metric entropy is the logarithm of the covering number. Let $\mathcal{B}_{\mathcal{H}_{(r,k)}}$ be the unit ball of the RKHS $\mathcal{H}_{(r,k)}$.

Assumption 2 There exists a real value $0 < s_{(r,k)} < 1$ and $C_0 > 0$ such that

$$\log N(\mathcal{B}_{\mathcal{H}_{(r,k)}}, \epsilon, \|\cdot\|_n) \leq C_0 \epsilon^{-2s_{(r,k)}} \quad (\epsilon > 0). \tag{7}$$

Moreover, the kernel function is bounded as $\sup_x k_{r,k}(x,x) \leq 1$.

An interesting fact is that the metric entropy condition in Eq. (7) controls the *small ball probability* of the corresponding Gaussian process as $-\log(\mathrm{GP}_{r,k}(\{f : \|f\|_n \leq \epsilon\})) = O\left(\epsilon^{-2s_{(r,k)}/(1-s_{(r,k)})}\right)$ [18, 20]. This assumption is usually satisfied by practically used kernels. For example, the Gaussian kernel satisfies this condition with an arbitrary $s_{(r,k)}$ with a different constant C_0 with high probability.

Next, we assume that the prior has a sufficient mass on bounded functions. This is a technical assumption and practically used kernels usually satisfy this assumption.

Assumption 3 There exists $c_1 > 0$ such that

$$-\log(\mathrm{GP}_{r,k}(\{f : \|f\|_\infty \leq 1\})) \leq c_1 \quad (\forall r, k).$$

Moreover, we assume the following condition on the true function f^*.

Assumption 4 $f^*_{(r,k)}$ is included in $\mathcal{H}_{r,k}$ for all $1 \leq r \leq d_{\max}$ and $1 \leq k \leq K$. There exists R such that $\max_{(r,k)} \|f^*_{(r,k)}\|_{\mathcal{H}_{(r,k)}} \leq R$. The true tensor is low rank, that is, there exists d such that $f^*_{(r,k)} = 0$ for all $r > d$ and $1 \leq k \leq K$.

Under these assumptions, we have the following estimation error bound.

Theorem 1 Suppose that Assumptions 1, 2, 3, and 4 are satisfied, and $\beta \geq 4\|g_\epsilon\|_\infty$. Then, there exists a constant $C > 0$ depending on β, C_0, c_1 and $s_{(r,k)}$ such that

$$\mathrm{E}_{Y_{1:n}|x_{1:n}}\left[\|\widehat{f} - f^*\|_n^2\right] \leq C\left\{(3R \vee 1)^{2(K-1)} \sum_{r=1}^{d}\sum_{k=1}^{K} n^{-\frac{1}{1+s_{(r,k)}}} + \frac{d}{n}\log\left(\frac{1}{\kappa}\right)\right\},$$

where $\mathrm{E}_{Y_{1:n}|x_{1:n}}$ indicates the expectation with respect to the outputs $Y_1, \ldots, Y_n$ conditioned by the inputs $x_1, \ldots, x_n$, and $\kappa = \xi(1-\xi)$.

Basically, the proof is obtained by using the PAC-Bayes bound [21, 22, 9] (the version we used was developed by [11]), and applying the small ball probability theorems of Gaussian processes [18, 20].

If $K = 1$, the convergence rate coincides with the usual one of the ordinary kernel ridge regression [26]. Note that we do not assume any (restricted) strong convexity on the design. Remarkably, the convergence

rate is determined by the true rank d (not $d_{\max}$). This implies that the posterior of the rank based on the prior in Eq. (3) properly concentrates around the true rank. The second term $\frac{d}{n}\log\left(\frac{1}{\kappa}\right)$ represents the complexity of the model selection. This term is almost negligible as $n \to \infty$. Moreover, if we know d beforehand and fix $d = d_{\max}$, then this term disappears. It will be shown that this convergence rate is actually minimax optimal (see Theorem 5 below).

We analyze the convergence rate more closely. To do so, let us consider a special case of the *Matérn prior*, and assume the domain of the input is a hypercube: $x^{(k)} \in [0,1]^{p_k}$. The Matérn prior is a Gaussian process prior corresponding to a kernel function that has a spectral density given as $\psi(s) = \frac{1}{(1+\|s\|^2)^{\alpha+p/2}}$, where α is a smoothness parameter and p is the dimension of the input. It is known that the corresponding RKHS is included in a Sobolev space $W^{\alpha+p/2}[0,1]^p$ with the smoothness $\alpha + p/2$ [33], and thus, the metric entropy exponent can be evaluated as $s \leq p/(2\alpha + p)$ (with high probability). We consider a simple situation where the Gaussian process prior on $f_r^{(k)}$ is the Matérn prior with the same smoothness parameter α for all r, k. Then, according to Theorem 1, we obtain the following convergence rate in this situation:

$$\mathrm{E}_{Y_{1:n}|x_{1:n}}\left[\|\widehat{f} - f^*\|_n^2\right] \leq C\left\{\sum_{r=1}^{d}\sum_{k=1}^{K} n^{-\frac{1}{1+\frac{p_k}{2\alpha+p_k}}}\right\}.$$

This could be much smaller than the optimal convergence rate $O(n^{-\frac{1}{1+p/(2\alpha+p)}})$ for the nive estimation of $f^* \in W^{\alpha+p/2}[0,1]^p$ on the whole space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_K$ because the full dimension $p = \sum_{k=1}^K p_k$ is larger than individual dimension p_k. However, an estimation fully utilizing the nonlinear tensor product model in Eq. (1) can alleviate the curse of dimensionality.

5 Convergence analysis of the alternating minimization procedure

Here, we give a statistical and algorithmic convergence analysis for the alternating minimization procedure (AMP, Algorithm 1).

5.1 Assumptions and problem settings for the convergence analysis

We prepare some assumptions for the theoretical analysis of AMP. First, we assume that the distribution $P(X)$ of the input feature $x \in \mathcal{X}$ is a product measure of $P_k(X)$ on each $\mathcal{X}_k$. That is, $P_{\mathcal{X}}(\mathrm{d}X) = P_1(\mathrm{d}X_1) \times \cdots \times P_K(\mathrm{d}X_K)$ for $X = (X_1, \ldots, X_K) \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_K$. This is typically assumed in the analysis of linear tensor estimation methods [15, 8, 4, 24, 2, 36, 28, 37]. Thus, the L_2-norm of a "rank-1" function $f(x) = \prod_{k=1}^K f_k(x^{(k)})$ can be decomposed into

$$\|f\|_{L_2(P_{\mathcal{X}})}^2 = \|f_1\|_{L_2(P_1)}^2 \times \cdots \times \|f_K\|_{L_2(P_K)}^2.$$

Hereafter, with a slight abuse of notations, we denote by $\|f\|_{L_2} = \|f\|_{L_2(P_k)}$ for a function $f : \mathcal{X}_k \to \mathbb{R}$. The inner product in the space L_2 is denoted by $\langle f, g\rangle_{L_2} := \int f(X)g(X)\mathrm{d}P_{\mathcal{X}}(X)$. Note that because of the construction of $P_{\mathcal{X}}$, it holds that $\langle f, g\rangle_{L_2} = \prod_{k=1}^K \langle f_k, g_k\rangle_{L_2}$ for functions $f(x) = \prod_{k=1}^K f_k(x^{(k)})$ and $g(x) = \prod_{k=1}^K g_k(x^{(k)})$ where $x = (x^{(1)}, \ldots, x^{(K)}) \in \mathcal{X}$.

Next, we assume that the norm of the true function is bounded away from zero and from above. Let the magnitude of the rth component of the true function be $v_r := \|\prod_{k=1}^K f_{(r,k)}^*\|_{L_2}$ and the normalized components be $f_{(r,k)}^{**} := f_{(r,k)}^*/\|f_{(r,k)}^*\|_{L_2}$ $(\forall(r,k))$.

Assumption 5 (Boundedness Assumption)

- (A5-1) There exist $0 < v_{\min} \leq v_{\max}$ such that $v_{\min} \leq v_r \leq v_{\max}$ $(\forall r = 1, \ldots, d)$.
- (A5-2) The true function $f_{(r,k)}^*$ is included in the RKHS $\mathcal{H}_{r,k}$, i.e., $f_{(r,k)}^* \in \mathcal{H}_{r,k}$ $(\forall(r,k))$, and there exists $R > 0$ such that $\max\{v_r, 1\}\|f_{(r,k)}^{**}\|_{\mathcal{H}_{r,k}} \leq R$ $(\forall(r,k))$.
- (A5-3) The kernel function $\mathrm{k}_{(r,k)}$ associated with the RKHS $\mathcal{H}_{r,k}$ is bounded as $\sup_{x\in\mathcal{X}_k} \mathrm{k}_{(r,k)}(x,x) \leq 1$ $(\forall(r,k))$.
- (A5-4) There exists $L > 0$ such that the noise is bounded as $|\epsilon_i| \leq L$ (a.s.).

Assumption 5 is a standard one for the analysis of the tensor model and the kernel regression model. Note that the boundedness condition of the kernel gives that $\|f\|_\infty = \sup_{x^{(k)}} |f(x^{(k)})| \leq \|f\|_{\mathcal{H}_{r,k}}$ for all $f \in \mathcal{H}_{r,k}$ because the Cauchy-Schwarz inequality gives $|\langle f, \mathrm{k}_{(r,k)}(\cdot, x^{(k)})\rangle_{\mathcal{H}_{r,k}}| \leq \mathrm{k}_{(r,k)}(x^{(k)}, x^{(k)})\|f\|_{\mathcal{H}_{r,k}}$ for all $x^{(k)}$. Thus, combining with (A5-2), we also have $\|f^{**}_{(r,k)}\|_\infty \leq R$. The last assumption (A5-4) is a bit restrictive. However, this assumption can be replaced with a Gaussian assumption. In that situation, we may use the Gaussian concentration inequality [19] instead of Talagrand's concentration inequality in the proof.

Next, we characterize the *complexity* of each RKHS $\mathcal{H}_{r,k}$ by using the *entropy number* [34, 26]. This is important because it directly determines the convergence rate. The *ϵ-covering number* $\mathcal{N}(\epsilon, \mathcal{G}, L_2(P_\mathcal{X}))$ with respect to $L_2(P_\mathcal{X})$ is the minimal number of balls with radius ϵ measured by $L_2(P_\mathcal{X})$ needed to cover a set $\mathcal{G} \subset L_2(P_\mathcal{X})$. The ith entropy number $e_i(\mathcal{G}, L_2(P_\mathcal{X}))$ is defined as the infimum of $\epsilon > 0$ such that $\mathcal{N}(\epsilon, \mathcal{G}, L_2) \leq 2^{i-1}$ [26]. Intuitively, if the entropy number is small, the space $\mathcal{G}$ is "simple"; otherwise, it is "complicated."

Assumption 6 (Complexity Assumption) Let $\mathcal{B}_{\mathcal{H}_{r,k}}$ be the unit ball of an RKHS $\mathcal{H}_{r,k}$. There exist $0 < s < 1$ and c such that

$$e_i(\mathcal{B}_{\mathcal{H}_{r,k}}, L_2(P_\mathcal{X})) \leq c i^{-\frac{1}{2s}}, \tag{8}$$

for all $1 \leq r \leq d$ and $1 \leq k \leq K$.

The optimal rate of the ordinary kernel ridge regression on the RKHS with Assumption 6 is given as $n^{-\frac{1}{1+s}}$ [27]. It is known that Assumption 6 is equivalent to the polynomial decay assumption on the spectrum of the integral operator associated with the kernel $\mathrm{k}_{(r,k)}$ (see [26] and Theorem 15 of [27] for more details).

Next, we give a technical assumption about the L_∞-norm.

Assumption 7 (Infinity Norm Assumption) There exist $0 < s_2 \leq 1$ and c_2 such that

$$\|f\|_\infty \leq c_2 \|f\|_{L_2}^{1-s_2} \|f\|_{\mathcal{H}_{r,k}}^{s_2} \quad (\forall f \in \mathcal{H}_{r,k}) \tag{9}$$

for all $1 \leq r \leq d$ and $1 \leq k \leq K$.

By Assumption 5, this assumption is always satisfied for $c_2 = 1$ and $s_2 = 1$. $s_2 < 1$ is a nontrivial situation and gives a tighter bound. We would like to note that this condition with $s_2 < 1$ is satisfied by many practically used kernels such as the Gaussian kernel. In particular, it is satisfied if the kernel is smooth so that $\mathcal{H}_{r,k}$ is included in a Sobolev space $W^{2,s_2}[0,1]$. More formal characterization of this condition using the notion of a *real interpolation space* can be found in [27] and Proposition 2.10 of [7].

Finally, we assume an *incoherence* condition on $\{f^*_{(r,k)}\}_{r,k}$. Roughly speaking, the incoherence property of a set of functions $\{f_{(r,k)}\}_{r,k}$ means that components $\{f_{(r,k)}\}_r$ are linearly independent across different $1 \leq r \leq d$ on the same mode k. This is required to distinguish each component. An analogous assumption has been assumed also in the literature of linear models [15, 8, 4, 24, 36, 28].

Definition 2 (Incoherence) A set of functions $\{f_{(r,k)}\}_{r,k}$, where $f_{(r,k)} \in L_2(P_k)$, is μ-incoherent if, for all $k = 1, \ldots, K$, it holds that

$$|\langle f_{(r,k)}, f_{(r',k)}\rangle_{L_2}| \leq \mu \|f_{(r,k)}\|_{L_2} \|f_{(r',k)}\|_{L_2} \quad (\forall r \neq r').$$

Assumption 8 (Incoherence Assumption) There exists $1 > \mu^* \geq 0$ such that the true function $\{f^*_{(r,k)}\}_{r,k}$ is μ^*-incoherent.

5.2 Linear convergence of alternating minimization procedure

In this section, we give the convergence analysis of the AMP algorithm. Under the assumptions presented in the previous section, it will be shown that the AMP algorithm shows linear convergence in the sense of optimization algorithm and achieves the minimax optimal rate in the sense of statistical performance. Roughly speaking, if the initial solution is sufficiently close to the true function (namely, in a distance of $O(1)$), then the solution generated by AMP linearly converges to the optimal solution and the estimation accuracy of the final solution is given as $O(dKn^{-\frac{1}{1+s}})$ up to $\log(dK)$ factor.

We analyze how close the updated estimator is to the true one when the (r,k)th component is updated from $\tilde{f}_{(r,k)}$ to $\tilde{f}'_{(r,k)}$. The tensor decomposition $\{f_{(r,k)}\}_{r,k}$ of a nonlinear tensor model has a freedom of scaling. Thus, we need to measure the accuracy based on a normalized representation to avoid the scaling factor uncertainty. Let the normalized components of the estimator be $\bar{f}_{(r',k')} = \tilde{f}_{(r',k')}/\|\tilde{f}_{(r',k')}\|_{L_2}$ $(\forall (r',k') \in [d] \times [K])$ and $\bar{v}_{r'} = \tilde{v}_{r'} \prod_{k'=1}^{K} \|\tilde{f}_{(r',k')}\|_{L_2}$ $(\forall r' \in [d])$. On the other hand, the newly updated (r,k)th element is denoted by $\tilde{f}'_{(r,k)}$ (see Eq. (5)) and we denote by $\bar{v}'_r$ the updated value of $\bar{v}_r$ correspondingly: $\bar{v}'_r = \|\tilde{f}'_{(r,k)}\|_{L_2} \prod_{k' \neq k} \|\tilde{f}_{(r,k')}\|_{L_2}$. The normalized newly updated element is denoted by $\bar{f}'_{(r,k)} = \tilde{f}'_{(r,k)}/\|\tilde{f}'_{(r,k)}\|_{L_2}$.

For an estimator $(\bar{f}, \bar{v}) = (\{\bar{f}_{(r',k')}\}_{r',k'}, \{\bar{v}_{r'}\}_{r'})$ which is a couple of the normalized component and the scaling factor, define

$$\mathrm{d}_\infty(\bar{f}, \bar{v}) := \max_{(r',k')} \{v_{r'} \|\bar{f}_{(r',k')} - f^{**}_{(r',k')}\|_{L_2} + |v_{r'} - \bar{v}_{r'}|\}.$$

For any $\lambda_{1,n} > 0$ and $\lambda_{2,n} > 0$ and $\tau > 0$, we let $a_\tau := \max\{1, L\} \max\{1, \tau\} \log(dK)$ and define $\xi_n = \xi_n(\lambda_{1,n}, \tau)$ and $\xi'_n = \xi'_n(\lambda_{2,n}, \tau)$ as *

$$\xi_n := a_\tau \left(\frac{K^{\frac{1+2s}{2}} \lambda_{1,n}^{-\frac{s}{2}}}{\sqrt{n}} \vee \frac{K^{\frac{1+2s}{1+s}}}{\lambda_{1,n}^{\frac{2s+(1-s)s_2}{2(1+s)}} n^{\frac{1}{1+s}}} \right), \quad \xi'_n := a_\tau \left(\frac{\lambda_{2,n}^{-\frac{s}{2}}}{\sqrt{n}} \vee \frac{1}{\lambda_{2,n}^{\frac{1}{2}} n^{\frac{1}{1+s}}} \right).$$

Theorem 3 Suppose that Assumptions 5–8 are satisfied, and the regularization parameter $\tilde{R}$ in Eq. (6) is set as $\tilde{R} = 2R$. Let $\hat{R} = 4\tilde{R}/\min\{v_{\min}, 1\}$ and suppose that we have already obtained an estimator $\tilde{f}$ satisfying the following conditions:

- The RKHS-norms of $\{\bar{f}_{(r',k')}\}_{r',k'}$ are bounded as $\|\bar{f}_{(r',k')}\|_{\mathcal{H}_{r',k'}} \leq \hat{R}$ $(\forall (r',k') \neq (r,k))$.
- The distance from the true one is bounded as $\mathrm{d}_\infty(\bar{f}, \bar{v}) \leq \gamma$.

Then, for a sufficiently small μ^* and γ (independent of n), there exists an event with probability greater than $1 - 3\exp(-\tau)$ where any $(\bar{f}, \bar{v})$ satisfying the above conditions gives

$$\left(v_r \|\bar{f}'_{(r,k)} - f^{**}_{(r,k)}\|_{L_2} + |\bar{v}'_r - v_r|\right)^2 \leq \frac{1}{2} \mathrm{d}_\infty(\bar{f}, \bar{v})^2 + S_n \hat{R}^{2K} \tag{10}$$

for any sufficiently large n, where S_n is defined for a constant C' depending on s, s_2, c, c_2 as

$$S_n := C' \left[(\hat{R}^K + 1)(\xi'_n \lambda_{2,n}^{1/2} + \xi'^2_n) + (\hat{R}^K + d)\xi_n \lambda_{1,n}^{1/2} + \hat{R}^{2(K-1)(\frac{1}{s_2} - 1)} (d\xi_n)^{2/s_2} \right].$$

Moreover, if we denote by η_n the right hand side of Eq. (10), then it holds that

$$\|\bar{f}'_{(r,k)}\|_{\mathcal{H}_{r,k}} \leq \frac{2}{v_r - \sqrt{\eta_n}} \tilde{R}.$$

The proof is given by using such techniques as the so-called peeling device [32] or, equivalently, the local Rademacher complexity [5], and by combining these techniques with the coordinate descent optimization argument. Theorem 3 states that, if the initial solution is sufficiently close to the true one, then the following updated estimator gets closer to the true one and its RKHS-norm is still bounded above by a constant. Importantly, it can be shown that the updated one still satisfies the conditions of Theorem 3 for large n. Since the bound given in Theorem 3 is uniform, the inequality (10) can be recursively applied to the sequence of $\widehat{f}^{(t)}$ $(t = 1, 2, \dots)$.

By substituting $\lambda_{1,n} = K^{-\frac{1+s}{1-s}} d^{-\frac{2}{1-s}} n^{-\frac{1}{1+s}}$ and $\lambda_{2,n} = n^{-\frac{1}{1+s}}$, we have that

$$S_n = O\left(n^{-\frac{1}{1+s}} \vee \left(n^{-\frac{1}{1+s} - (1-s_2)\min\{\frac{1-s}{4(1+s)}, \frac{1}{s_2(1+s)}\}} \mathrm{poly}(d, K) \right) \right) \log(dK),$$

where $\mathrm{poly}(d, K)$ means a polynomial of d, K. Thus, if $s_2 < 1$ and n is sufficiently large compared with d and K, then the second term is smaller than the first term and we have $S_n \leq C n^{-\frac{1}{1+s}}$ with a constant C. Furthermore, we can bound the L_2-norm from the true one as in the following theorem.

* The symbol $\vee$ indicates the max operation, that is, $a \vee b := \max\{a, b\}$.

Theorem 4 Let $(\widehat{f}^{(t)}, \hat{v}^{(t)})$ be the estimator at the tth iteration. In addition to the assumptions of Theorem 3, suppose that $(\widehat{f}^{(1)}, \hat{v}^{(1)})$ satisfies $\mathrm{d}_\infty(\widehat{f}^{(1)}, \hat{v}^{(1)})^2 \leq \frac{v_{\min}^2}{8}$ and $S_n \hat{R}^{2K} \leq \frac{v_{\min}^2}{8}$, $s_2 < 1$ and $n \gg d, K$, then $\check{f}^{(t)}(x) = \sum_{r=1}^d \hat{v}_r^{(t)} \prod_{k=1}^K \widehat{f}_{(r,k)}^{(t)}(x^{(k)})$ satisfies

$$\|\check{f}^{(t)} - f^*\|_{L_2}^2 = O\left(dKn^{-\frac{1}{1+s}}\log(dK) + dK\left(\frac{3}{4}\right)^t\right).$$

with probability $1 - 3\exp(-\tau)$.

This theorem indicates that after $T = O(\log(n))$ iterations, we obtain the estimation accuracy of $O(dKn^{-\frac{1}{1+s}}\log(dK))$. The estimation accuracy bound $O(dKn^{-\frac{1}{1+s}}\log(dK))$ is intuitively natural because we are estimating $d \times K$ functions $\{f_{(r,k)}^*\}_{r,k}$ and the optimal sample complexity to estimate one function $f_{(r,k)}^*$ is known as $n^{-\frac{1}{1+s}}$ [27]. Indeed, this accuracy bound is minimax optimal (see Section 6). A rough Bayes estimator would be a good initial solution satisfying the assumptions.

6 Minimax lower bound

Here, we give the minimax lower bound. To simplify the problem, we specify the structure of the problem. We assume that each component $x^{(k)} \in \mathcal{X}_k$ of the input $x = (x^{(1)}, \ldots, x^{(K)}) \in \mathcal{X}$ can be further decomposed as

$$x^{(k)} = (x_{(1,k)}, \ldots, x_{(d,k)}) \in \mathcal{X}_{(1,k)} \times \cdots \times \mathcal{X}_{(d,k)} = \mathcal{X}_k.$$

Then, each RKHS $\mathcal{H}_{r,k}$ takes $x_{(r,k)} \in \mathcal{X}_{(r,k)}$ as an input; that is, for any $f_{(r,k)} \in \mathcal{H}_{r,k}$, there is a function $\tilde{f}_{(r,k)} : \mathcal{X}_{(r,k)} \to \mathbb{R}$ such that $f_{(r,k)}(x_k) = \tilde{f}_{(r,k)}(x_{(r,k)})$. We assume that the distribution of the input $x_k \in \mathcal{X}_k$ is a product measure $P_{\mathcal{X}_k} = P_{\mathcal{X}_{(1,k)}} \times \cdots \times P_{\mathcal{X}_{(d,k)}}$ and the distribution of the whole input $x = (x^{(1)}, \ldots, x^{(K)}) \in \mathcal{X}$ is also a product of $P_{\mathcal{X}_k}$: $P_{\mathcal{X}} = P_{\mathcal{X}_1} \times \cdots \times P_{\mathcal{X}_K}$. We may assume that all functions $f_{(r,k)} \in \mathcal{H}_{r,k}$ have zero mean without loss of generality: $\mathrm{E}_{X \sim P_{\mathcal{X}_k}}[f_{(r,k)}(X)] = 0$. Then, by the set up of $P_{\mathcal{X}}$, we have that

$$\|f\|_{L_2(P_{\mathcal{X}})}^2 = \mathrm{E}_{X \sim P(X)}[f^2(X)] = \sum_{r=1}^d \prod_{k=1}^K \|f_{(r,k)}\|_{L_2(P_{\mathcal{X}_k})}^2$$

for $f = \sum_{r=1}^d \prod_{k=1}^K f_{(r,k)}$ where $f_{(r,k)} \in \mathcal{H}_{r,k}$. Moreover, we assume that the noise is distributed from a normal distribution: $\epsilon_i \sim N(0, \sigma^2)$ (i.i.d.).

To simplify the analysis, we assume that the complexities of all RKHSs $\mathcal{H}_{r,k}$ are the same and have the following lower bound of the metric entropy.

Assumption 9 There exists a real value $0 < s < 1$ such that

$$\log N(\mathcal{B}_{\mathcal{H}_{(r,k)}}, \epsilon, L_2(P_{\mathcal{X}_{(r,k)}})) \sim \epsilon^{-2s}. \tag{11}$$

Moreover, the kernel function is bounded as $\sup_x k_{r,k}(x,x) \leq 1$, and there exists $c_1 > 0$ such that $\exists \widehat{f}_{(r,k)} \in \mathcal{B}_{\mathcal{H}_{r,k}}$ satisfying $\|\widehat{f}_{(r,k)}\|_{L_2(P_{\mathcal{X}_k})} \geq c_1$ for all r, k.

Let $\mathcal{H}_{r,k}(R) := \{f \in \mathcal{H}_{r,k} \mid \|f\|_{\mathcal{H}_{r,k}} \leq R\}$ be the ball with radius R in $\mathcal{H}_{r,k}$. Then, we define a set of tensors as

$$\mathscr{H}_{(d,K)}(R) := \left\{ f = \sum_{r=1}^d \prod_{k=1}^K f_{(r,k)} \;\middle|\; f_{(r,k)} \in \mathcal{H}_{r,k}(R) \right\}.$$

Under these settings, we have the following minimax optimal lower bound of the estimation error.

Theorem 5 If every $\mathcal{X}_k$ is a compact metric space, every $\mathrm{k}_{(r,k)}$ is continuous, and the radius R of the tensor set $\mathscr{H}_{(d,K)}(R)$ satisfies $R \geq \frac{1+c_1}{c_1}$, then there is a constant $C_1 > 0$ independent of d, K, n such that

$$\inf_{\hat{f}} \sup_{f^* \in \mathscr{H}_{(d,K)}(R)} \mathrm{E}[\|f - \hat{f}\|_{L_2(P_{\mathcal{X}})}^2] \geq C_1 dKn^{-\frac{1}{1+s}},$$

where the inf is taken over all estimators $\hat{f}$.

In the proof, we utilize the information theoretic technique developed by [35]. This theorem states that the learning rates of the Gaussian process tensor estimator (Theorem 1) and the AMP (Theorem 4) is actually minimax-optimal up to constants.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[2] A. Aswani. Low-rank approximation and completion of positive tensors. *arXiv preprint arXiv:1412.0620*, 2014.

[3] M. T. Bahadori, Q. R. Yu, and Y. Liu. Fast multivariate spatio-temporal analysis via low rank tensor learning. In *Advances in Neural Information Processing Systems 27*.

[4] B. Barak and A. Moitra. Tensor prediction, rademacher complexity and random 3-xor. *arXiv preprint arXiv:1501.06521*, 2015.

[5] P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33:1487–1537, 2005.

[6] C. Bennett and R. Sharpley. *Interpolation of Operators*. Academic Press, Boston, 1988.

[7] C. Bennett and R. Sharpley. *Interpolation of Operators*. Academic Press, Boston, 1988.

[8] S. Bhojanapalli and S. Sanghavi. A new sampling technique for tensors. *arXiv preprint arXiv:1502.05023*, 2015.

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

[10] W. Chu and Z. Ghahramani. 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.

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

[12] D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge University Press, Cambridge, 1996.

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

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

[15] P. Jain and S. Oh. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems 27*, pages 1431–1439. Curran Associates, Inc., 2014.

[16] H. Kanagawa, T. Suzuki, H. Kobayashi, N. Shimizu, and Y. Tagami. Gaussian process nonparametric tensor estimator and its minimax optimality. In *International Conference on Machine Learning (ICML2016)*, pages 1632–1641, 2016.

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

[18] J. Kuelbs and W. V. Li. Metric entropy and the small ball problem for Gaussian measures. *Journal of Functional Analysis*, 116(1):133–157, 1993.

[19] M. Ledoux. *The concentration of measure phenomenon*. Number 89 in Mathematical Surveys and Monographs. American Mathematical Soc., 2005.

[20] W. V. Li and Q.-M. Shao. Gaussian processes: inequalities, small ball probabilities and applications. *Stochastic Processes: Theory and Methods*, 19:533–597, 2001.

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

[22] D. McAllester. PAC-Bayesian model averaging. In *the Anual Conference on Computational Learning Theory*, pages 164–170, 1999.

[23] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil. Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML2013)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1444–1452, 2013.

[24] P. Shah, N. Rao, and G. Tang. Optimal low-rank tensor recovery from separable measurements: Four contractions suffice. *arXiv preprint arXiv:1505.04085*, 2015.

[25] M. Signoretto, L. D. Lathauwer, and J. A. K. Suykens. Learning tensors in reproducing kernel Hilbert spaces with multilinear spectral penalties. *CoRR*, abs/1310.4977, 2013.

[26] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

[27] I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the Annual Conference on Learning Theory*, pages 79–93, 2009.

[28] W. Sun, Z. Wang, H. Liu, and G. Cheng. Non-convex statistical optimization for sparse tensor graphical model. In *Advances in Neural Information Processing Systems*, pages 1081–1089, 2015.

[29] T. Suzuki. Convergence rate of Bayesian tensor estimator and its minimax optimality. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1273–1282, 2015.

[30] T. Suzuki, H. Kanagawa, H. Kobayashi, N. Shimizu, and Y. Tagami. Minimax optimal alternating minimization for kernel nonparametric tensor learning. In *Annual Conference on Neural Information Processing Systems (NIPS2016)*, page to appear, 2016.

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

[32] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

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

[34] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

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

[36] Z. Zhang and S. Aeron. Exact tensor completion using t-svd. *arXiv preprint arXiv:1502.04689*, 2015.

[37] T. Zhao, Z. Wang, and H. Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems 28*, pages 559–567. Curran Associates, Inc., 2015. NIPS2015.