Improved Bounds for Multi-task Learning with Trace Norm Regularization

Weiwei Liu LIUWEIWEI863@GMAIL.COM

School of Computer Science, National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan, China.

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Compared with learning each task independently, multi-task learning (MTL) is able to learn with few training samples and achieves better prediction performance. Recently, Boursier et al. (2022) study the estimation error bound for MTL with trace norm regularizer and a few observations per task. However, their results rely on three assumptions: 1) The features are isotropic; 2) The task diversity assumption is enforced to the parameters matrix; 3) The number of tasks is larger than the features dimension. Whether it is possible to drop these three assumptions and improve the bounds in Boursier et al. (2022) has remained unknown. This paper provides an affirmative answer to this question. Specifically, we reduce their upper bounds from $\tilde{\mathcal{O}}(\sigma\sqrt{\frac{rd^2/m+rT}{m}}+\sqrt{\frac{rd^2/m+rdT/m}{m}})$ to $\mathcal{O}(\sigma\sqrt{\frac{r+rd/T}{m}})$ without three assumptions, where T is the number of tasks, d is the dimension of the feature space, m is the number of observations per task, r is the rank of ground truth matrix, σ is the standard deviation of the noise random variable. Moreover, we provide minimax lower bounds showing our upper bounds are rate optimal if $T = \mathcal{O}(d)$.

Keywords: Multi-task learning; Trace norm regularization; High dimensional statistics

1. Introduction

Multi-task Learning (MTL) aims to simultaneously learn multiple related tasks and achieve better performance than learning each task independently (Caruana, 1993; Ben-David and Schuller, 2003; Bakker and Heskes, 2003; Ando and Zhang, 2005; Liu et al., 2017, 2019; Mao et al., 2020a). It has achieved great success in various applications ranging from computer vision (Liu and Tsang, 2015, 2017; Kendall et al., 2018) to natural language processing (Xiao et al., 2018; Mao et al., 2020b, 2021).

The existing theoretical results (Rohde and Tsybakov, 2011) on MTL assume that the number of observations m per task is larger than the features dimension d ($m > d$). Recently, Boursier et al. (2022) study the following problem of MTL, which plays a vital role in meta-learning and few-shot learning (Vinyals et al., 2016; Finn et al., 2017) that aggregates knowledge among multiple tasks to learn a shared representation:

> *How can we learn across multiple tasks with a very limited number of observations for each of them?*

Tripuraneni et al. (2021) present a theoretical result on this problem with the requirement of spherically symmetric feature distribution. Without the assumption on the specific feature distributions, Boursier et al. (2022) bound the estimation error of the trace norm regularized estimator with a few observations per task ($m < d$). There are three main assumptions that behind the work of (Boursier et al., 2022):

1. The features are isotropic.
2. The task diversity assumption is enforced to the parameters matrix.
3. The number of tasks T is larger than the features dimension d ($T > d$).

It has remained unclear, however, whether it is possible to drop these three assumptions and improve the bounds in Boursier et al. (2022). This paper makes progress on this question.

Contributions. This paper studies the theory of MTL with trace norm regularizer, and improves the estimation error bounds of (Boursier et al., 2022) for the trace norm regularized estimator by removing the assumptions of isotropic features, task diversity and $T > d$. Moreover, we establish matching lower bounds on the minimax error, showing that our upper bounds cannot be improved beyond constant factors under some mild conditions. The analysis of our upper bounds rely on the restricted strong convexity (RSC) of the cost function and the dual norm bound (Wainwright, 2019). The main technical contribution of this paper is to prove that with high probability, a form of the RSC condition and dual norm bound hold for MTL. Our lower bounds are based on the Fano's inequality.

Theorem 1 (Informal upper bounds) *For any number of observations per task m, the trace norm regularized estimator* $\hat{\mathbf{M}}$ *satisfies with high probability*

$$||\hat{\mathbf{M}} - \mathbf{M}^*||_F \leq \mathcal{O}(\sigma\sqrt{\frac{r(T+d)}{mT}}) \tag{1}$$

where $\mathbf{M}^* \in \mathbb{R}^{d\times T}$ *is the ground truth matrix,* r *is its rank,* σ *is the standard deviation of the noise random variable.*

Note that we have derived our upper bounds by removing the assumptions behind the work of (Boursier et al., 2022). Boursier et al. (2022) prove that $||\hat{\mathbf{M}} - \mathbf{M}^*||_F \leq \tilde{\mathcal{O}}(\sigma\sqrt{\frac{rd^2/m+rT}{m}} + \sqrt{\frac{rd^2+rdT}{m^2}})$ holds with high probability under the conditions of $T > d$ and $d > m$, where $\tilde{\mathcal{O}}$ hides logarithmic terms in d, m and T. We can see that our upper bound $\mathcal{O}(\sigma\sqrt{\frac{r+rd/T}{m}})$ is significantly tighter than $\tilde{\mathcal{O}}(\sigma\sqrt{\frac{rd^2/m+rT}{m}} + \sqrt{\frac{rd^2+rdT}{m^2}})$, and thus we improve this estimation error bound by a factor $\mathcal{O}(\sqrt{\frac{rd^2+rdT}{m^2}})$ and a logarithmic term.

Theorem 2 (Informal lower bounds) *Let* ϕ^2 *be the maximum diagonal entry of the covariance matrix* $\mathbf{\Sigma}$*. The following bound holds.*

$$\inf_{\hat{\mathbf{M}}} \sup_{\mathbf{M}^* \in \{\mathbf{M}^* \in \mathbb{R}^{d\times T} | rank(\mathbf{M}^*) = r\}} E||\hat{\mathbf{M}} - \mathbf{M}^*||_F^2 \geq \mathcal{O}\Big(\frac{\sigma^2 r(T+d)}{m\phi^2(d-1)\ln(dT)}\Big) \tag{2}$$

Theorem 2 shows that the upper bounds obtained in Theorem 1 are minimax-optimal up to constant factors if $T = \mathcal{O}(d)$.

2. Main Results

Let $[n] := \{1, \dots, n\}$. $\langle \cdot, \cdot \rangle$ means the inner product both for vectors and matrices. We denote vectors as lowercase bold letters and matrices as uppercase bold letters, respectively. We denote the transpose of the vector/matrix by the superscript $'$, and the logarithm to base e by $\ln$, respectively. $|| \cdot ||_r$ represents the ℓ_r norm ($r \geq 1$).

Assume there are T tasks. Let $(i,t) \in [m] \times [T]$, each task has m samples $(\mathbf{x}_i^t, y_i^t) \in \mathbb{R}^d \times \mathbb{R}$. Given matrix $\mathbf{M} \in \mathbb{R}^{d \times T}$, $\mathbf{M}^{(t)} \in \mathbb{R}^d$ denotes its t-th column, $\lambda_i(\mathbf{M})$ its i-th largest singular value, $tr(\mathbf{M})$ its trace, $||\mathbf{M}||_*$ its trace norm, $||\mathbf{M}||_2$ its spectral norm and $||\mathbf{M}||_F$ its Frobenius norm, respectively. We consider the linear multi-task model

$$y_i^t = \langle \mathbf{M}^{*(t)}, \mathbf{x}_i^t \rangle + \epsilon_i^t \tag{3}$$

where $\mathbf{M}^* \in \mathbb{R}^{d \times T}$ denotes the ground truth matrix, and ϵ_i^t are independent and identically distributed (i.i.d.) from Gaussian distribution $\mathcal{N}(0, \sigma^2)$ with zero mean and standard deviation σ. Let $\mathbf{x}_i^t$ be drawn i.i.d. from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with d-dimensional zero mean vector and $d \times d$ covariance matrix $\mathbf{\Sigma}$. This paper assumes $rank(\mathbf{M}^*) = r < \min\{T, d\}$. Boursier et al. (2022) show that the trace norm is a natural choice when estimating low rank matrices, since the trace norm is well-known to be a convex surrogate to the matrix rank. Trace norm based methods have already been successfully used in numerous domains (Cheng et al., 2011; Harchaoui et al., 2012), such as computer vision, collaborative filtering and matrix completion. Therefore, Boursier et al. (2022) focus on the trace norm regularized estimator, and this paper aims to improve the results of Boursier et al. (2022). In the future work, we will explore whether our results can be further improved based on other regularization.

Given observations $(\mathbf{x}_i^t, y_i^t)$ from model (3), this paper considers the following trace norm regularized estimator

$$\hat{\mathbf{M}} = \arg \min_{\mathbf{M} \in \mathbb{R}^{d \times T}} \frac{1}{mT} \sum_{(i,t) \in [m] \times [T]} (y_i^t - \langle \mathbf{M}^{(t)}, \mathbf{x}_i^t \rangle)^2 + \lambda ||\mathbf{M}||_* \tag{4}$$

where $\lambda > 0$ is a user-defined regularization parameter.

2.1. Upper Bounds

We define a $m \times T$ matrix as $\mathbf{L}(\mathbf{M}) := (\langle \mathbf{M}^{(t)}, \mathbf{x}_i^t \rangle)_{(i,t) \in [m] \times [T]}$. Var represents the variance. Let $\rho^2 := \max_{\mathbf{a} \in \mathbb{R}^d, ||\mathbf{a}||_2 = 1} Var\langle \mathbf{a}, \mathbf{x}_i^t \rangle$. The following theorem provides the upper bounds for the error of the estimator defined in (4).

Theorem 3 *Assume $\sqrt{1/T} - \sqrt{1/m} \geq \frac{3}{4\sqrt{2}}$, $c_1 m \geq 128 T tr(\mathbf{\Sigma}) c_2 r$, where c_1 and c_2 are two positive constants satisfying that $c_1 < 1 < c_2$, any optimal solution to (4) satisfies the bound*

$$||\hat{\mathbf{M}} - \mathbf{M}^*||_F \leq \frac{\sqrt{18r}}{c_1} \Big(32\rho\sigma(\sqrt{\frac{T+d}{mT}} + \delta) \Big) \tag{5}$$

with probability at least $1 - \frac{e^{-mT/64}}{1 - e^{-mT/64}} - e^{-mT/16} - 2e^{-4mT\delta^2}$.

The upper bound shown in (Boursier et al., 2022) is $c\sigma\sqrt{\frac{rd^2/m + rT}{m}} + c\sqrt{\frac{(Crd^2 + CrdT)\ln(dT/m)}{m^2}}$ under the conditions of $T > d$ and $d > m$. Theorem 3 shows that our upper bound is significantly sharper than that in (Boursier et al., 2022), and we improve their estimation error bound by a factor $\mathcal{O}(\sqrt{\frac{rd^2+rdT}{m^2}})$ and a logarithmic term without the assumptions of isotropic features, task diversity and $T > d$.

2.2. Lower Bounds

To complement our upper bounds, this subsection lower bounds the minimax rates of the estimation error over classes of matrices with rank r. Let ϕ^2 be the maximum diagonal entry of the covariance matrix $\mathbf{\Sigma}$. Let

$$\mathbf{y} := (y_1^1, \dots, y_m^1, \dots, y_1^T, \dots, y_m^T)' \in \mathbb{R}^{mT}$$
$$\mathbf{X} := (\mathbf{x}_1^1, \dots, \mathbf{x}_m^1, \dots, \mathbf{x}_1^T, \dots, \mathbf{x}_m^T)' \in \mathbb{R}^{mT\times d}$$

Given the matrix classes $\Theta := \{\mathbf{M}^* \in \mathbb{R}^{d\times T} | rank(\mathbf{M}^*) = r\}$, we consider the following minimax risk in Frobenius norm

$$\mathfrak{M}(\Theta) := \inf_{\hat{\mathbf{M}}} \sup_{\mathbf{M}^*\in\Theta} E||\hat{\mathbf{M}} - \mathbf{M}^*||_F^2$$

where the infimum is taken over all estimators $\hat{\mathbf{M}}$ that are measurable functions of samples.

Theorem 4 *There is a universal numerical constant $c_0 > 0$ such that*

$$\mathfrak{M}(\Theta) \geq \frac{c_0\sigma^2 r(d+T)}{m\phi^2(1 + 32(d-1)\ln(dT))} \tag{6}$$

Let $c_3 > 0$ be a positive constant. If $T = \mathcal{O}(d)$, Theorem 4 establishes that the upper bounds obtained in Theorem 3 are minimax-optimal up to constant factors.

3. Proofs

In this section, we present the proofs of our main results.

3.1. Proof of Theorem 3

Our proof relies on Theorem 9.19 of Wainwright (2019) given the decomposability of regularizer, the restricted strong convexity (RSC) of the cost function and the dual norm bound. Negahban et al. (2009) have shown the decomposability of the trace norm regularizer. Showing both RSC condition and dual norm bound hold with high probability are the main technical challenges of this proof.

Proposition 5 $\mathbf{L}(\mathbf{M})$ *is drawn from the $\tilde{\mathbf{\Sigma}}(\mathbf{M})$-Gaussian ensemble, with rows sampled i.i.d. from a $\mathcal{N}(\boldsymbol{0}, \tilde{\mathbf{\Sigma}}(\mathbf{M}))$, where* $\tilde{\mathbf{\Sigma}}(\mathbf{M}) := \begin{bmatrix} \mathbf{M}^{1'}\mathbf{\Sigma}\mathbf{M}^1 & & \\ & \ddots & \\ & & \mathbf{M}^{T'}\mathbf{\Sigma}\mathbf{M}^T \end{bmatrix}_{T\times T}$.

Proposition 6 $\sqrt{tr(\tilde{\mathbf{\Sigma}}(\mathbf{M}))} = ||\sqrt{\mathbf{\Sigma}}\mathbf{M}||_F$.

Proof

$$||\sqrt{\mathbf{\Sigma}}\mathbf{M}||_F = \sqrt{tr(\mathbf{M}'\sqrt{\mathbf{\Sigma}}'\sqrt{\mathbf{\Sigma}}\mathbf{M})} = \sqrt{tr(\mathbf{M}'\mathbf{\Sigma}\mathbf{M})} = \sqrt{tr(\tilde{\mathbf{\Sigma}}(\mathbf{M}))}$$

■

We define the first-order Taylor-series error as below.

$$\varepsilon(\mathbf{M}) := \frac{1}{mT}\sum_{(i,t)\in[m]\times[T]}\Big((y_i^t - \langle\mathbf{M}^{(t)}+\mathbf{M}^{*(t)}, \mathbf{x}_i^t\rangle)^2 - (y_i^t - \langle\mathbf{M}^{*(t)}, \mathbf{x}_i^t\rangle)^2 - \langle\nabla(y_i^t - \langle\mathbf{M}^{*(t)}, \mathbf{x}_i^t\rangle)^2, \mathbf{M}\rangle\Big)$$
$$= \frac{1}{mT}\sum_{(i,t)\in[m]\times[T]}\langle\mathbf{M}^{(t)}, \mathbf{x}_i^t\rangle^2 = \frac{1}{mT}||\mathbf{L}(\mathbf{M})||_F^2$$

Our first step is to establish the restricted strong convexity condition for $\varepsilon(\mathbf{M})$ with high probability.

Lemma 7 (RSC condition) *Assume $\sqrt{1/T} - \sqrt{1/m} \geq \frac{3}{4\sqrt{2}}$. Let $\mathbf{L}(\mathbf{M})$ be drawn from the $\tilde{\mathbf{\Sigma}}(\mathbf{M})$-Gaussian ensemble, there are positive constants $c_1 < 1 < c_2$ such that*

$$\varepsilon(\mathbf{M}) \geq c_1||\sqrt{\mathbf{\Sigma}}\mathbf{M}||_F^2 - \frac{c_2 T tr(\mathbf{\Sigma})||\mathbf{M}||_*^2}{m}, \forall \mathbf{M} \in \mathbb{R}^{d\times T} \tag{7}$$

with probability at least $1 - \frac{e^{-mT/64}}{1-e^{-mT/64}}$.

Remark. Note that our RSC condition Lemma 7 is different from Lemma 1 of Boursier et al. (2022) in the following two aspects. Firstly, Lemma 1 in Boursier et al. (2022) holds only for the cone set of matrices, while our Lemma 7 holds for any matrices in $\mathbb{R}^{d\times T}$. Secondly, our Lemma 7 is consistent with the RSC condition defined in Definition 9.15 of Wainwright (2019), while Lemma 1 of Boursier et al. (2022) does not follow Definition 9.15 of Wainwright (2019).

Proof By a rescaling argument, we consider $\Gamma(\mathbf{\Sigma}) := \{\mathbf{M} \in \mathbb{R}^{d\times T} | ||\sqrt{\mathbf{\Sigma}}\mathbf{M}||_F = 1\}$. Let $g(t) := \frac{4\sqrt{2Ttr(\mathbf{\Sigma})}t}{\sqrt{m}}$ and $\Phi := \{\mathbf{x}_i^t, (i,t) \in [m]\times[T] | \inf_{\mathbf{M}\in\Gamma(\mathbf{\Sigma})} \frac{||\mathbf{L}(\mathbf{M})||_F}{\sqrt{mT}} \leq 1/4 - 2g(||\mathbf{M}||_*)\}$.

We first show that the lower bound (7) holds on the complementary set $\overline{\Phi}$. Given $\mathbf{M} \in \Gamma(\mathbf{\Sigma})$, define $a = 1/4$, $b = 2g(||\mathbf{M}||_*)$, $c = \frac{||\mathbf{L}(\mathbf{M})||_F}{\sqrt{mT}}$, and we obtain $c \geq \max\{a-b, 0\}$ on the event $\overline{\Phi}$. We first show that $c^2 \geq (1-\delta)^2a^2 - b^2/\delta^2$ for any $0 < \delta < 1$. If $b/\delta \geq a$, then the bound is trivial. Assume $b \leq a\delta$. $c \geq a - b$ implies that $c \geq (1-\delta)a$, and $c^2 \geq (1-\delta)^2a^2 - b^2/\delta^2$. Setting $\delta = 1/2$, then the lower bound (7) holds. Then, we need to upper bounding the probability of Φ.

Let $0 \leq r_1 \leq r_2$, we define the sets $K(r_1, r_2) := \{\mathbf{M} \in \Gamma(\mathbf{\Sigma}) | g(||\mathbf{M}||_*) \in [r_1, r_2]\}$, $Q(r_1, r_2) := \{\inf_{\mathbf{M}\in K(r_1,r_2)} \frac{||\mathbf{L}(\mathbf{M})||_F}{\sqrt{mT}} \leq 1/2 - r_2\}$, and present the key lemma as below.

Lemma 8 *Given $0 \leq r_1 \leq r_2$, we have*

$$\mathbb{P}(Q(r_1, r_2)) \leq e^{-mT/32}e^{-mTr_2^2/8}, \tag{8}$$

for $\mu = 1/4$, we have

$$\Phi \subseteq Q(0,\mu) \cup (\bigcup_{i=1}^{\infty} Q(2^{i-1}\mu, 2^i\mu)). \tag{9}$$

Combining (9) and the union bound, we have

$$\mathbb{P}(\Phi) \leq \mathbb{P}(Q(0,\mu)) + \sum_{i=1}^{\infty} \mathbb{P}(Q(2^{i-1}\mu, 2^i\mu))) \leq e^{-mT/32}(\sum_{i=0}^{\infty} e^{-mT2^{2i}\mu^2/8})$$

Since $\mu = 1/4$ and $2^{2i} \geq 2i$, we obtain

$$\mathbb{P}(\Phi) \leq e^{-mT/32}(\sum_{i=0}^{\infty} e^{-mT2^{2i}\mu^2/8}) \leq e^{-mT/32}(\sum_{i=0}^{\infty}(e^{-mT\mu^2/4})^i) \leq \frac{e^{-mT/64}}{1-e^{-mT/64}}$$

Thus, the lower bound (7) holds with probability at least $1 - \frac{e^{-m/64}}{1-e^{-m/64}}$. The rest part is devoted to prove Lemma 8.

Let $\mathbf{M} \in \Gamma(\mathbf{\Sigma})$ certify Φ. Then, it must belong either to the set $K(0,\mu)$ or to a set $K(2^{i-1}\mu, 2^i\mu)$ for some $i = 1, 2, \ldots$. First suppose that $\mathbf{M} \in K(0,\mu)$, so $g(||\mathbf{M}||_*) \leq \mu = 1/4$. Since $\mathbf{M}$ certifies the event Φ, we have $\frac{||\mathbf{L}(\mathbf{M})||_F}{\sqrt{mT}} \leq 1/4 - 2g(||\mathbf{M}||_*) \leq 1/4 = 1/2 - \mu$, showing that the event $Q(0,\mu)$ must happen. Otherwise, we must have $\mathbf{M} \in K(2^{i-1}\mu, 2^i\mu)$ for some $i = 1, 2, \ldots$, and we have $\frac{||\mathbf{L}(\mathbf{M})||_F}{\sqrt{mT}} \leq 1/4 - 2g(||\mathbf{M}||_*) \leq 1/4 - 22^{i-1}\mu \leq 1/2 - 2^i\mu$, showing that the event $Q(2^{i-1}\mu, 2^i\mu)$ must happen. Thus, (9) holds.

The following content shows that (8) holds. We consider the random variable $\Upsilon(r_1, r_2) := -\inf_{\mathbf{M}\in K(r_1,r_2)} \frac{||\mathbf{L}(\mathbf{M})||_F}{\sqrt{mT}}$, and the class of matrices $\Lambda_{m,T} := \{\mathbf{N} \in \mathbb{R}^{m\times T} | rank(\mathbf{N}) = 1, ||\mathbf{N}||_F = 1\}$. The spectral norm of matrix $\mathbf{L}(\mathbf{M})$ has the variational representation $||\mathbf{L}(\mathbf{M})||_2 = \sup_{\mathbf{N}\in\Lambda_{m,T}} \langle \mathbf{L}(\mathbf{M}), \mathbf{N}\rangle$. Let $\mathbf{W} \in \mathbb{R}^{m\times T}$ be the standard Gaussian ensemble with i.i.d. $\mathcal{N}(0,1)$ entries. $\mathbf{L}(\mathbf{M})$ can be written as $\mathbf{L}(\mathbf{M}) = \mathbf{W}\sqrt{\tilde{\mathbf{\Sigma}}(\mathbf{M})}$. We have

$$\begin{aligned}
\Upsilon(r_1, r_2) = -\inf_{\mathbf{M}\in K(r_1,r_2)} \frac{||\mathbf{L}(\mathbf{M})||_F}{\sqrt{mT}} &\leq -\inf_{\mathbf{M}\in K(r_1,r_2)} \frac{||\mathbf{L}(\mathbf{M})||_2}{\sqrt{mT}} \\
&= -\inf_{\mathbf{M}\in K(r_1,r_2)} \sup_{\mathbf{N}\in\Lambda_{m,T}} \frac{\langle \mathbf{L}(\mathbf{M}), \mathbf{N}\rangle}{\sqrt{mT}} \\
&= \sup_{\mathbf{M}\in K(r_1,r_2)} \inf_{\mathbf{N}\in\Lambda_{m,T}} \frac{\langle \mathbf{W}\sqrt{\tilde{\mathbf{\Sigma}}(\mathbf{M})}, \mathbf{N}\rangle}{\sqrt{mT}} \\
&= \sup_{\mathbf{M}\in K(r_1,r_2)} \inf_{\mathbf{N}\in\Lambda_{m,T}} \frac{\left\langle \mathbf{W}, \mathbf{N}\sqrt{\tilde{\mathbf{\Sigma}}(\mathbf{M})}\right\rangle}{\sqrt{mT}}
\end{aligned} \tag{10}$$

The following Gordon's inequality (Vershynin, 2018) plays the key role in the remainder of our proof.

Lemma 9 (Gordon's inequality) *Let $Z_{u,v}$ and $Y_{u,v}$ be two zero-mean Gaussian processes indexed by pairs of points (u,v) in a product set $\mathbb{T} = U \times V$. Suppose that $E(Z_{u,v} - Z_{\breve{u},\breve{v}})^2 \leq E(Y_{u,v} - Y_{\breve{u},\breve{v}})^2$ for all pairs $(u,v) \in \mathbb{T}$ and $(\breve{u},\breve{v}) \in \mathbb{T}$, and this inequality holds with equality whenever $v = \breve{v}$. Then we have*

$$E \sup_{v\in V} \inf_{u\in U} Z_{u,v} \leq E \sup_{v\in V} \inf_{u\in U} Y_{u,v}$$

Let $(\mathbf{M}, \mathbf{N})$ and $(\breve{\mathbf{M}}, \breve{\mathbf{N}})$ be any two pairs in $K(r_1, r_2) \times \Lambda_{m,T}$. The zero-mean Gaussian process is defined as $Z_{\mathbf{M},\mathbf{N}} := \Big\langle \mathbf{W}, \mathbf{N}\sqrt{\tilde{\boldsymbol{\Sigma}}(\mathbf{M})} \Big\rangle$. We find that

$$\begin{aligned} E(Z_{\mathbf{M},\mathbf{N}} - Z_{\breve{\mathbf{M}},\breve{\mathbf{N}}})^2 &= E\Big(\Big\langle \mathbf{W}, \mathbf{N}\sqrt{\tilde{\boldsymbol{\Sigma}}(\mathbf{M})} - \breve{\mathbf{N}}\sqrt{\tilde{\boldsymbol{\Sigma}}(\breve{\mathbf{M}})} \Big\rangle\Big)^2 \\ &= \Big|\Big|\mathbf{N}\sqrt{\tilde{\boldsymbol{\Sigma}}(\mathbf{M})} - \breve{\mathbf{N}}\sqrt{\tilde{\boldsymbol{\Sigma}}(\breve{\mathbf{M}})}\Big|\Big|_F^2 \\ &= \Big|\Big|\mathbf{N}\Big(\sqrt{\tilde{\boldsymbol{\Sigma}}(\mathbf{M})} - \sqrt{\tilde{\boldsymbol{\Sigma}}(\breve{\mathbf{M}})}\Big) + \Big(\mathbf{N} - \breve{\mathbf{N}}\Big)\sqrt{\tilde{\boldsymbol{\Sigma}}(\breve{\mathbf{M}})}\Big|\Big|_F^2 \\ &\leq 2||\mathbf{N}||_F^2\Big|\Big|\sqrt{\tilde{\boldsymbol{\Sigma}}(\mathbf{M})} - \sqrt{\tilde{\boldsymbol{\Sigma}}(\breve{\mathbf{M}})}\Big|\Big|_F^2 + 2||\mathbf{N} - \breve{\mathbf{N}}||_F^2\Big|\Big|\sqrt{\tilde{\boldsymbol{\Sigma}}(\breve{\mathbf{M}})}\Big|\Big|_F^2 \\ &= 2\Big|\Big|\sqrt{\tilde{\boldsymbol{\Sigma}}(\mathbf{M})} - \sqrt{\tilde{\boldsymbol{\Sigma}}(\breve{\mathbf{M}})}\Big|\Big|_F^2 + 2||\mathbf{N} - \breve{\mathbf{N}}||_F^2 \end{aligned} \tag{11}$$

where the last step follows from the definition of $\Lambda_{m,T}$, $K(r_1, r_2)$ and Proposition 6:

$$\Big|\Big|\sqrt{\tilde{\boldsymbol{\Sigma}}(\breve{\mathbf{M}})}\Big|\Big|_F^2 = tr(\tilde{\boldsymbol{\Sigma}}(\breve{\mathbf{M}})) = ||\sqrt{\boldsymbol{\Sigma}}\breve{\mathbf{M}}||_F^2 = 1$$

Motivated by (11), we define the zero-mean Gaussian processes $Y_{\mathbf{M},\mathbf{N}} := \sqrt{2}\Big\langle \mathbf{W}_1, \sqrt{\tilde{\boldsymbol{\Sigma}}(\mathbf{M})} \Big\rangle + \sqrt{2}\langle \mathbf{W}_2, \mathbf{N} \rangle$, where $\mathbf{W}_1 \in \mathbb{R}^{T \times T}$ and $\mathbf{W}_2 \in \mathbb{R}^{m \times T}$ are both standard Gaussian ensemble with i.i.d. $\mathcal{N}(0, 1)$ entries. We have $E(Y_{\mathbf{M},\mathbf{N}} - Y_{\breve{\mathbf{M}},\breve{\mathbf{N}}})^2 = 2\Big|\Big|\sqrt{\tilde{\boldsymbol{\Sigma}}(\mathbf{M})} - \sqrt{\tilde{\boldsymbol{\Sigma}}(\breve{\mathbf{M}})}\Big|\Big|_F^2 + 2||\mathbf{N} - \breve{\mathbf{N}}||_F^2$. Using Lemma 9, we find that

$$\begin{aligned} E\Upsilon(r_1, r_2) &\leq E\Big(\sup_{\mathbf{M} \in K(r_1, r_2)} \inf_{\mathbf{N} \in \Lambda_{m,T}} \frac{Z_{\mathbf{M},\mathbf{N}}}{\sqrt{mT}}\Big) \\ &\leq E\Big(\sup_{\mathbf{M} \in K(r_1, r_2)} \inf_{\mathbf{N} \in \Lambda_{m,T}} \frac{Y_{\mathbf{M},\mathbf{N}}}{\sqrt{mT}}\Big) \\ &= \sqrt{2}E\Big(\sup_{\mathbf{M} \in K(r_1, r_2)} \frac{\Big\langle \mathbf{W}_1, \sqrt{\tilde{\boldsymbol{\Sigma}}(\mathbf{M})} \Big\rangle}{\sqrt{mT}}\Big) + \sqrt{2}E\Big(\inf_{\mathbf{N} \in \Lambda_{m,T}} \frac{\langle \mathbf{W}_2, \mathbf{N} \rangle}{\sqrt{mT}}\Big) \end{aligned} \tag{12}$$

Lemma 10 *Consider a standard Gaussian ensemble* $\mathbf{W} \in \mathbb{R}^{m \times T}$ *generated with i.i.d.* $\mathcal{N}(0, 1)$ *entries. The expectation of the maximum singular value* $\sigma_{max}(\mathbf{W})$ *and minimum singular value* $\sigma_{min}(\mathbf{W})$ *satisfy the following upper and lower bounds:*

$$\begin{aligned} E\sigma_{max}(\mathbf{W}) &\leq \sqrt{m} + \sqrt{T} \\ E\sigma_{min}(\mathbf{W}) &\geq \sqrt{m} - \sqrt{T} \end{aligned}$$

The proof details of Lemma 10 can be found in the Appendix A. Combining Hölder's inequality, Proposition 6 and Lemma 10 imply that

$$\begin{aligned}
E\Big(\sup_{\mathbf{M}\in K(r_1,r_2)} \frac{\Big\langle \mathbf{W}_1, \sqrt{\tilde{\mathbf{\Sigma}}(\mathbf{M})}\Big\rangle}{\sqrt{mT}}\Big) &\leq E\Big(\sup_{\mathbf{M}\in K(r_1,r_2)} \frac{||\mathbf{W}_1||_2\Big|\Big|\sqrt{\tilde{\mathbf{\Sigma}}(\mathbf{M})}\Big|\Big|_*}{\sqrt{mT}}\Big) && \text{(Hölder's inequality)}\\
&\leq \frac{2\sqrt{T}}{\sqrt{mT}}\sup_{\mathbf{M}\in K(r_1,r_2)}\Big|\Big|\sqrt{\tilde{\mathbf{\Sigma}}(\mathbf{M})}\Big|\Big|_* && \text{(Lemma 10)}\\
&\leq \frac{2}{\sqrt{m}}\sup_{\mathbf{M}\in K(r_1,r_2)}\sum_{i=1}^{T}\sqrt{\mathbf{M}^{i'}\mathbf{\Sigma}\mathbf{M}^i}\\
&\leq \frac{2}{\sqrt{m}}\sup_{\mathbf{M}\in K(r_1,r_2)}\sqrt{T}\sqrt{\sum_{i=1}^{T}\mathbf{M}^{i'}\mathbf{\Sigma}\mathbf{M}^i}\\
&\leq \frac{2\sqrt{T}}{\sqrt{m}}\sup_{\mathbf{M}\in K(r_1,r_2)}||\sqrt{\mathbf{\Sigma}}\mathbf{M}||_F && \text{(Proposition 6)}\\
&\leq \frac{2\sqrt{T}}{\sqrt{m}}\sup_{\mathbf{M}\in K(r_1,r_2)}||\sqrt{\mathbf{\Sigma}}||_F||\mathbf{M}||_*\\
&\leq \frac{r_2}{2\sqrt{2}} && \text{(definition of } K(r_1,r_2))
\end{aligned} \tag{13}$$

Lemma 10 also implies that

$$E\Big(\inf_{\mathbf{N}\in\Lambda_{m,T}}\frac{\langle \mathbf{W}_2,\mathbf{N}\rangle}{\sqrt{mT}}\Big) = -E\Big(\sup_{\mathbf{N}\in\Lambda_{m,T}}\frac{\langle \mathbf{W}_2,\mathbf{N}\rangle}{\sqrt{mT}}\Big) = -E\frac{||\mathbf{W}_2||_2}{\sqrt{mT}} \leq \frac{\sqrt{T}-\sqrt{m}}{\sqrt{mT}} \tag{14}$$

(12), (13) and (14) show that

$$E\Upsilon(r_1,r_2) \leq \frac{(\sqrt{T}-\sqrt{m})\sqrt{2}}{\sqrt{mT}} + \frac{r_2}{2} \tag{15}$$

(10) shows that the random variable $\sqrt{mT}\Upsilon(r_1,r_2)$ is a 1-Lipschitz function of the standard Gaussian matrix $\mathbf{W}$. Theorem 2.26 in Wainwright (2019) implies that

$$\mathbb{P}(\Upsilon(r_1,r_2) \geq E\Upsilon(r_1,r_2) + t) \leq e^{-mTt^2/2}, \forall t > 0 \tag{16}$$

Assume $\sqrt{1/T} - \sqrt{1/m} \geq \frac{3}{4\sqrt{2}}$. Then the constant $C = \frac{(-\sqrt{T}+\sqrt{m})\sqrt{2}}{\sqrt{mT}} - 1/2 \geq 1/4$. Setting $t = C + \frac{r_2}{2}$, (15) and (16) show that

$$\mathbb{P}(\Upsilon(r_1,r_2) \geq -1/2 + r_2) \leq e^{-mTC^2/2}e^{-mTr_2^2/8} \leq e^{-mT/32}e^{-mTr_2^2/8} \tag{17}$$

Thus, (8) holds. ■

Let $\mathcal{L}_{mT}(\mathbf{M}^*) := \frac{1}{mT}\sum_{(i,t)\in[m]\times[T]}(y_i^t - \langle \mathbf{M}^{*(t)}, \mathbf{x}_i^t\rangle)^2$. The following Lemma shows the dual norm bound holds with high probability.

Lemma 11 (Dual norm bound) *With probability at least* $1 - e^{-mT/16} - 2e^{-4mT\delta^2}$*, we have*

$$||\nabla \mathcal{L}_{mT}(\mathbf{M}^*)||_2 \le 16\rho\sigma(\sqrt{\frac{T+d}{mT}} + \delta)$$

Proof

$$\begin{aligned}\nabla \mathcal{L}_{mT}(\mathbf{M}^*) &= \frac{-2}{mT}[\sum_{i\in[m]}(y_i^1 - \langle \mathbf{M}^{*(1)}, \mathbf{x}_i^1\rangle)\mathbf{x}_i^1, \ldots, \sum_{i\in[m]}(y_i^T - \langle \mathbf{M}^{*(T)}, \mathbf{x}_i^T\rangle)\mathbf{x}_i^T]_{d\times T} \\ &= \frac{-2}{mT}[\sum_{i\in[m]} \epsilon_i^1 \mathbf{x}_i^1, \ldots, \sum_{i\in[m]} \epsilon_i^T \mathbf{x}_i^T]_{d\times T}\end{aligned}$$

Based on the event $\mathscr{E} := \{\frac{||(\epsilon_i^t)_{(i,t)\in[m]\times[T]}||_F^2}{mT} \le 2\sigma^2\}$, we have

$$\mathbb{P}(||\nabla \mathcal{L}_{mT}(\mathbf{M}^*)||_2 \ge \lambda_{mT}/2) \le \mathbb{P}(\overline{\mathscr{E}}) + \mathbb{P}(||\nabla \mathcal{L}_{mT}(\mathbf{M}^*)||_2 \ge \lambda_{mT}/2|\mathscr{E}) \tag{18}$$

Since ϵ_i^t are drawn i.i.d. from Gaussian distribution $\mathcal{N}(0, \sigma^2)$, Exercise 2.6 in Wainwright (2019) implies that $\mathbb{P}(\overline{\mathscr{E}}) \le e^{-mT/16}$. It remains to upper bound $\mathbb{P}(||\nabla \mathcal{L}_{mT}(\mathbf{M}^*)||_2 \ge \lambda_{mT}/2|\mathscr{E})$.

Let $\{\mathbf{a}^1, \ldots, \mathbf{a}^I\}$ and $\{\mathbf{b}^1, \ldots, \mathbf{b}^J\}$ be $1/4$-covers in Euclidean norm of the spheres $\mathbb{S}^{T-1}$ and $\mathbb{S}^{d-1}$ respectively. Lemma 5.7 in Wainwright (2019) implies that $I \le 9^T$ and $J \le 9^d$. For any $\mathbf{b} \in \mathbb{S}^{d-1}$, we can write $\mathbf{b} = \mathbf{b}^j + \mathbf{c}$ for some vector $\mathbf{c}$ with ℓ_2 distance at most $1/4$, and we have

$$\begin{aligned}||\nabla \mathcal{L}_{mT}(\mathbf{M}^*)||_2 &= ||\nabla \mathcal{L}_{mT}(\mathbf{M}^*)'||_2 \\ &= \sup_{\mathbf{b}\in\mathbb{S}^{d-1}} ||\nabla \mathcal{L}_{mT}(\mathbf{M}^*)'\mathbf{b}||_2 \\ &\le 1/4||\nabla \mathcal{L}_{mT}(\mathbf{M}^*)'||_2 + \max_{j\in[J]} ||\nabla \mathcal{L}_{mT}(\mathbf{M}^*)'\mathbf{b}^j||_2\end{aligned} \tag{19}$$

Based on the cover of $\mathbb{S}^{T-1}$, we use the similar argument and obtain that

$$||\nabla \mathcal{L}_{mT}(\mathbf{M}^*)'\mathbf{b}^j||_2 \le 1/4||\nabla \mathcal{L}_{mT}(\mathbf{M}^*)'||_2 + \max_{i\in[I]} \langle \mathbf{a}^i, \mathcal{L}_{mT}(\mathbf{M}^*)'\mathbf{b}^j\rangle \tag{20}$$

(19) and (20) imply that $||\nabla \mathcal{L}_{mT}(\mathbf{M}^*)'||_2 \le 2\max_{j\in[J]} \max_{i\in[I]} |\vartheta^{i,j}|$, where $\vartheta^{i,j} = \langle \mathbf{a}^i, \mathcal{L}_{mT}(\mathbf{M}^*)'\mathbf{b}^j\rangle$. Fix some index pair (i, j, k), using the definition of ρ^2, we have

$$\begin{aligned}Var\langle \mathbf{a}^i, [\epsilon_k^1 \mathbf{x}_k^{1'}\mathbf{b}^j, \ldots, \epsilon_k^T \mathbf{x}_k^{T'}\mathbf{b}^j]'\rangle &= Var(\sum_{z=1}^T \mathbf{a}_z^i \epsilon_k^z \mathbf{x}_k^{z'}\mathbf{b}^j) \\ &= E(\sum_{z=1}^T (\mathbf{a}_z^i)^2 (\epsilon_k^z)^2 (\mathbf{x}_k^{z'}\mathbf{b}^j)^2) \\ &\le \rho^2 E(\sum_{z=1}^T (\epsilon_k^z)^2) \qquad \text{(definition of } \rho^2)\end{aligned} \tag{21}$$

Conditioning on event $\mathscr{E}$ and using (21), we have

$$\begin{aligned} Var(\vartheta^{i,j}|\mathscr{E}) &= \frac{4}{m^2T^2}\sum_{k=1}^{m} Var(\langle \mathbf{a}^i, [\epsilon_k^1 \mathbf{x}_k^{1'} \mathbf{b}^j, \dots, \epsilon_k^T \mathbf{x}_k^{T'} \mathbf{b}^j]' \rangle|\mathscr{E}) \\ &\leq \frac{4\rho^2}{m^2T^2} E(\sum_{k=1}^{m}\sum_{z=1}^{T}(\epsilon_k^z)^2|\mathscr{E}) \\ &\leq \frac{8\rho^2\sigma^2}{mT} \end{aligned} \tag{22}$$

(22) shows that $\vartheta^{i,j}$ is zero-mean Gaussian with variance at most $\frac{8\rho^2\sigma^2}{mT}$ conditioning on event $\mathscr{E}$, and implies that

$$\begin{aligned} \mathbb{P}(||\nabla \mathcal{L}_{mT}(\mathbf{M}^*)||_2 \geq \lambda_{mT}/2|\mathscr{E}) &\leq \sum_{j\in[J]}\sum_{i\in[I]} \mathbb{P}(|\vartheta^{i,j}| \geq \lambda_{mT}/4|\mathscr{E}) && \text{(union bound)} \\ &\leq 2e^{-\frac{mT\lambda_{mT}^2}{16^2\rho^2\sigma^2} + \ln I + \ln J} && \text{(sub-Gaussian tail bounds)} \\ &\leq 2e^{-\frac{mT\lambda_{mT}^2}{16^2\rho^2\sigma^2} + (T+d)\ln 9} \end{aligned} \tag{23}$$

Setting $\lambda_{mT} = 32\rho\sigma(\sqrt{\frac{T+d}{mT}} + \delta)$, (18) and (23) show that

$$\mathbb{P}(||\nabla \mathcal{L}_{mT}(\mathbf{M}^*)||_2 \geq 16\rho\sigma(\sqrt{\frac{T+d}{mT}} + \delta)) \leq e^{-mT/16} + 2e^{-4mT\delta^2}$$

■

Combining Theorem 9.19 of Wainwright (2019), Lemma 7 and Lemma 11, we obtain Theorem 3.

3.2. Proof of Theorem 4

Our proof is based on a standard reduction from lower bounding the probability of error to an N-ary hypothesis testing problem (Hasminskii, 1979; Yang and Barron, 1999) over a packing set of matrix pairs.

In particular, suppose that $\{\mathbf{M}^1, \cdots, \mathbf{M}^N\}$ is a 2ϱ-separated set contained in Θ, meaning a collection of elements $||\mathbf{M}^i - \mathbf{M}^j||_F \geq 2\varrho$ for all $i \neq j$. Let $\mathbb{P}_{\mathbf{M}^i}$ be the distribution that links to each $\mathbf{M}^i$. Given a class of distributions $\{\mathbb{P}_{\mathbf{M}^i}, i \in [N]\}$, a random matrix $\mathbf{Z} := (\mathbf{X}, \mathbf{y})$ is generated from the following procedure:(1) Sample a random integer B from the uniform distribution over the index set $[N]$. (2) Given $B = i$, sample $\mathbf{Z}$ from $\mathbb{P}_{\mathbf{M}^i}$. In this way, the observation follows the mixture distribution $\mathbb{Q} := 1/N\sum_{i=1}^{N}\mathbb{P}_{\mathbf{M}^i}$. We define a mapping function $\psi : \mathcal{I} \rightarrow [N]$. Our goal is to identify the index B of the probability distribution from which a given sample has been drawn. Proposition 15.1 in Wainwright (2019) implies that

$$\mathfrak{M}(\Theta) \geq \varrho^2 \inf_{\psi} \mathbb{P}(\psi(\mathbf{X}, \mathbf{y}) \neq B) \tag{24}$$

The remainder of the proof is devoted to low bounding $\mathbb{P}(\psi(\mathbf{X},\mathbf{y}) \neq B)$.

Conditioning on a particular instantiation $\mathbf{X} := (\mathbf{x}_1^1, \dots, \mathbf{x}_m^1, \dots, \mathbf{x}_1^T, \dots, \mathbf{x}_m^T)'$, we derive the following lower bound by using a form of Fano's inequality that involves the mutual information $\mathbf{I}_{\mathbf{X}}(\mathbf{y}; B)$ between the random vector $\mathbf{y}$ and the random index B with fixed $\mathbf{X}$.

$$\mathbb{P}(\psi(\mathbf{X},\mathbf{y}) \neq B|\mathbf{X}) \geq 1 - \frac{\mathbf{I}_{\mathbf{X}}(\mathbf{y}; B) + \ln 2}{\ln N} \tag{25}$$

By taking averages over $\mathbf{X}$, we obtain lower bound on $\mathbb{P}(\psi(\mathbf{X},\mathbf{y}) \neq B)$ that involve the quantity $E_{\mathbf{X}}(\mathbf{I}_{\mathbf{X}}(\mathbf{y}; B))$. $||\mathbf{M}||_\infty := \max_{j\in[d]} \max_{k\in[T]} |\mathbf{M}_{jk}|$ denotes the element-wise maximum of matrix $\mathbf{M}$. We first present the following Lemma from Agarwal et al. (2011).

Lemma 12 *For $d, T \geq 10$, $\varrho > 0$, and for each $r = 1, \dots, \min\{T, d\}$, there exists a set of $d \times T$-dimensional matrices $\{\mathbf{M}^1, \cdots, \mathbf{M}^N\}$ with cardinality $N \geq 1/4e^{\frac{r(d+T)}{128}}$ such that each matrix has rank r, and moreover*

$$\begin{aligned} ||\mathbf{M}^i||_F &= 2\varrho && \textit{for all } i = 1, \dots, N \\ ||\mathbf{M}^i - \mathbf{M}^j||_F &\geq 2\varrho && \textit{for all } i \neq j \\ ||\mathbf{M}^i||_\infty &\leq 2\varrho\sqrt{\frac{32\ln dT}{dT}} && \textit{for all } i = 1, \dots, N \end{aligned} \tag{26}$$

Lemma 12 shows that $||\mathbf{M}^i - \mathbf{M}^j||_F \leq 4\varrho$ for all $i \neq j$. Let $\mathbf{I}_{mT\times mT}$ be the $mT \times mT$ identity matrix. We define $\aleph := 2\varrho\sqrt{\frac{32\ln dT}{dT}}$. Let $\mathbb{P}_{\mathbf{M}^i}$ denote the distribution of $\mathbf{y}$ given ground truth matrix $\mathbf{M}^i$ and $\mathbf{X}$. Under $\mathbb{P}_{\mathbf{M}^i}$, our model (3) shows that $\mathbf{y}$ follows a distribution $\mathcal{N}(\xi_{\mathbf{M}^i}, \sigma^2\mathbf{I}_{mT\times mT})$ with mT-dimensional mean vector

$$\xi_{\mathbf{M}^i} = (\langle \mathbf{M}^{i(1)}, \mathbf{x}_1^1\rangle, \dots, \langle \mathbf{M}^{i(1)}, \mathbf{x}_m^1\rangle, \dots, \langle \mathbf{M}^{i(T)}, \mathbf{x}_1^T\rangle, \dots, \langle \mathbf{M}^{i(T)}, \mathbf{x}_m^T\rangle)' \in \mathbb{R}^{mT}$$

and $mT \times mT$ covariance matrix $\sigma^2\mathbf{I}_{mT\times mT}$. Let $D(\mathbb{P}_{\mathbf{M}^i}||\mathbb{P}_{\mathbf{M}^j})$ denote the Kullback-Leibler divergence between the distributions of $\mathbb{P}_{\mathbf{M}^i}$ and $\mathbb{P}_{\mathbf{M}^j}$.

A simple upper bound on the mutual information can be derived by using the convexity of the Kullback-Leibler divergence: $\mathbf{I}_{\mathbf{X}}(\mathbf{y}; B) \leq 1/N^2 \sum_{i,j} D(\mathbb{P}_{\mathbf{M}^i}||\mathbb{P}_{\mathbf{M}^j})$. Let $\underset{(i,j)}{\triangle}\mathbf{M} := \mathbf{M}^i - \mathbf{M}^j$. Exercise 15.13 in Wainwright (2019) ensures that

$$\begin{aligned} D(\mathbb{P}_{\mathbf{M}^i}||\mathbb{P}_{\mathbf{M}^j}) &= \frac{1}{2\sigma^2}||\xi_{\mathbf{M}^i} - \xi_{\mathbf{M}^j}||_2^2 \\ &= \frac{1}{2\sigma^2}||(\langle \underset{(i,j)}{\triangle}\mathbf{M}^{(1)}, \mathbf{x}_1^1\rangle, \dots, \langle \underset{(i,j)}{\triangle}\mathbf{M}^{(1)}, \mathbf{x}_m^1\rangle, \dots, \langle \underset{(i,j)}{\triangle}\mathbf{M}^{(T)}, \mathbf{x}_1^T\rangle, \dots, \langle \underset{(i,j)}{\triangle}\mathbf{M}^{(T)}, \mathbf{x}_m^T\rangle)||_2^2 \end{aligned}$$

Using Lemma 12, we have

$$\begin{aligned}
E_{\mathbf{X}}(\mathbf{I}_{\mathbf{X}}(\mathbf{y};B)) &\leq 1/N^2 \sum_{i,j} E_{\mathbf{X}}(D(\mathbb{P}_{\mathbf{M}^i}||\mathbb{P}_{\mathbf{M}^j})) \\
&= \frac{1}{2N^2\sigma^2} \sum_{i,j} E_{\mathbf{X}}(\langle \underset{(i,j)}{\triangle} \mathbf{M}^{(1)}, \mathbf{x}_1^1\rangle^2 + \ldots + \langle \underset{(i,j)}{\triangle} \mathbf{M}^{(1)}, \mathbf{x}_m^1\rangle^2 + \ldots + \langle \underset{(i,j)}{\triangle} \mathbf{M}^{(T)}, \mathbf{x}_m^T\rangle^2) \\
&\leq \frac{1}{2N^2\sigma^2} \sum_{i,j} \Big(m\phi^2 (\sum_{k=1}^{d} \underset{(i,j)}{\triangle} \mathbf{M}_k^{(1)} + \sum_{k \neq l} \underset{(i,j)}{\triangle} \mathbf{M}_k^{(1)} \underset{(i,j)}{\triangle} \mathbf{M}_l^{(1)}) + \ldots \\
&+ m\phi^2 (\sum_{k=1}^{d} \underset{(i,j)}{\triangle} \mathbf{M}_k^{(T)} + \sum_{k \neq l} \underset{(i,j)}{\triangle} \mathbf{M}_k^{(T)} \underset{(i,j)}{\triangle} \mathbf{M}_l^{(T)}) \Big) \\
&\leq \frac{1}{2N^2\sigma^2} \sum_{i,j} (m\phi^2 || \underset{(i,j)}{\triangle} \mathbf{M}||_F^2 + 4m\phi^2 Td(d-1)\aleph^2) \\
&\leq \frac{1}{2N^2\sigma^2} \sum_{i,j} (16m\phi^2\varrho^2 + 4m\phi^2 Td(d-1)\aleph^2) \\
&\leq \frac{1}{2\sigma^2} (16m\phi^2\varrho^2 + 4m\phi^2 Td(d-1)\aleph^2)
\end{aligned} \tag{27}$$

Assume that $r(d+T) > 512 \ln 4$. We obtain

$$\begin{aligned}
\mathbb{P}(\psi(\mathbf{X},\mathbf{y}) \neq B) &= E_{\mathbf{X}}(\mathbb{P}(\psi(\mathbf{X},\mathbf{y}) \neq B|\mathbf{X})) \\
&\geq 1 - \frac{E_{\mathbf{X}}(\mathbf{I}_{\mathbf{X}}(\mathbf{y};B)) + \ln 2}{\ln N} && (25) \\
&\geq 1 - \frac{\frac{1}{2\sigma^2}(16m\phi^2\varrho^2 + 4m\phi^2 Td(d-1)\aleph^2) + \ln 2}{\frac{r(d+T)}{128} - \ln 4} && (27) \text{ and Lemma } 12 \\
&\geq 1 - \frac{\frac{256}{2\sigma^2}(16m\phi^2\varrho^2 + 4m\phi^2 Td(d-1)\aleph^2) + 256 \ln 2}{r(d+T)} && r(d+T) > 512 \ln 4
\end{aligned} \tag{28}$$

Setting $\varrho^2 = \frac{\sigma^2 r(d+T)}{8192 m\phi^2(1+32(d-1)\ln(dT))}$, (28) implies that

$$\mathbb{P}(\psi(\mathbf{X},\mathbf{y}) \neq B) \geq 1 - \frac{\frac{r(d+T)}{4} + 256 \ln 2}{r(d+T)} \geq \frac{1}{2} \tag{29}$$

The claim follows from (24) and (29).

4. Conclusion

In this work, we establish a sharper estimation error bound for MTL with trace norm regularizer. We improve the estimation error bound in Boursier et al. (2022) by a factor $\mathcal{O}(\sqrt{\frac{rd^2/m+rdT/m}{m}})$ and a logarithmic term without three assumptions. Moreover, we establish matching lower bounds on the minimax error, showing that our upper bounds cannot be improved beyond constant factors under some mild conditions.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 61976161, and the Fundamental Research Funds for the Central Universities under Grant 2042022rc0016.

References

Alekh Agarwal, Sahand N. Negahban, and Martin J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. In *ICML*, pages 1129–1136, 2011.

Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research*, 4:83–99, 2003.

Shai Ben-David and Reba Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pages 567–580, 2003.

Etienne Boursier, Mikhail Konobeev, and Nicolas Flammarion. Trace norm regularization for multi-task learning with scarce data. In *COLT*, volume 178, pages 1303–1327, 2022.

Rich Caruana. Multitask learning: A knowledge-based source of inductive bias. In *ICML*, pages 41–48, 1993.

Bin Cheng, Guangcan Liu, Jingdong Wang, ZhongYang Huang, and Shuicheng Yan. Multi-task low-rank affinity pursuit for image segmentation. In *ICCV*, pages 2439–2446, 2011.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, volume 70, pages 1126–1135, 2017.

Zaïd Harchaoui, Matthijs Douze, Mattis Paulin, Miroslav Dudík, and Jérôme Malick. Large-scale image classification with trace-norm regularization. In *CVPR*, pages 3386–3393, 2012.

R. Z. Hasminskii. A lower bound on the risks of non-parametric estimates of densities in the uniform metric. *Theory of Probability & Its Applications*, 23(4):794–798, 1979.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, pages 7482–7491, 2018.

Weiwei Liu and Ivor W. Tsang. Large margin metric learning for multi-label prediction. In *AAAI*, pages 2800–2806, 2015.

Weiwei Liu and Ivor W. Tsang. Making decision trees feasible in ultrahigh feature and label dimensions. *Journal of Machine Learning Research*, 18:81:1–81:36, 2017.

Weiwei Liu, Ivor W. Tsang, and Klaus-Robert Müller. An easy-to-hard learning paradigm for multiple classes and multiple labels. *Journal of Machine Learning Research*, 18:94:1–94:38, 2017.

Weiwei Liu, Donna Xu, Ivor W. Tsang, and Wenjie Zhang. Metric learning for multi-output tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):408–422, 2019.

Yuren Mao, Weiwei Liu, and Xuemin Lin. Adaptive adversarial multi-task representation learning. In *ICML*, volume 119, pages 6724–6733, 2020a.

Yuren Mao, Shuang Yun, Weiwei Liu, and Bo Du. Tchebycheff procedure for multi-task text classification. In *ACL*, pages 4217–4226, 2020b.

Yuren Mao, Zekai Wang, Weiwei Liu, Xuemin Lin, and Wenbin Hu. Banditmtl: Bandit-based multi-task learning for text classification. In *ACL*, pages 5506–5516, 2021.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *NeurIPS*, pages 1348–1356, 2009.

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

Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. Provable meta-learning of linear representations. In *ICML*, volume 139, pages 10434–10443, 2021.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NeurIPS*, pages 3630–3638, 2016.

Martin Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

Liqiang Xiao, Honglun Zhang, and Wenqing Chen. Gated multi-task network for text classification. In *NAACL-HLT*, pages 726–731, 2018.

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

Appendix A. Proof of Lemma 10

The proof of Lemma 10 is based on Sudakov-Fernique Theorem (Vershynin, 2018) and Gordons inequality Lemma 9. Let $\mathbb{S}^{d-1}$ be the Euclidean unit sphere in $\mathbb{R}^d$.

$$\begin{aligned}\sigma_{max}(\mathbf{W}) &= \sup_{\mathbf{v}\in\mathbb{S}^{T-1}} ||\mathbf{W}\mathbf{v}||_2 = \sup_{\mathbf{v}\in\mathbb{S}^{T-1}} \sup_{\mathbf{u}\in\mathbb{S}^{m-1}} tr(\mathbf{u}'\mathbf{W}\mathbf{v}) \\ &= \sup_{\mathbf{v}\in\mathbb{S}^{T-1}} \sup_{\mathbf{u}\in\mathbb{S}^{m-1}} tr(\mathbf{W}\mathbf{v}\mathbf{u}') = \sup_{\mathbf{v}\in\mathbb{S}^{T-1}} \sup_{\mathbf{u}\in\mathbb{S}^{m-1}} \langle \mathbf{W}, \mathbf{u}\mathbf{v}'\rangle\end{aligned}$$

Let $(\mathbf{u}, \mathbf{v})$ and $(\breve{\mathbf{u}}, \breve{\mathbf{v}})$ be any two pairs in $\mathbb{S}^{m-1} \times \mathbb{S}^{T-1}$. The zero-mean Gaussian process is defined as $Z_{\mathbf{u},\mathbf{v}} := \langle \mathbf{W}, \mathbf{u}\mathbf{v}' \rangle$.

$$\begin{aligned} E(Z_{\mathbf{u},\mathbf{v}} - Z_{\breve{\mathbf{u}},\breve{\mathbf{v}}})^2 &= E(\langle \mathbf{W}, \mathbf{u}\mathbf{v}' - \breve{\mathbf{u}}\breve{\mathbf{v}}' \rangle)^2 = ||\mathbf{u}\mathbf{v}' - \breve{\mathbf{u}}\breve{\mathbf{v}}'||_F^2 = ||\mathbf{u}(\mathbf{v} - \breve{\mathbf{v}})' + (\mathbf{u} - \breve{\mathbf{u}})\breve{\mathbf{v}}'||_F^2 \\ &= ||\mathbf{u}(\mathbf{v} - \breve{\mathbf{v}})'||_F^2 + ||(\mathbf{u} - \breve{\mathbf{u}})\breve{\mathbf{v}}'||_F^2 + 2\langle \mathbf{u}(\mathbf{v} - \breve{\mathbf{v}})', (\mathbf{u} - \breve{\mathbf{u}})\breve{\mathbf{v}}' \rangle \\ &\leq ||\mathbf{u}||_2^2 ||\mathbf{v} - \breve{\mathbf{v}}||_2^2 + ||\mathbf{u} - \breve{\mathbf{u}}||_2^2 ||\breve{\mathbf{v}}||_2^2 + 2(||\mathbf{u}||_2^2 - \langle \mathbf{u}, \breve{\mathbf{u}} \rangle)(\langle \mathbf{v}, \breve{\mathbf{v}} \rangle - ||\mathbf{v}||_2^2) \\ &\leq ||\mathbf{v} - \breve{\mathbf{v}}||_2^2 + ||(\mathbf{u} - \breve{\mathbf{u}})||_2^2 \end{aligned}$$

The zero-mean Gaussian processes is defined as $Y_{\mathbf{u},\mathbf{v}} := \langle \mathbf{w}_1, \mathbf{u} \rangle + \langle \mathbf{w}_2, \mathbf{v} \rangle$, where $\mathbf{w}_1 \in \mathbb{R}^m$ and $\mathbf{w}_2 \in \mathbb{R}^T$ are both standard Gaussian random vectors with i.i.d. $\mathcal{N}(0, 1)$ entries. $E(Y_{\mathbf{u},\mathbf{v}} - Y_{\breve{\mathbf{u}},\breve{\mathbf{v}}})^2 = ||\mathbf{u} - \breve{\mathbf{u}}||_2^2 + ||\mathbf{v} - \breve{\mathbf{v}}||_2^2$. The Sudakov-Fernique Theorem implies that

$$\begin{aligned} E\sigma_{max}(\mathbf{W}) &\leq E \sup_{\mathbf{v} \in \mathbb{S}^{T-1}} \sup_{\mathbf{u} \in \mathbb{S}^{m-1}} (\langle \mathbf{w}_1, \mathbf{u} \rangle + \langle \mathbf{w}_2, \mathbf{v} \rangle) \leq E \sup_{\mathbf{v} \in \mathbb{S}^{T-1}} \langle \mathbf{w}_2, \mathbf{v} \rangle + E \sup_{\mathbf{u} \in \mathbb{S}^{m-1}} \langle \mathbf{w}_1, \mathbf{u} \rangle \\ &= E||\mathbf{w}_2||_2 + E||\mathbf{w}_1||_2 \leq \sqrt{m} + \sqrt{T} \end{aligned}$$

where the last step follows from the Jensen's inequality. The lower bound on the expectation of the minimum singular value is based on a similar argument with Gordons inequality Lemma 9.