Learning from Similar Linear Representations: Adaptivity, Minimaxity, and Robustness

Ye Tian   YE.T@COLUMBIA.EDU
Department of Statistics
Columbia University
New York, NY 10027, USA

Yuqi Gu   YUQI.GU@COLUMBIA.EDU
Department of Statistics
Columbia University
New York, NY 10027, USA

Yang Feng   YANG.FENG@NYU.EDU
Department of Biostatistics, School of Global Public Health
New York University
New York, NY 10003, USA

Editor: Ji Zhu

Abstract

Representation multi-task learning (MTL) has achieved tremendous success in practice. However, the theoretical understanding of these methods is still lacking. Most existing theoretical works focus on cases where all tasks share the same representation, and claim that MTL almost always improves performance. Nevertheless, as the number of tasks grows, assuming all tasks share the same representation is unrealistic. Furthermore, empirical findings often indicate that a shared representation does not necessarily improve single-task learning performance. In this paper, we aim to understand how to learn from tasks with *similar but not exactly the same* linear representations, while dealing with outlier tasks. Assuming a known intrinsic dimension, we propose a penalized empirical risk minimization method and a spectral method that are *adaptive* to the similarity structure and *robust* to outlier tasks. Both algorithms outperform single-task learning when representations across tasks are sufficiently similar and the proportion of outlier tasks is small. Moreover, they always perform at least as well as single-task learning, even when the representations are dissimilar. We provide information-theoretic lower bounds to demonstrate that both methods are nearly *minimax* optimal in a large regime, with the spectral method being optimal in the absence of outlier tasks. Additionally, we introduce a thresholding algorithm to adapt to an unknown intrinsic dimension. We conduct extensive numerical experiments to validate our theoretical findings.

1. Introduction

1.1 Representation Multi-task Learning

With the increased computational power, machine learning systems can now process datasets on a large scale. However, for each machine learning task, we may not have access to a large

amount of data due to data privacy restrictions and the high cost of data acquisition. This motivated the idea of multi-task learning (MTL), where we jointly learn many tasks that are similar but not identical to enhance model performance (Zhang and Yang, 2018, 2021). Related concepts include transfer learning (TL), learning-to-learn, and meta-learning, where model structures learned from multiple tasks can be transferred to new incoming tasks to improve their performance (Weiss et al., 2016; Hospedales et al., 2021). Among numerous multi-task and transfer learning approaches, representation learning has been one of the most popular and successful methods over the past few years, where a data representation is jointly learned from multiple similar data sets and can be shared across them (Rostami et al., 2022). A successful example of multi-task and transfer representation learning is learning the weights of a few initial layers of neural networks from ImageNet pre-training, then retraining final layers on new image classification tasks (Donahue et al., 2014; Goyal et al., 2019). Other applications include multilingual knowledge graph completion (Chen et al., 2020) and reinforcement learning (Gupta et al., 2017).

While representation learning has been successful in practice, its theoretical understanding in the context of multi-task and transfer learning remains limited. Most existing theoretical works assume that the same representation is shared across all tasks, which is not always realistic in scenarios with a large number of tasks (Rostami et al., 2022). Furthermore, empirical studies have shown that freezing a representation across tasks from different contexts may not improve model performance and can even be harmful. For example, Raghu et al. (2019) found that pre-training on ImageNet offered little help to target medical tasks, and Wang et al. (2019) found that different target tasks might benefit from different pre-training in natural language understanding. These studies suggested that a frozen representation may not always work well. Additionally, there may be outlier tasks that are dissimilar to other tasks (Zhang and Yang, 2021) or may be contaminated with adversarial attacks on the data (Qiao, 2018; Qiao and Valiant, 2018; Konstantinov et al., 2020). If left unaddressed, such issues could severely impact the machine learning system's overall performance.

This paper investigates the effective learning of tasks with *similar representations* in the presence of potential outlier tasks or adversarial attacks. Specifically, we consider the following linear model with linear representations. Suppose there are T tasks in total, and we have collected a sample $\{x_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ from the t-th task, where $x_i^{(t)} \in \mathbb{R}^p$, $y_i^{(t)} \in \mathbb{R}$, and $t \in [T] = \{1, 2, \ldots, T\}$. There exists an unknown subset $S \subseteq [T]$, such that for all $t \in S$,

$$y_i^{(t)} = (x_i^{(t)})^\top \beta^{(t)*} + \epsilon_i^{(t)}, \quad i = 1 : n, \tag{1}$$

where the regression coefficient $\beta^{(t)*} = A^{(t)*}\theta^{(t)*}$, the representation $A^{(t)*} \in \mathcal{O}^{p \times r} = \{A \in \mathbb{R}^{p \times r} : A^\top A = I_r\}$, low-dimensional parameter $\theta^{(t)*} \in \mathbb{R}^r$, $r \leq p$, and $\{\epsilon_i^{(t)}\}_{i=1}^n$ are random noises. Here r represents the intrinsic dimension of the problem, which is usually much smaller than p. The data $\{x_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ for $t \notin S$ can be arbitrarily distributed in the worst case, and the corresponding tasks in $S^c = [T] \backslash S$ are outlier or contaminated tasks. We call $\epsilon := |S^c|/T$ the *contamination proportion* or the *proportion of outlier tasks*. To ensure effective learning from similar representations, we assume that $\{A^{(t)*}\}_{t \in S}$ are similar to each other, in the sense that $\min_{\overline{A} \in \mathcal{O}^{p \times r}} \max_{t \in S} \|A^{(t)*}(A^{(t)*})^\top - \overline{A}(\overline{A})^\top\|_2 \leq h$, where $\overline{A} \in \mathcal{O}^{p \times r}$ achieving the minimum can be understood as a "central representation" and

h is the similarity measure. Our goal is to explore the upper and lower error bounds in estimating $\{\boldsymbol{\beta}^{(t)*}\}_{t\in S}$ for all possible cases of S under certain conditions. Furthermore, when the tasks in S^c also satisfy the linear model (1), we aim to ensure the effective estimation of $\{\boldsymbol{\beta}^{(t)*}\}_{t\in S^c}$ as well.

It is worth pointing out that we allow the scales of $\{\boldsymbol{\theta}^{(t)*}\}_{t\in S}$, i.e. $\{\|\boldsymbol{\theta}^{(t)*}\|_2\}_{t\in S}$, to differ across tasks in S, and $\|\boldsymbol{\theta}^{(t)*}\|_2$ can also diverge as $n\to\infty$. Here, $\|\boldsymbol{\theta}^{(t)*}\|_2 = \|\boldsymbol{\beta}^{(t)*}\|_2$ can be viewed as the signal strength of the t-th task. It turns out that the performance of representation MTL on each task is highly relevant to $\|\boldsymbol{\theta}^{(t)*}\|_2$. In contrast, existing literature generally assumes $\|\boldsymbol{\theta}^{(t)*}\|_2 \lesssim 1$ for all $t \in S = [T]$ and ignores the impact of $\|\boldsymbol{\theta}^{(t)*}\|_2$ on the model performance.

1.2 Related Works

1.2.1 MULTI-TASK AND TRANSFER LEARNING

Representation MTL and TL: Baxter (2000) is among the earliest works to study the theory of representation MTL under general function classes, where all tasks are generated from the same distribution. Maurer et al. (2016) improved their results by using the analysis based on Rademacher complexity. Ando et al. (2005) explored the case of semi-supervised learning. More recently, Du et al. (2020) and Tripuraneni et al. (2021) studied linear model (1) with $S = [T]$ and $h = 0$, i.e., under the assumption that there are no outlier tasks and all tasks share the same representation. They proposed the so-called task diversity condition, under which the learning rate can be significantly improved by a non-convex empirical risk minimization (ERM) algorithm. Tripuraneni et al. (2020) extended the analysis to general non-linear models and provided general results. Thekumparampil et al. (2021) proposed a polynomial-time alternating gradient descent algorithm that achieves similar performance as ERM but avoids solving the non-convex optimization directly. Meunier et al. (2023) characterizes the shared representation via a mapping into a finite-dimensional subspace of a reproducing kernel Hilbert space (RKHS). Other related works include federated representation learning (Collins et al., 2021; Duchi et al., 2022), tensor representation meta-learning (Deng et al., 2022), conditional meta-learning (Denevi et al., 2020), and matrix completion via representation MTL (Zhou et al., 2021). Note that MTL under the assumption that $\boldsymbol{\beta}^{(t)*}$'s in (1) share the same or similar support sets (Lounici et al., 2009, 2011; Jalali et al., 2010; Li et al., 2021; Xu and Bastani, 2021) can also be viewed as a special case of the general representation MTL.

Distance-based MTL and TL: There has been much literature in the statistics community studying model (1) under the assumption that Euclidean distance or ℓ_1-distance between $\boldsymbol{\beta}^{(t)*}$'s are small (Bastani, 2021; Li et al., 2022b; Duan and Wang, 2023; Gu et al., 2023), which is called "distance-based" MTL and TL in Gu et al. (2024). Some extensions include high-dimensional GLMs (Tian and Feng, 2022), graphical models (Li et al., 2022a), functional regression (Lin and Reimherr, 2022), semi-supervised classification (Zhou et al., 2022), and unsupervised mixture models (Tian et al., 2022, 2024). Recently, Gu et al. (2024) proposed the "angle-based" TL where they assume the angle between every pair of $\boldsymbol{\beta}^{(t)*}$'s is small. As we will discuss in the next section, their setting is a special case of (1) when $r = 1$.

Other related literature: Other relevant literature includes the non-parametric TL (Cai and Wei, 2021; Kpotufe and Martinet, 2021), the hardness of MTL (Hanneke and

Kpotufe, 2019, 2022), adversarial robustness of MTL or distributed learning (Chen et al., 2017; Alistarh et al., 2018; Yin et al., 2018; Qiao, 2018; Qiao and Valiant, 2018; Konstantinov et al., 2020; Zhu et al., 2023; Guerraoui et al., 2024), gradient-based meta-learning (Finn et al., 2017; Nichol et al., 2018; Finn et al., 2019), and theory of MTL based on distributional measure (Ben-David and Borbely, 2008; Ben-David et al., 2010).

To help readers better understand the difference between some settings in literature with our setting under the linear model (1), we drew Figure 1 as a simple visualization corresponding to the case where $p = 3$ and $r = 2$.
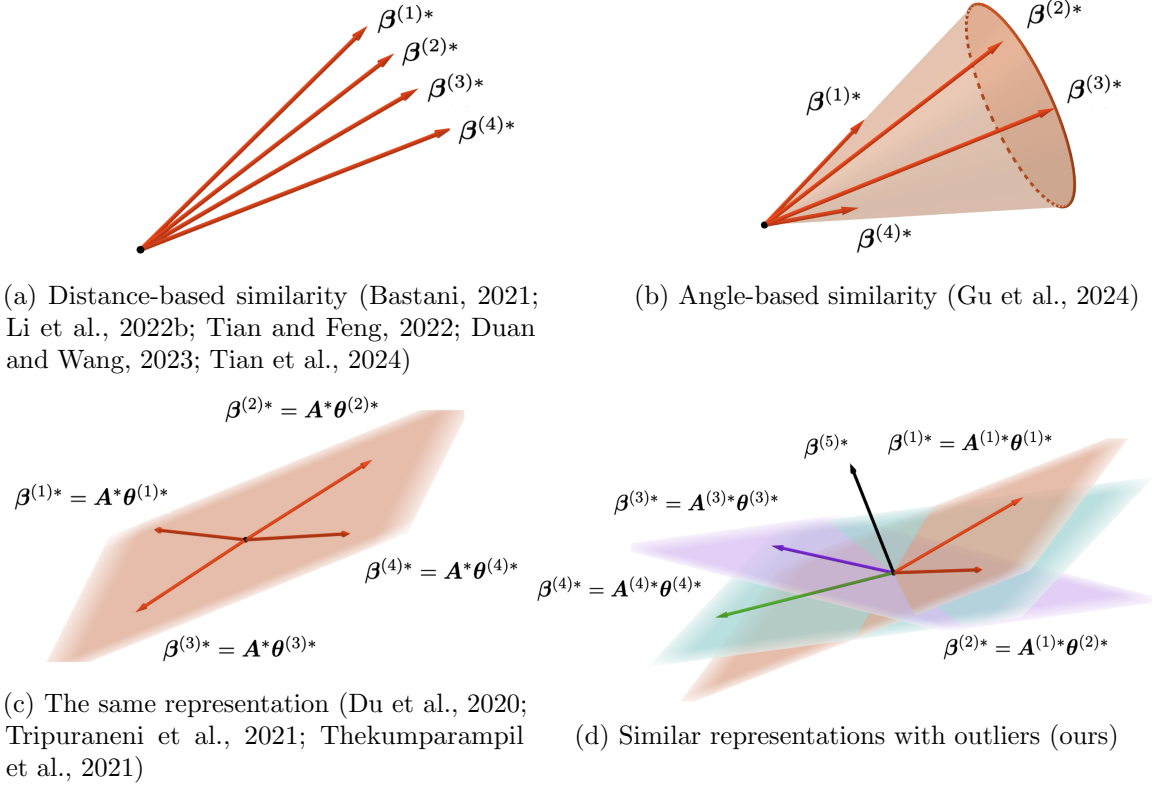

(a) Distance-based similarity (Bastani, 2021; Li et al., 2022b; Tian and Feng, 2022; Duan and Wang, 2023; Tian et al., 2024)

(b) Angle-based similarity (Gu et al., 2024)

(c) The same representation (Du et al., 2020; Tripuraneni et al., 2021; Thekumparampil et al., 2021)

(d) Similar representations with outliers (ours)

Figure 1: A simple visualization of four different settings under the linear model (1).

1.2.2 BEYOND THE ASSUMPTION OF THE SAME REPRESENTATION

Several works have studied similar problems to the current work, but in different formulations. Chua et al. (2021) explored linear model (1) but with the assumption that $\boldsymbol{A}^{(t)*} = \overline{\boldsymbol{A}} + \boldsymbol{\Delta}^{(t)*}$ with some $\overline{\boldsymbol{A}} \in \mathbb{R}^{p \times r}$ and $\|\boldsymbol{\Delta}^{(t)*}\|_{\mathrm{F}} \leq h''$ and $S = [T]$. Duan and Wang (2023) considered the same model with $\boldsymbol{\beta}^{(t)*} = \overline{\boldsymbol{A}}\bar{\boldsymbol{\theta}}^{(t)*} + \boldsymbol{\delta}^{(t)*}$ with $\|\boldsymbol{\delta}^{(t)*}\|_2 \leq h'$ and $S = [T]$. The following theorem establishes the equivalence between these alternative formulations and our proposed setting.

Theorem 1 *Consider the following three settings:* [a]

---

a. $\overline{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$ in the three settings are the same.

1. $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$, where $\|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \le h$ for some $\overline{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$;

2. $\boldsymbol{\beta}^{(t)*} = \overline{\boldsymbol{A}}\bar{\boldsymbol{\theta}}^{(t)*} + \boldsymbol{\delta}^{(t)*}$, with $\|\boldsymbol{\delta}^{(t)*}\|_2 \le h'$, $(\overline{\boldsymbol{A}})^\top \boldsymbol{\delta}^{(t)*} = \boldsymbol{0}$;

3. $\boldsymbol{\beta}^{(t)*} = (\overline{\boldsymbol{A}} + \boldsymbol{\Delta}^{(t)*})\tilde{\boldsymbol{\theta}}^{(t)*}$, with $\|\boldsymbol{\Delta}^{(t)*}\|_{\mathrm{F}} \le h''$.

They are equivalent in the following sense:

(i) *If Setting 1 holds, then there exist* $\bar{\boldsymbol{\theta}}^{(t)*} \in \mathbb{R}^r$, $\boldsymbol{\delta}^{(t)*} \in \mathbb{R}^p$ *with* $\|\boldsymbol{\delta}^{(t)*}\|_2 \le h\|\boldsymbol{\theta}^{(t)*}\|_2$ *satisfying Setting 2, and there exist* $\boldsymbol{\Delta}^{(t)*} \in \mathbb{R}^{p \times r}$, $\tilde{\boldsymbol{\theta}}^{(t)*} \in \mathbb{R}^p$ *with* $\|\boldsymbol{\Delta}^{(t)*}\|_{\mathrm{F}} \le \frac{h}{\sqrt{1-h^2}}$ *satisfying Setting 3.*

(ii) *If Setting 2 holds, then there exists* $\boldsymbol{A}^{(t)*} \in \mathcal{O}^{p \times r}$ *satisfying Setting 1 with* $\|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \le (10\sqrt{2} + 4)\frac{h'}{\|\boldsymbol{\beta}^{(t)*}\|_2}$;

(iii) *If Setting 3 holds, then there exists* $\boldsymbol{A}^{(t)*} \in \mathcal{O}^{p \times r}$ *and* $\boldsymbol{\theta}^{(t)*} \in \mathbb{R}^r$ *satisfying Setting 1 with* $\|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \le (10\sqrt{2} + 4)\frac{h''}{1-h''}$ *when* $h'' \in [0, 1)$.

It is worth emphasizing that, despite the equivalence of the three settings, both Chua et al. (2021) and Duan and Wang (2023) require $S = [T]$ and impose the constraints $\|\bar{\boldsymbol{\theta}}^{(t)*}\|_2 \lesssim 1$ and $\|\tilde{\boldsymbol{\theta}}^{(t)*}\|_2 \lesssim 1$ for all $t \in [T]$, respectively. In contrast, our setting allows for $S \ne [T]$ and for $\|\boldsymbol{\theta}^{(t)*}\|_2$ to diverge. Moreover, when $S = [T]$ and $h' = h'' = 0$, all three settings reduce to the shared representation case studied in Du et al. (2020); Thekumparampil et al. (2021); Tripuraneni et al. (2021). [b]

Thus, our work addresses a more general scenario compared to these existing formulations. Despite the equivalence established in Theorem 1, we focus on the setting defined in Section 1.1, as this formulation naturally facilitates the development of algorithms that adapt to the unknown similarity between tasks and enjoy robustness against adversarial contamination. More concretely, compared to the settings in Chua et al. (2021) and Duan and Wang (2023), our formulation leads to simpler learning algorithms, which can adapt to the unknown similarity level h (unlike Chua et al., 2021, where the algorithm requires tuning parameters depending on the similarity level h), with stronger theoretical results. Our formulation also inspires algorithms that remain robust to a small fraction of outlier tasks, a merit not shared by the methods following the formulations in Chua et al. (2021) and Duan and Wang (2023).

Moreover, our framework is easier to generalize to an unsupervised learning setting (e.g., multiple linear/nonlinear latent factor models with similar factor loading matrices or similar factor score matrices). Additionally, a representation-based reweighting strategy was proposed in Chen et al. (2021a), which is motivated by the concern of assuming the same representation. Their similarity metric between tasks depends on the weight assigned to the objective function of each task, while our similarity metric depends on the difference between representations explicitly, which is more intuitive. Moreover, their approach can suffer from a negative transfer in the worst case, while our approaches do not. Furthermore, none of Chen et al. (2021a); Chua et al. (2021); Duan and Wang (2023) considered the presence of outlier tasks.

---

b. Du et al. (2020) also studied the non-linear representations; here we are referring to their setting of linear representations.

Finally, we present a diagram in Figure 2 to summarize the relationship between the different regimes studied in various papers mentioned earlier. To compare algorithms across different settings more clearly, we focus on a specific regime, enclosed by a dashed line in Figure 2. This regime allows for simpler, more intuitive, and explicit results, making it easier to compare the estimation errors of $\boldsymbol{\beta}^{(t)*}$ across different approaches. In this regime, defined for rate comparison, we have our setting introduced in Section 1.1 with $\min_{\overline{\boldsymbol{A}} \in \mathcal{O}^{p \times r}} \max_{t \in S} \|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \le h$ and $\max_{t \in S} \|\boldsymbol{\theta}^{(t)*}\|_2 \lesssim 1$, and $\sqrt{|S|^{-1} \sum_{t \in S} \|\boldsymbol{\theta}^{(t)*}\|_2^2} \gtrsim 1$. We summarize the estimation errors of different approaches in Table 1, highlighting our two proposed algorithms: penalized ERM ("pERM") and the spectral method ("Spectral"). We show that the spectral method is minimax optimal with computational efficiency when there is no contamination ($\epsilon = 0$), and pERM can handle the contaminated case more effectively.
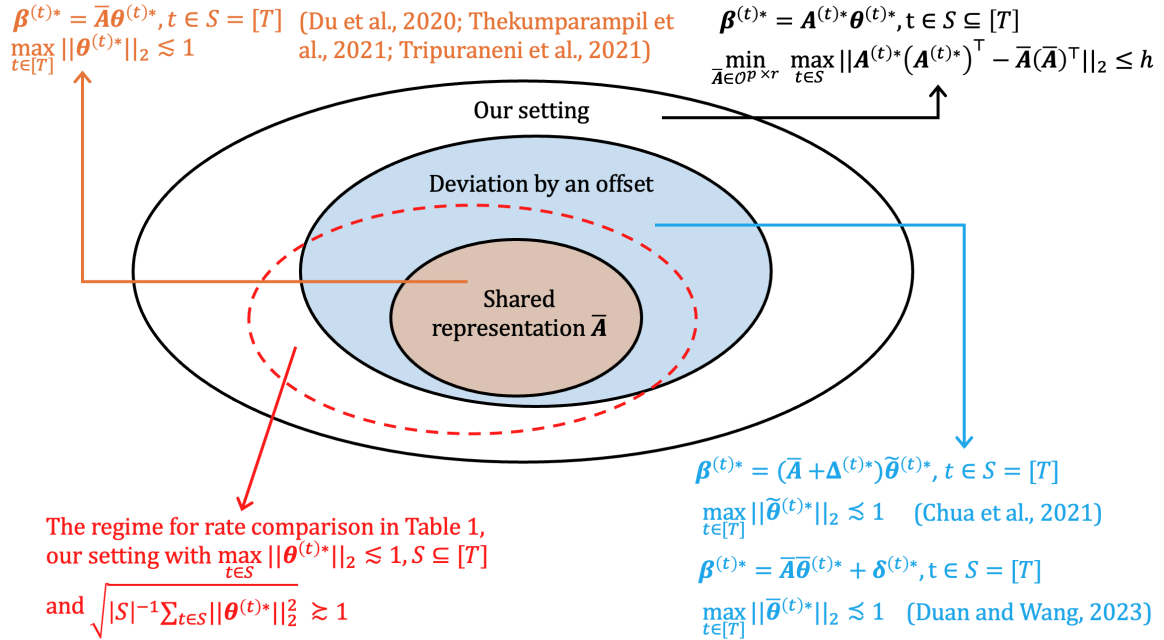
Figure 2: The diagram illustrating the relationship between different regimes studied in different papers.

1.3 Our Contributions

Our contributions can be summarized below.

(i) Compared to most literature on representation MTL, we considered a more general framework. Here, the linear representations can vary across tasks, signal strengths may differ between tasks, and there can be a small fraction of unknown outlier tasks.

(ii) We proposed two algorithms, the penalized ERM and the spectral method, to learn the regression coefficients and the representations from multiple tasks. Our algorithms were shown to have the following properties:

Regime	Algorithm	$\max_{t\in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$	Optimal ($\epsilon = 0$)?	Poly-time?
$h = \epsilon = 0$	ERM	$r\sqrt{\frac{p}{nT}} + r\sqrt{\frac{1}{n}}$	No	No
	MoM	$r\sqrt{\frac{p}{nT}} + \sqrt{\frac{\tau}{n}}$	No	Yes
	AltMinGD	$r\sqrt{\frac{p}{nT}} + \sqrt{\frac{\tau}{n}}$	No	Yes
$h \neq 0$	AdaptRep	$\left[r\sqrt{\frac{p}{nT}} + r\sqrt{\frac{1}{n}} + \sqrt{rh}(\frac{p}{n})^{1/4}\right] \wedge \sqrt{\frac{p}{n}}$	No	No
$\epsilon = 0$	ARMUL	$\left(r\sqrt{\frac{p}{nT}} + r\sqrt{\frac{1}{n}} + rh\right) \wedge \left(r\sqrt{\frac{p}{n}}\right)$	No	No
	pERM	$\left(r\sqrt{\frac{p}{nT}} + r\sqrt{\frac{1}{n}} + \sqrt{r}h + \epsilon\frac{r^{3/2}\sqrt{p}}{\sqrt{n}}\right) \wedge \sqrt{\frac{p}{n}}$	No	No
$h, \epsilon \neq 0$	Spectral	$\left(\sqrt{\frac{pr}{nT}} + \sqrt{\frac{\tau}{n}} + h + \sqrt{\epsilon r}\right) \wedge \sqrt{\frac{p}{n}}$	Yes	Yes
	Single-task	$\sqrt{\frac{p}{n}}$	No	Yes
Lower-bound		$\left(\sqrt{\frac{pr}{nT}} + \sqrt{\frac{\tau}{n}} + h + \epsilon\frac{r}{\sqrt{n}}\right) \wedge \sqrt{\frac{p}{n}}$	–	–

Table 1: A summary of the high-probability estimation error of $\boldsymbol{\beta}^{(t)*}$'s for different methods in different regimes, where the two algorithms we proposed, penalized ERM ("pERM", Algorithm 1) and the spectral method ("Spectral", Algorithm 2), are highlighted. Denote the contamination proportion $\epsilon = |S^c|/T$. Except for the lower bound, all the other rates are upper bounds. All the rates are up to logarithmic factors. "ERM": Du et al. (2020); Tripuraneni et al. (2021). "MoM": method-of-moments, Tripuraneni et al. (2021). "AltMinGD": alternating minimization gradient descent, Thekumparampil et al. (2021). "AdaptRep": adaptive representation learning, Chua et al. (2021). "ARMUL": adaptive and robust multi-task learning, Duan and Wang (2023). "Single-task": single-task regression by only using the local data from each task.

- They outperform single-task learning when the representations of different tasks are sufficiently similar, and the proportion of outlier tasks is low.

- They guarantee no worse performance than single-task learning (safe-net guarantee), even when task representations are dissimilar.

(iii) We thoroughly analyzed the relationship between different regimes of representation MTL studied in the literature. Our derived upper bounds improve over existing rates, particularly in scenarios without task contamination. Furthermore, in the context of representation MTL, we are the first to examine the scenario where a small proportion of the tasks is contaminated.

(iv) We derived the lower bounds for model (1). To our knowledge, these are the first lower bound results for regression coefficient estimation under the representation MTL. Prior works such as Duchi et al. (2022) and Tripuraneni et al. (2021) provided analogous lower bounds for the subspace recovery problem, assuming identical representations without outlier tasks. In a TL setup, Chua et al. (2021) showed that assuming the same representation can lead to worse performance than target-only learning when source representations differ from each other, but they did not provide a full lower bound that relates to the representation difference. Comparing the upper and lower

bounds, we demonstrated that both proposed algorithms are nearly minimax optimal, with the spectral method being optimal in uncontaminated settings.

(v) We extended our analysis from linear model (1) to generalized linear models (GLMs) and non-linear regression models, and obtained similar theoretical guarantees in these settings.

(vi) We proposed a thresholding algorithm based on singular value decomposition (SVD) to estimate the unknown intrinsic dimension r. This adaptation enables our penalized ERM and spectral methods to handle cases where r is unknown, addressing a common challenge where r is not a priori known in most prior works.

1.4 Notations and Organization

Throughout the paper, we use bold capitalized and lower-case letters to denote matrices and vectors, respectively. For a real number a, $|a|$ stands for its absolute value. For a vector $\boldsymbol{u}$, $\|\boldsymbol{u}\|_2$ stands for its Euclidean norm. For a matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_2$ and $\|\boldsymbol{A}\|_\mathrm{F}$ represent its spectral and Frobenius norm, respectively. $\boldsymbol{A}^\top$ denotes its transpose. $\sigma_j(\boldsymbol{A})$, $\sigma_{\max}(\boldsymbol{A})$, $\sigma_{\min}(\boldsymbol{A})$ are its j-th largest singular value, maximum singular value, and minimum (non-zero) singular value, respectively. When $\boldsymbol{A}$ is a square matrix, we denote its maximum and minimum eigenvalues as $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$, respectively. For a function $\psi : \mathcal{X} \to \mathbb{R}$, $\|\psi\|_\infty$ is defined to be $\max_{x \in \mathcal{X}} |\psi(x)|$. For two real numbers a and b, we denote their minimum by $\min\{a, b\}$ or $a \wedge b$ and their maximum by $\max\{a, b\}$ or $a \vee b$, respectively. For two positive real sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ means there exists a universal constant $C > 0$ such that $a_n \leq C b_n$ for all n, and $a_n \ll b_n$ or $b_n \gg a_n$ means that $a_n/b_n \to 0$ as $n \to \infty$. $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold simultaneously. Sometimes, we abbreviate "with probability" as "w.p." and "with respect to" as "w.r.t.". For any $N \in \mathbb{N}_+$, $[N]$ and $1 : N$ are defined to be $\{1, \ldots, N\}$. $\mathbb{P}$ and $\mathbb{E}$ are the probability measure and expectation taken over all randomness. We use $C, C', C'', C_1, C_2, C_3, c, c'$ to represent universal constants that could change from place to place.

The rest of this paper is organized as follows. In Section 2, we first propose a penalized ERM algorithm for the linear model (1), establish the upper bound of estimation error for global minimizers of the ERM, study properties of the local minimizers, and discuss implementation details. Subsequently, we introduce a novel spectral method that is computationally more efficient and achieves sharper estimation error upper bounds than the penalized ERM when there is no contamination. Next, we present lower bound results for the representation MTL problem and conclude Section 2 with a brief discussion on extensions to generalized linear models (GLMs) and non-linear regression models. In Section 3, we propose a thresholding algorithm to estimate the intrinsic dimension r and adapt our penalized ERM algorithm and the spectral method to the case where r is unknown. We conduct extensive simulation studies and analyze a real-world dataset to demonstrate our theoretical findings in Section 4. Finally, we summarize our contributions and outline a few potential avenues for future research in Section 5.

Due to space constraints, certain results are deferred to the appendix. In Section A of the appendix, we provide details on the extension of our methods and theory to GLMs and non-linear regression models. In Section B of the appendix, we study how to transfer

the knowledge to an unknown task, i.e., under the setting of transfer learning (TL) or learning-to-learn. We propose an algorithm that leverages outputs from MTL algorithms to adapt to a new target task, presenting corresponding upper and lower bounds. All the proofs are also provided in the appendix.

2. Multi-task Learning with Similar Representations

2.1 Problem Set-up

Let us describe the problem setting introduced in Section 1.1 in more detail. Suppose there are T tasks, and we have collected sample $\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ from the t-th task, where $\boldsymbol{x}_i^{(t)} \in \mathbb{R}^p$, $y_i^{(t)} \in \mathbb{R}$, and $t \in [T]$. There exists an unknown subset $S \subseteq [T]$, such that for all $t \in S$,

$$y_i^{(t)} = (\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*} + \epsilon_i^{(t)}, \quad i = 1:n, \tag{2}$$

where $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$, $\boldsymbol{A}^{(t)*} \in \mathcal{O}^{p \times r} = \{\boldsymbol{A} \in \mathbb{R}^{p \times r} : \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_r\}$, low-dimensional parameter $\boldsymbol{\theta}^{(t)*} \in \mathbb{R}^r$, $r \leq p$, and $\{\epsilon_i^{(t)}\}_{i=1}^n$ are i.i.d. zero-mean sub-Gaussian variables independent of $\{\boldsymbol{x}_i^{(t)}\}_{i=1}^n$ [c]. Throughout this and the next sections, we assume the intrinsic dimension r is known. The case that r is unknown will be addressed in Section 3.

Here the T tasks are divided into two groups, S and S^c. The tasks in S have "similar" representations (similarity to be defined in the following), while the tasks in S^c can be understood as *outlier* tasks or *contaminated* tasks with *an arbitrary distribution*. Our goal is twofold:

1. Improve the learning performance simultaneously on the tasks in S, when they share "similar" representations and the proportion of outlier tasks in S^c among all T tasks is small;

2. Maintain the single-task learning performance when the "similarity" between tasks in S is low.

It should be emphasized that if we allow the outlier tasks in S^c to be arbitrarily distributed, no guarantee can be obtained for these tasks in the worst case. However, as we will discuss later, if these tasks still follow linear models (2) (without a low-dimensional representation), then the single-task linear regression estimation rate can be achieved for $\{\boldsymbol{\beta}^{(t)*}\}_{t \in S^c}$.

We also want to point out that the set S is unknown. We will show that our penalized ERM algorithm and spectral method can perform well across all potential sets $S \subseteq [T]$ under certain conditions. This flexibility is crucial from both perspectives of outlier tasks and adversarial attacks. From the perspective of outlier tasks, we expect an algorithm to succeed for all possible outlier task index sets S^c as long as $|S^c|/T$ is small. In other words, the algorithm should not only work for a specific S^c, but also not rely on task indices (otherwise, we can always drop the data from tasks in S^c to avoid the impact of outliers). From the perspective of adversarial attacks, the attacker can choose to corrupt the data from any

---

c. This is assumed for simplicity and can be relaxed. In fact, it suffices to require $\{\epsilon_i^{(t)} | \boldsymbol{x}_i^{(t)} = \boldsymbol{x}_i\}_{i=1}^n$ to be independent zero-mean sub-Gaussian variables for almost surely $\{\boldsymbol{x}_i\}_{i=1}^n$ w.r.t. the product probability measure induced by the distribution of $\boldsymbol{x}_i^{(t)}$'s.

task, which usually happens after the release of the machine learning system. Therefore, a robust learner should achieve ideal performance for all possible sets S. See Figure 3 for an illustration of these two points of view.
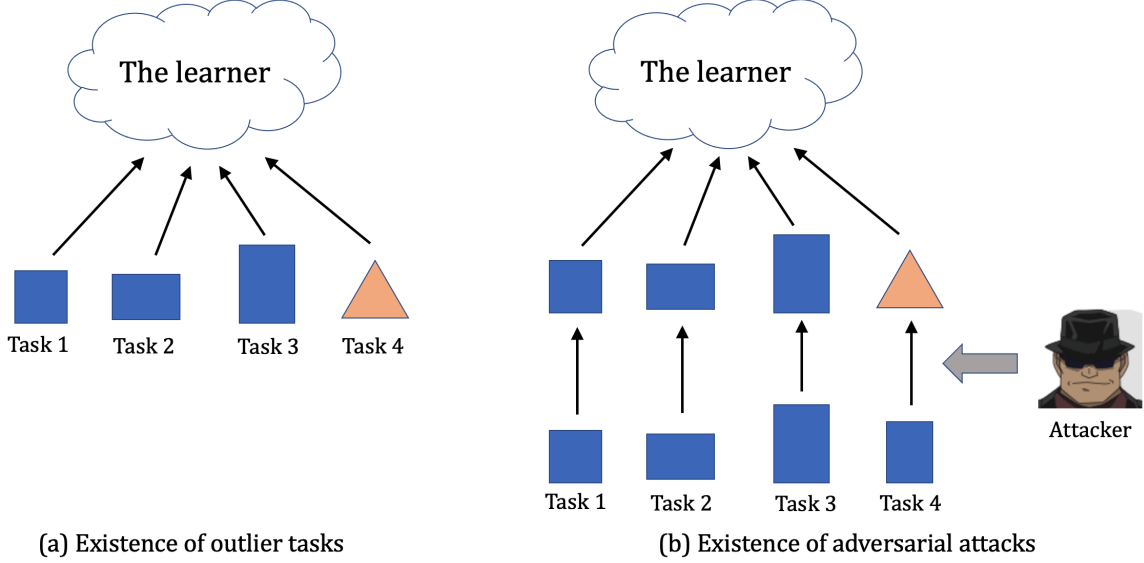

(a) Existence of outlier tasks

(b) Existence of adversarial attacks

Figure 3: An illustration for two perspectives of viewing model (2). In both cases, $S^c = \{4\}$.

The same setting when $\boldsymbol{A}^{(t)*}$'s are the same and $S = [T]$ (i.e. no outlier tasks) has been studied in Du et al. (2020), where they argued that when r is much smaller than p, a better estimation error rate of $\boldsymbol{\beta}^{(t)*}$ can be achieved compared to the single-task learning. Our framework is more general and realistic because it is difficult for all tasks to be embedded in precisely the same subspace as the number of tasks T grows (Rostami et al., 2022), and the prevalence of outlier tasks is common (Zhang and Yang, 2018).

To mathematically quantify the similarity between representations $\{\boldsymbol{A}^{(t)*}\}_{t \in S}$, we consider the maximum principal angle between subspaces spanned by the columns of these representation matrices. More specifically, we assume that there exists $h \in [0, 1]$ such that

$$\min_{\overline{\boldsymbol{A}} \in \mathcal{O}^{p \times r}} \max_{t \in S} \|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \le h. \quad \text{d}$$

A small h means the representations are more similar. The case when $h = 0$ reduces to the setting of the same representations in literature (Du et al., 2020; Tripuraneni et al., 2021). In the literature, $\|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2$ is often referred to as the *maximum principal angle* between column spaces of $\boldsymbol{A}^{(t)*}$ and $\overline{\boldsymbol{A}}$. This concept and its variations have been widely used to measure the difference between subspaces in perturbation theory (e.g., Wedin, 1972; Cai et al., 2013; Kato, 2013; Yu et al., 2015; Chen et al., 2021b). The case when $r = 1$ (where all representations are $p \times 1$ vectors and $\boldsymbol{\theta}^{(t)*}$'s are scalars) reduces to the setting

---

d. Note that the LHS of the following inequality is always less than or equal to 1, because $\boldsymbol{A}^{(t)*}$ and $\overline{\boldsymbol{A}}$ are orthonormal matrices. Here, the minimum is achievable because of the Weierstrass Theorem. Note that the objective function is continuous (over $\overline{\boldsymbol{A}}$) under $\|\cdot\|_2$ norm and the space $(\mathcal{O}^{p \times r}, \|\cdot\|_2)$ is compact.

of Gu et al. (2024). In this case, the principal angle between subspaces becomes the angle between regression coefficient vectors.

We now make some assumptions. Without loss of generality, suppose $\boldsymbol{x}^{(t)}$ is mean-zero. Denote the covariance matrix $\boldsymbol{\Sigma}^{(t)} = \mathbb{E}[\boldsymbol{x}^{(t)}(\boldsymbol{x}^{(t)})^{\top}]$ and the joint distribution of $\{\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n\}_{t\in S^c}$ as $\mathbb{Q}_{S^c}$. For the convenience of description, define a coefficient matrix $\boldsymbol{B}_S^* \in \mathbb{R}^{p\times|S|}$, each column of which is a coefficient vector in $\{\boldsymbol{\beta}^{(t)*}\}_{t\in S}$. Denote the ℓ_2-norm $\zeta^{(t)} = \|\boldsymbol{\theta}^{(t)*}\|_2 = \|\boldsymbol{\beta}^{(t)*}\|_2$, the average ℓ_2-norm $\bar{\zeta} = \sqrt{|S|^{-1}\sum_{t\in S}(\zeta^{(t)})^2}$, and assume $\min_{t\in S}\zeta^{(t)} \gtrsim \sqrt{\frac{p+\log T}{n}}$. [e] $\zeta^{(t)}$ can be viewed as the signal strength of the t-th task. In almost all the existing literature, it is assumed that $\zeta^{(t)} \lesssim 1$ for all $t \in S = [T]$ and $\bar{\zeta} \gtrsim 1$, and the impact of $\zeta^{(t)}$ on the model performance is ignored. In this work, however, we allow $\zeta^{(t)}$'s to vary across tasks, and we will show later how the performance of representation MTL on each task depends on $\zeta^{(t)}$.

Assumption 1 *For any $t \in S$, $\boldsymbol{x}^{(t)}$ is sub-Gaussian in the sense that for any $\boldsymbol{u} \in \mathbb{R}^p$ and $\lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda\boldsymbol{u}^{\top}\boldsymbol{x}^{(t)}}] \leq e^{C\lambda^2\|\boldsymbol{u}\|_2^2}$ with some constant $C > 0$. And there exist constants c, C such that $0 < c \leq \lambda_{\min}(\boldsymbol{\Sigma}^{(t)}) \leq \lambda_{\max}(\boldsymbol{\Sigma}^{(t)}) \leq C < \infty$, for all $t \in S$.*

Assumption 2 *There exists a constant $c > 0$ such that $\sigma_r(\boldsymbol{B}_S^*/\sqrt{|S|}) \geq \frac{c}{\sqrt{r}}\bar{\zeta}$, where each column of $\boldsymbol{B}_S^*$ is a coefficient vector in $\{\boldsymbol{\beta}^{(t)*}\}_{t\in S}$.*

Assumption 3 *$n \geq C(p + \log T)$ with a sufficiently large constant $C > 0$.*

Assumptions 1 and 3 are standard conditions in literature (Du et al., 2020; Duan and Wang, 2023). Assumption 2 is often called the task diversity condition. When $\boldsymbol{A}^{(t)*}$'s are the same, $\sigma_r(\boldsymbol{B}_S^*/\sqrt{|S|}) = \sigma_r(\boldsymbol{\Theta}_S^*/\sqrt{|S|})$, where each column of $\boldsymbol{\Theta}_S^*$ is a coefficient vector in $\{\boldsymbol{\theta}^{(t)*}\}_{t\in S}$, which means that the low-dimensional task-specific parameters $\boldsymbol{\theta}^{(t)*}$'s are diverse. Such a task diversity condition has been adopted in other related studies (Du et al., 2020; Chua et al., 2021; Tripuraneni et al., 2021; Duchi et al., 2022) [f], to obtain a parametric rate which is faster than the rate without this condition (Maurer et al., 2016). The benefit of this condition is intuitive, because a full exploration of all directions in the subspace is necessary to learn representations well, which is the key to representation MTL.

It is important to note that the presented assumptions are imposed on both the set of non-outlier tasks S and the model parameters. These assumptions are made from the perspective of outlier tasks. However, when considering adversarial attacks, we can replace S with the set of all tasks $[T]$ and assume that each task follows the linear model (2). An attacker can then adversarially select a subset $S^c \subseteq [T]$ and distort the data distribution for these tasks. For simplicity, we do not distinguish between these two perspectives in the following parts of this paper.

---

e. In general, we have the same results hold by defining $\zeta^{(t)} = \|\boldsymbol{\theta}^{(t)*}\|_2 \vee \sqrt{\frac{p+\log T}{n}}$. Here we define $\zeta^{(t)} = \|\boldsymbol{\theta}^{(t)*}\|_2$ and assume $\min_{t\in S}\zeta^{(t)} \gtrsim \sqrt{\frac{p+\log T}{n}}$ for presentation simplicity.

f. In Duchi et al. (2022) and Tripuraneni et al. (2021), the lower bound of $\sigma_r(\boldsymbol{B}_S^*/\sqrt{|S|})$ is defined as a parameter and appears in the estimation error. Here we follow Du et al. (2020) and impose an explicit bound on it to obtain a cleaner result, but our analysis can carry over to the analysis where the lower bound of $\sigma_r(\boldsymbol{B}_S^*/\sqrt{|S|})$ is denoted as a parameter.

In the following subsections, we will present two algorithms, study their properties, and derive a lower bound of the estimation error.

2.2 The First Algorithm: Penalized ERM

2.2.1 Algorithm and Upper Bounds

In a special case of our setting, when $\boldsymbol{A}^{(t)*}$'s are the same, $S = [T]$, $\max_{t \in S} \zeta^{(t)} \lesssim 1$, and $\bar{\zeta} \gtrsim 1$, Du et al. (2020) proposed an algorithm by combining the objective functions of all tasks and solving the optimization problem. A more general version accommodating various loss functions has been explored in Tripuraneni et al. (2020). Under our setting, where the representations are similar but not exactly the same, and with the inclusion of potential outlier tasks, the objective function needs to be properly adjusted. When h is large, learning $\{\boldsymbol{\beta}^{(t)*}\}_{t \in S}$ by presuming similar representations may lead to a negative transfer effect. Considering these differences, we proposed a two-step learning approach in Algorithm 1 that addresses these issues. Note that we will apply the same algorithm to some extended models in Section 2.5, so for description convenience, we introduce the algorithm with generic loss functions $f^{(t)}$ for the t-th task. For the linear model (2), specifically, we set $f^{(t)}(\boldsymbol{\beta}) = \frac{1}{2n}\|\boldsymbol{Y}^{(t)} - \boldsymbol{X}^{(t)}\boldsymbol{\beta}\|_2 = \frac{1}{2n}\sum_{i=1}^{n}[y_i^{(t)} - (\boldsymbol{x}_i^{(t)})^\top\boldsymbol{\beta}]^2$ for $\boldsymbol{\beta} \in \mathbb{R}^p$, where $\boldsymbol{X}^{(t)} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{Y}^{(t)} \in \mathbb{R}^n$ are corresponding matrix representations of the data from the t-th task.

Algorithm 1: Penalized ERM

    Input: Data $\{\boldsymbol{X}^{(t)}, \boldsymbol{Y}^{(t)}\}_{t=1}^{T} = \{\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^{n}\}_{t=1}^{T}$, penalty parameters λ and γ,
             intrinsic dimension r
    Output: Estimators $\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^{T}, \widehat{\overline{\boldsymbol{A}}}$

1 Step 1: (Aggregation) $\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^{T}, \{\widehat{\boldsymbol{\theta}}^{(t)}\}_{t=1}^{T}, \widehat{\overline{\boldsymbol{A}}} \in$

$\arg\min_{\{\boldsymbol{A}^{(t)}\}_{t=1}^{T} \subseteq \mathcal{O}^{p \times r}, \overline{\boldsymbol{A}} \in \mathcal{O}^{p \times r}, \{\boldsymbol{\theta}^{(t)}\}_{t=1}^{T} \subseteq \mathbb{R}^r} \left\{ \frac{1}{T}\sum_{t=1}^{T} \left[f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}) \right. \right.$
$\left. \left. + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \right] \right\}$

2 Step 2: (Biased regularization) $\widehat{\boldsymbol{\beta}}^{(t)} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ f^{(t)}(\boldsymbol{\beta}) + \frac{\gamma}{\sqrt{n}}\|\boldsymbol{\beta} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2 \right\}$ for
    $t \in [T]$

In Algorithm 1, Step 1 aims to learn all tasks by aggregating the data, where the penalty $\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2$ is added to force the subspaces represented by $\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^{T}$ to be similar. This penalty is motivated by the connection between penalized over-parameterized models and robustified empirical risk minimization (ERM) (e.g., Gannaz, 2007; She and Owen, 2011; Donoho and Montanari, 2016; Duan and Wang, 2023). Specifically, adding such a penalty is equivalent to employing a robustified loss function to estimate the central subspace $\overline{\boldsymbol{A}}$, which is then adapted to obtain individual estimators for each $\boldsymbol{A}^{(t)*}$. As discussed in Section 1.2.2, formulating the problem in our setting naturally leads to this penalty structure, which provides an advantage over existing approaches such as those in Chua et al. (2021) and Duan and Wang (2023). Moreover, note that when $\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 = 0$, it can be shown that $\boldsymbol{A}^{(t)}$ and $\overline{\boldsymbol{A}}$ are identical up to a rotation. And this distance is equivalent to other subspace distances like sin-Θ distance or the distance between two matrices up to a rotation (Chen et al., 2021b). Step 2 uses data from each task to make proper corrections to

prevent negative transfer, which is often referred to as *biased regularization* in the literature (Schölkopf et al., 2001; Kuzborskij and Orabona, 2013, 2017). Such two-step methods are widely used in the distance-based MTL and TL literature to alleviate the adverse effect of negative transfer (Bastani, 2021; Li et al., 2022b; Lin and Reimherr, 2022; Tian and Feng, 2022). It is important to point out that $\boldsymbol{A}^{(t)*}$'s and $\boldsymbol{\theta}^{(t)*}$'s are not uniquely identifiable in model (2), which does not pose an issue because our focus is on estimating $\boldsymbol{\beta}^{(t)*}$'s.

Next, we proceed to present the upper bound on estimation errors of $\{\boldsymbol{\beta}^{(t)*}\}_{t \in S}$ incurred by Algorithm 1.

Theorem 2 (Upper bound for Algorithm 1) *Suppose Assumptions 1-3 hold with a subset $S \subseteq [T]$ satisfying $\epsilon = \frac{|S^c|}{T} \leq cr^{-3/2}\frac{\bar{\zeta}^2}{\max_{t \in S}(\zeta^{(t)})^2}$, where $c > 0$ is a small constant. By setting $\lambda = C\frac{\max_{t \in S}(\zeta^{(t)})^2}{\min_{t \in S}\zeta^{(t)}}\sqrt{r(p + \log T)}$ and $\gamma = C'\sqrt{p + \log T}$ with sufficiently large positive constants C and C', for an arbitrary distribution $\mathbb{Q}_{S^c}$ of $\{\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n\}_{t \in S^c}$, w.p. at least $1 - e^{-C''(r + \log T)}$, we have*

$$\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \left\{ \frac{\zeta^{(t)}}{\bar{\zeta}}r\sqrt{\frac{p}{nT}} + \left(\frac{\zeta^{(t)}\sqrt{r}}{\bar{\zeta}} \vee 1\right)\sqrt{\frac{r + \log T}{n}} + \zeta^{(t)}\sqrt{r}h \right.$$

$$\left. + \frac{\zeta^{(t)}\max_{t \in S}(\zeta^{(t)})^2}{\bar{\zeta}^2 \min_{t \in S}\zeta^{(t)}}r^{3/2}\sqrt{\frac{p + \log T}{n}}\epsilon \right\} \wedge \sqrt{\frac{p + \log T}{n}}, \quad \forall t \in S.$$

Furthermore, if the data from tasks in S^c satisfies the linear model (2) (without any latent structure assumption) and Assumption 1, then w.p. at least $1 - e^{-C'(p + \log T)}$, we also have

$$\max_{t \in S^c}\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \sqrt{\frac{p + \log T}{n}}.$$

In Theorem 2, the upper bound of $\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ for $t \in S$ is the minimum of two terms, where the first term corresponds to the rate obtained through data aggregation and the second term corresponds to the single-task rate. This result shows that our algorithm is automatically **adaptive** to the optimal situation, whether or not aggregating data across tasks is beneficial. Furthermore, it demonstrates that Algorithm 1 is **robust** to a small fraction of outlier tasks, in the sense that representation MTL remains beneficial when the outlier proportion ϵ is small.

Analyzing the components of the upper bound of $\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ for $t \in S$ further elucidates their interpretability. The term $\frac{\zeta^{(t)}}{\bar{\zeta}}r\sqrt{\frac{p}{nT}} + \zeta^{(t)}\sqrt{r}h$ arises from learning similar representations, $\left(\frac{\zeta^{(t)}\sqrt{r}}{\bar{\zeta}} \vee 1\right)\sqrt{\frac{r + \log T}{n}}$ is due to learning the representations and task-specific parameters, $\frac{\zeta^{(t)}\max_{t \in S}(\zeta^{(t)})^2}{\bar{\zeta}^2 \min_{t \in S}\zeta^{(t)}}r^{3/2}\sqrt{\frac{p + \log T}{n}}\epsilon$ accounts for outlier tasks, and $\sqrt{\frac{p + \log T}{n}}$ is the error rate of single-task learning. Note that the term $\zeta^{(t)}\sqrt{r}h$ does not explicitly depend on n (although $\zeta^{(t)}, r, h$ might depend on n). At first glance, this may appear overly restrictive, but the lower bound in Section 2.4 contains a similar term that does not involve n explicitly. This phenomenon was also observed in many distance-based multi-task and transfer learning studies, such as Li et al. (2022b,a); Lin and Reimherr (2022); Tian and Feng (2022); Duan and Wang (2023).

To understand when Algorithm 1 improves upon single-task learning, consider the scenario $\max_{t \in S} \zeta^{(t)} \lesssim 1$ and $\bar{\zeta} \gtrsim 1$, where the upper bound simplifies to $\left(r\sqrt{\frac{p}{nT}} + \sqrt{r}\sqrt{\frac{r+\log T}{n}} + \sqrt{r}h + r^{3/2}\sqrt{\frac{p+\log T}{n}}\epsilon\right) \wedge \sqrt{\frac{p+\log T}{n}}$. When $T \gg r^2$ (many tasks), $h \ll \sqrt{\frac{p+\log T}{nr}}$ (similar representations), $p \gg r(r \vee \log T)$ (low intrinsic dimension), and $\epsilon \ll r^{-3/2}$ (a small fraction of outlier tasks), the rate is faster than the single-task error rate $\sqrt{\frac{p+\log T}{n}}$.

Furthermore, contrary to the results in existing literature (Du et al., 2020; Chua et al., 2021; Tripuraneni et al., 2021; Thekumparampil et al., 2021; Duan and Wang, 2023), our results indicate that the performance of representation MTL on each task depends critically on the norm $\zeta^{(t)} = \|\boldsymbol{\theta}^{(t)*}\|_2$. This insight suggests that tasks characterized by smaller coefficients may benefit more from representation MTL. We will verify this phenomenon through numerical experiments in Section 4.1.4.

It is also worth noting that Assumption 3 can be relaxed to $n \gtrsim r + \log T$ for penalized ERM, provided that we are willing to forego the safe-net guarantee $\sqrt{\frac{p+\log T}{n}}$. However, omitting the minimum with $\sqrt{\frac{p+\log T}{n}}$ in the estimation error can expose penalized ERM to negative transfer effects when either h or ϵ is large.

Before closing this subsection, we present the following theorem, which demonstrates that the personalized estimators $\{\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\}_{t=1}^T$ obtained in Step 1 already achieve strong performance, albeit with a loss of a factor of $\sqrt{r}$ and other scaling factors compared to the estimation error of $\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T$ in Theorem 2. In this sense, Step 2 in Algorithm 1 can be seen as a refinement that further improves estimation accuracy.

Theorem 3 (Step 1 only) *Under the same assumptions imposed in Theorem 2, for any $S \subseteq [T]$ satisfying $\epsilon = \frac{|S^c|}{T} \leq cr^{-3/2}\frac{\bar{\zeta}^2}{\max_{t \in S}(\zeta^{(t)})^2}$ with a small constant $c > 0$, and an arbitrary distribution $\mathbb{Q}_{S^c}$ of $\{\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n\}_{t \in S^c}$, w.p. at least $1 - e^{-C'(r+\log T)}$, we have*

$$\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \left\{ \frac{\zeta^{(t)}}{\bar{\zeta}}\sqrt{\frac{p}{nT}} + \left(\frac{\zeta^{(t)}\sqrt{r}}{\bar{\zeta}} \vee 1\right)\sqrt{\frac{r+\log T}{n}} + \zeta^{(t)}\sqrt{r}h \right.$$

$$\left. + \frac{\zeta^{(t)}\max_{t \in S}(\zeta^{(t)})^2}{\bar{\zeta}^2\min_{t \in S}\zeta^{(t)}}\sqrt{\frac{p+\log T}{n}}\epsilon \right\} \wedge \left\{ \sqrt{\frac{r(p+\log T)}{n}}\frac{\max_{t \in S}(\zeta^{(t)})^2}{\zeta^{(t)}\min_{t \in S}\zeta^{(t)}} \right\},$$

$$\forall t \in S.$$

As pointed out in Section 1.2, our setting reduces to the setting in Gu et al. (2024) when $r = 1$, where they considered a ridge regression by penalizing the angle between different regression coefficients. The main idea of our Algorithm 1 resembles their approach, but they only considered the case when the angles between different regression coefficients are in $[0, \pi/2]$. Our result shows that representation MTL helps as long as the subspaces spanned by each coefficient (i.e., the straight line) are similar, where the angles can be either close to 0 or π. Moreover, Theorem 3 shows that when r is a constant, the estimators obtained from Step 1 of Algorithm 1 already achieve a desired rate (despite an inflation term in the single-task rate $\sqrt{\frac{p+\log T}{n}}$), which means Step 2 may be omitted.

2.2.2 Discussions on Local Minimizers

It can be observed that the optimization problem in Step 1 of Algorithm 1 is non-convex. Therefore, in addition to studying the global minimizers as previously examined, it is also crucial to investigate the properties of local minimizers. Specifically, if only a local minimizer is found in Step 1, does Algorithm 1 still achieve the same upper bound in Theorem 2?

Our first result below focuses on a special case where $S = [T]$ and $\lambda = +\infty$. In this scenario, Step 1 of Algorithm 1 reduces to the ERM algorithm introduced in Du et al. (2020). The following theorem shows that even if we replace the global minimizer with a local minimizer in Step 1, Algorithm 1 still achieves the same upper bound as stated in Theorem 2. In other words, there is no "bad" local minimizer in this case.

Theorem 4 (No bad local minimizer when $S = [T]$ and $\lambda = +\infty$) *When $S = [T]$ and $\lambda = +\infty$, replacing the global minimizer in Step 1 of Algorithm 1 with any local minimizer* [g] *delivers the same upper bound as in Theorem 2, i.e. w.p. at least $1 - e^{-C'(r+\log T)}$,*

$$\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \left\{ \frac{\zeta^{(t)}}{\bar{\zeta}} r\sqrt{\frac{p}{nT}} + \left(\frac{\zeta^{(t)}\sqrt{r}}{\bar{\zeta}} \vee 1\right)\sqrt{\frac{r + \log T}{n}} + \zeta^{(t)}\sqrt{r}h \right\} \wedge \sqrt{\frac{p + \log T}{n}}, \ \forall t \in [T].$$

Tripuraneni et al. (2021) proves a similar result for the ERM when $S = [T]$, $h = 0$, and $\lambda = +\infty$. However, their result only holds for ERM with an additional regularization term, and the optimization must be conducted within a constrained set. In contrast, Theorem 4 shows that there is no bad local minimizer across the *entire optimization landscape*. This result significantly supplements the existing literature for the case $S = [T]$.

Next, we present a more general result for the generic case $S \neq [T]$ with the same λ value as in Theorem 2. Unfortunately, similar to the result in Tripuraneni et al. (2021), in this general case, we can only establish that there is no bad local minimizer within a large regime instead of the entire landscape. This implies that we can replace the global minimizer in Step 1 with a local minimizer under some constraints to obtain the same upper bound.

Theorem 5 (No bad local minimizer in a large regime, for general $S \neq [T]$) *Consider the same λ value taken in Theorem 2. Replacing the global minimizer in Step 1 of Algorithm 1 with any local minimizer satisfying*

$$4(1 + \sqrt{2}) \max_{t \in S} \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 + 4\frac{\epsilon}{1 - \epsilon} < 1 - \gamma;$$

$$\sqrt{2} \max_{t \in S} \left\{ \frac{\sigma_{\max}(\boldsymbol{\Sigma}^{(t)})}{\sigma_{\min}(\boldsymbol{\Sigma}^{(t)})} \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top\|_2 \right\} < 1 - \gamma,$$

with a universal constant $\gamma > 0$ delivers the same upper bound as in Theorem 2, i.e. w.p. at least $1 - e^{-C'(r+\log T)}$, for all $t \in S$,

$$\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \left\{ \frac{\zeta^{(t)}}{\bar{\zeta}} r\sqrt{\frac{p}{nT}} + \left(\frac{\zeta^{(t)}\sqrt{r}}{\bar{\zeta}} \vee 1\right)\sqrt{\frac{r + \log T}{n}} + \zeta^{(t)}\sqrt{r}h \right.$$

---

g. Here we say $\{\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}}, \{\widehat{\boldsymbol{\theta}}^{(t)}\}_{t=1}^T\}$ is a local minimizer of function $G(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \overline{\boldsymbol{A}}, \{\boldsymbol{\theta}^{(t)}\}_{t=1}^T) = \sum_{t=1}^T f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}) + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2$, if there exists a constant $\delta > 0$, such that for any $\{\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \boldsymbol{A}, \{\boldsymbol{\theta}^{(t)}\}_{t=1}^T\}$ with $\max_{t \in [T]} \min_{\boldsymbol{R} \in \mathcal{O}^{p \times r}} \{\|\widehat{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)}\boldsymbol{R}\|_2 + \|\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{R}^\top\boldsymbol{\theta}^{(t)}\|_2\} + \min_{\boldsymbol{R} \in \mathcal{O}^{p \times r}} \|\widehat{\overline{\boldsymbol{A}}} - \boldsymbol{A}\boldsymbol{R}\|_2 \leq \delta$, we must have $G(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}}, \{\widehat{\boldsymbol{\theta}}^{(t)}\}_{t=1}^T) < G(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \boldsymbol{A}, \{\boldsymbol{\theta}^{(t)}\}_{t=1}^T)$. The same definition is used in Theorem 5.

$$+ \frac{\zeta^{(t)} \max_{t \in S}(\zeta^{(t)})^2}{\bar{\zeta}^2 \min_{t \in S} \zeta^{(t)}} r^{3/2} \sqrt{\frac{p + \log T}{n}} \cdot \epsilon \Bigg\} \wedge \sqrt{\frac{p + \log T}{n}}.$$

Theorem 5 is similar to the results from other non-convex problems such as the EM algorithm for the mixture models (e.g., Balakrishnan et al., 2017; Cai et al., 2019), where a contraction basin exists within which the only local minimizer is the global one. This suggests that a good initialization can be used in the optimization problem of Step 1 to avoid bad local minima. We conjecture a similar result in Theorem 4 holds in the general case as well, i.e. there is no bad local minimizer in the entire optimization landscape. This conjecture remains a direction for future research.

The diagrams in Figure 4 below summarize the results discussed in this subsection.

(a) The case $S = [T]$ and $\lambda = +\infty$ (b) General case $S \neq [T]$

Area with no local minimizer

Desired estimation error range

Area which we conjecture has no local minimizer
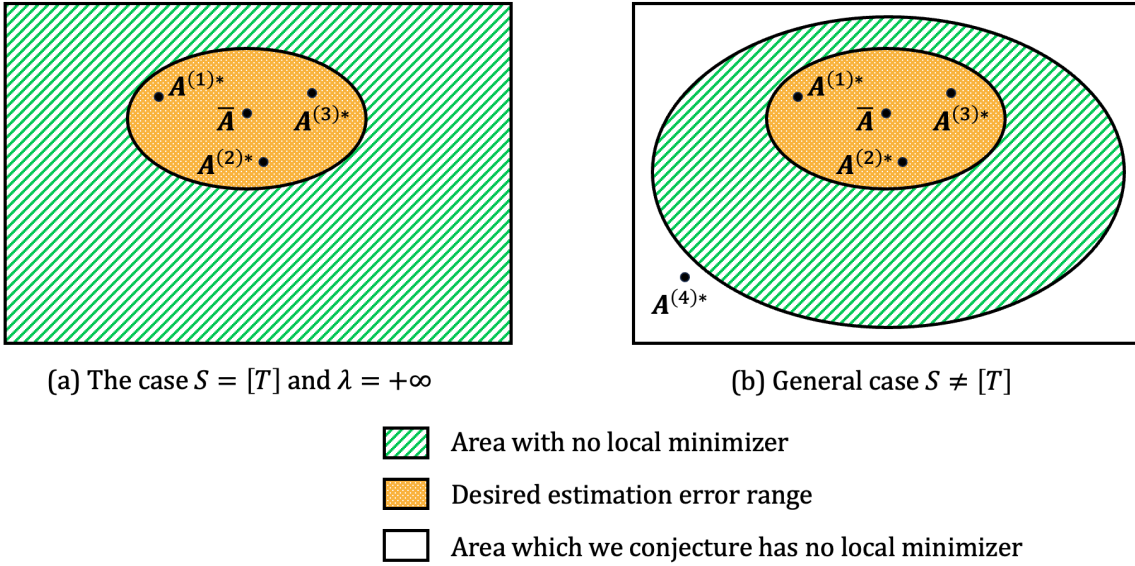
Figure 4: The optimization landscape of Algorithm 1 proved in Theorems 4 and 5. The big rectangle is the full optimization landscape.

2.2.3 IMPLEMENTATION

While we have provided desired upper bounds of estimation error with both local and global minimizers in Step 1 of Algorithm 1, solving the optimization problem there is challenging due to the constraint that representation matrices $\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T$ and $\widehat{\overline{\boldsymbol{A}}}$ must belong to the orthonormal space $\mathcal{O}^{p \times r} = \{\boldsymbol{A} \in \mathbb{R}^{p \times r} : \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_r\}$. In this subsection, we want to point out that this restriction is adopted mainly for clarity of exposition. In practice, we can formulate the optimization problem in $\mathbb{R}^{p \times r}$ with a modified penalty term, achieving the same theoretical guarantees as with $\mathcal{O}^{p \times r}$.

Before discussing the practical formulation of the optimization problem in Step 1, we first introduce several new notations. For any non-zero matrix $\boldsymbol{A} \in \mathbb{R}^{p \times r}$ with $p \geq r$, we define the projection matrix onto the column space of $\boldsymbol{A}$ as $\boldsymbol{P_A} = \boldsymbol{U}\boldsymbol{U}^\top$, where $\boldsymbol{U} \in \mathcal{O}^{p \times r'}$ is the left singular matrix in the SVD of $\boldsymbol{A} = \boldsymbol{U}_{p \times r'} \boldsymbol{\Lambda}_{r' \times r'} \boldsymbol{V}_{r' \times r'}^\top$, $1 \leq r' \leq r$, and $\boldsymbol{\Lambda}_{r' \times r'}$ is

diagonal with positive entries. When $\text{rank}(\boldsymbol{A}) = r$, we have $\boldsymbol{P_A} = \boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top$. When $\boldsymbol{A}$ is zero, we define $\boldsymbol{P_A} = \boldsymbol{0}_{p \times p}$. Note that while $\boldsymbol{U}$ in the SVD of $\boldsymbol{A}$ may not be unique, $\boldsymbol{P_A}$ is unique by the Hilbert projection theorem.

In practice, Step 1 of Algorithm 1, i.e.

$$\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \{\widehat{\boldsymbol{\theta}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}} \in \underset{\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \overline{\boldsymbol{A}} \subseteq \mathcal{O}^{p \times r}, \{\boldsymbol{\theta}^{(t)}\}_{t=1}^T \subseteq \mathbb{R}^r}{\arg\min} \left\{ \frac{1}{T} \sum_{t=1}^T \left[f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}) \right. \right.$$
$$\left. \left. + \frac{\lambda}{\sqrt{n}} \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \right] \right\},$$

is equivalent to solving the following relaxed problem

$$\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \{\widehat{\boldsymbol{\theta}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}} \in \underset{\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \overline{\boldsymbol{A}} \subseteq \mathbb{R}^{p \times r}, \{\boldsymbol{\theta}^{(t)}\}_{t=1}^T \subseteq \mathbb{R}^r}{\arg\min} \left\{ \frac{1}{T} \sum_{t=1}^T \left[f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}) + \frac{\lambda}{\sqrt{n}} \|\boldsymbol{P}_{\boldsymbol{A}^{(t)}} - \boldsymbol{P}_{\overline{\boldsymbol{A}}}\|_2 \right] \right\}.$$
$$(3)$$

This replacement relaxes the constraint that $\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T$ and $\widehat{\overline{\boldsymbol{A}}}$ must belong to the orthonormal space $\mathcal{O}^{p \times r} = \{\boldsymbol{A} \in \mathbb{R}^{p \times r} : \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_r\}$, allowing them to reside in $\mathbb{R}^{p \times r}$ with a modified penalty term. Intuitively, we can see why this relaxation preserves our previous theoretical guarantees. For any $\boldsymbol{A}^{(t)} \in \mathbb{R}^{p \times r}$ and $\boldsymbol{\theta}^{(t)} \in \mathbb{R}^r$, it can be shown that there exist $\widetilde{\boldsymbol{A}}^{(t)} \in \mathcal{O}^{p \times r}$ and $\widetilde{\boldsymbol{\theta}}^{(t)} \in \mathbb{R}^r$ such that $\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} = \widetilde{\boldsymbol{A}}^{(t)}\widetilde{\boldsymbol{\theta}}^{(t)}$ and $\boldsymbol{P}_{\boldsymbol{A}^{(t)}} = \boldsymbol{P}_{\widetilde{\boldsymbol{A}}^{(t)}}$. Since we are interested in the product instead of $\boldsymbol{A}^{(t)} \in \mathbb{R}^{p \times r}$ and $\boldsymbol{\theta}^{(t)} \in \mathbb{R}^r$ individually, the replacement does not affect the estimation error of $\boldsymbol{\beta}^{(t)*}$. The following theorem formally presents this result.

Theorem 6 *We replace the Step 1 of Algorithm 1 with the real-matrix optimization problem (3), and replace $\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)}) - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2$, $\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)}) - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top\|_2$ in Theorem 5 with $\|\boldsymbol{P}_{\widehat{\boldsymbol{A}}^{(t)}} - \boldsymbol{P}_{\overline{\boldsymbol{A}}}\|_2$ and $\|\boldsymbol{P}_{\widehat{\boldsymbol{A}}^{(t)}} - \boldsymbol{P}_{\boldsymbol{A}^{(t)*}}\|_2$. All theoretical results in Sections 2.2.1 and 2.2.2 remain valid.*

In all the numerical experiments, we used the automatic differentiation implemented in PyTorch (Paszke et al., 2019) along with the Adam optimizer (Kingma and Ba, 2015) to solve the optimization problem (3). Further implementation details will be discussed in Section 4.

It is also possible to conduct optimization directly on the Stiefel manifold $\mathcal{O}^{p \times r}$ in Step 1, which is recently explored in Chen and Feng (2025).

2.3 The Second Algorithm: Spectral Method

In this subsection, we propose another algorithm for the representation MTL, which is based on singular value decomposition (SVD), and we refer to this approach as the *Spectral Method.*

The motivation of the spectral method arises from the special case $h = 0$ and $S = [T]$, implying $\boldsymbol{A}^{(t)*} \equiv \overline{\boldsymbol{A}}$ for all $t \in [T]$. In this scenario, $\boldsymbol{B}_S^* = (\boldsymbol{\beta}^{(1)*}, \ldots, \boldsymbol{\beta}^{(T)*}) = \overline{\boldsymbol{A}}_{p \times r}\boldsymbol{\Theta}_{r \times T}^*$ with $\boldsymbol{\Theta}^* = (\boldsymbol{\theta}^{(1)*}, \ldots, \boldsymbol{\theta}^{(T)*})$ is a rank-r matrix, and the column spaces of $\boldsymbol{B}_S^*$ and $\overline{\boldsymbol{A}}$ are the same. Thus, the left singular matrix in the SVD of $\boldsymbol{B}_S^*$ can exactly recover $\overline{\boldsymbol{A}}$. In

practice, we can use the single-task estimator to estimate each column of $\boldsymbol{B}_S^*$ and apply SVD to estimate $\overline{\boldsymbol{A}}$. With the estimated $\overline{\boldsymbol{A}}$, we can then perform a single-task regression to estimate $\boldsymbol{\theta}^{(t)*}$. When h is not necessarily zero and $S \neq [T]$, the same idea can still be applied. However, to avoid negative transfer when h is large, we can follow the SVD with a biased regularization step similar to Algorithm 1. Additionally, we need to conduct SVD on a robust estimator of $\boldsymbol{B}_S^*$ to accommodate outlier tasks.

We use $\prod_R$ to denote the projection operator to an ℓ_2-ball centered at zero of radius R in $\mathbb{R}^p$, and use $\texttt{quantile}(\{a_t\}_{t=1}^T, 1 - \bar{\epsilon})$ to denote the lower $(1 - \bar{\epsilon})$-quantile of a sequence $\{a_t\}_{t=1}^T \subseteq \mathbb{R}$, where $\bar{\epsilon} \in [0, 1]$. We formalize the intuition into the spectral method in Algorithm 2.

Algorithm 2: Spectral Method

   Input: Data $\{\boldsymbol{X}^{(t)}, \boldsymbol{Y}^{(t)}\}_{t=1}^T = \{\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n\}_{t=1}^T$, penalty parameter γ, an upper
          bound $\bar{\epsilon}$ (for ϵ), intrinsic dimension r

   Output: Estimators $\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}}$

1 <u>Step 1:</u> (Single-task regression) $\widetilde{\boldsymbol{\beta}}^{(t)} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \{f^{(t)}(\boldsymbol{\beta})\}$ for $t \in [T]$

2 <u>Step 2:</u> (Projection and concatenation) Create a $p \times T$ matrix $\widehat{\boldsymbol{B}}$ of which the t-th
    column is $\prod_R(\widetilde{\boldsymbol{\beta}}^{(t)})$, where $R = \texttt{quantile}(\{\|\widetilde{\boldsymbol{\beta}}^{(t)}\|_2\}_{t=1}^T, 1 - \bar{\epsilon})$

3 <u>Step 3:</u> (SVD) Conduct SVD $\widehat{\boldsymbol{B}} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{V}}^\top$ with $\widehat{\boldsymbol{U}} \in \mathcal{O}^{p \times T}$, let $\widehat{\overline{\boldsymbol{A}}}$ be the first r
    columns of $\widehat{\boldsymbol{U}}$, and set $\widehat{\boldsymbol{\theta}}^{(t)} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^r} f^{(t)}(\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta})$

4 <u>Step 4:</u> (Biased regularization) $\widehat{\boldsymbol{\beta}}^{(t)} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \{f^{(t)}(\boldsymbol{\beta}) + \frac{\gamma}{\sqrt{n}}\|\boldsymbol{\beta} - \widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\theta}}^{(t)}\|_2\}$ for
    $t \in [T]$

Steps 1 and 2 of Algorithm 2 construct a robust estimator of $\boldsymbol{B}_S^*$ by concatenating projected single-task estimators. The projection limits the impact of outlier tasks by setting the projection radius to a quantile of the ℓ_2-norms of all single-task estimators. This projection technique, sometimes referred to as truncation or winsorization, is widely used in robust statistics (e.g., Lugosi and Mendelson, 2021) and differential privacy (e.g., Dwork, 2006; Dwork et al., 2014). To determine the quantile percentage, we need an upper bound $\bar{\epsilon}$ of ϵ. If we have prior knowledge of ϵ (but not S and S^c), we can set $\bar{\epsilon} = \epsilon$. Without this information, $\bar{\epsilon}$ can be chosen as a small constant, such as 0.05. In the simulations, we set $\bar{\epsilon}$ as the true ϵ value, and in the real-data study, we set $\bar{\epsilon} = 0.05$. Step 3 performs the SVD to estimate the central representation $\overline{\boldsymbol{A}}$. Step 4 applies biased regularization as in Algorithm 1 to prevent negative transfer. The spectral method avoids the complicated non-convex optimization problems present in the penalized ERM algorithm, making it easy to implement in practice. In fact, the spectral method can be solved in polynomial time since the objective function in Step 4, which consists of a sum of a smooth and strongly convex component and an ℓ_2-Lipschitz component, can be minimized in polynomial time (e.g., via stochastic gradient descent; see Shamir and Zhang, 2013). Moreover, SVD can also be computed in polynomial time. Besides its simplicity and efficiency, we can show that it achieves a better upper bound of estimation error for $\boldsymbol{\beta}^{(t)*}$'s when $S = [T]$.

Theorem 7 (Upper bound for spectral method) *Suppose Assumptions 1-3 hold with a subset $S \subseteq [T]$ satisfying $\epsilon = \frac{|S^c|}{T} \leq cr^{-1} \cdot \left(\frac{\bar{\zeta}}{\max_{t \in S} \zeta^{(t)}} \right)^2$, where $c > 0$ is a small constant. By setting $\gamma = C'\sqrt{p + \log T}$ with a sufficiently large positive constant C', w.p. at least $1 - e^{-C''(r + \log T)}$ and $\bar{\epsilon}$ satisfying $\epsilon \leq \bar{\epsilon} \leq \frac{c''}{r} \cdot \left(\frac{\bar{\zeta}}{\max_{t \in S} \zeta^{(t)}} \right)^2$ with a sufficiently small positive constant c'', we have* [h]*

$$\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \left\{ \frac{\zeta^{(t)}}{\bar{\zeta}} \sqrt{\frac{pr}{nT}} + \left(\frac{\zeta^{(t)}}{\bar{\zeta}} \vee 1 \right) \sqrt{\frac{r}{n}} + \sqrt{\frac{\log T}{n}} + \zeta^{(t)} h \cdot \left[\frac{\sigma_{\max}((\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{B}_S^*)}{\sigma_{\min}((\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{B}_S^*)} \wedge \sqrt{r} \right] \right.$$

$$\left. + \frac{\zeta^{(t)}}{\bar{\zeta}} \max_{t \in S} \zeta^{(t)} \cdot \sqrt{r\bar{\epsilon}} \right\} \wedge \sqrt{\frac{p + \log T}{n}}, \qquad \forall t \in S,$$

where $\boldsymbol{B}_S^* \in \mathbb{R}^{p \times |S|}$ is the coefficient matrix whose columns are $\{\boldsymbol{\beta}^{(t)*}\}_{t \in S}$, $\overline{\boldsymbol{A}} \in \arg\min_{\boldsymbol{A} \in \mathcal{O}^{p \times r}} \max_{t \in [T]} \|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \boldsymbol{A}\boldsymbol{A}^\top\|_2$ is the central representation, and $\overline{\boldsymbol{A}}^\perp$ is orthogonal to $\overline{\boldsymbol{A}}$ in the sense that $\overline{\boldsymbol{A}}^\perp \in \mathcal{O}^{p \times (p-r)}$ and $(\overline{\boldsymbol{A}}^\perp)^\top \overline{\boldsymbol{A}} = \boldsymbol{0}_{(p-r) \times r}$.

Furthermore, if we assume the data from tasks in S^c satisfies the linear model (2) (without any latent structure assumption) and Assumption 1, then w.p. at least $1 - e^{-C'(p + \log T)}$, we also have

$$\max_{t \in [T]} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \sqrt{\frac{p + \log T}{n}}.$$

We can similarly discuss when Algorithm 2 improves single-task learning as in Theorem 2, which we do not repeat here. However, it is worth emphasizing that, like Algorithm 1, Algorithm 2 is also **adaptive** to the unknown similarity structure and it is also **robust** to a small fraction of outlier tasks. Comparing the rates in Theorems 2 and 7, we can see that the estimation error rate of the spectral method has a better dependence on r when $S = [T]$, although the last term related to outlier tasks is worse. In the next subsection, we will see that the spectral method is minimax optimal when $S = [T]$ and the condition number $\sigma_{\max}((\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{B}_S^*)/\sigma_{\min}((\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{B}_S^*)$ is bounded by a constant.

As mentioned, Algorithm 2 requires an upper bound $\bar{\epsilon}$ of ϵ for determining the quantile percentage, and $\bar{\epsilon}$ also appears in the upper bound of estimation error. If $\bar{\epsilon} \asymp \epsilon$, then we can replace $\bar{\epsilon}$ in the upper bound with the true ϵ.

Besides the intuition provided at the beginning of this subsection, Algorithm 2 also has connections to average derivative estimation (ADE) and expected gradient outer product (EGOP) methods used to estimate the index space in single-index models and multi-index models (e.g., Härdle and Stoker, 1989; Samarov, 1993; Hristache et al., 2001; Yang et al., 2017; Yuan et al., 2023). For instance, in the multi-index model $\mathbb{E}[Y|X = \boldsymbol{x}] := g(\boldsymbol{x}) = f((\boldsymbol{B}^*)^\top \boldsymbol{x}) \in \mathbb{R}$ with $\boldsymbol{B}^* \in \mathbb{R}^{p \times T}$, $\boldsymbol{x} \in \mathbb{R}^p$, g continuously differentiable and unknown, and $\mathbb{E}[\nabla g(X)(\nabla g(X))^\top]$ existing, we have:

$$\mathbb{E}[\nabla g(X)(\nabla g(X))^\top] = (\boldsymbol{B}^*)^\top \mathbb{E}\{\nabla f((\boldsymbol{B}^*)^\top X)[\nabla f((\boldsymbol{B}^*)^\top X)]^\top\} \boldsymbol{B}^*.$$

---

h. For convention, we define $0/0 = 0$, which is mainly for the case $(\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{B}^* = \boldsymbol{0}_{(p-r) \times r}$.

Therefore, it is possible to use SVD or PCA on some estimator of $\mathbb{E}[\nabla g(X)(\nabla g(X))^\top]$ to recover the column space of $\boldsymbol{B}_S^*$, and many estimators of the expected gradient outer product have been proposed in the literature. In the context of representation MTL, when $S = [T]$, we can view g as a multivariate function $\boldsymbol{x} \mapsto ((\boldsymbol{\beta}^{(1)*})^\top \boldsymbol{x}, \ldots, (\boldsymbol{\beta}^{(T)*})^\top \boldsymbol{x})^\top$ with f the identity function from $\mathbb{R}^T$ to $\mathbb{R}^T$, implying that $\mathbb{E}[\nabla g(X)(\nabla g(X))^\top] = (\boldsymbol{B}_S^*)^\top \boldsymbol{B}_S^*$. Thus, our spectral method, which performs SVD on the estimated $\boldsymbol{B}_S^*$, can be viewed as a multi-task variant of the EGOP framework.

The SVD step in our method is also closely related to the SVD-based approaches used in Kong et al. (2020) and Meunier et al. (2023), where the former considers a mixture model setting and the latter operates in an infinite-dimensional RKHS framework. Translated into our terminology, the key difference is that their methods rely on data splitting to estimate two instances of $\boldsymbol{B}_S^*$'s separately in the product $(\boldsymbol{B}_S^*)^\top \boldsymbol{B}_S^*$. In contrast, our approach applies SVD directly to the unbiased estimator $\widehat{\boldsymbol{B}}$ rather than to the potentially biased $\widehat{\boldsymbol{B}}^\top \widehat{\boldsymbol{B}}$ and thereby eliminates the need for sample splitting.

When there are outlier tasks or contaminations, i.e., $S \neq [T]$, besides the simple projection technique, we can also borrow ideas from robust principal component analysis (PCA) literature (e.g., Wright et al., 2009; Candès et al., 2011; Vidal et al., 2016) to robustify the SVD procedure in Algorithm 2. For example, instead of conducting SVD on $\widehat{\boldsymbol{B}}$ in Step 3 of Algorithm 2, we may conduct SVD on another $p \times T$ matrix $\widehat{\boldsymbol{L}}$ which is the solution of the convex optimization (Liu et al., 2012; Xu et al., 2012)

$$\min_{\boldsymbol{L},\boldsymbol{Z}} \|\boldsymbol{L}\|_* + \lambda \|\boldsymbol{Z}\|_{2,1} \quad \text{s.t. } \widehat{\boldsymbol{B}} = \boldsymbol{L} + \boldsymbol{Z},$$

where $\|\boldsymbol{L}\|_* = \sum_{j=1}^{p \wedge T} \sigma_j(\boldsymbol{L})$ is the nuclear norm, and $\|\boldsymbol{Z}\|_{2,1} = \sum_{t=1}^T \|\boldsymbol{z}_t\|_2$ with $\boldsymbol{z}_t$ the t-th column of $\boldsymbol{Z}$ is the $L_{2,1}$-norm. This approach might lead to better performance when $S \neq [T]$. Given the extensive scope of the current paper, we leave the study of this approach for future research.

2.4 Lower Bound

In this subsection, we derive lower bounds to explore the information-theoretic hardness of the representation MTL problem. Consider a collection of all subsets $S \subseteq [T]$ as

$$\mathcal{S} = \{S \subseteq [T] : |S^c|/T \leq \epsilon\}.$$

Given the subset S, define a coefficient matrix $\boldsymbol{B}_S \in \mathbb{R}^{p \times |S|}$, where each column corresponds to a coefficient vector in $\{\boldsymbol{\beta}^{(t)}\}_{t \in S}$. Consider the parameter space for the coefficient vectors $\{\boldsymbol{\beta}^{(t)}\}_{t \in S}$ as

$$\mathscr{B}(S, h) = \left\{ \{\boldsymbol{\beta}^{(t)}\}_{t \in S} : \boldsymbol{\beta}^{(t)} = \boldsymbol{A}^{(t)} \boldsymbol{\theta}^{(t)} \text{ for all } t \in S, \{\boldsymbol{A}^{(t)}\}_{t \in S} \subseteq \mathcal{O}^{p \times r}, \|\boldsymbol{\theta}^{(t)}\|_2 \leq \zeta^{(t)}, \right.$$

$$\left. \min_{\overline{\boldsymbol{A}} \in \mathcal{O}^{p \times r}} \max_{t \in S} \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2 \leq h, \sigma_r\left(|S|^{-1/2}\boldsymbol{B}_S\right) \geq \frac{c}{\sqrt{r}}\sqrt{\frac{1}{|S|}\sum_{t \in S}\|\boldsymbol{\theta}^{(t)}\|_2^2} \right\}$$

where c can be any fixed positive constant such that $\mathscr{B}(S, h) \neq \emptyset$. Given a set $S \subseteq [T]$, denote $\bar{\zeta} := \bar{\zeta}(S) = \sqrt{\frac{1}{|S|}\sum_{t \in S}(\zeta^{(t)})^2}$. We have the following lower bound.

Theorem 8 (Lower bound for MTL) *Suppose $p \geq 2r$, $T \geq r^{1.01}$, $\min_{t \in [T]} \zeta^{(t)} \geq C$, and $\epsilon \leq c/r$ where C and c are some positive constants. We have the following lower bound hold:*

$$
\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^{T}} \sup_{S \subseteq \mathcal{S}} \sup_{\substack{\{\boldsymbol{\beta}^{(t)}\}_{t \in S} \in \mathscr{B}(S,h) \\ \mathbb{Q}_{S^c}}} \mathbb{P}\left(\bigcup_{t \in S} \left\{ \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \gtrsim \left[\frac{\zeta^{(t)}}{\bar{\zeta}} \sqrt{\frac{pr}{nT}} + \left(\frac{\zeta^{(t)}}{\bar{\zeta}} \vee 1 \right) \sqrt{\frac{r}{n}} + \frac{\log T}{n} \right. \right.\right.
$$

$$
\left.\left.\left. + \zeta^{(t)} h + \frac{\zeta^{(t)}}{\bar{\zeta}} \frac{\epsilon r}{\sqrt{n}} \right] \wedge \sqrt{\frac{p + \log T}{n}} \right\} \right) \geq \frac{1}{10},
$$

where for any given $S \subseteq \mathcal{S}$, $\mathbb{P} = \mathbb{P}_S \otimes \mathbb{Q}_{S^c}$, and $\mathbb{P}_S$, $\mathbb{Q}_{S^c}$ are the probability measures on sample space of tasks in S, S^c, respectively. Furthermore, if tasks in S^c also follow the linear model (2), then we have the following lower bound, where $\mathbb{P}$ is the probability measure on sample space of all tasks:

$$
\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^{T}} \sup_{S \subseteq \mathcal{S}} \sup_{\substack{\{\boldsymbol{\beta}^{(t)}\}_{t \in S} \in \mathscr{B}(S,h) \\ \{\boldsymbol{\beta}^{(t)}\}_{t \in S^c}}} \mathbb{P}\left(\max_{t \in [T]} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \gtrsim \sqrt{\frac{p + \log T}{n}} \right) \geq \frac{1}{10}.
$$

Similar to the upper bound in Theorem 2, the lower bound of $\max_{t \in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ contains several terms reflecting the difficulty of learning different components. For example, $\frac{\zeta^{(t)}}{\bar{\zeta}} \sqrt{\frac{pr}{nT}} + \zeta^{(t)} h$ arises from learning the similar representations in S; $\left(\frac{\zeta^{(t)}}{\bar{\zeta}} \vee 1 \right) \sqrt{\frac{r}{n}} + \sqrt{\frac{\log T}{n}}$ is due to learning the task-specific parameters; $\frac{\zeta^{(t)}}{\bar{\zeta}} \frac{\epsilon r}{\sqrt{n}}$ is caused by outlier tasks; and $\sqrt{\frac{p + \log T}{n}}$ is the single-task rate.

To our knowledge, this is the first lower bound for learning regression parameters in the context of representation MTL. Tripuraneni et al. (2021) and Duchi et al. (2022) derived lower bounds for the subspace recovery when $\zeta^{(t)} \lesssim 1$ for all $t \in S = [T]$, with no outlier tasks ($\epsilon = 0$) and all tasks sharing the same representation ($h = 0$).

Comparing the upper bound for penalized ERM in Theorem 2 with the lower bound, the upper bound exhibits a sub-optimal dependence on r compared to the information-theoretic lower bound. This phenomenon has been noted in Du et al. (2020) and Tripuraneni et al. (2021) for both the ERM estimator and a method-of-moments estimator. The upper bounds of estimation errors for both estimators have sub-optimal dependence on r. They related this to a similar phenomenon observed in other works on linear regression models (Raskutti et al., 2011), where the upper bounds of estimation errors have sub-optimal dependence on eigenvalues of design matrices. In addition, the last term in our lower bound, $\frac{\zeta^{(t)}}{\bar{\zeta}} \frac{\epsilon r}{\sqrt{n}}$, does not depend on the full dimension p, whereas the counterpart in the upper bound does. A similar phenomenon has been noted in several papers (e.g. Tian et al., 2022; Duan and Wang, 2023). An open question is whether the dependence on p can be removed. Unfortunately, our ongoing work shows that this is impossible for the penalized ERM method like Algorithm 1 for a broad class of commonly used regularizers. As pointed out by Tian et al. (2022), estimators based on techniques in robust statistics like Tukey's depth function have been shown to achieve minimax rate under Huber's contamination model for location and covariance estimation (Chen et al., 2018), which might help improve the upper bound

in our setting. In summary, when r is bounded, $\epsilon = 0$, and $\zeta^{(t)} \lesssim \bar{\zeta}$ for all $t \in S = [T]$, the penalized ERM is optimal. An adaptation of techniques in robust statistics to improve our estimation algorithm will be an interesting future research direction.

On the other hand, the upper bound for the spectral method in Theorem 7 matches the lower bound when $\epsilon = 0$ and the condition number $\sigma_{\max}((\overline{\boldsymbol{A}}^{\perp})^{\top} \boldsymbol{B}_S^*)/\sigma_{\min}((\overline{\boldsymbol{A}}^{\perp})^{\top} \boldsymbol{B}_S^*)$ in Theorem 7 is bounded by a constant. To our knowledge, when there is no contamination ($\epsilon = 0$), no algorithm in the literature could achieve the optimal estimation error rate of $\boldsymbol{\beta}^{(t)*}$'s before our work, even in the special case $h = 0$. This demonstrates the power of the spectral method in the representation MTL. However, the term related to the outlier tasks does not involve n and is not optimal. It remains unknown what algorithm can achieve the optimal estimation error rate when $\epsilon \neq 0$.

Finally, we would like to emphasize that the condition $\epsilon r \leq c$ for some constant $c > 0$ is necessary for representation multi-task learning. As discussed at the end of the proof of Theorem 8 in Section D.9 of the appendix, when $\epsilon r > 1$, the lower bound immediately deteriorates to $\sqrt{p/n}$. This reduces the multi-task rate to the single-task rate, thereby eliminating the benefits of data integration in the worst-case scenario.

2.5 Extensions to Generalized Linear Models and Non-linear Regression

We can extend the proposed methods and theoretical framework beyond the linear model (2). For example, we can consider generalized linear models (GLMs) (McCullagh and Nelder, 1989), where the conditional distribution of Y given $X = \boldsymbol{x}$ for task t has density

$$p(y_i^{(t)} = y | \boldsymbol{x}_i^{(t)} = \boldsymbol{x}) = \rho(y) \exp \left\{ y \cdot \boldsymbol{x}^{\top} \boldsymbol{\beta}^{(t)*} - \psi(\boldsymbol{x}^{\top} \boldsymbol{\beta}^{(t)*}) \right\}, \quad i = 1 : n, \tag{4}$$

for $t \in S$, w.r.t. some measure μ on a subset of $\mathbb{R}$, where ψ is second-order continuously differentiable on $\mathbb{R}$, and ψ' is often called the inverse link function.

In addition to GLMs, we can also extend the linear model (2) to a non-linear regression model (Yang et al., 2015), where

$$y_i^{(t)} = g\big((\boldsymbol{x}_i^{(t)})^{\top} \boldsymbol{\beta}^{(t)*}\big) + \epsilon_i^{(t)}, \quad i = 1 : n, \tag{5}$$

for $t \in S$, where g is a monotone function with a continuous second-order derivative on $\mathbb{R}$ and $\{\epsilon_i^{(t)}\}_{i=1}^n$ are i.i.d. zero-mean sub-Gaussian variables independent of $\{\boldsymbol{x}_i^{(t)}\}_{i=1}^n$. In literature, g is often referred to as the link function.

Given the extensive coverage in the main text, we defer the details of these extensions to Section A of the appendix.

3. Adaptation to Unknown Intrinsic Dimension r

In Sections 2 and A, the intrinsic dimension r is assumed to be known a priori. To our knowledge, this assumption is standard in almost all related theoretical literature on representation multi-task and transfer learning (e.g., Ando et al., 2005; Maurer et al., 2016; Du et al., 2020; Thekumparampil et al., 2021; Tripuraneni et al., 2021; Chua et al., 2021; Collins et al., 2021; Deng et al., 2022; Duchi et al., 2022; Duan and Wang, 2023), despite being potentially unrealistic in practice. In this section, we propose a simple yet effective

algorithm to adapt the previous MTL algorithms to the case of an unknown r. Recall the notations $\zeta^{(t)} = \|\boldsymbol{\theta}^{(t)*}\|_2$ and $\bar{\zeta} = \sqrt{|S|^{-1}\sum_{t\in S}(\zeta^{(t)})^2}$. Similar to the notations used in Algorithm 2, we use $\prod_R$ to denote the projection operator to an ℓ_2-ball centered at zero of radius R in $\mathbb{R}^p$, and use $\texttt{quantile}(\{a_t\}_{t=1}^T, 1-\bar{\epsilon})$ to denote the lower $(1-\bar{\epsilon})$-quantile of $\{a_t\}_{t=1}^T \subseteq \mathbb{R}$, where $\bar{\epsilon} \in [0,1]$.

The algorithm is based on SVD, and the details are summarized in Algorithm 3.

Algorithm 3: Adaptation to unknown intrinsic dimension r

    Input: Data from tasks $\{\boldsymbol{X}^{(t)}, \boldsymbol{Y}^{(t)}\}_{t=1}^T = \{\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n\}_{t=1}^T$, threshold parameters
        $T_1, T_2 > 0$, an upper bound $\bar{\epsilon}$ (for ϵ)

    Output: An estimate $\hat{r}$

1 <u>Step 1:</u> (Single-task regression) $\widetilde{\boldsymbol{\beta}}^{(t)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{f^{(t)}(\boldsymbol{\beta})\}$ for $t \in [T]$

2 <u>Step 2:</u> (Projection and concatenation) Create a $p \times T$ matrix $\widehat{\boldsymbol{B}}$ whose t-th column
    is $\prod_R(\widetilde{\boldsymbol{\beta}}^{(t)})$, where $R = \texttt{quantile}(\{\|\widetilde{\boldsymbol{\beta}}^{(t)}\|_2\}_{t=1}^T, 1-\bar{\epsilon})$

3 <u>Step 3:</u> (Thresholding) Set $\hat{r} = \max\left\{r' \in [T] : \sigma_{r'}(\widehat{\boldsymbol{B}}/\sqrt{T}) \geq T_1\sqrt{\frac{p+\log T}{n}} + T_2 R\sqrt{\bar{\epsilon}}\right\}$

Algorithm 3 leverages Assumption 2 to determine an appropriate value for r. The underlying rationale is that when h is small, a significant spectral gap often exists between the r-th largest singular value and the $(r+1)$-th largest singular value of $\boldsymbol{B}_S^*/\sqrt{T}$. For example, when $h = 0$, $\boldsymbol{B}_S^*$ is a rank-r matrix, implying that $\sigma_{r'}(\boldsymbol{B}_S^*/\sqrt{T}) = 0$ for $r' \geq r+1$. Therefore, thresholding on singular values of an empirical version of $\boldsymbol{B}_S^*$ can be an effective strategy to estimate r. In Section 2.3, a similar approach based on SVD of the estimated $\boldsymbol{B}^*$ was used to develop the spectral method. In fact, Steps 1-2 of Algorithms 2 and 3 are the same, which construct a robust estimate of $\boldsymbol{B}_S^*$ through projected single-task estimators. Algorithm 3 is also conceptually similar to the thresholding method often used to determine the intrinsic dimension in principal component analysis (Onatski, 2010; Fan et al., 2021).

Under almost identical assumptions imposed in previous sections, with proper choices of tuning parameters, Algorithm 3 is shown to be consistent in estimating the true intrinsic dimension r, when representation matrices are similar (i.e., h is small). As we will elaborate, this suffices to ensure the same upper bounds of estimation error for Algorithms 1 and 2 when r is unknown.

Theorem 9 (Consistency of the intrinsic dimension estimation) *Suppose we choose an $\bar{\epsilon}$ such that $\epsilon \leq \bar{\epsilon} \leq \frac{c}{r}\left(\frac{\bar{\zeta}}{\max_{t\in S}\zeta^{(t)}}\right)^2$ with a small constant $c > 0$. Assume $\min_{t\in S}\zeta^{(t)} \cdot h\left[\frac{\sigma_{\max}((\overline{\boldsymbol{A}}^\perp)^\top\boldsymbol{B}_S^*)}{\sigma_{\min}((\overline{\boldsymbol{A}}^\perp)^\top\boldsymbol{B}_S^*)} \wedge \sqrt{r}\right] \leq c'\sqrt{\frac{p+\log T}{n}}$ with a small constant $c' > 0$, where $\boldsymbol{B}_S^* \in \mathbb{R}^{p\times|S|}$ is the coefficient matrix whose columns are $\{\boldsymbol{\beta}^{(t)*}\}_{t\in S}$, $\overline{\boldsymbol{A}} \in \arg\min_{\boldsymbol{A}\in\mathcal{O}^{p\times r}}\max_{t\in[T]}\|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \boldsymbol{A}\boldsymbol{A}^\top\|_2$ is the central representation, and $\overline{\boldsymbol{A}}^\perp$ is orthogonal to $\overline{\boldsymbol{A}}$ in the sense that $\overline{\boldsymbol{A}}^\perp \in \mathcal{O}^{p\times(p-r)}$ and $(\overline{\boldsymbol{A}}^\perp)^\top\overline{\boldsymbol{A}} = \boldsymbol{0}_{(p-r)\times r}$. Further assume that:* [i]

    (i) For the linear model (2), Assumptions 1, 2, and 5 hold;

---

i. Assumptions 4, 5, 6, and 7 for GLMs and non-linear regression models are presented in Section A of the appendix.

(ii) For the GLM (4), Assumptions 1, 2, 4, and 5 hold;

(iii) For the non-linear regression model (5), Assumptions 1, 2, 6, 7 hold;

Then there exist constants $T_1, T_2 > 0$ such that the output of Algorithm 3 satisfies $\hat{r} = r$ w.p. at least $1 - e^{-C'(p+\log T)}$ with some constant $C' > 0$.

Remark 10 *When $\min_{t \in S} \zeta^{(t)} \cdot h\left[\frac{\sigma_{\max}((\bar{\boldsymbol{A}}^\perp)^\top \boldsymbol{B}_S^*)}{\sigma_{\min}((\bar{\boldsymbol{A}}^\perp)^\top \boldsymbol{B}_S^*)} \wedge \sqrt{r}\right] \gtrsim \sqrt{\frac{p+\log T}{n}}$ in the MTL problem, estimating r becomes unnecessary as the upper bounds of MTL estimation errors are dominated by the single-task rate $\sqrt{\frac{p+\log T}{n}}$ for both Algorithms 1 and 2. In such cases, single-task learning is sufficient to achieve the minimax rate. Specifically, by the proofs of these upper bounds, the single-task rate is always guaranteed by biased regularization in Step 2 of Algorithm 1 and Step 4 of Algorithm 2, regardless of the performance achieved in other steps.*

According to Theorem 9 and Remark 10, when r is unknown, we can first run Algorithm 3 to obtain an estimate $\hat{r}$, then run Algorithms 1 and 2 with $\hat{r}$. All the previous results remain valid. This confirms that our full procedure is **adaptive** to an unknown intrinsic dimension r.

Similar to Algorithm 2, Algorithm 3 also requires $\bar{\epsilon}$ as an upper bound of the proportion of outlier tasks ϵ. If we have prior knowledge of ϵ (but not S and S^c), we can set $\bar{\epsilon} = \epsilon$. Without such information, a small constant such as 0.05 can be chosen for $\bar{\epsilon}$. Similar to Algorithm 2, we set $\bar{\epsilon}$ to the true ϵ value in simulations, while setting $\bar{\epsilon} = 0.05$ in the real-data study. We set the tuning parameters to $T_1 = 0.5$ and $T_2 = 0.25$ in the numerical experiments. Generally, T_1, T_2 can be chosen through cross-validation.

Before concluding this section, we would like to highlight that, in addition to the thresholding method, other approaches have been proposed for selecting the number of factors in factor models, which may also be beneficial in our context. For example, information criterion-based methods (Bai and Ng, 2002; Bunea et al., 2011) provide an alternative approach. Algorithm 3 is naturally motivated by the singular value gap condition in Assumption 2. However, we believe that an information criterion can also be developed to consistently select r, leveraging the estimation error bounds we have established. Furthermore, this selection process can be framed as a hypothesis testing problem (Onatski, 2009). While all these methods appear promising, it remains unclear how to effectively address heterogeneity across tasks and mitigate task contamination. We leave these challenges for future exploration.

4. Numerical Experiments

To validate the theoretical insights discussed in previous sections, we conducted extensive simulations and one real-data study, and the results are presented in this section.

All the experiments were implemented in Python. For penalized ERM ("pERM", Algorithm 1), we used the automatic differentiation implemented in PyTorch (Paszke et al., 2019) along with the Adam optimizer (Kingma and Ba, 2015) to solve the optimization problem (3) in Step 1. We set the learning rate equal to 0.01 in `torch.optim.Adam` function and kept all the other parameter choices as in default. Step 2 of pERM and Step 4 of the

spectral method ("Spectral", Algorithm 2) were also solved by the Adam optimizer with a learning rate 0.01. Consistent with our theory, we set penalty parameters $\lambda = \sqrt{r(p + \log T)}$ and $\gamma = \sqrt{p + \log T}$ in pERM, and $\gamma = 0.5\sqrt{p + \log T}$ in the spectral method. As mentioned in Section 2.3, Spectral requires an upper bound $\bar{\epsilon}$ of the contamination proportion ϵ. We set $\bar{\epsilon} = \epsilon$ in simulations and $\bar{\epsilon} = 0.05$ in the real-data study. Besides pERM and Spectral, we included the following approaches as benchmarks.

- Empirical risk minimization ("ERM") in Du et al. (2020); Tripuraneni et al. (2021): The optimization was also solved by the Adam solver in PyTorch with a learning rate 0.01;

- Method-of-moments ("MoM") in Tripuraneni et al. (2021);

- Adaptive representation learning ("AdaptRep") in Chua et al. (2021): We used the code included in the original paper and kept all the parameter settings as default.

- Adaptive and robust multi-task learning ("ARMUL") in Duan and Wang (2023): We used the code included in their paper (`https://github.com/kw2934/ARMUL`) and retained all default parameter settings. The tuning parameters were chosen using 5-fold cross-validation as default.

- Group Lasso ("GLasso") in Yuan and Lin (2006); Lounici et al. (2009, 2011): We grouped the same coordinate of coefficients from different tasks and applied an $L_{2,1}$-matrix penalty. The method was implemented in an R package `RMTL` (Cao et al., 2019), and we used the Python package `rpy2` to call functions `cv.MTL` and `MTL` in the R package `RMTL`. The penalty parameter was chosen by a 5-fold cross-validation as default.

- Data pooling or pooled regression ("Pooled") in Crammer et al. (2008); Ben-David et al. (2010): We fitted the linear regression and logistic regression models on the pooled data from all tasks. Both models were implemented in the Python module `sklearn.linear_models`.

- Single-task regression (Single-task): In simulations and the real-data study, we run single-task linear regression and logistic regression on the data of each task, respectively.

In Section 4.1, we present the performance of different approaches under different simulation settings, such as different heterogeneity parameter h (Section 4.1.1), different contamination proportion ϵ (Section 4.1.2), and different number of tasks T (Section 4.1.3). We also change the full dimension p, the intrinsic dimension r, and the per-task sample size n from setting to setting. To verify the intuition we obtained in Section 2, where we mentioned that the performance of each task in ERM, pERM, and Spectral depends on the signal strength $\zeta^{(t)} = \|\boldsymbol{\theta}^{(t)*}\|_2$, we conduct a simulation with different $\|\boldsymbol{\theta}^{(t)*}\|_2$ values across tasks, and the results are presented in Section 4.1.4. Finally, in Section 4.1.5, we evaluate Algorithm 3 for estimating r, compare the performance of pERM and Spectral with the estimated r and the true r, and demonstrate the effectiveness of Algorithm 3.

In Section 4.2, we compare the performance of different approaches on a real dataset.

The code to reproduce the results is available at `https://github.com/ytstat/RL-MTL-TL`.

4.1 Simulations

4.1.1 Simulation with Different Heterogeneity Parameter h

In this subsection, we investigate the linear model (1) with various values of h. We explore four distinct settings with different combinations of (n, p, r, T):

(i) $n = 100, p = 30, r = 5, T = 50$;

(ii) $n = 100, p = 50, r = 5, T = 50$;

(iii) $n = 100, p = 80, r = 5, T = 50$;

(iv) $n = 150, p = 80, r = 10, T = 50$.

In all settings, no outlier tasks are included, implying $S = [T]$ and $\epsilon = |S^c|/T = 0$. Given each (n, p, r, T) combination, we generated $\boldsymbol{x}_i^{(t)}$ i.i.d. from $N(\boldsymbol{0}_p, \boldsymbol{I}_p)$, $\epsilon_i^{(t)}$ i.i.d. from $N(0, 1)$, and a random $p \times r$ matrix $\boldsymbol{C}$ with i.i.d. standard normal entries. We defined $\overline{\boldsymbol{A}}$ as the first r columns of the left singular matrix of $\boldsymbol{C}$, $\widetilde{\boldsymbol{A}}^{(t)} = \overline{\boldsymbol{A}} + a^{(t)}(\boldsymbol{I}_{r \times r}, \boldsymbol{0}_{r \times (p-r)})^\top$, and $\boldsymbol{A}^{(t)*} = \widetilde{\boldsymbol{A}}^{(t)}[(\widetilde{\boldsymbol{A}}^{(t)})^\top \widetilde{\boldsymbol{A}}^{(t)}]^{-1}(\widetilde{\boldsymbol{A}}^{(t)})^\top$ for $t \in [T]$, where $a^{(t)}$'s are i.i.d. sampled from Unif($[-h, h]$). We generated each coordinate of $\boldsymbol{\theta}^{(t)*} \in \mathbb{R}^r$ from Unif($[-2, 2]$) independently. We considered h from 0 to 0.8 in increments of 0.1 and replicated each setting 100 times.

Figure 5 presents the simulation results. Across all settings, pooled regression and MoM perform worse than or on par with single-task regression, with their performance deteriorating as h increases. AdaptRep exhibits slightly better performance than single-task regression initially, but declines as h increases. GLasso performs similarly to single-task regression in settings (i), (ii), and (iv), while notably improving upon it in setting (iii). ERM significantly enhances single-task regression performance when h is small, but performs worse than single-task regression for large h. In contrast, ARMUL, pERM, and Spectral can improve the performance of single-task regression when h is small, and their performance will be comparable to single-task regression for large h.

As one reviewer pointed out, Spectral performs worse than pERM when h is small, particularly when p/n is large, despite having a sharper theoretical estimation upper bound. There are several possible explanations for this phenomenon. First, comparing the estimation errors in Theorems 2 and 7, Spectral outperforms pERM only in the term that depends on the similarity level h. When h is small, this advantage may be overshadowed by other terms, making the benefit of Spectral less apparent. However, as h increases, Spectral can catch up and eventually surpass pERM in performance. Second, our theoretical analysis primarily focuses on convergence rates while ignoring constant factors, which can significantly impact practical performance. Third, our analysis does not explicitly account for the effect of p/n on Spectral's performance, although we believe it plays a crucial role, particularly when p/n is large. In such cases, replacing OLS with single-task ridge regression may improve performance. On the other hand, as discussed at the end of Section 2.2.1, Spectral requires a more stringent sample size condition, namely $n \gtrsim p + \log T$, compared to $n \gtrsim r + \log T$ for pERM when h is small. An even stronger requirement, $n \gtrsim r^2(p + \log T)$, is needed for ARMUL. A more refined theoretical analysis in the proportional regime $p/n \to \gamma$ with some constant $\gamma > 0$ could provide deeper insights into the behaviors of Spectral and ARMUL. We leave these investigations for future work.
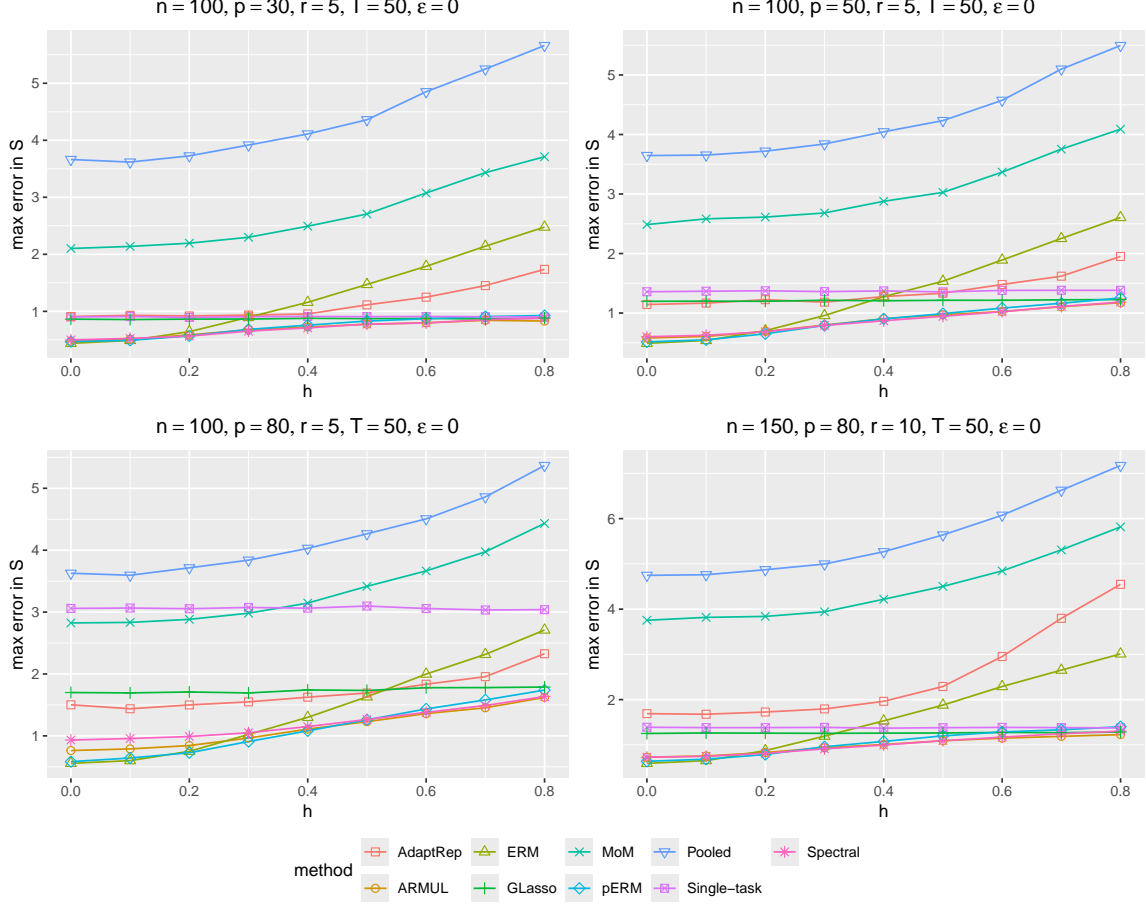
Figure 5: Simulation with different heterogeneity parameters h: estimation error $\max_{t \in [T]} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ of different algorithms with different (n, p, r, T) settings. "max error in S" in the y-axis stands for $\max_{t \in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ with $S = [T]$. Each point represents the average over 100 replications.

4.1.2 SIMULATION WITH DIFFERENT CONTAMINATION PROPORTION ϵ

In this subsection, we considered the linear model (1) with varying contamination proportions $\epsilon = |S^c|/T$. We explored two settings with different values of (n, p, r, T):

(i) $n = 100, p = 50, r = 5, T = 100$;

(ii) $n = 150, p = 80, r = 10, T = 100$.

For each replication in each setting, we randomly selected a subset of size $T(1 - \epsilon)$ from $[T]$ without replacement to form S. We set $h = 0$ and generated tasks in S using the same mechanism as in Section 4.1.1. The outlier tasks in S^c were generated by the linear model with $\boldsymbol{x}_i^{(t)}$ i.i.d. from $N(\boldsymbol{0}_p, 2\boldsymbol{I}_p)$ and each coordinate of the coefficient $\boldsymbol{\beta}^{(t)*}$ was generated i.i.d. from Unif$([-3, 3])$. We varied ϵ from 0 to 10% in increments of 2% and replicated each setting 100 times.
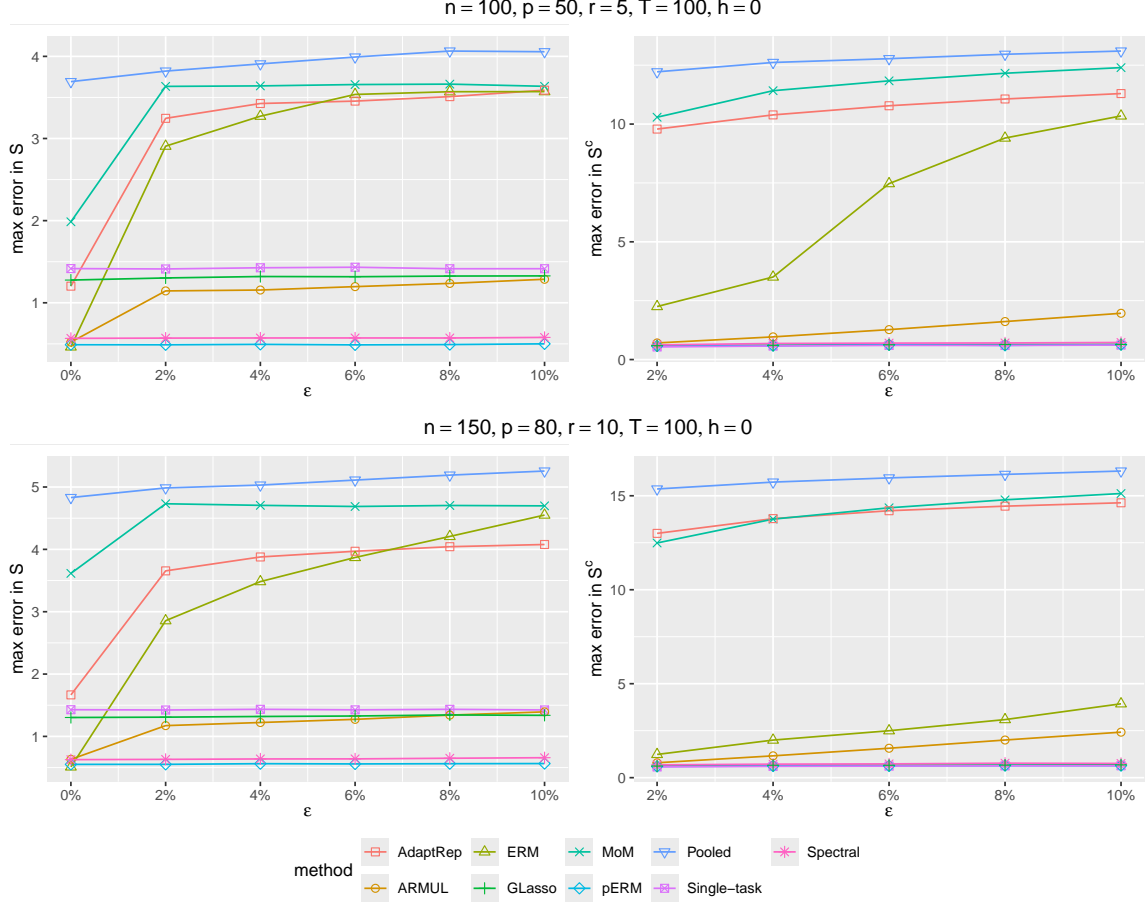
Figure 6: Simulation with different contamination proportions ϵ: within-S estimation error $\max_{t \in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ and outlier estimation error $\max_{t \in S^c} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ of different algorithms with different (n, p, r, T) settings. "max error in S" and "max error in S^c" in the y-axis stand for $\max_{t \in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ and $\max_{t \in S^c} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$, respectively. Each point represents the average over 100 replications.

The results are summarized in Figure 6, where we evaluated each method by both $\max_{t \in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ and $\max_{t \in S^c} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$. As the outlier proportion ϵ increases, the performance of most algorithms deteriorates rapidly. Even with just 2% outlier tasks, all methods perform similarly to or worse than single-task regression on tasks in S, except for pERM and Spectral. This demonstrates the robustness of pERM and Spectral against outlier tasks. On the other hand, ARMUL does perform worse than single-task regression, even when ϵ is large. Since the outlier tasks were also generated from the linear model, our theory guarantees that pERM and Spectral can match the single-task performance on outlier tasks in S^c, which is indeed observed. In contrast, ARMUL performs much worse on outlier tasks in S^c compared to single-task regression.

4.1.3 SIMULATION WITH DIFFERENT NUMBER OF TASKS T

In this subsection, we explore the impact of the number of tasks T on the performance of different methods in this subsection. We considered the linear model (1) in two settings, with different values of (n, p, r, ϵ):

(i) $n = 100, p = 50, r = 5, \epsilon = 0$;

(ii) $n = 100, p = 50, r = 5, \epsilon = 4\%$.

For each (n, p, r, ϵ) setting and T value, we set $h = 0$ and generated data in the same way as in Section 4.1.2. We increased T from 10 to 190 with increments of 15. The performance of different methods on tasks in S is summarized in Figure 7.
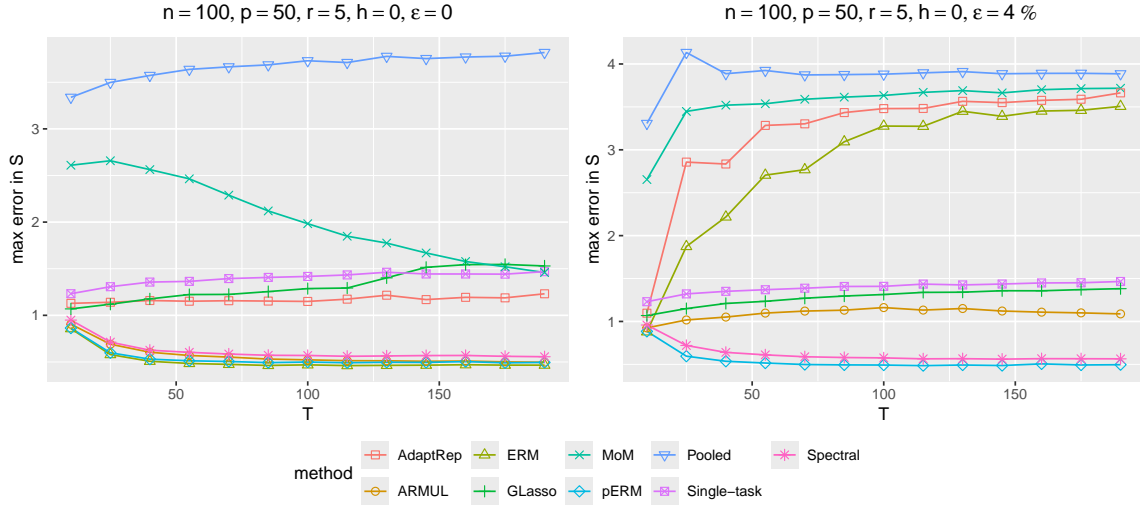
Figure 7: Simulation with different number of tasks T: estimation error $\max_{t \in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ of different algorithms with different (n, p, r, T) settings. "max error in S" in the y-axis stands for $\max_{t \in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$. Each point represents the average over 100 replications.

In the absence of outlier tasks, ERM, pERM, ARMUL, and Spectral exhibit comparable performance. As T increases, their estimation errors first decrease gradually and then stabilize, aligning with the theoretical result. For example, when $\zeta^{(t)} = \|\boldsymbol{\theta}^{(t)*}\|_2 \lesssim 1$ for all $t \in S = [T]$, $p \gtrsim r^2$, and $h = \epsilon = 0$, Theorem 2 implies that $\max_{t \in [T]} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim r\sqrt{\frac{p}{nT}} + r\sqrt{\frac{1}{n}}$, up to logarithmic factors with high probability. When T becomes large, the error is dominated by the second term $r\sqrt{\frac{1}{n}}$, which is independent of T. Besides ERM, pERM, ARMUL, and Spectral, the performance of MoM also improves as T increases. Similar to our findings in Sections 4.1.1 and 4.1.2, MoM's performance improves only when T is sufficiently large. This aligns with the empirical observations in Tripuraneni et al. (2021), where they found that ERM always outperforms MoM until T is very large, and the underlying reason is unclear.

When $\epsilon = 4\%$ tasks are contaminated, only the performance of pERM and Spectral improves as T increases, and they outperform all other methods, demonstrating their

robustness against outlier tasks. ARMUL and GLasso perform slightly better than single-task regression, while the other benchmark methods suffer from severe negative transfer and are significantly impacted by the outliers.
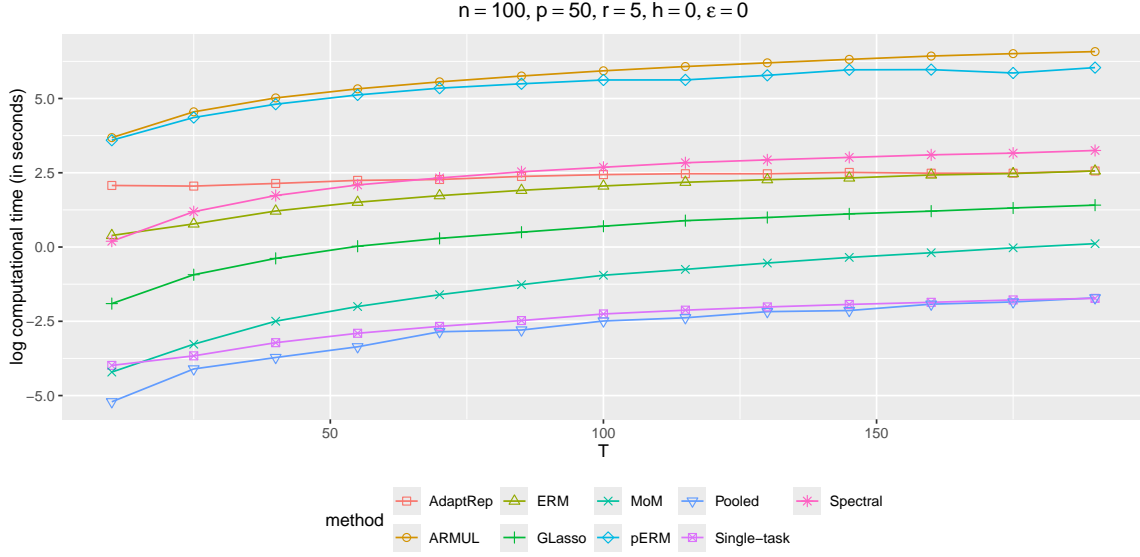
Figure 8: Computational time of different methods with various numbers of tasks T. The y-axis is in the log-scale, and the unit is seconds. Each point represents the average over 100 replications.

Finally, we recorded the computational time for different methods in setting (i). The experiments were run on the Terremoto HPC Cluster of Columbia University with a CPU Intel Xeon Gold 6126 2.6 GHz. We used a single core with 3 GB of memory when running each method. The computational time for different methods with different T values is plotted on a logarithmic scale in Figure 8. We can see that ARMUL and pERM are the most time-consuming methods, with pERM slightly faster than ARMUL, taking approximately $e^{6.25} \approx 500$ seconds for $T = 190$. In contrast, all the other methods can be run within $e^{3.5} \approx 30$ seconds for all values of T. This demonstrates the computational efficiency of the spectral method.

4.1.4 Relationship between Task Performance and Signal Strength $\|\boldsymbol{\theta}^{(t)*}\|_2$

In this subsection, we aim to verify our theoretical findings that the performance of ERM, pERM, and Spectral on each task can depend on the signal strength in terms of $\|\boldsymbol{\theta}^{(t)*}\|_2$, a relationship not previously discussed in the literature.

We generated data by a mechanism similar to that in Section 4.1.1, with $n = 100$, $p = 50$, $r = 5$, $T = 10$, $h = \epsilon = 0$. The only change in this section's data generation mechanism is that each $\boldsymbol{\theta}^{(t)*}$ is uniformly generated from $0.5t\mathcal{S}^{r-1}$, i.e., the sphere centered at $\mathbf{0}$ in $\mathbb{R}^r$ with a radius of $0.5t$. We replicated this setting 100 times and summarized the average estimation error of the single-task regression, ERM, pERM, and Spectral on each task with different $\|\boldsymbol{\theta}^{(t)*}\|_2$ values in Figure 9.
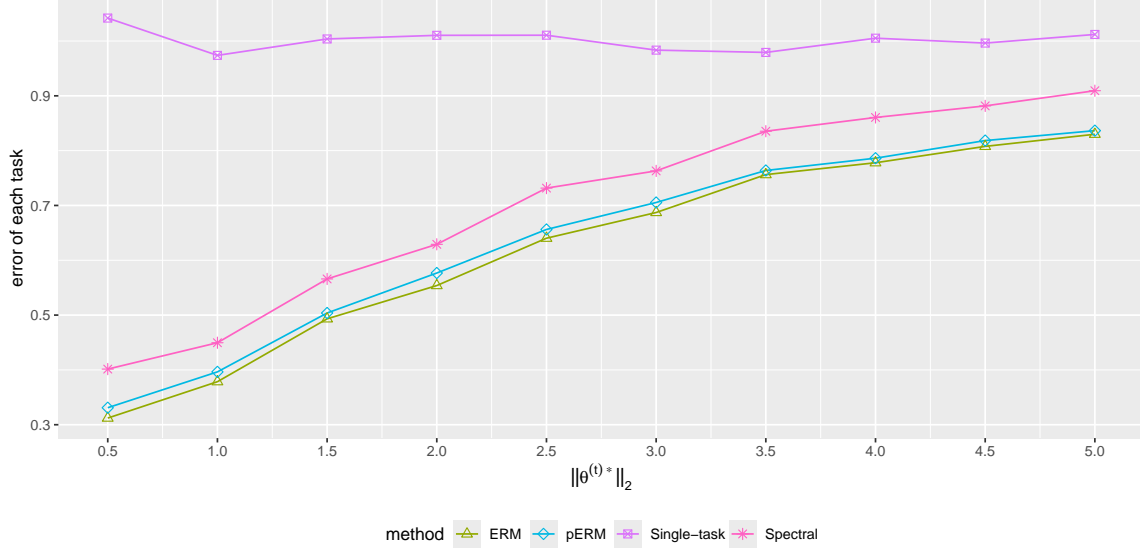
Figure 9: Simulation verifying the relationship between the estimation error $\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ and signal strength $\|\boldsymbol{\theta}^{(t)*}\|_2$ on each task for different methods. "error of each task" in the y-axis stands for $\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$. Each point represents the average over 100 replications.

We can observe that the estimation error of ERM, pERM, and Spectral is approximately proportional to $\|\boldsymbol{\theta}^{(t)*}\|_2$, which matches our theoretical findings. Specifically, the Pearson correlation coefficients between $\|\boldsymbol{\theta}^{(t)*}\|_2$ and the average $\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ for ERM, pERM, and Spectral are 0.977, 0.975, and 0.976, respectively. In contrast, single-task regression has comparable performance across different tasks, regardless of the signal strength $\|\boldsymbol{\theta}^{(t)*}\|_2$. This indicates that the benefit each individual task derives from representation MTL is highly dependent on the signal strength in terms of $\|\boldsymbol{\theta}^{(t)*}\|_2$, unlike in single-task linear regression where the ℓ_2-estimation error does not depend on the scale of the coefficient $\|\boldsymbol{\beta}^{(t)*}\|_2$.

4.1.5 ADAPTIVITY TO THE INTRINSIC DIMENSION r

In the previous simulations, we used the true intrinsic dimension r in different representation MTL methods. In this subsection, we want to test the performance of our Algorithm 3 for estimating r and how it enables pERM and Spectral to adapt to unknown r in practice.

We consider the linear model (1) under four settings, with different (n, p, r, T, ϵ) values:

(i) $n = 100, p = 50, r = 5, T = 50, \epsilon = 0$;

(ii) $n = 100, p = 50, r = 5, T = 50, \epsilon = 4\%$;

(iii) $n = 150, p = 80, r = 10, T = 50, \epsilon = 0$;

(iv) $n = 150, p = 80, r = 10, T = 50, \epsilon = 4\%$.

Given each (n, p, r, T, ϵ), we increased h from 0 to 0.8 in increments of 0.1, generated data for tasks in S following the same mechanism used in Section 4.1.1, and generated data for

tasks in S^c following the same mechanism used in Section 4.1.2. We first ran Algorithm 3 to obtain an estimate $\hat{r}$ of r, then ran pERM and Spectral with $\hat{r}$. We denote these versions as pERM-adaptive and Spectral-adaptive, respectively. We also ran pERM and Spectral with the true r value as benchmarks, and we call them pERM-oracle and Spectral-oracle, respectively. Each setting was replicated 100 times, and the average estimation error on tasks in S of different approaches, as well as the average $\hat{r}$, was plotted in Figure 10.
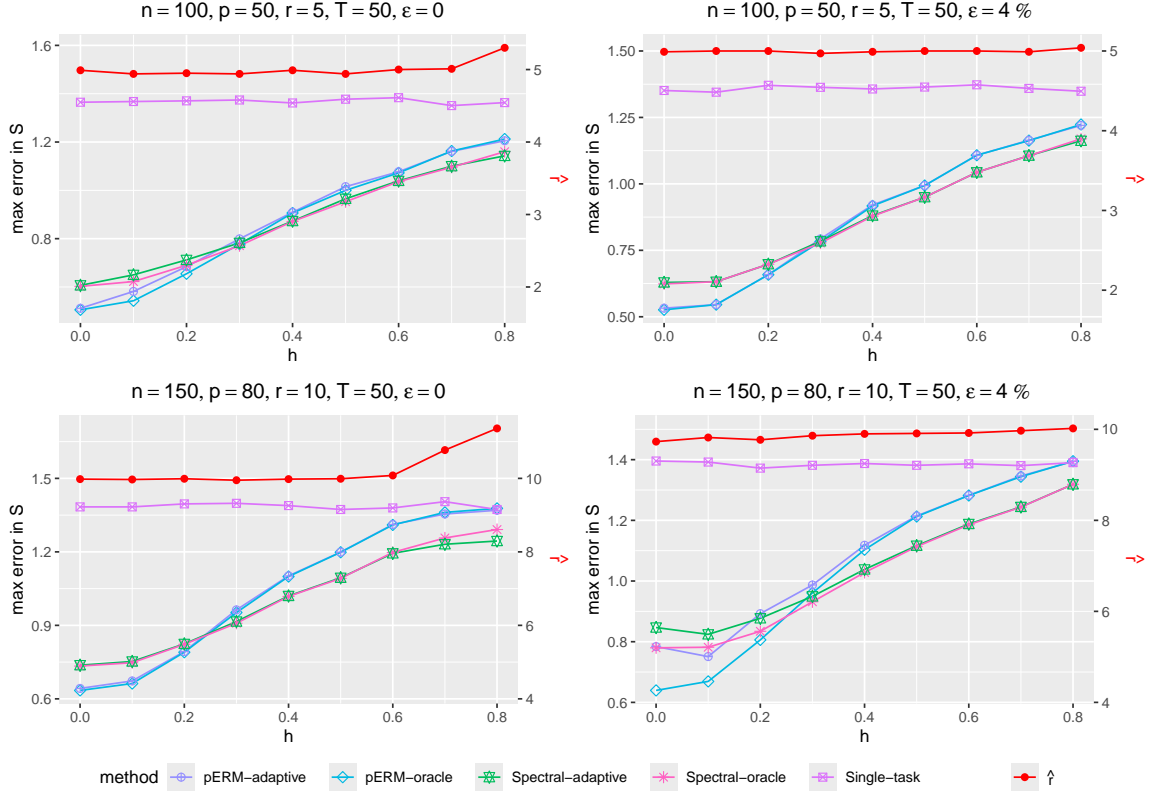
Figure 10: Simulation verifying the performance of Algorithm 3 for estimating r and its impact on helping pERM and Spectral adapt to unknown r. "max error in S" in the left y-axis stands for $\max_{t \in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$. The red $\hat{r}$ in the right y-axis stands for the estimated r value. Each point represents the average over 100 replications.

We can see that the estimate $\hat{r}$ is very close to the true r in all four settings when h is small, ensuring that pERM and Spectral using $\hat{r}$ perform comparably to their oracle counterparts using the true r. When h becomes large, the estimate $\hat{r}$ may deviate from the true r. However, as we mentioned in Remark 10 of Section 3, there is no need to estimate r precisely when h is large, and biased regularization in Step 2 of pERM and Step 4 of Spectral can always guarantee the single-task performance.

4.2 A Real-data Study

In this subsection, we applied different approaches to a real data set, Human Activity Recognition (HAR) Using a Smartphones Data Set. This data set includes data collected

from 30 volunteers performing six activities (walking, walking upstairs, walking downstairs, sitting, standing, and laying) with a smartphone (Anguita et al., 2013). Each observation has $p = 561$ time and frequency domain variables. We treated each volunteer as a task, with the sample size per task ranging from 281 to 409. The original data set is available at UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones`.

We focused on the binary classification problem of discriminating between walking and standing postures (i.e., walking, walking upstairs, walking downstairs, and standing) and the others (i.e., sitting and laying). We standardized the data of each task before training different algorithms. For each task, in each replication, we used 50% of the samples as training data and held 50% of the sample as test data.

We ran single-task logistic regression, pooled logistic regression, ERM, ARMUL, pERM, the spectral method, and GLasso on this problem, replicated it 100 times, and summarized the average misclassification test error rates of different methods over $T = 30$ tasks with different r values in Table 2. We can see that pERM consistently achieved the lowest error rate (tied with ERM when $r = 15$) among all the methods for different r values.

r/Method	Single-task	Pooled	ERM	ARMUL	pERM	Spectral	GLasso
$r = 5$	1.66 (0.20)	1.79 (0.21)	1.62 (1.42)	2.12 (0.27)	1.33 (0.23)	1.85 (0.27)	1.44 (0.25)
$r = 10$	1.66 (0.20)	1.79 (0.21)	1.42 (0.23)	1.77 (0.23)	1.25 (0.20)	1.47 (0.18)	1.44 (0.25)
$r = 15$	1.66 (0.20)	1.79 (0.21)	1.36 (0.23)	1.68 (0.21)	1.36 (1.07)	1.50 (0.19)	1.44 (0.25)

Table 2: The average mis-classification test error rates (standard deviations) of different methods over $T = 30$ tasks with different r values. All values are in percentages.

5. Discussions

In this work, we investigated the representation multi-task learning (MTL) problem, where most tasks share *similar* linear representations, and a small fraction of tasks can be *arbitrarily* contaminated. To address this problem, we proposed a penalized empirical risk minimization (ERM) method and a spectral method, and derived upper bounds for the estimation error. Our theory demonstrated that both algorithms are *adaptive* to the unknown similarity level between tasks and *robust* to a small fraction of outlier tasks. Additionally, the spectral method achieves a sharper estimation error bound than the penalized ERM when there is no contamination. Our theory also reveals the relationship between the performance of representation MTL methods on each task and the signal strength, which is usually ignored in the literature. We also presented the first lower bound results for estimating regression coefficients in the context of representation MTL. Our new spectral method is minimax optimal when there is no outlier task, and our penalized ERM is nearly optimal in a large regime with little impact from outlier tasks. We extended the algorithms and theory to generalized linear models and non-linear regression models. We also proposed a simple thresholding algorithm to adapt our MTL algorithms to the case of an unknown intrinsic dimension r. Finally, we conducted extensive numerical experiments to empirically validate our theoretical findings.

A recent paper (Niu et al., 2024) conducts a sophisticated and sharp analysis for the case where $h = \epsilon = 0$, accurately characterizing how the estimation error could depend on the singular values of $\boldsymbol{B}_S^* = (\boldsymbol{\beta}^{(1)*} \ \ldots \ \boldsymbol{\beta}^{(T)*}) \in \mathbb{R}^{p \times T}$. It would be interesting to explore how their analysis extends to two of our proposed algorithms in the more general regime considered in our study.

Acknowledgments

References

D. Alistarh, Z. Allen-Zhu, and J. Li. Byzantine stochastic gradient descent. *Advances in neural information processing systems*, 31, 2018.

R. K. Ando, T. Zhang, and P. Bartlett. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(11), 2005.

D. Anguita, A. Ghio, L. Oneto, X. Parra, J. Reyes-Ortiz, et al. A public domain dataset for human activity recognition using smartphones. In *21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 437–442. CIACO, 2013.

J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

M. Bakhshizadeh, A. Maleki, and V. H. de la Pena. Sharp concentration results for heavy-tailed distributions. *arXiv preprint arXiv:2003.13819*, 2020.

S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

H. Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984, 2021.

J. Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12: 149–198, 2000.

S. Ben-David and R. S. Borbely. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine learning*, 73:273–287, 2008.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

F. Bunea, Y. She, and M. H. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282, 2011.

T. T. Cai and H. Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1), 2021.

T. T. Cai, Z. Ma, and Y. Wu. Sparse pca: Optimal rates and adaptive estimation. *Annals of Statistics*, 41(6):3074–3110, 2013.

T. T. Cai, J. Ma, and L. Zhang. Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality 1. *Annals of Statistics*, 47(3):1234–1267, 2019.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

H. Cao, J. Zhou, and E. Schwarz. Rmtl: an r library for multi-task learning. *Bioinformatics*, 35(10):1797–1798, 2019.

A. Chen and Y. Feng. GeoERM: Geometry-aware multi-task representation learning on riemannian manifolds. *arXiv preprint arXiv:2505.02972*, 2025.

M. Chen, C. Gao, and Z. Ren. Robust covariance and scatter matrix estimation under huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

S. Chen, K. Crammer, H. He, D. Roth, and W. J. Su. Weighted training for cross-task learning. *arXiv preprint arXiv:2105.14095*, 2021a.

X. Chen, M. Chen, C. Fan, A. Uppunda, Y. Sun, and C. Zaniolo. Multilingual knowledge graph completion via ensemble knowledge transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3227–3238, 2020.

Y. Chen, L. Su, and J. Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017.

Y. Chen, Y. Chi, J. Fan, C. Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021b.

K. Chua, Q. Lei, and J. D. Lee. How fine-tuning allows for effective meta-learning. *Advances in Neural Information Processing Systems*, 34:8871–8884, 2021.

L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pages 2089–2099. PMLR, 2021.

K. Crammer, M. Kearns, and J. Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(8), 2008.

G. Denevi, M. Pontil, and C. Ciliberto. The advantage of conditional meta-learning for biased regularization and fine tuning. *Advances in Neural Information Processing Systems*, 33:964–974, 2020.

S. Deng, Y. Guo, D. Hsu, and D. Mandal. Learning tensor representations for meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 11550–11580. PMLR, 2022.

J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR, 2014.

D. Donoho and A. Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166:935–969, 2016.

S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Y. Duan and K. Wang. Adaptive and robust multi-task learning. *The Annals of Statistics*, 51(5):2015–2039, 2023.

J. Duchi, V. Feldman, L. Hu, and K. Talwar. Subspace recovery from heterogeneous data with non-isotropic noise. *arXiv preprint arXiv:2210.13497*, 2022.

C. Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

J. Fan, K. Wang, Y. Zhong, and Z. Zhu. Robust high dimensional factor models with applications to statistical machine learning. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 36(2):303, 2021.

C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

C. Finn, A. Rajeswaran, S. Kakade, and S. Levine. Online meta-learning. In *International Conference on Machine Learning*, pages 1920–1930. PMLR, 2019.

I. Gannaz. Robust estimation and wavelet thresholding in partially linear models. *Statistics and Computing*, 17:293–310, 2007.

P. Goyal, D. Mahajan, A. Gupta, and I. Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the ieee/cvf International Conference on computer vision*, pages 6391–6400, 2019.

T. Gu, P. H. Lee, and R. Duan. Commute: Communication-efficient transfer learning for multi-site risk prediction. *Journal of Biomedical Informatics*, 137:104243, 2023.

T. Gu, Y. Han, and R. Duan. Robust angle-based transfer learning in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae111, 2024.

R. Guerraoui, N. Gupta, and R. Pinot. Byzantine machine learning: A primer. *ACM Computing Surveys*, 56(7):1–39, 2024.

A. Gupta, C. Devin, Y. Liu, P. Abbeel, and S. Levine. Learning invariant feature spaces to transfer skills with reinforcement learning. *arXiv preprint arXiv:1703.02949*, 2017.

S. Hanneke and S. Kpotufe. On the value of target data in transfer learning. *Advances in Neural Information Processing Systems*, 32, 2019.

S. Hanneke and S. Kpotufe. A no-free-lunch theorem for multitask learning. *The Annals of Statistics*, 50(6):3119–3143, 2022.

W. Härdle and T. M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association*, 84(408):986–995, 1989.

T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

M. Hristache, A. Juditsky, and V. Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001.

A. Jalali, S. Sanghavi, C. Ruan, and P. Ravikumar. A dirty model for multi-task learning. *Advances in neural information processing systems*, 23, 2010.

T. Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.

D. P. Kingma and J. L. Ba. Adam: A method for stochastic gradient descent. In *ICLR: international conference on learning representations*, pages 1–15. ICLR US., 2015.

W. Kong, R. Somani, Z. Song, S. Kakade, and S. Oh. Meta-learning for mixed linear regression. In *International Conference on Machine Learning*, pages 5394–5404. PMLR, 2020.

N. Konstantinov, E. Frantar, D. Alistarh, and C. Lampert. On the sample complexity of adversarial multi-source pac learning. In *International Conference on Machine Learning*, pages 5416–5425. PMLR, 2020.

S. Kpotufe and G. Martinet. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021.

I. Kuzborskij and F. Orabona. Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pages 942–950. PMLR, 2013.

I. Kuzborskij and F. Orabona. Fast rates by transferring from auxiliary hypotheses. *Machine Learning*, 106:171–195, 2017.

S. Li, T. Cai, and R. Duan. Targeting underrepresented populations in precision medicine: A federated transfer learning approach. *arXiv preprint arXiv:2108.12112*, 2021.

S. Li, T. T. Cai, and H. Li. Transfer learning in large-scale gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association*, pages 1–13, 2022a.

S. Li, T. T. Cai, H. Li, et al. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B*, 84(1):149–173, 2022b.

H. Lin and M. Reimherr. On transfer learning in functional linear regression. *arXiv preprint arXiv:2206.04277*, 2022.

G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):171–184, 2012.

P.-L. Loh and M. J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16: 559–616, 2015.

K. Lounici, M. Pontil, A. Tsybakov, and S. Van De Geer. Taking advantage of sparsity in multi-task learning. In *COLT 2009-The 22nd Conference on Learning Theory*, 2009.

K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.

G. Lugosi and S. Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021.

A. Maurer, M. Pontil, and B. Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*, volume 37. CRC Press, 1989.

D. Meunier, Z. Li, A. Gretton, and S. Kpotufe. Nonlinear meta-learning can guarantee faster rates. *arXiv preprint arXiv:2307.10870*, 2023.

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

A. Nichol, J. Achiam, and J. Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

X. Niu, L. Su, J. Xu, and P. Yang. Collaborative learning with shared linear representations: Statistical rates and optimal algorithms. *arXiv preprint arXiv:2409.04919*, 2024.

A. Onatski. Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77(5):1447–1479, 2009.

A. Onatski. Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016, 2010.

A. Pajor. Metric entropy of the grassmann manifold. *Convex Geometric Analysis*, 34:181–188, 1998.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

M. Qiao. Do outliers ruin collaboration? In *International Conference on Machine Learning*, pages 4180–4187. PMLR, 2018.

M. Qiao and G. Valiant. Learning discrete distributions from untrusted batches. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

M. Rostami, H. He, M. Chen, and D. Roth. Transfer learning via representation learning. In *Federated and Transfer Learning*, pages 233–257. Springer International Publishing Cham, 2022.

A. M. Samarov. Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association*, 88(423):836–847, 1993.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pages 71–79. PMLR, 2013.

Y. She and A. B. Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011.

K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Statistically and computationally efficient linear meta-representation learning. *Advances in Neural Information Processing Systems*, 34:18487–18500, 2021.

Y. Tian and Y. Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, pages 1–14, 2022.

Y. Tian, H. Weng, L. Xia, and Y. Feng. Unsupervised multi-task and transfer learning on gaussian mixture models. *arXiv preprint arXiv:2209.15224*, 2022.

Y. Tian, H. Weng, and Y. Feng. Towards the theory of unsupervised federated learning: Non-asymptotic analysis of federated em algorithms. *Proceedings of Machine Learning Research*, 235:48226–48279, 2024.

N. Tripuraneni, M. Jordan, and C. Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.

N. Tripuraneni, C. Jin, and M. Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer, New York, 2009.

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

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

R. Vidal, Y. Ma, S. S. Sastry, R. Vidal, Y. Ma, and S. S. Sastry. Principal component analysis. *Generalized principal component analysis*, pages 25–62, 2016.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

A. Wang, J. Hula, P. Xia, R. Pappagari, R. T. McCoy, R. Patel, N. Kim, I. Tenney, Y. Huang, K. Yu, et al. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476, 2019.

P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Advances in neural information processing systems*, 22, 2009.

H. Xu, C. Caramanis, and S. Sanghavi. Robust pca via outlier pursuit. *IEEE Transactions on Information Theory*, 58(5):3047–3064, 2012.

K. Xu and H. Bastani. Learning across bandits in high dimension via robust statistics. *arXiv preprint arXiv:2112.14233*, 2021.

Z. Yang, Z. Wang, H. Liu, Y. C. Eldar, and T. Zhang. Sparse nonlinear regression: Parameter estimation and asymptotic inference. *arXiv preprint arXiv:1511.04514*, 2015.

Z. Yang, K. Balasubramanian, Z. Wang, and H. Liu. Learning non-gaussian multi-index model via second-order stein's method. *Advances in Neural Information Processing Systems*, 30:6097–6106, 2017.

D. Yin, Y. Chen, R. Kannan, and P. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, pages 5650–5659. Pmlr, 2018.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

G. Yuan, M. Xu, S. Kpotufe, and D. Hsu. Efficient estimation of the central mean subspace via smoothed gradient outer products. *arXiv preprint arXiv:2312.15469*, 2023.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

Y. Zhang and Q. Yang. An overview of multi-task learning. *National Science Review*, 5(1): 30–43, 2018.

Y. Zhang and Q. Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

D. Zhou, T. Cai, and J. Lu. Multi-source learning via completion of block-wise overlapping noisy matrices. *arXiv preprint arXiv:2105.10360*, 2021.

D. Zhou, M. Liu, M. Li, and T. Cai. Doubly robust augmented model accuracy transfer inference with high dimensional features. *arXiv preprint arXiv:2208.05134*, 2022.

B. Zhu, L. Wang, Q. Pang, S. Wang, J. Jiao, D. Song, and M. I. Jordan. Byzantine-robust federated learning with optimal statistical rates. In *International Conference on Artificial Intelligence and Statistics*, pages 3151–3178. PMLR, 2023.

Content of Appendices

Appendix A. Extensions to More General Models

In this section, we consider two extensions to the linear model (2).

A.1 Generalized Linear Models

A generalization to generalized linear models (GLMs) from the linear model (2) is as follows. Suppose that the conditional distribution of Y given $X = \boldsymbol{x}$ for task t is

$$p(y_i^{(t)} = y | \boldsymbol{x}_i^{(t)} = \boldsymbol{x}) = \rho(y) \exp\left\{ y \cdot \boldsymbol{x}^\top \boldsymbol{\beta}^{(t)*} - \psi(\boldsymbol{x}^\top \boldsymbol{\beta}^{(t)*}) \right\}, \quad i = 1 : n, \qquad (6)$$

for $t \in S$, w.r.t. some measure μ on a subset of $\mathbb{R}$, where ψ is second-order continuously differentiable on $\mathbb{R}$, and ψ' is often called the inverse link function. More discussions on GLMs can be found in McCullagh and Nelder (1989).

Example 1 *Some canonical examples of GLMs include:*

(i) *Linear models: $\psi(u) = \frac{1}{2}u^2$, $\psi'(u) = u$, and μ is Lebesgue measure;*

(ii) *Logistic regression models: $\psi(u) = u + \log(1 + e^{-u})$, $\psi'(u) = \frac{1}{1+e^{-u}}$, and μ is the counting measure on $\{0,1\}$;*

(iii) *Poisson regression models: $\psi(u) = e^u$, $\psi'(u) = e^u$, and μ is the counting measure on $\mathbb{N} = \{0, 1, 2, \ldots\}$;*

We replace the linear model (2) with the GLM (6) and keep all the other settings the same as in Section 2. Moreover, we impose the following extra conditions for GLMs.

Assumption 4 *ψ satisfies the following three conditions:*

(i) *ψ is strictly convex;*

(ii) *$\max_{t \in S} \max_{\|\boldsymbol{\Delta}\|_2 \leq 1} \psi''((\boldsymbol{\beta}^{(t)*})^\top \boldsymbol{x}^{(t)} + \boldsymbol{\Delta}^\top \boldsymbol{x}^{(t)}) \leq C$ a.s. with a constant $C > 0$;*

(iii) *One of the following conditions holds:*

    (a) *There exists a large constant $C' > 0$ such that $\min_{t \in S} \min_{|u| \leq C'\zeta^{(t)}} \psi''(u) \geq C'$, where $C'' > 0$ is another universal constant.*

    (b) *$\min_{t \in S} \min_{\|\boldsymbol{\Delta}\|_2 \leq 1} \psi''((\boldsymbol{\beta}^{(t)*})^\top \boldsymbol{x}^{(t)} + \boldsymbol{\Delta}^\top \boldsymbol{x}^{(t)}) \geq C'''$ a.s., where $C''' > 0$ is a constant.*

Assumption 5 *$n \geq Cr(p + \log T)$ with a sufficiently large constant $C > 0$.*

Assumption 4 is commonly used in the non-asymptotic analysis of GLMs (e.g., see Negahban et al., 2012; Loh and Wainwright, 2015). The sample size requirement in Assumption 5 is more stringent than that for linear regression, serving as a technical condition needed for our proof. Specifically, the Hessian matrices of GLMs may not exhibit good spectral controls when evaluated far from the true parameter value. To ensure that our estimators avoid these problematic regions, we require the penalty term in Step 1 of the penalized ERM method in Algorithm 1 (which is upper bounded by $\lambda/\sqrt{n}$) to remain reasonably small. For more details, see the proof of Lemma 51 in Appendix. Note that for the spectral method in Algorithm 2, $n \gtrsim p + \log T$ in Assumption 3 is sufficient. For simplicity, we use the stronger condition in Assumption 4 for both methods.

It is noteworthy that the function ψ in (6) is allowed to be different across tasks. Here we assume ψ is the same for different tasks for simplicity. Additionally, we assume $\sqrt{r} \gtrsim \zeta^{(t)} \gtrsim 1$ for all $t \in S$. For the GLM (6), we apply Algorithms 1 and 2 with $f^{(t)}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n [-y_i^{(t)} \cdot (\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta} + \psi((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta})]$ for $\boldsymbol{\beta} \in \mathbb{R}^p$. With these GLM assumptions in place, Algorithms 1 and 2 achieve the same upper bounds of estimation error as in linear regression models.

Theorem 11 (Upper bound for MTL under GLMs) *Suppose Assumptions 4, 5 and other conditions imposed in Theorems 2 and 7 hold, the same high-probability upper bounds in Theorems 2 and 7 hold for GLMs.*

A.2 Non-linear Regression

In addition to GLMs, we can extend the linear model (2) to a non-linear regression model as follows. Suppose

$$y_i^{(t)} = g\big((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*}\big) + \epsilon_i^{(t)}, \quad i = 1 : n, \tag{7}$$

for $t \in S$, where g is a monotone function with a continuous second-order derivative on $\mathbb{R}$ and $\{\epsilon_i^{(t)}\}_{i=1}^n$ are i.i.d. zero-mean sub-Gaussian variables independent of $\{\boldsymbol{x}_i^{(t)}\}_{i=1}^n$. In literature, g is often referred to as the link function. More discussions for this model under a single-task learning setting can be found in Yang et al. (2015). Note that Yang et al. (2015) considered the case of a fixed design while we considered the random design case, which is more challenging. Hence, stronger conditions are necessary to guarantee the desired rate.

We replace the linear model (2) with the non-linear regression model (7), and keep all the other settings the same as in Section 2. Furthermore, we impose the following assumptions for non-linear regression models. Recall the notation $\boldsymbol{\Sigma}^{(t)} = \mathbb{E}[\boldsymbol{x}^{(t)}(\boldsymbol{x}^{(t)})^T]$ and denote $\boldsymbol{\Sigma}_\epsilon^{(t)} = \mathbb{E}[\boldsymbol{\epsilon}^{(t)}(\boldsymbol{\epsilon}^{(t)})^T]$.

Assumption 6 $0 < C_1 \leq [g'(u)]^2 \leq C_2 < \infty$, for all $u \in \mathbb{R}$, and $\|g''\|_\infty \leq C_3 < \infty$, where $C_3 \leq cC_1^2 \cdot \min_{t \in S}\{\lambda_{\min}(\boldsymbol{\Sigma}^{(t)})/\lambda_{\max}(\boldsymbol{\Sigma}^{(t)})\} \min_{t \in S}\{\lambda_{\max}^{-1/2}(\boldsymbol{\Sigma}_\epsilon^{(t)})\}$ with a small constant c.

Assumption 7 $n \geq C[(p + \log T)^{3/2}] \vee [r(p + \log T)]$ with a sufficiently large constant C.

The sample size requirement is stronger than that for linear regression. The first term $(p + \log T)^{3/2}$ arises due to the heavy-tailed distributions appearing in the analysis. A similar requirement for the high-dimensional sparse non-linear regression can be found in Yang et al. (2015). The reason for the term $r(p + \log T)$ is the same as in Section A.1 for GLMs. More details can be found in the proof of Lemma 59 in Appendix.

Similar to the case of GLMs, the link function g in (7) is also allowed to vary across tasks. Here we assume g is the same for different tasks for simplicity. We further assume $\sqrt{r} \gtrsim \zeta^{(t)} \gtrsim 1$ for all $t \in S$. We apply Algorithms 1 and 2 for the non-linear regression model (7), by setting $f^{(t)}(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^n [y_i^{(t)} - g((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta})]^2$ for $\boldsymbol{\beta} \in \mathbb{R}^p$. We have the same upper bounds of estimation error for Algorithms 1 and 2 as in linear regression models.

Theorem 12 (Upper bound for MTL under non-linear regression models) *Suppose Assumptions 6, 7 and other conditions imposed in Theorems 2 and 7 hold, then the same high-probability upper bounds in Theorems 2 and 7 hold for non-linear regression models.*

Appendix B. Transferring to New Tasks (Learning-to-learn)

In this section, we extend the MTL framework discussed in the main text to a transfer learning (TL) setting.

B.1 Problem Set-up

In this section, in addition to data from the T tasks, suppose we also observe data $\{(\boldsymbol{x}_i^{(0)}, y_i^{(0)})\}_{i=1}^{n_0}$ from a new task

$$y_i^{(0)} = (\boldsymbol{x}_i^{(0)})^\top \boldsymbol{\beta}^{(0)*} + \epsilon_i^{(0)}, \quad i = 1 : n_0, \tag{8}$$

where $\boldsymbol{\beta}^{(0)*} = \boldsymbol{A}^{(0)*}\boldsymbol{\theta}^{(0)*}$, $\boldsymbol{A}^{(0)*} \in \mathcal{O}^{p \times r} = \{\boldsymbol{A} \in \mathbb{R}^{p \times r} : \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_r\}$, $\boldsymbol{\theta}^{(0)*} \in \mathbb{R}^r$, and $\{\epsilon_i^{(0)}\}_{i=1}^{n_0}$ are i.i.d. zero-mean sub-Gaussian variables independent of $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^{n_0}$. Similar to Section 2, the intrinsic dimension r is assumed to be known. Under such a transfer learning (TL) or learning-to-learn setting, the new task is often called the *target* task, and the T tasks are called *source* tasks. Our goal is two-fold:

1. Transfer knowledge from source tasks to improve the learning performance on the *target* task, when the source and target share "similar" representations and the number of outlier source tasks is small;

2. Ensure the learning performance is no worse than the target-only learning performance to avoid the negative transfer.

To describe the similarity between target and source representations, we assume

$$\max_{t \in S} \|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \boldsymbol{A}^{(0)*}(\boldsymbol{A}^{(0)*})^\top\|_2 \leq h,$$

where S is a subset of $[T]$. Similar to the setting in the last section, the joint distribution $\mathbb{Q}_{S^c}$ of data from source tasks in S^c, i.e., $\{\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n\}_{t \in S^c}$, is allowed to be arbitrary. For simplicity, we focus on the regime that $\|\boldsymbol{\theta}^{(t)*}\|_2 \leq C < \infty$ for all $t \in S$. Denote $\boldsymbol{\Sigma}^{(0)} = \mathbb{E}[\boldsymbol{x}^{(0)}(\boldsymbol{x}^{(0)})^\top]$. We impose the following assumptions on the target task.

Assumption 8 *For $\boldsymbol{u} \in \mathbb{R}^p$, $\boldsymbol{u}^\top \boldsymbol{x}^{(0)}$ is sub-Gaussian in the sense that $\mathbb{E}[e^{\lambda \boldsymbol{u}^\top \boldsymbol{x}^{(0)}}] \leq e^{C\lambda^2\|\boldsymbol{u}\|_2^2}$ for any $\lambda \in \mathbb{R}$ with some constant $C > 0$. And there exist constants c, C such that $0 < c \leq \lambda_{\min}(\boldsymbol{\Sigma}^{(0)}) \leq \lambda_{\max}(\boldsymbol{\Sigma}^{(0)}) \leq C < \infty$.*

Assumption 9 *$n_0 \geq Cp$ with a sufficiently large constant $C > 0$.*

Remark 13 *Similar to our discussion in Section 2.2.1 for the penalized ERM in MTL, Assumption 9 is imposed to guarantee the target-only rate $\sqrt{p/n_0}$. If we do not care about this safe-net guarantee, then it suffices to require $n_0 \geq Cr$, the same as the condition imposed in literature (Tripuraneni et al., 2021; Thekumparampil et al., 2021). The RHS of the following Theorem 14 shall be replaced by $r\sqrt{\frac{p}{nT}} + \sqrt{r}h + \sqrt{r}\sqrt{\frac{r + \log T}{n}} + \sqrt{\frac{p}{n}} \cdot \frac{|S^c|}{T}r^{3/2} + \sqrt{\frac{r}{n_0}}$, which allows us for a few-shot learning when both h and $|S^c|/T$ are sufficiently small. See more details in the proof of Theorem 14 in Appendix.*

B.2 Upper Bounds

Similar to the Algorithm 1, a two-step transfer learning method is proposed in Algorithm 4. We introduce the algorithm with a general loss function $f^{(0)}$ for the target since the same algorithm with different losses will be extended to other models later. For linear model (8), define $f^{(0)}(\boldsymbol{\beta}) = \frac{1}{2n_0}\|\boldsymbol{Y}^{(0)} - \boldsymbol{X}^{(0)}\boldsymbol{\beta}\|_2 = \frac{1}{2n_0}\sum_{i=1}^{n_0}[y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^\top\boldsymbol{\beta}]^2$ for $\boldsymbol{\beta} \in \mathbb{R}^p$, where $\boldsymbol{X}^{(0)} \in \mathbb{R}^{n_0 \times p}$ and $\boldsymbol{Y}^{(0)} \in \mathbb{R}^{n_0}$ are corresponding matrix/vector representations of target data.

In Algorithm 4, the "central representation" $\widehat{\overline{\boldsymbol{A}}}$ learned by Algorithm 1 is passed to Step 1 to obtain the estimator $\widehat{\boldsymbol{\theta}}^{(0)}$ of the target-specific low-dimensional parameter. The same

---

Algorithm 4: Transferring to new tasks

Input: Data from a new task $(\boldsymbol{X}^{(0)}, \boldsymbol{Y}^{(0)}) = \{\boldsymbol{x}_i^{(0)}, y_i^{(0)}\}_{i=1}^{n_0}$, estimator $\widehat{\overline{\boldsymbol{A}}}$ from
       Algorithm 1 or 2, penalty parameter γ

Output: Estimator $\widehat{\boldsymbol{\beta}}^{(0)}$

1 <u>Step 1:</u> $\widehat{\boldsymbol{\theta}}^{(0)} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^r} \{f^{(0)}(\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta})\}$

2 <u>Step 2:</u> $\widehat{\boldsymbol{\beta}}^{(0)} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \{f^{(0)}(\boldsymbol{\beta}) + \frac{\gamma}{\sqrt{n_0}}\|\boldsymbol{\beta} - \widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\theta}}^{(0)}\|_2\}$

---

step has appeared in literature when there are no outliers, and the target and source share the same representations (Du et al., 2020; Tripuraneni et al., 2021). Step 2 is similar to Step 2 of Algorithm 1, which guarantees the target-only rate even when the representations of the target and sources are dissimilar.

We have the following upper bounds of target estimation error for Algorithm 4.

Theorem 14 (Upper bound for TL) *Suppose Assumptions 1, 2, 3, 8, and 9 hold.*

(i) (TL with penalized ERM) By setting $\lambda = Cr^{3/4}\sqrt{p + \log T}$ and $\gamma = C'\sqrt{p + \log T}$ with sufficiently large constants $C, C' > 0$ in Algorithm 4 along with Algorithm 1 to learn $\widehat{\overline{\boldsymbol{A}}}$, for all $S \subseteq [T]$ satisfying $\epsilon = \frac{|S^c|}{T} \leq cr^{-3/2}$ with a small constant $c > 0$ and an arbitrary distribution $\mathbb{Q}_{S^c}$ of $\{\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n\}_{t \in S^c}$, w.p. at least $1 - e^{-C''(r + \log T)}$, we have

$$\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)*}\|_2 \lesssim \left(r\sqrt{\frac{p}{nT}} + \sqrt{r}h + \sqrt{r}\sqrt{\frac{r + \log T}{n}} + \sqrt{\frac{p}{n}} \cdot \epsilon r^{3/2}\right) \wedge \sqrt{\frac{p}{n_0}} + \sqrt{\frac{r}{n_0}}.$$

(ii) (TL with the spectral method) By setting $\gamma = C'\sqrt{p + \log T}$ with a sufficiently large positive constant C', for any subset $S \subseteq [T]$ satisfying $\epsilon = \frac{|S^c|}{T} \leq cr^{-1} \cdot \left(\frac{\bar{\zeta}}{\max_{t \in S} \zeta^{(t)}}\right)^2$ with $c > 0$ a small constant, w.p. at least $1 - e^{-C''(r + \log T)}$ and $\bar{\epsilon}$ satisfying $\epsilon \leq \bar{\epsilon} \leq \frac{c''}{r} \cdot \left(\frac{\bar{\zeta}}{\max_{t \in S} \zeta^{(t)}}\right)^2$ with a sufficiently small positive constant c'', we have

$$\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)*}\|_2 \lesssim \left\{\sqrt{\frac{pr}{nT}} + \sqrt{\frac{r}{n}} + h \cdot \left[\frac{\sigma_{\max}((\overline{\boldsymbol{A}}^{\perp})^\top \boldsymbol{B}_S^*)}{\sigma_{\min}((\overline{\boldsymbol{A}}^{\perp})^\top \boldsymbol{B}_S^*)} \wedge \sqrt{r}\right] + \sqrt{r\bar{\epsilon}}\right\} \wedge \sqrt{\frac{p}{n_0}} + \sqrt{\frac{r}{n_0}},$$

where $\boldsymbol{B}_S^* \in \mathbb{R}^{p \times |S|}$ is the coefficient matrix whose columns are $\{\boldsymbol{\beta}^{(t)*}\}_{t \in S}$, $\overline{\boldsymbol{A}} \in \arg\min_{\boldsymbol{A} \in \mathcal{O}^{p \times r}} \max_{t \in [T]} \|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \boldsymbol{A}\boldsymbol{A}^\top\|_2$ is the central representation, and $\overline{\boldsymbol{A}}^{\perp}$ is orthogonal to $\overline{\boldsymbol{A}}$ in the sense that $\overline{\boldsymbol{A}}^{\perp} \in \mathcal{O}^{p \times (p-r)}$ and $(\overline{\boldsymbol{A}}^{\perp})^\top \overline{\boldsymbol{A}} = \boldsymbol{0}_{(p-r) \times r}$.

Both upper bounds can be seen as the minimum of two terms which represent the rate of learning target model via data aggregation and the target-only rate $\sqrt{\frac{p}{n_0}}$, respectively. This rate entails that our algorithm is **adaptive** to the optimal situation regardless of whether transferring from source to target is beneficial. Moreover, it is **robust** to a small fraction of outlier source tasks, in the sense that TL is still helpful when the outlier proportion ϵ is sufficiently small.

B.3 Lower Bounds

In this subsection, we explore the lower bound of the TL problem. Consider the space for all subsets $S \subseteq [T]$ as

$$\mathcal{S} = \{S \subseteq [T] : |S^c|/T \leq \epsilon\}.$$

Given the subset S, consider the parameter spaces for the coefficient vectors $\{\boldsymbol{\beta}^{(t)}\}_{t \in \{0\} \cup S}$ as

$$\mathscr{B}_0(S, h) = \Bigg\{ \{\boldsymbol{\beta}^{(t)}\}_{t \in \{0\} \cup S} : \boldsymbol{\beta}^{(t)} = \boldsymbol{A}^{(t)} \boldsymbol{\theta}^{(t)} \text{ for all } t \in \{0\} \cup S, \{\boldsymbol{A}^{(t)}\}_{t \in \{0\} \cup S} \subseteq \mathcal{O}^{p \times r},$$

$$\max_{t \in \{0\} \cup S} \|\boldsymbol{\theta}^{(t)}\|_2 \leq C, \max_{t \in S} \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \boldsymbol{A}^{(0)}(\boldsymbol{A}^{(0)})^\top\|_2 \leq h, \sigma_r\left(|S|^{-1/2} \boldsymbol{B}_S^*\right) \geq \frac{c}{\sqrt{r}} \Bigg\}$$

where C and c can be any fixed positive constants such that $\mathscr{B}_0(S, h) \neq \emptyset$.

Theorem 15 (Lower bound for TL) *Suppose $p \geq 2r$ and $\epsilon \leq c/r$ where c is a small constant. We have the following lower bound:*

$$\inf_{\widehat{\boldsymbol{\beta}}^{(0)}} \sup_{S \subseteq \mathcal{S}} \sup_{\{\boldsymbol{\beta}^{(t)}\}_{t \in \{0\} \cup S} \in \mathscr{B}_0(S, h)} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)*}\|_2 \gtrsim \left(\sqrt{\frac{pr}{nT}} + h \right) \wedge \sqrt{\frac{p}{n_0}} + \sqrt{\frac{r}{n_0}} + \frac{\epsilon r}{\sqrt{n}} \wedge \sqrt{\frac{1}{n_0}} \right) \geq \frac{1}{10}.$$

To our knowledge, this is the first lower bound for learning regression parameters under representation transfer learning. Comparing the upper and lower bounds of the representation TL problem, we can see that the upper bound of penalized ERM has suboptimal dependence on r and $\log T$. The spectral method has a sharper upper bound when $S = [T]$ (i.e. $\epsilon = 0$) with $\bar{\epsilon} = 0$, and is minimax optimal when the condition number $\frac{\sigma_{\max}((\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{B}_S^*)}{\sigma_{\min}((\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{B}_S^*)}$ is bounded and $n \gtrsim n_0$.

Similar to the MTL case, we can extend the upper bound from the linear model to the GLMs and non-linear regression models.

Theorem 16 (Upper bound for TL under GLMs) *Suppose Assumptions 4, 5 and other conditions imposed in Theorem 14 hold, then the same high-probability upper bounds in Theorem 14 hold for GLMs.*

Theorem 17 (Upper bound for TL under non-linear regression models) *Suppose Assumptions 6, 7, and other conditions imposed in Theorem 14 hold. Let $n_0 \geq Cp^{3/2}$ with $C > 0$ a sufficiently large constant. The same high-probability upper bounds in Theorem 14 hold for non-linear regression models.*

Appendix C. General Lemmas

C.1 Lemmas

Lemma 18 (Theorem 6.5 in Wainwright (2019)) *Suppose that Assumptions 1 and 3 hold. Then for any $\delta > 0$ and any $t \in [T]$, w.p. at least $1 - C_1 e^{-nC_2(\delta \wedge \delta^2)}$,*

$$\|\widehat{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma}^{(t)}\|_2 \leq C_3 \sqrt{\frac{p}{n}} + \delta,$$

48

with some constants $C_1, C_2, C_3 > 0$. Note that C_1, C_2, C_3 are universal in the sense that they do not depend on t or δ. As a consequence, we have

$$\max_{t \in [T]} \|\widehat{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma}^{(t)}\|_2 \lesssim \sqrt{\frac{p + \log T}{n}},$$

w.p. at least $1 - e^{-C(p + \log T)}$.

Lemma 19 (A variant of Theorem 6.5 in Wainwright (2019)) *Suppose that Assumptions 8 and 9 hold. Then for any $\delta > 0$, for any fixed $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{O}^{p \times r}$, w.p. at least $1 - C_1 e^{-n_0 C_2 (\delta \wedge \delta)}$,*

$$\|\boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{B} - \boldsymbol{A}^\top \boldsymbol{\Sigma}^{(t)} \boldsymbol{B}\|_2 \leq C_3 \sqrt{\frac{r}{n_0}} + \delta,$$

with some constants $C_1, C_2, C_3 > 0$. Note that C_1, C_2, C_3 are universal in the sense that they do not depend on δ or t. As a consequence, for any fixed $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{O}^{p \times r}$, we have

$$\|\boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(0)} \boldsymbol{B} - \boldsymbol{A}^\top \boldsymbol{\Sigma}^{(0)} \boldsymbol{B}\|_2 \lesssim \sqrt{\frac{r}{n_0}},$$

w.p. at least $1 - e^{-C(r + \log T)}$.

Lemma 20 (Lemmas 2.5 and 2.6 in Chen et al. (2021b)) *Suppose $p \geq r$. Consider two matrics $\boldsymbol{A}, \widetilde{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$. Suppose that $(\boldsymbol{A}, \boldsymbol{A}^\perp), (\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{A}}^\perp) \in \mathbb{R}^{p \times p}$ are both orthonormal matrices, which means that $\boldsymbol{A}^\perp$ and $\widetilde{\boldsymbol{A}}^\perp$ are orthonormal complements of $\boldsymbol{A}$ and $\widetilde{\boldsymbol{A}}$, respectively. Then*

$$\|\boldsymbol{A}\boldsymbol{A}^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2 = \|\boldsymbol{A}^\top(\widetilde{\boldsymbol{A}}^\perp)\|_2 = \|(\widetilde{\boldsymbol{A}})^\top \boldsymbol{A}^\perp\|_2,$$

$$\frac{1}{\sqrt{2}} \|\boldsymbol{A}\boldsymbol{A}^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_F = \|\boldsymbol{A}^\top(\widetilde{\boldsymbol{A}}^\perp)\|_F = \|(\widetilde{\boldsymbol{A}})^\top \boldsymbol{A}^\perp\|_F,$$

$$\|\boldsymbol{A}\boldsymbol{A}^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2 \leq \min_{\boldsymbol{R} \in \mathcal{O}^{r \times r}} \|\boldsymbol{A} - \widetilde{\boldsymbol{A}}\boldsymbol{R}\|_2 \leq \sqrt{2} \|\boldsymbol{A}\boldsymbol{A}^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2,$$

$$\frac{1}{\sqrt{2}} \|\boldsymbol{A}\boldsymbol{A}^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_F \leq \min_{\boldsymbol{R} \in \mathcal{O}^{r \times r}} \|\boldsymbol{A} - \widetilde{\boldsymbol{A}}\boldsymbol{R}\|_F \leq \|\boldsymbol{A}\boldsymbol{A}^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_F.$$

Lemma 21 (Proposition 8 in Pajor (1998)) *When $p \geq 2r$, for any $\delta > 0$, we have*

(i) $(C1 \frac{\sqrt{r}}{\delta})^{r(p-r)} \leq N(\mathcal{O}^{p \times r}, \text{dist}_F, \delta) \leq C_2(\frac{\sqrt{r}}{8})^{r(p-r)}$;

(ii) $(C_1 \frac{1}{\delta})^{r(p-r)} \leq N(\mathcal{O}^{p \times r}, \text{dist}_2, \delta) \leq (C_2 \frac{1}{\delta})^{r(p-r)}$;

(iii) *(Consequence of (i))* $M(\mathcal{O}^{p \times r}, \text{dist}_F, \delta) \geq C(\frac{\sqrt{r}}{\delta})^{r(p-r)}$.

Lemma 22 *When $p \geq 2r$, for any $\delta > 0$ and $\alpha \in (0, 1)$, $\exists \overline{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$, such that*

$$M(\mathcal{B}_\delta(\overline{\boldsymbol{A}}, \mathcal{O}^{p \times r}, \text{dist}_2), \text{dist}_F, \alpha\delta) \geq \left(\frac{C\sqrt{r}}{\alpha}\right)^{r(p-r)}.$$

Lemma 23 (Fano's lemma, see Tsybakov (2009) and Wainwright (2019)) *Suppose* (Θ, d) *is a metric space and each* θ *in this space is associated with a probability measure* $\mathbb{P}_\theta$. *If* $\{\theta_j\}_{j=1}^N$ *is an s-separated set (i.e.* $d(\theta_j, \theta_k) \geq s$ *for any* $j \neq k$*), and* $\mathrm{KL}(\mathbb{P}_{\theta_j}, \mathbb{P}_{\theta_k}) \leq \alpha \log N$, *then*

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta(d(\widehat{\theta}, \theta) \geq s/2) \geq 1 - \alpha - \frac{\log 2}{\log N}.$$

Lemma 24 *Consider the following generative model:*

$$y|\boldsymbol{x} \sim \mathbb{P}_{y|\boldsymbol{x}, \boldsymbol{\beta}} = N(\boldsymbol{x}^\top \boldsymbol{\beta}, 1), \quad \boldsymbol{x} \sim \mathbb{P}_{\boldsymbol{x}},$$

where $\mathbb{P}_{\boldsymbol{x}}$ *is sub-Gaussian with* $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top)$. *Suppose there exist constants* c *and* C *such that* $0 < c \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq C < \infty$. *Define two distributions* $\mathbb{P}$ *and* $\widetilde{\mathbb{P}}$ *of* $(\boldsymbol{x}, y)$ *as* $\mathbb{P}_{\boldsymbol{x}, y} = \mathbb{P}_{y|\boldsymbol{x}; \boldsymbol{\beta}} \cdot \mathbb{P}_{\boldsymbol{x}}$ *and* $\widetilde{\mathbb{P}}_{\boldsymbol{x}, y} = \mathbb{P}_{y|\boldsymbol{x}; \widetilde{\boldsymbol{\beta}}} \cdot \mathbb{P}_{\boldsymbol{x}}$. *Then, the KL divergence between* $\mathbb{P}_{\boldsymbol{x}, y}$ *and* $\widetilde{\mathbb{P}}_{\boldsymbol{x}, y}$ *can be bounded as*

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{x}, y} \| \widetilde{\mathbb{P}}_{\boldsymbol{x}, y}) \leq C \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_2^2.$$

Lemma 25 (Theorem 5.1 in Chen et al. (2018)) *Given a family of distributions* $\{\mathbb{P}_\theta : \theta \in \Theta\}$, *which is indexed by a parameter* $\theta \in \Theta$. *Consider* $\boldsymbol{x}^{(t)} \sim (1 - \epsilon')\mathbb{P}_\theta + \epsilon'\mathbb{Q}$ *independently for* $t \in [T]$, *and* $\widetilde{x} \sim \mathbb{P}_{\widetilde{\boldsymbol{x}}}$. *Denote the joint distribution of* $\{\boldsymbol{x}^{(t)}\}_{t=1}^T$ *and* $\widetilde{x}$ *as* $\mathbb{P}_{(\epsilon', \theta, \mathbb{Q})} \cdot \mathbb{P}_{\widetilde{\boldsymbol{x}}}$. *Then*

$$\inf_{\widehat{\theta}} \sup_{\substack{\theta \in \Theta \\ \mathbb{Q}}} (\mathbb{P}_{(\epsilon', \theta, \mathbb{Q})} \cdot \mathbb{P}_{\widetilde{\boldsymbol{x}}}) \left(\|\widehat{\theta} - \theta\| \geq C\varpi(\epsilon', \Theta) \right) \geq \frac{1}{2},$$

where $\varpi(\epsilon', \Theta) := \sup\{\|\theta_1 - \theta_2\| : \mathrm{TV}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) \leq \epsilon'/(1 - \epsilon'), \theta_1, \theta_2 \in \Theta\}$.

Lemma 26 (Lemma 22 in Tian et al. (2022)) *Consider two data generating mechanisms:*

(i) $\boldsymbol{x}^{(t)} \sim (1 - \epsilon')\mathbb{P}_\theta + \epsilon'\mathbb{Q}$ *independently for* $t \in [T]$, *where* $\epsilon' = \frac{|S^c|}{T}$, *and* $\widetilde{x} \sim \mathbb{P}_{\widetilde{\boldsymbol{x}}}$;

(ii) *With a preserved set* $S \subseteq [T]$, *generate* $\{\boldsymbol{x}^{(t)}\}_{t \in S^c} \sim \mathbb{Q}_{S^c}$ *and* $\boldsymbol{x}^{(t)} \sim \mathbb{P}_\theta$ *independently for* $t \in S$, *and* $\widetilde{x} \sim \mathbb{P}_{\widetilde{\boldsymbol{x}}}$.

Denote the joint distributions of $\{\boldsymbol{x}^{(t)}\}_{t=1}^T$ *and* $\widetilde{x}$ *in (i) and (ii) as* $\mathbb{P}_{(\epsilon, \theta, \mathbb{Q})} \cdot \mathbb{P}_{\widetilde{\boldsymbol{x}}}$ *and* $\mathbb{P}_{(S, \theta, \mathbb{Q}_{S^c})} \cdot \mathbb{P}_{\widetilde{\boldsymbol{x}}}$, *respectively. We claim that if*

$$\inf_{\widehat{\theta}} \sup_{\substack{\theta \in \Theta \\ \mathbb{Q}}} (\mathbb{P}_{(\epsilon'/50, \theta, \mathbb{Q})} \cdot \mathbb{P}_{\widetilde{\boldsymbol{x}}}) \left(\|\widehat{\theta} - \theta\| \geq C\varpi(\epsilon'/50, \Theta) \right) \geq \frac{1}{2}$$

then

$$\inf_{\widehat{\theta}} \sup_{S: |S| \geq T(1 - \epsilon')} \sup_{\substack{\theta \in \Theta \\ \mathbb{Q}_{S^c}}} (\mathbb{P}_{(S, \theta, \mathbb{Q}_{S^c})} \cdot \mathbb{P}_{\widetilde{\boldsymbol{x}}}) \left(\|\widehat{\theta} - \theta\| \geq C\varpi\left(\epsilon'/50, \Theta\right) \right) \geq \frac{1}{10},$$

where $\varpi(\epsilon', \Theta) := \sup\{\|\theta_1 - \theta_2\| : \mathrm{TV}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) \leq \epsilon'/(1 - \epsilon'), \theta_1, \theta_2 \in \Theta\}$.

Lemma 27 (Lemma 33 in Tian et al. (2022)) *Given a family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$, which is indexed by a parameter $\theta \in \Theta$. Consider $\boldsymbol{x}^{(t)} \sim (1-\epsilon')\mathbb{P}_\theta + \epsilon'\mathbb{Q}$ independently for $t \in [T]$, and $\widetilde{x} \sim \mathbb{P}_{\widetilde{x}}$. Consider another family of distributions $\{\mathbb{P}_\theta^{(0)} : \theta \in \Theta\}$ indexed by the same parameter set, and $\boldsymbol{x}^{(0)} \sim \mathbb{P}_\theta^{(0)}$. Denote the joint distribution of $\{\boldsymbol{x}^{(t)}\}_{t=0}^T$ and $\widetilde{x}$ as $\mathbb{P}_{(\epsilon',\theta,\mathbb{Q})} \cdot \mathbb{P}_{\widetilde{\boldsymbol{x}}}$. Then*

$$\inf_{\widehat{\theta}} \sup_{\substack{\theta \in \Theta \\ \mathbb{Q}}} (\mathbb{P}_{(\epsilon',\theta,\mathbb{Q})} \cdot \mathbb{P}_{\widetilde{\boldsymbol{x}}}) \left(\|\widehat{\theta} - \theta\| \geq C\varpi(\epsilon',\Theta) \right) \geq \frac{9}{20},$$

where $\varpi(\epsilon',\Theta) := \sup \left\{ \|\theta_1 - \theta_2\| : \mathrm{TV}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) \leq \epsilon'/(1-\epsilon'), \mathrm{TV}(\mathbb{P}_{\theta_1}^{(0)}, \mathbb{P}_{\theta_2}^{(0)}) \leq 1/20, \theta_1, \theta_2 \in \Theta \right\}$.

Lemma 28 (Lemma 34 in Tian et al. (2022)) *Consider two data generating mechanisms:*

(i) *$\boldsymbol{x}^{(t)} \sim (1 - \epsilon')\mathbb{P}_\theta + \epsilon'\mathbb{Q}$ independently for $t \in [T]$, $\boldsymbol{x}^{(0)} \sim \mathbb{P}_\theta^{(0)}$, and $\widetilde{x} \sim \mathbb{P}_{\widetilde{\boldsymbol{x}}}$, where $\epsilon' = \frac{|S^c|}{T}$;*

(ii) *With a preserved set $S \subseteq [T]$, generate $\{\boldsymbol{x}^{(t)}\}_{t \in S^c} \sim \mathbb{Q}_{S^c}$ and $\boldsymbol{x}^{(t)} \sim \mathbb{P}_\theta$ independently for $t \in S$, $\boldsymbol{x}^{(0)} \sim \mathbb{P}_\theta^{(0)}$, and $\widetilde{x} \sim \mathbb{P}_{\widetilde{\boldsymbol{x}}}$.*

Denote the joint distributions of $\{\boldsymbol{x}^{(t)}\}_{t=0}^T$ and $\widetilde{x}$ in (i) and (ii) as $\mathbb{P}_{(\epsilon,\theta,\mathbb{Q})} \cdot \mathbb{P}_{\widetilde{\boldsymbol{x}}}$ and $\mathbb{P}_{(S,\theta,\mathbb{Q}_{S^c})} \cdot \mathbb{P}_{\widetilde{\boldsymbol{x}}}$, respectively. We claim that if

$$\inf_{\widehat{\theta}} \sup_{\substack{\theta \in \Theta \\ \mathbb{Q}}} (\mathbb{P}_{(\epsilon'/50,\theta,\mathbb{Q})} \cdot \mathbb{P}_{\widetilde{\boldsymbol{x}}}) \left(\|\widehat{\theta} - \theta\| \geq C\varpi'\left(\epsilon'/50,\Theta\right) \right) \geq \frac{9}{20},$$

then

$$\inf_{\widehat{\theta}} \sup_{S : |S| \geq T(1-\epsilon')} \sup_{\substack{\theta \in \Theta \\ \mathbb{Q}_S}} (\mathbb{P}_{(S,\theta,\mathbb{Q}_S)} \cdot \mathbb{P}_{\widetilde{\boldsymbol{x}}}) \left(\|\widehat{\theta} - \theta\| \geq C\varpi'\left(\epsilon'/50,\Theta\right) \right) \geq \frac{1}{10},$$

where $\varpi(\epsilon',\Theta) := \sup \left\{ \|\theta_1 - \theta_2\| : \mathrm{TV}(\mathbb{P}_{\theta_1}\|\mathbb{P}_{\theta_2}) \leq \epsilon'/(1-\epsilon'), \mathrm{TV}(\mathbb{P}_{\theta_1}^{(0)}\|\mathbb{P}_{\theta_2}^{(0)}) \leq 1/20, \theta_1, \theta_2 \in \Theta \right\}$.

C.2 Proofs of Lemmas

Denote $\widehat{\boldsymbol{\Sigma}}^{(t)} = \sum_{i=1}^n \frac{1}{n} \boldsymbol{x}_i^{(t)} (\boldsymbol{x}_i^{(t)})^\top$.

C.2.1 PROOF OF LEMMA 19

Note that for any fixed $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{O}^{p \times r}$,

$$\|\boldsymbol{A}^\top(\widehat{\boldsymbol{\Sigma}}^{(0)} - \boldsymbol{\Sigma}^{(0)})\boldsymbol{B}\|_2 = \sup_{\|\boldsymbol{u}\|_2, \|\boldsymbol{v}\|_2 \leq 1} (\boldsymbol{A}\boldsymbol{u})^\top (\widehat{\boldsymbol{\Sigma}}^{(0)} - \boldsymbol{\Sigma}^{(0)})(\boldsymbol{B}\boldsymbol{v}),$$

where $\boldsymbol{A}\boldsymbol{u}, \boldsymbol{B}\boldsymbol{v} \in \mathbb{R}^p$ and $\|\boldsymbol{A}\boldsymbol{u}\|_2 = \|\boldsymbol{B}\boldsymbol{v}\|_2 = 1$. Note that both $\boldsymbol{A}\boldsymbol{u}$ and $\boldsymbol{B}\boldsymbol{v}$ live in r-dimensional space (isomorphic to the unit ball in $\mathbb{R}^r$). Therefore, by Example 5.8 in

Wainwright (2019), there exist two 1/8-covers (whose components are inside the set to be covered) of $\{\boldsymbol{\beta} \in \mathbb{R} : \boldsymbol{\beta} = \boldsymbol{Au}, \|\boldsymbol{u}\|_2 \le 1\}$ and $\{\boldsymbol{\beta} \in \mathbb{R} : \boldsymbol{\beta} = \boldsymbol{Bu}, \|\boldsymbol{u}\|_2 \le 1\}$ under Euclidean norm, denoted as $\{\boldsymbol{\beta}_j^{\boldsymbol{u}}\}_{j=1}^{N_1}$ and $\{\boldsymbol{\beta}_j^{\boldsymbol{u}}\}_{j=1}^{N_2}$, respectively, such that $|N_1| = |N_2| \le 17^r$. Then we can proceed in the same steps as in the proof of Theorem 6.5 in Wainwright (2019) to finish the proof.

C.2.2 Proof of Lemma 22

We apply the same trick as in the proof of Proposition 3 in Cai et al. (2013). Suppose the minimum cover of $\mathcal{O}^{p \times r}$ under $\mathrm{dist}_2(\cdot)$ corresponding to $N(\mathcal{O}^{p \times r}, \mathrm{dist}_2, \delta)$ is $\mathcal{N}$. By pigeonhole theorem, for any $\alpha \in (0, 1)$,

$$N(\mathcal{O}^{p \times r}, \mathrm{dist}_2, \alpha\delta) \le N(\cup_{\boldsymbol{A} \in \mathcal{N}} \mathcal{B}_\delta(\boldsymbol{A}, \mathcal{O}^{p \times r}, \mathrm{dist}_2), \mathrm{dist}_{\mathrm{F}}, \alpha\delta)$$
$$\le |\mathcal{N}| \max_{\boldsymbol{A} \in \mathcal{N}} N(\mathcal{B}_\delta(\boldsymbol{A}, \mathcal{O}^{p \times r}, \mathrm{dist}_2), \mathrm{dist}_{\mathrm{F}}, \alpha\delta).$$

Then applying Lemma 21 leads to

$$\max_{\boldsymbol{A} \in \mathcal{N}} N(\mathcal{B}_\delta(\boldsymbol{A}, \mathcal{O}^{p \times r}, \mathrm{dist}_2), \mathrm{dist}_{\mathrm{F}}, \alpha\delta) \ge \frac{N(\mathcal{O}^{p \times r}, \mathrm{dist}_2, \alpha\delta)}{N(\mathcal{O}^{p \times r}, \mathrm{dist}_2, \delta)} \ge \left(\frac{C\sqrt{r}}{\alpha}\right)^{r(p-r)}.$$

Since $\mathcal{N}$ has only finite elements, there must exist one $\boldsymbol{A} \in \mathcal{N}$ achieving the maximum of LHS. Because the packing number is always larger than or equal to the covering number, the proof is done.

C.2.3 Proof of Lemma 24

By the form of Gaussian density function, it is straightfoward to see that

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{x},y} \| \widetilde{\mathbb{P}}_{\boldsymbol{x},y}) = \int \log\left(\frac{\mathbb{P}_{y|\boldsymbol{x},\boldsymbol{\beta}}}{\mathbb{P}_{y|\boldsymbol{x},\widetilde{\boldsymbol{\beta}}}}\right) \mathrm{d}\mathbb{P}_{y|\boldsymbol{x},\boldsymbol{\beta}} \mathrm{d}\mathbb{P}_{\boldsymbol{x}}$$
$$\le \underbrace{\int (y - \boldsymbol{x}^\top \boldsymbol{\beta}) \cdot \boldsymbol{x}^\top (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) \mathrm{d}\mathbb{P}_{y|\boldsymbol{x},\boldsymbol{\beta}} \mathrm{d}\mathbb{P}_{\boldsymbol{x}}}_{=0} + \int [\boldsymbol{x}^\top (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})]^2 \mathrm{d}\mathbb{P}_{\boldsymbol{x}}$$
$$\lesssim \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_2^2.$$

Appendix D. Proofs for Linear Regression Models

Denote $\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} \in \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^r} f^{(t)}(\boldsymbol{A}\boldsymbol{\theta})$ and $f^{(t)}(\boldsymbol{\beta}) = \frac{1}{n}\|\boldsymbol{Y}^{(t)} - \boldsymbol{X}^{(t)}\boldsymbol{\beta}\|_2^2$. For $\boldsymbol{A}, \widetilde{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$, define metrics $\mathrm{dist}_2(\boldsymbol{A}, \widetilde{\boldsymbol{A}}) = \|\boldsymbol{A}\boldsymbol{A}^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2$, $\mathrm{dist}_{\mathrm{F}}(\boldsymbol{A}, \widetilde{\boldsymbol{A}}) = \|\boldsymbol{A}\boldsymbol{A}^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_{\mathrm{F}}$. In any metric space $(\mathcal{X}, \rho)$, denote the ball of radius δ with center x under metric ρ as $\mathcal{B}_\delta(x, \mathcal{X}, \rho)$. Denote $\zeta^{(t)} = \|\boldsymbol{\theta}^{(t)*}\|_2 \vee \sqrt{\frac{p + \log T}{n}}$ and $\bar{\zeta} = \sqrt{|S|^{-1} \sum_{t \in S} (\zeta^{(t)})^2}$.

D.1 Lemmas

Lemma 29 *Suppose Assumptions 1-3 hold. Then w.p. at least $1 - e^{-C(r + \log T)}$, for all $t \in [T]$ and $\boldsymbol{A} \in \mathcal{O}^{p \times r}$,*

$$\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \lesssim \|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top\|_2 \cdot \zeta^{(t)} + \sqrt{\frac{r + \log T}{n}}.$$

Lemma 30 *Suppose Assumptions 1-3 hold. Then w.p. at least $1 - e^{-C(r+\log T)}$, for all $t \in S$ and $\boldsymbol{A} \in \mathcal{O}^{p \times r}$, there exists a universal constant $C' > 0$ (which does not depend on t or $\boldsymbol{A}$), such that $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}\|_2 = \|\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}\|_2 \leq C'\zeta^{(t)}$.*

Lemma 31 *Suppose Assumptions 1-3 hold. Then w.p. at least $1 - e^{-C(r+\log T)}$, for all $t \in [T]$ and $\boldsymbol{A}, \widetilde{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$,*

$$\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}\|_2 \lesssim \|\boldsymbol{A}\boldsymbol{A}^\top - \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{A}}^\top\|_2 (\|\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}\|_2 + \|\nabla f^{(t)}(\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)})\|_2).$$

Lemma 32 *Suppose Assumptions 1-3 hold. For any $\overline{\boldsymbol{A}} \in \arg\min_{\overline{\boldsymbol{A}} \in \mathcal{O}^{p \times r}} \max_{t \in S}\{\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2\}$, denote $\overline{\boldsymbol{G}} = \{\nabla f^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})\}_{t \in S}$, then w.p. at least $1 - e^{-C(p+\log T)}$,*

$$\|\overline{\boldsymbol{G}}\|_2 \lesssim \sqrt{\frac{p}{n}} + \sqrt{T}\left(h\bar{\zeta} + \sqrt{\frac{r + \log T}{n}}\right).$$

Lemma 33 *Suppose $\boldsymbol{G} = \{\boldsymbol{g}_j\}_{j=1}^d \in \mathbb{R}^{p \times d}$ and $\overline{\boldsymbol{G}} = \{\overline{\boldsymbol{g}}_j\}_{j=1}^d \in \mathbb{R}^{p \times d}$ with each $\boldsymbol{g}_j, \overline{\boldsymbol{g}}_j \in \mathbb{R}^p$ and $\max_{j \in [d]} \|\boldsymbol{g}_j - \overline{\boldsymbol{g}}_j\|_2 \leq \Delta$. Then*

$$\|\overline{\boldsymbol{G}} - \boldsymbol{G}\|_2 \leq \sqrt{d}\Delta.$$

Lemma 34 *Suppose Assumptions 1-3 hold. Consider any $\widehat{\overline{\boldsymbol{A}}} \in \mathcal{O}^{p \times r}$.*

(i) *When $\frac{\lambda}{\sqrt{n}} \geq C\zeta^{(t)}\left[\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 + \|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2\right]$ with a sufficiently large C and the minimizer $(\boldsymbol{A}^{(t)}, \boldsymbol{\theta}^{(t)}) = \arg\min_{\boldsymbol{A} \in \mathcal{O}^{p \times r}, \boldsymbol{\theta} \in \mathbb{R}^r}\{\frac{1}{T}f^{(t)}(\boldsymbol{A}\boldsymbol{\theta}) + \frac{\sqrt{n}}{nT}\lambda\|\boldsymbol{A}\boldsymbol{A}^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2\}$, we must have $\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top = \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top$.*

(ii) *As a consequence, by Lemma 29, when $\frac{\lambda}{\sqrt{n}} \geq C\zeta^{(t)}\left[\zeta^{(t)}\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top\|_2 + \sqrt{\frac{p+\log T}{n}}\right]$, w.p. at least $1 - e^{-C(r+\log T)}$, we have $\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top = \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top$.*

Lemma 35 *Suppose $\{\boldsymbol{\theta}^{(t)}\}_{t \in S} \subseteq \mathbb{R}^r$ satisfying $|S|^{-1}\sum_{t \in S}\boldsymbol{\theta}^{(t)}(\boldsymbol{\theta}^{(t)})^\top \succeq \frac{c}{r}\bar{\zeta}\boldsymbol{I}_r$, where c is a positive constant. For any subset $S' \subseteq S$ with $|S'| \geq (1-\alpha)|S|$ and*

$$\alpha \leq \frac{c}{r} \cdot \frac{\bar{\zeta}^2}{\max_{t \in S}(\zeta^{(t)})^2},$$

where c is the same positive constant above and $\zeta^{(t)} = \|\boldsymbol{\theta}^{(t)}\|_2$, we have

$$\frac{1}{|S'|}\sum_{t \in S'}\boldsymbol{\theta}^{(t)}(\boldsymbol{\theta}^{(t)})^\top \succeq \frac{c}{2r}\bar{\zeta}^2\boldsymbol{I}_r.$$

Lemma 36 *Under Assumptions 1-3, w.p. at least $1 - e^{-C(r+\log T)}$, for all $t \in S$,*

$$\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \lesssim \sqrt{\frac{p + \log T}{n}} + \frac{\lambda}{\sqrt{n}\zeta^{(t)}}.$$

Lemma 37 *Consider the case that $r = 1$. Suppose that $\min_{t \in S} |\boldsymbol{\theta}^{(t)*}| \geq c > 0$ with some constant $c > 0$. Under Assumptions 1-3, w.p. at least $1 - e^{-C'(1+\log T)}$, for all $\boldsymbol{A} \in \mathcal{O}^{p \times r}$ and $t \in S$,*

$$\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \geq C'\zeta^{(t)}\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top\|_2 - C\sqrt{\frac{p + \log T}{n}}.$$

Lemma 38 (For MTL) *Under Assumptions 1-3, we have:*

(i) *For all $t \in S$, when $\frac{\gamma}{\sqrt{n}} \geq \|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 + C\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2$, it holds that $\widehat{\boldsymbol{\beta}}^{(t)} = \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}$ w.p. at least $1 - e^{-C(p+\log T)}$;*

(ii) *$\max_{t \in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \leq C\frac{\gamma}{\sqrt{n}} + \max_{t \in S} \|\widetilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ w.p. at least $1 - e^{-C(p+\log T)}$, where $\widetilde{\boldsymbol{\beta}}^{(t)} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} f^{(t)}(\boldsymbol{\beta})$;*

(iii) *If the data from tasks in S^c satisfies the linear model (2) (without any latent structure assumption) and Assumption 1, then w.p. at least $1 - e^{-C(p+\log T)}$, $\max_{t \in S^c} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \leq C\frac{\gamma}{\sqrt{n}} + \max_{t \in S^c} \|\widetilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$, where $\widetilde{\boldsymbol{\beta}}^{(t)} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} f^{(t)}(\boldsymbol{\beta})$.*

Lemma 39 *Suppose Assumption 8 hold and $n \geq Cr$. Then for **any fixed** $\boldsymbol{A} \in \mathcal{O}^{p \times r}$, w.p. at least $1 - e^{-C(r+\log T)}$,*

$$\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(0)} - \boldsymbol{A}^{(0)*}\boldsymbol{\theta}^{(0)*}\|_2 \lesssim \|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{A}^{(0)*}(\boldsymbol{A}^{(0)*})^\top\|_2 + \sqrt{\frac{r}{n}}.$$

Lemma 40 *Suppose Assumptions 8-9 hold. Then w.p. at least $1 - e^{-C(r+\log T)}$, for **all** $\boldsymbol{A} \in \mathcal{O}^{p \times r}$,*

$$\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(0)} - \boldsymbol{A}^{(0)*}\boldsymbol{\theta}^{(0)*}\|_2 \lesssim \|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{A}^{(0)*}(\boldsymbol{A}^{(0)*})^\top\|_2 + \sqrt{\frac{r}{n}}.$$

Lemma 41 (For TL) *Under Assumptions 8-9, we have:*

(i) *When $\frac{\gamma}{\sqrt{n_0}} \geq \|\nabla f^{(t)}(\boldsymbol{A}^{(0)*}\boldsymbol{\theta}^{(0)*})\|_2 + C\|\widehat{\boldsymbol{\Sigma}}^{(0)}\|_2 \cdot \|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(0)} - \boldsymbol{A}^{(0)*}\boldsymbol{\theta}^{(0)*}\|_2$, it holds that $\widehat{\boldsymbol{\beta}}^{(0)} = \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(0)}$ w.p. at least $1 - e^{-C(p+\log T)}$; [j]*

(ii) *$\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)*}\|_2 \leq C\frac{\gamma}{\sqrt{n_0}} + \|\widetilde{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)*}\|_2$ w.p. at least $1 - e^{-C(p+\log T)}$, where $\widetilde{\boldsymbol{\beta}}^{(0)} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} f^{(0)}(\boldsymbol{\beta})$.*

D.2 Proof of Theorem 1

(i) $\boldsymbol{\beta}^{(t)*} = \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*} + \overline{\boldsymbol{A}}^\perp(\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*} = \overline{\boldsymbol{A}}\bar{\boldsymbol{\theta}}^{(t)*} + \boldsymbol{\delta}^{(t)*}$, where $\bar{\boldsymbol{\theta}}^{(t)*} = (\overline{\boldsymbol{A}})^\top \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$ and $\boldsymbol{\delta}^{(t)*} = \overline{\boldsymbol{A}}^\perp(\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$. Then $\|\boldsymbol{\delta}^{(t)*}\|_2 \leq \|\boldsymbol{\theta}^{(t)*}\|_2\|(\overline{\boldsymbol{A}})^\top \boldsymbol{A}^{(t)*}\|_2 \leq h\zeta^{(t)}$. This shows how we can transform Setting 1 to Setting 2.

On the other hand, if we denote $\boldsymbol{\Delta}^{(t)*} = \boldsymbol{\delta}^{(t)*}(\bar{\boldsymbol{\theta}}^{(t)*})^\top \frac{1}{\|\bar{\boldsymbol{\theta}}^{(t)*}\|_2}$, then $\boldsymbol{\Delta}^{(t)*}\bar{\boldsymbol{\theta}}^{(t)*} = \boldsymbol{\delta}^{(t)*}$, hence $\boldsymbol{\beta}^{(t)*} = (\overline{\boldsymbol{A}} + \boldsymbol{\Delta}^{(t)*})\bar{\boldsymbol{\theta}}^{(t)*}$. Note that we have shown that $\|\boldsymbol{\delta}^{(t)*}\|_2 \leq h\zeta^{(t)}$. Also,

---

j. For this result to hold, it suffices to require $n_0 \geq Cr$, which is important for Remark 13 to be true.

$\|\bar{\boldsymbol{\theta}}^{(t)*}\|_2 \geq \sigma_{\min}((\overline{\boldsymbol{A}})^\top \boldsymbol{A}^{(t)*})\|\boldsymbol{\theta}^{(t)*}\|_2 \geq \sqrt{1 - h^2}\zeta^{(t)}$. Therefore $\|\boldsymbol{\Delta}^{(t)*}\|_{\mathrm{F}} = \|\boldsymbol{\Delta}^{(t)*}\|_2 \leq \|\boldsymbol{\delta}^{(t)*}\|_2/\|\bar{\boldsymbol{\theta}}^{(t)*}\|_2 \leq h/\sqrt{1 - h^2}$. This shows how we can transform Setting 1 to Setting 3.

(ii) Let $\bar{\boldsymbol{\beta}}^{(t)*} = \overline{\boldsymbol{A}}\bar{\boldsymbol{\theta}}^{(t)*}$, $\widetilde{\boldsymbol{A}} = (\frac{\boldsymbol{\beta}^{(t)*}}{\|\boldsymbol{\beta}^{(t)*}\|_2}, \widetilde{\boldsymbol{A}}_{-1})$ with $(\boldsymbol{\beta}^{(t)*})^\top \widetilde{\boldsymbol{A}}_{-1} = \boldsymbol{0}$, $\widetilde{\boldsymbol{A}}_{-1}^\top \widetilde{\boldsymbol{A}}_{-1} = \boldsymbol{I}$. WLOG, consider $\overline{\boldsymbol{A}} = (\frac{\bar{\boldsymbol{\beta}}^{(t)*}}{\|\bar{\boldsymbol{\beta}}^{(t)*}\|_2}, \overline{\boldsymbol{A}}_{-1})$ with $(\bar{\boldsymbol{\beta}}^{(t)*})^\top \overline{\boldsymbol{A}}_{-1} = \boldsymbol{0}$, $\overline{\boldsymbol{A}}_{-1}^\top \overline{\boldsymbol{A}}_{-1} = \boldsymbol{I}$. Also, we have

$$\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{A}}^\top = \frac{\boldsymbol{\beta}^{(t)*}(\boldsymbol{\beta}^{(t)*})^\top}{\|\boldsymbol{\beta}^{(t)*}\|_2^2} + \widetilde{\boldsymbol{A}}_{-1}\widetilde{\boldsymbol{A}}_{-1}^\top, \quad \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top = \frac{\bar{\boldsymbol{\beta}}^{(t)*}(\bar{\boldsymbol{\beta}}^{(t)*})^\top}{\|\bar{\boldsymbol{\beta}}^{(t)*}\|_2^2} + \overline{\boldsymbol{A}}_{-1}\overline{\boldsymbol{A}}_{-1}^\top.$$

And

$$\frac{\boldsymbol{\beta}^{(t)*}(\boldsymbol{\beta}^{(t)*})^\top}{\|\boldsymbol{\beta}^{(t)*}\|_2^2} - \frac{\bar{\boldsymbol{\beta}}^{(t)*}(\bar{\boldsymbol{\beta}}^{(t)*})^\top}{\|\bar{\boldsymbol{\beta}}^{(t)*}\|_2^2}$$
$$= \frac{\bar{\boldsymbol{\beta}}^{(t)*}(\bar{\boldsymbol{\beta}}^{(t)*})^\top + \boldsymbol{\delta}^{(t)*}(\boldsymbol{\delta}^{(t)*})^\top + \bar{\boldsymbol{\beta}}^{(t)*}(\boldsymbol{\delta}^{(t)*})^\top + \boldsymbol{\delta}^{(t)*}(\bar{\boldsymbol{\beta}}^{(t)*})^\top}{\|\bar{\boldsymbol{\beta}}^{(t)*}\|_2^2 + \|\boldsymbol{\delta}^{(t)*}\|_2^2} - \frac{\bar{\boldsymbol{\beta}}^{(t)*}(\bar{\boldsymbol{\beta}}^{(t)*})^\top}{\|\bar{\boldsymbol{\beta}}^{(t)*}\|_2^2}$$
$$= \frac{-\|\boldsymbol{\delta}^{(t)*}\|_2^2 \cdot \bar{\boldsymbol{\beta}}^{(t)*}(\bar{\boldsymbol{\beta}}^{(t)*})^\top + \|\bar{\boldsymbol{\beta}}^{(t)*}\|_2^2[\boldsymbol{\delta}^{(t)*}(\boldsymbol{\delta}^{(t)*})^\top + \bar{\boldsymbol{\beta}}^{(t)*}(\boldsymbol{\delta}^{(t)*})^\top + \boldsymbol{\delta}^{(t)*}(\bar{\boldsymbol{\beta}}^{(t)*})^\top]}{(\|\bar{\boldsymbol{\beta}}^{(t)*}\|_2^2 + \|\boldsymbol{\delta}^{(t)*}\|_2^2)\|\bar{\boldsymbol{\beta}}^{(t)*}\|_2^2},$$

which implies that

$$\left\|\frac{\boldsymbol{\beta}^{(t)*}(\boldsymbol{\beta}^{(t)*})^\top}{\|\boldsymbol{\beta}^{(t)*}\|_2^2} - \frac{\bar{\boldsymbol{\beta}}^{(t)*}(\bar{\boldsymbol{\beta}}^{(t)*})^\top}{\|\bar{\boldsymbol{\beta}}^{(t)*}\|_2^2}\right\|_2 \leq \max\left\{\frac{\|\boldsymbol{\delta}^{(t)*}\|_2^2}{\|\boldsymbol{\beta}^{(t)*}\|_2^2} + 2\frac{\|\bar{\boldsymbol{\beta}}^{(t)*}\|_2\|\boldsymbol{\delta}^{(t)*}\|_2}{\|\boldsymbol{\beta}^{(t)*}\|_2^2}, \frac{\|\boldsymbol{\delta}^{(t)*}\|_2^2}{\|\boldsymbol{\beta}^{(t)*}\|_2^2}\right\}$$
$$\leq \frac{\|\boldsymbol{\delta}^{(t)*}\|_2^2}{\|\boldsymbol{\beta}^{(t)*}\|_2^2} + 2\frac{\|\boldsymbol{\delta}^{(t)*}\|_2}{\|\boldsymbol{\beta}^{(t)*}\|_2} \cdot \sqrt{1 - \left(\frac{\|\boldsymbol{\delta}^{(t)*}\|_2}{\|\boldsymbol{\beta}^{(t)*}\|_2}\right)^2}$$
$$= \xi^2 + 2\xi\sqrt{1 - \xi^2}$$
$$\leq 2\sqrt{2}\xi,$$

where $\xi := \|\boldsymbol{\delta}^{(t)*}\|_2/\|\boldsymbol{\beta}^{(t)*}\|_2$.

Let $\boldsymbol{P}_{\boldsymbol{\beta}} = \frac{\boldsymbol{\beta}\boldsymbol{\beta}^\top}{\|\boldsymbol{\beta}\|_2^2}$ as a projection matrix onto the linear space spanned by $\boldsymbol{\beta} \in \mathbb{R}^p \backslash \{\boldsymbol{0}\}$, and $\widetilde{\boldsymbol{A}}_{-1} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}})\overline{\boldsymbol{A}}_{-1}[(\overline{\boldsymbol{A}}_{-1})^\top(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}})\overline{\boldsymbol{A}}_{-1}]^{-1/2}$. Then

$$\widetilde{\boldsymbol{A}}_{-1}\widetilde{\boldsymbol{A}}_{-1}^\top - \overline{\boldsymbol{A}}_{-1}(\overline{\boldsymbol{A}}_{-1})^\top = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}})\overline{\boldsymbol{A}}_{-1}[(\overline{\boldsymbol{A}}_{-1})^\top(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}})\overline{\boldsymbol{A}}_{-1}]^{-1}(\overline{\boldsymbol{A}}_{-1})^\top(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}}) - \overline{\boldsymbol{A}}_{-1}(\overline{\boldsymbol{A}}_{-1})^\top.$$

By noticing that $\overline{\boldsymbol{A}}_{-1} = (\boldsymbol{I} - \boldsymbol{P}_{\bar{\boldsymbol{\beta}}^{(t)*}})\overline{\boldsymbol{A}}$, after some simplifications, we have

$$\|\widetilde{\boldsymbol{A}}_{-1}\widetilde{\boldsymbol{A}}_{-1}^\top - \overline{\boldsymbol{A}}_{-1}(\overline{\boldsymbol{A}}_{-1})^\top\|_2 \leq 2\|\boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}}(\boldsymbol{I} - \boldsymbol{P}_{\bar{\boldsymbol{\beta}}^{(t)*}})\|_2\|[(\overline{\boldsymbol{A}}_{-1})^\top(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}})\overline{\boldsymbol{A}}_{-1}]^{-1}\|_2$$
$$+ \|\boldsymbol{I} - [(\overline{\boldsymbol{A}}_{-1})^\top(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}})\overline{\boldsymbol{A}}_{-1}]^{-1}\|_2,$$

where

$$\|\boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}}(\boldsymbol{I} - \boldsymbol{P}_{\bar{\boldsymbol{\beta}}^{(t)*}})\|_2 = \|\boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}} - \boldsymbol{P}_{\bar{\boldsymbol{\beta}}^{(t)*}}\|_2 \leq 2\sqrt{2} \cdot \frac{\|\boldsymbol{\delta}^{(t)*}\|_2}{\|\boldsymbol{\beta}^{(t)*}\|_2} = 2\sqrt{2}\xi,$$
$$\|\boldsymbol{I} - [(\overline{\boldsymbol{A}}_{-1})^\top(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}})\overline{\boldsymbol{A}}_{-1}]^{-1}\|_2 = \|\boldsymbol{I} - (\boldsymbol{I} - (\overline{\boldsymbol{A}}_{-1})^\top\boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}}\overline{\boldsymbol{A}}_{-1})^{-1}\|_2$$

$$= \left\| \boldsymbol{I} - \left(\boldsymbol{I} + \frac{(\overline{\boldsymbol{A}}_{-1})^\top \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}} \overline{\boldsymbol{A}}_{-1}}{1 - \frac{(\boldsymbol{\beta}^{(t)*})^\top \overline{\boldsymbol{A}}_{-1}(\overline{\boldsymbol{A}}_{-1})^\top \boldsymbol{\beta}^{(t)*}}{\|\boldsymbol{\beta}^{(t)*}\|_2^2}} \right) \right\|_2$$

$$= \left\| \frac{(\overline{\boldsymbol{A}}_{-1})^\top \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}} \overline{\boldsymbol{A}}_{-1}}{1 - \frac{(\boldsymbol{\beta}^{(t)*})^\top \overline{\boldsymbol{A}}_{-1}(\overline{\boldsymbol{A}}_{-1})^\top \boldsymbol{\beta}^{(t)*}}{\|\boldsymbol{\beta}^{(t)*}\|_2^2}} \right\|_2,$$

$$\|(\overline{\boldsymbol{A}}_{-1})^\top \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}} \overline{\boldsymbol{A}}_{-1}\|_2 = \|(\overline{\boldsymbol{A}}_{-1})^\top (\boldsymbol{I} - \boldsymbol{P}_{\bar{\boldsymbol{\beta}}^{(t)*}}) \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}} (\boldsymbol{I} - \boldsymbol{P}_{\bar{\boldsymbol{\beta}}^{(t)*}}) \overline{\boldsymbol{A}}_{-1}\|_2$$
$$\leq \|\boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}} (\boldsymbol{I} - \boldsymbol{P}_{\bar{\boldsymbol{\beta}}^{(t)*}})\|_2^2,$$

$$\left| \frac{(\boldsymbol{\beta}^{(t)*})^\top \overline{\boldsymbol{A}}_{-1}(\overline{\boldsymbol{A}}_{-1})^\top \boldsymbol{\beta}^{(t)*}}{\|\boldsymbol{\beta}^{(t)*}\|_2^2} \right| = \|(\overline{\boldsymbol{A}}_{-1})^\top \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}} \overline{\boldsymbol{A}}_{-1}\|_2.$$

Therefore, we have

$$\|\boldsymbol{I} - [(\overline{\boldsymbol{A}}_{-1})^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}}) \overline{\boldsymbol{A}}_{-1}]^{-1}\|_2 \leq \frac{\|\boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}} (\boldsymbol{I} - \boldsymbol{P}_{\bar{\boldsymbol{\beta}}^{(t)*}})\|_2^2}{1 - \|\boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}} (\boldsymbol{I} - \boldsymbol{P}_{\bar{\boldsymbol{\beta}}^{(t)*}})\|_2^2},$$

$$\|[(\overline{\boldsymbol{A}}_{-1})^\top (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}}) \overline{\boldsymbol{A}}_{-1}]^{-1}\|_2 = 1 + \left\| \frac{(\overline{\boldsymbol{A}}_{-1})^\top \boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}} \overline{\boldsymbol{A}}_{-1}}{1 - \frac{(\boldsymbol{\beta}^{(t)*})^\top \overline{\boldsymbol{A}}_{-1}(\overline{\boldsymbol{A}}_{-1})^\top \boldsymbol{\beta}^{(t)*}}{\|\boldsymbol{\beta}^{(t)*}\|_2^2}} \right\|_2 \leq 1 + \frac{\|\boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}} (\boldsymbol{I} - \boldsymbol{P}_{\bar{\boldsymbol{\beta}}^{(t)*}})\|_2^2}{1 - \|\boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}} (\boldsymbol{I} - \boldsymbol{P}_{\bar{\boldsymbol{\beta}}^{(t)*}})\|_2^2}.$$

Combining all of the results, we have

$$\|\widetilde{\boldsymbol{A}}_{-1} \widetilde{\boldsymbol{A}}_{-1}^\top - \overline{\boldsymbol{A}}_{-1}(\overline{\boldsymbol{A}}_{-1})^\top\|_2 \leq 2 \cdot 2\sqrt{2}\xi \cdot \left[1 + \frac{(2\sqrt{2}\xi)^2}{1 - (2\sqrt{2}\xi)^2} \right] + \frac{(2\sqrt{2}\xi)^2}{1 - (2\sqrt{2}\xi)^2}$$
$$= 4\sqrt{2}\xi \left(1 + \frac{8\xi^2}{1 - 8\xi^2} \right) + \frac{8\xi^2}{1 - 8\xi^2},$$

which implies that

$$\|\widetilde{\boldsymbol{A}} \widetilde{\boldsymbol{A}}^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \leq \|\boldsymbol{P}_{\boldsymbol{\beta}^{(t)*}} \boldsymbol{P}_{\bar{\boldsymbol{\beta}}^{(t)*}}\|_2 + \|\widetilde{\boldsymbol{A}}_{-1} \widetilde{\boldsymbol{A}}_{-1}^\top - \overline{\boldsymbol{A}}_{-1}(\overline{\boldsymbol{A}}_{-1})^\top\|_2$$
$$\leq 2\sqrt{2}\xi + 4\sqrt{2}\xi \left(1 + \frac{8\xi^2}{1 - 8\xi^2} \right) + \frac{8\xi^2}{1 - 8\xi^2}. \qquad (*)$$

- When $\xi \leq 1/4$: $(*) \leq 2\sqrt{2}\xi + 4\sqrt{2}\xi \cdot (1 + 1) + 8\xi \cdot \frac{1/4}{1 - 1/2} = (10\sqrt{2} + 4)\xi$.

- When $\xi > 1/4$: $\|\widetilde{\boldsymbol{A}} \widetilde{\boldsymbol{A}}^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \leq 1 \leq 4\xi$.

Therefore, $\|\widetilde{\boldsymbol{A}} \widetilde{\boldsymbol{A}}^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \leq (10\sqrt{2} + 4)\xi$.

(iii) $\boldsymbol{\beta}^{(t)*} = (\overline{\boldsymbol{A}} + \boldsymbol{\Delta}^{(t)*})\tilde{\boldsymbol{\theta}}^{(t)*} = \overline{\boldsymbol{A}}(\boldsymbol{I} + (\overline{\boldsymbol{A}})^\top \boldsymbol{\Delta}^{(t)*})\tilde{\boldsymbol{\theta}}^{(t)*} + \overline{\boldsymbol{A}}^\perp (\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{\Delta}^{(t)*}\tilde{\boldsymbol{\theta}}^{(t)*}$. Let $\bar{\boldsymbol{\theta}}^{(t)*} = (\boldsymbol{I} + (\overline{\boldsymbol{A}})^\top \boldsymbol{\Delta}^{(t)*})\tilde{\boldsymbol{\theta}}^{(t)*}$ and $\boldsymbol{\delta}^{(t)*} = \overline{\boldsymbol{A}}^\perp (\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{\Delta}^{(t)*}\tilde{\boldsymbol{\theta}}^{(t)*}$. By (ii), there exists $\boldsymbol{A}^{(t)*} \in \mathcal{O}^{p \times r}$ such that $\|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \leq (10\sqrt{2} + 4)\frac{\|\boldsymbol{\delta}^{(t)*}\|_2}{\|\boldsymbol{\beta}^{(t)*}\|_2}$.

Since

$$\|\boldsymbol{\delta}^{(t)*}\|_2 \leq \|(\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{\Delta}^{(t)*}\|_2 \|\tilde{\boldsymbol{\theta}}^{(t)*}\|_2,$$

$$\|\boldsymbol{\beta}^{(t)*}\|_2 = \|\tilde{\boldsymbol{\theta}}^{(t)*}\|_2^2 + \|\boldsymbol{\Delta}^{(t)*}\tilde{\boldsymbol{\theta}}^{(t)*}\|_2^2 + 2(\tilde{\boldsymbol{\theta}}^{(t)*})^\top \overline{\boldsymbol{A}}^\top \boldsymbol{\Delta}^{(t)*}\tilde{\boldsymbol{\theta}}^{(t)*}$$

$$\geq \|\tilde{\boldsymbol{\theta}}^{(t)*}\|_2^2 + \|\overline{\boldsymbol{A}}^\perp(\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{\Delta}^{(t)*}\tilde{\boldsymbol{\theta}}^{(t)*}\|_2^2 - 2\|\tilde{\boldsymbol{\theta}}^{(t)*}\|_2^2\|\overline{\boldsymbol{A}}^\top \boldsymbol{\Delta}^{(t)*}\|_2$$

$$= \|\tilde{\boldsymbol{\theta}}^{(t)*}\|_2^2 + \|\boldsymbol{\delta}^{(t)*}\|_2^2 - 2\|\tilde{\boldsymbol{\theta}}^{(t)*}\|_2^2\|\overline{\boldsymbol{A}}^\top \boldsymbol{\Delta}^{(t)*}\|_2,$$

we have

$$\frac{\|\boldsymbol{\delta}^{(t)*}\|_2}{\|\boldsymbol{\beta}^{(t)*}\|_2} \leq \frac{\|(\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{\Delta}^{(t)*}\|_2}{\sqrt{1 + \|(\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{\Delta}^{(t)*}\|_2^2 - 2\|\boldsymbol{\Delta}^{(t)*}\|_2}} \leq \frac{\|\boldsymbol{\Delta}^{(t)*}\|_2}{1 - \|\boldsymbol{\Delta}^{(t)*}\|_2} \leq \frac{\|\boldsymbol{\Delta}^{(t)*}\|_F}{1 - \|\boldsymbol{\Delta}^{(t)*}\|_F},$$

when $\|\boldsymbol{\Delta}^{(t)*}\|_F < 1$. Hence there exists $\boldsymbol{A}^{(t)*} \in \mathcal{O}^{p \times r}$ such that $\|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \leq (10\sqrt{2}+4)\frac{\|\boldsymbol{\delta}^{(t)*}\|_2}{\|\boldsymbol{\beta}^{(t)*}\|_2} \leq (10\sqrt{2}+4)\frac{\|\boldsymbol{\Delta}^{(t)*}\|_F}{1-\|\boldsymbol{\Delta}^{(t)*}\|_F}$ when $\|\boldsymbol{\Delta}^{(t)*}\|_F < 1$.

D.3 Proof of Theorem 2

First, we have the following proposition holds.

Proposition 42 *Suppose Assumptions 1-3 hold. Further assume*

(i) $\min_{t \in S}\left\{\frac{\zeta^{(t)}}{\overline{\zeta}}r\sqrt{\frac{p+\log T}{nT}} + \frac{\zeta^{(t)}}{\overline{\zeta}}\sqrt{r}\sqrt{\frac{r+\log T}{n}} + \zeta^{(t)}\sqrt{r}h\right\} \lesssim \sqrt{\frac{p+\log T}{n}}$;

(ii) $|S^c|/T \leq cr^{-3/2}\frac{\overline{\zeta}}{\max_{t\in S}\zeta^{(t)}}$ *with a small constant* $c > 0$.

Let $\lambda \geq C\frac{\max_{t\in S}(\zeta^{(t)})^2}{\min_{t\in S}\zeta^{(t)}}\sqrt{r(p+\log T)}$ *with a sufficiently large constant* C *and* $\lambda \leq C'\frac{\max_{t\in S}(\zeta^{(t)})^2}{\min_{t\in S}\zeta^{(t)}}\sqrt{r(p+\log T)}$ *with another constant* C'. *Then w.p. at least* $1 - e^{-C'(r+\log T)}$,

$$\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2 \lesssim r\overline{\zeta}^{-1}\sqrt{\frac{p}{nT}} + \sqrt{r}h + \overline{\zeta}^{-1}\sqrt{r}\sqrt{\frac{r+\log T}{n}} + \frac{\lambda r}{\sqrt{n}\overline{\zeta}^2}\epsilon.$$

Proposition 42 provides the bound when the first term is faster than the second term in the upper bound of Theorem 2. In the other case, a direct application of Lemma 38 leads to the second term in the upper bound. Combining two situations gives us the disired result for tasks in S.

Denote $\eta^{(t)} = \frac{\zeta^{(t)}}{\overline{\zeta}}r\sqrt{\frac{p+\log T}{nT}} + \frac{\zeta^{(t)}}{\overline{\zeta}}\sqrt{r}\sqrt{\frac{r+\log T}{n}} + \zeta^{(t)}\sqrt{r}h + \frac{\lambda r\zeta^{(t)}}{\sqrt{n}\overline{\zeta}^2}\epsilon$.

(i) For any t satisfying $\eta^{(t)} \leq C\sqrt{\frac{p+\log T}{n}}$: By Proposition 42, w.p. at least $1 - e^{-C'(r+\log T)}$,

$$\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top\|_2 \lesssim r\overline{\zeta}^{-1}\sqrt{\frac{p}{nT}} + \sqrt{r}h + \overline{\zeta}^{-1}\sqrt{r}\sqrt{\frac{r+\log T}{n}} + \frac{\lambda r}{\sqrt{n}\overline{\zeta}^2}\epsilon.$$

By Lemma 34, since $\frac{\lambda}{\sqrt{n}} \geq C\frac{\max_{t\in S}(\zeta^{(t)})^2}{\min_{t\in S}\zeta^{(t)}}\sqrt{\frac{r(p+\log T)}{n}} \geq C\max_{t\in S}\zeta^{(t)}\eta^{(t)} \geq C\zeta^{(t)}\left[\zeta^{(t)}\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top\|_2 + \sqrt{\frac{p+\log T}{n}}\right]$, we have $\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top = \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top$ w.p. at least $1 - e^{-C'(r+\log T)}$, which combining with Lemma 29 implies that

$$\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 = \|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2$$

$$\lesssim \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top\|_2 \zeta^{(t)} + \sqrt{\frac{r + \log T}{n}}$$

$$\lesssim \eta^{(t)} + \sqrt{\frac{r + \log T}{n}},$$

w.p. at least $1 - e^{-C'(r + \log T)}$. Therefore by our choice $\gamma = C'\sqrt{p + \log T}$ with a large constant $C' > 0$:

$$\|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 + C\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \lesssim \frac{\gamma}{\sqrt{n}}.$$

Then by Lemma 38.(i) and Proposition 42, w.p. at least $1 - e^{-C'(r + \log T)}$, $\max_{t \in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \eta^{(t)} + \sqrt{\frac{r + \log T}{n}}$.

(ii) For any t satisfying $\eta^{(t)} > C\sqrt{\frac{p + \log T}{n}}$: by Lemma 38.(ii), w.p. at least $1 - e^{-C'(p + \log T)}$, $\max_{t \in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \frac{\gamma}{\sqrt{n}} + \max_{t \in S} \|\widetilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \sqrt{\frac{p + \log T}{n}}$, where $\widetilde{\boldsymbol{\beta}}^{(t)} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} f^{(t)}(\boldsymbol{\beta})$. The fact that $\max_{t \in S} \|\widetilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \sqrt{\frac{p + \log T}{n}}$ w.p. at least $1 - e^{-C'(p + \log T)}$ is the standard linear regression result.

(iii) When data from tasks in S^c also satisfies the linear model (2), the result comes from the same argument as in (ii).

D.4 Proof of Theorem 3

The conclusion follows directly by combining Lemma 36 with the upper bound on $\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2$ established in the proof of Theorem 2.

D.5 Proof of Theorem 4

First, note that when $\lambda = +\infty$, we have $\widehat{\boldsymbol{A}}^{(t)} \equiv \widehat{\overline{\boldsymbol{A}}}$ for all $t \in S = [T]$. We will prove that any local minimizer $(\widehat{\overline{\boldsymbol{A}}}, \{\widehat{\boldsymbol{\theta}}^{(t)}\}_{t=1}^T)$ which minimizes

$$f(\boldsymbol{A}, \boldsymbol{\Theta}) := \frac{1}{T}\sum_{t=1}^T f^{(t)}(\boldsymbol{A}\boldsymbol{\theta}^{(t)}).$$

over $\boldsymbol{A} \in \mathcal{O}^{p \times r}$ and $\boldsymbol{\Theta} = (\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}) \in \mathbb{R}^{r \times T}$ must satisfy

$$\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_F \lesssim (\bar{\zeta}^{-1})r\sqrt{\frac{p}{nT}} + (\bar{\zeta}^{-1})\sqrt{r}\sqrt{\frac{r + \log T}{n}} + \sqrt{r}h,$$

simulteneously w.p. at least $1 - e^{-C'(r + \log T)}$. The following the argument in the proof of Theorem 2, we will obtain the desired result.

We prove the claim above by construction. Consider a local minimizer $(\widehat{\overline{\boldsymbol{A}}}, \{\widehat{\boldsymbol{\theta}}^{(t)}\}_{t=1}^T)$ of f and $\widehat{\overline{\boldsymbol{A}}}$'s neightbor $\widetilde{\boldsymbol{A}} = a_1\overline{\boldsymbol{A}} + a_2\widehat{\overline{\boldsymbol{A}}}\boldsymbol{D}$, where $a_1 + a_2 = 1$, $a_1, a_2 \geq 0$, and $\boldsymbol{D} \in \mathbb{R}^{r \times r}$ is invertible. Note that

$$\|((\widehat{\overline{\boldsymbol{A}}})^\perp)^\top \widetilde{\boldsymbol{A}}\|_2 = a_1\|((\widehat{\overline{\boldsymbol{A}}})^\perp)^\top \overline{\boldsymbol{A}}\|_2 \leq a_1.$$

Define $\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^r} f^{(t)}(\boldsymbol{A}\boldsymbol{\theta})$, $\widehat{\boldsymbol{\Theta}} = \{\boldsymbol{\theta}_{\widehat{\boldsymbol{A}}}^{(t)}\}_{t=1}^T$, $\widetilde{\boldsymbol{R}} \in \mathcal{O}^{r \times r}$ by

$$\widetilde{\boldsymbol{A}}^\top \widetilde{\boldsymbol{A}} = a_1^2 \boldsymbol{I}_p + a_2^2 \boldsymbol{D}^\top \boldsymbol{D} + a_1 a_2 ((\overline{\boldsymbol{A}})^\top \widehat{\overline{\boldsymbol{A}}} \boldsymbol{D} + \boldsymbol{D}^\top (\widehat{\overline{\boldsymbol{A}}})^\top \overline{\boldsymbol{A}}) = (\widetilde{\boldsymbol{R}}^{-1})^\top \widetilde{\boldsymbol{R}}^{-1}.$$

Then define $\boldsymbol{A} = \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{R}}$, hence $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_r$. Let $\boldsymbol{D}$ satisfy $\boldsymbol{D}\widetilde{\boldsymbol{R}}\overline{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Theta}}$, where $\overline{\boldsymbol{\Theta}} = \{\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t=1}^T \in \mathbb{R}^{r \times T}$. Such $\boldsymbol{D}$ must exist. For example, we can take $\boldsymbol{D} = (\widetilde{\boldsymbol{R}}\overline{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Theta}}^\top)^{-1}\widehat{\boldsymbol{\Theta}}\overline{\boldsymbol{\Theta}}^\top$. Consider $\boldsymbol{\Theta}$ s.t. $\widetilde{\boldsymbol{R}}\boldsymbol{\Theta} = \overline{\boldsymbol{\Theta}}$, then

$$\boldsymbol{A}\boldsymbol{\Theta} = a_1 \overline{\boldsymbol{A}}\widetilde{\boldsymbol{R}}\boldsymbol{\Theta} + a_2 \widehat{\overline{\boldsymbol{A}}}\boldsymbol{D}\widetilde{\boldsymbol{R}}\boldsymbol{\Theta} = a_1 \overline{\boldsymbol{A}}\overline{\boldsymbol{\Theta}} + a_2 \widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\Theta}}. \tag{9}$$

Since $f^{(t)}$ is defined through square loss, we have

$$f(\boldsymbol{A}, \boldsymbol{\Theta}) - f(\widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}}) = \frac{1}{2T}\sum_{t=1}^T (\boldsymbol{A}\boldsymbol{\theta}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\theta}}^{(t)})^\top \widehat{\boldsymbol{\Sigma}}^{(t)}(\boldsymbol{A}\boldsymbol{\theta}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\theta}}^{(t)})$$

$$+ \underbrace{\frac{1}{T}\sum_{t=1}^T \left[\nabla f^{(t)}(\widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\theta}}^{(t)})\right]^\top (\boldsymbol{A}\boldsymbol{\theta}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\theta}}^{(t)})}_{(*)}.$$

Therefore, (9) implies that

$$(*) \leq \frac{1}{T}\sum_{t=1}^T \left[\nabla f^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})\right]^\top (\boldsymbol{A}\boldsymbol{\theta}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\theta}}^{(t)}) - \frac{1}{T}a_1 \sum_{t=1}^T (\widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\theta}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})^\top \widehat{\boldsymbol{\Sigma}}^{(t)}(\widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\theta}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})$$

$$\leq \frac{1}{T}\|\{\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\}_{t=1}^T\|_2 \cdot \sqrt{2r}\|\boldsymbol{A}\boldsymbol{\Theta} - \widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\Theta}}\|_{\mathrm{F}}$$

$$+ \frac{1}{T}\|\{\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}) - \nabla f^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})\}_{t=1}^T\|_{\mathrm{F}}\|\boldsymbol{A}\boldsymbol{\Theta} - \widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\Theta}}\|_{\mathrm{F}} - \frac{C}{T}a_1\|\widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{A}}\overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2$$

$$\leq \frac{1}{T}\sqrt{2r}\sqrt{\frac{p+T}{n}}\|\boldsymbol{A}\boldsymbol{\Theta} - \widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\Theta}}\|_{\mathrm{F}} + \frac{C}{T}\left(\sqrt{T}\sqrt{\frac{r+\log T}{n}} + h\sqrt{\sum_{t=1}^T (\zeta^{(t)})^2}\right)\|\boldsymbol{A}\boldsymbol{\Theta} - \widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\Theta}}\|_{\mathrm{F}}$$

$$- \frac{C}{T}a_1\|\widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{A}}\overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2$$

$$\leq \frac{C}{T}a_1\left(\sqrt{\frac{pr}{n}} + \sqrt{\frac{r+\log T}{n}}\sqrt{T} + h\sqrt{T}\bar{\zeta}\right)\|\widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{A}}\overline{\boldsymbol{\Theta}}\|_{\mathrm{F}} - \frac{C}{T}a_1\|\widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{A}}\overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2,$$

w.p. at least $1 - e^{-C'(r+\log T)}$. Hence w.p. at least $1 - e^{-C'(r+\log T)}$,

$$f(\boldsymbol{A}, \boldsymbol{\Theta}) - f(\widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}}) \leq \frac{C''}{T}a_1^2 + \frac{C}{T}a_1\left(\sqrt{\frac{pr}{n}} + \sqrt{\frac{r+\log T}{n}}\sqrt{T} + h\sqrt{T}\bar{\zeta}\right)\|\widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{A}}\overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}$$

$$- \frac{C}{T}a_1\|\widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{A}}\overline{\boldsymbol{\Theta}}\|_{\mathrm{F}}^2. \tag{10}$$

By (9), as $a_1 \to 0+$,

$$\boldsymbol{\Theta} = a_1 \boldsymbol{A}^\top \overline{\boldsymbol{A}} \overline{\boldsymbol{\Theta}} + a_2 \boldsymbol{A}^\top \widehat{\overline{\boldsymbol{A}}} \widehat{\boldsymbol{\Theta}} \to \widetilde{\boldsymbol{R}}^\top \boldsymbol{D}^\top (\widehat{\overline{\boldsymbol{A}}})^\top \widehat{\overline{\boldsymbol{A}}} \widehat{\boldsymbol{\Theta}} = \widetilde{\boldsymbol{R}}^\top \boldsymbol{D}^\top \widehat{\boldsymbol{\Theta}},$$

where "$\to$" is in the sense of $\|\cdot\|_{\max}$ (similar below). On the other hand, as $a_1 \to 0+$,

$$\boldsymbol{A} = a_1 \overline{\boldsymbol{A}} \widetilde{\boldsymbol{R}} + a_2 \widehat{\overline{\boldsymbol{A}}} \boldsymbol{D} \widetilde{\boldsymbol{R}} \to \widehat{\overline{\boldsymbol{A}}} \boldsymbol{D} \widetilde{\boldsymbol{R}},$$

and $\widetilde{\boldsymbol{R}}^\top \boldsymbol{D}^\top \boldsymbol{D} \widetilde{\boldsymbol{R}} \to \boldsymbol{I}_r$. Therefore, by continuity, as $a_1 \to 0+$,

$$\max_{t \in [T]} \min_{\boldsymbol{R} \in \mathcal{O}^{r \times r}} \left\{ \|\boldsymbol{A} - \widehat{\overline{\boldsymbol{A}}} \boldsymbol{R}\|_2 + \|\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{\theta}}^{(t)} \boldsymbol{R}\|_2 \right\} \to 0.$$

Due to the local optimality of $(\widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})$, when a_1 is very close to 0, we must have $f(\boldsymbol{A}, \boldsymbol{\Theta}) - f(\widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}}) \geq 0$, hence the RHS of (10) is non-negative. Let $a_1 \lesssim \sqrt{\frac{pr}{n}} + \sqrt{\frac{r + \log T}{n}} \sqrt{T} + h\sqrt{T}\bar{\zeta}$, we have w.p. at least $1 - e^{-C'(r + \log T)}$,

$$\|\widehat{\overline{\boldsymbol{A}}} \widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{A}} \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}} \lesssim a_1 + \sqrt{\frac{pr}{n}} + \sqrt{\frac{r + \log T}{n}} \sqrt{T} + h\sqrt{T}\bar{\zeta} \lesssim \sqrt{\frac{pr}{n}} + \sqrt{\frac{r + \log T}{n}} \sqrt{T} + h\sqrt{T}\bar{\zeta}.$$

By Wedin's Theorem, w.p. at least $1 - e^{-C'(r + \log T)}$,

$$\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_{\mathrm{F}} \lesssim \sqrt{\frac{T}{r}} \bar{\zeta} \|\widehat{\overline{\boldsymbol{A}}} \widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{A}} \overline{\boldsymbol{\Theta}}\|_{\mathrm{F}} \lesssim \bar{\zeta}^{-1} r \sqrt{\frac{p}{nT}} + \bar{\zeta}^{-1} \sqrt{r} \sqrt{\frac{r + \log T}{n}} + \sqrt{r} h.$$

This proves our previous claim. The remaining argument is the same as in the proof of Theorem 2.

D.6 Proof of Theorem 5

We prove a slightly stronger version of Theorem 5 by replacing the constraint set

$$4(1 + \sqrt{2}) \max_{t \in S} \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)}) - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 + 4\frac{\epsilon}{1 - \epsilon} < 1 - \gamma,$$

$$\sqrt{2} \max_{t \in S} \left\{ \frac{\sigma_{\max}(\boldsymbol{\Sigma}^{(t)})}{\sigma_{\min}(\boldsymbol{\Sigma}^{(t)})} \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top\|_2 \right\} < 1 - \gamma. \tag{11}$$

with

$$2(2 + \sqrt{2}) \max_{t \in S} \min_{\boldsymbol{R} \in \mathcal{O}^{r \times r}} \|\widehat{\boldsymbol{A}}^{(t)} - \overline{\boldsymbol{A}} \boldsymbol{R}\|_2 + 4\frac{\epsilon}{1 - \epsilon} < 1 - \gamma,$$

$$\max_{t \in S} \left\{ \frac{\sigma_{\max}(\boldsymbol{\Sigma}^{(t)})}{\sigma_{\min}(\boldsymbol{\Sigma}^{(t)})} \min_{\boldsymbol{R} \in \mathcal{O}^{r \times r}} \|\widehat{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*} \boldsymbol{R}\|_2 \right\} < 1 - \gamma. \tag{12}$$

Note that for any matrices $\boldsymbol{A}, \boldsymbol{A}' \in \mathcal{O}^{p \times r}$, Lemma 2.6 in Chen et al. (2021b) tells us that

$$\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{A}'(\boldsymbol{A}')^\top\|_2 \leq \min_{\boldsymbol{R} \in \mathcal{O}^{r \times r}} \|\boldsymbol{A} - \boldsymbol{A}' \boldsymbol{R}\|_2 \leq \sqrt{2} \|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{A}'(\boldsymbol{A}')^\top\|_2.$$

Therefore the current constraint set (11) is indeed weaker than the constraint set (12) in the main text.

The proof follows the construction idea in the proof of Theorem 4 and the argument in the proof of Proposition 42. The difference is that here we will first consider a relaxed optimization problem where we require $\widehat{\boldsymbol{A}}^{(t)}$ and $\widehat{\overline{\boldsymbol{A}}}$ to satisfy $\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \boldsymbol{I}_r\|_2$, $\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \boldsymbol{I}_r\|_2 \leq \delta$ with a small constant $\delta > 0$ instead of $\widehat{\boldsymbol{A}}^{(t)}, \widehat{\overline{\boldsymbol{A}}} \in \mathcal{O}^{p \times r}$. Finally, we will let $\delta \to 0+$ to obtain the desired original result.

Denote

$$G(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \overline{\boldsymbol{A}}, \boldsymbol{\Theta}) = \frac{1}{T}\sum_{t=1}^T f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}) + \sum_{t=1}^T \frac{\lambda}{\sqrt{nT}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2,$$

with $\boldsymbol{\Theta} = (\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}) \in \mathbb{R}^{r \times T}$. First, consider any local minimizer $\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T$ and $\widehat{\overline{\boldsymbol{A}}}$ of

$$\min_{\{\boldsymbol{\theta}^{(t)}\}_{t=1}^T \subseteq \mathbb{R}^r} G(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \overline{\boldsymbol{A}}, \boldsymbol{\Theta}),$$

where $\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \boldsymbol{I}_r\|_2$, $\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \boldsymbol{I}_r\|_2 \leq \delta$ with a small constant $\delta > 0$. Define $\boldsymbol{A}^{(t)} = a_1 \overline{\boldsymbol{A}}\boldsymbol{D}^{(t)} + a_2\widehat{\boldsymbol{A}}^{(t)}$ with $a_1 + a_2 = 1$ and $a_1, a_2 \geq 0$, $\boldsymbol{D}^{(t)} \in \mathcal{O}^{p \times r}$. Define a "normalization" operator $\mathcal{N}(\cdot)$ such that $\mathcal{N}(\boldsymbol{A}) := \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1/2} \in \mathcal{O}^{p \times r}$ for any $\boldsymbol{A} \in \mathbb{R}^{p \times r}$ with full column rank. Define two distances between subspaces spanned by columns of $\boldsymbol{A}$ and $\boldsymbol{B} \in \mathbb{R}^{p \times r}$ as

$$d_1(\boldsymbol{A}, \boldsymbol{B}) = \|\mathcal{N}(\boldsymbol{A})(\mathcal{N}(\boldsymbol{A}))^\top - \mathcal{N}(\boldsymbol{B})(\mathcal{N}(\boldsymbol{B}))^\top\|_2,$$
$$d_2(\boldsymbol{A}, \boldsymbol{B}) = \min_{\boldsymbol{R} \in \mathcal{O}^{r \times r}} \|\mathcal{N}(\boldsymbol{A}) - \mathcal{N}(\boldsymbol{B})\boldsymbol{R}\|_2.$$

By Lemma 2.6 of Chen et al. (2021b), $d_1(\boldsymbol{A}, \boldsymbol{B}) \leq d_2(\boldsymbol{A}, \boldsymbol{B}) \leq \sqrt{2}d_1(\boldsymbol{A}, \boldsymbol{B})$.

Note that

$$\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top = a_1^2\boldsymbol{I}_r + a_2^2\boldsymbol{I}_r + a_1a_2[(\boldsymbol{D}^{(t)})^\top(\overline{\boldsymbol{A}})^\top\widehat{\boldsymbol{A}}^{(t)} + (\widehat{\boldsymbol{A}}^{(t)})^\top\overline{\boldsymbol{A}}\boldsymbol{D}^{(t)}].$$

In addition, there exists a rotation matrix $\boldsymbol{R}^{(t)} \in \mathcal{O}^{r \times r}$ s.t.

$$\widehat{\boldsymbol{\theta}}^{(t)} = \boldsymbol{R}^{(t)}\frac{\|\widehat{\boldsymbol{\theta}}^{(t)}\|_2}{\|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}.$$

Consider $\boldsymbol{\theta}^{(t)} \in \mathbb{R}^r$ s.t.

$$\boldsymbol{\theta}^{(t)} = \alpha \cdot \frac{\widehat{\boldsymbol{\theta}}^{(t)}}{\|\widehat{\boldsymbol{\theta}}^{(t)}\|_2},$$

then by taking $\boldsymbol{D}^{(t)} = (\boldsymbol{R}^{(t)})^{-1}$, we have

$$\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} = a_1\overline{\boldsymbol{A}}(\boldsymbol{R}^{(t)})^{-1}\boldsymbol{\theta}^{(t)} + a_2\widehat{\boldsymbol{A}}^{(t)}\boldsymbol{\theta}^{(t)} = a_1\frac{\alpha}{\|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2}\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} + a_2\frac{\alpha}{\|\widehat{\boldsymbol{\theta}}^{(t)}\|_2}\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}.$$

Let $a_1 \frac{\alpha}{\|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2} + a_2 \frac{\alpha}{\|\widehat{\boldsymbol{\theta}}^{(t)}\|_2} = 1$, we obtain that $\alpha = \big(\frac{a_1}{\|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2} + \frac{a_2}{\|\widehat{\boldsymbol{\theta}}^{(t)}\|_2}\big)^{-1}$, which implies that

$$\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} = \underbrace{\frac{a_1/\|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2}{a_1/\|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2 + a_2/\|\widehat{\boldsymbol{\theta}}^{(t)}\|_2}}_{:=a_1^{(t)}} (\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}).$$

Also define $a_2^{(t)} = 1 - a_1^{(t)}$. We claim that

$$C \geq \frac{a_1^{(t)}}{a_1} = \frac{1}{a_1 + a_2 \cdot \frac{\|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2}{\|\widehat{\boldsymbol{\theta}}^{(t)}\|_2}} \gtrsim \frac{\|\widehat{\boldsymbol{\theta}}^{(t)}\|_2}{\|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2} \geq C' > 0, \quad \forall t \in S, \tag{13}$$

with some constants $C, C' > 0$. To see this, note that

$$\widehat{\boldsymbol{\theta}}^{(t)} = ((\widehat{\boldsymbol{A}}^{(t)})^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \widehat{\boldsymbol{A}}^{(t)})^{-1}(\widehat{\boldsymbol{A}}^{(t)})^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*} - ((\widehat{\boldsymbol{A}}^{(t)})^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \widehat{\boldsymbol{A}}^{(t)})^{-1}(\widehat{\boldsymbol{A}}^{(t)})^\top \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}).$$

Also note that $((\widehat{\boldsymbol{A}}^{(t)})^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \widehat{\boldsymbol{A}}^{(t)})^{-1}$ and $(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1}$ share the same eigenvalues. Since $\sigma_{\max}((\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1}) \asymp \sigma_{\min}((\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1}) \asymp 1$ w.p. at least $1 - e^{-C'(p+\log T)}$ by Lemma 18, we have

$$\sigma_{\max}\big(((\widehat{\boldsymbol{A}}^{(t)})^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \widehat{\boldsymbol{A}}^{(t)})^{-1}\big) \asymp \sigma_{\min}\big(((\widehat{\boldsymbol{A}}^{(t)})^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \widehat{\boldsymbol{A}}^{(t)})^{-1}\big) \asymp 1,$$

w.p. at least $1 - e^{-C'(p+\log T)}$. Furthermore, by Weyl's inequality and Lemma 18, with $\boldsymbol{R} \in \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\mathcal{N}(\widehat{\boldsymbol{A}}^{(t)}) - \boldsymbol{A}^{(t)*}\boldsymbol{R}\|_2$,

$$\sigma_r((\widehat{\boldsymbol{A}}^{(t)})^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A}^{(t)*})$$
$$\geq \sigma_r((\widehat{\boldsymbol{A}}^{(t)})^\top \boldsymbol{\Sigma}^{(t)} \boldsymbol{A}^{(t)*}) - \|\widehat{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma}^{(t)}\|_2$$
$$\geq \sigma_r(\boldsymbol{R}^\top(\boldsymbol{A}^{(t)*})^\top \boldsymbol{\Sigma}^{(t)} \boldsymbol{A}^{(t)*}) - \sigma_{\max}((\boldsymbol{R}^\top(\boldsymbol{A}^{(t)*})^\top - \mathcal{N}(\widehat{\boldsymbol{A}}^{(t)})^\top)\boldsymbol{\Sigma}^{(t)} \boldsymbol{A}^{(t)*}) - C\delta - C\sqrt{\frac{p+\log T}{n}}$$

$$\geq \sigma_r((\boldsymbol{A}^{(t)*})^\top \boldsymbol{\Sigma}^{(t)} \boldsymbol{A}^{(t)*}) - \|\mathcal{N}(\widehat{\boldsymbol{A}}^{(t)}) - \boldsymbol{A}^{(t)*}\boldsymbol{R}\|_2 \|\boldsymbol{\Sigma}^{(t)}\|_2 - C\delta - C\sqrt{\frac{p+\log T}{n}}$$
$$\geq \sigma_{\min}(\boldsymbol{\Sigma}^{(t)}) - \|\mathcal{N}(\widehat{\boldsymbol{A}}^{(t)}) - \boldsymbol{A}^{(t)*}\boldsymbol{R}\|_2 \cdot \sigma_{\max}(\boldsymbol{\Sigma}^{(t)}) - C\delta - C\sqrt{\frac{p+\log T}{n}}$$
$$> 0,$$

w.p. at least $1 - e^{-C'(p+\log T)}$, if $d_2(\widehat{\boldsymbol{A}}^{(t)}, \boldsymbol{A}^{(t)*}) \leq \frac{\sigma_{\min}(\boldsymbol{\Sigma}^{(t)})}{\sigma_{\max}(\boldsymbol{\Sigma}^{(t)})}(1-\gamma)$ with some $\gamma > 0$. On the other hand, by Lemma 18, w.p. at least $1 - e^{-C'(p+\log T)}$, for all $t \in S$,

$$\sigma_{\max}((\widehat{\boldsymbol{A}}^{(t)})^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A}^{(t)*}) \lesssim 1.$$

Hence w.p. at least $1 - e^{-C'(p+\log T)}$,

$$\|\widehat{\boldsymbol{\theta}}^{(t)}\|_2 \geq C\sigma_r((\widehat{\boldsymbol{A}}^{(t)})^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A}^{(t)*})\|\boldsymbol{\theta}^{(t)*}\|_2 - C'\|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2$$

62

$$\geq C''\|\boldsymbol{\theta}^{(t)*}\|_2 - C'''\sqrt{\frac{p + \log T}{n}}$$

$$\geq \frac{C''}{2}\|\boldsymbol{\theta}^{(t)*}\|_2.$$

In addition, note that w.p. at least $1 - e^{-C'(r+\log T)}$,

$$\left|\|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2 - \|\boldsymbol{\theta}^{(t)*}\|_2\right| \leq \|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \boldsymbol{\theta}^{(t)*}\|_2 \lesssim h\zeta^{(t)} + \sqrt{\frac{r + \log T}{n}},$$

implying that w.p. at least $1 - e^{-C'(r+\log T)}$,

$$\|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2 \leq \|\boldsymbol{\theta}^{(t)*}\|_2 + C\left(h\zeta^{(t)} + \sqrt{\frac{r + \log T}{n}}\right) \leq \|\boldsymbol{\theta}^{(t)*}\|_2 + C'\left(r^{-1/2}\zeta^{(t)} + \sqrt{\frac{r + \log T}{n}}\right) \lesssim \|\boldsymbol{\theta}^{(t)*}\|_2,$$

$$\|\boldsymbol{\theta}^{(t)*}\|_2 \lesssim \|\boldsymbol{\theta}^{(t)*}\|_2 - C\left(h\zeta^{(t)} + \sqrt{\frac{r + \log T}{n}}\right) \leq \|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2.$$

Hence w.p. at least $1 - e^{-C'(p+\log T)}$,

$$\|\widehat{\boldsymbol{\theta}}^{(t)}\|_2 \geq \frac{C''}{2}\|\boldsymbol{\theta}^{(t)*}\|_2 \gtrsim \|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2.$$

This proves the second half of (13). On the other hand, when $a_1 \leq 1/2$, we must have

$$\frac{a_1^{(t)}}{a_1} \leq 2\frac{\|\widehat{\boldsymbol{\theta}}^{(t)}\|_2}{\|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2},$$

and it is easy to see that w.p. at least $1 - e^{-C'(r+\log T)}$,

$$\|\widehat{\boldsymbol{\theta}}^{(t)}\|_2 \leq C\|\boldsymbol{\theta}^{(t)*}\|_2 + C\sqrt{\frac{p + \log T}{n}} \lesssim \|\boldsymbol{\theta}^{(t)*}\|_2 \lesssim \|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2,$$

which proves the first half of (13). Therefore, our claim (13) holds, and we will use it later in our proof.

Now similar to the proof of Proposition 42, we divide S into the following two index sets

$$\mathcal{A}_1 = \left\{t \in S : \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 \geq c\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2 \cdot \frac{1}{\sqrt{r}}\right\},$$

$$\mathcal{A}_2 = \left\{t \in S : \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 < c\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2 \cdot \frac{1}{\sqrt{r}}\right\},$$

where $c > 0$ is a small constant. For any index set $\mathcal{A} \subseteq [T]$, define

$$G_{\mathcal{A}}(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \overline{\boldsymbol{A}}, \boldsymbol{\Theta}) = \frac{1}{T}\sum_{t \in \mathcal{A}} f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}) + \sum_{t \in \mathcal{A}} \frac{\lambda}{\sqrt{nT}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2$$

$$= \frac{1}{2T}\sum_{t\in\mathcal{A}}\|\boldsymbol{Y}^{(t)} - \boldsymbol{X}^{(t)}\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}\|_2^2 + \frac{\lambda}{T\sqrt{n}}\sum_{t\in\mathcal{A}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \boldsymbol{A}\boldsymbol{A}^\top\|_2.$$

Note that by basic algebra,

$$\frac{1}{2T}\sum_{t\in\mathcal{A}}\|\boldsymbol{Y}^{(t)} - \boldsymbol{X}^{(t)}\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}\|_2^2 - \frac{1}{2T}\sum_{t\in\mathcal{A}}\|\boldsymbol{Y}^{(t)} - \boldsymbol{X}^{(t)}\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2^2$$

$$= \frac{1}{2T}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}) + \frac{1}{T}\sum_{t\in\mathcal{A}}[\nabla f^{(t)}(\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)})]^\top(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)})$$

$$= \frac{1}{2T}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}) + \frac{1}{T}\sum_{t\in\mathcal{A}}[\nabla f^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})]^\top(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)})$$

$$+ \frac{1}{T}\sum_{t\in\mathcal{A}}(\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}).$$

Hence

$$G_{\mathcal{A}_1}(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \boldsymbol{A}, \boldsymbol{\Theta}) - G_{\mathcal{A}_1}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})$$

$$\leq \frac{1}{2T}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)})$$

$$+ \frac{1}{T}\sum_{t\in\mathcal{A}_1}\|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 a_1^{(t)}(\|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\|_2 + \|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2)$$

$$+ \frac{1}{T}\left\|\left\{\sqrt{a_1^{(t)}}[\nabla f^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}) - \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})]\right\}_{t\in\mathcal{A}_1}\right\|_{\mathrm{F}} \cdot \sqrt{\sum_{t\in\mathcal{A}_1}a_1^{(t)}\|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2^2}$$

$$- \frac{C}{T}\sum_{t\in\mathcal{A}_1}a_1^{(t)}\|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2^2 + \frac{\lambda}{T\sqrt{n}}\sum_{t\in\mathcal{A}_1}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \boldsymbol{A}\boldsymbol{A}^\top\|_2$$

$$- \frac{\lambda}{T\sqrt{n}}\sum_{t\in\mathcal{A}_1}\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2, \tag{14}$$

with

$$[1] = \frac{1}{2T}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)})$$

$$+ \frac{1}{T}\sum_{t\in\mathcal{A}_1}\|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 a_1^{(t)}(\|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\|_2 + \|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2)$$

$$+ \frac{1}{T}\left\|\left\{\sqrt{a_1^{(t)}}[\nabla f^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}) - \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})]\right\}_{t\in\mathcal{A}_1}\right\|_{\mathrm{F}} \cdot \sqrt{\sum_{t\in\mathcal{A}_1}a_1^{(t)}\|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2^2}$$

$$- \frac{C}{T}\sum_{t\in\mathcal{A}_1}a_1^{(t)}\|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2^2,$$

and

$$[2] = \frac{\lambda}{T\sqrt{n}} \sum_{t \in \mathcal{A}_1} \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \boldsymbol{A}\boldsymbol{A}^\top\|_2 - \frac{\lambda}{T\sqrt{n}} \sum_{t \in \mathcal{A}_1} \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2. \quad (15)$$

Recall that $\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} = a_1^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)})$ and our previous conclusion that w.p. at least $1 - e^{-C'(r+\log T)}$, $a_1^{(t)}/a_1 \asymp 1$ for all $t \in S$, hence w.p. at least $1 - e^{-C'(r+\log T)}$,

$$\begin{aligned}
[1] &\leq \frac{C}{T} \sum_{t \in \mathcal{A}_1} a_1^2 \|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2^2 \\
&\quad + \frac{1}{T} \sum_{t \in \mathcal{A}_1} \|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 a_1 \zeta^{(t)} \left[d_1(\widehat{\overline{\boldsymbol{A}}}, \overline{\boldsymbol{A}}) + d_1(\widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{A}}^{(t)}) \right] \\
&\quad + \frac{1}{T} \left(\sqrt{T}\sqrt{a_1}\sqrt{\frac{r+\log T}{n}} + h\sqrt{\sum_{t \in \mathcal{A}_1}(\zeta^{(t)})^2 a_1} \right) \sqrt{\sum_{t \in \mathcal{A}_1} a_1 \|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2^2} \\
&\quad - \frac{C}{T} \sum_{t \in \mathcal{A}_1} a_1 \|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2^2 \\
&\leq \frac{C}{T} \sum_{t \in \mathcal{A}_1} a_1^2 \|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2^2 \\
&\quad + \frac{1}{T} \sum_{t \in \mathcal{A}_1} \|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 a_1 \zeta^{(t)} (\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 + \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top\|_2 + \delta) \\
&\quad + \frac{1}{T} \left(\sqrt{T}\sqrt{a_1}\sqrt{\frac{r+\log T}{n}} + h\sqrt{\sum_{t \in \mathcal{A}_1}(\zeta^{(t)})^2 a_1} \right) \sqrt{\sum_{t \in \mathcal{A}_1} a_1 \|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2^2} \\
&\quad - \frac{C}{T} \sum_{t \in \mathcal{A}_1} a_1 \|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2^2 \\
&\leq \frac{C}{T} \sum_{t \in \mathcal{A}_1} \sqrt{\frac{p+\log T}{n}} a_1 \zeta^{(t)} \sqrt{r} \cdot \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + C a_1 \left(\frac{r+\log T}{n} + h^2 \cdot \frac{1}{T} \sum_{t \in \mathcal{A}_1}(\zeta^{(t)})^2 \right) \\
&\quad + C \frac{a_1}{T} \delta \sqrt{\frac{p+\log T}{n}} \sum_{t \in \mathcal{A}_1} \zeta^{(t)}
\end{aligned}$$

To bound [2], we need to do some preparations. Note that

$$\begin{aligned}
\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top &= (a_1 \overline{\boldsymbol{A}}\boldsymbol{D}^{(t)} + a_2 \widehat{\boldsymbol{A}}^{(t)})(a_1 \overline{\boldsymbol{A}}\boldsymbol{D}^{(t)} + a_2 \widehat{\boldsymbol{A}}^{(t)})^\top \\
&= a_2^2 \widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top + a_1 a_2 [\overline{\boldsymbol{A}}\boldsymbol{D}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top + \widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top(\overline{\boldsymbol{A}})^\top] + \mathcal{O}(a_1^2) \\
&= (1 - 2a_1)\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top + a_1 [\overline{\boldsymbol{A}}\boldsymbol{D}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top + \widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top(\overline{\boldsymbol{A}})^\top] + \mathcal{O}(a_1^2),
\end{aligned}$$

where $\mathcal{O}(a_1^2)$ here refers to a $p \times p$ matrix with each entry of order $\mathcal{O}(a_1^2)$. Let $\boldsymbol{A} = a_1 \overline{\boldsymbol{A}} + a_2 \widehat{\overline{\boldsymbol{A}}}$, then similarly we have

$$\boldsymbol{A}\boldsymbol{A}^\top = (1 - 2a_1)\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top + a_1[\overline{\boldsymbol{A}}(\widehat{\overline{\boldsymbol{A}}})^\top + \widehat{\overline{\boldsymbol{A}}}(\overline{\boldsymbol{A}})^\top] + \mathcal{O}(a_1^2). \tag{16}$$

This implies that

$$\begin{aligned}
&\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \boldsymbol{A}\boldsymbol{A}^\top \\
&= (1 - a_1)[\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top] - \frac{a_1}{2}(\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top + \widehat{\overline{\boldsymbol{A}}})(\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}})^\top \\
&\quad - \frac{a_1}{2}(\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}})(\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top + \widehat{\overline{\boldsymbol{A}}})^\top + a_1\overline{\boldsymbol{A}}[\boldsymbol{D}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - (\widehat{\overline{\boldsymbol{A}}})^\top] \\
&\quad + a_1[\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}](\overline{\boldsymbol{A}})^\top + \mathcal{O}(a_1^2) \\
&= (1 - a_1)[\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top] + a_1\left[\overline{\boldsymbol{A}} - \frac{1}{2}(\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top + \widehat{\overline{\boldsymbol{A}}})\right](\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}})^\top \\
&\quad + a_1[\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}]\left[\overline{\boldsymbol{A}} - \frac{1}{2}(\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top + \widehat{\overline{\boldsymbol{A}}})\right]^\top + \mathcal{O}(a_1^2) \\
&= (1 - a_1)[\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top] + a_1(\overline{\boldsymbol{A}} - \widehat{\overline{\boldsymbol{A}}})[\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}]^\top \\
&\quad - a_1[\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}][\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}]^\top + a_1[\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}](\overline{\boldsymbol{A}} - \widehat{\overline{\boldsymbol{A}}})^\top + \mathcal{O}(a_1^2).
\end{aligned}$$

Note that we can multiply $\overline{\boldsymbol{A}}$, $\widehat{\boldsymbol{A}}^{(t)}$, and $\widehat{\overline{\boldsymbol{A}}}$ by a rotation matrix from the right without changing their definitions

$$\overline{\boldsymbol{A}} \in \underset{\boldsymbol{A} \in \mathcal{O}^{p \times r}}{\arg\min} \|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \boldsymbol{A}\boldsymbol{A}^\top\|_2,$$

$$\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}} \in \underset{\substack{\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \overline{\boldsymbol{A}} \in \mathbb{R}^{p \times r} \\ \|(\boldsymbol{A}^{(t)})^\top \boldsymbol{A}^{(t)} - \boldsymbol{I}_r\|_2, \|(\overline{\boldsymbol{A}})^\top \overline{\boldsymbol{A}} - \boldsymbol{I}_r\|_2 \le \delta}}{\arg\min} \underset{\boldsymbol{\Theta} \in \mathbb{R}^{r \times T}}{\min} G_{[T]}(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \overline{\boldsymbol{A}}, \boldsymbol{\Theta}).$$

Therefore, WLOG, we assume

$$\|\overline{\boldsymbol{A}} - \mathcal{N}(\widehat{\overline{\boldsymbol{A}}})\|_2 = \underset{\boldsymbol{R} \in \mathcal{O}^{r \times r}}{\min} \|\overline{\boldsymbol{A}} - \mathcal{N}(\widehat{\overline{\boldsymbol{A}}})\boldsymbol{R}\|_2 = d_2(\overline{\boldsymbol{A}}, \widehat{\overline{\boldsymbol{A}}}), \tag{17}$$

$$\|\mathcal{N}(\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top) - \mathcal{N}(\widehat{\overline{\boldsymbol{A}}})\|_2 = \underset{\boldsymbol{R} \in \mathcal{O}^{r \times r}}{\min} \|\mathcal{N}(\widehat{\boldsymbol{A}}^{(t)}) - \mathcal{N}(\widehat{\overline{\boldsymbol{A}}})\boldsymbol{R}\|_2 = d_2(\widehat{\boldsymbol{A}}^{(t)}, \widehat{\overline{\boldsymbol{A}}}).$$

Therefore, by the triangle inequality,

$$\begin{aligned}
\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \boldsymbol{A}\boldsymbol{A}^\top\|_2 &\le (1 - a_1)\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + \max\left\{2a_1\|\overline{\boldsymbol{A}} - \widehat{\overline{\boldsymbol{A}}}\|_2\|\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}\|_2,\right. \\
&\qquad \left. a_1\|\widehat{\boldsymbol{A}}^{(t)}(\boldsymbol{D}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}\|_2^2\right\} + C\sqrt{p}a_1^2 \\
&\le (1 - a_1)\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + \max\left\{2\sqrt{2}a_1 \cdot d_2(\overline{\boldsymbol{A}}, \widehat{\overline{\boldsymbol{A}}})d_1(\widehat{\boldsymbol{A}}^{(t)}, \widehat{\overline{\boldsymbol{A}}}),\right.
\end{aligned}$$

$$\sqrt{2}a_1 \cdot d_2(\widehat{\boldsymbol{A}}^{(t)}, \widehat{\overline{\boldsymbol{A}}})d_1(\widehat{\boldsymbol{A}}^{(t)}, \widehat{\overline{\boldsymbol{A}}})\big\} + Ca_1\delta + C\sqrt{p}a_1^2$$
$$\le \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + \Big[-1 + \big(2\sqrt{2}d_2(\overline{\boldsymbol{A}}, \widehat{\overline{\boldsymbol{A}}})\big) \vee \big(\sqrt{2}d_2(\widehat{\boldsymbol{A}}^{(t)}, \widehat{\overline{\boldsymbol{A}}})\big)\Big]$$

$$\cdot a_1 d_1(\widehat{\boldsymbol{A}}^{(t)}, \widehat{\overline{\boldsymbol{A}}}) + Ca_1\delta + C\sqrt{p}a_1^2$$
$$\le \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + \Big[-2^{-\frac{1}{2}} + \big(2d_2(\overline{\boldsymbol{A}}, \widehat{\overline{\boldsymbol{A}}})\big) \vee \big(d_2(\overline{\boldsymbol{A}}, \widehat{\overline{\boldsymbol{A}}}) + d_2(\widehat{\boldsymbol{A}}^{(t)}, \overline{\boldsymbol{A}})\big)\Big]$$

$$\cdot \sqrt{2}a_1 d_1(\widehat{\boldsymbol{A}}^{(t)}, \widehat{\overline{\boldsymbol{A}}}) + Ca_1\delta + C\sqrt{p}a_1^2$$
$$\le \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + \Big[-2^{-\frac{1}{2}} + \big(2d_2(\overline{\boldsymbol{A}}, \widehat{\overline{\boldsymbol{A}}})\big) \vee \big(d_2(\overline{\boldsymbol{A}}, \widehat{\overline{\boldsymbol{A}}}) + d_2(\widehat{\boldsymbol{A}}^{(t)}, \overline{\boldsymbol{A}})\big)\Big]$$

$$\cdot \sqrt{2}a_1 \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + Ca_1\delta + C\sqrt{p}a_1^2.$$

To further bound the RHS, we claim that $d_2(\overline{\boldsymbol{A}}, \widehat{\overline{\boldsymbol{A}}}) \le (1+\sqrt{2}) \max_{t \in S} d_2(\widehat{\boldsymbol{A}}^{(t)}, \overline{\boldsymbol{A}}) + \sqrt{2}\frac{\epsilon}{1-\epsilon} + C'\delta$ with some constant $C' > 0$. To see this, denote $q = \max_{t \in S} d_2(\widehat{\boldsymbol{A}}^{(t)}, \overline{\boldsymbol{A}})$, and suppose $d_2(\widehat{\overline{\boldsymbol{A}}}, \overline{\boldsymbol{A}}) \ge q + b$ with some $b \ge 0$, then $\min_{t \in S} d_2(\widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{A}}^{(t)}) \ge b$. This leads to

$$\sum_{t=1}^T \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top\|_2 \ge \frac{1}{\sqrt{2}}\sum_{t \in S} d_2(\widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{A}}^{(t)}) - C\delta \ge \frac{|S|}{\sqrt{2}}b - CT\delta.$$

On the other hand, we have

$$\sum_{t \in S} d_2(\widehat{\boldsymbol{A}}^{(t)}, \overline{\boldsymbol{A}}) \le |S|q, \quad \sum_{t \in S^c} d_2(\widehat{\boldsymbol{A}}^{(t)}, \overline{\boldsymbol{A}}) \le |S^c|,$$

which implies that

$$\sum_{t=1}^T \|\overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top - \widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top\|_2 \le |S|q + |S^c| + CT\delta.$$

By the optimization procedure, we know that $\widehat{\overline{\boldsymbol{A}}} \in \underset{\boldsymbol{A}\in\mathbb{R}^{p\times r}:\|\boldsymbol{A}^\top\boldsymbol{A}-\boldsymbol{I}_r\|_2\le\delta}{\arg\min} \sum_{t=1}^T \|\boldsymbol{A}\boldsymbol{A}^\top - \widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top\|_2$. Therefore we must have $|S|q + |S^c| + CT\delta \ge \frac{|S|}{\sqrt{2}}b - CT\delta$, i.e.

$$b \le \sqrt{2}q + \sqrt{2}\frac{|S^c|}{|S|} + C'\delta \le \sqrt{2}q + \sqrt{2}\frac{\epsilon}{1-\epsilon} + C'\delta.$$

In other words,

$$d_2(\overline{\boldsymbol{A}}, \widehat{\overline{\boldsymbol{A}}}) \le q + b \le (1+\sqrt{2}) \max_{t \in S} d_2(\widehat{\boldsymbol{A}}^{(t)}, \overline{\boldsymbol{A}}) + \sqrt{2}\frac{\epsilon}{1-\epsilon} + C'\delta,$$

which proves the claim. Therefore, using this claim, we have

$$\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \boldsymbol{A}\boldsymbol{A}^\top\|_2$$

$$\leq \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + \Big[-2^{-\frac{1}{2}} + 2(1+\sqrt{2})\max_{t\in S} d_2(\widehat{\boldsymbol{A}}^{(t)}, \overline{\boldsymbol{A}}) + 2\sqrt{2}\frac{\epsilon}{1-\epsilon}\Big]$$

$$\cdot \sqrt{2}a_1\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + Ca_1\delta + C\sqrt{p}a_1^2$$

$$\leq \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 - \gamma a_1\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + Ca_1\delta + C\sqrt{p}a_1^2,$$

when $2(2+\sqrt{2})\max_{t\in S} d_2(\widehat{\boldsymbol{A}}^{(t)}, \overline{\boldsymbol{A}}) + 4\frac{\epsilon}{1-\epsilon} < 1 - \gamma$.

Therefore, we can upper bound the term in (15) as

$$[2] \leq -\frac{\lambda}{T\sqrt{n}} \cdot \gamma \cdot a_1 \cdot \sum_{t\in\mathcal{A}_1} \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + C\frac{\lambda}{\sqrt{n}}\gamma \cdot a_1\delta + C\frac{\lambda}{\sqrt{n}} \cdot \gamma \cdot \sqrt{p}a_1^2.$$

Therefore, using the bounds we have proved for [1] and [2], plugging them back in (14), we have Hence

$$G_{\mathcal{A}_1}(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \boldsymbol{A}, \boldsymbol{\Theta}) - G_{\mathcal{A}_1}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})$$

$$\leq \frac{C}{T}\sum_{t\in\mathcal{A}_1} \sqrt{\frac{p+\log T}{n}}a_1\zeta^{(t)}\sqrt{r} \cdot \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + Ca_1\Big(\frac{r+\log T}{n} + h^2 \cdot \frac{1}{T}\sum_{t\in\mathcal{A}_1}(\zeta^{(t)})^2\Big)$$

$$+ C\frac{a_1}{T}\delta\sqrt{\frac{p+\log T}{n}}\sum_{t\in\mathcal{A}_1}\zeta^{(t)} - \frac{\lambda}{T\sqrt{n}} \cdot \gamma \cdot a_1 \cdot \sum_{t\in\mathcal{A}_1}\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$+ C\frac{\lambda}{\sqrt{n}}\gamma \cdot a_1\delta + C\frac{\lambda}{\sqrt{n}} \cdot \gamma \cdot \sqrt{p}a_1^2$$

$$\leq Ca_1\Big(\frac{r+\log T}{n} + h^2 \cdot \frac{1}{T}\sum_{t\in\mathcal{A}_1}(\zeta^{(t)})^2\Big) + C\frac{a_1}{T}\delta\sqrt{\frac{p+\log T}{n}}\sum_{t\in\mathcal{A}_1}\zeta^{(t)}$$

$$- \frac{\lambda}{T\sqrt{n}} \cdot \gamma \cdot a_1 \cdot \sum_{t\in\mathcal{A}_1}\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + C\frac{\lambda}{\sqrt{n}}\gamma \cdot a_1\delta + C\frac{\lambda}{\sqrt{n}} \cdot \gamma \cdot \sqrt{p}a_1^2.$$

Considering $G_{\mathcal{A}_2}(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \overline{\boldsymbol{A}}, \boldsymbol{\Theta}) - G_{\mathcal{A}_2}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})$, first we have

$$\frac{1}{T}\sum_{t\in\mathcal{A}_2}[\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})]^\top a_1^{(t)} \cdot (\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} + \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)})$$

$$\leq \frac{1}{T}\Big\|\Big\{\sqrt{a_1^{(t)}}\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\Big\}_{t\in\mathcal{A}_2}\Big\|_2 \sqrt{2r} \cdot \Big\|\Big\{\sqrt{a_1^{(t)}}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)})\Big\}_{t\in\mathcal{A}_2}\Big\|_F$$

$$+ \frac{1}{T}\sum_{t\in\mathcal{A}_2}\|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 \cdot a_1^{(t)} \cdot \|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2$$

$$\leq \frac{1}{T}\sqrt{2r}C \cdot \sqrt{\frac{p+T}{n}}a_1\Big\|\Big\{(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)})\Big\}_{t\in\mathcal{A}_2}\Big\|_F + \frac{1}{T}\sum_{t\in\mathcal{A}_2}\sqrt{\frac{p+\log T}{n}}a_1 \cdot \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$\leq C\frac{r(p+T)}{nT}a_1 + \frac{C}{T}a_1\Big\|\Big\{(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)})\Big\}_{t\in\mathcal{A}_2}\Big\|_F^2 + \frac{1}{T}\sum_{t\in\mathcal{A}_2}\sqrt{\frac{p+\log T}{n}}a_1 \cdot \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2,$$

where the second inequality holds because

$$\left\|\left\{\sqrt{a_1^{(t)}}\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\right\}_{t\in\mathcal{A}_2}\right\|_2 \lesssim \max_{t\in\mathcal{A}_2}\sqrt{a_1^{(t)}}\cdot\sqrt{\frac{p+T}{n}},$$

w.p. at least $1-e^{-C(p+\log T)}$, which is due to Lemma 5.39 in Vershynin (2010). Therefore, by using the same arguments to bound $G_{\mathcal{A}_1}(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \overline{\boldsymbol{A}}, \boldsymbol{\Theta}) - G_{\mathcal{A}_1}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})$, we can prove that

$$G_{\mathcal{A}_2}(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \boldsymbol{A}, \boldsymbol{\Theta}) - G_{\mathcal{A}_2}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})$$

$$\leq \frac{1}{2T}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)})$$

$$+ C\frac{r(p+T)}{nT}a_1^2 + \frac{C}{T}\left\|\left\{(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)})\right\}_{t\in\mathcal{A}_2}\right\|_F^2 + \frac{1}{T}\sum_{t\in\mathcal{A}_2}\sqrt{\frac{p+\log T}{n}}a_1\cdot\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$+ \frac{C}{T}\left(\sqrt{T}\sqrt{a_1}\sqrt{\frac{r+\log T}{n}} + h\sqrt{\sum_{t\in\mathcal{A}_2}(\zeta^{(t)})^2 a_1}\right)\sqrt{\sum_{t\in\mathcal{A}_2}a_1\|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2^2}$$

$$- \frac{C}{T}\sum_{t\in\mathcal{A}_2}a_1\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2^2 - \frac{\lambda}{T\sqrt{n}}\cdot\gamma a_1\cdot\sum_{t\in\mathcal{A}_2}\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 \qquad (18)$$

$$\leq -\frac{C}{T}\sum_{t\in\mathcal{A}_2}a_1\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2^2 - \frac{\lambda}{T\sqrt{n}}\cdot\gamma a_1\cdot\sum_{t\in\mathcal{A}_2}\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$+ Ca_1\left(\frac{pr}{nT} + \frac{r+\log T}{n} + h^2\cdot\frac{1}{T}\sum_{t\in\mathcal{A}_2}(\zeta^{(t)})^2\right).$$

Since $\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\|_2 \leq \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2(\|\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\|_2 + \|\nabla f^{(t)}(\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)})\|_2) \leq C\zeta^{(t)}\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 \leq C\zeta^{(t)}\|\overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2\frac{c}{\sqrt{r}}$, we have

$$\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2^2 \geq \frac{1}{2}\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2^2 - 2\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\|_2^2$$

$$\geq \frac{1}{2}\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2^2 - \frac{2C^2c}{r}(\zeta^{(t)})^2\|\overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2^2.$$

This implies that

$$G_{\mathcal{A}_2}(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \boldsymbol{A}, \boldsymbol{\Theta}) - G_{\mathcal{A}_2}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})$$

$$\leq -\frac{C'}{T}a_1\|\widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{A}}\overline{\boldsymbol{\Theta}}\|_F^2 + \frac{2C^2c}{r}a_1(\zeta^{(t)})^2\|\overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2^2$$

$$+ Ca_1\left(\frac{pr}{nT} + \frac{r+\log T}{n} + h^2\cdot\frac{1}{T}\sum_{t\in\mathcal{A}_2}(\zeta^{(t)})^2\right).$$

In addition, by triangle inequality,

$$G_{S^c}(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \boldsymbol{A}, \boldsymbol{\Theta}) - G_{S^c}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}}) \le \frac{\lambda}{\sqrt{nT}}|S^c| \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \boldsymbol{A}\boldsymbol{A}^\top\|_2.$$

Recall (16):

$$\boldsymbol{A}\boldsymbol{A}^\top = (1 - 2a_1)\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top + a_1[\overline{\boldsymbol{A}}(\widehat{\overline{\boldsymbol{A}}})^\top + \widehat{\overline{\boldsymbol{A}}}(\overline{\boldsymbol{A}})^\top] + \mathcal{O}(a_1^2).$$

Hecne we have

$$\begin{aligned}
\|\boldsymbol{A}\boldsymbol{A}^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 &\le a_1\|\widehat{\overline{\boldsymbol{A}}}(\overline{\boldsymbol{A}} - \widehat{\overline{\boldsymbol{A}}})^\top\|_2 + a_1\|\overline{\boldsymbol{A}}(\widehat{\overline{\boldsymbol{A}}} - \overline{\boldsymbol{A}})^\top\|_2 + Ca_1^2 \\
&\le 2a_1 d_2(\overline{\boldsymbol{A}}, \widehat{\overline{\boldsymbol{A}}}) + Ca_1\delta + Ca_1^2 \\
&\le 2\sqrt{2}a_1\|\overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + Ca_1\delta + Ca_1^2,
\end{aligned}$$

where the second inequality is due to (17). Therefore,

$$G_{S^c}(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \boldsymbol{A}, \boldsymbol{\Theta}) - G_{S^c}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}}) \le 2\sqrt{2}a_1\frac{\lambda}{\sqrt{nT}}|S^c|\|\overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + Ca_1\delta + Ca_1^2.$$

Case 1: If $|\mathcal{A}_2| \ge \left(1 - cr^{-1}\frac{\bar{\zeta}^2}{\max_{t\in S}(\zeta^{(t)})^2}\right)|S|$:

By Wedin's $\sin\Theta$-theorem, we have By Assumption 2, Wedin's $\sin\Theta$-Theorem, and applying Lemma 35 on $\{\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}$,

$$\|\mathcal{N}(\widehat{\overline{\boldsymbol{A}}})(\mathcal{N}(\widehat{\overline{\boldsymbol{A}}}))^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \le \|\mathcal{N}(\widehat{\overline{\boldsymbol{A}}})(\mathcal{N}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_{\mathrm{F}} \lesssim \sqrt{\frac{r}{|S|}}\bar{\zeta}^{-1}\cdot\|\{\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\|_{\mathrm{F}}.$$

Here we used the fact that $\frac{1}{|S|}\sum_{t\in S}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}(\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})^\top \succeq \frac{c}{r}\bar{\zeta}^2\boldsymbol{I}_p$. To see this, notice that

$$\begin{aligned}
\sigma_r\left(\frac{1}{\sqrt{|S|}}\{\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in S}\right) &= \sigma_r\left(\frac{1}{\sqrt{|S|}}\{\boldsymbol{\beta}^{(t)*}\}_{t\in S}\right) - \frac{1}{\sqrt{|S|}}\|\{\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in S} - \{\boldsymbol{\beta}^{(t)*}\}_{t\in S}\|_2 \\
&\ge \frac{c}{\sqrt{r}}\bar{\zeta} - \left(\bar{\zeta}h + \sqrt{\frac{r + \log T}{n}}\right) \\
&\ge \frac{c'}{\sqrt{r}}\bar{\zeta},
\end{aligned}$$

where by Lemmas 29 and 33,

$$\frac{1}{\sqrt{|S|}}\|\{\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in S} - \{\boldsymbol{\beta}^{(t)*}\}_{t\in S}\|_2 \le \sqrt{\sum_{t\in S}\|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2} \lesssim h\bar{\zeta} + \sqrt{\frac{r + \log T}{n}},$$

and the last step comes from conditions (i) and (ii) by noticing that

$$\sqrt{\frac{r + \log T}{n}} \lesssim \frac{\bar{\zeta}}{\sqrt{r}}\cdot\sqrt{\frac{p + \log T}{n}}\cdot\frac{1}{\zeta^{(t)}} \lesssim \frac{\bar{\zeta}}{\sqrt{r}}.$$

This leads to

$$\sigma_r\left(\frac{1}{|S|}\sum_{t\in S}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}(\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})^\top\right) = \sigma_r\left(\frac{1}{|S|}\sum_{t\in S}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}(\overline{\boldsymbol{A}})^\top\overline{\boldsymbol{A}}(\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})^\top\right) \ge \frac{c'}{r}\bar{\zeta}^2.$$

Hence

$$G_{[T]}(\{\boldsymbol{A}^{(t)}\}_{t=1}^{T}, \boldsymbol{A}, \boldsymbol{\Theta}) - G_{[T]}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^{T}, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})$$

$$\leq [G_{\mathcal{A}_1}(\{\boldsymbol{A}^{(t)}\}_{t=1}^{T}, \boldsymbol{A}, \boldsymbol{\Theta}) - G_{\mathcal{A}_1}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^{T}, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})] + [G_{\mathcal{A}_2}(\{\boldsymbol{A}^{(t)}\}_{t=1}^{T}, \boldsymbol{A}, \boldsymbol{\Theta}) - G_{\mathcal{A}_2}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^{T}, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})]$$

$$+ [G_{S^c}(\{\boldsymbol{A}^{(t)}\}_{t=1}^{T}, \boldsymbol{A}, \boldsymbol{\Theta}) - G_{S^c}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^{T}, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})]$$

$$\leq Ca_1\left(\frac{r + \log T}{n} + h^2 \cdot \frac{1}{T}\sum_{t \in \mathcal{A}_1}(\zeta^{(t)})^2\right) + C\frac{a_1}{T}\delta\sqrt{\frac{p + \log T}{n}}\sum_{t \in \mathcal{A}_1}\zeta^{(t)} + C\frac{\lambda}{\sqrt{n}}\gamma \cdot a_1\delta + C\frac{\lambda}{\sqrt{n}} \cdot \gamma \cdot \sqrt{p}a_1^2$$

$$- \frac{C}{T}\sum_{t \in \mathcal{A}_2}a_1\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2^2 + Ca_1\left(\frac{pr}{nT} + \frac{r + \log T}{n} + h^2 \cdot \frac{1}{T}\sum_{t \in \mathcal{A}_2}(\zeta^{(t)})^2\right)$$

$$+ 2\sqrt{2}a_1\frac{\lambda}{\sqrt{nT}}|S^c|\|\overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$\leq Ca_1\left(\frac{pr}{nT} + \frac{r + \log T}{n} + h^2\bar{\zeta}^2\right) - Ca_1\frac{\bar{\zeta}^2}{r}\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2^2 + 2\sqrt{2}a_1\frac{\lambda}{\sqrt{nT}}|S^c|\|\overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$+ Ca_1\delta^2 + C\frac{a_1}{T}\delta\sqrt{\frac{p + \log T}{n}}\sum_{t \in S}\zeta^{(t)} + C\frac{\lambda}{\sqrt{n}}\gamma \cdot a_1\delta + C\frac{\lambda}{\sqrt{n}} \cdot \gamma \cdot \sqrt{p}a_1^2$$

$$< 0, \tag{19}$$

w.p. at least $1 - e^{-C'(r + \log T)}$, if $\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 > C\bar{\zeta}^{-1}\left(\sqrt{\frac{pr}{nT}} + \sqrt{\frac{r + \log T}{n}}\right) + Ch + C\frac{r\lambda}{\bar{\zeta}^2\sqrt{n}}\frac{|S^c|}{T}$ and $Ca_1\delta^2 + C\frac{a_1}{T}\delta\sqrt{\frac{p + \log T}{n}}\sum_{t \in S}\zeta^{(t)} + C\frac{\lambda}{\sqrt{n}}\gamma \cdot a_1\delta + C\frac{\lambda}{\sqrt{n}} \cdot \gamma \cdot \sqrt{p}a_1^2 < Ca_1\frac{pr}{nT} + Ca_1\frac{r + \log T}{n}$. Note that the latter can easily hold because we are free to choose an arbitrary small a_1 and δ. And as $a_1 \to 0+$, we have $\boldsymbol{A}^{(t)} \to \widehat{\boldsymbol{A}}^{(t)}$, $\boldsymbol{A} \to \widehat{\overline{\boldsymbol{A}}}$, $\boldsymbol{\theta}^{(t)} \to \widehat{\boldsymbol{\theta}}^{(t)}$, in the sense of $\|\cdot\|_{\max}$. Therefore (19) contradicts with the local optimality of $(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^{T}, \widehat{\overline{\boldsymbol{A}}}, \{\widehat{\boldsymbol{\theta}}^{(t)}\}_{t=1}^{T})$ in terms of $G_{[T]}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^{T}, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})$. Therefore, w.p. at least $1 - e^{-C'(r + \log T)}$, we must have

$$\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \leq C\bar{\zeta}^{-1}\left(r\sqrt{\frac{p}{nT}} + \sqrt{r}\sqrt{\frac{r + \log T}{n}}\right) + C\sqrt{r}h + C\frac{r\lambda}{\bar{\zeta}^2\sqrt{n}}\frac{|S^c|}{T},$$

when $Ca_1\delta^2 + C\frac{a_1}{T}\delta\sqrt{\frac{p + \log T}{n}}\sum_{t \in S}\zeta^{(t)} + C\frac{\lambda}{\sqrt{n}}\gamma \cdot a_1\delta + C\frac{\lambda}{\sqrt{n}} \cdot \gamma \cdot \sqrt{p}a_1^2 < Ca_1\frac{pr}{nT} + Ca_1\frac{r + \log T}{n}$.

Case 2: If $|\mathcal{A}_1| \geq cr^{-1}\frac{\bar{\zeta}^2}{\max_{t \in S}(\zeta^{(t)})^2}|S|$: Note that by (18), we have

$$G_{\mathcal{A}_2}(\{\boldsymbol{A}^{(t)}\}_{t=1}^{T}, \boldsymbol{A}, \boldsymbol{\Theta}) - G_{\mathcal{A}_2}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^{T}, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})$$

$$\leq \frac{1}{2T}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)})$$

$$+ \frac{C}{T}a_1\sqrt{\frac{p + T}{n}}\sqrt{r}\left\|\left\{\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\right\}_{t \in \mathcal{A}_2}\right\|_{\mathrm{F}} + \frac{1}{T}\sum_{t \in \mathcal{A}_2}\sqrt{\frac{p + \log T}{n}}a_1 \cdot \|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$+ \frac{C}{T}\left(\sqrt{T}\sqrt{a_1}\sqrt{\frac{r+\log T}{n}} + h\sqrt{\sum_{t\in\mathcal{A}_2}(\zeta^{(t)})^2 a_1}\right)\sqrt{\sum_{t\in\mathcal{A}_2} a_1\|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2^2}$$

$$- \frac{C}{T}\sum_{t\in\mathcal{A}_2} a_1\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2^2 - \frac{\lambda}{T\sqrt{n}}\cdot\gamma a_1\cdot\sum_{t\in\mathcal{A}_2}\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$\leq -\frac{C}{T}\sum_{t\in\mathcal{A}_2} a_1\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2^2 - \frac{\lambda}{T\sqrt{n}}\cdot\gamma a_1\cdot\sum_{t\in\mathcal{A}_2}\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$+ \frac{C}{T}\left(\sqrt{T}\sqrt{a_1}\sqrt{\frac{r+\log T}{n}} + h\sqrt{\sum_{t\in\mathcal{A}_2}(\zeta^{(t)})^2 a_1}\right)\sqrt{\sum_{t\in\mathcal{A}_2} a_1\|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2^2}$$

$$+ \frac{C}{T}a_1\sqrt{\frac{p+T}{n}}\sqrt{r}\left\|\{\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\}_{t\in\mathcal{A}_2}\right\|_{\mathrm{F}} + \frac{C}{T}a_1\sqrt{\frac{p+T}{n}}\sqrt{r}\left\|\{\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\}_{t\in\mathcal{A}_2}\right\|_{\mathrm{F}}$$

$$\leq C a_1\left(\frac{pr}{nT} + \frac{r+\log T}{n} + h^2\bar{\zeta}^2\right) + \frac{C}{T}a_1\sqrt{\frac{p+T}{n}}\sqrt{r}\sqrt{\sum_{t\in\mathcal{A}_2}\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\|_2^2}$$

$$\leq C a_1\left(\frac{pr}{nT} + \frac{r+\log T}{n} + h^2\bar{\zeta}^2\right) + C a_1\sqrt{\frac{p+T}{nT}}\bar{\zeta}\cdot\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2.$$

Therefore,

$$G_{[T]}(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \boldsymbol{A}, \boldsymbol{\Theta}) - G_{[T]}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})$$

$$\leq [G_{\mathcal{A}_1}(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \boldsymbol{A}, \boldsymbol{\Theta}) - G_{\mathcal{A}_1}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})] + [G_{\mathcal{A}_2}(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \boldsymbol{A}, \boldsymbol{\Theta}) - G_{\mathcal{A}_2}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})]$$

$$+ [G_{S^c}(\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \boldsymbol{A}, \boldsymbol{\Theta}) - G_{S^c}(\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}}, \widehat{\boldsymbol{\Theta}})]$$

$$\leq C a_1\left(\frac{r+\log T}{n} + h^2\cdot\frac{1}{T}\sum_{t\in\mathcal{A}_1}(\zeta^{(t)})^2\right) + C\frac{a_1}{T}\delta\sqrt{\frac{p+\log T}{n}}\sum_{t\in\mathcal{A}_1}\zeta^{(t)} + C\frac{\lambda}{\sqrt{n}}\gamma\cdot a_1\delta + C\frac{\lambda}{\sqrt{n}}\cdot\gamma\cdot\sqrt{p}a_1^2$$

$$- \frac{\lambda}{T\sqrt{n}}\cdot\gamma\cdot a_1\cdot\sum_{t\in\mathcal{A}_1}\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + C a_1\left(\frac{pr}{nT} + \frac{r+\log T}{n} + h^2\bar{\zeta}^2\right)$$

$$+ 2\sqrt{2}a_1\frac{\lambda}{\sqrt{n}T}|S^c|\|\overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + C a_1\sqrt{\frac{p+T}{nT}}\bar{\zeta}\cdot\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2$$

$$\leq C a_1\left(\frac{pr}{nT} + \frac{r+\log T}{n} + h^2\bar{\zeta}^2\right) + 2\sqrt{2}a_1\frac{\lambda}{\sqrt{n}T}|S^c|\|\overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$- \frac{\lambda}{T\sqrt{n}}\cdot\gamma\cdot a_1\cdot\sum_{t\in\mathcal{A}_1}\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + C\frac{a_1}{T}\delta\sqrt{\frac{p+\log T}{n}}\sum_{t\in S}\zeta^{(t)}$$

$$+ C\frac{\lambda}{\sqrt{n}}\gamma\cdot a_1\delta + C\frac{\lambda}{\sqrt{n}}\cdot\gamma\cdot\sqrt{p}a_1^2 + C a_1\sqrt{\frac{p+T}{nT}}\bar{\zeta}\cdot\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2$$

$$\leq C a_1\left(\frac{pr}{nT} + \frac{r+\log T}{n} + h^2\bar{\zeta}^2\right) + \frac{\lambda}{\sqrt{n}T}a_1\left(2\sqrt{2}|S^c| - |\mathcal{A}_1|\gamma\right)\cdot\|\overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$+ C\frac{\lambda}{\sqrt{n}}\gamma \cdot a_1\delta + C\frac{\lambda}{\sqrt{n}} \cdot \gamma \cdot \sqrt{p}a_1^2 + C\frac{a_1}{T}\delta\sqrt{\frac{p+\log T}{n}}\sum_{t\in S}\zeta^{(t)} + Ca_1\sqrt{\frac{p+T}{nT}}\bar{\zeta} \cdot \|\widehat{\overline{\bm{A}}}(\widehat{\overline{\bm{A}}})^\top - \overline{\bm{A}}(\overline{\bm{A}})^\top\|_2$$

$$\leq Ca_1\frac{p+\log T}{n} - C'\frac{\lambda}{\sqrt{n}Tr} \cdot \gamma\frac{\bar{\zeta}^2}{\max_{t\in S}(\zeta^{(t)})^2} \cdot \|\overline{\bm{A}}(\overline{\bm{A}})^\top - \widehat{\overline{\bm{A}}}(\widehat{\overline{\bm{A}}})^\top\|_2 + C\frac{\lambda}{\sqrt{n}}\gamma \cdot a_1\delta + C\frac{\lambda}{\sqrt{n}} \cdot \gamma \cdot \sqrt{p}a_1^2$$

$$+ C\frac{a_1}{T}\delta\sqrt{\frac{p+\log T}{n}}\sum_{t\in S}\zeta^{(t)} + Ca_1\sqrt{\frac{p+T}{nT}}\bar{\zeta} \cdot \|\widehat{\overline{\bm{A}}}(\widehat{\overline{\bm{A}}})^\top - \overline{\bm{A}}(\overline{\bm{A}})^\top\|_2$$

$$\leq Ca_1\frac{p+\log T}{n} - \frac{C'}{2}\frac{\lambda}{\sqrt{n}Tr} \cdot \gamma\frac{\bar{\zeta}^2}{\max_{t\in S}(\zeta^{(t)})^2} \cdot \|\overline{\bm{A}}(\overline{\bm{A}})^\top - \widehat{\overline{\bm{A}}}(\widehat{\overline{\bm{A}}})^\top\|_2 + C\frac{\lambda}{\sqrt{n}}\gamma \cdot a_1\delta + C\frac{\lambda}{\sqrt{n}} \cdot \gamma \cdot \sqrt{p}a_1^2$$

$$+ C\frac{a_1}{T}\delta\sqrt{\frac{p+\log T}{n}}\sum_{t\in S}\zeta^{(t)}$$

$$< 0,$$

w.p. at least $1 - e^{-C'(r+\log T)}$, when $\|\overline{\bm{A}}(\overline{\bm{A}})^\top - \widehat{\overline{\bm{A}}}(\widehat{\overline{\bm{A}}})^\top\|_2 \geq C\eta$, $C\frac{\lambda}{\sqrt{n}}\gamma \cdot a_1\delta + C\frac{\lambda}{\sqrt{n}} \cdot \gamma \cdot \sqrt{p}a_1^2 < Ca_1\frac{p+\log T}{n}$, with a large $C'' > 0$, and $\min_{t\in S}\zeta^{(t)} \gtrsim \sqrt{\frac{p+\log T}{n}}$. Note that

$$\frac{\min_{t\in S}\zeta^{(t)}}{\bar{\zeta}} \cdot r \lesssim \sqrt{p+\log T}, \quad \frac{\min_{t\in S}\zeta^{(t)}}{\bar{\zeta}} \cdot \frac{r}{\sqrt{T}} \lesssim 1. \tag{20}$$

We used the fact that

$$\lambda \asymp \sqrt{r(p+\log T)} \cdot \frac{\max_{t\in S}(\zeta^{(t)})^2}{\min_{t\in S}\zeta^{(t)}}$$

$$\gtrsim \left[\sqrt{pr}\frac{\bar{\zeta}}{\min_{t\in S}\zeta^{(t)}} + \sqrt{r(p+\log T)} \cdot \frac{\bar{\zeta}}{\min_{t\in S}\zeta^{(t)}}\right] \cdot \frac{\max_{t\in S}(\zeta^{(t)})^2}{\bar{\zeta}},$$

where we used (20) in the second inequality. This contradicts with the **local** optimality of $(\{\widehat{\bm{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\bm{A}}}, \{\widehat{\bm{\theta}}^{(t)}\}_{t=1}^T)$ in terms of $G_{[T]}(\{\widehat{\bm{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\bm{A}}}, \widehat{\bm{\Theta}})$.

Therefore, w.p. at least $1 - e^{-C'(r+\log T)}$, we must have

$$\|\widehat{\overline{\bm{A}}}(\widehat{\overline{\bm{A}}})^\top - \overline{\bm{A}}(\overline{\bm{A}})^\top\|_2 \leq C\bar{\zeta}^{-1}\left(r\sqrt{\frac{p}{nT}} + \sqrt{r}\sqrt{\frac{r+\log T}{n}}\right) + C\sqrt{r}h + C\frac{r\lambda}{\bar{\zeta}^2\sqrt{n}}\frac{|S^c|}{T},$$

when $C\frac{a_1}{T}\delta\sqrt{\frac{p+\log T}{n}}\sum_{t\in S}\zeta^{(t)} + C\frac{\lambda}{\sqrt{n}}\gamma \cdot a_1\delta + C\frac{\lambda}{\sqrt{n}} \cdot \gamma \cdot \sqrt{p}a_1^2 < Ca_1\frac{pr}{nT} + Ca_1\frac{r+\log T}{n}$.

Until now, we have proved that

$$\|\widehat{\overline{\bm{A}}}(\widehat{\overline{\bm{A}}})^\top - \overline{\bm{A}}(\overline{\bm{A}})^\top\|_2 \leq C\bar{\zeta}^{-1}\left(r\sqrt{\frac{p}{nT}} + \sqrt{r}\sqrt{\frac{r+\log T}{n}}\right) + C\sqrt{r}h + C\frac{r\lambda}{\bar{\zeta}^2\sqrt{n}}\frac{|S^c|}{T},$$

for any local minimizers $(\{\widehat{\bm{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\bm{A}}}, \{\widehat{\bm{\theta}}^{(t)}\}_{t=1}^T)$ of $G_{[T]}(\{\widehat{\bm{A}}^{(t)}\}_{t=1}^T, \widehat{\overline{\bm{A}}}, \widehat{\bm{\Theta}})$ satisfying

$$2(2+\sqrt{2})\max_{t\in S}d_2(\widehat{\bm{A}}^{(t)}, \overline{\bm{A}}) + 4\frac{\epsilon}{1-\epsilon} < 1 - \gamma,$$

$$\max_{t \in S} \left\{ \frac{\sigma_{\max}(\mathbf{\Sigma}^{(t)})}{\sigma_{\min}(\mathbf{\Sigma}^{(t)})} d_2(\widehat{\mathbf{A}}^{(t)}, \mathbf{A}^{(t)*}) \right\} < 1 - \gamma.$$

The remaining proof follows the arguments in the proof of Theorem 2 and pushing $\delta \to 0$.

D.7 Proof of Theorem 6

We briefly point out how the arguments in the proofs of theorems in Section 2.2.1 can be modified to prove the same results for the revised Algorithm 1 which solves Step 1 in the entire $\mathbb{R}^{p \times r}$ space with the penalty on projection matrices.

   We first discuss the results of global minimizers (Theorems 2 and 3).

(i) If $\mathrm{rank}(\widehat{\mathbf{A}}^{(t)}) = \mathrm{rank}(\widehat{\overline{\mathbf{A}}}) = r$ for all $t \in S$, then we can replace $\widehat{\mathbf{A}}^{(t)} \boldsymbol{\theta}^{(t)}_{\widehat{\mathbf{A}}^{(t)}}$ and $\widehat{\overline{\mathbf{A}}} \boldsymbol{\theta}^{(t)}_{\widehat{\overline{\mathbf{A}}}}$ with $\boldsymbol{\theta}^{(t)}_{\mathbf{A}} := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^r} f^{(t)}(\mathbf{A}\boldsymbol{\theta})$ used in the proofs by $\mathcal{N}(\widehat{\mathbf{A}}^{(t)}) \boldsymbol{\theta}^{(t)}_{\mathcal{N}(\widehat{\mathbf{A}}^{(t)})}$ and $\mathcal{N}(\widehat{\overline{\mathbf{A}}}) \boldsymbol{\theta}^{(t)}_{\mathcal{N}(\widehat{\overline{\mathbf{A}}})}$, where $\mathcal{N}(\widehat{\mathbf{A}}^{(t)}) := \widehat{\mathbf{A}}^{(t)}[(\widehat{\mathbf{A}}^{(t)})^\top \widehat{\mathbf{A}}^{(t)}]^{-1}(\widehat{\mathbf{A}}^{(t)})^\top$, and $\mathcal{N}(\widehat{\overline{\mathbf{A}}}) = \widehat{\overline{\mathbf{A}}}[(\widehat{\overline{\mathbf{A}}})^\top \widehat{\overline{\mathbf{A}}}]^{-1}(\widehat{\overline{\mathbf{A}}})^\top$. It is straightforward to see that $\mathcal{N}(\widehat{\mathbf{A}}^{(t)}) \boldsymbol{\theta}^{(t)}_{\mathcal{N}(\widehat{\mathbf{A}}^{(t)})} = \widehat{\mathbf{A}}^{(t)} \boldsymbol{\theta}^{(t)}_{\widehat{\mathbf{A}}^{(t)}}$ and $\mathcal{N}(\widehat{\overline{\mathbf{A}}}) \boldsymbol{\theta}^{(t)}_{\mathcal{N}(\widehat{\overline{\mathbf{A}}})} = \widehat{\overline{\mathbf{A}}} \boldsymbol{\theta}^{(t)}_{\widehat{\overline{\mathbf{A}}}}$. We can also see that $\mathbf{P}_{\widehat{\mathbf{A}}^{(t)}} = \mathcal{N}(\widehat{\mathbf{A}}^{(t)})(\mathcal{N}(\widehat{\mathbf{A}}^{(t)}))^\top$ and $\mathbf{P}_{\widehat{\overline{\mathbf{A}}}} = \mathcal{N}(\widehat{\overline{\mathbf{A}}})(\mathcal{N}(\widehat{\overline{\mathbf{A}}}))^\top$. We can proceed the same proofs of Theorems 2 and 3, with $\mathcal{N}(\widehat{\mathbf{A}}^{(t)})$, $\mathcal{N}(\widehat{\overline{\mathbf{A}}})$, $\boldsymbol{\theta}^{(t)}_{\widehat{\mathbf{A}}^{(t)}}$, and $\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\mathbf{A}}}}$.

(ii) If $\mathrm{rank}(\widehat{\mathbf{A}}^{(t)}) < r$ or $\mathrm{rank}(\widehat{\overline{\mathbf{A}}}) < r$ for some $t \in S$, then for those t, we can define $\mathcal{N}(\widehat{\mathbf{A}}^{(t)})$ and $\mathcal{N}(\widehat{\overline{\mathbf{A}}})$ in a different way. Consider a matrix $\mathbf{A} \in \mathbb{R}^{p \times r}$ with $\mathrm{rank}(\mathbf{A}) = r' < r$ and its SVD $\mathbf{A}_{p \times r} = \mathbf{U}_{p \times r'} \mathbf{\Lambda}_{r' \times r'} \mathbf{V}^\top_{r' \times r'}$, where $\mathbf{\Lambda}_{r' \times r'}$ is diagonal with positive entries and $\mathbf{U} \in \mathcal{O}^{p \times r'}$. Then we can define $\mathcal{N}(\mathbf{A}) = (\mathbf{U}_{p \times r'} \quad \mathbf{0}_{p \times (r-r')}) \in \mathbb{R}^{p \times r}$. Note that $\mathcal{N}(\mathbf{A})(\mathcal{N}(\mathbf{A}))^\top = \mathbf{U}\mathbf{U}^\top = \mathbf{P}_{\mathbf{A}} = \mathbf{P}_{\mathcal{N}(\mathbf{A})}$, and this definition of $\mathcal{N}(\mathbf{A})$ can be seen as an extension of the full-rank case. Then we can proceed the same proofs of Theorems 2 and 3, with $\mathcal{N}(\widehat{\mathbf{A}}^{(t)})$, $\mathcal{N}(\widehat{\overline{\mathbf{A}}})$, $\boldsymbol{\theta}^{(t)}_{\widehat{\mathbf{A}}^{(t)}}$, and $\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\mathbf{A}}}}$.

Then let us discuss the results of local minimizers (Theorems 4 and 5). Recall that we proved Theorems 4 and 5 by construction based on the local minimizers $\widehat{\overline{\mathbf{A}}}$ and $\widehat{\mathbf{A}}^{(t)}$. And it suffices to replace $\widehat{\overline{\mathbf{A}}}$ and $\widehat{\mathbf{A}}^{(t)}$ in the construction with $\mathcal{N}(\widehat{\overline{\mathbf{A}}})$ and $\mathcal{N}\widehat{\mathbf{A}}^{(t)})$ as defined above. In addition, if we want to show the stronger version of Theorem 5 we proved in the proof, we need to rewrite the constraint set by replacing

$$2(2 + \sqrt{2}) \max_{t \in S} \min_{\mathbf{R} \in \mathcal{O}^{r \times r}} \|\widehat{\mathbf{A}}^{(t)} - \overline{\mathbf{A}} \mathbf{R}\|_2 + 4\frac{\epsilon}{1-\epsilon} < 1 - \gamma,$$

$$\max_{t \in S} \left\{ \frac{\sigma_{\max}(\mathbf{\Sigma}^{(t)})}{\sigma_{\min}(\mathbf{\Sigma}^{(t)})} \min_{\mathbf{R} \in \mathcal{O}^{r \times r}} \|\widehat{\mathbf{A}}^{(t)} - \mathbf{A}^{(t)*} \mathbf{R}\|_2 \right\} < 1 - \gamma.$$

with

$$2(2 + \sqrt{2}) \max_{t \in S} \min_{\mathbf{R} \in \mathcal{O}^{r \times r}} \|\mathcal{N}(\widehat{\mathbf{A}}^{(t)}) - \overline{\mathbf{A}} \mathbf{R}\|_2 + 4\frac{\epsilon}{1-\epsilon} < 1 - \gamma,$$

$$\max_{t \in S} \left\{ \frac{\sigma_{\max}(\mathbf{\Sigma}^{(t)})}{\sigma_{\min}(\mathbf{\Sigma}^{(t)})} \min_{\mathbf{R} \in \mathcal{O}^{r \times r}} \|\mathcal{N}(\widehat{\mathbf{A}}^{(t)}) - \mathbf{A}^{(t)*} \mathbf{R}\|_2 \right\} < 1 - \gamma.$$

All the remaining arguments still hold.

D.8 Proof of Theorem 7

Let us rewrite $\boldsymbol{\beta}^{(t)*} = \overline{\boldsymbol{A}}\bar{\boldsymbol{\theta}}^{(t)*} + \boldsymbol{\delta}^{(t)*}$, where $\bar{\boldsymbol{\theta}}^{(t)*} = \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$ and $\boldsymbol{\delta}^{(t)*} = \overline{\boldsymbol{A}}^\perp(\overline{\boldsymbol{A}}^\perp)^\top \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$. Denote $D_S^* = \{\boldsymbol{\delta}^{(t)*}\}_{t\in S} \in \mathbb{R}^{d\times|S|}$.

First, let us prove a useful result: When $nT \geq Cpr$, $n \geq Cp$, $\frac{\min_{t\in S}\zeta^{(t)}}{\bar{\zeta}}\sqrt{\frac{r}{T}}\|\boldsymbol{D}_S^*\|_2 \leq c'\sqrt{\frac{p+\log T}{n}}$, and $\min_{t\in S}\|\boldsymbol{\delta}^{(t)*}\|_2 \leq c'\sqrt{\frac{p+\log T}{n}}$, with a large constant C and a small constant c', we have

$$\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \lesssim \bar{\zeta}^{-1}\sqrt{\frac{pr}{nT}} + \bar{\zeta}^{-1}\sqrt{\frac{r}{n}} + \bar{\zeta}^{-1}\sqrt{\frac{r}{T}}\|\boldsymbol{D}_S^*\|_2 + \frac{\max_{t\in S}\zeta^{(t)}}{\bar{\zeta}}\sqrt{r\bar{\epsilon}}, \quad (21)$$

w.p. at least $1 - e^{-C(p+\log T)}$, and

$$\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \frac{\zeta^{(t)}}{\bar{\zeta}}\sqrt{\frac{pr}{nT}} + \left(\frac{\zeta^{(t)}}{\bar{\zeta}} \vee 1\right)\sqrt{\frac{r}{n}} + \frac{\zeta^{(t)}}{\bar{\zeta}}\sqrt{\frac{r}{T}}\|\boldsymbol{D}_S^*\|_2 + \frac{\zeta^{(t)}}{\bar{\zeta}}\max_{t\in S}\zeta^{(t)}\sqrt{r\bar{\epsilon}}$$
$$+ \|\boldsymbol{\delta}^{(t)*}\|_2 + \sqrt{\frac{\log T}{n}}, \quad \forall t \in S, \quad (22)$$

w.p. at least $1 - e^{-C(r+\log T)}$.

Denote $\overline{\boldsymbol{B}}_S = \{\overline{\boldsymbol{A}}\bar{\boldsymbol{\theta}}^{(t)*}\}_{t\in S} \in \mathbb{R}^{p\times|S|}$ and $\widetilde{\boldsymbol{B}}_S = \{\widetilde{\boldsymbol{\beta}}^{(t)}\}_{t\in S}$.

Note that when $\frac{\min_{t\in S}\zeta^{(t)}}{\bar{\zeta}}\sqrt{\frac{r}{T}}\|\boldsymbol{D}_S^*\|_2 \leq c'\sqrt{\frac{p+\log T}{n}}$, and $\min_{t\in S}\|\boldsymbol{\delta}^{(t)*}\|_2 \leq c'\sqrt{\frac{p+\log T}{n}}$, we have

$$\frac{1}{|S|}\sigma_r(\overline{\boldsymbol{B}}_S)^2 = \frac{1}{|S|}\sigma_r\left(\sum_{t\in S}\boldsymbol{\beta}^{(t)*}\overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top(\boldsymbol{\beta}^{(t)*})^\top\right)$$
$$= \frac{1}{|S|}\sigma_r\left(\sum_{t\in S}\boldsymbol{\beta}^{(t)*}(\boldsymbol{\beta}^{(t)*})^\top\right) - \frac{1}{|S|}\sigma_r\left(\sum_{t\in S}\boldsymbol{\beta}^{(t)*}\overline{\boldsymbol{A}}^\perp(\overline{\boldsymbol{A}}^\perp)^\top(\boldsymbol{\beta}^{(t)*})^\top\right)$$
$$\geq c\frac{\bar{\zeta}^2}{r} - \frac{1}{T(1-\epsilon)}\|\boldsymbol{D}_S^*(\boldsymbol{D}_S^*)^\top\|_2$$
$$> \frac{c}{2}\frac{\bar{\zeta}^2}{r},$$

because $\sqrt{\frac{p+\log T}{n}}\frac{1}{\bar{\zeta}}\sqrt{\frac{r}{T}}\|\boldsymbol{D}_S^*\|_2 \lesssim \frac{\min_{t\in S}\zeta^{(t)}}{\bar{\zeta}}\sqrt{\frac{r}{T}}\|\boldsymbol{D}_S^*\|_2 \leq c'\sqrt{\frac{p+\log T}{n}}$.

Also, $\widetilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*} = (\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1}\frac{1}{n}(\boldsymbol{X}^{(t)})^\top\boldsymbol{\epsilon}^{(t)}$ is a sub-Gaussian vector with variance proxy $n^{-1}\sigma_{\max}((\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1})$ given $\boldsymbol{X}^{(t)}$, and $\{\widetilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\}_{t\in[T]}$ are independent of each other. By Lemma 5.39 in Vershynin (2010), conditioned on $\{\boldsymbol{X}^{(t)}\}_{t\in S}$, we have

$$\|\widetilde{\boldsymbol{B}}_S - \boldsymbol{B}_S^*\|_2 \lesssim \sqrt{\frac{p+T}{n}}\max_{t\in S}\sigma_{\max}((\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1}),$$

w.p. at least $1 - e^{-C(p+\log T)}$. Then combining

$$\|(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1}\|_2 \leq \|(\boldsymbol{\Sigma}^{(t)})^{-1}\|_2 + \|(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1}\|_2\|\widehat{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma}^{(t)}\|_2\|(\boldsymbol{\Sigma}^{(t)})^{-1}\|_2$$

with Lemma 18, we have $\sigma_{\max}((\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1}) = \|(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1}\|_2 \lesssim 1$. Therefore, w.p. at least $1 - e^{-C(p+\log T)}$,

$$\|\widetilde{\boldsymbol{B}}_S - \boldsymbol{B}_S^*\|_2 \lesssim \sqrt{\frac{p+T}{n}}.$$

Note that $\widehat{\boldsymbol{B}}_S - \overline{\boldsymbol{B}}_S = (\widehat{\boldsymbol{B}}_S - \widetilde{\boldsymbol{B}}_S) + (\widetilde{\boldsymbol{B}}_S - \boldsymbol{B}_S^*) + (\boldsymbol{B}_S^* - \overline{\boldsymbol{B}}_S)$. By triangle inequality,

$$\|\widehat{\boldsymbol{B}} - \begin{pmatrix}\overline{\boldsymbol{B}}_S & \boldsymbol{0}\end{pmatrix}_{p\times T}\|_2 \leq \|\widehat{\boldsymbol{B}}_{S^c}\|_2 + \|\widehat{\boldsymbol{B}}_S - \widetilde{\boldsymbol{B}}_S\|_2 + \|\widetilde{\boldsymbol{B}}_S - \boldsymbol{B}_S^*\|_2 + \|\boldsymbol{B}_S^* - \overline{\boldsymbol{B}}_S\|_2$$

$$\leq \|\widehat{\boldsymbol{B}}_{S^c}\|_{\mathrm{F}} + \|\widehat{\boldsymbol{B}}_S - \widetilde{\boldsymbol{B}}_S\|_{\mathrm{F}} + \|\widetilde{\boldsymbol{B}}_S - \boldsymbol{B}_S^*\|_2 + \|\boldsymbol{B}_S^* - \overline{\boldsymbol{B}}_S\|_2$$

$$\leq R\sqrt{\epsilon T} + \sqrt{\sum_{t\in S}\|\Pi_R(\widetilde{\boldsymbol{\beta}}^{(t)}) - \widetilde{\boldsymbol{\beta}}^{(t)}\|_2^2} + C\sqrt{\frac{p+T}{n}} + \|\boldsymbol{D}_S^*\|_2,$$

w.p. at least $1 - e^{-C'(p+\log T)}$. Since $R = \texttt{quantile}(\{\|\widetilde{\boldsymbol{\beta}}^{(t)}\|_2\}_{t=1}^T, 1 - \bar{\epsilon})$, $\bar{\epsilon} \geq \epsilon$, and

$$\max_{t\in S}\|\widetilde{\boldsymbol{\beta}}^{(t)}\|_2 \leq \max_{t\in S}\|\widetilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 + \max_{t\in S}\|\boldsymbol{\beta}^{(t)*}\|_2 \leq C\sqrt{\frac{p+\log T}{n}} + \max_{t\in S}\|\boldsymbol{\beta}^{(t)*}\|_2 \leq C\max_{t\in S}\|\boldsymbol{\beta}^{(t)*}\|_2,$$

w.p. at least $1 - e^{-C'(p+\log T)}$, we have

$$R \leq C\max_{t\in S}\|\boldsymbol{\beta}^{(t)*}\|_2,$$

w.p. at least $1 - e^{-C'(p+\log T)}$. And there are at most $\lceil\bar{\epsilon}T\rceil$ number of $t \in S$ satisfying $\Pi_R(\widetilde{\boldsymbol{\beta}}^{(t)}) \neq \widetilde{\boldsymbol{\beta}}^{(t)}$. Therefore, w.p. at least $1 - e^{-C'(p+\log T)}$,

$$\|\widehat{\boldsymbol{B}} - \begin{pmatrix}\overline{\boldsymbol{B}}_S & \boldsymbol{0}\end{pmatrix}_{p\times T}\|_2 \leq R\sqrt{\epsilon T} + C\max_{t\in S}\|\boldsymbol{\beta}^{(t)*}\|_2\sqrt{\epsilon T} + C\sqrt{\frac{p+T}{n}} + \|\boldsymbol{D}_S^*\|_2$$

$$\leq 2C\max_{t\in S}\|\boldsymbol{\beta}^{(t)*}\|_2\sqrt{\epsilon T} + C\sqrt{\frac{p+T}{n}} + \|\boldsymbol{D}_S^*\|_2.$$

Similarly, by Weyl's inequality, w.p. at least $1 - e^{-C'(p+\log T)}$,

$$\sigma_r(\widehat{\boldsymbol{B}}) - \sigma_{r+1}(\begin{pmatrix}\overline{\boldsymbol{B}}_S & \boldsymbol{0}\end{pmatrix}) = \sigma_r(\widehat{\boldsymbol{B}})$$

$$\geq \sigma_r(\overline{\boldsymbol{B}}_S) - \|\widehat{\boldsymbol{B}}_{S^c}\|_2 - \|\widehat{\boldsymbol{B}}_S - \widetilde{\boldsymbol{B}}_S\|_2 - \|\widetilde{\boldsymbol{B}}_S - \boldsymbol{B}_S^*\|_2 - \|\boldsymbol{B}_S^* - \overline{\boldsymbol{B}}_S\|_2$$

$$\geq c\bar{\zeta}\sqrt{\frac{T}{r}} - 2C\max_{t\in S}\|\boldsymbol{\beta}^{(t)*}\|_2\sqrt{\epsilon T} - C\sqrt{\frac{p+T}{n}} - \|\boldsymbol{D}_S^*\|_2$$

$$\geq \frac{c}{2}\bar{\zeta}\sqrt{\frac{T}{r}},$$

since $\max_{t\in S}\|\boldsymbol{\beta}^{(t)*}\|_2\sqrt{\bar{\epsilon}} \leq \frac{c}{6C}\frac{\bar{\zeta}}{\sqrt{r}}$, $\sqrt{\frac{p+T}{n}} \leq \frac{c}{6C}\sqrt{\frac{p+\log T}{n}} \cdot \bar{\zeta}\sqrt{\frac{T}{r}} \leq \bar{\zeta}\frac{c}{6C}\sqrt{\frac{T}{r}}$, and $\|\boldsymbol{D}_S^*\|_2 \leq \frac{c}{6}\bar{\zeta}\sqrt{\frac{T}{r}}$, because $\bar{\epsilon} \leq \frac{c''}{r} \cdot \left(\frac{\bar{\zeta}}{\max_{t\in S}\|\boldsymbol{\beta}^{(t)*}\|_2}\right)^2$, $nT \geq C''pr$, $p \geq r$, $n \geq C''r$, and $\frac{1}{\bar{\zeta}}\sqrt{\frac{r}{T}}\|\boldsymbol{D}_S^*\|_2 \leq \frac{1}{\bar{\zeta}}\sqrt{\frac{r}{T}}\|\boldsymbol{D}_S^*\|_2 \cdot \frac{\min_{t\in S}\zeta^{(t)}}{\sqrt{(p+\log T)/n}} \leq c''$, where $c'' > 0$ is a small constant and $C'' > 0$ is a large constant.

Then by the SVD and Wedin's $\sin\Theta$-Theorem:

$$\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \lesssim \frac{\|\widehat{\boldsymbol{B}} - (\overline{\boldsymbol{B}}_S \quad \boldsymbol{0})_{p\times T}\|_2}{\sigma_r(\widehat{\boldsymbol{B}}) - \sigma_{r+1}((\overline{\boldsymbol{B}}_S \quad \boldsymbol{0}))}$$

$$\lesssim \frac{\max_{t\in S}\|\boldsymbol{\beta}^{(t)*}\|_2\sqrt{\bar{\epsilon}T} + \sqrt{\frac{p+T}{n}} + \|\boldsymbol{D}_S^*\|_2}{\bar{\zeta}\sqrt{\frac{T}{r}}}$$

$$\lesssim \bar{\zeta}^{-1}\sqrt{\frac{pr}{nT}} + \bar{\zeta}^{-1}\sqrt{\frac{r}{n}} + \bar{\zeta}^{-1}\sqrt{\frac{r}{T}}\|\boldsymbol{D}_S^*\|_2 + \frac{\max_{t\in S}\|\boldsymbol{\beta}^{(t)*}\|_2}{\bar{\zeta}}\sqrt{r\bar{\epsilon}},$$

w.p. at least $1 - e^{-C(p+\log T)}$. Hence (21) holds.

Next, let us prove (22). By noticing that

$$\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} = ((\widehat{\overline{\boldsymbol{A}}})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\widehat{\overline{\boldsymbol{A}}})^{-1}(\widehat{\overline{\boldsymbol{A}}})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{\beta}^{(t)*} - ((\widehat{\overline{\boldsymbol{A}}})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\widehat{\overline{\boldsymbol{A}}})^{-1}(\widehat{\overline{\boldsymbol{A}}})^\top\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*}),$$

we have

$$\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \boldsymbol{\beta}^{(t)*} = \left[\widehat{\overline{\boldsymbol{A}}}((\widehat{\overline{\boldsymbol{A}}})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\widehat{\overline{\boldsymbol{A}}})^{-1}(\widehat{\overline{\boldsymbol{A}}})^\top\widehat{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{I}\right]\boldsymbol{\beta}^{(t)*} - ((\widehat{\overline{\boldsymbol{A}}})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\widehat{\overline{\boldsymbol{A}}})^{-1}(\widehat{\overline{\boldsymbol{A}}})^\top\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})$$

$$= (\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1/2}\left[(\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\widehat{\overline{\boldsymbol{A}}}((\widehat{\overline{\boldsymbol{A}}})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\widehat{\overline{\boldsymbol{A}}})^{-1}(\widehat{\overline{\boldsymbol{A}}})^\top(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1/2} - \boldsymbol{I}\right](\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\boldsymbol{\beta}^{(t)*}$$
$$- ((\widehat{\overline{\boldsymbol{A}}})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\widehat{\overline{\boldsymbol{A}}})^{-1}(\widehat{\overline{\boldsymbol{A}}})^\top\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*}).$$

Denote the projection matrix of a matrix $\boldsymbol{A} \in \mathbb{R}^{p_1\times p_2}$ (projection onto the column space of $\boldsymbol{A}$) with $\text{rank}(\boldsymbol{A}) = p_2 \leq p_1$ as

$$\boldsymbol{P_A} = \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top.$$

Denote $\widehat{\overline{\boldsymbol{B}}} = (\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\widehat{\overline{\boldsymbol{A}}} \in \mathbb{R}^{p\times r}$ and $(\widehat{\overline{\boldsymbol{B}}})^\perp = (\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1/2}(\widehat{\overline{\boldsymbol{A}}})^\perp \in \mathbb{R}^{p\times(p-r)}$. Notice that $(\widehat{\overline{\boldsymbol{B}}})^\top(\widehat{\overline{\boldsymbol{B}}})^\perp = \boldsymbol{0}_{r\times(p-r)}$. By definition $\boldsymbol{P}_{\widehat{\overline{\boldsymbol{B}}}} + \boldsymbol{P}_{(\widehat{\overline{\boldsymbol{B}}})^\perp} = \boldsymbol{I}_p$. Hence

$$\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \boldsymbol{\beta}^{(t)*} = (\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1/2}(\boldsymbol{P}_{\widehat{\overline{\boldsymbol{B}}}} - \boldsymbol{I})(\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}(\overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)*} + \boldsymbol{\delta}^{(t)*}) - ((\widehat{\overline{\boldsymbol{A}}})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\widehat{\overline{\boldsymbol{A}}})^{-1}(\widehat{\overline{\boldsymbol{A}}})^\top\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})$$

$$= -(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1/2}\boldsymbol{P}_{(\widehat{\overline{\boldsymbol{B}}})^\perp}(\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)*} - (\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1/2}\boldsymbol{P}_{(\widehat{\overline{\boldsymbol{B}}})^\perp}(\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\boldsymbol{\delta}^{(t)*}$$
$$- ((\widehat{\overline{\boldsymbol{A}}})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\widehat{\overline{\boldsymbol{A}}})^{-1}(\widehat{\overline{\boldsymbol{A}}})^\top\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*}).$$

Note that since $\|\widehat{\boldsymbol{\Sigma}}^{(t)}\|_2$ and $\|(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1}\|_2$ are bounded from above by constants w.p. at least $1 - e^{-C'(p+\log T)}$, we have

$$\|(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1/2}\boldsymbol{P}_{(\widehat{\overline{\boldsymbol{B}}})^\perp}(\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)*}\|_2 \lesssim \|((\widehat{\overline{\boldsymbol{A}}})^\perp)^\top\overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)*}\|_2 + \|\boldsymbol{\delta}^{(t)*}\|_2 \lesssim \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2\zeta^{(t)} + \|\boldsymbol{\delta}^{(t)*}\|_2,$$

$$\|(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1/2}\boldsymbol{P}_{(\widehat{\overline{\boldsymbol{B}}})^\perp}(\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\boldsymbol{\delta}^{(t)*}\|_2 \lesssim \|\boldsymbol{\delta}^{(t)*}\|_2,$$

w.p. at least $1 - e^{-C'(p + \log T)}$. Therefore w.p. at least $1 - e^{-C'(p + \log T)}$,

$$\|\widehat{\overline{A}}\theta^{(t)}_{\widehat{\overline{A}}} - \beta^{(t)*}\|_2 \lesssim \|\widehat{\overline{A}}(\widehat{\overline{A}})^\top - \overline{A}(\overline{A})^\top\|_2 \zeta^{(t)} + \|\delta^{(t)*}\|_2 + \|((\widehat{\overline{A}})^\top \widehat{\Sigma}^{(t)} \widehat{\overline{A}})^{-1}(\widehat{\overline{A}})^\top \nabla f^{(t)}(\beta^{(t)*})\|_2.$$

Let us handle the last term. Denote $\widetilde{R} = \arg\min_{R \in \mathcal{O}^{r \times r}} \|\widehat{\overline{A}} - \overline{A}R\|_2$. Therefore

$$\begin{aligned}
&\|((\widehat{\overline{A}})^\top \widehat{\Sigma}^{(t)} \widehat{\overline{A}})^{-1}(\widehat{\overline{A}})^\top \nabla f^{(t)}(\beta^{(t)*})\|_2 \\
&\le \|((\widehat{\overline{A}})^\top \widehat{\Sigma}^{(t)} \widehat{\overline{A}})^{-1}(\widehat{\overline{A}} - \overline{A}\widetilde{R})^\top \nabla f^{(t)}(\beta^{(t)*})\|_2 + \|((\widehat{\overline{A}})^\top \widehat{\Sigma}^{(t)} \widehat{\overline{A}})^{-1}(\overline{A}\widetilde{R})^\top \nabla f^{(t)}(\beta^{(t)*})\|_2 \\
&\le \|\widehat{\overline{A}} - \overline{A}\widetilde{R}\|_2 \|\nabla f^{(t)}(\beta^{(t)*})\|_2 + \|(\overline{A}\widetilde{R})^\top \nabla f^{(t)}(\beta^{(t)*})\|_2 \\
&\lesssim \|\widehat{\overline{A}}(\widehat{\overline{A}})^\top - \overline{A}(\overline{A})^\top\|_2 \sqrt{\frac{p + \log T}{n}} + \sqrt{\frac{r + \log T}{n}},
\end{aligned}$$

w.p. at least $1 - e^{-C'(r + \log T)}$. Hence, w.p. at least $1 - e^{-C'(r + \log T)}$,

$$\begin{aligned}
\|\widehat{\overline{A}}\theta^{(t)}_{\widehat{\overline{A}}} - \beta^{(t)*}\|_2 &\lesssim \|\widehat{\overline{A}}(\widehat{\overline{A}})^\top - \overline{A}(\overline{A})^\top\|_2 \zeta^{(t)} + \|\delta^{(t)*}\|_2 + \sqrt{\frac{r + \log T}{n}} \\
&\lesssim \left(\bar{\zeta}^{-1}\sqrt{\frac{pr}{nT}} + \bar{\zeta}^{-1}\sqrt{\frac{r}{n}} + \bar{\zeta}^{-1}\sqrt{\frac{r}{T}}\|D_S^*\|_2 + \frac{\max_{t \in S}\|\beta^{(t)*}\|_2}{\bar{\zeta}}\sqrt{r\bar{\epsilon}}\right)\zeta^{(t)} \\
&\quad + \|\delta^{(t)*}\|_2 + \sqrt{\frac{r + \log T}{n}} \\
&\lesssim \frac{\zeta^{(t)}}{\bar{\zeta}}\sqrt{\frac{pr}{nT}} + \left(\frac{\zeta^{(t)}}{\bar{\zeta}} \vee 1\right)\sqrt{\frac{r}{n}} + \frac{\zeta^{(t)}}{\bar{\zeta}}\sqrt{\frac{r}{T}}\|D_S^*\|_2 + \frac{\zeta^{(t)}}{\bar{\zeta}}\max_{t \in S}\zeta^{(t)}\sqrt{r\bar{\epsilon}} \\
&\quad + \|\delta^{(t)*}\|_2 + \sqrt{\frac{\log T}{n}} \\
&:= \eta^{(t)}.
\end{aligned}$$

Therefore, (22) holds. Now let us use (21) and (22) to complete the proof of the theorem.
(i) For $t \in S$ with $\eta^{(t)} \le c'\sqrt{\frac{p + \log T}{n}}$ with a small $c' > 0$: we must have

$$\|\nabla f^{(t)}(\beta^{(t)*})\|_2 + \|\widehat{\overline{A}}\theta^{(t)}_{\widehat{\overline{A}}} - \beta^{(t)*}\|_2 \lesssim \sqrt{\frac{p + \log T}{n}} + \eta^{(t)} \lesssim \frac{\gamma}{\sqrt{n}},$$

hence by Lemma 38.(i), $\widehat{\beta}^{(t)} = \widehat{\overline{A}}\theta^{(t)}_{\widehat{\overline{A}}}$ and $\|\widehat{\beta}^{(t)} - \beta^{(t)*}\|_2 \lesssim \eta^{(t)}$ w.p. at least $1 - e^{-C'(r + \log T)}$. Note that we used (22) to obtain the bound $\|\widehat{\overline{A}}\theta^{(t)}_{\widehat{\overline{A}}} - \beta^{(t)*}\|_2 \lesssim \eta^{(t)}$, w.p. at least $1 - e^{-C'(r + \log T)}$. We can use it because $\eta^{(t)} \le c'\sqrt{\frac{p + \log T}{n}}$ implies that $\frac{\min_{t \in S}\zeta^{(t)}}{\bar{\zeta}}\sqrt{\frac{r}{T}}\|D_S^*\|_2 \le c'\sqrt{\frac{p + \log T}{n}}$, hence the condition of (22) applies and (22) holds. Next, let us explain why (22) implies the bound in our theorem.

Note that

$$\|\delta^{(t)*}\|_2 \le \|(\overline{A})^\perp A^{(t)*}\theta^{(t)*}\|_2 \lesssim \|A^{(t)*}(A^{(t)*})^\top - \overline{A}(\overline{A})^\top\|_2 \zeta^{(t)} \lesssim h\zeta^{(t)},$$

$$\frac{1}{\sqrt{T}}\|\boldsymbol{D}_S^*\|_2 \leq \frac{1}{\sqrt{T}}\|\boldsymbol{D}_S^*\|_{\mathrm{F}} \cdot \frac{\sigma_{\max}(\boldsymbol{D}_S^*)}{\sqrt{r}\sigma_{\min}(\boldsymbol{D}_S^*)} \lesssim \bar{\zeta} \cdot \frac{\sigma_{\max}(\boldsymbol{D}_S^*)}{\sqrt{r}\sigma_{\min}(\boldsymbol{D}_S^*)} \cdot h, \tag{23}$$

where $\boldsymbol{D}_S^* = (\overline{\boldsymbol{A}})^{\perp}\boldsymbol{B}_S^*$ and $\|\boldsymbol{D}_S^*\|_{\mathrm{F}} \leq h\sqrt{\sum_{t\in S}\|\boldsymbol{\delta}^{(t)*}\|_2^2} \leq h\sqrt{\sum_{t\in S}(\zeta^{(t)})^2}$. Regarding the last inequality (23), there is an alternative way to bound the LHS:

$$\frac{1}{\sqrt{T}}\|\boldsymbol{D}_S^*\|_2 \leq \frac{1}{\sqrt{T}}\|\boldsymbol{D}_S^*\|_{\mathrm{F}} \leq \bar{\zeta}h.$$

Therefore,

$$\frac{1}{\sqrt{T}}\|\boldsymbol{D}_S^*\|_2 \lesssim \bar{\zeta}h \cdot \left[\frac{\sigma_{\max}(\boldsymbol{D}_S^*)}{\sqrt{r}\sigma_{\min}(\boldsymbol{D}_S^*)} \wedge 1\right].$$

Combing all these facts, we obtain the high-probability bound $\frac{\zeta^{(t)}}{\bar{\zeta}}\sqrt{\frac{pr}{nT}} + \left(\frac{\zeta^{(t)}}{\bar{\zeta}} \vee 1\right)\sqrt{\frac{r}{n}} +$

$\sqrt{\frac{\log T}{n}} + \zeta^{(t)}h \cdot \left[\frac{\sigma_{\max}((\overline{\boldsymbol{A}}^{\perp})^{\top}\boldsymbol{B}_S^*)}{\sigma_{\min}((\overline{\boldsymbol{A}}^{\perp})^{\top}\boldsymbol{B}_S^*)} \wedge \sqrt{r}\right] + \frac{\zeta^{(t)}}{\bar{\zeta}}\max_{t\in S}\zeta^{(t)} \cdot \sqrt{r\bar{\epsilon}}$ for $\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$.

(ii) For any $t \in S$, by Lemma 38, $\max_{t\in S}\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \leq C\frac{\gamma}{\sqrt{n}} + \max_{t\in S}\|\widetilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ w.p. at least $1 - e^{-C(p+\log T)}$, where $\widetilde{\boldsymbol{\beta}}^{(t)} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} f^{(t)}(\boldsymbol{\beta})$. Since $\gamma \asymp \sqrt{p+\log T}$, this implies that

$$\max_{t\in S}\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \sqrt{\frac{p+\log T}{n}},$$

w.p. at least $1 - e^{-C'(p+\log T)}$.

Combining (i) and (ii), we get the desired bound:

$$\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \left\{\frac{\zeta^{(t)}}{\bar{\zeta}}\sqrt{\frac{pr}{nT}} + \left(\frac{\zeta^{(t)}}{\bar{\zeta}} \vee 1\right)\sqrt{\frac{r}{n}} + \sqrt{\frac{\log T}{n}} + \zeta^{(t)}h \cdot \left[\frac{\sigma_{\max}((\overline{\boldsymbol{A}}^{\perp})^{\top}\boldsymbol{B}_S^*)}{\sigma_{\min}((\overline{\boldsymbol{A}}^{\perp})^{\top}\boldsymbol{B}_S^*)} \wedge \sqrt{r}\right]\right.$$

$$\left. + \frac{\zeta^{(t)}}{\bar{\zeta}}\max_{t\in S}\zeta^{(t)}\sqrt{r\bar{\epsilon}}\right\} \wedge \sqrt{\frac{p+\log T}{n}},$$

for all $t \in S$, w.p. at least $1 - e^{-C'(r+\log T)}$.

When the tasks in S^c are generated by the linear model, we can get

$$\max_{t\in S^c}\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \sqrt{\frac{p+\log T}{n}},$$

w.p. at least $1 - e^{-C'(p+\log T)}$, similar to the argument in the proof of Theorem 2.

D.9 Proof of Theorem 8

First, when tasks in S^c come from linear models, the term $\sqrt{p/n}$ in the lower bound of $\max_{t\in[T]}\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ is the standard result in linear regression. The other part $\sqrt{\log T/n}$ can be shown using the same arguments as in Tian et al. (2022) and Duan and Wang (2023). See the proof of Theorem 4.3 in Duan and Wang (2023) and the proof of Theorem 2 in Tian

et al. (2022). We omit the proof of lower bound for $\max_{t \in [T]} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ when tasks in S^c come from linear models, and only show the lower bound of $\max_{t \in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$ when tasks in S^c can be arbitrarily distributed.

Throughout this subsection, we assume the following generative model for tasks in S:

$$y|\boldsymbol{x} \sim \mathbb{P}_{y|\boldsymbol{x},\boldsymbol{\beta}} = N(\boldsymbol{x}^\top \boldsymbol{\beta}, 1), \quad \boldsymbol{x} \sim \mathbb{P}_{\boldsymbol{x}},$$

where $\mathbb{P}_{\boldsymbol{x}}$ is sub-Gaussian with $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top)$. Suppose there exist constants c, C such that $0 < c \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq C < \infty$. Then any joint distribution $\mathbb{P}$ of $(\boldsymbol{x}, y)$ can be written as $\mathbb{P}_{\boldsymbol{x},y} = \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}} \cdot \mathbb{P}_{\boldsymbol{x}}$.

Recall the parameter space for the coefficient vectors $\{\boldsymbol{\beta}^{(t)}\}_{t \in S}$ as

$$\mathscr{B}(S, h) = \left\{ \{\boldsymbol{\beta}^{(t)}\}_{t \in S} : \boldsymbol{\beta}^{(t)} = \boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} \text{ for all } t \in S, \{\boldsymbol{A}^{(t)}\}_{t \in S} \subseteq \mathcal{O}^{p \times r}, \|\boldsymbol{\theta}^{(t)}\|_2 \leq \zeta^{(t)}, \right.$$

$$\left. \min_{\overline{\boldsymbol{A}} \in \mathcal{O}^{p \times r}} \max_{t \in S} \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \overline{\boldsymbol{A}}\overline{\boldsymbol{A}}^\top\|_2 \leq h, \sigma_r\left(|S|^{-1/2}\boldsymbol{B}_S\right) \geq \frac{c}{\sqrt{r}} \sqrt{\frac{1}{|S|} \sum_{t \in S} \|\boldsymbol{\theta}^{(t)}\|_2^2} \right\}$$

where c can be any fixed positive constants such that $\mathscr{B}(S, h) \neq \emptyset$.

Given $S \subseteq [T]$, define

$$\mathcal{A}(S, h) = \left\{ \{\boldsymbol{A}^{(t)}\}_{t \in S} \subseteq \mathcal{O}^{p \times r} : \min_{\overline{\boldsymbol{A}} \in \mathcal{O}^{p \times r}} \max_{t \in S} \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \overline{\boldsymbol{A}}\overline{\boldsymbol{A}}^\top\|_2 \leq h \right\},$$

$$\Theta(S) = \left\{ \{\boldsymbol{\theta}^{(t)}\}_{t \in S} \subseteq \mathbb{R}^r : \|\boldsymbol{\theta}^{(t)}\|_2 \leq \zeta^{(t)}, \frac{1}{|S|} \sum_{t \in S} \boldsymbol{\theta}^{(t)}(\boldsymbol{\theta}^{(t)})^\top \succeq \frac{c}{r} \frac{1}{|S|} \sum_{t \in S} \|\boldsymbol{\theta}^{(t)}\|_2^2 \cdot \boldsymbol{I}_r \right\},$$

$$\bar{\zeta}_S = \sqrt{\frac{1}{|S|} \sum_{t \in S} (\zeta^{(t)})^2}.$$

The proof proceeds as follows. We will first show

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T} \sup_{S \subseteq \mathcal{S}} \sup_{\substack{\{\boldsymbol{\beta}^{(t)}\}_{t \in S} \in \mathscr{B}(S,h) \\ \mathbb{Q}_{S^c}}} \mathbb{P}\left(\bigcup_{t \in S} \left\{ \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \gtrsim \left[\frac{\zeta^{(t)}}{\bar{\zeta}_S} \sqrt{\frac{pr}{nT}} + \left(\frac{\zeta^{(t)}}{\bar{\zeta}_S} \vee 1 \right) \sqrt{\frac{r}{n}} + \zeta^{(t)}h \right. \right.$$

$$\left. \left. + \frac{\zeta^{(t)}}{\bar{\zeta}_S} \frac{\epsilon r}{\sqrt{n}} \right] \wedge \sqrt{\frac{p}{n}} \right\} \right) \geq \frac{1}{10}, \tag{24}$$

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T} \sup_{S \subseteq \mathcal{S}} \sup_{\substack{\{\boldsymbol{\beta}^{(t)}\}_{t \in S} \in \mathscr{B}([T],0) \\ \mathbb{Q}_{S^c}}} \mathbb{P}\left(\max_{t \in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \gtrsim \sqrt{\frac{\log T}{n}} \right) \geq \frac{1}{10}. \tag{25}$$

Then we will discuss how to get the final lower bound by the same arguments.
Part 1: First, let us prove (24) first.

- If $h \geq \frac{1}{\zeta_{[T]}} \sqrt{\frac{pr}{nT}} + \frac{1}{\zeta_{[T]}} \sqrt{\frac{r}{n}} + \frac{1}{\zeta_{[T]}} \frac{\epsilon r}{\sqrt{n}}$: then $\zeta^{(t)} h \geq \frac{\zeta^{(t)}}{\zeta_{[T]}} \sqrt{\frac{pr}{nT}} + \frac{\zeta^{(t)}}{\zeta_{[T]}} \sqrt{\frac{r}{n}} + \frac{\zeta^{(t)}}{\zeta_{[T]}} \frac{\epsilon r}{\sqrt{n}}$ for all $t \in [T]$.

  ○ If $\sqrt{\frac{r}{n}} \leq (\zeta^{(t_0)} h) \wedge \sqrt{\frac{p}{n}}$ for some $t_0 \in [T]$: then we take $S = [T]$ and prove the lower bound

  $$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T} \sup_{\{\boldsymbol{\beta}^{(t)}\}_{t\in[T]}\in\mathscr{B}([T],h)} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(t_0)} - \boldsymbol{\beta}^{(t_0)}\|_2 \gtrsim (\zeta^{(t_0)} h) \wedge \sqrt{\frac{p}{n}} \right) \geq \frac{1}{10}. \quad (26)$$

  WLOG, assume $t_0 = 1$. For all $t \in [T]$, fix $\boldsymbol{A}^{(t)} =$ an arbitrary $\overline{\boldsymbol{A}} \in \mathcal{O}^{p\times r}$, and fix $\{\boldsymbol{\theta}^{(t)}\}_{t=2}^T$ s.t. $\|\boldsymbol{\theta}^{(t)}\|_2 \leq \zeta^{(t)}$. Let $\boldsymbol{\theta}^{(1)} = \frac{1}{\sqrt{r}} \zeta^{(1)} \mathbf{1}_r$. Set $\delta = \frac{1}{12}\left[\left(\frac{1}{\zeta^{(1)}} \sqrt{\frac{p}{n}} \right) \wedge h \right]$. Consider a $\delta/2$-packing of the ball $\mathcal{B}_\delta(\mathbf{0}, \mathbb{R}^p, \|\cdot\|_2)$ (denoted as $\mathcal{M}$). By Example 5.8 in Wainwright (2019), $|\mathcal{M}| \geq 5^p$. Consider

  $$\boldsymbol{A} = \overline{\boldsymbol{A}} + \frac{c'}{\sqrt{r}} \cdot \boldsymbol{u}\mathbf{1}_r^\top, \quad \widetilde{\boldsymbol{A}} = \overline{\boldsymbol{A}} + \frac{c'}{\sqrt{r}} \cdot \widetilde{\boldsymbol{u}}\mathbf{1}_r^\top,$$

  where $\boldsymbol{u} \neq \widetilde{\boldsymbol{u}} \in \mathcal{M}$ and c' is a small constant, then

  $$\|(\boldsymbol{A} - \widetilde{\boldsymbol{A}})\boldsymbol{\theta}^{(1)}\|_2 = \frac{1}{r}\zeta^{(t)} \|(\boldsymbol{u} - \widetilde{\boldsymbol{u}})\mathbf{1}_r^\top \mathbf{1}_r\|_2 = \|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\|_2 \in \left[\frac{c'}{2}\zeta^{(1)}\delta, 2c'\zeta^{(1)}\delta \right] \quad (27)$$

  $$\|(\boldsymbol{A} - \widetilde{\boldsymbol{A}})\boldsymbol{\theta}^{(t)}\|_2 = \|(\boldsymbol{A} - \widetilde{\boldsymbol{A}})\boldsymbol{e}_{j_k}\|_2 = \frac{1}{\sqrt{r}}\zeta^{(1)} \|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\|_2 \leq \frac{\delta}{\sqrt{r}}\zeta^{(1)},$$

  where $\boldsymbol{\theta}^{(t)} = \boldsymbol{e}_{j_k}$. Hence $\mathscr{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\theta}^{(1)}, \boldsymbol{A} = \overline{\boldsymbol{A}} + \frac{c'}{\sqrt{r}} \cdot \boldsymbol{u}\mathbf{1}_r^\top, \boldsymbol{u} \in \mathcal{M}\}$ is a $c'\delta/2$-packing in $\mathbb{R}^p$ with $|\mathscr{B}| = |\mathcal{M}| \geq 5^p$. On the other hand, for any $\boldsymbol{A} \in \mathbb{R}^{p\times r}$ with $1 + 2c'\delta \geq \sigma_1(\boldsymbol{A}) \geq \sigma_2(\boldsymbol{A}) \geq \cdots \geq \sigma_r(\boldsymbol{A}) \geq 1 - 2c'\delta > 0$ and $\|\boldsymbol{A} - \overline{\boldsymbol{A}}\|_2 \leq \delta$, consider its SVD where $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{V}$ where $\boldsymbol{Q} \in \mathcal{O}^{p\times r}$ and $\boldsymbol{V} \in \mathcal{O}^{r\times r}$. Denote $\boldsymbol{R} = \boldsymbol{\Lambda}\boldsymbol{V}$. Then $0 < 1 - 2c'\delta \leq \sigma_{\min}(\boldsymbol{R}) \leq \sigma_{\max}(\boldsymbol{R}) \leq 1 + 2c'\delta < \infty$, and $\boldsymbol{A}\boldsymbol{\theta}^{(1)} = \boldsymbol{Q}\boldsymbol{R}\boldsymbol{\theta}^{(1)} = \boldsymbol{Q}\widetilde{\boldsymbol{\theta}}^{(1)}$. Note that $\boldsymbol{Q} \in \mathcal{O}^{p\times r}$ and $\widetilde{\boldsymbol{\theta}}^{(1)} = \boldsymbol{R}\boldsymbol{\theta}^{(1)}$ with $\|\widetilde{\boldsymbol{\theta}}^{(1)}\|_2 \leq \|\boldsymbol{R}\|_2 \|\boldsymbol{\theta}^{(1)}\|_2 \leq C[1 + 2c']$. And $\|\boldsymbol{Q}\boldsymbol{Q}^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \leq \|(\boldsymbol{Q}\boldsymbol{R})(\boldsymbol{Q}\boldsymbol{R})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 + \|\boldsymbol{Q}(\boldsymbol{I}_r - \boldsymbol{R}\boldsymbol{R}^\top)\boldsymbol{Q}^\top\|_2 \leq 4\|\boldsymbol{A} - \overline{\boldsymbol{A}}\|_2 + 4\|\boldsymbol{I}_r - \boldsymbol{R}\|_2 \leq 4\delta + 8c'\delta \leq 12\delta \leq h$. Therefore,

  $$\inf_{\widehat{\boldsymbol{\beta}}^{(1)}} \sup_{\{\boldsymbol{\beta}^{(t)}\}_{t\in S}\in\mathscr{B}([T],h)} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)}\|_2 \gtrsim (\zeta^{(1)} h) \wedge \sqrt{\frac{p}{n}} \right)$$

  $$\geq \inf_{\widehat{\boldsymbol{\beta}}^{(1)}} \sup_{\substack{0 < c \leq \sigma_{\max}(\boldsymbol{A}) \leq \sigma_{\max}(\boldsymbol{A}) \leq C \\ \|\boldsymbol{\theta}^{(1)}\|_2 \leq C'}} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{A}\boldsymbol{\theta}^{(1)}\|_2 \gtrsim (\zeta^{(1)} h) \wedge \sqrt{\frac{p}{n}} \right)$$

  $$:= (*),$$

  where all $\boldsymbol{A}^{(t)} =$ an arbitrary $\overline{\boldsymbol{A}} \in \mathcal{O}^{p\times r}$, and $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^T \subseteq \{\boldsymbol{e}_j\}_{j=1}^r$ are fixed. For any $\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\theta}^{(1)}$ and $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{A}}\boldsymbol{\theta}^{(1)} \in \mathscr{B}$ with $\boldsymbol{A} \neq \widetilde{\boldsymbol{A}}$, by Lemma 24 and equation (27),

  $$\mathrm{KL}\left(\prod_{t=2}^T \mathbb{P}_{y|\boldsymbol{x};\overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)}}^{(t)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(t)\otimes n} \cdot \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\boldsymbol{\theta}^{(1)}}^{(1)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(1)\otimes n} \middle\| \prod_{t=2}^T \mathbb{P}_{y|\boldsymbol{x};\overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)}}^{(t)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(t)\otimes n} \cdot \mathbb{P}_{y|\boldsymbol{x};\widetilde{\boldsymbol{A}}\boldsymbol{\theta}^{(1)}}^{(1)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(1)\otimes n} \right)$$

$$= \text{KL}\left(\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\boldsymbol{\theta}^{(1)}}^{(1)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(1)\otimes n} \middle\| \mathbb{P}_{y|\boldsymbol{x};\widetilde{\boldsymbol{A}}\boldsymbol{\theta}^{(1)}}^{(1)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(1)\otimes n}\right)$$

$$\lesssim n\|(\boldsymbol{A} - \widetilde{\boldsymbol{A}})\boldsymbol{\theta}^{(1)}\|_2^2$$

$$\lesssim n(\zeta^{(1)})^2\delta^2$$

$$\leq c\log|\mathscr{B}|.$$

Then by Fano's lemma (Lemma 23), $(*) \geq 1 - \frac{\log 2}{\log|\mathscr{B}|} - c \geq 1/10$.

○ If $\sqrt{\frac{r}{n}} > (\zeta^{(t)}h) \wedge \sqrt{\frac{p}{n}}$ for all $t \in [T]$: then we take $S = [T]$ and prove the lower bound

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T} \sup_{\{\boldsymbol{\beta}^{(t)}\}_{t\in[T]}\in\mathscr{B}([T],0)} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(t_0)} - \boldsymbol{\beta}^{(t_0)}\|_2 \gtrsim \sqrt{\frac{r}{n}}\right) \geq \frac{1}{10}, \qquad (28)$$

where $t_0 \in \arg\min_{t\in[T]} \zeta^{(t)}$. WLOG, assume $t_0 = 1$. Fix all $\boldsymbol{A}^{(t)} = $ some $\boldsymbol{A} \in \mathcal{O}^{p\times r}$. Fixing $\boldsymbol{\Theta} = \{\boldsymbol{\theta}^{(t)}\}_{t=2}^T$ such that $\frac{1}{T}\sum_{t=2}^T \boldsymbol{\theta}^{(t)}(\boldsymbol{\theta}^{(t)})^\top \succeq \frac{c}{r}\boldsymbol{I}_r$ (Hence for any $\boldsymbol{\theta}^{(1)}$ with $\|\boldsymbol{\theta}^{(1)}\|_2 \leq C \leq \zeta^{(1)}$, we must have $\frac{1}{T}\sum_{t=1}^T \boldsymbol{\theta}^{(t)}(\boldsymbol{\theta}^{(t)})^\top \succeq \frac{c}{r}\boldsymbol{I}_r$). We want to show

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T} \sup_{\|\boldsymbol{\theta}^{(1)}\|_2\leq C} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{A}\boldsymbol{\theta}^{(1)}\|_2 \geq c\sqrt{\frac{r}{n}}\right) \geq \frac{1}{10}.$$

Denote $\delta = \sqrt{\frac{r}{n}}$. Consider a $c\delta$-packing of $\mathcal{B}_\delta(\boldsymbol{0}, \mathbb{R}^r, \|\cdot\|_2)$ (denoted as $\mathscr{T}$). By Example 5.8 in Wainwright (2019), we know that $\log|\mathscr{T}| \gtrsim r$. Note that this also defines a $c\delta$-packing of the space of $\boldsymbol{\beta}^{(1)}$ as $\mathscr{M} = \{\boldsymbol{\beta}^{(1)} \in \mathbb{R}^p : \boldsymbol{\beta}^{(1)} = \boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{\theta} \in \mathscr{T}\}$, because for any $\boldsymbol{\theta}^{(1)} \neq \widetilde{\boldsymbol{\theta}}^{(1)} \in \mathscr{T}$ we have $\|\boldsymbol{A}(\boldsymbol{\theta}^{(1)} - \widetilde{\boldsymbol{\theta}}^{(1)})\|_2 = \|\boldsymbol{\theta}^{(1)} - \widetilde{\boldsymbol{\theta}}^{(1)}\|_2 \geq c\delta$. Denote the distribution of $\{y_i^{(t)}\}_{i=1}^n$ given $\{\boldsymbol{x}_i^{(t)}\}_{i=1}^n$ and $\{\boldsymbol{\beta}^{(t)}\}_{t=1}^T$ as $\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}^{(t)}}^{(t)\otimes n}$ and the distribution of $\{\boldsymbol{x}_i^{(t)}\}_{i=1}^n$ as $\mathbb{P}_{\boldsymbol{x}}^{(t)\otimes n}$. For any $\boldsymbol{\beta}^{(1)} \neq \widetilde{\boldsymbol{\beta}}^{(1)} \in \mathscr{M}$:

$$\text{KL}\left(\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}^{(1)}}^{(1)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(1)\otimes n} \middle\| \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\widetilde{\boldsymbol{\beta}}^{(1)}}^{(1)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(1)\otimes n}\right) \lesssim n\|\boldsymbol{\beta}^{(1)} - \widetilde{\boldsymbol{\beta}}^{(1)}\|_2^2$$

$$\lesssim n\delta^2$$

$$\lesssim r$$

$$\leq c\log|\mathscr{M}|,$$

where c is a small constant. Finally, applying Fano's Lemma (Lemma 23), we have

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T} \sup_{\|\boldsymbol{\theta}^{(1)}\|_2\leq\delta} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{A}\boldsymbol{\theta}^{(1)}\|_2 \geq c\delta\right) \geq 1 - \frac{\log 2}{\log|\mathscr{M}|} - c \geq \frac{1}{10}.$$

• If $h < \frac{1}{\zeta_{[T]}}\sqrt{\frac{pr}{nT}} + \frac{1}{\zeta_{[T]}}\sqrt{\frac{r}{n}} + \frac{1}{\zeta_{[T]}}\frac{\epsilon r}{\sqrt{n}}$: then $\zeta^{(t)}h < \frac{\zeta^{(t)}}{\zeta_{[T]}}\sqrt{\frac{pr}{nT}} + \frac{\zeta^{(t)}}{\zeta_{[T]}}\sqrt{\frac{r}{n}} + \frac{\zeta^{(t)}}{\zeta_{[T]}}\frac{\epsilon r}{\sqrt{n}}$ for all $t \in [T]$.

82

○ If $\sqrt{\frac{r}{n}} \geq \frac{\zeta^{(t_0)}}{\bar{\zeta}_{[T]}} \sqrt{\frac{pr}{nT}} + \frac{\zeta^{(t_0)}}{\bar{\zeta}_{[T]}} \sqrt{\frac{r}{n}} + \frac{\zeta^{(t_0)}}{\bar{\zeta}_{[T]}} \frac{\epsilon r}{\sqrt{n}}$ for $t_0 \in \arg\min_{t \in [T]} \zeta^{(t)}$: then we take $S = [T]$ and it suffices to prove the lower bound

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T} \sup_{\{\boldsymbol{\beta}^{(t)}\}_{t \in [T]} \in \mathscr{B}([T], 0)} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(t_0)} - \boldsymbol{\beta}^{(t_0)}\|_2 \gtrsim \sqrt{\frac{r}{n}} \right) \geq \frac{1}{10}.$$

This has already been proved in the previous analysis.

○ If $\sqrt{\frac{r}{n}} < \frac{\zeta^{(t)}}{\bar{\zeta}_{[T]}} \sqrt{\frac{pr}{nT}} + \frac{\zeta^{(t)}}{\bar{\zeta}_{[T]}} \sqrt{\frac{r}{n}} + \frac{\zeta^{(t)}}{\bar{\zeta}_{[T]}} \frac{\epsilon r}{\sqrt{n}}$ for all $t \in [T]$:

⋆ If $\sqrt{\frac{r}{n}} \geq \sqrt{\frac{pr}{nT}} + \epsilon \frac{r}{\sqrt{n}}$: then we take $S = [T]$ and it suffices to prove

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T} \sup_{\{\boldsymbol{\beta}^{(t)}\}_{t \in [T]} \in \mathscr{B}([T], 0)} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(t_0)} - \boldsymbol{\beta}^{(t_0)}\|_2 \gtrsim \frac{\zeta^{(t_0)}}{\bar{\zeta}} \sqrt{\frac{r}{n}} \right) \geq \frac{1}{10},$$

with $t_0 \in \arg\min_{t \in [T]} \zeta^{(t)}$, which is automatically true because $\zeta^{(t_0)}/\bar{\zeta} \leq 1$ and (28).

⋆ If $\sqrt{\frac{r}{n}} < \sqrt{\frac{pr}{nT}} + \epsilon \frac{r}{\sqrt{n}}$:

▷ If $\sqrt{\frac{pr}{nT}} \geq \epsilon \frac{r}{\sqrt{n}}$: then we take $S = [T]$ and prove

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T} \sup_{\{\boldsymbol{\beta}^{(t)}\}_{t \in [T]} \in \mathscr{B}([T], 0)} \mathbb{P}\left(\bigcup_{t \in [T]} \left\{ \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \gtrsim \frac{\zeta^{(t)}}{\bar{\zeta}_{[T]}} \sqrt{\frac{pr}{nT}} \right\} \right) \geq \frac{1}{10}.$$

Note that it suffices to prove

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T} \sup_{\{\boldsymbol{\beta}^{(t)}\}_{t \in [T]} \in \mathscr{B}([T], 0)} \mathbb{P}\left(\frac{1}{T} \sum_{t \in [T]} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2^2 \gtrsim \frac{pr}{nT} \right) \geq \frac{1}{10},$$

because if $\frac{1}{T} \sum_{t \in [T]} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2^2 \gtrsim \frac{pr}{nT}$, then there must exist $t \in [T]$ such that $\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2^2 \gtrsim (\frac{\zeta^{(t)}}{\bar{\zeta}_{[T]}})^2 \frac{pr}{nT}$.

Now let us prove it. Consider an $\overline{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$ such that the packing number $M(\mathcal{B}_\delta(\overline{\boldsymbol{A}}, \mathcal{O}^{p \times r}, \text{dist}_2), \text{dist}_F, \alpha\delta) \geq (\frac{C\sqrt{r}}{\alpha})^{r(p-r)}$ in Lemma 22 with $\delta = \frac{1}{\bar{\zeta}_{[T]}} \sqrt{\frac{pr}{nT}}$. Let $\mathcal{M} = \{\boldsymbol{A}_j\}_{j=1}^{|\mathcal{M}|}$ be the maximum packing of $\mathcal{B}_\delta(\overline{\boldsymbol{A}}, \mathcal{O}^{p \times r}, \text{dist}_2)$ corresponding to the packing number $M(\mathcal{B}_\delta(\overline{\boldsymbol{A}}, \mathcal{O}^{p \times r}, \text{dist}_2), \text{dist}_F, c\sqrt{r}\delta)$ with a very small constant $c > 0$. This leads to a $c'\bar{\zeta}_{[T]}\delta$-packing $\mathscr{M} = \{\{\boldsymbol{\beta}^{(t)}\}_{t=1}^T : \boldsymbol{\beta}^{(t)} := \boldsymbol{A}_j \boldsymbol{R}_j \boldsymbol{\theta}^{(t)}, \boldsymbol{R}_j = \arg\min_{\boldsymbol{R} \in \mathcal{O}^{r \times r}} \|\overline{\boldsymbol{A}} - \boldsymbol{A}_j \boldsymbol{R}\|_2$, for all $t \in [T]$ and the same $j \in [|\mathcal{M}|]\}$ of the space $\mathscr{B}([T], 0)$ w.r.t. distance $\sqrt{\frac{1}{T} \sum_{t \in [T]} \|\boldsymbol{\beta}^{(t)} - \widetilde{\boldsymbol{\beta}}^{(t)}\|_2^2}$. To verify this, notice that for any $j_1 \neq j_2 \in [|\mathcal{M}|]$, $\boldsymbol{\beta}^{(t)} = \boldsymbol{A}_{j_1} \boldsymbol{R}_{j_1} \boldsymbol{\theta}^{(t)}$, $\widetilde{\boldsymbol{\beta}}^{(t)} = \boldsymbol{A}_{j_2} \boldsymbol{R}_{j_2} \boldsymbol{\theta}^{(t)}$, with $\boldsymbol{R}_{j_1} = \arg\min_{\boldsymbol{R} \in \mathcal{O}^{r \times r}} \|\overline{\boldsymbol{A}} - \boldsymbol{A}_{j_1} \boldsymbol{R}\|_2$, $\boldsymbol{R}_{j_2} = \arg\min_{\boldsymbol{R} \in \mathcal{O}^{r \times r}} \|\overline{\boldsymbol{A}} - \boldsymbol{A}_{j_2} \boldsymbol{R}\|_2$,

by Wedin's $\sin\Theta$-Theorem and Assumption 2,

$$\frac{1}{T}\sum_{t=1}^{T}\|\boldsymbol{\beta}^{(t)}-\widetilde{\boldsymbol{\beta}}^{(t)}\|_2^2 \geq \frac{1}{T}\|\boldsymbol{A}_{j_1}\boldsymbol{\Theta}_{j_1}-\boldsymbol{A}_{j_2}\boldsymbol{\Theta}_{j_2}\|_{\mathrm{F}}^2 \geq \frac{c}{r}\|\boldsymbol{A}_{j_1}\boldsymbol{A}_{j_1}^\top-\boldsymbol{A}_{j_2}\boldsymbol{A}_{j_2}^\top\|_{\mathrm{F}}^2 \geq (c')^2\bar{\zeta}^2\delta^2,$$

where c' is a small constant, $\boldsymbol{\Theta}_{j_1}=\{\boldsymbol{R}_{j_1}\boldsymbol{\theta}^{(t)}\}_{t\in[T]}$, and $\boldsymbol{\Theta}_{j_2}=\{\boldsymbol{R}_{j_2}\boldsymbol{\theta}^{(t)}\}_{t\in[T]}$. On the other hand, notice that for any $j\in[|\mathcal{M}|]$, $\boldsymbol{\beta}^{(t)}=\boldsymbol{A}_j\boldsymbol{R}_j\boldsymbol{\theta}^{(t)}$ with $\boldsymbol{R}_j=\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\|\bar{\boldsymbol{A}}-\boldsymbol{A}_j\boldsymbol{R}\|_2$, by Lemma 20, for all $t\in[T]$, we have

$$\begin{aligned}
\|\boldsymbol{\beta}^{(t)}-\bar{\boldsymbol{A}}\boldsymbol{\theta}^{(t)}\|_2 &\leq \|\boldsymbol{A}_j\boldsymbol{R}_j-\bar{\boldsymbol{A}}\|_2\|\boldsymbol{\theta}^{(t)}\|_2\\
&= \zeta^{(t)}\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\|\boldsymbol{A}_j\boldsymbol{R}-\bar{\boldsymbol{A}}\|_2\\
&\leq \sqrt{2}\zeta^{(t)}\|\boldsymbol{A}_j\boldsymbol{A}_j^\top-\bar{\boldsymbol{A}}(\bar{\boldsymbol{A}})^\top\|_2\\
&\leq \sqrt{2}\zeta^{(t)}\delta,
\end{aligned}$$

which by triangle inequality leads to

$$\frac{1}{T}\sum_{t\in[T]}\|\boldsymbol{\beta}^{(t)}-\widetilde{\boldsymbol{\beta}}^{(t)}\|_2^2 \leq 4C^2\cdot\frac{1}{T}\sum_{t\in[T]}(\zeta^{(t)})^2\cdot\delta^2 \leq 4C^2\bar{\zeta}_{[T]}^2\delta^2, \qquad (29)$$

for any $\{\boldsymbol{\beta}^{(t)}\}_{t=1}^T$ and $\{\widetilde{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T\in\mathcal{M}$.

By Lemma 21, $|\mathcal{M}|=|\mathcal{M}|\gtrsim C^{r(p-r)}\geq C^{rp/2}$. Denote the distribution of $\{y_i^{(t)}\}_{i=1}^n$ given $\{\boldsymbol{x}_i^{(t)}\}_{i=1}^n$ as $\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}}^{(t)\otimes n}$ and the distribution of $\{\boldsymbol{x}_i^{(t)}\}_{i=1}^n$ as $\mathbb{P}_{\boldsymbol{x}}^{(t)\otimes n}$. Then for any $\{\boldsymbol{\beta}^{(t)}\}_{t=1}^T,\{\widetilde{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T\in\mathcal{M}$, by Lemma 24 and equation (29),

$$\begin{aligned}
\mathrm{KL}\left(\prod_{t=1}^T\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}^{(t)}}^{(t)\otimes n}\cdot\mathbb{P}_{\boldsymbol{x}}^{(t)\otimes n}\,\bigg\|\,\prod_{t=1}^T\mathbb{P}_{y|\boldsymbol{x};\widetilde{\boldsymbol{\beta}}^{(t)}}^{(t)\otimes n}\cdot\mathbb{P}_{\boldsymbol{x}}^{(t)\otimes n}\right) &\lesssim n\sum_{t=1}^T\|\boldsymbol{\beta}^{(t)}-\widetilde{\boldsymbol{\beta}}^{(t)}\|_2^2\\
&\lesssim nT\bar{\zeta}_{[T]}^2\delta^2\\
&\leq c\log|\mathcal{M}|,
\end{aligned}$$

where c is a small constant. Finally, applying Fano's Lemma (Lemma 23), we have

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T}\sup_{\{\boldsymbol{\beta}^{(t)}\}_{t=1}^T\in\mathscr{B}([T],0)}\mathbb{P}\left(\frac{1}{T}\sum_{t\in[T]}\|\widehat{\boldsymbol{\beta}}^{(t)}-\boldsymbol{\beta}^{(t)}\|_2^2\geq c^2\delta^2\right)\geq 1-\frac{\log 2}{\log|\mathscr{B}|}-c$$

$$\geq\frac{1}{10}.$$

$\triangleright$ If $\sqrt{\frac{pr}{nT}}<\epsilon\frac{r}{\sqrt{n}}$: then we consider $t_0\in\arg\min_{t\in[T]}\zeta^{(t)}$, $S\ni t_0$, and prove

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T}\sup_{S\subseteq\mathcal{S}}\sup_{\substack{\{\boldsymbol{\beta}^{(t)}\}_{t\in[T]}\in\mathscr{B}(S,0)\\\mathbb{Q}_{S^c}}}\mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(t_0)}-\boldsymbol{\beta}^{(t_0)}\|_2\gtrsim\frac{\zeta^{(t_0)}}{\bar{\zeta}_S}\epsilon\frac{r}{\sqrt{n}}\right)\geq\frac{1}{10}.$$

Note that $\frac{\zeta^{(t_0)}}{\bar{\zeta}_S} \leq 1$ for any $S \subseteq [T]$. Therefore it suffices to prove

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T} \sup_{S \subseteq \mathcal{S}} \sup_{\substack{\{\boldsymbol{\beta}^{(t)}\}_{t \in [T]} \in \mathscr{B}(S,0) \\ \mathbb{Q}_{S^c}}} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(t_0)} - \boldsymbol{\beta}^{(t_0)}\|_2 \gtrsim \epsilon \frac{r}{\sqrt{n}}\right) \geq \frac{1}{10}.$$

WLOG, assume $t_0 = 1$. Fix all $\boldsymbol{A}^{(t)} = $ some $\boldsymbol{A} \in \mathcal{O}^{p \times r}$ and $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^T \subseteq \min_{t \in [T]} \zeta^{(t)} \cdot \{\boldsymbol{e}_j\}_{j=1}^r \subseteq \mathbb{R}^r$ s.t. $\sigma_r(T^{-1/2}\{\boldsymbol{\theta}^{(t)}\}_{t=1}^T) \geq \frac{c}{\sqrt{r}}$ with $\#\{t : \boldsymbol{\theta}^{(t)} = \boldsymbol{e}_1\} = \lfloor T/r \rfloor$. Without loss of generality, suppose $\boldsymbol{\theta}^{(t)} = \min_{t \in [T]} \zeta^{(t)} \cdot \boldsymbol{e}_1$ when $t \in [\lfloor T/r \rfloor]$. Denote $\widetilde{\boldsymbol{\beta}} = \min_{t \in [T]} \zeta^{(t)} \cdot \boldsymbol{A}\boldsymbol{e}_1$, and $\boldsymbol{\beta}^{(t)} = \boldsymbol{A}\boldsymbol{\theta}^{(t)}$ for $t \geq \lfloor T/r \rfloor + 1$. Consider two data generating mechanisms in Lemma 26:

(I) $\{(\boldsymbol{x}_i^{(t)}, y_i^{(t)})\}_{i=1}^n \sim (1 - \epsilon')(\mathbb{P}_{y|\boldsymbol{x};\widetilde{\boldsymbol{\beta}}}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n}) + \epsilon'\mathbb{Q}$ independently for $t \in [\lfloor T/r \rfloor]$, where $\epsilon' = \frac{T\epsilon}{50\lfloor T/r \rfloor}$, and $\{\{\boldsymbol{x}_i^{(t)}\}_{i=1}^n\}_{t=\lfloor T/r \rfloor+1}^T \sim \mathbb{D} = \prod_{t=\lfloor T/r \rfloor+1}^T (\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}^{(t)}}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n})$;

(II) With a preserved set $S^c \subseteq [\lfloor T/r \rfloor]$, generate $\{\boldsymbol{x}^{(t)}\}_{t \in S^c} \sim \mathbb{Q}_{S^c}$ and $\{(\boldsymbol{x}_i^{(t)}, y_i^{(t)})\}_{i=1}^n \sim \mathbb{P}_{y|\boldsymbol{x};\widetilde{\boldsymbol{\beta}}}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n}$ independently for $t \in S \cap [\lfloor T/r \rfloor]$, and $\{\{\boldsymbol{x}_i^{(t)}\}_{i=1}^n\}_{t=\lfloor T/r \rfloor+1}^T \sim \mathbb{D} = \prod_{t=\lfloor T/r \rfloor+1}^T (\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}^{(t)}}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n})$.

Denote the joint distributions of $\{\boldsymbol{x}_i^{(t)}\}_{t \in [\lfloor T/r \rfloor]}$ in (I) and (II) as $\mathbb{P}_{(\epsilon,\theta,\mathbb{Q})}$ and $\mathbb{P}_{(S,\theta,\mathbb{Q}_{S^c})}$, respectively.

Note that by Lemma 24,

$$\varpi(\epsilon', \Theta) := \sup\{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2 : \mathrm{TV}\big(\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}_1}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n}, \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}_2}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n}\big) \leq \epsilon'/(1 - \epsilon')\}$$

$$\geq \sup\{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2 : \mathrm{KL}\big(\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}_1}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n}, \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}_2}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n}\big) \leq 2[\epsilon'/(1 - \epsilon')]^2\}$$

$$\geq \sup\{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2 : n\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 \leq c[\epsilon'/(1 - \epsilon')]^2\}$$

$$= c' \frac{\epsilon'}{\sqrt{n}}$$

$$\asymp \frac{r\epsilon}{\sqrt{n}}.$$

Then by Lemma 25,

$$\inf_{\widehat{\boldsymbol{\beta}}^{(1)}} \sup_{\widetilde{\boldsymbol{\beta}}, \mathbb{Q}} (\mathbb{P}_{(\epsilon,\theta,\mathbb{Q})} \cdot \mathbb{D})\left(\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)*}\|_2 \geq \varpi(\epsilon', \Theta)\right) \geq \frac{1}{2}.$$

Therefore, by Lemma 26, it follows that

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T} \sup_{S \subseteq \mathcal{S}} \sup_{\substack{\{\boldsymbol{\beta}^{(t)}\}_{t \in S} \in \mathscr{B}(S,0) \\ \mathbb{Q}_{S^c}}} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)*}\|_2 \geq c\frac{\epsilon r}{\sqrt{n}}\right)$$

$$\geq \inf_{\widehat{\boldsymbol{\beta}}^{(1)}} \sup_{S:|S|\geq T(1-\epsilon)} \sup_{\substack{\{\boldsymbol{\beta}^{(t)}\}_{t\in S}\in\mathscr{B}(S,0) \\ \mathbb{Q}_{S^c}}} (\mathbb{P}_{(S,\theta,\mathbb{Q}_{S^c})} \cdot \mathbb{D})\left(\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)*}\|_2 \geq \varpi(\epsilon/50,\Theta)\right)$$

$$\geq \frac{1}{10}.$$

Part 2: Next, we want to show (25). Consider the case that $S = [T]$ and $h = 0$. Fix all $\boldsymbol{A}^{(t)} = $ some $\boldsymbol{A} \in \mathcal{O}^{p\times r}$. Fix $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^T \subseteq \{\sqrt{c}\boldsymbol{e}_j\}_{j=1}^r$ satisfying $\frac{1}{T}\sum_{t=1}^T \boldsymbol{\theta}^{(t)}(\boldsymbol{\theta}^{(t)})^\top \succeq \frac{c}{r}\boldsymbol{I}_r$. We want to show

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T} \sup_{\{\boldsymbol{\theta}^{(t)}\}_{t=1}^T\in\Theta([T])} \mathbb{P}\left(\max_{t\in[T]} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{A}\boldsymbol{\theta}^{(t)}\|_2 \geq c\sqrt{\frac{\log T}{n}}\right) \geq \frac{1}{10},$$

where

$$\Theta([T]) = \left\{\{\boldsymbol{\theta}^{(t)}\}_{t\in[T]} \subseteq \mathbb{R}^r : \max_{t\in[T]} \|\boldsymbol{\theta}^{(t)}\|_2 \leq C, \frac{1}{T}\sum_{t=1}^T \boldsymbol{\theta}^{(t)}(\boldsymbol{\theta}^{(t)})^\top \succeq \frac{c}{r}\boldsymbol{I}_r\right\}.$$

Denote $\delta = \sqrt{\frac{\log T}{n}}$. Then $\mathscr{T} = \{\{\widetilde{\boldsymbol{\theta}}^{(t)}\}_{t=1}^T : \exists t_0 \text{ s.t. } \widetilde{\boldsymbol{\theta}}^{(t_0)} = \boldsymbol{\theta}^{(t_0)}(1 + \delta/\sqrt{c}), \widetilde{\boldsymbol{\theta}}^{(t)} = \boldsymbol{\theta}^{(t)} \text{ for } t \neq t_0\}$ is a $\sqrt{2}\delta$-packing of $\Theta(S)$ w.r.t. distance $\max_{t\in[T]} \|\bar{\boldsymbol{\theta}}^{(t)} - \widetilde{\boldsymbol{\theta}}^{(t)}\|_2$ with $\{\bar{\boldsymbol{\theta}}^{(t)}\}_{t=1}^T$ and $\{\widetilde{\boldsymbol{\theta}}^{(t)}\}_{t=1}^T \in \mathscr{T}$. Hence $\mathscr{B} = \{\{\boldsymbol{A}\widetilde{\boldsymbol{\theta}}^{(t)}\}_{t=1}^T : \exists t_0 \text{ s.t. } \widetilde{\boldsymbol{\theta}}^{(t_0)} = \boldsymbol{\theta}^{(t_0)}(1 + \delta/\sqrt{c}), \widetilde{\boldsymbol{\theta}}^{(t)} = \boldsymbol{\theta}^{(t)} \text{ for } t \neq t_0\}$ is a $\sqrt{2}\delta$-packing in $\mathbb{R}^p$ w.r.t. distance $\max_{t\in[T]} \|\boldsymbol{\beta}^{(t)} - \widetilde{\boldsymbol{\beta}}^{(t)}\|_2$ with $\{\boldsymbol{\beta}^{(t)}\}_{t=1}^T$ and $\{\widetilde{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T \in \mathscr{T}$. Apparently $|\mathscr{T}| = |\mathscr{B}| = T$. And for any $\{\boldsymbol{\beta}^{(t)}\}_{t=1}^T$ and $\{\widetilde{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T \in \mathscr{T}$, they only differ by two components. WLOG, suppose the indices of different components are t_1 and t_2. Then we have

$$\text{KL}\left(\prod_{t=1}^T \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}^{(t)}}^{(t)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(t)\otimes n} \middle\| \prod_{t=1}^T \mathbb{P}_{y|\boldsymbol{x};\widetilde{\boldsymbol{\beta}}^{(t)}}^{(t)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(t)\otimes n}\right)$$

$$= \text{KL}\left(\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\boldsymbol{\theta}^{(t_1)}(1+\delta/\sqrt{c})}^{(t_1)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(t_1)\otimes n} \cdot \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\boldsymbol{\theta}^{(t_2)}}^{(t_2)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(t_2)\otimes n} \middle\| \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\widetilde{\boldsymbol{\theta}}^{(t_1)}}^{(t_1)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(t_1)\otimes n} \cdot \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\boldsymbol{\theta}^{(t_2)}(1+\delta/\sqrt{c})}^{(t_2)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(t_2)\otimes n}\right)$$

$$\lesssim n\|\boldsymbol{A}\boldsymbol{\theta}^{(t_1)}(1 + \delta/\sqrt{c}) - \boldsymbol{A}\boldsymbol{\theta}^{(t_1)}\|_2^2 + n\|\boldsymbol{A}\boldsymbol{\theta}^{(t_2)} - \boldsymbol{A}\boldsymbol{\theta}^{(t_2)}(1 + \delta/\sqrt{c})\|_2^2$$
$$\lesssim n\delta^2$$
$$\leq c'\log|\mathscr{T}|,$$

where c is a small constant. By Fano's Lemma (Lemma 23), we have

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T} \sup_{\{\boldsymbol{\theta}^{(t)}\}_{t=1}^T\in\Theta([T])} \mathbb{P}\left(\max_{t\in[T]} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{A}\boldsymbol{\theta}^{(t)}\|_2 \geq c\sqrt{\frac{\log T}{n}}\right) \geq 1 - \frac{\log 2}{\log|\mathscr{T}|} - c' \geq \frac{1}{10}.$$

Part 3: Finally, let us discuss how to obtain the final desired lower bound by similar arguments to prove (24) and (25). Denote S' as the index set of $t \in [T]$ where $\zeta^{(t)}$ is among the largest $T(1 - \frac{c'}{r})$ ones of $\{\zeta^{(t)}\}_{t=1}^T$.

- If $\sqrt{\frac{\log T}{n}} \leq \min_{t \in S'} \left\{ \left[\frac{\zeta^{(t)}}{\bar{\zeta}_{S'}} \sqrt{\frac{pr}{nT}} + \left(\frac{\zeta^{(t)}}{\bar{\zeta}_{S'}} \vee 1 \right) \sqrt{\frac{r}{n}} + \zeta^{(t)} h + \frac{\zeta^{(t)}}{\bar{\zeta}_{S'}} \epsilon \frac{r}{\sqrt{n}} \right] \wedge \sqrt{\frac{p}{n}} \right\}$: then by the same arguments we used to prove (24), we can get a lower bound similar to (24) by replacing $[T]$ and $\bar{\zeta}_{[T]} = \sqrt{\frac{1}{T} \sum_{t \in [T]} (\zeta^{(t)})^2}$ with S' and $\bar{\zeta}_{S'} = \sqrt{\frac{1}{|S'|} \sum_{t \in S'} (\zeta^{(t)})^2}$. Then notice that by definition of S', $\bar{\zeta}_{S'} \gtrsim \bar{\zeta}_{[T]}$, hence we can get the bound

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^{T}} \sup_{S \subseteq \mathcal{S}} \sup_{\substack{\{\boldsymbol{\beta}^{(t)}\}_{t \in S} \in \mathscr{B}(S,h) \\ \mathbb{Q}_{S^c}}} \mathbb{P} \left(\bigcup_{t \in S'} \left\{ \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)}\|_2 \gtrsim \left[\frac{\zeta^{(t)}}{\bar{\zeta}_{S'}} \sqrt{\frac{pr}{nT}} + \left(\frac{\zeta^{(t)}}{\bar{\zeta}_{S'}} \vee 1 \right) \sqrt{\frac{r}{n}} + \zeta^{(t)} h \right. \right.$$

$$\left. \left. + \frac{\zeta^{(t)}}{\bar{\zeta}_{S'}} \frac{\epsilon r}{\sqrt{n}} \right] \wedge \sqrt{\frac{p}{n}} \right\} \right) \geq \frac{1}{10},$$

which implies the desired lower bound because $\sqrt{\frac{\log T}{n}} \leq \min_{t \in S'} \left\{ \left[\frac{\zeta^{(t)}}{\zeta} \sqrt{\frac{pr}{nT}} + \left(\frac{\zeta^{(t)}}{\zeta} \vee 1 \right) \sqrt{\frac{r}{n}} + \zeta^{(t)} h + \frac{\zeta^{(t)}}{\zeta} \epsilon \frac{r}{\sqrt{n}} \right] \wedge \sqrt{\frac{p}{n}} \right\}$.

- If $\sqrt{\frac{\log T}{n}} > \min_{t \in S'} \left\{ \left[\frac{\zeta^{(t)}}{\bar{\zeta}_{S'}} \sqrt{\frac{pr}{nT}} + \left(\frac{\zeta^{(t)}}{\bar{\zeta}_{S'}} \vee 1 \right) \sqrt{\frac{r}{n}} + \zeta^{(t)} h + \frac{\zeta^{(t)}}{\bar{\zeta}_{S'}} \epsilon \frac{r}{\sqrt{n}} \right] \wedge \sqrt{\frac{p}{n}} \right\}$: by definition of S', automatically we have $\sqrt{\frac{\log T}{n}} > \min_{t \in [T] \backslash S'} \left\{ \left[\frac{\zeta^{(t)}}{\bar{\zeta}_{S'}} \sqrt{\frac{pr}{nT}} + \left(\frac{\zeta^{(t)}}{\bar{\zeta}_{S'}} \vee 1 \right) \sqrt{\frac{r}{n}} + \zeta^{(t)} h + \frac{\zeta^{(t)}}{\bar{\zeta}_{S'}} \epsilon \frac{r}{\sqrt{n}} \right] \wedge \sqrt{\frac{p}{n}} \right\}$. Then by the same argument we used to prove (25), we can get the following bound by replacing $[T]$ in (25) with $[T] \backslash S'$:

$$\inf_{\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^{T}} \sup_{\{\boldsymbol{\beta}^{(t)}\}_{t \in [T]} \in \mathscr{B}([T],0)} \mathbb{P} \left(\max_{t \in [T] \backslash S'} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)}\|_2 \gtrsim \sqrt{\frac{\log(T - |S'|)}{n}} \right) \geq \frac{1}{10}.$$

This implies the desired lower bound, because when $T \geq r^{1.01}$ we have $\log(T - |S'|) \gtrsim \log(T/r) \gtrsim \log T^{\frac{0.01}{1.01}} \gtrsim \log T$.

A comment: In Part 1, when we derive the term $\frac{\zeta^{(t)}}{\zeta} \epsilon \frac{r}{\sqrt{n}}$ in the lower bound, if $\epsilon r > 1$, the lower bound can be strengthened to $\sqrt{\frac{p}{n}}$. This follows by contaminating all but one of the $\lfloor T/r \rfloor$ tasks such that $\boldsymbol{\theta}^{(t)} = \min_{t \in [T]} \zeta^{(t)} \cdot \boldsymbol{e}_1$. We can then apply the same reasoning used in Part 1 to derive (26), which leads to the lower bound $\sqrt{\frac{p}{n}}$. This result is quite intuitive: the problem effectively reduces to estimating a p-dimensional parameter using only n samples, since the coefficients of the other tasks are orthogonal to the current task. This phenomenon justifies the condition $\epsilon r \lesssim 1$ for the lower and upper bounds of both proposed methods.

D.10 Proof of Theorem 14

If Algorithm 4 is coupled with Algorithm 1, the proof idea is very similar to the proof of Theorem 2. When the representation learning helps, an argument based on Lemma 39, Proposition 42, and Lemma 41.(i) lead to the corresponding term. In the other case, Lemma 41.(ii) guarantees that the single-task rate holds all the time. A combination of these two situations entails the final TL upper bound. The details are as follows.

Denote $\eta = r\sqrt{\frac{p}{nT}} + \sqrt{r}h + \sqrt{r}\sqrt{\frac{r+\log T}{n}} + \frac{|S^c|}{T} \cdot r \cdot \frac{\lambda}{\sqrt{n}}$.

(i) When $\eta \leq C\sqrt{\frac{p}{n_0}}$: note that $\widehat{\overline{A}}$ is **independent** of $\{x_i^{(0)}, y_i^{(0)}\}_{i=1}^{n_0}$, which is the key for part (i) to be correct by only requiring $n \geq Cr$ (see Remark 13). Hence by Lemma 39 and an argument by first conditioning on $\widehat{\overline{A}}$ then taking the expectation,

$$\|\widehat{\overline{A}}\theta_{\widehat{\overline{A}}}^{(0)} - A^{(0)*}\theta^{(0)*}\|_2 \lesssim \|\widehat{\overline{A}}(\widehat{\overline{A}})^\top - A^{(0)*}(A^{(0)*})^\top\|_2 + \sqrt{\frac{r}{n_0}}$$

$$\leq \|\widehat{\overline{A}}(\widehat{\overline{A}})^\top - \overline{A}(\overline{A})^\top\|_2 + \|\overline{A}(\overline{A})^\top - A^{(0)*}(A^{(0)*})^\top\|_2 + \sqrt{\frac{r}{n_0}}$$

$$\lesssim \eta + \sqrt{\frac{r}{n_0}}$$

w.p. at least $1 - e^{-C(r+\log T)}$, where we used Proposision 42 in the last step. Then by Lemma 41.(i), $\frac{\gamma}{\sqrt{n_0}} \gtrsim \sqrt{\frac{p}{n_0}} \geq \|\nabla f^{(t)}(A^{(0)*}\theta^{(0)*})\|_2 + C\|\widehat{\overline{A}}\theta_{\widehat{\overline{A}}}^{(0)} - A^{(0)*}\theta^{(0)*}\|_2$, which implies that

$$\|\widehat{\beta}^{(0)} - \beta^{(0)*}\|_2 = \|\widehat{\overline{A}}\theta_{\widehat{\overline{A}}}^{(0)} - A^{(0)*}\theta^{(0)*}\|_2 \leq \eta,$$

w.p. at least $1 - e^{-C(r+\log T)}$.

(ii) By Lemma 41.(ii), we always have $\|\widehat{\beta}^{(0)} - \beta^{(0)*}\|_2 \lesssim \sqrt{p/n_0}$ w.p. at least $1 - e^{-C(p+\log T)}$.

If Algorithm 4 is coupled with Algorithm 2, we can follow the discussions above and the proof of Theorem 7 to get the desired result.

D.11 Proof of Theorem 15

Similar to the proof of Theorem 8, we prove the following parts one by one. Combining them together entails the lower bound. Throughout this subsection, we assume the following generative model for tasks in $\{0\} \cup S$:

$$y|x \sim \mathbb{P}_{y|x,\beta} = N(x^\top\beta, 1), \quad x \sim \mathbb{P}_x,$$

where $\mathbb{P}_x$ is sub-Gaussian with $\Sigma = \mathbb{E}(xx^\top)$. Suppose there exist constants c, C such that $0 < c \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C < \infty$. Then any joint distribution $\mathbb{P}$ of (x, y) can be written as $\mathbb{P}_{x,y} = \mathbb{P}_{y|x;\beta} \cdot \mathbb{P}_x$.

Denote

$$\mathcal{A}_0(S, h) = \left\{\{A^{(t)}\}_{t\in\{0\}\cup S} \subseteq \mathcal{O}^{p\times r} : \max_{t\in S} \|A^{(t)}(A^{(t)})^\top - A^{(0)}(A^{(0)})^\top\|_2 \leq h\right\},$$

$$\Theta_0(S) = \left\{\{\theta^{(t)}\}_{t\in\{0\}\cup S} \subseteq \mathbb{R}^r : \max_{t\in\{0\}\cup S} \|\theta^{(t)}\|_2 \leq C, \frac{1}{|S|}\sum_{t\in S}\theta^{(t)}(\theta^{(t)})^\top \succeq \frac{c}{r}I_r\right\}.$$

(i) Consider the case $S = [T]$ and $h = 0$. For any $A \in \mathbb{R}^{p\times r}$ with $C \geq \sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_r(A) \geq c > 0$, consider its SVD where $A = Q\Lambda V$ where $Q \in \mathcal{O}^{p\times r}$ and $V \in \mathcal{O}^{r\times r}$. Denote $R = \Lambda V$. Then $0 < c \leq \sigma_{\min}(R) \leq \sigma_{\max}(R) \leq C < \infty$, and $A\theta^{(t)} = QR\theta^{(t)} = Q\widetilde{\theta}^{(t)}$. Note that $Q \in \mathcal{O}^{p\times r}$ and $\widetilde{\theta}^{(t)} = R\theta^{(t)}$ with $\|\widetilde{\theta}^{(t)}\|_2 \leq \|R\|_2\|\theta^{(t)}\|_2 \leq C'$

and $\frac{1}{T}\sum_{t=1}^{T}\widetilde{\boldsymbol{\theta}}^{(t)}(\widetilde{\boldsymbol{\theta}}^{(t)})^{\top} = \boldsymbol{R}\big[\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\theta}^{(t)}(\boldsymbol{\theta}^{(t)})^{\top}\big]\boldsymbol{R}^{\top} \succeq \frac{cc'}{r}\boldsymbol{I}_r$ if $\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\theta}^{(t)}(\boldsymbol{\theta}^{(t)})^{\top} \succeq \frac{c'}{r}\boldsymbol{I}_r$. Therefore, fixing $\boldsymbol{\theta}^{(t)} \in \{\boldsymbol{e}_j\}_{j=1}^{r}$ and $\boldsymbol{\theta}^{(0)} = \frac{1}{\sqrt{r}}\boldsymbol{1}_r$ s.t. $\{\boldsymbol{\theta}^{(t)}\}_{t=0}^{T} \in \Theta_0([T])$, we know that

$$\inf_{\widehat{\boldsymbol{\beta}}^{(0)}} \sup_{\{\boldsymbol{\beta}^{(t)}\}_{t=1}^{T}\in\mathscr{B}_0([T],0)} \mathbb{P}\bigg(\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{A}^{(0)}\boldsymbol{\theta}^{(0)}\|_2 \gtrsim \sqrt{\frac{pr}{nT}} \wedge \sqrt{\frac{p}{n_0}}\bigg)$$

$$\geq \inf_{\widehat{\boldsymbol{\beta}}^{(0)}} \sup_{\substack{0<c\leq\sigma_{\max}(\boldsymbol{A})\leq\sigma_{\max}(\boldsymbol{A})\leq C \\ \|\boldsymbol{\theta}^{(0)}\|_2\leq C'}} \mathbb{P}\bigg(\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{A}\boldsymbol{\theta}^{(0)}\|_2 \gtrsim \sqrt{\frac{pr}{nT}} \wedge \sqrt{\frac{p}{n_0}}\bigg)$$

$$:= (*),$$

where $\boldsymbol{A}^{(t)} = \boldsymbol{A}$ for all $t \in \{0\} \cup [T]$. Let $\delta = \sqrt{\frac{pr}{nT}} \wedge \sqrt{\frac{p}{n_0}}$. Consider a $\delta/2$-packing of the ball $\mathcal{B}_\delta(\boldsymbol{0}, \mathbb{R}^p, \|\cdot\|_2)$ (denoted as $\mathcal{M}$). By Example 5.8 in Wainwright (2019), $|\mathcal{M}| \geq 5^p$. Consider

$$\boldsymbol{A} = \overline{\boldsymbol{A}} + \frac{c'}{\sqrt{r}}\cdot\boldsymbol{u}\boldsymbol{1}_r^{\top}, \quad \widetilde{\boldsymbol{A}} = \overline{\boldsymbol{A}} + \frac{c'}{\sqrt{r}}\cdot\widetilde{\boldsymbol{u}}\boldsymbol{1}_r^{\top},$$

where $\boldsymbol{u} \neq \widetilde{\boldsymbol{u}} \in \mathcal{M}$, $c' > 0$ is a small constant such that $\sigma_{\min}(\boldsymbol{A}), \sigma_{\min}(\widetilde{\boldsymbol{A}}) \geq 1 - c' \geq c$ and $\sigma_{\max}(\boldsymbol{A}), \sigma_{\max}(\widetilde{\boldsymbol{A}}) \leq 1 + c' \leq C$. Then

$$\|(\boldsymbol{A} - \widetilde{\boldsymbol{A}})\boldsymbol{\theta}^{(0)}\|_2 = \frac{1}{r}\|(\boldsymbol{u} - \widetilde{\boldsymbol{u}})\boldsymbol{1}_r^{\top}\boldsymbol{1}_r\|_2 = \|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\|_2 \in \bigg[\frac{1}{2}\delta, 2\delta\bigg], \tag{30}$$

$$\|(\boldsymbol{A} - \widetilde{\boldsymbol{A}})\boldsymbol{\theta}^{(t)}\|_2 = \|(\boldsymbol{A} - \widetilde{\boldsymbol{A}})\boldsymbol{e}_{j_k}\|_2 = \frac{1}{\sqrt{r}}\|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\|_2 \leq \frac{\delta}{\sqrt{r}}. \tag{31}$$

where $\boldsymbol{\theta}^{(t)} = \boldsymbol{e}_{j_k}$ with $j_t \in [r]$. Therefore $\mathscr{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\theta}^{(0)}, \boldsymbol{A} = \overline{\boldsymbol{A}} + \frac{1}{\sqrt{r}}\cdot\boldsymbol{u}\boldsymbol{1}_r^{\top}, \boldsymbol{u} \in \mathcal{M}\}$ becomes a $\delta/2$-packing in $\mathbb{R}^p$ with $|\mathscr{B}| = |\mathcal{M}| \geq 5^p$. Furthermore, for any $\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\theta}^{(0)}$ and $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{A}}\boldsymbol{\theta}^{(0)} \in \mathscr{B}$ with $\boldsymbol{A} \neq \widetilde{\boldsymbol{A}}$, by Lemma 24 and equations (30) and (31),

$$\text{KL}\left(\prod_{t=1}^{T}\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\boldsymbol{\theta}^{(t)}}^{(t)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(t)\otimes n} \cdot \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\boldsymbol{\theta}^{(0)}}^{(0)\otimes n_0} \cdot \mathbb{P}_{\boldsymbol{x}}^{(0)\otimes n_0} \,\bigg\|\, \prod_{t=1}^{T}\mathbb{P}_{y|\boldsymbol{x};\widetilde{\boldsymbol{A}}\boldsymbol{\theta}^{(t)}}^{(t)\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(t)\otimes n} \cdot \mathbb{P}_{y|\boldsymbol{x};\widetilde{\boldsymbol{A}}\boldsymbol{\theta}^{(0)}}^{(0)\otimes n_0} \cdot \mathbb{P}_{\boldsymbol{x}}^{(0)\otimes n_0}\right)$$

$$\lesssim n\sum_{t=1}^{T}\|(\boldsymbol{A} - \widetilde{\boldsymbol{A}})\boldsymbol{\theta}^{(t)}\|_2^2 + n_0\|(\boldsymbol{A} - \widetilde{\boldsymbol{A}})\boldsymbol{\theta}^{(0)}\|_2^2$$

$$\lesssim nT\cdot\frac{\delta^2}{r} + n_0\delta^2$$

$$\leq c\log|\mathscr{B}|.$$

Then by Fano's lemma (Lemma 23), $(*) \geq 1 - \frac{\log 2}{\log|\mathscr{B}|} - c \geq 1/10$.

(ii) Consider the case $S = [T]$. For all $t \in [T]$, fix $\boldsymbol{A}^{(t)} =$ an arbitrary $\overline{\boldsymbol{A}} \in \mathcal{O}^{p\times r}$. Fix $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{T} \subseteq \{\boldsymbol{e}_j\}_{j=1}^{r}$ and $\boldsymbol{\theta}^{(0)} = \frac{1}{\sqrt{r}}\boldsymbol{1}_r$ s.t. $\{\boldsymbol{\theta}^{(t)}\}_{t=0}^{T} \in \Theta_0([T])$. Set $\delta = \frac{h}{12} \wedge \sqrt{\frac{p}{n_0}}$. Consider a $\delta/2$-packing of the ball $\mathcal{B}_\delta(\boldsymbol{0}, \mathbb{R}^p, \|\cdot\|_2)$ (denoted as $\mathcal{M}$). By Example 5.8 in Wainwright (2019), $|\mathcal{M}| \geq 5^p$. Consider

$$\boldsymbol{A} = \overline{\boldsymbol{A}} + \frac{c'}{\sqrt{r}}\cdot\boldsymbol{u}\boldsymbol{1}_r^{\top}, \quad \widetilde{\boldsymbol{A}} = \overline{\boldsymbol{A}} + \frac{c'}{\sqrt{r}}\cdot\widetilde{\boldsymbol{u}}\boldsymbol{1}_r^{\top},$$

where $\boldsymbol{u} \neq \widetilde{\boldsymbol{u}} \in \mathcal{M}$ and c' is a small constant, then

$$\|(\boldsymbol{A} - \widetilde{\boldsymbol{A}})\boldsymbol{\theta}^{(0)}\|_2 = \frac{1}{r}\|(\boldsymbol{u} - \widetilde{\boldsymbol{u}})\mathbf{1}_r^\top \mathbf{1}_r\|_2 = \|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\|_2 \in \left[\frac{c'}{2}\delta, 2c'\delta\right], \tag{32}$$

hence $\mathcal{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\theta}^{(0)}, \boldsymbol{A} = \overline{\boldsymbol{A}} + \frac{c'}{\sqrt{r}} \cdot \boldsymbol{u}\mathbf{1}_r^\top, \boldsymbol{u} \in \mathcal{M}\}$ is a $c'\delta/2$-packing in $\mathbb{R}^p$ with $|\mathcal{B}| = |\mathcal{M}| \geq 5^p$. On the other hand, for any $\boldsymbol{A} \in \mathbb{R}^{p \times r}$ with $1 + 2c'\delta \geq \sigma_1(\boldsymbol{A}) \geq \sigma_2(\boldsymbol{A}) \geq \cdots \geq \sigma_r(\boldsymbol{A}) \geq 1 - 2c'\delta > 0$ and $\|\boldsymbol{A} - \overline{\boldsymbol{A}}\|_2 \leq \delta$, consider its SVD where $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{V}$ where $\boldsymbol{Q} \in \mathcal{O}^{p \times r}$ and $\boldsymbol{V} \in \mathcal{O}^{r \times r}$. Denote $\boldsymbol{R} = \boldsymbol{\Lambda}\boldsymbol{V}$. Then $0 < 1 - 2c'\delta \leq \sigma_{\min}(\boldsymbol{R}) \leq \sigma_{\max}(\boldsymbol{R}) \leq 1 + 2c'\delta < \infty$, and $\boldsymbol{A}\boldsymbol{\theta}^{(0)} = \boldsymbol{Q}\boldsymbol{R}\boldsymbol{\theta}^{(0)} = \boldsymbol{Q}\widetilde{\boldsymbol{\theta}}^{(0)}$. Note that $\boldsymbol{Q} \in \mathcal{O}^{p \times r}$ and $\widetilde{\boldsymbol{\theta}}^{(0)} = \boldsymbol{R}\boldsymbol{\theta}^{(0)}$ with $\|\widetilde{\boldsymbol{\theta}}^{(0)}\|_2 \leq \|\boldsymbol{R}\|_2\|\boldsymbol{\theta}^{(0)}\|_2 \leq C[1 + 2c']$. And $\|\boldsymbol{Q}\boldsymbol{Q}^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \leq \|(\boldsymbol{Q}\boldsymbol{R})(\boldsymbol{Q}\boldsymbol{R})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 + \|\boldsymbol{Q}(\boldsymbol{I}_r - \boldsymbol{R}\boldsymbol{R}^\top)\boldsymbol{Q}^\top\|_2 \leq 4\|\boldsymbol{A} - \overline{\boldsymbol{A}}\|_2 + 4\|\boldsymbol{I}_r - \boldsymbol{R}\|_2 \leq 4\delta + 8c'\delta \leq 12\delta \leq h$. Therefore,

$$\inf_{\substack{\widehat{\boldsymbol{\beta}}^{(0)} \\ \{\boldsymbol{\theta}^{(t)}\}_{t \in \{0\} \cup S} \in \Theta_0([T])}} \sup_{\{\boldsymbol{A}^{(t)}\}_{t \in \{0\} \cup S} \in \mathcal{A}_0([T], h)} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{A}^{(0)}\boldsymbol{\theta}^{(0)}\|_2 \gtrsim h \wedge \sqrt{\frac{p}{n_0}}\right)$$

$$\geq \inf_{\widehat{\boldsymbol{\beta}}^{(0)}} \sup_{\substack{0 < c \leq \sigma_{\max}(\boldsymbol{A}) \leq \sigma_{\max}(\boldsymbol{A}) \leq C \\ \|\boldsymbol{\theta}^{(0)}\|_2 \leq C'}} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{A}\boldsymbol{\theta}^{(0)}\|_2 \gtrsim h \wedge \sqrt{\frac{p}{n_0}}\right)$$

$$:= (*),$$

where all $\boldsymbol{A}^{(t)} =$ an arbitrary $\overline{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$, and $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^T \subseteq \{\boldsymbol{e}_j\}_{j=1}^r$ are fixed.

For any $\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\theta}^{(0)}$ and $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{A}}\boldsymbol{\theta}^{(0)} \in \mathcal{B}$ with $\boldsymbol{A} \neq \widetilde{\boldsymbol{A}}$, by Lemma 24 and equation (32),

$$\mathrm{KL}\left(\prod_{t=1}^T \mathbb{P}_{y|\boldsymbol{x};\overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)}}^{(t) \otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(t) \otimes n} \cdot \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\boldsymbol{\theta}^{(0)}}^{(0) \otimes n_0} \cdot \mathbb{P}_{\boldsymbol{x}}^{(0) \otimes n_0} \middle\| \prod_{t=1}^T \mathbb{P}_{y|\boldsymbol{x};\overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)}}^{(t) \otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(t) \otimes n} \cdot \mathbb{P}_{y|\boldsymbol{x};\widetilde{\boldsymbol{A}}\boldsymbol{\theta}^{(0)}}^{(0) \otimes n_0} \cdot \mathbb{P}_{\boldsymbol{x}}^{(0) \otimes n_0}\right)$$

$$= \mathrm{KL}\left(\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\boldsymbol{\theta}^{(0)}}^{(0) \otimes n_0} \cdot \mathbb{P}_{\boldsymbol{x}}^{(0) \otimes n_0} \middle\| \mathbb{P}_{y|\boldsymbol{x};\widetilde{\boldsymbol{A}}\boldsymbol{\theta}^{(0)}}^{(0) \otimes n_0} \cdot \mathbb{P}_{\boldsymbol{x}}^{(0) \otimes n_0}\right)$$

$$\lesssim n_0 \|(\boldsymbol{A} - \widetilde{\boldsymbol{A}})\boldsymbol{\theta}^{(0)}\|_2^2$$

$$\lesssim n_0 \delta^2$$

$$\leq c \log |\mathcal{B}|.$$

Then by Fano's lemma (Lemma 23), $(*) \geq 1 - \frac{\log 2}{\log |\mathcal{B}|} - c \geq 1/10$.

(iii) Consider the case that $S = [T]$ and $h = 0$. Fix all $\boldsymbol{A}^{(t)} =$ some $\boldsymbol{A} \in \mathcal{O}^{p \times r}$. Fixing $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^T$ such that $\frac{1}{T}\sum_{t=1}^T \boldsymbol{\theta}^{(t)}(\boldsymbol{\theta}^{(t)})^\top \succeq \frac{c}{r}\boldsymbol{I}_r$. We want to show

$$\inf_{\widehat{\boldsymbol{\beta}}^{(0)}} \sup_{\|\boldsymbol{\theta}^{(0)}\|_2 \leq C} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{A}\boldsymbol{\theta}^{(0)}\|_2 \geq c\sqrt{\frac{r}{n_0}}\right) \geq \frac{1}{10}.$$

Denote $\delta = \sqrt{\frac{r}{n_0}}$. Consider a $c\delta$-packing of $\mathcal{B}_\delta(\boldsymbol{0}, \mathbb{R}^r, \|\cdot\|_2)$ (denoted as $\mathcal{T}$). By Example 5.8 in Wainwright (2019), we know that $\log |\mathcal{T}| \gtrsim r$. Denote the distribution of $\{y_i^{(0)}\}_{i=1}^{n_0}$ given $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^{n_0}$ as $\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}}^{(0) \otimes n_0}$ and the distribution of $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^{n_0}$ as $\mathbb{P}_{\boldsymbol{x}}^{(0) \otimes n_0}$. For any $\boldsymbol{\theta}^{(0)} \neq \widetilde{\boldsymbol{\theta}}^{(0)} \in \mathcal{T}$:

$$\mathrm{KL}\left(\prod_{t=1}^T \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\boldsymbol{\theta}^{(t)}}^{(t) \otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(t) \otimes n} \cdot \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\boldsymbol{\theta}^{(0)}}^{(0) \otimes n_0} \cdot \mathbb{P}_{\boldsymbol{x}}^{(0) \otimes n_0} \middle\| \prod_{t=1}^T \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\boldsymbol{\theta}^{(t)}}^{(t) \otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{(t) \otimes n} \cdot \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\widetilde{\boldsymbol{\theta}}^{(0)}}^{(0) \otimes n_0} \cdot \mathbb{P}_{\boldsymbol{x}}^{(0) \otimes n_0}\right)$$

90

$$\text{KL}\left(\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\boldsymbol{\theta}^{(0)}}^{(0)\otimes n_0} \cdot \mathbb{P}_{\boldsymbol{x}}^{(0)\otimes n_0} \middle\| \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{A}\widetilde{\boldsymbol{\theta}}^{(0)}}^{(0)\otimes n_0} \cdot \mathbb{P}_{\boldsymbol{x}}^{(0)\otimes n_0}\right)$$

$$\lesssim n_0 \|\boldsymbol{A}\boldsymbol{\theta}^{(0)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}^{(0)}\|_2^2$$

$$\lesssim r$$

$$\leq c\log|\mathscr{T}|,$$

where c is a small constant. Finally, applying Fano's Lemma (Lemma 23), we have

$$\inf_{\widehat{\boldsymbol{\beta}}^{(0)}} \sup_{\|\boldsymbol{\theta}^{(0)}\|_2 \leq \delta} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{A}\boldsymbol{\theta}^{(0)}\|_2 \geq c\delta\right) \geq 1 - \frac{\log 2}{\log|\mathscr{T}|} - c \geq \frac{1}{10}.$$

(iv) We follow a similar analysis in part (v) of the proof of Theorem 8. Consider the case $h = 0$. Fix all $\boldsymbol{A}^{(t)} = $ some $\boldsymbol{A} \in \mathcal{O}^{p\times r}$ and $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^T \subseteq \{\boldsymbol{e}_j\}_{j=1}^r \subseteq \mathbb{R}^r$ s.t. $\sigma_r(T^{-1/2}\{\boldsymbol{\theta}^{(t)}\}_{t=1}^T) \geq \frac{c}{\sqrt{r}}$ with $\#\{t : \boldsymbol{\theta}^{(t)} = \boldsymbol{e}_1\} = \lfloor T/r\rfloor$. Without loss of generality, suppose $\boldsymbol{\theta}^{(t)} = \boldsymbol{e}_1$ when $t \in \{0\} \cup \lfloor\lfloor T/r\rfloor\rfloor$. Denote $\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{e}_1$, and $\boldsymbol{\beta}^{(t)} = \boldsymbol{A}\boldsymbol{\theta}^{(t)}$ for $t \geq \lfloor T/r\rfloor + 1$. Consider two data generating mechanisms in Lemma 26:

(I) $\{(\boldsymbol{x}_i^{(t)}, y_i^{(t)})\}_{i=1}^n \sim (1-\epsilon')(\mathbb{P}_{y|\boldsymbol{x};\widetilde{\boldsymbol{\beta}}}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n}) + \epsilon'\mathbb{Q}$ independently for $t \in \lfloor\lfloor T/r\rfloor\rfloor$, where $\epsilon' = \frac{T\epsilon}{50\lfloor T/r\rfloor}$, $\{(\boldsymbol{x}_i^{(0)}, y_i^{(0)})\}_{i=1}^{n_0}$, and $\{\{\boldsymbol{x}_i^{(t)}\}_{i=1}^n\}_{t=\lfloor T/r\rfloor+1}^T \sim \mathbb{D} = \prod_{t=\lfloor T/r\rfloor+1}^T(\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}^{(t)}}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n})$;

(II) With a preserved set $S^c \subseteq \lfloor\lfloor T/r\rfloor\rfloor$, generate $\{\boldsymbol{x}^{(t)}\}_{t\in S^c} \sim \mathbb{Q}_{S^c}$ and $\{(\boldsymbol{x}_i^{(t)}, y_i^{(t)})\}_{i=1}^n \sim \mathbb{P}_{y|\boldsymbol{x};\widetilde{\boldsymbol{\beta}}}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n}$ independently for $t \in \{0\} \cup (S \cap \lfloor\lfloor T/r\rfloor\rfloor)$, and $\{\{\boldsymbol{x}_i^{(t)}\}_{i=1}^n\}_{t=\lfloor T/r\rfloor+1}^T \sim \mathbb{D}\{\{\boldsymbol{x}_i^{(t)}\}_{i=1}^n\}_{t=\lfloor T/r\rfloor+1}^T \sim \mathbb{D} = \prod_{t=\lfloor T/r\rfloor+1}^T(\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}^{(t)}}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n})$.

Denote the joint distributions of $\{\boldsymbol{x}_i^{(t)}\}_{t\in\{0\}\cup\lfloor\lfloor T/r\rfloor\rfloor}$ in (I) and (II) as $\mathbb{P}_{(\epsilon,\theta,\mathbb{Q})}$ and $\mathbb{P}_{(S,\theta,\mathbb{Q}_{S^c})}$, respectively.

Note that by Lemma 24,

$$\varpi(\epsilon', \Theta)$$
$$:= \sup\{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2 : \text{TV}\left(\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}_1}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n}, \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}_2}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n}\right) \leq \epsilon'/(1-\epsilon'),$$
$$\text{TV}\left(\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}_1}^{\otimes n_0} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n}, \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}_2}^{\otimes n_0} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n_0}\right) \leq 1/20\}$$
$$\geq \sup\{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2 : \text{KL}\left(\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}_1}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n}, \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}_2}^{\otimes n} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n}\right) \leq 2[\epsilon'/(1-\epsilon')]^2,$$
$$\text{KL}\left(\mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}_1}^{\otimes n_0} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n}, \mathbb{P}_{y|\boldsymbol{x};\boldsymbol{\beta}_2}^{\otimes n_0} \cdot \mathbb{P}_{\boldsymbol{x}}^{\otimes n_0}\right) \leq 1/200\}$$
$$\geq \sup\{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2 : n\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 \leq c[\epsilon'/(1-\epsilon')]^2, n_0\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 \leq c\}$$
$$= c'\frac{\epsilon'}{\sqrt{n}} \wedge \frac{1}{\sqrt{n_0}}$$
$$\asymp \frac{r\epsilon}{\sqrt{n}} \wedge \frac{1}{\sqrt{n_0}}.$$

Then by Lemma 27,

$$\inf_{\widehat{\boldsymbol{\beta}}^{(0)}} \sup_{\widetilde{\boldsymbol{\beta}},\mathbb{Q}}(\mathbb{P}_{(\epsilon,\theta,\mathbb{Q})} \cdot \mathbb{D})\left(\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)*}\|_2 \geq \varpi(\epsilon', \Theta)\right) \geq \frac{1}{2}.$$

91

Therefore, by Lemma 28, it follows that

$$\inf_{\widehat{\boldsymbol{\beta}}^{(0)}} \sup_{S \subseteq \mathcal{S}} \sup_{\substack{\{\boldsymbol{\beta}^{(t)}\}_{t \in \{0\} \cup S} \in \mathscr{B}(S,0) \\ \mathbb{Q}_{S^c}}} \mathbb{P}\left(\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)*}\|_2 \geq c \frac{\epsilon r}{\sqrt{n}} \wedge \frac{1}{\sqrt{n_0}} \right)$$

$$\geq \inf_{\widehat{\boldsymbol{\beta}}^{(0)}} \sup_{S:|S| \geq T(1-\epsilon)} \sup_{\substack{\{\boldsymbol{\beta}^{(t)}\}_{t \in \{0\} \cup S} \in \mathscr{B}(S,0) \\ \mathbb{Q}_{S^c}}} (\mathbb{P}_{(S,\theta,\mathbb{Q}_{S^c})} \cdot \mathbb{D}) \left(\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)*}\|_2 \geq \varpi(\epsilon/50, \Theta) \right)$$

$$\geq \frac{1}{10}.$$

D.12 Proofs of Lemmas and Propositions

Denote $\widehat{\boldsymbol{\Sigma}}^{(t)} = \sum_{i=1}^{n} \frac{1}{n} \boldsymbol{x}_i^{(t)} (\boldsymbol{x}_i^{(t)})^{\top}$.

D.12.1 PROOF OF LEMMA 29

By optimality of $\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}$ and $\boldsymbol{\theta}_{\boldsymbol{A}^{(t)*}}^{(t)}$,

$$\boldsymbol{A}^{\top} \nabla f^{(t)}(\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}) = (\boldsymbol{A}^{(t)*})^{\top} \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}_{\boldsymbol{A}^{(t)*}}^{(t)}) = \boldsymbol{0}_r, \tag{33}$$

which entails

$$\boldsymbol{A}^{\top} \left[\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}) + \widehat{\boldsymbol{\Sigma}}^{(t)}(\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}) \right] = \boldsymbol{0},$$

$$\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} = (\boldsymbol{A}^{\top}\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1} \boldsymbol{A}^{\top}\widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*} - (\boldsymbol{A}^{\top}\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1} \boldsymbol{A}^{\top} \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}).$$

Therefore,

$$\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq \underbrace{\|[\boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^{\top}\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1}\boldsymbol{A}^{\top}\widehat{\boldsymbol{\Sigma}}^{(t)}]\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2}_{[1]}$$

$$+ \underbrace{\|\boldsymbol{A}(\boldsymbol{A}^{\top}\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1}\boldsymbol{A}^{\top}\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2}_{[2]}.$$

Note that

$$[1] \lesssim \|[\boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^{\top}\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1}\boldsymbol{A}^{\top}\widehat{\boldsymbol{\Sigma}}^{(t)}]\boldsymbol{A}^{(t)*}\|_2 \|\boldsymbol{\theta}^{(t)*}\|_2$$

$$\lesssim \|(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1/2}\|_2 \cdot \|(\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\boldsymbol{A}^{(t)*} - (\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\boldsymbol{A}(\boldsymbol{A}^{\top}\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1}\boldsymbol{A}^{\top}(\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}(\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\boldsymbol{A}^{(t)*}\|_2 \|\boldsymbol{\theta}^{(t)*}\|_2 \quad \text{k}$$

$$\lesssim \|[\boldsymbol{I} - \boldsymbol{B}(\boldsymbol{B}^{\top}\boldsymbol{B})^{-1}\boldsymbol{B}^{\top}]\boldsymbol{B}^{(t)*}\|_2 \|\boldsymbol{\theta}^{(t)*}\|_2$$

$$\lesssim \|P_{\boldsymbol{B}^{\perp}}\boldsymbol{B}^{(t)*}\|_2 \|\boldsymbol{\theta}^{(t)*}\|_2, \tag{34}$$

w.p. at least $1 - e^{-Cp}$, where $\boldsymbol{B} = (\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\boldsymbol{A}$, $\boldsymbol{B}^{(t)*} = (\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\boldsymbol{A}^{(t)*}$, $\boldsymbol{A}^{\perp} \in \{\tilde{\boldsymbol{A}} \in \mathcal{O}^{p \times (p-r)}\} : \tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^{\top} + \boldsymbol{A}\boldsymbol{A}^{\top} = \boldsymbol{I}_p\}$, $\boldsymbol{B}^{\perp} = (\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1/2}\boldsymbol{A}^{\perp}$, $P_{\boldsymbol{B}^{\perp}} = \boldsymbol{B}^{\perp}((\boldsymbol{B}^{\perp})^{\top}\boldsymbol{B}^{\perp})^{-1}(\boldsymbol{B}^{\perp})^{\top} = \boldsymbol{I} - P_{\boldsymbol{B}}$. Further, notice that

$$\|P_{\boldsymbol{B}^{\perp}}\boldsymbol{B}^{(t)*}\|_2 = \|(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1/2}\boldsymbol{A}^{\perp}[(\boldsymbol{A}^{\perp})^{\top}(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1}\boldsymbol{A}^{\perp}]^{-1}(\boldsymbol{A}^{\perp})^{\top}\boldsymbol{A}^{(t)*}\|_2$$

---

k. This step holds because of Lemma 18, which shows that $\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}^{(t)}) > 0$ (so that $\widehat{\boldsymbol{\Sigma}}^{(t)}$ is invertible) w.p. at least $1 - e^{-C(p+\log T)}$.

$$\leq \|(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1/2}\|_2 \cdot \lambda_{\min}^{-1}((\boldsymbol{A}^{\perp})^{\top}(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1}\boldsymbol{A}^{\perp}) \cdot \|(\boldsymbol{A}^{\perp})^{\top}\boldsymbol{A}^{(t)*}\|_2$$

$$\lesssim \|(\boldsymbol{A}^{\perp})^{\top}\boldsymbol{A}^{(t)*}\|_2$$

$$\lesssim \|\boldsymbol{A}\boldsymbol{A}^{\top} - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^{\top}\|_2, \tag{35}$$

w.p. at least $1 - e^{-C(p+\log T)}$. And

$$[2] \leq \|\boldsymbol{A}(\boldsymbol{A}^{\top}\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1}\boldsymbol{A}^{\top}\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}_{\boldsymbol{A}^{(t)*}}^{(t)})\|_2$$

$$+ \|\boldsymbol{A}(\boldsymbol{A}^{\top}\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1}\boldsymbol{A}^{\top}[\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}) - \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}_{\boldsymbol{A}^{(t)*}}^{(t)})]\|_2$$

$$\lesssim \|\boldsymbol{A}^{\top}(\boldsymbol{A}^{(t)*})^{\perp}\|_2 \cdot \|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}_{\boldsymbol{A}^{(t)*}}^{(t)})\|_2 + \|\boldsymbol{\theta}^{(t)*} - \boldsymbol{\theta}_{\boldsymbol{A}^{(t)*}}^{(t)}\|_2 \tag{36}$$

$$\lesssim \|\boldsymbol{A}\boldsymbol{A}^{\top} - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^{\top}\|_2(\|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 + C\|\boldsymbol{\theta}^{(t)*} - \boldsymbol{\theta}_{\boldsymbol{A}^{(t)*}}^{(t)}\|_2) + \|\boldsymbol{\theta}^{(t)*} - \boldsymbol{\theta}_{\boldsymbol{A}^{(t)*}}^{(t)}\|_2$$

$$\lesssim \|\boldsymbol{A}\boldsymbol{A}^{\top} - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^{\top}\|_2\sqrt{\frac{p + \log T}{n}} + \sqrt{\frac{r + \log T}{n}}, \tag{37}$$

w.p. at least $1 - e^{-C(r+\log T)}$, where we used the fact that $\|\boldsymbol{\theta}^{(t)*} - \boldsymbol{\theta}_{\boldsymbol{A}^{(t)*}}^{(t)}\|_2 \lesssim \sqrt{\frac{r+\log T}{n}}$ for all $t \in S$ w.p. at least $1 - e^{-C(r+\log T)}$ (which is the standard rate of r-dimensional linear regression). Inequality (36) holds because there exists $\boldsymbol{a} \in \mathbb{R}^r$ such that $\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}) = (\boldsymbol{A}^{(t)*})^{\perp}\boldsymbol{a}$ (due to (33)) and $\|\boldsymbol{a}\|_2 = \|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 \lesssim \sqrt{\frac{p+\log T}{n}}$ w.p. at least $1 - e^{C(p+\log T)}$. Putting (34), (35), and (37) together, we complete the proof.

D.12.2 PROOF OF LEMMA 30

By Lemma 29, w.p. at least $1 - e^{-C(r+\log T)}$,

$$\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}\|_2 - \|\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq \|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \cdot \zeta^{(t)}$$

$$\lesssim \|\boldsymbol{A}\boldsymbol{A}^{\top} - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^{\top}\|_2 \cdot \zeta^{(t)} + \sqrt{\frac{r + \log T}{n}}$$

$$\leq C'\zeta^{(t)}.$$

Note that $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}\|_2 = \|\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}\|_2$ and $\|\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 = \|\boldsymbol{\theta}^{(t)*}\|_2 \leq \zeta^{(t)}$.

D.12.3 PROOF OF LEMMA 31

The proof is almost the same as the proof of Lemma 29, thus omitted.

D.12.4 PROOF OF LEMMA 32

Denote $\boldsymbol{G} = \{\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\}_{t \in S}$ and $\Delta^{(t)} = \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$. By Lemma 29, w.p. at least $1 - e^{-C(r+\log T)}$, for all $t \in S$,

$$\|\Delta^{(t)}\|_2 \lesssim h\zeta^{(t)} + \sqrt{\frac{r + \log T}{n}},$$

leading to

$$\|\{\boldsymbol{\Delta}^{(t)}\}_{t\in S}\|_2 = \sup_{\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1} \left| \sum_{t\in S} u_t(\boldsymbol{\Delta}^{(t)})^\top \boldsymbol{v} \right|$$

$$\leq \sup_{\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1} \|\boldsymbol{u}\|_2 \cdot \sqrt{\sum_{t\in S}((\boldsymbol{\Delta}^{(t)})^\top \boldsymbol{v})^2}$$

$$\lesssim \sqrt{T}\left(h\bar{\zeta} + \sqrt{\frac{r + \log T}{n}} \right). \tag{38}$$

By Lemma D.3 in Duan and Wang (2023) or Theorem 5.39 in Vershynin (2010), w.p. at least $1 - e^{-C(p+\log T)}$,

$$\|\boldsymbol{G}\|_2 \lesssim \sqrt{\frac{p+T}{n}}. \tag{39}$$

Combining (38), (39), and the fact that $\|\overline{\boldsymbol{G}}\|_2 \leq \|\boldsymbol{G}\|_2 + \|\{\boldsymbol{\Delta}^{(t)}\}_{t\in S}\|_2$ implies the desired result.

D.12.5 PROOF OF LEMMA 33

Note that

$$\|\overline{\boldsymbol{G}} - \boldsymbol{G}\|_2 = \sup_{\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1} \left| \sum_{j=1}^{d} u_j(\overline{\boldsymbol{g}}_j - \boldsymbol{g}_j)^\top \boldsymbol{v} \right|$$

$$\leq \sup_{\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1} \|\boldsymbol{u}\|_2 \cdot \sqrt{\sum_{j=1}^{d}[(\overline{\boldsymbol{g}}_j - \boldsymbol{g}_j)^\top \boldsymbol{v}]^2}$$

$$\leq \sqrt{d}\Delta,$$

which completes the proof.

D.12.6 PROOF OF LEMMA 34

By convexity of $f^{(t)}$,

$$f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}_{\boldsymbol{A}^{(t)}}) - f^{(t)}(\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}}) + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$\geq \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + \nabla f^{(t)}(\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}})^\top(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}_{\boldsymbol{A}^{(t)}} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}})$$

$$\geq \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 - \|\nabla f^{(t)}(\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}})^\top\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}_{\boldsymbol{A}^{(t)}}\|_2 \tag{40}$$

$$\geq \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 - \|\nabla f^{(t)}(\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}})\|_2 \cdot \|((\widehat{\overline{\boldsymbol{A}}})^\perp)^\top\boldsymbol{A}^{(t)}\|_2 \cdot \|\boldsymbol{\theta}^{(t)}_{\boldsymbol{A}^{(t)}}\|_2 \tag{41}$$

$$\geq \left(\frac{\lambda}{\sqrt{n}} - C\|\nabla f^{(t)}(\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}})\|_2 \cdot \zeta^{(t)} \right) \cdot \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 \tag{42}$$

$$> 0,$$

when $\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top \neq \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top$, where (40) and (41) are due to the definition of $\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}}$ (which leads to $\nabla f^{(t)}(\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}})^\top \widehat{\overline{\boldsymbol{A}}} = \boldsymbol{0}$ and $\nabla f^{(t)}(\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}}) = (\widehat{\overline{\boldsymbol{A}}})^\perp \boldsymbol{u}$ with $\|\boldsymbol{u}\|_2 = \|f^{(t)}(\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}})\|_2$), and (42) is due to Lemma 20. Note that the above strict inequality leads to a contradiction with the fact that $(\boldsymbol{A}^{(t)}, \boldsymbol{\theta}^{(t)})$ is a minimizer, therefore we must have $\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top \neq \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top$ for all $t \in S$.

D.12.7 PROOF OF LEMMA 35

It is straightforward to see that

$$\frac{1}{|S'|}\sum_{t \in S'} \boldsymbol{\theta}^{(t)*}(\boldsymbol{\theta}^{(t)*})^\top = \frac{|S|}{|S'|}\left(\frac{1}{|S|}\sum_{t \in S}\boldsymbol{\theta}^{(t)*}(\boldsymbol{\theta}^{(t)*})^\top - \frac{1}{|S|}\sum_{t \in S \setminus S'}\boldsymbol{\theta}^{(t)*}(\boldsymbol{\theta}^{(t)*})^\top\right)$$

$$\succeq \frac{|S|}{|S'|}\left(\frac{c}{r}\bar{\zeta}^2 - \alpha\max_{t \in S}(\zeta^{(t)})^2\right)\boldsymbol{I}_r$$

$$\succeq \frac{c}{2r}\bar{\zeta}^2,$$

because $|S|/|S'| \geq 1$ and $\alpha \leq c/r \cdot \bar{\zeta}^2/[2\max_{t \in S}(\zeta^{(t)})^2]$ with a very small $c'' > 0$.

D.12.8 PROOF OF LEMMA 36

By definition, we have

$$0 \leq f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}) - f^{(t)}(\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}) + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 - \frac{\lambda}{\sqrt{n}}\|\widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2,$$

which implies

$$\frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top\|_2 \geq f^{(t)}(\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}) - f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})$$

$$\geq \frac{1}{2}(\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})^\top\widehat{\boldsymbol{\Sigma}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})$$

$$- \|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 \cdot \|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2$$

$$\geq c\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2 - C\sqrt{\frac{p + \log T}{n}}\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2,$$

$$(43)$$

w.p. at least $1 - e^{-C'(p+\log T)}$.

Next, we will argue how to upper bound $\|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top\|_2$ by $\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2$ so that solving the inequality can lead to the desired conclusion. Note that upper bounding $\|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \widehat{\boldsymbol{A}}^{(t)}(\widehat{\boldsymbol{A}}^{(t)})^\top\|_2$ by $\|\widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2$ is not always applicable. Fortunately, for the analysis purpose, we can pick any $\boldsymbol{A}^{(t)*}$ whose column space is the same and the corresponding $\boldsymbol{\theta}^{(t)*}$ without changing their product $\boldsymbol{\beta}^{(t)*}$. The same argument holds for $\widehat{\boldsymbol{A}}^{(t)}$ and $\widehat{\boldsymbol{\theta}}^{(t)}$ as well. So WLOG, we denote $\widehat{\boldsymbol{A}}^{(t)} = (\widehat{\boldsymbol{a}}_1, \widehat{\boldsymbol{A}}_{-1})$,

$A^{(t)*} = (a_1^*, A_{-1}^*)$, and consider $\widehat{\theta} = (\widehat{\theta}_1, \mathbf{0}_{r-1}^\top)^\top$, $\theta^{(t)*} = (\theta_1^*, \mathbf{0}_{r-1}^\top)^\top$, where $\widehat{\theta}_1 \widehat{a}_1 = \widehat{\beta}^{(t)}$ and $\theta_1^* a_1^* = \beta^{(t)*}$, no matter which $\widehat{A}_{-1}$ and A_{-1}^* we pick to encode $\mathrm{Col}(\widehat{A}^{(t)})$ and $\mathrm{Col}(A^{(t)*})$.

Consider $A_{-1}^* = (I - P_{a_1^*})\widehat{A}_{-1}R$, where $R = (\widehat{A}_{-1}^\top(I - P_{a_1^*})\widehat{A}_{-1})^{-1/2}$ serving as a normalization matrix which makes $(A_{-1}^*)^\top A_{-1}^* = I_r$. Here $P_{a_1^*} = a_1^*(a_1^*)^\top / \|a_1^*\|_2^2$ is the projection matrix corresponding to a_1^*.

Putting everything together, note that

$$\|\widehat{A}^{(t)}(\widehat{A}^{(t)})^\top - A^{(t)*}(A^{(t)*})^\top\|_2 \le \|a_1^*(a_1^*)^\top - \widehat{a}_1(\widehat{a}_1)^\top\|_2 + \|\widehat{A}_{-1}(\widehat{A}_{-1})^\top - A_{-1}^*(A_{-1}^*)^\top\|_2, \tag{44}$$

where

$$\begin{aligned}
&\|\widehat{A}_{-1}(\widehat{A}_{-1})^\top - A_{-1}^*(A_{-1}^*)^\top\|_2 \\
&\le \|(I - P_{\widehat{a}_1})\widehat{A}_{-1}(\widehat{A}_{-1})^\top(I - P_{\widehat{a}_1}) - (I - P_{a_1^*})\widehat{A}_{-1}RR^\top(\widehat{A}_{-1})^\top(I - P_{a_1^*})\|_2 \\
&\le \|P_{a_1^*}(I - P_{\widehat{a}_1})\|_2 + \|(I - P_{\widehat{a}_1})P_{a_1^*}\|_2 + \|(I - P_{a_1^*})\widehat{A}_{-1}(I - RR^\top)(\widehat{A}_{-1})^\top(I - P_{a_1^*})\|_2 \\
\\
&\lesssim \|a_1^*(a_1^*)^\top - \widehat{a}_1(\widehat{a}_1)^\top\|_2 + \|I - RR^\top\|_2.
\end{aligned}$$

Recall that $RR^\top = [\widehat{A}_{-1}^\top(I - P_{a_1^*})\widehat{A}_{-1}]^{-1} = (I - \widehat{A}_{-1}^\top P_{a_1^*}\widehat{A}_{-1})^{-1}$. Therefore,

$$\|I - RR^\top\|_2 \lesssim \|(I - \widehat{A}_{-1}^\top P_{a_1^*}\widehat{A}_{-1})^{-1}\|_2 \cdot \|I - \widehat{A}_{-1}^\top P_{a_1^*}\widehat{A}_{-1} - I\|_2 \lesssim \|a_1^*(a_1^*)^\top - \widehat{a}_1(\widehat{a}_1)^\top\|_2.$$

Plugging these bounds back to (44), we have

$$\|\widehat{A}^{(t)}(\widehat{A}^{(t)})^\top - A^{(t)*}(A^{(t)*})^\top\|_2 \lesssim \|a_1^*(a_1^*)^\top - \widehat{a}_1(\widehat{a}_1)^\top\|_2.$$

Then by Lemma 37, w.p. at least $1 - e^{-C'(p + \log T)}$, for all $t \in S$,

$$\zeta^{(t)}\|a_1^*(a_1^*)^\top - \widehat{a}_1(\widehat{a}_1)^\top\|_2 \lesssim \|\widehat{\theta}_1\widehat{a}_1 - \theta_1^* a_1^*\|_2 + \sqrt{\frac{p + \log T}{n}} = \|\widehat{A}^{(t)}\widehat{\theta}^{(t)} - A^{(t)*}\theta^{(t)*}\|_2 + \sqrt{\frac{p + \log T}{n}}.$$

Plugging it back to (43), we have

$$\|\widehat{A}^{(t)}\widehat{\theta}^{(t)} - A^{(t)*}\theta^{(t)*}\|_2 \lesssim \sqrt{\frac{p + \log T}{n}} + \frac{\lambda}{\sqrt{n}\zeta^{(t)}} + \left(\frac{p + \log T}{n}\right)^{1/4}\left(\frac{\lambda}{\zeta^{(t)}\sqrt{n}}\right)^{1/2}$$

for all $t \in S$, w.p. at least $1 - e^{-C'(p + \log T)}$.

D.12.9 PROOF OF LEMMA 37

The proof idea is very similar to the proof of Lemma 29. By the optimality condition, $A^\top \nabla f^{(t)}(A\theta_A^{(t)}) = 0$, implying that

$$\begin{aligned}
&A^\top\left[\nabla f^{(t)}(A^{(t)*}\theta^{(t)*}) + \widehat{\Sigma}^{(t)}(A\theta_A^{(t)} - A^{(t)*}\theta^{(t)*})\right] = 0, \\
&\theta_A^{(t)} = (A^\top\widehat{\Sigma}^{(t)}A)^{-1}A^\top\widehat{\Sigma}^{(t)}A^{(t)*}\theta^{(t)*} - (A^\top\widehat{\Sigma}^{(t)}A)^{-1}A^\top\nabla f^{(t)}(A^{(t)*}\theta^{(t)*}).
\end{aligned}$$

Note that $\boldsymbol{\theta}^{(t)*} \in \mathbb{R}$ and $\boldsymbol{A}^{(t)*} \in \mathbb{R}^p$ when $r = 1$. Hence, w.p. at least $1 - e^{-C'(p+\log T)}$, we have

$$\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(A)} - \boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)*}\|_2$$
$$\geq \|[\boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\widehat{\boldsymbol{\Sigma}}^{(t)}]\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 - \|\boldsymbol{A}(\boldsymbol{A}^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2$$
$$\geq \|[\boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\widehat{\boldsymbol{\Sigma}}^{(t)}]\boldsymbol{A}^{(t)*}\|_2\zeta^{(t)} - \|\boldsymbol{A}\|_2 \cdot \|\boldsymbol{A}(\boldsymbol{A}^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1}\|_2 \cdot \|\boldsymbol{A}\|_2 \cdot \|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2$$

$$\geq \zeta^{(t)}\lambda_{\min}((\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1/2}) \cdot \|(\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\boldsymbol{A}^{(t)*} - (\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\boldsymbol{A}(\boldsymbol{A}^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A}^{(t)*}\|_2 - C\sqrt{\frac{p+\log T}{n}}$$

$$\geq C'\zeta^{(t)}\|[\boldsymbol{I} - \boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B})^{-1})\boldsymbol{B}^\top]\boldsymbol{B}^{(t)*}\|_2 - C\sqrt{\frac{p+\log T}{n}}, \tag{45}$$

where $\boldsymbol{B} = (\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\boldsymbol{A}$, $\boldsymbol{B}^{(t)*} = (\widehat{\boldsymbol{\Sigma}}^{(t)})^{1/2}\boldsymbol{A}^{(t)*}$, $\boldsymbol{A}^\perp \in \{\tilde{\boldsymbol{A}} \in \mathcal{O}^{p\times(p-1)}\} : \tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^\top + \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I}_p\}$, $\boldsymbol{B}^\perp = (\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1/2}\boldsymbol{A}^\perp$, $P_{\boldsymbol{B}^\perp} = \boldsymbol{B}^\perp((\boldsymbol{B}^\perp)^\top\boldsymbol{B}^\perp)^{-1}(\boldsymbol{B}^\perp)^\top = \boldsymbol{I} - P_{\boldsymbol{B}}$. Further, notice that

$$\|P_{\boldsymbol{B}^\perp}\boldsymbol{B}^{(t)*}\|_2 = \|(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1/2}\boldsymbol{A}^\perp[(\boldsymbol{A}^\perp)^\top(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1}\boldsymbol{A}^\perp]^{-1}(\boldsymbol{A}^\perp)^\top\boldsymbol{A}^{(t)*}\|_2$$
$$\geq \lambda_{\min}^{1/2}([(\boldsymbol{A}^\perp)^\top(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1}\boldsymbol{A}^\perp]^{-1}) \cdot \|(\boldsymbol{A}^\perp)^\top\boldsymbol{A}^{(t)*}\|_2$$
$$= \|(\boldsymbol{A}^\perp)^\top(\widehat{\boldsymbol{\Sigma}}^{(t)})^{-1}\boldsymbol{A}^\perp\|_2^{-1/2} \cdot \|(\boldsymbol{A}^\perp)^\top\boldsymbol{A}^{(t)*}\|_2$$
$$\gtrsim \|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top\|_2, \tag{46}$$

w.p. at least $1 - e^{-C(p+\log T)}$. Combining (45) and (46) completes the proof.

D.12.10 PROOF OF LEMMA 38

First, by Lemma 18, $0 < c \leq \min_{t\in S}\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}^{(t)}) \leq \max_{t\in S}\lambda_{\max}(\widehat{\boldsymbol{\Sigma}}^{(t)}) \leq C < \infty$ w.p. at least $1 - e^{-C'(p+\log T)}$. Then (i) is the direct consequence of Lemmas E.3 in Duan and Wang (2023). (ii) and (iii) come from Lemma E.2 in Duan and Wang (2023).

D.12.11 PROOF OF LEMMA 39

We modify a few steps in the proof of Lemma 29. The key point here is that $\boldsymbol{A} \in \mathcal{O}^{p\times r}$ is fixed.

By optimality of $\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}$ and $\boldsymbol{\theta}_{\boldsymbol{A}^{(t)*}}^{(t)}$,

$$\boldsymbol{A}^\top\nabla f^{(t)}(\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}) = (\boldsymbol{A}^{(t)*})^\top\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}_{\boldsymbol{A}^{(t)*}}^{(t)}) = \boldsymbol{0}_r,$$

which entails

$$\boldsymbol{A}^\top\left[\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}) + \widehat{\boldsymbol{\Sigma}}^{(t)}(\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\right] = \boldsymbol{0},$$
$$\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} = (\boldsymbol{A}^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*} - (\boldsymbol{A}^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}).$$

Therefore,

$$\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq \underbrace{\|[\boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^\top\widehat{\boldsymbol{\Sigma}}^{(t)}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\widehat{\boldsymbol{\Sigma}}^{(t)}]\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2}_{[1]}$$

$$+ \underbrace{\|\boldsymbol{A}(\boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A})^{-1} \boldsymbol{A}^\top \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2}_{[2]}.$$

Note that

$$\begin{aligned}
[1] &\lesssim \|\boldsymbol{A}^{(t)*} - \boldsymbol{A}(\boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A})^{-1} \boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A}^{(t)*}\|_2 \\
&= \|\boldsymbol{A}^{(t)*} - \boldsymbol{A}(\boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A})^{-1} \boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} (\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{A}^\perp (\boldsymbol{A}^\perp)^\top) \boldsymbol{A}^{(t)*}\|_2 \\
&\leq \|(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^\top)\boldsymbol{A}^{(t)*}\|_2 + \|\boldsymbol{A}(\boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A})^{-1} \boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A}^\perp (\boldsymbol{A}^\perp)^\top \boldsymbol{A}^{(t)*}\|_2 \\
&\lesssim \|(\boldsymbol{A}^\perp)^\top \boldsymbol{A}^{(t)*}\|_2 + \|(\boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A})^{-1}\|_2 \cdot \|\boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A}^\perp\|_2 \cdot \|(\boldsymbol{A}^\perp)^\top \boldsymbol{A}^{(t)*}\|_2
\end{aligned}$$

By Lemma 19 and the condition $n_0 \geq Cr$,

$$\begin{aligned}
\|(\boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A})^{-1}\|_2 &= \lambda_{\min}^{-1}(\boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A}) \\
&\leq [\lambda_{\min}(\boldsymbol{A}^\top \boldsymbol{\Sigma}^{(t)} \boldsymbol{A}) - \|\boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A} - \boldsymbol{A}^\top \boldsymbol{\Sigma}^{(t)} \boldsymbol{A}\|_2]^{-1} \\
&\leq [\lambda_{\min}(\boldsymbol{A}^\top \boldsymbol{\Sigma}^{(t)} \boldsymbol{A}) - C\sqrt{r/n_0}]^{-1} \\
&\lesssim C'',
\end{aligned}$$

w.p. at least $1 - e^{-C'(r+\log T)}$. And similarly

$$\|\boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A}^\perp\|_2 \lesssim \|\boldsymbol{A}^\top \boldsymbol{\Sigma}^{(t)} \boldsymbol{A}^\perp\|_2 + \|\boldsymbol{A}^\top(\widehat{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma}^{(t)})\boldsymbol{A}^\perp\|_2 \lesssim C,$$

w.p. at least $1 - e^{-C'(r+\log T)}$. Therefore by Lemma 20, w.p. at least $1 - e^{-C(r+\log T)}$,

$$[1] \lesssim \|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top\|_2. \tag{47}$$

And w.p. at least $1 - e^{-C(r+\log T)}$,

$$[2] \lesssim \|(\boldsymbol{A}^\top \widehat{\boldsymbol{\Sigma}}^{(t)} \boldsymbol{A})^{-1}\|_2 \cdot \|\boldsymbol{A}^\top \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 \lesssim \sqrt{\frac{r}{n_0}}, \tag{48}$$

Denote $\{\boldsymbol{\beta}_j\}_{j=1}^N$ as a 1/2-cover (whose components are inside the set to be covered) of $\mathscr{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{u}, \|\boldsymbol{u}\|_2 \leq 1\}$ (which is isomorphic to the unit ball in $\mathbb{R}^r$), with $N \leq 5^r$ (by Example 5.8 in Wainwright (2019)). The second inequality in (48) holds because

$$\begin{aligned}
\|\boldsymbol{A}^\top \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 &= \sup_{\|\boldsymbol{u}\|_2 \leq 1} (\boldsymbol{A}\boldsymbol{u})^\top \left[\frac{2}{n_0}(\boldsymbol{X}^{(0)})^\top \boldsymbol{\epsilon}^{(0)}\right] \\
&= \sup_{\boldsymbol{\beta} \in \mathscr{B}} \boldsymbol{\beta}^\top \left[\frac{2}{n_0}(\boldsymbol{X}^{(0)})^\top \boldsymbol{\epsilon}^{(0)}\right] \\
&\leq \sup_{\boldsymbol{\beta} \in \mathscr{B}} (\boldsymbol{\beta} - \boldsymbol{\beta}_{j_0}^\top)^\top \left[\frac{2}{n_0}(\boldsymbol{X}^{(0)})^\top \boldsymbol{\epsilon}^{(0)}\right] + \max_{j \in [N]} \boldsymbol{\beta}_j^\top \left[\frac{2}{n_0}(\boldsymbol{X}^{(0)})^\top \boldsymbol{\epsilon}^{(0)}\right] \\
&\leq \sup_{\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{u} \in \mathscr{B}} (\boldsymbol{u} - \boldsymbol{u}_{j_0})^\top \boldsymbol{A} \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}) + \max_{j \in [N]} \boldsymbol{\beta}_j^\top \left[\frac{2}{n_0}(\boldsymbol{X}^{(0)})^\top \boldsymbol{\epsilon}^{(0)}\right] \\
&\leq \frac{1}{2}\|\boldsymbol{A} \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 + \max_{j \in [N]} \boldsymbol{\beta}_j^\top \left[\frac{2}{n_0}(\boldsymbol{X}^{(0)})^\top \boldsymbol{\epsilon}^{(0)}\right],
\end{aligned}$$

leading to

$$\|\boldsymbol{A}^\top \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 \leq 2 \max_{j\in[N]} \boldsymbol{\beta}_j^\top \left[\frac{2}{n_0}(\boldsymbol{X}^{(0)})^\top \boldsymbol{\epsilon}^{(0)}\right],$$

where $\boldsymbol{\beta}_{j_0} = \boldsymbol{A}\boldsymbol{u}_{j_0}$ with some $\|\boldsymbol{u}_{j_0}\|_2 \leq 1$ is the one which is closest to $\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{u}$ in the cover (depending on $\boldsymbol{\beta}$). Hence, $\|\boldsymbol{u}_{j_0}-\boldsymbol{u}\|_2 = \|\boldsymbol{\beta}_{j_0}-\boldsymbol{\beta}\|_2 \leq 1/2$. Showing $\max_{j\in[N]} \boldsymbol{\beta}_j^\top \left[\frac{2}{n_0}(\boldsymbol{X}^{(0)})^\top \boldsymbol{\epsilon}^{(0)}\right] \lesssim \sqrt{r/n_0}$ w.p. at least $1 - e^{-C(r+\log T)}$ is standard.

Finally, combining (47) and (48), we complete the proof.

D.12.12 PROOF OF LEMMA 40

The proof is exactly the same as the proof of Lemma 29, so omitted.

D.12.13 PROOF OF LEMMA 41

The proof is the same as the proof of Lemma 38, so we omit it here.

D.12.14 PROOF OF PROPOSITION 42

Define the random event $\mathcal{E}_1$ as the event that Lemmas 29, 30, 31, 32, and 34 hold, hence $\mathbb{P}(\mathcal{E}_1) \geq 1 - Ce^{-(r+\log T)}$ with some constant $C > 0$. Define another event

$$\mathcal{E}_2 = \left\{ \max_{t\in[T]} \|\widehat{\boldsymbol{\Sigma}}^{(t)} - \boldsymbol{\Sigma}^{(t)}\|_2 \lesssim \sqrt{\frac{p+\log T}{n}} \right\}.$$

By Lemma 18, we have $\mathbb{P}(\mathcal{E}_2) \geq 1 - Ce^{-(p+\log T)}$ with some constant $C > 0$. It suffices to prove the upper bounds in Proposition 42 conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2$, therefore we condition on $\mathcal{E}_1 \cap \mathcal{E}_2$ in the remaining proof without stating it explicitly and all the arguments are deterministic.

Define

$$G(\boldsymbol{A}) = \sum_{t=1}^T \min_{\boldsymbol{A}^{(t)}\in\mathcal{O}^{p\times r},\boldsymbol{\theta}^{(t)}\in\mathbb{R}^r} \left\{ \frac{1}{T}f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}) + \frac{\sqrt{n}}{nT}\lambda\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \boldsymbol{A}\boldsymbol{A}^\top\|_2 \right\},$$

$$G_S(\boldsymbol{A}) = \sum_{t\in S} \min_{\boldsymbol{A}^{(t)}\in\mathcal{O}^{p\times r},\boldsymbol{\theta}^{(t)}\in\mathbb{R}^r} \left\{ \frac{1}{T}f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}) + \frac{\sqrt{n}}{nT}\lambda\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \boldsymbol{A}\boldsymbol{A}^\top\|_2 \right\},$$

$$G_{S^c}(\boldsymbol{A}) = \sum_{t\in S^c} \min_{\boldsymbol{A}^{(t)}\in\mathcal{O}^{p\times r},\boldsymbol{\theta}^{(t)}\in\mathbb{R}^r} \left\{ \frac{1}{T}f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}) + \frac{\sqrt{n}}{nT}\lambda\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \boldsymbol{A}\boldsymbol{A}^\top\|_2 \right\},$$

$$\widehat{\overline{\boldsymbol{A}}} \in \arg\min_{\boldsymbol{A}\in\mathcal{O}^{p\times r}} G(\boldsymbol{A}).$$

Our first step is to prove the same bound for $\|\widehat{\overline{\boldsymbol{A}}}\widehat{\overline{\boldsymbol{A}}}^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2$. Let's prove it by contradiction. Denote the desired rate as

$$\eta = \bar{\zeta}^{-1}r\sqrt{\frac{p}{nT}} + \bar{\zeta}^{-1}\sqrt{r}\sqrt{\frac{r+\log T}{n}} + \sqrt{r}h + \frac{\lambda r}{\sqrt{n}\bar{\zeta}^2}\epsilon.$$

Suppose $\|\widehat{\overline{\boldsymbol{A}}}\widehat{\overline{\boldsymbol{A}}}^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2 > C\eta$ with some constant $C > 0$.

Consider any minimizer $(\boldsymbol{A}^{(t)}, \boldsymbol{\theta}^{(t)}) = \arg\min_{\boldsymbol{A} \in \mathcal{O}^{p \times r}, \boldsymbol{\theta} \in \mathbb{R}^r} \{\frac{1}{T} f^{(t)}(\boldsymbol{A}\boldsymbol{\theta}) + \frac{\sqrt{n}}{nT} \lambda \|\boldsymbol{A}\boldsymbol{A}^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2\}$. Define two index sets

$$\mathcal{A}_1 = \left\{ t \in S : \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 \geq c\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\overline{\boldsymbol{A}}^\top\|_2 \cdot \frac{1}{\sqrt{r}} \right\},$$

$$\mathcal{A}_2 = \left\{ t \in S : \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 < c\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\overline{\boldsymbol{A}}^\top\|_2 \cdot \frac{1}{\sqrt{r}} \right\},$$

where $c > 0$ is a small constant. Then we have

$$\sum_{t \in \mathcal{A}_1} \frac{1}{T}[f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}) - f^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})] + \sum_{t \in \mathcal{A}_1} \frac{\sqrt{n}}{nT}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$\geq -\frac{1}{T}\left|\left\langle \{\nabla f^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})\}_{t \in \mathcal{A}_1}, \{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t \in \mathcal{A}_1}\right\rangle\right|$$

$$+ \frac{1}{2}\frac{1}{T}\|\{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t \in \mathcal{A}_1}\|_F^2 + \sum_{t \in \mathcal{A}_1} \frac{1}{\sqrt{n}T}\lambda\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$\geq -\frac{1}{T}\left|\left\langle \{\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\}_{t \in \mathcal{A}_1}, \{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} + \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t \in \mathcal{A}_1}\right\rangle\right|$$

$$- \frac{1}{T}\|\{\nabla f^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}) - \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\}\|_F \|\{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t \in \mathcal{A}_1}\|_F$$

$$+ \frac{1}{2}\frac{1}{T}\|\{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t \in \mathcal{A}_1}\|_F^2 + \sum_{t \in \mathcal{A}_1} \frac{1}{\sqrt{n}T}\lambda\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$\geq -\frac{1}{T}\sum_{t \in \mathcal{A}_1}\|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 \cdot (\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\|_2 + \|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2)$$

$$- \frac{1}{T}C\left[\sqrt{T}\sqrt{\frac{r + \log T}{n}} + h\sqrt{\sum_{t \in \mathcal{A}_1}(\zeta^{(t)})^2}\right] \cdot \|\{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t \in \mathcal{A}_1}\|_F$$

$$+ \frac{1}{2}\frac{1}{T}\|\{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t \in \mathcal{A}_1}\|_F^2 + \sum_{t \in \mathcal{A}_1} \frac{1}{\sqrt{n}T}\lambda\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 \qquad (49)$$

$$\geq -\frac{C}{T}\sum_{t \in \mathcal{A}_1}\sqrt{\frac{p + \log T}{n}} \cdot \zeta^{(t)}(\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\overline{\boldsymbol{A}}^\top\|_2)$$

$$- \frac{1}{T}C\left[\sqrt{T}\sqrt{\frac{r + \log T}{n}} + h\sqrt{\sum_{t \in \mathcal{A}_1}(\zeta^{(t)})^2}\right] \cdot \|\{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t \in \mathcal{A}_1}\|_F$$

$$+ \frac{1}{2}\frac{1}{T}\|\{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t \in \mathcal{A}_1}\|_F^2 + \sum_{t \in \mathcal{A}_1} \frac{1}{\sqrt{n}T}\lambda\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 \qquad (50)$$

$$\geq -C\left(h^2 \cdot \frac{1}{T}\sum_{t \in \mathcal{A}_1}(\zeta^{(t)})^2 + \frac{r + \log T}{n}\right) + \frac{|\mathcal{A}_1|}{2\sqrt{n}T}\lambda \cdot \frac{1}{\sqrt{r}}\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\overline{\boldsymbol{A}}^\top\|_2, \qquad (51)$$

where (49) holds because $\|\{\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\}_{t \in \mathcal{A}_1}\|_2 \leq \sqrt{\frac{p}{n}} + h\sqrt{\sum_{t \in \mathcal{A}_1}(\zeta^{(t)})^2} + \sqrt{T}\sqrt{\frac{r + \log T}{n}}$ (Lemma 32) and $\|\{\nabla f^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}) - \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\}\|_F =$

$$\sqrt{\sum_{t\in\mathcal{A}_1}\|\widehat{\boldsymbol{\Sigma}}^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}-\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2^2}\lesssim\sqrt{\sum_{t\in\mathcal{A}_1}\|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}-\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2}\lesssim\sqrt{h^2\sum_{t\in\mathcal{A}_1}(\zeta^{(t)})^2+T\cdot\frac{r+\log T}{n}}$$

(Lemma 29). Inequality (50) holds because

$$\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}\|_2\lesssim\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top-\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2(\|\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\|_2+\|\nabla f^{(t)}(\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)})\|_2)$$

$$\lesssim\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top-\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2\Big(\zeta^{(t)}+\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}-\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2+\|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2\Big)$$

$$\lesssim\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top-\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2\zeta^{(t)},$$

where we used Lemmas 29, 30, and 31. Inequality (51) holds because

$$\frac{1}{\sqrt{nT}}\lambda\sum_{t\in\mathcal{A}_1}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top-\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$-\frac{1}{T}\sum_{t\in\mathcal{A}_1}\sqrt{\frac{p+\log T}{n}}\cdot\zeta^{(t)}(\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top-\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2+\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top-\overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2)$$

$$\geq\frac{1}{\sqrt{nT}}\lambda\sum_{t\in\mathcal{A}_1}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top-\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2-\frac{1}{T}\sum_{t\in\mathcal{A}_1}C\sqrt{\frac{r(p+\log T)}{n}}\cdot\zeta^{(t)}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top-\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$\geq\frac{1}{2\sqrt{nT}}\lambda\sum_{t\in\mathcal{A}_1}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top-\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$\geq\frac{|\mathcal{A}_1|}{2\sqrt{nT}}\lambda\cdot\frac{1}{\sqrt{r}}\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top-\overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2,$$

when $\frac{\lambda}{\sqrt{n}}\geq 2C\max_{t\in S}\zeta^{(t)}\cdot\sqrt{\frac{r(p+\log T)}{n}}$.

For any $t\in\mathcal{A}_2$, since $\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}_{\boldsymbol{A}^{(t)}}^{(t)}-\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\|_2\leq\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top-\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2(\|\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\|_2+\|\nabla f^{(t)}(\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)})\|_2)\leq$
$c\zeta^{(t)}\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top-\overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2\frac{1}{\sqrt{r}}$, by triangle inequality, we have

$$\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}_{\boldsymbol{A}^{(t)}}^{(t)}-\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2\geq\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}-\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2-\frac{c}{\sqrt{r}}\zeta^{(t)}\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top-\overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2.$$

With similar arguments to obtain (51), here we have

$$\sum_{t\in\mathcal{A}_2}\frac{1}{T}[f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)})-f^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})]+\sum_{t\in\mathcal{A}_2}\frac{\sqrt{n}}{nT}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top-\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$\geq\frac{1}{2T}\|\{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}-\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\|_{\mathrm{F}}^2-\frac{1}{T}\sqrt{\frac{p+T}{n}}\cdot\sqrt{2r}\left\|\{\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}-\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\right\|_{\mathrm{F}}$$

$$-\frac{C'}{T}\left(\sqrt{\frac{pr}{n}}+\sqrt{T}\sqrt{\frac{r+\log T}{n}}+h\sqrt{\sum_{t\in\mathcal{A}_2}(\zeta^{(t)})^2}\right)\cdot\|\{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}-\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\|_{\mathrm{F}}$$

$$+\frac{1}{T}\sum_{t\in\mathcal{A}_2}\frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top-\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2-\frac{C}{T}\sum_{t\in\mathcal{A}_2}\sqrt{\frac{p+\log T}{n}}\cdot\zeta^{(t)}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top-\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2$$

$$\tag{52}$$

$$\geq -C\left(\frac{pr}{nT} + h^2 \cdot \frac{1}{T}\sum_{t\in\mathcal{A}_2}(\zeta^{(t)})^2 + \frac{r+\log T}{n}\right) + \frac{C}{T}\|\{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\|_{\mathrm{F}}^2$$

$$- \frac{1}{T}\sqrt{\frac{p+T}{n}} \cdot \sqrt{2r}\left\|\{\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\right\|_{\mathrm{F}}$$

$$= -C\left(\frac{pr}{nT} + h^2 \cdot \frac{1}{T}\sum_{t\in\mathcal{A}_2}(\zeta^{(t)})^2 + \frac{r+\log T}{n}\right) + \frac{C}{T}\sum_{t\in\mathcal{A}_2}\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2^2$$

$$- \frac{1}{T}\sqrt{\frac{p+T}{n}} \cdot \sqrt{2r}\left\|\{\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\right\|_{\mathrm{F}} \tag{53}$$

$$\geq -C\left(\frac{pr}{nT} + h^2 \cdot \frac{1}{T}\sum_{t\in\mathcal{A}_2}(\zeta^{(t)})^2 + \frac{r+\log T}{n}\right) + \frac{C'}{T}\sum_{t\in\mathcal{A}_2}\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2^2$$

$$- \frac{1}{T}\sqrt{\frac{p+T}{n}} \cdot \sqrt{2r}\left\|\{\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\right\|_{\mathrm{F}} - \frac{c}{Tr}\sum_{t\in\mathcal{A}_2}(\zeta^{(t)})^2 \cdot \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2^2$$

$$\geq -C\left(\frac{pr}{nT} + h^2 \cdot \frac{1}{T}\sum_{t\in\mathcal{A}_2}(\zeta^{(t)})^2 + \frac{r+\log T}{n}\right) + \frac{C'}{2T}\sum_{t\in\mathcal{A}_2}\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2^2$$

$$+ \frac{C'}{2T}\left\|\{\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\right\|_{\mathrm{F}}^2 - \frac{1}{T}\sqrt{\frac{p+T}{n}} \cdot \sqrt{2r}\left\|\{\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\right\|_{\mathrm{F}}$$

$$- \frac{c}{Tr}\sum_{t\in\mathcal{A}_2}(\zeta^{(t)})^2 \cdot \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2^2.$$

$$\geq -C\left(\frac{pr}{nT} + h^2 \cdot \frac{1}{T}\sum_{t\in\mathcal{A}_2}(\zeta^{(t)})^2 + \frac{r+\log T}{n}\right) + \frac{C'}{2T}\sum_{t\in\mathcal{A}_2}\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2^2$$

$$- \frac{c}{Tr}\sum_{t\in\mathcal{A}_2}(\zeta^{(t)})^2 \cdot \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2^2. \tag{54}$$

Compared to the derivation of (51), here we used a different way to bound the intermediate term $\frac{1}{T}|\langle\{\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\}_{t\in\mathcal{A}_2}, \{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} + \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\rangle|$ to obtain (52). More specifically,

$$\frac{1}{T}\left|\left\langle\{\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\}_{t\in\mathcal{A}_2}, \{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} + \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\right\rangle\right|$$

$$\leq \frac{1}{T}\sum_{t\in\mathcal{A}_2}\|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\|_2 + \frac{1}{T}\|\{\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\}_{t\in\mathcal{A}_2}\|_2\|\{\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\|_*$$

$$\leq \frac{1}{T}\sum_{t\in\mathcal{A}_2}\sqrt{\frac{p+\log T}{n}}\zeta^{(t)}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + \frac{1}{T}\sqrt{\frac{p+T}{n}} \cdot \sqrt{2r}\left\|\{\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\right\|_{\mathrm{F}},$$

where we used the fact that $|\langle \boldsymbol{A}, \boldsymbol{B}\rangle| \leq \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_* \leq \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_F \cdot \text{rank}(\boldsymbol{B})$ for any matrix $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same dimension and $\text{rank}\big(\{\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\big) \leq 2r$.

Finally, notice that by triangle inequality,

$$G_{S^c}(\widehat{\overline{\boldsymbol{A}}}) - G_{S^c}(\overline{\boldsymbol{A}}) \geq \sum_{t\in S^c} \min_{\boldsymbol{A}^{(t)}\in\mathcal{O}^{p\times r}, \boldsymbol{\theta}^{(t)}\in\mathbb{R}^r} \left\{ \frac{\sqrt{n}}{nT}\lambda(\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 - \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2) \right\}$$

$$\geq -\frac{\sqrt{n}}{nT}\lambda|S^c| \cdot \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2. \tag{55}$$

Case 1: $|\mathcal{A}_2| \geq (1 - c/r \cdot \bar\zeta^2/\max_{t\in S}(\zeta^{(t)})^2)|S|$ with some small constant $c > 0$:

Then by (51), (54), and (55),

$$G(\widehat{\overline{\boldsymbol{A}}}) - G(\overline{\boldsymbol{A}}) = G_S(\widehat{\overline{\boldsymbol{A}}}) - G_S(\overline{\boldsymbol{A}}) + G_{S^c}(\widehat{\overline{\boldsymbol{A}}}) - G_{S^c}(\overline{\boldsymbol{A}})$$

$$\geq -C\left(\frac{pr}{nT} + h^2\bar\zeta^2 + \frac{r + \log T}{n}\right) + \frac{C'}{T}\sum_{t\in\mathcal{A}_2}\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2^2$$

$$-\frac{c}{Tr}\sum_{t\in\mathcal{A}_2}(\zeta^{(t)})^2 \cdot \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2^2$$

$$-\frac{\sqrt{n}}{nT}\lambda|S^c| \cdot \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2. \tag{56}$$

By Assumption 2, Wedin's $\sin\Theta$-Theorem, and applying Lemma 35 on $\{\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}$,

$$\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \leq \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_F \lesssim \sqrt{\frac{r}{|S|}}\bar\zeta^{-1} \cdot \|\{\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in\mathcal{A}_2}\|_F. \tag{57}$$

Here we used the fact that $\frac{1}{|S|}\sum_{t\in S}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}(\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})^\top \succeq \frac{c}{r}\bar\zeta^2\boldsymbol{I}_p$. To see this, notice that

$$\sigma_r\left(\frac{1}{\sqrt{|S|}}\{\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in S}\right) = \sigma_r\left(\frac{1}{\sqrt{|S|}}\{\boldsymbol{\beta}^{(t)*}\}_{t\in S}\right) - \frac{1}{\sqrt{|S|}}\|\{\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in S} - \{\boldsymbol{\beta}^{(t)*}\}_{t\in S}\|_2$$

$$\geq \frac{c}{\sqrt{r}}\bar\zeta - \left(\bar\zeta h + \sqrt{\frac{r + \log T}{n}}\right)$$

$$\geq \frac{c'}{\sqrt{r}}\bar\zeta,$$

where by Lemmas 29 and 33,

$$\frac{1}{\sqrt{|S|}}\|\{\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t\in S} - \{\boldsymbol{\beta}^{(t)*}\}_{t\in S}\|_2 \leq \sqrt{\sum_{t\in S}\|\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2} \lesssim h\bar\zeta + \sqrt{\frac{r + \log T}{n}},$$

and the last step comes from conditions (i) and (ii) by noticing that

$$\sqrt{\frac{r + \log T}{n}} \lesssim \frac{\bar\zeta}{\sqrt{r}} \cdot \sqrt{\frac{p + \log T}{n}} \cdot \frac{1}{\zeta^{(t)}} \lesssim \frac{\bar\zeta}{\sqrt{r}}.$$

103

This leads to

$$\lambda_r \left(\frac{1}{|S|} \sum_{t \in S} \boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} (\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})^\top \right) = \lambda_r \left(\frac{1}{|S|} \sum_{t \in S} \boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} (\overline{\boldsymbol{A}})^\top \overline{\boldsymbol{A}} (\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})^\top \right) \geq \frac{c'}{r} \bar{\zeta}^2.$$

Plugging (57) into (56), we have

$$G(\widehat{\overline{\boldsymbol{A}}}) - G(\overline{\boldsymbol{A}}) \geq -C \left(h^2 \bar{\zeta}^2 + \frac{r + \log T}{n} \right) + \frac{C'}{T} \cdot \frac{|S|}{r} \bar{\zeta}^2 \| \widehat{\overline{\boldsymbol{A}}} (\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}} \overline{\boldsymbol{A}}^\top \|_2^2$$

$$- \frac{c|S|}{Tr} \bar{\zeta}^2 \| \widehat{\overline{\boldsymbol{A}}} (\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}} \overline{\boldsymbol{A}}^\top \|_2^2 - C \sqrt{\frac{r(p+T)}{nT}} \cdot \| \widehat{\overline{\boldsymbol{A}}} (\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}} \overline{\boldsymbol{A}}^\top \|_2 \sqrt{\frac{1}{T} \sum_{t \in \mathcal{A}_2} (\zeta^{(t)})^2}$$

$$- \frac{\sqrt{n}}{nT} \lambda |S^c| \cdot \| \widehat{\overline{\boldsymbol{A}}} (\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}} \overline{\boldsymbol{A}}^\top \|_2$$

$$\geq \frac{C}{T} \cdot \frac{|S|}{r} \bar{\zeta}^2 \| \widehat{\overline{\boldsymbol{A}}} (\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}} (\overline{\boldsymbol{A}})^\top \|_2^2 - \frac{\sqrt{n}}{nT} \lambda |S^c| \cdot \| \widehat{\overline{\boldsymbol{A}}} (\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}} (\overline{\boldsymbol{A}})^\top \|_2$$

$$- C \left(\frac{pr}{nT} + h^2 \bar{\zeta}^2 + \frac{r + \log T}{n} \right) - C \sqrt{\frac{r(p+T)}{nT}} \bar{\zeta}.$$

$$> 0,$$

because $\| \widehat{\overline{\boldsymbol{A}}} (\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}} \overline{\boldsymbol{A}}^\top \|_2 \geq C\eta$ with some constant $C > 0$.

Case 2: $|\mathcal{A}_1| \geq c/r \cdot \bar{\zeta}^2 / \max_{t \in S} (\zeta^{(t)})^2 \cdot |S|$ with some small constant $c > 0$:

By inequality (53) and triangle inequality, we have

$$\sum_{t \in \mathcal{A}_2} \frac{1}{T} [f^{(t)}(\boldsymbol{A}^{(t)} \boldsymbol{\theta}^{(t)}) - f^{(t)}(\overline{\boldsymbol{A}} \boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})] + \sum_{t \in \mathcal{A}_2} \frac{\sqrt{n}}{nT} \| \boldsymbol{A}^{(t)} (\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}} (\widehat{\overline{\boldsymbol{A}}})^\top \|_2$$

$$= -C \left(\frac{pr}{nT} + h^2 \cdot \frac{1}{T} \sum_{t \in \mathcal{A}_2} (\zeta^{(t)})^2 + \frac{r + \log T}{n} \right) + \frac{C}{T} \sum_{t \in \mathcal{A}_2} \| \boldsymbol{A}^{(t)} \boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}} \boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} \|_2^2$$

$$- \frac{1}{T} \sqrt{\frac{p+T}{n}} \cdot \sqrt{2r} \left\| \{ \widehat{\overline{\boldsymbol{A}}} \boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}} \boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} \}_{t \in \mathcal{A}_2} \right\|_F$$

$$\geq -C \left(\frac{pr}{nT} + h^2 \cdot \frac{1}{T} \sum_{t \in \mathcal{A}_2} (\zeta^{(t)})^2 + \frac{r + \log T}{n} \right) + \frac{C}{T} \sum_{t \in \mathcal{A}_2} \| \boldsymbol{A}^{(t)} \boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}} \boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} \|_2^2$$

$$- \frac{1}{T} \sqrt{\frac{p+T}{n}} \cdot \sqrt{2r} \left\| \{ \boldsymbol{A}^{(t)} \boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}} \boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} \}_{t \in \mathcal{A}_2} \right\|_F - \frac{1}{T} \sqrt{\frac{p+T}{n}} \cdot \sqrt{2r} \left\| \{ \boldsymbol{A}^{(t)} \boldsymbol{\theta}^{(t)} - \widehat{\overline{\boldsymbol{A}}} \boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} \}_{t \in \mathcal{A}_2} \right\|_F$$

$$\geq -C \left(\frac{pr}{nT} + h^2 \cdot \frac{1}{T} \sum_{t \in \mathcal{A}_2} (\zeta^{(t)})^2 + \frac{r + \log T}{n} \right) + \frac{C}{T} \left\| \{ \boldsymbol{A}^{(t)} \boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}} \boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} \}_{t \in \mathcal{A}_2} \right\|_F^2$$

$$- \frac{1}{T} \sqrt{\frac{p+T}{n}} \cdot \sqrt{2r} \left\| \{ \boldsymbol{A}^{(t)} \boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}} \boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)} \}_{t \in \mathcal{A}_2} \right\|_F$$

$$
-\sqrt{\frac{p+T}{nT}} \cdot \sqrt{2r} \cdot \sqrt{\frac{1}{T} \sum_{t \in \mathcal{A}_2} \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2^2 (\zeta^{(t)})^2}
$$

$$
\geq -C \left(\frac{pr}{nT} + h^2 \cdot \frac{1}{T} \sum_{t \in \mathcal{A}_2} (\zeta^{(t)})^2 + \frac{r + \log T}{n} \right) + \frac{C}{T} \left\| \{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t \in \mathcal{A}_2} \right\|_F^2
$$

$$
- \frac{1}{T}\sqrt{\frac{p+T}{n}} \cdot \sqrt{2r} \left\| \{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t \in \mathcal{A}_2} \right\|_F - \sqrt{\frac{p+T}{nT}} \cdot \bar{\zeta} \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2
$$

$$
\geq -C \left(\frac{pr}{nT} + h^2 \bar{\zeta}^2 + \frac{r + \log T}{n} \right) - \sqrt{\frac{p+T}{nT}} \cdot \bar{\zeta} \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2 \tag{58}
$$

By the condition of the proposition, we must have

$$
\frac{\min_{t \in S} \zeta^{(t)}}{\bar{\zeta}} \cdot r \lesssim \sqrt{p + \log T}, \quad \frac{\min_{t \in S} \zeta^{(t)}}{\bar{\zeta}} \cdot \frac{r}{\sqrt{T}} \lesssim 1. \tag{59}
$$

By (51), (55), and (58), we have

$$
G(\widehat{\overline{\boldsymbol{A}}}) - G(\overline{\boldsymbol{A}})
$$

$$
\geq \frac{C\lambda}{\sqrt{nT}} \left(\frac{|\mathcal{A}_1|}{\sqrt{r}} - |S^c| \right) \cdot \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2 - \sqrt{\frac{p+T}{nT}} \cdot \bar{\zeta} \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2
$$

$$
- C \left(\frac{pr}{nT} + h^2 \cdot \bar{\zeta}^2 + \frac{r + \log T}{n} \right)
$$

$$
\geq \frac{C\lambda}{\sqrt{nT}} \left(\frac{|S|}{r^{3/2}} \cdot \frac{\bar{\zeta}^2}{\max_{t \in S}(\zeta^{(t)})^2} - |S^c| \right) \cdot \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2 - \sqrt{\frac{p+T}{nT}} \cdot \bar{\zeta} \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2
$$

$$
- C \left(\frac{pr}{nT} + h^2 \cdot \bar{\zeta}^2 + \frac{r + \log T}{n} \right)
$$

$$
\geq \frac{C\lambda}{\sqrt{n}} r^{-3/2} \cdot \frac{\bar{\zeta}^2}{\max_{t \in S}(\zeta^{(t)})^2} \cdot \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2 - \sqrt{\frac{p+T}{nT}} \cdot \bar{\zeta} \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2
$$

$$
- C \left(\frac{pr}{nT} + h^2 \cdot \bar{\zeta}^2 + \frac{r + \log T}{n} \right)
$$

$$
\geq \frac{C\lambda}{2\sqrt{n}} r^{-3/2} \cdot \frac{\bar{\zeta}^2}{\max_{t \in S}(\zeta^{(t)})^2} \cdot \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2 - C \left(\frac{pr}{nT} + h^2 \cdot \bar{\zeta}^2 + \frac{r + \log T}{n} \right)
$$

due to the assumption that $|S^c|/T \leq cr^{-3/2}\bar{\zeta}/\max_{t \in S} \zeta^{(t)}$ with a small constant $c > 0$ and

$$
\lambda \asymp \sqrt{r(p + \log T)} \cdot \frac{\max_{t \in S}(\zeta^{(t)})^2}{\min_{t \in S} \zeta^{(t)}}
$$

$$
\gtrsim \left[\sqrt{pr} \frac{\bar{\zeta}}{\min_{t \in S} \zeta^{(t)}} + \sqrt{r(p + \log T)} \cdot \frac{\bar{\zeta}}{\min_{t \in S} \zeta^{(t)}} \right] \cdot \frac{\max_{t \in S}(\zeta^{(t)})^2}{\bar{\zeta}},
$$

where we used (59) in the second inequality. Furthermore, by $\min_{t \in S} \zeta^{(t)} \gtrsim \sqrt{\frac{p + \log T}{n}}$, we have

$$
G(\widehat{\overline{\boldsymbol{A}}}) - G(\overline{\boldsymbol{A}})
$$

$$\geq C''\sqrt{\frac{p+\log T}{n}}r^{-1} \cdot \frac{\bar{\zeta}^2}{\min_{t\in S}\zeta^{(t)}} \cdot \|\widehat{\overline{A}}(\widehat{\overline{A}})^\top - \overline{A}\,\overline{A}^\top\|_2 - C\left(\frac{pr}{nT} + h^2\cdot\bar{\zeta}^2 + \frac{r+\log T}{n}\right)$$

$$\geq C''r^{-1}\cdot\bar{\zeta}^2\cdot\|\widehat{\overline{A}}(\widehat{\overline{A}})^\top - \overline{A}\,\overline{A}^\top\|_2 - C\left(\frac{pr}{nT} + h^2\cdot\bar{\zeta}^2 + \frac{r+\log T}{n}\right).$$

Therefore, if $\|\widehat{\overline{A}}(\widehat{\overline{A}})^\top - \overline{A}\,\overline{A}^\top\|_2 \geq C\eta$, we must have

$$G(\widehat{\overline{A}}) - G(\overline{A}) > 0,$$

which implies that $\widehat{\overline{A}}$ is not a minimizer of G and contradicts with the optimality of $\widehat{\overline{A}}$. Hence we must have $\|\widehat{\overline{A}}(\widehat{\overline{A}})^\top - \overline{A}\,\overline{A}^\top\|_2 \leq C\eta \asymp \bar{\zeta}^{-1}r\sqrt{\frac{p}{nT}} + \bar{\zeta}^{-1}\sqrt{r}\sqrt{\frac{r+\log T}{n}} + \sqrt{r}h + \frac{\lambda r}{\sqrt{n}\bar{\zeta}^2}\epsilon$ with some constant $C > 0$. This completes our proof.

Appendix E. Proofs for GLMs

E.1 Lemmas

Lemma 43 *Suppose Assumptions 1 and 4 hold, and $n \geq C(p+\log T)$ with a sufficiently large C. Then when $\delta \leq C$ with some constant $C > 0$,*

$$\|\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})\|_2 \lesssim \delta,$$

w.p. at least $1 - \exp\{-Cn\delta^2 + C'p\}$. By taking $\delta \asymp \sqrt{\frac{p+\log T}{n}}$, this implies that:

(i) $\max_{t\in S}\|\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})\|_2 \lesssim \sqrt{\frac{p+\log T}{n}}$, *w.p. at least $1 - e^{-C(p+\log T)}$;*

(ii) $\|\nabla f^{(0)}(\boldsymbol{\beta}^{(0)*})\|_2 \lesssim \sqrt{\frac{p}{n_0}}$, *w.p. at least $1 - e^{-C(p+\log T)}$.*

Lemma 44 *Suppose Assumptions 1 and 4 hold, and $n \geq C(p+\log T)$ with a sufficiently large C. Then when $t \leq C'$ with some constant $C' > 0$, for any $S \subseteq [T]$,*

$$\|\{\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})\}_{t\in S}\|_2 \lesssim \sqrt{\frac{p+|S|+\delta}{n}},$$

w.p. at least $1 - e^{-C'\delta}$.

Lemma 45 (A revised version of Proposition 1 in Loh and Wainwright (2015))
Suppose Assumptions 1 and 4 hold, and $n \geq C(p+\log T)$ with a sufficiently large C. Then there exist constants C_1, C_2, such that w.p. at least $1 - e^{-C'(p+\log T)}$, for all $\boldsymbol{\beta} \in \mathbb{R}^p$, $t \in [T]$, we have

$$f^{(t)}(\boldsymbol{\beta}) - f^{(t)}(\boldsymbol{\beta}^{(t)*}) - \nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)*}) \geq \begin{cases} C_1\|\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)*}\|_2^2, & \text{when } \|\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)*}\|_2 \leq 1; \\ C_2\|\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)*}\|_2, & \text{when } \|\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)*}\|_2 > 1. \end{cases}$$
$$(60)$$

Lemma 46 *Suppose Assumptions 1 and 4 hold, and $n \geq C(p + \log T)$ with a sufficiently large C. Then there exist constants C_1, C_2, C_3, C_4, such that for all $\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}} \in \mathbb{R}^p$, $t \in [T]$, w.p. at least $1 - e^{-C(p + \log T)}$,*

$$f^{(t)}(\boldsymbol{\beta}) - f^{(t)}(\widetilde{\boldsymbol{\beta}}) - \nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})^\top (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) \geq \begin{cases} C_1 \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_2^2 - C_2 \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(t)*}\|_2^2, & \text{if } \|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*}\|_2 \leq 1; \\ C_3 \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_2 - C_4 \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(t)*}\|_2^2, & \text{if } \|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*}\|_2 > 1. \end{cases}$$

Lemma 47 *Suppose Assumptions 1 and 4 hold, and $n \geq C(p + \log T)$ with a sufficiently large C. Then w.p. at least $1 - e^{-C'(p + \log T)}$, for all $t \in S$, for all $\boldsymbol{A} \in \mathcal{O}^{p \times r}$ with $\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top\|_2 \leq c(\zeta^{(t)})^{-1}$, where $c > 0$ is a small constant, we have*

(i) $\|\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}\|_2 \leq C'' \zeta^{(t)}$;

(ii) $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq c'$ *with a small constant $c' > 0$.*

Lemma 48 *Suppose Assumptions 1 and 4 hold, and $n \geq C(p + \log T)$ with a sufficiently large C. Assume there exists $\widetilde{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$ such that $\|\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq c$ with a small constant c. Then w.p. at least $1 - e^{-C'(p + \log T)}$, for all $t \in S$, for all $\boldsymbol{A} \in \mathcal{O}^{p \times r}$ with $\|\boldsymbol{A}\boldsymbol{A}^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2 \leq c'(\zeta^{(t)})^{-1}$, where $c' > 0$ is a small constant, we have*

(i) $\|\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}\|_2 \leq C'' \zeta^{(t)}$;

(ii) $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq c''$ *with a small constant $c'' > 0$.*

Lemma 49 *Suppose Assumptions 1 and 4 hold, and $n \geq C(p + \log T)$ with a sufficiently large C. Then w.p. at least $1 - e^{-C'(p + \log T)}$, for all $t \in S$, for all $\boldsymbol{A}, \widetilde{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$ with $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2, \|\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq c'$, where $c' > 0$ is a small constant, we have*

$$\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}\|_2 \lesssim \|\boldsymbol{A}\boldsymbol{A}^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2 (\|\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}\|_2 + \|\nabla f^{(t)}(\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)})\|_2).$$

Lemma 50 *Suppose Assumptions 1 and 4 hold, and $n \geq C(p + \log T)$ with a sufficiently large C. Then w.p. at least $1 - e^{-C'(p + \log T)}$, for all $t \in S$, for all $\boldsymbol{A}, \widetilde{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$ with $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq c'$, where $c' > 0$ is a small constant, we have*

$$\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \lesssim \|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top\|_2 \zeta^{(t)} + \sqrt{\frac{r + \log T}{n}}.$$

Lemma 51 *Suppose Assumptions 1 and 4 hold, and $n \geq C[(p + \log T) \vee \lambda^2]$ with a sufficiently large C. Given $\widetilde{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$, denote $(\boldsymbol{A}^{(t)}, \boldsymbol{\theta}^{(t)}) = \arg\min_{\boldsymbol{A} \in \mathcal{O}^{p \times r}, \boldsymbol{\theta} \in \mathbb{R}^r} \{f^{(t)}(\boldsymbol{A}\boldsymbol{\theta}) + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}\boldsymbol{A}^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2\}$. Then w.p. at least $1 - e^{-C(p + \log T)}$, for any $\widetilde{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$, we have $\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \leq c$, where $c > 0$ is a small constant.*

E.2 Proof of Theorem 11

For the penalized ERM, the proof logic is almost the same as the logic underlying the proofs of Proposition 42 and Theorem 2. We only point out the difference and skip the details here. We need to bound $\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\boldsymbol{A}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2$ first. Then we know that $\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\boldsymbol{A}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2$ and $\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2$ are small w.h.p. Then a direct application of Lemma 48 implies that $\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)*}_{\widehat{\overline{\boldsymbol{A}}}} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2$ is small w.h.p. Finally, applying Lemma 49, we have $\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)*}_{\widehat{\overline{\boldsymbol{A}}}} - \overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)}_{\overline{\boldsymbol{A}}}\|_2 \lesssim \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\boldsymbol{A}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 (\zeta^{(t)} + \|\nabla f^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)}_{\overline{\boldsymbol{A}}})\|_2) \lesssim \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\boldsymbol{A}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \zeta^{(t)}$ w.h.p. which gives us the ideal bound for $\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)*}_{\widehat{\overline{\boldsymbol{A}}}} - \overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)}_{\overline{\boldsymbol{A}}}\|_2$ for all $t \in S$.

(i) In the proof of Proposition 42, for $t \in \mathcal{A}_1$, we know $\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq$ a small constant c w.h.p. by Lemma 51. Then by Lemmas 44, 46, and 50, w.p. at least $1 - e^{-C'(r+\log T)}$,

$$\frac{1}{T}\sum_{t \in \mathcal{A}_1}\left[f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}) - f^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)}_{\overline{\boldsymbol{A}}}) \right]$$

$$\geq -\frac{1}{T}\left| \left\langle \{\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\}_{t \in \mathcal{A}_1}, \{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}} + \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}} - \overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)}_{\overline{\boldsymbol{A}}}\}_{t \in \mathcal{A}_1} \right\rangle \right|$$

$$+ \frac{C_1}{T}\|\{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)}_{\overline{\boldsymbol{A}}}\}_{t \in \mathcal{A}_1}\|_F^2 - \frac{C_2}{T}\|\{\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*} - \overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)}_{\overline{\boldsymbol{A}}}\}_{t \in \mathcal{A}_1}\|_F^2$$

$$\geq -\frac{C}{T}\sum_{t \in \mathcal{A}_1}\sqrt{\frac{p+\log T}{n}}\cdot\zeta^{(t)}(\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\boldsymbol{A}})^\top\|_2 + \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\boldsymbol{A}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2)$$

$$+ \frac{C_1}{T}\|\{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)}_{\overline{\boldsymbol{A}}}\}_{t \in \mathcal{A}_1}\|_F^2 - C\left(h^2 \cdot \frac{1}{T}\sum_{t \in \mathcal{A}_1}(\zeta^{(t)})^2 + \frac{r+\log T}{n} \right)$$

$$\geq -C\left(h^2 \cdot \frac{1}{T}\sum_{t \in \mathcal{A}_1}(\zeta^{(t)})^2 + \frac{r+\log T}{n} \right) - \frac{C'}{T}\sum_{t \in \mathcal{A}_1}\sqrt{\frac{p+\log T}{n}}\cdot\zeta^{(t)}\sqrt{r}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\boldsymbol{A}})^\top\|_2.$$

Therefore, the calculations in (51) are still correct under GLMs.

(ii) In the proof of Proposition 42, for $t \in \mathcal{A}_2$, we used the fact that w.p. at least $1 - e^{-C(r+\log T)}$,

$$\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}_{\boldsymbol{A}^{(t)}} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}}\|_2 \lesssim \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\boldsymbol{A}})^\top\|_2(\|\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}}\|_2 + \|\nabla f^{(t)}(\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}})\|_2). \tag{61}$$

Here, this still holds because

- $\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}_{\boldsymbol{A}^{(t)}} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq$ a small constant c by Lemma 51;

- For $t \in \mathcal{A}_2$, we know that $\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\boldsymbol{A}})^\top\|_2 \leq \frac{c}{\sqrt{r}} \leq c'(\zeta^{(t)})^{-1}$ with a small constant $c' > 0$. Then by Lemma 48, $\|\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}}\|_2 \leq C\zeta^{(t)}$ and $\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq Cc'$ with a small constant $c' > 0$.

Then (61) holds by Lemma 49.

Finally, Lemma 38 still holds for the case of GLMs. Note that in Lemma 38, (i) only requires the maximum eigenvalue of the Hessian to be upper bounded (see Lemma E.3 in Duan and Wang (2023)), which is true by Assumptions 1 and 4. By Lemma E.2 in Duan and Wang (2023), (ii) and (iii) require the minimum eigenvalue of Hessian to be lower bounded when $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \leq C$ with some constant $C > 0$ w.h.p. and $\gamma/\sqrt{n_0} \leq C$, both of which are true.

For the spectral method, the proof is almost the same as the proof of Theorem 7, hence we do not repeat it here.

E.3 Proof of Theorem 16

We only outline the proof when Algorithm 4 is coupled with Algorithm 1. If Algorithm 4 is coupled with Algorithm 2, we can follow a similar argument with the proof of Theorem 7.

Denote $\eta = r\sqrt{\frac{p}{nT}} + \sqrt{r}h + \sqrt{r}\sqrt{\frac{r+\log T}{n}} + \frac{|S^c|}{T} \cdot r \cdot \frac{\lambda}{\sqrt{n}}$.

(i) When $\eta \leq C\sqrt{\frac{p}{n_0}}$: since $\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \boldsymbol{A}^{(0)*}(\boldsymbol{A}^{(0)*})^\top\|_2 \leq$ a small constant c, by Lemma 51, w.p. at least $1 - e^{-C(p+\log T)}$, $\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}} - \boldsymbol{A}^{(0)*}\boldsymbol{\theta}^{(0)*}\|_2 \leq$ a small constant c'. Then by Lemma 50 and Theorem 11, w.p. at least $1 - e^{-C(r+\log T)}$, $\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}^{(t)}_{\widehat{\overline{\boldsymbol{A}}}} - \boldsymbol{A}^{(0)*}\boldsymbol{\theta}^{(0)*}\|_2 \lesssim \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \boldsymbol{A}^{(0)*}(\boldsymbol{A}^{(0)*})^\top\|_2 \lesssim \eta$. As we commented in the proof of Theorem 11, Lemma 41.(i) still holds for GLMs because the maximum eigenvalue of the Hessian is upper bounded w.h.p.

(ii) When $\eta > C\sqrt{\frac{p}{n_0}}$: the result comes from Lemma 41.(ii), which still holds when $\gamma/\sqrt{n_0} \lesssim$ some constant C because of Lemma E.2 in Duan and Wang (2023).

E.4 Proofs of Lemmas

E.4.1 Proof of Lemma 43

First, note that $\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*}) = \frac{2}{n}\sum_{i=1}^{n}\boldsymbol{x}^{(t)}_i[y^{(t)}_i - \psi'((\boldsymbol{x}^{(t)}_i)^\top\boldsymbol{\beta}^{(t)*})]$. Consider a 1/2-cover of the unit ball in $\mathbb{R}^p$ w.r.t Euclidean norm $\{\boldsymbol{u}_j\}_{j=1}^{N}$ with $N \leq 5^p$ (Example 5.8 in Wainwright (2019)). By a standard argument (see the proof of Theorem 6.5 in Wainwright (2019)), we have

$$\|\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})\|_2 \leq 2\max_{j\in[N]}\left\{\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{u}_j^\top\boldsymbol{x}^{(t)}_i[y^{(t)}_i - \psi'((\boldsymbol{x}^{(t)}_i)^\top\boldsymbol{\beta}^{(t)*})]\right\}.$$

By union bound,

$$\mathbb{E}\exp\{\lambda\|\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})\|_2\} \leq N\max_{j\in[N]}\mathbb{E}\exp\left\{\frac{4\lambda}{n}\sum_{i=1}^{n}\boldsymbol{u}_j^\top\boldsymbol{x}^{(t)}_i[y^{(t)}_i - \psi'((\boldsymbol{x}^{(t)}_i)^\top\boldsymbol{\beta}^{(t)*})]\right\} = (*).$$

Note that by the GLM density function, for any $a \in \mathbb{R}$,

$$\mathbb{E}\exp\left\{a[y^{(t)}_i - \psi'((\boldsymbol{x}^{(t)}_i)^\top\boldsymbol{\beta}^{(t)*})]\right\} \leq \int \rho(y)\exp\left\{[a + (\boldsymbol{x}^{(t)}_i)^\top\boldsymbol{\beta}^{(t)*}]y^{(t)}_i - \psi(a + (\boldsymbol{x}^{(t)}_i)^\top\boldsymbol{\beta}^{(t)*})\right\}\mathrm{d}\mu(y)$$

$$\cdot \exp\left\{\psi(a + (\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*}) - \psi((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*}) - a\psi'((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*})\right\}$$

$$\leq \exp\left\{\frac{1}{2}a^2 \psi''\left(\delta[a + (\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*}] + (1-\delta)(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*}\right)\right\}$$

$$\leq \exp\left\{Ca^2\right\}, \tag{62}$$

where $\delta \in [0, 1]$. Therefore, for any positive $\lambda \leq Cn$ with some $C > 0$,

$$(*) \leq N \max_{j \in [N]} \mathbb{E}\exp\left\{C'\frac{\lambda^2}{n^2}\sum_{i=1}^n (\boldsymbol{u}_j^\top \boldsymbol{x}_i^{(t)})^2\right\} \leq \exp\left\{\frac{C''\lambda^2}{n} + C'''p\right\},$$

where the second inequality comes from the property of sub-exponential variables (see property 3 in Proposition 2.7.1 of Vershynin (2018)). Finally, for any $\delta > 0$, by Chernoff's bound,

$$\mathbb{P}(\|\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})\|_2 > \delta) \leq e^{-\lambda t}\mathbb{E}\exp\{\lambda\|\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})\|_2\} \leq \exp\left\{\frac{C''\lambda^2}{n} + C'''p - \lambda\delta\right\}.$$

Let $\lambda = C''n\delta/2$ ($\delta <$ some constant C), leading to

$$\mathbb{P}(\|\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})\|_2 > \delta) \leq \exp\left\{-Cn\delta^2 + C'p\right\},$$

for all $t \in S$.

E.4.2 Proof of Lemma 44

Note that $\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*}) = \frac{2}{n}\sum_{i=1}^n \boldsymbol{x}_i^{(t)}[y_i^{(t)} - \psi'((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*})] := \boldsymbol{g}^{(t)} \in \mathbb{R}^p$. Denote $\{\boldsymbol{g}^{(t)}\}_{t \in S}$ as $\boldsymbol{G} \in \mathbb{R}^{p \times |S|}$. Consider two $1/4$-covers of the unit ball in $\mathbb{R}^p$ and $\mathbb{R}^{|S|}$ w.r.t Euclidean norm, as $\{\boldsymbol{u}^{(j_1)}\}_{j_1=1}^{N_1}$ and $\{\boldsymbol{v}^{(j_2)}\}_{j_2=1}^{N_2}$ respectively, with $N_1 \leq 9^p$ and $N_2 \leq 9^{|S|}$ (Example 5.8 in Wainwright (2019)). By a standard argument (see the proof of Theorem 6.5 in Wainwright (2019)) , we have

$$\|\boldsymbol{G}\|_2 \lesssim \max_{\substack{j_1 \in [N_1] \\ j_2 \in [N_2]}} |(\boldsymbol{u}^{(j_1)})^\top \boldsymbol{G}\boldsymbol{v}^{(j_2)}| = \max_{\substack{j_1 \in [N_1] \\ j_2 \in [N_2]}} \left|\sum_{t \in S}(\boldsymbol{u}^{(j_1)})^\top \boldsymbol{g}^{(t)} v_k^{j_2}\right|.$$

It follows that for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}e^{\lambda\|\boldsymbol{G}\|_2} \lesssim \sum_{j_1=1}^{N_1}\sum_{j_2=1}^{N_2}\prod_{t=1}^{|S|}\mathbb{E}\exp\{\lambda|(\boldsymbol{u}^{(j_1)})^\top \boldsymbol{g}^{(t)} v_k^{j_2}|\}$$

$$\leq \sum_{j_1=1}^{N_1}\sum_{j_2=1}^{N_2}\prod_{t=1}^{|S|}\left[\mathbb{E}\exp\{\lambda(\boldsymbol{u}^{(j_1)})^\top \boldsymbol{g}^{(t)} v_k^{j_2}\} + \mathbb{E}\exp\{-\lambda(\boldsymbol{u}^{(j_1)})^\top \boldsymbol{g}^{(t)} v_k^{j_2}\}\right].$$

Similar to the calculations in (62), we can obtain that

$$\mathbb{E}\exp\{\lambda(\boldsymbol{u}^{(j_1)})^\top \boldsymbol{g}^{(t)} v_k^{(j_2)}\}, \mathbb{E}\exp\{-\lambda(\boldsymbol{u}^{(j_1)})^\top \boldsymbol{g}^{(t)} v_k^{(j_2)}\} \lesssim \exp\left\{\frac{C}{n}\lambda^2 \cdot |v_k^{(j_2)}|^2\right\},$$

hence

$$\mathbb{E}e^{\lambda\|\boldsymbol{G}\|_2} \lesssim \sum_{j_1=1}^{N_1}\sum_{j_2=1}^{N_2} 2^{|S|} \cdot \exp\left\{\frac{C}{n}\lambda^2\right\} \lesssim \exp\left\{\frac{C}{n}\lambda^2 + C'p + C'|S|\right\}.$$

Then similar to the proof of Lemma 43, the proof can be finished by Chernoff's bound.

E.4.3 Proof of Lemma 45

The proof of part (i) follows the idea in the proof of Theorem 9.36 in Wainwright (2019), and the proof of part (ii) is the same as the proof of equation (39b) in Proposition 1 of Loh and Wainwright (2015).

(i) If $\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*}\|_2 \leq 1$: Denote $\boldsymbol{\Delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*}$, $\varphi_\tau(u) = u^2\mathbb{1}(|u| \leq \tau)$ with any $\tau > 0$. By Taylor expansion, with some $\delta \in [0, 1]$ and any constant $T > 0$,

$$\begin{aligned}
\text{LHS of (60)} &= \frac{1}{n}\sum_{i=1}^n \psi''((\boldsymbol{\beta}^{(t)*})^\top \boldsymbol{x}_i^{(t)} + \delta\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(t)}) \cdot (\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(t)})^2 \\
&\geq \frac{1}{n}\sum_{i=1}^n \psi''((\boldsymbol{\beta}^{(t)*})^\top \boldsymbol{x}_i^{(t)} + \delta\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(t)})\varphi_\tau(\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(t)})\mathbb{1}(|(\boldsymbol{\beta}^{(t)*})^\top \boldsymbol{x}_i^{(t)}| \leq T) \\
&\geq \min_{|u| \leq \tau+T}\psi''(u) \cdot \underbrace{\frac{1}{n}\sum_{i=1}^n \varphi_\tau(\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(t)})\mathbb{1}(|(\boldsymbol{\beta}^{(t)*})^\top \boldsymbol{x}_i^{(t)}| \leq T)}_{(*)}
\end{aligned}$$

It suffices to show that by fixing some $\tau, T > 0$, $(*) \geq C\|\boldsymbol{\Delta}\|_2^2$ w.p. at least $1 - e^{-C'n}$. In fact, we only need to consider the case $\|\boldsymbol{\Delta}\|_2 = 1$. If the results hold when $\|\boldsymbol{\Delta}\|_2 = 1$, then when $\|\boldsymbol{\Delta}\|_2 \in (0, 1)$, since φ_τ is non-decreasing on τ, plugging in $\boldsymbol{\Delta}/\|\boldsymbol{\Delta}\|_2$ implies $(*) \geq C\|\boldsymbol{\Delta}\|_2^2$. So in the following analysis, we consider $\|\boldsymbol{\Delta}\|_2 = 1$.

Define

$$Z_n = \sup_{\|\boldsymbol{\Delta}\|_2=1}\left\{\frac{1}{n}\sum_{i=1}^n \varphi_\tau(\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(t)})\mathbb{1}(|(\boldsymbol{\beta}^{(t)*})^\top \boldsymbol{x}_i^{(t)}| \leq T) - \mathbb{E}[\varphi_\tau(\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(t)})\mathbb{1}(|(\boldsymbol{\beta}^{(t)*})^\top \boldsymbol{x}_i^{(t)}| \leq T)]\right\}.$$

By bounded difference inequality (Corollary 2.21 in Wainwright (2019)) or functional Hoeffding inequality (Theorem 3.26 in Wainwright (2019)),

$$\mathbb{P}(Z_n \geq \mathbb{E}Z_n + \delta) \leq \exp\{-Cn\delta^2/\tau^4\}. \tag{63}$$

Denote $\{\epsilon_i\}_{i=1}^n$ as independent Rademacher variables. By symmetrization,

$$\begin{aligned}
\mathbb{E}Z_n &\lesssim \mathbb{E}_{\boldsymbol{x}^{(t)},\epsilon}\left[\sup_{\|\boldsymbol{\Delta}\|_2=1}\left|\frac{1}{n}\sum_{i=1}^n \varphi_\tau(\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(t)})\mathbb{1}(|(\boldsymbol{\beta}^{(t)*})^\top \boldsymbol{x}_i^{(t)}| \leq T) \cdot \epsilon_i\right|\right] \\
&\lesssim \tau\mathbb{E}_{\boldsymbol{x}^{(t)},\epsilon}\left[\sup_{\|\boldsymbol{\Delta}\|_2=1}\left|\frac{1}{n}\sum_{i=1}^n (\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(t)})\epsilon_i\right|\right] \\
&\lesssim \tau \cdot \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(t)},\epsilon}\sqrt{\sup_{\|\boldsymbol{\Delta}\|_2=1}\sum_{i=1}^n (\boldsymbol{\Delta}^\top \boldsymbol{x}_i^{(t)})^2\epsilon_i^2}
\end{aligned} \tag{64}$$

$$\lesssim \tau \cdot \frac{1}{n}\sqrt{n\mathbb{E}\|\boldsymbol{x}_i^{(t)}\|_2^2} \tag{65}$$

$$\lesssim \tau\sqrt{\frac{p}{n}}, \tag{66}$$

where (64) and (65) are due to Rademacher contraction inequality (equation (5.61) in Wainwright (2019), because the function $\varphi_\tau(\cdot)\mathbb{1}(|(\boldsymbol{\beta}^{(t)*})^\top\boldsymbol{x}_i^{(t)}| \leq T)$ is 2τ-Lipschitz) and Jensen's inequality. On the other hand,

$$\sup_{\|\boldsymbol{\Delta}\|_2=1} \mathbb{E}[\varphi_\tau(\boldsymbol{\Delta}^\top\boldsymbol{x}_i^{(t)})\mathbb{1}(|(\boldsymbol{\beta}^{(t)*})^\top\boldsymbol{x}_i^{(t)}|] \leq T)]$$

$$\geq \sup_{\|\boldsymbol{\Delta}\|_2=1}\left\{\mathbb{E}[(\boldsymbol{\Delta}^\top\boldsymbol{x}_i^{(t)})^2 \cdot \mathbb{1}(|(\boldsymbol{\beta}^{(t)*})^\top\boldsymbol{x}_i^{(t)}|] \leq T)] - \mathbb{E}[(\boldsymbol{\Delta}^\top\boldsymbol{x}_i^{(t)})^2 \cdot \mathbb{1}(|\boldsymbol{\Delta}^\top\boldsymbol{x}_i^{(t)}| > \tau) \cdot \mathbb{1}(|(\boldsymbol{\beta}^{(t)*})^\top\boldsymbol{x}_i^{(t)}|] \leq T)]\right\}$$

$$\geq \sup_{\|\boldsymbol{\Delta}\|_2=1} \mathbb{E}(\boldsymbol{\Delta}^\top\boldsymbol{x}_i^{(t)})^2 - \mathbb{P}(|(\boldsymbol{\beta}^{(t)*})^\top\boldsymbol{x}_i^{(t)}| > T) - \sup_{\|\boldsymbol{\Delta}\|_2=1}\sqrt{\mathbb{E}(\boldsymbol{\Delta}^\top\boldsymbol{x}_i^{(t)})^4} \cdot \sqrt{\mathbb{P}(|\boldsymbol{\Delta}^\top\boldsymbol{x}_i^{(t)}| > \tau)}$$

$$\geq C - \exp\{-C'T^2/(\zeta^{(t)})^2\} - e^{-C''\tau^2}$$
$$\geq C/2, \tag{67}$$

when T and τ are large.

Putting (63), (66), and (67) together, and setting $\delta = $ a small constant c, w.p. at least $1 - e^{-C''n}$, we have

$$(*) \geq \sup_{\|\boldsymbol{\Delta}\|_2=1} \mathbb{E}[\varphi_\tau(\boldsymbol{\Delta}^\top\boldsymbol{x}_i^{(t)})\mathbb{1}(|(\boldsymbol{\beta}^{(t)*})^\top\boldsymbol{x}_i^{(t)}|] \leq T)] - Z_n$$

$$\geq C/2 - C'\sqrt{\frac{p}{n}} - \delta$$

$$\geq C/4,$$

when $n \gtrsim p + \log T$.

(ii) If $\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*}\|_2 > 1$: Denote $\boldsymbol{\Delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*}$, $\delta = \frac{1}{\|\boldsymbol{\Delta}\|_2}$. By convexity of $f^{(t)}$,

$$f^{(t)}(\delta\boldsymbol{\beta} + (1-\delta)\boldsymbol{\beta}^{(t)*}) \leq \delta f^{(t)}(\boldsymbol{\beta}) + (1-\delta)f^{(t)}(\boldsymbol{\beta}^{(t)*}),$$

which combining with part (i) implies that w.p. at least $1 - e^{-C''n}$,

$$f^{(t)}(\boldsymbol{\beta}) - f^{(t)}(\boldsymbol{\beta}^{(t)*}) - \nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*}) \geq \frac{f^{(t)}(\boldsymbol{\beta}^{(t)*} + \delta\boldsymbol{\Delta}) - f^{(t)}(\boldsymbol{\beta}^{(t)*}) - \delta\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})^\top\boldsymbol{\Delta}}{\delta}$$

$$\geq \delta C\|\boldsymbol{\Delta}\|_2^2$$
$$= C\|\boldsymbol{\Delta}\|_2,$$

because $\|\delta\boldsymbol{\Delta}\|_2 = 1$.

E.4.4 Proof of Lemma 46

It is easy to see that

$$f^{(t)}(\boldsymbol{\beta}) - f^{(t)}(\widetilde{\boldsymbol{\beta}}) - \nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})^\top(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) = \underbrace{f^{(t)}(\boldsymbol{\beta}) - f^{(t)}(\boldsymbol{\beta}^{(t)*}) - \nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*})}_{[1]}$$

$$+ \underbrace{f^{(t)}(\boldsymbol{\beta}^{(t)*}) - f^{(t)}(\widetilde{\boldsymbol{\beta}}) + \nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})^\top(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(t)*})}_{[2]}.$$

By Lemma 45, w.p. at least $1 - e^{-C''n}$,

$$[1] \geq C\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*}\|_2^2.$$

And with some $\delta, \delta' \in [0, 1]$, by Assumption 4, w.p. at least $1 - e^{-C(p+\log T)}$,

$$[2] \leq \nabla f^{(t)}(\delta\boldsymbol{\beta}^{(t)*} + (1-\delta)\widetilde{\boldsymbol{\beta}})^\top(\boldsymbol{\beta}^{(t)*} - \widetilde{\boldsymbol{\beta}}) + \nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})^\top(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(t)*})$$
$$= (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(t)*})^\top\nabla^2 f^{(t)}(\delta'[\delta\boldsymbol{\beta}^{(t)*} + (1-\delta)\widetilde{\boldsymbol{\beta}}] + (1-\delta')\boldsymbol{\beta}^{(t)*})^\top(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(t)*})$$
$$\lesssim \frac{1}{n}\sum_{i=1}^n [(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(t)*})^\top\boldsymbol{x}_i^{(t)}]^2$$
$$\lesssim \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(t)*}\|_2^2,$$

where the last inequality comes from the fact that $\|\widehat{\boldsymbol{\Sigma}}^{(t)}\|_2 \leq C$ w.p. at least $1 - e^{-C'(p+\log T)}$ (by Lemma 18). Putting all pieces together leads to the desired conclusion.

E.4.5 Proof of Lemma 47

Define $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} = \boldsymbol{A}^\top\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$. It is easy to see that

$$\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq \|(\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{I})\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \lesssim \|(\boldsymbol{A}^\perp)^\top\boldsymbol{A}^{(t)*}\|_2\zeta^{(t)} \leq c.$$

If $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 > 1$, then by Lemma 46, w.p. at least $1 - e^{-C(p+\log T)}$,

$$f^{(t)}(\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}) - f^{(t)}(\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}) \geq -\|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 \cdot \|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2 + C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2$$

$$- C'\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2$$
$$\geq (C - \|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2) \cdot \|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2$$
$$- C\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 - C'\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2$$
$$\geq \frac{1}{2}C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2 - Cc - C'c^2$$
$$\geq \frac{1}{2}C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 - \frac{1}{2}Cc - Cc - C'c^2$$
$$> 0,$$

which is contradicted with the definition of $\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}$. Therefore, we must have $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq 1$. By Lemma 46, w.p. at least $1 - e^{-C(p+\log T)}$,

$$f^{(t)}(\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}) - f^{(t)}(\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}) \geq -\|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 \cdot \|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2 + C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2^2$$

$$- C'\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2$$
$$\geq -C''\sqrt{\frac{p + \log T}{n}} + C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2^2 - C'c^2.$$

We know that $f^{(t)}(\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}) - f^{(t)}(\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}) \geq 0$ by optimality, so we must have

$$-C''\sqrt{\frac{p + \log T}{n}} + C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2^2 - C'c^2 \leq 0,$$

i.e., $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2 \leq \sqrt{\frac{C'}{C}}c + (C'')^{1/2}(\frac{p+\log T}{n})^{1/4} \leq$ a small constant c', w.p. at least $1 - e^{-C(p+\log T)}$. This implies that $\|\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}\|_2 = \|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}\|_2 \leq \|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2 + \|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 + \|\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq C\zeta^{(t)}$, w.p. at least $1 - e^{-C(p+\log T)}$, which completes the proof.

E.4.6 Proof of Lemma 48

Define $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} = \boldsymbol{A}^{\top}\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}$. It is easy to see that

$$\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}\|_2 \leq \|(\boldsymbol{A}\boldsymbol{A}^{\top} - \boldsymbol{I})\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}\|_2 \lesssim \|(\boldsymbol{A}^{\perp})^{\top}\widetilde{\boldsymbol{A}}\|_2\zeta^{(t)} \leq c.$$

If $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 > 1$, then by Lemma 46, w.p. at least $1 - e^{-C(p+\log T)}$,

$$f^{(t)}(\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}) - f^{(t)}(\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}) \geq -\|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 \cdot \|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2 + C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2$$

$$- C'\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2$$
$$\geq (C - \|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2) \cdot \|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2$$
$$- C\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 - C'\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2$$
$$\geq \frac{1}{2}C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2 - C\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}\|_2 - C\|\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2$$

$$- C'\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}\|_2^2 - C'\|\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2$$
$$\geq \frac{1}{2}C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2 - Cc - C'c^2$$
$$\geq \frac{1}{2}C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 - \frac{1}{2}Cc - Cc - C'c^2$$
$$> 0,$$

which is contradicted with the definition of $\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}$. Therefore, we must have $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq 1$. By Lemma 46, w.p. at least $1 - e^{-C(p+\log T)}$,

$$f^{(t)}(\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}) - f^{(t)}(\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}) \geq -\|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 \cdot \|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2 + C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2^2$$

$$- C'\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2$$

$$\geq -\|\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\|_2 \cdot \|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2 + C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2^2$$

$$- \frac{1}{2}C\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}\|_2^2 - C''\|\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2$$

$$\geq -C''\sqrt{\frac{p+\log T}{n}} + \frac{1}{2}C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2^2 - 4Cc^2.$$

We know that $f^{(t)}(\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}) - f^{(t)}(\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}) \geq 0$ by optimality, so we must have

$$-C''\sqrt{\frac{p+\log T}{n}} + C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2^2 - C'c^2 \leq 0,$$

i.e., $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2 \leq \sqrt{\frac{C'}{C}}c + (C'')^{1/2}(\frac{p+\log T}{n})^{1/4} \leq$ a small constant c', w.p. at least $1 - e^{-C(p+\log T)}$. This implies that $\|\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}\|_2 = \|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}\|_2 \leq \|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2 + \|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 + \|\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq C\zeta^{(t)}$, w.p. at least $1 - e^{-C(p+\log T)}$, which completes the proof.

E.4.7 Proof of Lemma 49

Similar to the proof of Lemma 31, by the optimality of $\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}$,

$$\boldsymbol{A}^\top(\nabla f^{(t)}(\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}) + \widetilde{\boldsymbol{\Sigma}}^{(t)}(\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)})) = \boldsymbol{0},$$

where $\widetilde{\boldsymbol{\Sigma}}^{(t)} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^{(t)}(\boldsymbol{x}_i^{(t)})^\top \psi''((\boldsymbol{x}_i^{(t)})^\top \widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)} + \delta(\boldsymbol{x}_i^{(t)})^\top(\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}))$, $\delta \in [0,1]$. With similar arguments used in the proof of Lemma 45, it can be shown that $\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}^{(t)}) > 0$ w.p. at least $1 - e^{-C(p+\log T)}$. The remaining steps are the same as the steps in the proof of Lemma 31.

E.4.8 Proof of Lemma 50

The proof is almost the same as the proof of Lemma 49, so we omit it here.

E.4.9 Proof of Lemma 51

If $\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 > 1$, by Lemmas 43 and 45, w.p. at least $1 - e^{-C'(p+\log T)}$, for any $\widetilde{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$,

$$f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}) + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2 - f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}) - \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2$$

$$\geq \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})^\top(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}) - 2\frac{\lambda}{\sqrt{n}} + C\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2$$

$$\geq \left(C - C'\sqrt{\frac{p+\log T}{n}}\right)\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 - 2\frac{\lambda}{\sqrt{n}}$$

$$\geq \frac{1}{2}C - \frac{1}{4}C$$

$$> 0,$$

which contradicts to the definitions of $\boldsymbol{A}^{(t)}$ and $\boldsymbol{\theta}^{(t)}$. Hence $\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \leq 1$, w.p. at least $1 - e^{-C'(p+\log T)}$. Then by Lemmas 43 and 45 again, w.p. at least $1 - e^{-C'(p+\log T)}$, for any $\widetilde{\boldsymbol{A}} \in \mathcal{O}^{p\times r}$,

$$f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}) + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2 - f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}) - \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2$$

$$\geq \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})^\top(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}) - 2\frac{\lambda}{\sqrt{n}} + C\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2$$

$$\geq C\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2 - C'\sqrt{\frac{p+\log T}{n}}\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 - 2\frac{\lambda}{\sqrt{n}}.$$

By the optimality of $\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}$, we must have the RHS of above ≤ 0 w.p. at least $1 - e^{-C'(p+\log T)}$, which entails that $\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \lesssim \sqrt{\frac{p+\log T}{n}} + \lambda^{1/2}n^{-1/4} \leq$ a small constant c, which completes the proof.

Appendix F. Proofs for Non-linear Regression Models

F.1 Lemmas

Lemma 52 *Suppose Assumptions 1, 6, and 7 hold. Then w.p. at least $1 - e^{-C(p+\log T)}$, for any $\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}} \in \mathbb{R}^p$, for any $t \in S$,*

$$f^{(t)}(\boldsymbol{\beta}) - f^{(t)}(\widetilde{\boldsymbol{\beta}}) - \nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})^\top(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) \geq C'\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_2^2 - C''\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(t)*}\|_2^2.$$

Lemma 53 *Suppose Assumptions 1, 6, and 7 hold. Then when $\delta \leq C$ with some constant $C > 0$,*

$$\|\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})\|_2 \lesssim \delta,$$

w.p. at least $1 - \exp\{-Cn\delta^2 + C'p\}$. By taking $t \asymp \sqrt{\frac{p+\log T}{n}}$, this implies that:

(i) $\max_{t\in S}\|\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})\|_2 \lesssim \sqrt{\frac{p+\log T}{n}}$, w.p. at least $1 - e^{-C(p+\log T)}$;

(ii) $\|\nabla f^{(0)}(\boldsymbol{\beta}^{(0)*})\|_2 \lesssim \sqrt{\frac{p}{n_0}}$, w.p. at least $1 - e^{-C(p+\log T)}$.

Lemma 54 *Suppose Assumptions 1, 6, and 7 hold. Then when $t \leq C$ with some constant $C > 0$, for any $S \subseteq [T]$,*

$$\|\{\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})\}_{t\in S}\|_2 \lesssim \sqrt{\frac{p + |S| + \delta}{n}},$$

w.p. at least $1 - e^{-C'\delta}$.

Lemma 55 *Suppose Assumptions 1, 6, and 7 hold. Then w.p. at least $1 - e^{-C'(p+\log T)}$, for all $t \in S$, for all $\boldsymbol{A}, \widetilde{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$ with $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2, \|\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \le c'$, where $c' > 0$ is a small constant, we have*

$$\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}\|_2 \lesssim \|\boldsymbol{A}\boldsymbol{A}^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2 (\|\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}\|_2 + \|\nabla f^{(t)}(\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)})\|_2).$$

Lemma 56 *Suppose Assumptions 1, 6, and 7 hold. Then w.p. at least $1 - e^{-C'(p+\log T)}$, for all $t \in S$, for all $\boldsymbol{A}, \widetilde{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$ with $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \le c'$, where $c' > 0$ is a small constant, we have*

$$\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \lesssim \|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top\|_2 \zeta^{(t)} + \sqrt{\frac{r + \log T}{n}}.$$

Lemma 57 *Suppose Assumptions 1, 6, and 7 hold. Then w.p. at least $1 - e^{-C'(p+\log T)}$, for all $t \in S$, for all $\boldsymbol{A} \in \mathcal{O}^{p \times r}$ with $\|\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top\|_2 \le c(\zeta^{(t)})^{-1}$, where $c > 0$ is a small constant, we have*

(i) $\|\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}\|_2 \le C''\zeta^{(t)}$;

(ii) $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \le c'$ with a small constant $c' > 0$.

Lemma 58 *Suppose Assumptions 1, 6, and 7 hold. Assume there exists $\widetilde{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$ such that $\|\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \le c$, where $c > 0$ is a small constant. Then w.p. at least $1 - e^{-C'(p+\log T)}$, for all $t \in S$, for all $\boldsymbol{A} \in \mathcal{O}^{p \times r}$ with $\|\boldsymbol{A}\boldsymbol{A}^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2 \le c'(\zeta^{(t)})^{-1}$, where $c' > 0$ is a small constant, we have*

(i) $\|\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}\|_2 \le C''\zeta^{(t)}$;

(ii) $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \le c''$ with a small constant $c'' > 0$.

Lemma 59 *Suppose Assumptions 1, 6, and 7 hold. Let $n \ge C\lambda^2$ with a sufficiently large C. Given $\widetilde{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$, denote $(\boldsymbol{A}^{(t)}, \boldsymbol{\theta}^{(t)}) = \arg\min_{\boldsymbol{A} \in \mathcal{O}^{p \times r}, \boldsymbol{\theta} \in \mathbb{R}^r} \{f^{(t)}(\boldsymbol{A}\boldsymbol{\theta}) + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}\boldsymbol{A}^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2\}$. Then w.p. at least $1 - e^{-C(p+\log T)}$, for any $\widetilde{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$, we have $\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \le c$ with a small constant $c > 0$.*

F.2 Proof of Theorem 12

For the penalized ERM, the proof logic is almost the same as the logic underlying the proofs of Proposition 42 and Theorem 2. Similar to the proof of Theorem 11, we only point out the differences and skip the details. We need to bound $\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2$ first. Then we know that $\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2$ and $\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2$ are small w.h.p. Then a direct application of Lemma 58 implies that $\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)*} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2$ is small w.h.p. Finally, applying Lemma 55, we have $\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)*} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2 \lesssim \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 (\|\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2 + \|\nabla f^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)})\|_2) \lesssim$

$\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \zeta^{(t)}$ w.h.p. which gives us the ideal bound for $\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)*} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\|_2$ for all $t \in S$.

(i) In the proof of Proposition 42, for $t \in \mathcal{A}_1$, we know $\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \le c$ with a small constant $c > 0$ w.h.p. by Lemma 59. Then by Lemmas 52, 54, and 56, w.p. at least $1 - e^{-C'(r+\log T)}$,

$$\frac{1}{T} \sum_{t \in \mathcal{A}_1} \left[f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}) - f^{(t)}(\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}) \right]$$

$$\ge -\frac{1}{T}\left| \left\langle \{\nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})\}_{t \in \mathcal{A}_1}, \{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} + \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t \in \mathcal{A}_1} \right\rangle \right|$$

$$+ \frac{C_1}{T}\|\{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t \in \mathcal{A}_1}\|_{\mathrm{F}}^2 - \frac{C_2}{T}\|\{\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t \in \mathcal{A}_1}\|_{\mathrm{F}}^2$$

$$\ge -\frac{C}{T} \sum_{t \in \mathcal{A}_1} \sqrt{\frac{p + \log T}{n}} \cdot \zeta^{(t)} (\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 + \|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \overline{\boldsymbol{A}}\,\overline{\boldsymbol{A}}^\top\|_2)$$

$$+ \frac{C_1}{T}\|\{\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)}\}_{t \in \mathcal{A}_1}\|_{\mathrm{F}}^2 - C\left(h^2 \cdot \frac{1}{T} \sum_{t \in \mathcal{A}_1} (\zeta^{(t)})^2 + \frac{r + \log T}{n} \right)$$

$$\ge -C\left(h^2 \cdot \frac{1}{T} \sum_{t \in \mathcal{A}_1} (\zeta^{(t)})^2 + \frac{r + \log T}{n} \right) - \frac{C'}{T} \sum_{t \in \mathcal{A}_1} \sqrt{\frac{p + \log T}{n}} \cdot \zeta^{(t)} \sqrt{r} \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2.$$

Therefore the calculations in (51) are still correct under non-linear regression model (7).

(ii) In the proof of Proposition 42, for $t \in \mathcal{A}_2$, we used the fact that w.p. at least $1 - e^{-C(r+\log T)}$,

$$\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}_{\boldsymbol{A}^{(t)}}^{(t)} - \widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\|_2 \lesssim \|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 (\|\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\|_2 + \|\nabla f^{(t)}(\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)})\|_2). \tag{68}$$

Here this still holds because

- $\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}_{\boldsymbol{A}^{(t)}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \le c$ with a small constant $c > 0$, by Lemma 59;

- For $t \in \mathcal{A}_2$, we know that $\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top\|_2 \le \frac{c}{\sqrt{r}} \le c'(\zeta^{(t)})^{-1}$ with a small constant $c' > 0$. Then by Lemma 58, $\|\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)}\|_2 \le C\zeta^{(t)}$ and $\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \le Cc'$ with a small constant $c' > 0$.

Then (68) holds by Lemma 55.

Finally, Lemma 38 still holds for non-linear regression model (7). Note that in Lemma 38, (i) only requires the maximum eigenvalue of Hessian to be upper bounded when the evaluation point is close to the true $\boldsymbol{\beta}^{(t)*}$ (see Lemma E.3 in Duan and Wang (2023)), which is true by Assumptions 1 and 6 (see our analysis in the proof of Lemma 55). By Lemma E.2 in Duan and Wang (2023), (ii) and (iii) require the minimum eigenvalue of Hessian to be lower bounded when $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \le$ some constant C w.h.p. and $\gamma/\sqrt{n_0} \le C$, both of which are true.

For the spectral method, the proof is almost the same as the proof of Theorem 7, hence we do not repeat it here..

F.3 Proof of Theorem 17

We only outline the proof when Algorithm 4 is coupled with Algorithm 1. If Algorithm 4 is coupled with Algorithm 2, we can follow a similar argument with the proof of Theorem 7.

Denote $\eta = r\sqrt{\frac{p}{nT}} + \sqrt{r}h + \sqrt{r}\sqrt{\frac{r+\log T}{n}} + \frac{|S^c|}{T} \cdot r \cdot \frac{\lambda}{\sqrt{n}}$.

(i) When $\eta \leq C\sqrt{\frac{p}{n_0}}$: since $\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \boldsymbol{A}^{(0)*}(\boldsymbol{A}^{(0)*})^\top\|_2 \leq$ a small constant c, by Lemma 59, w.p. at least $1 - e^{-C(p+\log T)}$, $\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \boldsymbol{A}^{(0)*}\boldsymbol{\theta}^{(0)*}\|_2 \leq$ a small constant c'. Then by Lemma 56 and Theorem 12, w.p. at least $1 - e^{-C(r+\log T)}$, $\|\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}_{\widehat{\overline{\boldsymbol{A}}}}^{(t)} - \boldsymbol{A}^{(0)*}\boldsymbol{\theta}^{(0)*}\|_2 \lesssim$ $\|\widehat{\overline{\boldsymbol{A}}}(\widehat{\overline{\boldsymbol{A}}})^\top - \boldsymbol{A}^{(0)*}(\boldsymbol{A}^{(0)*})^\top\|_2 \lesssim \eta$. As we commented in the proof of Theorem 12, Lemma 41.(i) still holds for non-linear regression model (7) because the maximum eigenvalue of Hessian is upper bounded w.h.p. when the evaluation point is close to the true $\boldsymbol{\beta}^{(t)*}$

(ii) When $\eta > C\sqrt{\frac{p}{n_0}}$: the result comes from Lemma 41.(ii), which still holds when $\gamma/\sqrt{n_0} \lesssim$ some constant C because of Lemma E.2 in Duan and Wang (2023).

F.4 Proofs of Lemmas

F.4.1 PROOF OF LEMMA 52

First,

$$f^{(t)}(\boldsymbol{\beta}) - f^{(t)}(\widetilde{\boldsymbol{\beta}}) - \nabla f^{(t)}(\widetilde{\boldsymbol{\beta}})^\top(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) \geq \underbrace{f^{(t)}(\boldsymbol{\beta}) - f^{(t)}(\boldsymbol{\beta}^{(t)*}) - \nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*})}_{[1]}$$
$$+ \underbrace{f^{(t)}(\boldsymbol{\beta}^{(t)*}) - f^{(t)}(\widetilde{\boldsymbol{\beta}}) - \nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})^\top(\boldsymbol{\beta}^{(t)*} - \widetilde{\boldsymbol{\beta}})}_{[2]}.$$

By Taylor expansion,

$$[1] = \left\langle \int_0^1 \nabla f^{(t)}(\delta\boldsymbol{\beta} + (1-\delta)\boldsymbol{\beta}^{(t)*})\mathrm{d}t - \nabla f^{(t)}(\boldsymbol{\beta}^{(t)*}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*} \right\rangle$$

$$= \frac{2}{n}\sum_{i=1}^{n}\int_0^1 [g(\delta(\boldsymbol{x}_i^{(t)})^\top\boldsymbol{\beta} + (1-\delta)(\boldsymbol{x}_i^{(t)})^\top\boldsymbol{\beta}^{(t)*}) - g((\boldsymbol{x}_i^{(t)})^\top\boldsymbol{\beta}^{(t)*})]g'((\boldsymbol{x}_i^{(t)})^\top\delta\boldsymbol{\beta} + (1-\delta)\boldsymbol{\beta}^{(t)*})\mathrm{d}\delta$$

$$\cdot [(\boldsymbol{x}_i^{(t)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*})] - \frac{2}{n}\sum_{i=1}^{n}\epsilon_i^{(t)} \cdot \int_0^1 [g'(\delta(\boldsymbol{x}_i^{(t)})^\top\boldsymbol{\beta} + (1-\delta)(\boldsymbol{x}_i^{(t)})^\top\boldsymbol{\beta}^{(t)*}) - g'((\boldsymbol{x}_i^{(t)})^\top\boldsymbol{\beta}^{(t)*})]\mathrm{d}\delta$$

$$\cdot [(\boldsymbol{x}_i^{(t)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*})]$$

$$= \frac{2}{n}\sum_{i=1}^{n}\int_0^1 \underbrace{g'((\boldsymbol{x}_i^{(t)})^\top\boldsymbol{\beta}_i)}_{\geq C} \cdot \underbrace{g'(\delta(\boldsymbol{x}_i^{(t)})^\top\boldsymbol{\beta} + (1-\delta)(\boldsymbol{x}_i^{(t)})^\top\boldsymbol{\beta}^{(t)*})}_{\geq C}\delta\mathrm{d}\delta \cdot [(\boldsymbol{x}_i^{(t)})^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*})]^2$$

$$- \frac{2}{n} \sum_{i=1}^{n} \epsilon_i^{(t)} \cdot \int_0^1 \int_0^1 g''(w(\delta(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta} + (1-\delta)(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*}) + (1-w)(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*}) \delta \mathrm{d}\delta \mathrm{d}w$$

$$\cdot [(\boldsymbol{x}_i^{(t)})^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*})]^2$$

$$\geq C(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*})^\top \widehat{\boldsymbol{\Sigma}}^{(t)} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*}) - \frac{2}{n} \sum_{i=1}^{n} \epsilon_i^{(t)} b_i^{(t)} [(\boldsymbol{x}_i^{(t)})^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*})]^2,$$

where $b_i^{(t)} = \int_0^1 \int_0^1 g''(w(\delta(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta} + (1-\delta)(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*}) + (1-w)(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*}) \delta \mathrm{d}\delta \mathrm{d}w$ and $|b_i^{(t)}| \leq C'$. Therefore, by Hölder's inequality,

$$\left| \frac{2}{n} \sum_{i=1}^{n} \epsilon_i^{(t)} b_i^{(t)} [(\boldsymbol{x}_i^{(t)})^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*})]^2 \right| \leq \frac{2C'}{n} \sum_{i=1}^{n} |\epsilon_i^{(t)}| [(\boldsymbol{x}_i^{(t)})^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*})]^2$$

$$\leq \frac{2C'}{n} \sup_{\|\boldsymbol{u}\|_2 \leq 1} \left\| \sum_{i=1}^{n} |\epsilon_i^{(t)}| \cdot [(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}]^2 \right\|_2 \cdot \|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*}\|_2^2.$$

Suppose $\{\boldsymbol{u}_j\}_{j=1}^{N}$ is a $1/64$-packing of the unit ball in $\mathbb{R}^p$, with $N \lesssim 129^p$. By standard arguments,

$$\frac{1}{n} \sup_{\|\boldsymbol{u}\|_2 \leq 1} \left\| \sum_{i=1}^{n} |\epsilon_i^{(t)}| \cdot [(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}]^2 \right\|_2 \lesssim \frac{1}{n} \max_{j \in [N]} \left\{ \sum_{i=1}^{n} |\epsilon_i^{(t)}| \cdot [(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}_j]^2 \right\}.$$

It is easy to see that for any $\delta > 0$,

$$\mathbb{P} \left(|\epsilon_i^{(t)}| \cdot [(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}_j]^2 - \mathbb{E}\{|\epsilon_i^{(t)}| \cdot [(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}_j]^2\} > \delta \right) \leq \mathbb{P} \left(|\epsilon_i^{(t)}| \cdot [(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}_j]^2 > \delta \right)$$

$$\leq \mathbb{P}(|\epsilon_i^{(t)}| > \delta^{1/3}) + \mathbb{P}(|(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}_j| > \delta^{1/3})$$

$$\lesssim \exp\{-C\delta^{2/3}\},$$

and $\mathbb{E} \left\{ \{|\epsilon_i^{(t)}|[(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}_j]^2 - \mathbb{E}[|\epsilon_i^{(t)}|[(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}_j]^2]\}^2 \cdot \mathbb{1}\left(|\epsilon_i^{(t)}|[(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}_j]^2 - \mathbb{E}[|\epsilon_i^{(t)}|[(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}_j]^2] < 0 \right) \right\} \leq$ $\sqrt{\mathbb{E}\{|\epsilon_i^{(t)}|^2[(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}_j]^4\}} \leq C < \infty$. By Corollary 2 in Bakhshizadeh et al. (2020),

$$\mathbb{P} \left(|\epsilon_i^{(t)}| \cdot [(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}_j]^2 - \mathbb{E}\{|\epsilon_i^{(t)}| \cdot [(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}_j]^2\} > \delta \right) \leq \exp\{-Cn\delta\} + \exp\{-C(n\delta)^{2/3}\},$$

which entails that w.p. at least $1 - e^{-C'(p+\log T)}$,

$$\frac{1}{n} \max_{j \in [N]} \left\{ \sum_{i=1}^{n} |\epsilon_i^{(t)}| \cdot [(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}_j]^2 \right\} \lesssim \max_{j \in [N]} \mathbb{E}\{|\epsilon_i^{(t)}| \cdot [(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}_j]^2\} + \sqrt{\frac{p + \log T}{n}} \vee \frac{(p + \log T)^{3/2}}{n}$$

$$\leq C''.$$

Hence by Lemma 18 and the assumption on g'', w.p. at least $1 - e^{-C'(p + \log T)}$,

$$[1] \geq C\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*}\|_2^2 \geq \frac{C}{2}\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_2^2 - C\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(t)*}\|_2^2. \tag{69}$$

On the other hand, similarly, we can show that w.p. at least $1 - e^{-C'(p + \log T)}$,

$$[2] \lesssim \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(t)*}\|_2^2. \tag{70}$$

Combining (69) and (70) completes the proof.

F.4.2 Proof of Lemma 53

Note that $\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*}) = -\frac{2}{n}\sum_{i=1}^{n} \epsilon_i^{(t)} g'((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*})\boldsymbol{x}_i^{(t)}$, where $\epsilon_i^{(t)} g'((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*})\boldsymbol{x}_i^{(t)}$ is sub-exponential in the sense that $\mathbb{E}\exp\{\lambda \epsilon_i^{(t)} g'((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*})(\boldsymbol{x}_i^{(t)})^\top \boldsymbol{u}\} \leq \exp\{C^2\lambda^2\|\boldsymbol{u}\|_2^2\}$ for any $\boldsymbol{u} \in \mathbb{R}^p$ and $\lambda \in \mathbb{R}$ with $|\lambda|\|\boldsymbol{u}\|_2 \leq C'$. Then the proof is very similar to the proof of Lemma 43, so we omit the details.

F.4.3 Proof of Lemma 54

The proof is very similar to the proof of Lemma 44, so omitted.

F.4.4 Proof of Lemma 55

Denote $\boldsymbol{\beta}_\delta = \delta\boldsymbol{\beta} + (1 - \delta)\widetilde{\boldsymbol{\beta}}$ with some $t \in [0, 1]$. It suffices to show that w.h.p, for all $t \in [0, 1]$ and all $t \in S$, the minimum and maximum eigenvalues of $\nabla^2 f^{(t)}(\boldsymbol{\beta}_\delta)$ are bounded away from zero and infinity.

Note that

$$\nabla^2 f^{(t)}(\boldsymbol{\beta}_\delta) = \frac{1}{n}\sum_{i=1}^{n}[g'((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}_\delta)]^2 \cdot \boldsymbol{x}_i^{(t)}(\boldsymbol{x}_i^{(t)})^\top$$

$$+ \frac{1}{n}\sum_{i=1}^{n}[g((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}_\delta) - y_i^{(t)}] \cdot g''((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}_\delta) \cdot \boldsymbol{x}_i^{(t)}(\boldsymbol{x}_i^{(t)})^\top$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}[g'((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}_\delta)]^2 \cdot \boldsymbol{x}_i^{(t)}(\boldsymbol{x}_i^{(t)})^\top}_{[1]}$$

$$+ \underbrace{\frac{1}{n}\sum_{i=1}^{n}[g((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}_\delta) - g((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*})] \cdot g''((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}_\delta) \cdot \boldsymbol{x}_i^{(t)}(\boldsymbol{x}_i^{(t)})^\top}_{[2]}$$

$$- \underbrace{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^{(t)} \cdot g''((\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}_\delta) \cdot \boldsymbol{x}_i^{(t)}(\boldsymbol{x}_i^{(t)})^\top}_{[3]}.$$

It is straightforward to see that w.p. at least $1 - e^{-C'(p + \log T)}$,

$$0 < C \leq \min_{t \in S} \lambda_{\min}([1]) \leq \max_{t \in S} \lambda_{\max}([1]) \leq C' < \infty,$$

121

by the condition on g' and Lemma 18. And similar to the analysis in the proof of Lemma 52, w.p. at least $1 - e^{-C'(p+\log T)}$,

$$\lambda_{\max}([3]) \le c,$$

where c is small. Similarly, it can be shown that

$$[2] = \frac{1}{n}\sum_{i=1}^{n}\int_{0}^{1}g'(w(\boldsymbol{x}_i^{(t)})^\top\boldsymbol{\beta}_\delta + (1-w)(\boldsymbol{x}_i^{(t)})^\top\boldsymbol{\beta}^{(t)*})\mathrm{d}w \cdot g''((\boldsymbol{x}_i^{(t)})^\top\boldsymbol{\beta}_\delta)\cdot\boldsymbol{x}_i^{(t)}(\boldsymbol{x}_i^{(t)})^\top\cdot(\boldsymbol{x}_i^{(t)})^\top(\boldsymbol{\beta}_\delta-\boldsymbol{\beta}^{(t)*}),$$

$$\lambda_{\max}([2]) \le C\|\boldsymbol{\beta}_\delta - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)*}\|_2 + \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(t)*}\|_2 \le c,$$

w.p. at least $1 - e^{-C'(p+\log T)}$, where c is small. Putting all the pieces together, we have

$$0 < C \le \min_{t\in S}\lambda_{\min}(\nabla^2 f^{(t)}(\boldsymbol{\beta}_\delta)) \le \max_{t\in S}\lambda_{\max}(\nabla^2 f^{(t)}(\boldsymbol{\beta}_\delta)) \le C' < \infty,$$

w.p. at least $1 - e^{-C(p+\log T)}$. The remaining steps are the same as in the proof of Lemma 31.

F.4.5 Proof of Lemma 56

The proof is similar to the proofs of Lemmas 55 and 29, so omitted.

F.4.6 Proof of Lemma 57

Define $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} = \boldsymbol{A}^\top\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$. It is easy to see that

$$\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \le \|(\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{I})\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \lesssim \|(\boldsymbol{A}^\perp)^\top\boldsymbol{A}^{(t)*}\|_2\zeta^{(t)} \le c.$$

Then by Lemma 52, w.p. at least $1 - e^{-C(p+\log T)}$,

$$f^{(t)}(\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}) - f^{(t)}(\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}) \ge C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2^2 - C'\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2^2 - C'\|\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})\|_2\cdot\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2,$$

where the RHS must be non-positive, otherwise it is contradicted by the definition of $\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}$. Therefore, $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2 \lesssim \|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 + \|\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})\|_2^{1/2} \le c$, where $c > 0$ is a small constant, w.p. at least $1 - e^{-C(p+\log T)}$. This also implies that $\|\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}\|_2 \le C\zeta^{(t)}$ w.p. at least $1 - e^{-C'(p+\log T)}$.

F.4.7 Proof of Lemma 58

Define $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} = \boldsymbol{A}^\top\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}$. It is easy to see that

$$\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}\|_2 \le \|(\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{I})\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}\|_2 \lesssim \|(\boldsymbol{A}^\perp)^\top\widetilde{\boldsymbol{A}}\|_2\zeta^{(t)} \le c.$$

Then by Lemma 52, w.p. at least $1 - e^{-C(p+\log T)}$,

$$f^{(t)}(\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}) - f^{(t)}(\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}) \ge C\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2^2 - C'\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2^2 - C'\|\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})\|_2\cdot\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2,$$

where the RHS must be non-positive, otherwise it is contradicted by the definition of $\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}$. Therefore, $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2 \lesssim \|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 + \|\nabla f^{(t)}(\boldsymbol{\beta}^{(t)*})\|_2^{1/2} \leq$ a small constant c w.p. at least $1 - e^{-C(p+\log T)}$. This also implies that $\|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \leq \|\boldsymbol{A}\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)} - \boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)}\|_2 +$ $\|\boldsymbol{A}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{A}}^{(t)} - \widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)}\|_2 + \|\widetilde{\boldsymbol{A}}\boldsymbol{\theta}_{\widetilde{\boldsymbol{A}}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \leq$ a small constant c and $\|\boldsymbol{\theta}_{\boldsymbol{A}}^{(t)}\|_2 \leq C$ w.p. at least $1 - e^{-C'(p+\log T)}$, which completes the proof.

F.4.8 PROOF OF LEMMA 59

By Lemmas 52 and 54, w.p. at least $1 - e^{-C(p+\log T)}$, for any $\widetilde{\boldsymbol{A}} \in \mathcal{O}^{p \times r}$,

$$f^{(t)}(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}) + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2 - f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}) - \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{A}})^\top\|_2$$

$$\geq \nabla f^{(t)}(\boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*})^\top(\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}) - 2\frac{\lambda}{\sqrt{n}} + C\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2$$

$$\geq C\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2^2 - C'\sqrt{\frac{p + \log T}{n}}\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 - 2\frac{\lambda}{\sqrt{n}}.$$

By definition of $\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}$, the RHS must be non-positive, which gives us

$$\|\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)} - \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}\|_2 \lesssim \sqrt{\frac{p + \log T}{n}} + \frac{\lambda^{1/2}}{n^{1/4}} \leq c,$$

where $c > 0$ is a small constant, w.p. at least $1 - e^{-C'(p+\log T)}$.

Appendix G. Proofs for Estimation of Intrinsic Dimension r

G.1 Proof of Theorem 9

We already mentioned the motivation of the thresholding strategy briefly in Section 3. Intuitively, there is a gap between $\sigma_r(\boldsymbol{B}_S^*)$ and $\sigma_{r+1}(\boldsymbol{B}_S^*)$ when h and $|S^c|$ are small. But in practice, we do not have access to the singular values of $\boldsymbol{B}_S^*$. We have to estimate $\boldsymbol{B}_S^*$ first, then use the singular value of this estimate. As long as the singular values of the estimate are not far from the singular values of $\boldsymbol{B}_S^*$, thresholding can work well. We give a rigorous proof in the following. Most parts of the proof follow the arguments in the proof of Theorem 7.

Consider the decomposition

$$\boldsymbol{\beta}^{(t)*} = \boldsymbol{P}_{\overline{\boldsymbol{A}}}\boldsymbol{\beta}^{(t)*} + \boldsymbol{P}_{\overline{\boldsymbol{A}}^\perp}\boldsymbol{\beta}^{(t)*} = \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\boldsymbol{\beta}^{(t)*} + \overline{\boldsymbol{A}}^\perp(\overline{\boldsymbol{A}}^\perp)^\top\boldsymbol{\beta}^{(t)*} := \overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)*} + \boldsymbol{\delta}^{(t)*}, \quad \forall t \in S.$$

Denote $\overline{\boldsymbol{\Theta}}_S = \{\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)*}\}_{t \in S}$, $\overline{\boldsymbol{B}}_S = \{\overline{\boldsymbol{A}}\boldsymbol{\theta}_{\overline{\boldsymbol{A}}}^{(t)*}\}_{t \in S}$, $\overline{\boldsymbol{B}} = \overline{\boldsymbol{A}}\overline{\boldsymbol{\Theta}} = (\overline{\boldsymbol{B}}_S \quad \boldsymbol{0}_{p \times |S^c|})$, $\boldsymbol{B}_S^* = \{\boldsymbol{\beta}^{(t)*}\}_{t \in S}$, $\widetilde{\boldsymbol{B}}_S = \{\widetilde{\boldsymbol{\beta}}^{(t)}\}_{t \in S}$, $\widehat{\boldsymbol{B}}_S = \{\prod_R(\widetilde{\boldsymbol{\beta}}^{(t)})\}_{t \in S}$, $\widehat{\boldsymbol{B}} = \{\prod_R(\widetilde{\boldsymbol{\beta}}^{(t)})\}_{t=1}^T$, and $\boldsymbol{D}_S^* = \{\boldsymbol{\delta}^{(t)*}\}_{t \in S}$. By a similar argument as in the proof of Theorem 7,

$$\|\overline{\boldsymbol{B}}_S/\sqrt{T} - \boldsymbol{B}_S^*/\sqrt{T}\|_2 = \frac{1}{\sqrt{T}}\|\boldsymbol{D}_S^*\|_2$$

123

$$\lesssim \bar\zeta h \cdot \left[\frac{\sigma_{\max}(\boldsymbol{D}_S^*)}{\sqrt{r}\sigma_{\min}(\boldsymbol{D}_S^*)} \wedge 1\right]$$

$$\lesssim \frac{\bar\zeta}{\sqrt{r}}\left[\frac{\sigma_{\max}(\boldsymbol{D}_S^*)}{\sigma_{\min}(\boldsymbol{D}_S^*)} \wedge \sqrt{r}\right] \cdot h \cdot \frac{\min_{t\in S}\zeta^{(t)}}{\sqrt{(p+\log T)/n}}$$

$$\lesssim \frac{\bar\zeta}{\sqrt{r}}$$

$$\le \widetilde{C}\sqrt{\frac{p+\log T}{n}},$$

when $\min_{t\in S}\zeta^{(t)} \cdot h\left[\frac{\sigma_{\max}(\boldsymbol{D}_S^*)}{\sigma_{\min}(\boldsymbol{D}_S^*)} \wedge \sqrt{r}\right] \lesssim \sqrt{\frac{p+\log T}{n}}$. Because $n \gtrsim p + \log T$, we have

$$\|\overline{\boldsymbol{B}}_S/\sqrt{T} - \boldsymbol{B}_S^*/\sqrt{T}\|_2 \le \widetilde{C}\sqrt{\frac{p+\log T}{n}}, \tag{71}$$

w.p. at least $1-e^{-C'(p+\log T)}$. Since $R = \texttt{quantile}(\{\widetilde{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T, 1-\bar\epsilon)$, we have $\prod_R(\widetilde{\boldsymbol{\beta}}^{(t)}) = \widetilde{\boldsymbol{\beta}}^{(t)}$ for $\bar\epsilon$-proportion of S among all $t \in S$. In addition, by the fact that $\max_{t\in S}\|\widetilde{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \sqrt{\frac{p+\log T}{n}}$, w.p. at least $1 - e^{-C'(p+\log T)}$, the assumption $\min_{t\in S}\zeta^{(t)} \gtrsim \sqrt{\frac{p+\log T}{n}}$, and $\bar\epsilon \ge \epsilon$, we have

$$R \le \max_{t\in S}\|\widetilde{\boldsymbol{\beta}}^{(t)}\|_2 \lesssim \max_{t\in S}\zeta^{(t)}, \tag{72}$$

w.p. at least $1 - e^{-C'(p+\log T)}$. Therefore,

$$\|\widehat{\boldsymbol{B}}_S/\sqrt{T} - \widetilde{\boldsymbol{B}}_S/\sqrt{T}\|_2 \le \|\widehat{\boldsymbol{B}}_S/\sqrt{T} - \widetilde{\boldsymbol{B}}_S/\sqrt{T}\|_{\mathrm{F}} = \sqrt{\frac{1}{T}\sum_{t\in S}\|\Pi_R(\widetilde{\boldsymbol{\beta}}^{(t)}) - \widetilde{\boldsymbol{\beta}}^{(t)}\|_2^2} \le C\sqrt{\bar\epsilon}R.$$

By Lemma 5.39 in Vershynin (2010), we have

$$\|\widetilde{\boldsymbol{B}}_S/\sqrt{T} - \boldsymbol{B}_S^*/\sqrt{T}\|_2 \le C''\sqrt{\frac{p+T}{nT}} \le C\sqrt{\frac{p+\log T}{n}},$$

w.p. at least $1 - e^{-C'(p+\log T)}$. By the equation (71), and the single-task rate, w.p. at least $1 - e^{-C'(p+\log T)}$,

$$\|\widehat{\boldsymbol{B}}_S/\sqrt{T} - \boldsymbol{B}_S^*/\sqrt{T}\|_2 \le \|\widehat{\boldsymbol{B}}_S/\sqrt{T} - \widetilde{\boldsymbol{B}}_S/\sqrt{T}\|_2 + \|\widetilde{\boldsymbol{B}}_S/\sqrt{T} - \boldsymbol{B}_S^*/\sqrt{T}\|_2$$

$$\le C\sqrt{\bar\epsilon}R + C\sqrt{\frac{p+\log T}{n}}.$$

Moreover, by the projection, $\|\widehat{\boldsymbol{B}}_{S^c}/\sqrt{T}\|_2 \le \|\widehat{\boldsymbol{B}}_{S^c}/\sqrt{T}\|_{\mathrm{F}} \le \sqrt{\frac{|S^c|}{T}} \cdot R \le \sqrt{\bar\epsilon}R$, w.p. at least $1 - e^{-C'(p+\log T)}$. Hence w.p. at least $1 - e^{-C'(p+\log T)}$,

$$\|\widehat{\boldsymbol{B}}/\sqrt{T} - \overline{\boldsymbol{B}}/\sqrt{T}\|_2 \le \|\widehat{\boldsymbol{B}}_S/\sqrt{T} - \overline{\boldsymbol{B}}_S/\sqrt{T}\|_2 + \|\widehat{\boldsymbol{B}}_{S^c}/\sqrt{T}\|_2$$

$$\le \|\widehat{\boldsymbol{B}}_S/\sqrt{T} - \boldsymbol{B}_S^*/\sqrt{T}\|_2 + \|\boldsymbol{B}_S^*/\sqrt{T} - \overline{\boldsymbol{B}}_S/\sqrt{T}\|_2 + \|\widehat{\boldsymbol{B}}_{S^c}/\sqrt{T}\|_2$$

$$\le (\widetilde{C} + C)\sqrt{\frac{p+\log T}{n}} + (C+1)\sqrt{\bar\epsilon}R.$$

Then by Weyl's inequality, for all $r' \leq r$,

$$\sigma_{r'}(\widehat{\boldsymbol{B}}/\sqrt{T}) \geq \sigma_{r'}(\overline{\boldsymbol{B}}/\sqrt{T}) - (\widetilde{C} + C)\sqrt{\frac{p + \log T}{n}} - (C+1)\sqrt{\bar{\epsilon}}R$$

$$\geq \sigma_{r'}(\boldsymbol{B}_S^*/\sqrt{T}) - (2\widetilde{C} + C)\sqrt{\frac{p + \log T}{n}} - (C+1)\sqrt{\bar{\epsilon}}R$$

$$\geq \frac{c}{\sqrt{r}}\bar{\zeta} - (2\widetilde{C} + C)\sqrt{\frac{p + \log T}{n}} - (C+1)\sqrt{\bar{\epsilon}}R$$

$$\geq 2(2\widetilde{C} + C)\sqrt{\frac{p + \log T}{n}} + 2(C+1)\sqrt{\bar{\epsilon}}R,$$

w.p. at least $1 - e^{-C'(p + \log T)}$, when $n \geq \frac{36(2\widetilde{C}+C)^2}{c^2\bar{\zeta}^2}r(p + \log T)$ and $\frac{c}{2\sqrt{r}}\bar{\zeta} \geq 3(C+1)\sqrt{\bar{\epsilon}}R$. The second condition holds because of $\bar{\epsilon} \leq \frac{c'}{r}\left(\frac{\bar{\zeta}}{\max_{t \in S}\zeta^{(t)}}\right)^2$ and (72).

On the other hand, by Weyl's inequality, for all $r' > r$, w.p. at least $1 - e^{-C'(p + \log T)}$,

$$\sigma_{r'}(\widehat{\boldsymbol{B}}/\sqrt{T}) \leq \sigma_{r'}(\overline{\boldsymbol{B}}/\sqrt{T}) + (\widetilde{C}+C)\sqrt{\frac{p + \log T}{n}} + (C+1)\sqrt{\bar{\epsilon}}R = (2\widetilde{C}+C)\sqrt{\frac{p + \log T}{n}} + (C+1)\sqrt{\bar{\epsilon}}R,$$

because $\sigma_{r'}(\overline{\boldsymbol{B}}/\sqrt{T}) = 0$ when $r' > r$. Finally, by setting $T_1 \in (2\widetilde{C} + C, 4\widetilde{C} + 2C]$ and $T_2 \in (C + 1, 2C + 2]$, we complete the proof.