Enhancing Federated Learning Efficiency: Accelerating Expectation-Maximization for Non-IID Data

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

October 3, 2024

Abstract

Federated learning (FL) has become a promising framework for training machine learning models on decentralized data while preserving privacy. One fundamental challenge in FL is handling non-independent and identically distributed (non-i.i.d.) data across clients, which complicates the convergence of learning algorithms. In this paper, we examine the convergence properties of the Expectation-Maximization (EM) algorithm in the federated setting, specifically under a Federated Mixture of Linear Regressions (FMLR) model. We provide a detailed analysis of the algorithm's performance across different regimes of the number of clients and the signal-to-noise ratio (SNR), demonstrating that the EM algorithm achieves minimax convergence under certain initialization conditions. We also show that federated EM converges more efficiently than centralized EM, particularly in scenarios with highly heterogeneous data. These theoretical results are supported by experiments on synthetic datasets. Our findings open new directions for improving optimization algorithms in federated systems.

Keywords: Federated Learning, Expectation-Maximization, Convergence Analysis, Mixture Models, Data Heterogeneity

1 Introduction

In today's era of big data, the ability to train models on decentralized data sources is increasingly essential due to privacy concerns and the high cost of data transfer. Federated learning (FL) provides an effective solution by enabling collaborative training of models without requiring raw data to be shared among clients. However, one of the major obstacles in FL is dealing with non-independent and identically distributed (non-i.i.d.) data. Each client in a federated system often generates data from a different distribution, posing significant challenges for traditional learning algorithms.

The Expectation-Maximization (EM) algorithm is a popular method for dealing with latent variable models, such as the mixture of linear regressions (MLR), in centralized settings. The federated version of the EM algorithm, when applied to a Federated Mixture of Linear Regressions (FMLR) model, offers a promising approach for addressing statistical heterogeneity across clients. In this paper, we explore the convergence behavior of the federated EM algorithm under different client and data heterogeneity settings. Specifically, we study how the number of clients and the signal-to-noise ratio (SNR) influence the convergence rate of the algorithm.

While leveraging large datasets can enhance estimation accuracy in today's digital age, the challenges of curating these datasets can complicate practical implementation. Firstly, the computational and time cost of storing and processing large centrally-stored datasets can be prohibitive. Secondly, there are privacy concerns in central collection and storage of raw data. When using individual level data, in particular, it is difficult not only to obtain the data, but also to ensure each client's privacy. In order to resolve both of these issues, recent machine learning efforts have been directed towards distributed storage of data with central processing that can still levarage the larger volume of data to provide more accurate estimation, termed as Federated Learning [MMR+17].

Federated Learning (FL) is a machine learning paradigm that enables clients to collaboratively train a global model without sharing the raw data. This approach decouples the model training from the raw data, which preserves the privacy of the clients and reduces computational costs [MMR+17]. One fundamental challenge in FL is the presence of non-independent and identically (non-i.i.d) distributed data. The non-i.i.d data presents challenges in achieving the original goal of training a single global model on the union of all clients' datasets [KMA+21]. One common cause of non-i.i.d data is that each client may have a different data generating process (denoted as P_j for client j). In other words, $P_j \neq P_{j'}$ for clients $j \neq j'$ [YFD+23]. See Section 2 for more details on existing approaches to address issues with non-i.i.d data.

In this paper, in order to quantify the level of data heterogeneity among clients, we assume the Federated Mixture of Linear Regressions (FMLR) model with K components. More specifically, we assume each client j has a latent variable $z_j \in [K]$ which serves to indicate that the data for client j is distributed according to the z_j-th component of a mixture of linear regressions. The data for the j-th client, conditional on z_j, is i.i.d. In this setup, all of the data heterogeneity is captured by the different latent variables of the clients. In the traditional centralized machine learning setting (where the number of clients is equal to 1), the Expectation-Maximization (EM) algorithm has been one of the most popular methods for studying the mixture of linear regressions (MLR) [DV89, FS10]. Thus, one natural question to pose is: ***Can the EM algorithm be used successfully to analyze the MLR model under the federated setting?***

1.1 Main Results

In this section we provide an overview of our results and the major contributions of this paper to the Federated Learning literature as well as the mixture model literature. In Federated Learning, the number of clients is often a limiting hyperparameter that may be exogenous to all other parameters of the model. As such, we fully characterize the behavior of the EM algorithm under all regimes of the number of clients, m. Another parameter that can often be limiting in the ability of various learning models to achieve the minimax rate of convergence is the signal-to-noise ratio (SNR). In this paper, we show given that the SNR is of order $\sqrt{K}$, a well-initialized EM algorithm always converges to within the minimax distance of the ground truth. Moreover, if the number of clients, m, grows exponentially in the number of data points for each client, n, the EM algorithm only requires a constant number of iterations to converge. In all settings, our results directly generalize that of [RGO23], which shows convergence rates only for the 2-mixture model ($K = 2$) case. We use experiments on synthetic datasets to validate our theoretical results and investigate tightness of our assumptions.

A point of contention in the federated learning literature has been the role of the maximum separation of the mixture models. Through a refined analysis of the EM algorithm, we are able to show that contrary to popular belief, the maximum separation between elements of the mixture model need not always help the EM algorithm converge faster. Furthermore, our results show that the EM converges much faster in the Federated Learning setting when compared to the traditional centralized setting such as [KC20b]. To the best of our knowledge, this is the first paper establishing the rates of convergence of the EM algorithm for the FMLR model.

Our main contributions are as follows:

- We analyze the performance of the federated EM algorithm under varying numbers of clients and SNR conditions, demonstrating that it achieves minimax convergence when appropriately initialized.

- We extend existing results in the literature by providing convergence rates for a general K-component mixture model, generalizing prior results that focused only on the case where $K = 2$.

- We show that the federated EM algorithm exhibits significantly faster convergence compared to its centralized counterpart, particularly in high-heterogeneity scenarios.

- Through synthetic experiments, we validate the theoretical results and explore the tightness of our assumptions, showing that the EM algorithm can achieve efficient convergence even with substantial data heterogeneity.

The remainder of this paper is organized as follows. Section 2 provides an overview of the literature and existing results. Section 3 sets up the FMLR model we study and introduces all relevant notations. Section 4 presents and discusses the main results. We conclude in Section 5.

2 Related Work

2.1 Data Heterogeneity

There has been a lot of recent work on mitigating the issue of data heterogeneity in Federated Learning. In the prototypical setting, the goal is to learn a single global model that minimizes the empirical risk over the union of datasets from all the clients. The non-i.i.d nature of these datasets hinders the convergence of federated learning optimization algorithms [KMA+21]. There is a line of work focusing on designing new optimization algorithms to accelerate the convergence with non-i.i.d data [WLL+20, KKM+20, LSZ+20, WTBR19, LSL+19, RCZ+20]. For example, [KKM+20] proposed an algorithm called SCAFFOLD which estimates the drift of clients' updates from the global update due to data dependency and adjusts the estimator accordingly. [LSZ+20] proposed FedProx which adds a proximal term to the original objective function in order to restrict the movement of the local updates to be closer to the global update. [RCZ+20] proposed a collection of federated versions of adaptive central optimizers.

Another line of work focuses on multi-model learning. In many applications, treating all datasets equally to train a single global model is not reasonable. For example, in the next-word prediction task by Google Keyboard [HRM+18], different clients may use different languages, so personalization on each device is very crucial to the performance of the model. One popular framework, termed Personalized Federated Learning, aims to learn individual models for each client [HR20, HCZ+21, ALW20, LLZ+20]. For example, [SCST17] formulated a Federated Multi-Task Learning (MTL) framework to learn separate models for each client. [LHBS21] proposed an algorithm to simultaneously learn local and global models via a global-regularized MTL framework. [FMO20] proposed a Model-Agnostic Meta-Learning (MAML) framework to personalize models for each client.

Under multi-model learning, another important framework is the Clustered Federated Learning (Clustered FL) framework which partitions clients into different clusters where clients within the same cluster share the same optimal model [MMRS20, SMS20, MNB+21, MLZ+22, WHK+23]. For example, [GCYR20] assumed there are K underlying distinct data distributions and the m clients can be partitioned into the K disjoint clusters based on the parameters that determine their data generating process. [LXS+23] proposed a multi-center aggregation mechanism to group clients into different clusters by minimizing the distance between each client's local model and the nearest global model.

2.2 Mixture Models and EM Algorithm

In order to better capture the statistical heterogeneity, some literature in federated learning assumes that the data comes from a mixture model [RGO23, WZY+23]. For example, [MNB+21] assumed the true data generating process is a mixture of distributions such that each client generates data from one unique distribution in the mixture model. [SXY22] formulated a similar assumption with the added condition that each of the K distributions in the mixture is a linear regression model that is uniquely identified by a single hyperparameter. They focused specifically on the case when the local data volume is highly unbalanced. The authors proposed to estimate the parameters of the mixture model using a two-phase Federated Learning algorithm which consists of a coarse model estimation phase and a fine-tuning phase. [RGO23] assumed each client has data from one of the two regressions in a symmetric Mixture of Linear Regression Model. This is equivalent to the data generating process considered in this paper with $K = 2$. As we discuss in Section 4, our results contain the results of this paper as a special case.

Among different techniques to learn a mixture model such as the spectral method [KSV05] and Markov Chain Monte Carlo (MCMC) [Gew07], the EM algorithm is a popular one, specifically because of its computational efficiency. Recent advances have been made in establishing theoretical properties relating to the convergence of the EM algorithm both in the centralized setting [KC20b, KYB19, DTZ17, KC20a, YCS14, GK20, ZLS20] and in the federated setting [WZY+23, DFMR21, TWF23, MNB+21, LXS+23]. For example, in the traditional machine learning setting, [BWY17] established a theoretical connection between the analysis of the population EM and the empirical EM. They leveraged fixed-point analysis methods to identify regions of convergence for the EM algorithm in the 2-mixture model setting. [XHM16] and [WZ21] analyzed the EM algorithm for the symmetric two-component Gaussian Mixture Model (GMM) and provided minimax rates of convergence. [KQC+19] established a global convergence result for the two-component MLR, following which [KHC21] studied the convergence rates of the same model under all SNR regimes. For the centralized multiple-component MLR, [KC20b] provided a convergence result conditional on good initializations and identifiable lower bounds on the SNR.

In Federated Learning, the EM algorithm has been applied in a few different contexts. [DFMR21] proposed the first extension of EM to federated learning which supports communication compression. [WZY+23] used the EM algorithm to address the case when the data distribution of each client is a mixture of Gaussians. [MNB+21] also used an EM-like algorithm to study a similar problem but when the data distribution of each client is a mixture of unknown underlying distributions. [TWF23] introduced a Federated Gradient EM algorithm to solve a more complicated mixture of distributions with the existence of outliers.

3 Problem Setup and EM Algorithm

In this section, we introduce the FMLR model and provide a brief overview of the EM Algorithm.

3.1 Notation

We start by first introducing relevant notation.

- $x \in \mathbb{R}^d$: collection of d features, where d is known and fixed.
- $y \in \mathbb{R}$: observed variable.
- K: number of mixture components, known and fixed.
- m: number of clients.
- n: number of data points for each client.
- $j \in [1, m]$: index denoting client.
- $i \in [1, n]$: index denoting data point.
- $[n] = \{1, \ldots, n\}$
- $x_{[n]} = \{x_1, \ldots, x_n\}$
- $f_\Theta(\cdot)$: probability density function of a continuous random variable with parameter Θ
- $g_\Theta(\cdot)$: probability mass function of a discrete random variable with parameter Θ

We use $\|\cdot\|$ to denote the Euclidean distance. θ_k^* denotes the k-th ground truth coefficient vector. For the one-step analysis, we use θ_k and θ_k^+ to denote the current and the next estimate of θ_k^* for the population EM. We use $\widetilde{\theta}_k$ and $\widetilde{\theta}_k^+$ to denote the current and the next estimate of θ_k^* for the empirical EM.

$$\Delta_{\max} := \max_{k \neq k'} \|\theta_k^* - \theta_{k'}^*\|$$

denotes the maximum separation, and

$$\Delta_{\min} := \min_{k \neq k'} \|\theta_k^* - \theta_{k'}^*\|$$

denotes the minimum separation between any two true coefficient vectors. Then the signal-to-noise ratio (SNR) is given by

$$\frac{\Delta_{\min}}{\sigma}$$

where σ is the variance of the noise (specified in Assumption 1 below). We use $\mathbb{E}_{\mathcal{D}_k^*}[\cdot]$, to denote the expectation with respect to the joint distribution of (x, y) conditional on $z = k$. That is,

$$\mathbb{E}_{\mathcal{D}_k^*}[\cdot] = \mathbb{E}[\cdot \mid z = k]$$

3.2 The FMLR model

In this section we set up the Federated Mixture of Linear Regressions model. Suppose there are m clients, where j denotes the j-th client. Each of the m clients has a latent variable $z_j \in [K]$ and observes n pairs of data points $\{(x_i^j, y_i^j) : i = 1, \ldots, n\}$ generated from the z_j-th linear regression. Sequentially, the data generating process can be explained by Algorithm 1. This model inherently exhibits a clustered structure in the sense that clients can be grouped into K disjoint clusters based on their true latent variable z_j. Note that x_i^j and ε_i^j are independent of the cluster membership while y_i^j depends on the cluster structure through z_j. Furthermore, it is important to observe that

Algorithm 1 The FMLR Algorithm

Input: K, m, n, and $\theta^* = [\theta_1^*, \dots, \theta_K^*]$
Output: $\{x_i^j, y_i^j\}_{i=1,j=1}^{i=n,j=m}$

```
for j = 1, ... m do
    Sample z_j ~ Uniform([1, K])                 {latent variable for each client}
    for i = 1,... n do
        Sample x_i^j ~_{i.i.d} P_x^j                         {predictor variables}
        Sample ε_i^j ~_{i.i.d} P_ε^j                                     {noise}
        Generate y_i^j = ⟨x_i^j, θ*_{z_j}⟩ + ε_i^j              {observed variables}
    end for
end for
```

for each client j, there are n pairs of $(x_i^j, y_i^j)_{i=1}^n$ sharing the same latent variable z_j, which means $(x_i^j, y_i^j, z_j)_{i=1}^n$ are not i.i.d.

The FMLR model specifically takes data heterogeneity arising from a *concept shift* into consideration [KMA+21]. A concept shift occurs when $P_j(x,y) \neq P_{j'}(x,y)$ for $j \neq j'$ is due to $P_j(y|x) \neq P_{j'}(y|x)$ even if $P_j(x)$ is the same for all j. This can be understood in the context of user preferences as when presented with identical feature vectors, users may label items differently based on personal preferences that can be categorized based on more general features like regional or demographic variations.

3.3 EM Algorithm

While the EM algorithm has been a popular algorithm for learning latent variables in many statistical problems, here we focus on understanding the algorithm in the context of MLR models. We start by assuming the FMLR data generating model as described in Section 3.2.

In order to estimate the parameters $\{\theta_j^*\}_{j=1}^K$, the EM algorithm approximates maximizing the finite-sample log-likelihood:

$$\ell_m(\Theta) = \frac{1}{m}\sum_{j=1}^m \log \int_{\mathcal{Z}} f_\Theta(x_{[n]}^j, y_{[n]}^j, z_j) dz_j \tag{1}$$

which is generally non-concave and thus computationally very difficult to solve for. Instead, the EM algorithm uses an auxiliary function to lower bound the log-likelihood. This lower bound is given by Q-function. More specifically, let $g_\Theta(z|x_{[n]}, y_{[n]})$ denote the conditional probability mass function of z given $(x_{[n]}, y_{[n]})$. The finite-sample version of Q-function is given by

$$\begin{aligned} &Q_m(\Theta'|\Theta) \\ &= \frac{1}{m}\sum_{j=1}^m \int_{\mathcal{Z}} g_\Theta(z_j|x_{[n]}^j, y_{[n]}^j) \log f_{\Theta'}(x_{[n]}^j, y_{[n]}^j, z_j) dz_j \end{aligned}$$

Given an initialization $\Theta^0 \subset \Omega$, the empirical EM algorithm updates it by

$$\Theta^{(t+1)} = \arg\max_{\Theta \subset \Omega} Q_m(\Theta|\Theta^t), \quad t = 0, 1, \dots.$$

The analysis of the empirical EM and Q_m-function is based on the analysis of the population EM and the population Q-function [BWY17]. Formally, the population Q-function is defined as

$$Q(\Theta'|\Theta) = \int_{\mathcal{X}^n \times \mathcal{Y}^n} (\int_{\mathcal{Z}} g_{\Theta}(z|x_{[n]}, y_{[n]}) \cdot \log f_{\Theta'}(x_{[n]}, y_{[n]}, z) dz) f_{\Theta^*}(x_{[n]}, y_{[n]}) dx_{[n]} dy_{[n]}$$

Different from the empirical EM, the population level analysis considers the case when m goes to infinity. In other words, the computation of the population EM assumes the access to the joint distribution f_{Θ^*}.

In order to analyze the EM algorithm, we must make some assumptions on the underlying distributions. In this paper, we consider the following standard assumption in EM.

Assumption 1 (DGP)**.** *Assume that the covariates, x and noise ε follow standard Gaussian distributions. That is,*

$$x \sim \mathcal{N}(0, I_d),$$
$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

where σ is a constant.

We can now re-write both the population and empirical EM iterations as shown in the following two propositions. The proofs of all the results are deferred to the Appendix.

Proposition 1 (Population EM)**.** *Suppose Assumption 1 holds and $\{(x_i, y_i)\}_{i=1}^n$ are generated by the MLR model as given in Algorithm 1. Then one iteration of the population EM, $\forall k$, given the current estimates $\{\theta_k\}_{k=1}^K$, is given by*

$$\textit{E-Step:} \quad w(\theta)_k = \frac{\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i^j - \langle x_i^j, \theta_k \rangle)^2)}{\sum_{l=1}^K \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i^j - \langle x_i^j, \theta_l \rangle)^2)}$$
$$\textit{M-Step:} \quad \theta_k^+ = \mathbb{E}[w(\theta)_k \sum_{i=1}^n x_i x_i^T]^{-1} \mathbb{E}[w(\theta)_k \sum_{i=1}^n y_i x_i^T]$$

Proposition 2 (Empirical EM)**.** *Suppose Assumption 1 holds and $\{(x_i, y_i)\}_{i=1}^n$ are generated by the MLR model as given in Algorithm 1 for each of the m clients. Then, one iteration of the empirical EM, $\forall k$, given the current estimates $\{\widetilde{\theta}_k\}_{k=1}^K$, is given by*

$$\textit{E-Step:} \quad w(\widetilde{\theta})_k^j = \frac{\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i^j - \langle x_i^j, \widetilde{\theta}_k \rangle)^2)}{\sum_{l=1}^K \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i^j - \langle x_i^j, \widetilde{\theta}_l \rangle)^2)}$$
$$\textit{M-Step:} \quad \widetilde{\theta}_k^+ = (\sum_{j=1}^m w(\widetilde{\theta})_k^j \sum_{i=1}^n x_i^j x_i^{jT})^{-1} \sum_{j=1}^m w(\widetilde{\theta})_k^j \sum_{i=1}^n y_i^j x_i^j$$

4 Main Results

In this section, we present our main theoretical results. The analysis of the empirical EM relies on the population-level analysis by standard limiting arguments. Thus, it is natural to break up the main results into two distinct theorems, one for the population EM and one for the empirical

EM. Through these two results, we can better understand where the gains in rates of convergence appear in the federated setting.

We start by making an additional assumption on initialization that ensures identifiability of the solution.

Assumption 2 (Identifiability). *The initial estimates, $\{\theta_k\}_{k=1}^K$ are chosen such that*

$$\|\theta_k - \theta_k^*\| \leq \alpha \Delta_{\min} \ \forall \ k$$

where $\alpha \in (0, 1/4)$ is a constant.

We can now establish the uniform convergence result for the population EM, as shown in Theorem 1 below.

Theorem 1 (Uniform consistency). *Suppose Assumptions 1 and 2 hold. If $n = \Omega(\frac{1}{C_\alpha}(\log K + \log(\sigma + \Delta_{\max})))$ and SNR $\geq \mathcal{O}(\sqrt{K})$ where C_α is a constant depending on α, then, the estimates generated after one iteration of the Population EM algorithm (as given in Proposition 1) satisfy*

$$\max_{k \in [K]} \|\theta_k^+ - \theta_k^*\| \leq \mathcal{O}\left(\exp(-C_\alpha n) \max_{k \in [K]} \|\theta_k - \theta_k^*\|\right)$$

From this result, we note that conditional on initializing the algorithm reasonably well, one step of the population EM will converge to the true centers within precision error that depends on the magnitude of the problem.

In order to complete our analysis of the federated EM algorithm, we now present the convergence of the empirical EM algorithm. Theorem 2 shows that based on different assumptions on m, the algorithm converges with different rates.

Theorem 2 (Empirical uniform consistency). *Suppose Assumptions 1 and 2 hold. If $n = \Omega(\frac{1}{C_\alpha}(\log K + \log(\sigma + \Delta_{\max})))$ and SNR $\geq \Omega(\sqrt{K})$ where C_α is a constant depending on α, then, with probability at least $1 - \delta$, the estimates generated after one iteration of the empirical EM algorithm (as given in Proposition 2) concentrate polynomially in the number of clients, m, and number of data points, n:*

$$\max_{k \in [K]} \|\widetilde{\theta}_k^+ - \theta_k^*\| \precsim \begin{cases} \sqrt{\frac{1}{mn}}(\sigma + \frac{\Delta \max}{\exp(n)}) & \text{if } \Omega(\exp(n)) \leq m \\ \sqrt{\frac{1}{mn}}(\sigma + \frac{\Delta_{\max}}{\sqrt{n}}) & \text{if } \Omega(\sqrt{n}) \leq m \leq \mathcal{O}(\exp(n)) \\ \sqrt{\frac{1}{mn}}(\sigma + \Delta_{\max}) + \sqrt{\frac{1}{m\sqrt{n}}} & \text{if } m < \mathcal{O}(\sqrt{n}) \end{cases}$$

As we see in the statement of the theorem, the precise rate of convergence depends on the relationship between the two key variables m and n. The dependency of the convergence rate on all other parameters is provided in the proof of the theorem in the Appendix. A more detailed version of the theorem is also provided in the Appendix.

Theorem 2 also shows how the maximum separation $\Delta_{\max}$ affects the convergence rate in different regime of m. As opposed to the existing literature where either the effect of $\Delta_{\max}$ is ignored in the analysis of the error rate [KC20b] or the error rate is only guaranteed to be $\widetilde{\mathcal{O}}(\sqrt{\frac{1}{mn}}(\sigma + \Delta_{\max}))$ [BWY17, KYB19, ZLS20, YYS17, RGO23], we did a more refined analysis on $\Delta_{\max}$. When m grows at least exponentially in n, $\Delta_{\max}$ is controlled by $\exp(-n)$, and when m grows polynomially in n, $\Delta_{\max}$ is controlled by $\frac{1}{\sqrt{n}}$. Thus, in both two cases, the effect of $\Delta_{\max}$ is negligible if n is large enough. This means that exact recovery is guaranteed when $\sigma \to 0$. However, if m is even less than $\mathcal{O}(\sqrt{n})$, the effect of $\Delta_{\max}$ cannot be canceled out. Intuitively, the larger the $\Delta_{\max}$ is, the easier the clustering problem should be since the regression coefficients being further away from each other should allow clients to identity their cluster membership more easily. However, $\Delta_{\max}$ might also be related to the scale of the problem. We conjecture that all the data is used in every iteration of the algorithm and thus, individual center-level accuracy is sacrificed for worst-case error. Further investigation into this phenomenon across other clustering algorithms is required to better understand the trade-off between individual and collection level error. This phenomenon has also been observed and investigated in the experiments.

Comparing the result with that of the population EM convergence rate from Theorem 1 we can identify that the error between the empirical and population versions of the algorithm is larger than the error between the population level algorithm and the ground truth. This is in line with classical convergence results for clustering algorithms in the centralized setting. The following corollary states the number of iterations required to reach the desired error ε.

Corollary 1. *Let T be the number of iterations and $\varepsilon = \widetilde{\mathcal{O}}(\sqrt{\frac{1}{mn}(\sigma + \Delta_{\max})})$. Assume mn is large enough such that $\varepsilon \leq \frac{\alpha\Delta_{\min}}{2}$. Then after T iterations, either $\max_{k\in[K]} \|\widetilde{\theta}_k^{(t)} - \theta_k^*\| \leq \varepsilon$ for some $t = 0, 1, \ldots, T-1$, or $\max_{k\in[K]} \|\widetilde{\theta}_k^{(T)} - \theta_k^*\| \leq \varepsilon$ with probability $1-\delta$ where T is given by*

$$\begin{cases} \frac{1}{C_\alpha n}\mathcal{O}(\log(\frac{\max_{k\in[K]} \|\widetilde{\theta}_k^{(0)} - \theta_k^*\|}{\varepsilon})) & \text{if } \Omega(\exp(n)) \leq m \\ \mathcal{O}(1) & \text{if } \Omega(\sqrt{n}) \leq m \leq \mathcal{O}(\exp(n)) \\ \mathcal{O}(\log(\frac{\max_{k\in[K]} \|\widetilde{\theta}_k^{(0)} - \theta_k^*\|}{\varepsilon})) & \text{if } m < \mathcal{O}(\sqrt{n}) \end{cases}$$

As we can see from the above corollary, the estimates $\{\widetilde{\theta}_k^+\}$ are guaranteed to converge within the minimax distance of the ground truth. Specifically, when $\Omega(\sqrt{n}) \leq m \leq \mathcal{O}(\exp(n))$, the EM algorithm only requires a constant number of iterations to converge. Compared with the performance of the EM algorithm in the classical setting, the Federated EM exhibits a much faster convergence rate. The intuition is that data points on the same client all share the same latent variable, so there is no need to identity the cluster membership of every single data. This essentially makes the clustering task much easier. The constant number of iterations phenomenon has actually been observed in the experiments for even a broader range of m.

5 Conclusions and Discussions

This paper provides the first known convergence rates for the EM algorithm under all regimes of m and n in federated learning. The key findings show that when the data heterogeneity among clients can be described by the FMLR model, the well-initialized federated EM algorithm can find the true regression coefficients very fast. These results open up several different avenues for future work. For example, it is unclear what the tightest lower bound of SNR is for fast convergence. Additionally, all experiments showed a constant number of iterations required to reach convergence, which suggests that there is room to further loosen the assumptions of the theorems established in this paper. In fact, we conjecture that as long as the product $m \cdot n$ is large enough, the well-initialized EM algorithm will always converge after a constant number of iterations. Furthermore, we suspect that the initialization condition could also be improved.

References

[ALW20] Alekh Agarwal, John Langford, and Chen-Yu Wei. Federated residual learning. *arXiv preprint arXiv:2003.12880*, 2020.

[BWY17] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77 – 120, 2017.

[DFMR21] Aymeric Dieuleveut, Gersende Fort, Eric Moulines, and Geneviève Robin. Federated-em with heterogeneity mitigation and variance reduction. *Advances in Neural Information Processing Systems*, 34:29553–29566, 2021.

[DTZ17] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of em suffice for mixtures of two gaussians. In *Conference on Learning Theory*, pages 704–710. PMLR, 2017.

[DV89] Richard D De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8(3):227–245, 1989.

[FMO20] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.

[FS10] Susana Faria and Gilda Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010.

[GCYR20] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.

[Gew07] John Geweke. Interpretation and inference in mixture models: Simple mcmc works. *Computational Statistics & Data Analysis*, 51(7):3529–3550, 2007.

[GK20] Avishek Ghosh and Ramchandran Kannan. Alternating minimization converges super-linearly for mixed linear regression. In *International conference on artificial intelligence and statistics*, pages 1093–1103. PMLR, 2020.

[HCZ+21] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7865–7873, 2021.

[HR20] Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

[HRM+18] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

[KC20a] Jeongyeol Kwon and Constantine Caramanis. The em algorithm gives sample-optimality for learning mixtures of well-separated gaussians. In *Conference on Learning Theory*, pages 2425–2487. PMLR, 2020.

[KC20b] Jeongyeol Kwon and Constantine Caramanis. Em converges for a mixture of many linear regressions. In *International Conference on Artificial Intelligence and Statistics*, pages 1727–1736. PMLR, 2020.

[KHC21] Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the minimax optimality of the em algorithm for learning two-component mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1405–1413. PMLR, 2021.

[KKM+20] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[KMA+21] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

[KQC+19] Jeongyeol Kwon, Wei Qian, Constantine Caramanis, Yudong Chen, and Damek Davis. Global convergence of the em algorithm for mixtures of two component linear regression. In *Conference on Learning Theory*, pages 2055–2110. PMLR, 2019.

[KSV05] Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. In *International conference on computational learning theory*, pages 444–457. Springer, 2005.

[KYB19] Jason M Klusowski, Dana Yang, and WD Brinda. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, 65(6):3515–3524, 2019.

[LHBS21] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pages 6357–6368. PMLR, 2021.

[LLZ+20] Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

[LM00] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000.

[LSL+19] Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng. Variance reduced local sgd with lower communication complexity. *arXiv preprint arXiv:1912.12844*, 2019.

[LSZ+20] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[LXS+23] Guodong Long, Ming Xie, Tao Shen, Tianyi Zhou, Xianzhi Wang, and Jing Jiang. Multi-center federated learning: clients clustering for better personalization. *World Wide Web*, 26(1):481–500, 2023.

[MLZ+22] Jie Ma, Guodong Long, Tianyi Zhou, Jing Jiang, and Chengqi Zhang. On the convergence of clustered federated learning. *arXiv preprint arXiv:2202.06187*, 2022.

[MMR+17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[MMRS20] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

[MNB+21] Othmane Marfoq, Giovanni Neglia, Aurélien Bellet, Laetitia Kameni, and Richard Vidal. Federated multi-task learning under a mixture of distributions. *Advances in Neural Information Processing Systems*, 34:15434–15447, 2021.

[RCZ+20] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

[RGO23] Amirhossein Reisizadeh, Khashayar Gatmiry, and Asuman Ozdaglar. Em for mixture of linear regression with clustered data. *arXiv preprint arXiv:2308.11518*, 2023.

[SCST17] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

[SMS20] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

[SXY22] Lili Su, Jiaming Xu, and Pengkun Yang. Global convergence of federated learning for mixed regression. *Advances in Neural Information Processing Systems*, 35:29889–29902, 2022.

[TWF23] Ye Tian, Haolei Weng, and Yang Feng. Unsupervised federated learning: A federated gradient em algorithm for heterogeneous mixture models with robustness against adversarial attacks. *arXiv preprint arXiv:2310.15330*, 2023.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[WHK+23] Mariel Werner, Lie He, Sai Praneeth Karimireddy, Michael Jordan, and Martin Jaggi. Provably personalized and robust federated learning. *arXiv preprint arXiv:2306.08393*, 2023.

[WLL+20] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

[WTBR19] Jianyu Wang, Vinayak Tantia, Nicolas Ballas, and Michael Rabbat. Slowmo: Improving communication-efficient distributed sgd with slow momentum. *arXiv preprint arXiv:1910.00643*, 2019.

[WZ21] Yihong Wu and Harrison H Zhou. Randomly initialized em algorithm for two-component gaussian mixture achieves near optimality in $o(\sqrt{n})$ iterations. *Mathematical Statistics and Learning*, 4(3), 2021.

[WZY+23] Yue Wu, Shuaicheng Zhang, Wenchao Yu, Yanchi Liu, Quanquan Gu, Dawei Zhou, Haifeng Chen, and Wei Cheng. Personalized federated learning under mixture of distributions. In *International Conference on Machine Learning*, pages 37860–37879. PMLR, 2023.

[XHM16] Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. *Advances in Neural Information Processing Systems*, 29, 2016.

[YCS14] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pages 613–621. PMLR, 2014.

[YFD+23] Mang Ye, Xiuwen Fang, Bo Du, Pong C Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 56(3):1–44, 2023.

[YYS17] Bowei Yan, Mingzhang Yin, and Purnamrita Sarkar. Convergence of gradient em on multi-component mixture of gaussians. *Advances in Neural Information Processing Systems*, 30, 2017.

[ZLS20] Ruofei Zhao, Yuanzhi Li, and Yuekai Sun. Statistical convergence of the EM algorithm on Gaussian mixture models. *Electronic Journal of Statistics*, 14(1):632 – 660, 2020.

A Proof of Propositions

In this section, we derive the closed-form updates of the EM algorithm. Recall that we denote $f_\Theta(\cdot)$ as the probability density function of a continuous random variable and $g_\Theta(\cdot)$ as the probability mass function of a discrete random variable with parameter Θ. We also use $P(\cdot)$ to denote a generic probability density function of probability mass function without specifying any parameters.

A.1 Proof of Proposition 1: Population EM

The joint density of $(x_{[n]}, y_{[n]}, z)$ is given by

$$\begin{aligned} f_\Theta(x_{[n]}, y_{[n]}, z) &= P(z) f_\Theta(x_{[n]}, y_{[n]}|z) \\ &= \frac{1}{K} P(x_{[n]}) \prod_{i=1}^{n} \mathcal{N}(\langle x_i, \theta_z \rangle, \sigma^2) \\ &= \frac{1}{K} P(x_{[n]}) \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \langle x_i, \theta_z \rangle)^2\} \end{aligned}$$

and by the Law of Probability, the joint density of $(x_{[n]}, y_{[n]})$ is given by

$$\begin{aligned} f_\Theta(x_{[n]}, y_{[n]}) &= \sum_{l=1}^{K} f_\Theta(x_{[n]}, y_{[n]}, z = k) \\ &= \frac{1}{K} P(x_{[n]}) (\frac{1}{2\pi\sigma^2})^{n/2} \sum_{l=1}^{K} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \langle x_i, \theta_k \rangle)^2\} \end{aligned}$$

Then

$$\begin{aligned} g_\Theta(z|x_{[n]}, y_{[n]}) &= \frac{f_\Theta(x_{[n]}, y_{[n]}, z)}{f_\Theta(x_{[n]}, y_{[n]})} \\ &= \frac{P(x_{[n]}) (\frac{1}{2\pi\sigma^2})^{n/2} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \langle x_i, \theta_z \rangle)^2\} \frac{1}{K}}{\frac{1}{K} P(x_{[n]}) (\frac{1}{2\pi\sigma^2})^{n/2} \sum_{l=1}^{K} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \langle x_i, \theta_k \rangle)^2\}} \\ &= \frac{\exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \langle x_i, \theta_z \rangle)^2\}}{\sum_{l=1}^{K} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \langle x_i, \theta_k \rangle)^2\}} \\ &= w(\theta)_z \end{aligned}$$

Recall that

$$\begin{aligned} Q(\Theta'|\Theta) &= \int_{\mathcal{X}^n \times \mathcal{Y}^n} (\int_{\mathcal{Z}} g_\Theta(z|x_{[n]}, y_{[n]}) \log f_{\Theta'}(x_{[n]}, y_{[n]}, z) dz) f_{\Theta^*}(x_{[n]}, y_{[n]}) dx_{[n]} dy_{[n]} \\ &= \mathbb{E}_{x_{[n]}, y_{[n]}} [\int_{\mathcal{Z}} g_\Theta(z|x_{[n]}, y_{[n]}) \log f_{\Theta'}(x_{[n]}, y_{[n]}, z) dz] \end{aligned}$$

Notice that when maximizing $Q(\Theta'|\Theta)$ with respect to Θ', the only term that involves Θ' in $\log f_{\Theta'}(x_{[n]}, y_{[n]}, z)$ is $-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \langle x_i, \theta'_z \rangle)^2$. Therefore, it is equivalent to maximizing

$$\widehat{Q}(\Theta'|\Theta) = \mathbb{E}_{x_{[n]}, y_{[n]}} [\int_{\mathcal{Z}} g_\Theta(z|x_{[n]}, y_{[n]}) \{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \langle x_i, \theta'_z \rangle)^2\} dz]$$

$$
\begin{aligned}
&= \mathbb{E}_{x_{[n]},y_{[n]}}[\sum_{k=1}^{K} g_{\Theta}(z=k|x_{[n]},y_{[n]})\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\langle x_i,\theta'_k\rangle)^2\}] \\
&= \mathbb{E}_{x_{[n]},y_{[n]}}[\sum_{k=1}^{K} w(\theta)_k\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\langle x_i,\theta'_k\rangle)^2\}] \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\mathbb{E}_{y_{[n]}}[y_i^2]-\frac{1}{2\sigma^2}\sum_{k=1}^{K}\mathbb{E}_{x_{[n]},y_{[n]}}[w(\theta)_k\sum_{i=1}^{n}\langle x_i,\theta'_k\rangle^2]+\frac{1}{\sigma^2}\sum_{k=1}^{K}\mathbb{E}_{x_{[n]},y_{[n]}}[w(\theta)_k\sum_{i=1}^{n}y_i\langle x_i,\theta'_k\rangle]
\end{aligned}
$$

Taking derivative of $\widehat{Q}(\Theta'|\Theta)$ with respect to θ'_k and setting it to 0, we have

$$
\begin{aligned}
&-\frac{1}{\sigma^2}\mathbb{E}_{x_{[n]},y_{[n]}}[w(\theta_k)\sum_{i=1}^{n}x_i x_i^T\theta'_k]+\frac{1}{\sigma^2}\mathbb{E}_{x_{[n]},y_{[n]}}[w(\theta)_k\sum_{i=1}^{n}y_i x_i]=0 \\
&\Rightarrow \arg\max_{\theta'_k} Q(\Theta'|\Theta)=\mathbb{E}_{x_{[n]},y_{[n]}}[w(\theta_k)\sum_{i=1}^{n}x_i x_i^T]^{-1}\mathbb{E}_{x_{[n]},y_{[n]}}[w(\theta)_k\sum_{i=1}^{n}y_i x_i]
\end{aligned}
$$

Therefore, $\theta_k^+=\mathbb{E}_{x_{[n]},y_{[n]}}[w(\theta_k)\sum_{i=1}^{n}x_i x_i^T]^{-1}\mathbb{E}_{x_{[n]},y_{[n]}}[w(\theta)_k\sum_{i=1}^{n}y_i x_i]$.

A.2 Proof of Proposition 2: Empirical EM

Notice that the only difference between the derivation of the empirical EM and the population EM is that instead of using Q-function, we use the finite-sample Q_m-function, which is defined as

$$
Q_m(\widetilde{\Theta}'|\widetilde{\Theta})=\frac{1}{m}\sum_{j=1}^{m}\int_{\mathcal{Z}} g_{\widetilde{\Theta}}(z_j|x^j_{[n]},y^j_{[n]})\log f_{\widetilde{\Theta}'}(x^j_{[n]},y^j_{[n]},z_j)dz_j
$$

Similarly as before, this is equivalent to maximizing

$$
\begin{aligned}
\widehat{Q}_m(\widetilde{\Theta}'|\widetilde{\Theta}) &= \frac{1}{m}\sum_{j=1}^{m}\int_{\mathcal{Z}} g_{\widetilde{\Theta}}(z_j|x^j_{[n]},y^j_{[n]})\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^j-\langle x_i^j,\widetilde{\theta}'_z\rangle)^2\}dz \\
&= \frac{1}{m}\sum_{j=1}^{m}\sum_{k=1}^{K} g_{\widetilde{\Theta}}(z_j=k|x^j_{[n]},y^j_{[n]})\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^j-\langle x_i^j,\widetilde{\theta}'_k\rangle)^2\} \\
&= \frac{1}{m}\sum_{j=1}^{m}\sum_{k=1}^{K} w(\widetilde{\theta})_k^j\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^j-\langle x_i^j,\widetilde{\theta}'_k\rangle)^2\}
\end{aligned}
$$

Again, taking derivative of $\widehat{Q}_m(\widetilde{\Theta}'|\widetilde{\Theta})$ with respect to $\widetilde{\theta}'_k$ and setting it to 0, we get

$$
\widetilde{\theta}'_k=(\sum_{j=1}^{m}w(\widetilde{\theta})_k^j\sum_{i=1}^{n}x_i^j x_i^{jT})^{-1}(\sum_{j=1}^{m}w(\widetilde{\theta})_k^j\sum_{i=1}^{n}y_i^j x_i^j).
$$

B Proof of Theorem 1

This section proves the convergence rate for the population EM. We perform one step analysis. Suppose at the current step, we have estimates $\theta_1,\ldots,\theta_K$, and one step iteration of population EM generates new estimates $\theta_1^+, \ldots, \theta_K^+$. WLOG, we focus on θ_1^+.

$$\begin{aligned}
\theta_1^+ - \theta_1^* &= \mathbb{E}[w(\theta)_1 \sum_{i=1}^n x_i x_i^T]^{-1} \mathbb{E}[w(\theta)_1 \sum_{i=1}^n y_i x_i] - \theta_1^* \\
&= \mathbb{E}[w(\theta)_1 \sum_{i=1}^n x_i x_i^T]^{-1} \mathbb{E}[w(\theta)_1 \sum_{i=1}^n x_i (y_i - \langle x_i, \theta_1^* \rangle)]
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}[w(\theta^*)_1 \sum_{i=1}^n x_i (y_i - \langle x_i, \theta_1^* \rangle)] = \int \frac{1}{K} \{\sum_{i=1}^n x_i (y_i - \langle x_i, \theta_1^* \rangle)\} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \langle x_i, \theta_1^* \rangle)^2\} \\
(2\pi\sigma^2)^{-n/2} (2\pi)^{-nd/2} \exp\{-\frac{1}{2} \sum_{i=1}^n x_i^T x_i\} dx_{[n]} dy_{[n]} \\
= \frac{1}{K} \mathbb{E}_{(x_{[n]}, y_{[n]}) \sim \mathcal{D}_1^*} [\sum_{i=1}^n x_i (y_i - \langle x_i, \theta_1^* \rangle)] = 0
\end{aligned}$$

where $(x_{[n]}, y_{[n]}) \sim \mathcal{D}_1^*$ denotes $\{(x_i, y_i) : i = 1, \ldots, n\}$ being generated from the first component in the mixture with ground truth coefficient θ_1^*. Therefore,

$$\theta_1^+ - \theta_1^* = \underbrace{\mathbb{E}[w(\theta)_1 \sum_{i=1}^n x_i x_i^T]^{-1}}_{A} \underbrace{\mathbb{E}[(w(\theta)_1 - w(\theta^*)_1) \sum_{i=1}^n x_i (y_i - \langle x_i, \theta_1^* \rangle)]}_{B} .$$

B.1 Bound on B

$$\begin{aligned}
\|B\| &= \sup_{s \in \mathcal{S}^{d-1}} |\mathbb{E}[(w(\theta)_1 - w(\theta^*)_1) \sum_{i=1}^n (y_i - \langle x_i, \theta_1^* \rangle) \langle x_i, s \rangle]| \\
&\leq \frac{1}{K} |\underbrace{\mathbb{E}_{\mathcal{D}_1^*}[(w(\theta)_1 - w(\theta^*)_1) \sum_{i=1}^n (y_i - \langle x_i, \theta_1^* \rangle) \langle x_i, s \rangle]}_{T_1}| \\
&+ \frac{1}{K} \sum_{k \neq 1} |\underbrace{\mathbb{E}_{\mathcal{D}_k^*}[(w(\theta)_1 - w(\theta^*)_1) \sum_{i=1}^n (y_i - \langle x_i, \theta_1^* \rangle) \langle x_i, s \rangle]}_{T_k}|
\end{aligned}$$

For simplicity, denote $\Delta_1 = w(\theta)_1 - w(\theta^*)_1$. Notice that $|\Delta_1|$ is uniformly bounded by 1.

B.1.1 Bound on T_k

First, we bound T_k $\forall k \neq 1$.

$$T_k = |\mathbb{E}_{\mathcal{D}_k^*, \varepsilon_i}[\Delta_1 \sum_{i=1}^n (\varepsilon_i + \langle x_i, \theta_k^* - \theta_1^* \rangle) \langle x_i, s \rangle]|$$

$$\leq \underbrace{|\mathbb{E}_{\mathcal{D}_k^*,\varepsilon_i}[\Delta_1 \sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle \langle x_i, s\rangle]|}_{T_{k1}} + \underbrace{|\mathbb{E}_{\mathcal{D}_k^*,\varepsilon_i}[\Delta_1 \sum_{i=1}^n \varepsilon_i \langle x_i, s\rangle]|}_{T_{k2}}$$

In order to bound both two terms, we define the following good events

$$G_{k,1} = \{\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle^2 \geq \frac{320\sigma^2 n}{3}\},$$

$$G_{k,2} = \{\max\{\sum_{i=1}^n \langle x_i, \theta_k - \theta_k^* \rangle^2, \sum_{i=1}^n \langle x_i, \theta_1 - \theta_1^* \rangle^2\} \leq \frac{1}{16} \sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle^2\},$$

$$G_3 = \{\sum_{i=1}^n \varepsilon_i^2 \leq 2\sigma^2 n\}.$$

Define $G_k = G_{k,1} \cap G_{k,2} \cap G_3$ as the intersection of all three good events. Then

$$\begin{aligned} T_{k1} \leq\, & \mathbb{E}_{\mathcal{D}_k^*}[|\Delta_1 \sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle \langle x_i, s\rangle| \mathbb{1}_{G_k}] + \mathbb{E}_{\mathcal{D}_k^*}[|\Delta_1 \sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle \langle x_i, s\rangle| \mathbb{1}_{G_{k,1}^c}] \\ & + \mathbb{E}_{\mathcal{D}_k^*}[|\Delta_1 \sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle \langle x_i, s\rangle| \mathbb{1}_{G_{k,2}^c}] + \mathbb{E}_{\mathcal{D}_k^*}[|\Delta_1 \sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle \langle x_i, s\rangle| \mathbb{1}_{G_3^c}] \end{aligned}$$

and

$$\begin{aligned} T_{k2} \leq\, & \mathbb{E}_{\mathcal{D}_k^*,\varepsilon_i}[|\Delta_1 \sum_{i=1}^n \varepsilon_i \langle x_i, s\rangle| \mathbb{1}_{G_k}] + \mathbb{E}_{\mathcal{D}_k^*,\varepsilon_i}[|\Delta_1 \sum_{i=1}^n \varepsilon_i \langle x_i, s\rangle| \mathbb{1}_{G_{k,1}^c}] \\ & + \mathbb{E}_{\mathcal{D}_k^*,\varepsilon_i}[|\Delta_1 \sum_{i=1}^n \varepsilon_i \langle x_i, s\rangle| \mathbb{1}_{G_{k,2}^c}] + \mathbb{E}_{\mathcal{D}_k^*,\varepsilon_i}[|\Delta_1 \sum_{i=1}^n \varepsilon_i \langle x_i, s\rangle| \mathbb{1}_{G_3^c}] \end{aligned}$$

• On event G_k, we bound $w(\theta)_1$ and $w(\theta^*)_1$. As we take expectation with respect to $\mathcal{D}_k^*$, $y_i = \langle x_i, \theta_k^* \rangle + \varepsilon_i$. Then we have

$$\begin{aligned} w(\theta)_1 &= \frac{\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \langle x_i, \theta_1 \rangle)^2\}}{\sum_{l=1}^K \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \langle x_i, \theta_l \rangle)^2\}} \\ &\leq \exp\{\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \langle x_i, \theta_k \rangle)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \langle x_i, \theta_1 \rangle)^2\}\} \\ &= \exp\{\frac{1}{2\sigma^2}\sum_{i=1}^n (\varepsilon_i + \langle x_i, \theta_k^* - \theta_k \rangle)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (\varepsilon_i + \langle x_i, \theta_k^* - \theta_1^* \rangle - \langle x_i, \theta_1 - \theta_1^* \rangle)^2\} \end{aligned}$$

Note that

$$\begin{aligned} \sum_{i=1}^n (\varepsilon_i + \langle x_i, \theta_k^* - \theta_k \rangle)^2 &\leq 2\sum_{i=1}^n \varepsilon_i^2 + 2\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_k \rangle^2 \\ &\leq 2\sum_{i=1}^n \varepsilon_i^2 + \frac{1}{8}\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle^2 \end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^n(\varepsilon_i+\langle x_i,\theta_k^*-\theta_1^*\rangle-\langle x_i,\theta_1-\theta_1^*\rangle)^2 &\geq \sum_{i=1}^n\frac{1}{2}(\langle x_i,\theta_k^*-\theta_1^*\rangle-\langle x_i,\theta_1-\theta_1^*\rangle)^2-\varepsilon_i^2\\ &\geq \sum_{i=1}^n\frac{1}{4}\langle x_i,\theta_k^*-\theta_1^*\rangle^2-\sum_{i=1}^n\frac{1}{2}\langle x_i,\theta_1-\theta_1^*\rangle^2-\sum_{i=1}^n\varepsilon_i^2\\ &\geq \frac{7}{32}\sum_{i=1}^n\langle x_i,\theta_k^*-\theta_1^*\rangle^2-\sum_{i=1}^n\varepsilon_i^2\end{aligned}$$

Therefore,

$$w(\theta)_1\leq \exp\{\frac{3}{2\sigma^2}\sum_{i=1}^n\varepsilon_i^2-\frac{1}{2\sigma^2}\frac{3}{32}\sum_{i=1}^n\langle x_i,\theta_k^*-\theta_1^*\rangle^2\}\leq \exp\{-2n\}$$

Similarly,

$$\begin{aligned}w(\theta^*)_1&=\frac{\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\langle x_i,\theta_1^*\rangle)^2\}}{\sum_{l=1}^K\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\langle x_i,\theta_l^*\rangle)^2\}}\\ &\leq \exp\{\frac{1}{2\sigma^2}\sum_{i=1}^n\varepsilon_i^2-\frac{1}{2\sigma^2}\sum_{i=1}^n(\varepsilon_i^2+\langle x_i,\theta_k^*-\theta_1^*\rangle)^2\}\\ &\leq \exp\{\frac{1}{2\sigma^2}\sum_{i=1}^n2\varepsilon_i^2-\frac{1}{2\sigma^2}\sum_{i=1}^n\frac{1}{2}\langle x_i,\theta_k^*-\theta_1^*\rangle^2\}\\ &\leq \exp\{2n-\frac{80}{3}n\}\leq \exp(-2n)\end{aligned}$$

Therefore, $|\Delta_1|\leq|w(\theta)_1|+|w(\theta^*)_1|\leq 2\exp(-2n)$.
Then

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_k^*}[|\Delta_1\sum_{i=1}^n\langle x_i,\theta_k^*-\theta_1^*\rangle\langle x_i,s\rangle|\mathbb{1}_{G_k}]\\ &\leq 2\exp(-2n)\mathbb{E}_{x_i}[|\sum_{i=1}^n\langle x_i,\theta_k^*-\theta_1^*\rangle\langle x_i,s\rangle|]\\ &\leq 2\exp(-2n)\mathbb{E}[(\sum_{i=1}^n\langle x_i,\theta_k^*-\theta_1^*\rangle\langle x_i,s\rangle)^2]^{1/2}\\ &\leq 2\exp(-2n)\mathbb{E}[(\sum_{i=1}^n\langle x_i,\theta_k^*-\theta_1^*\rangle^2)(\sum_{i=1}^n\langle x_i,s\rangle^2)]^{1/2}\text{ by Cauchy-Schwarz}\\ &=2\exp(-2n)\{\mathbb{E}[\sum_{i=1}^n\langle x_i,\theta_k^*-\theta_1^*\rangle^2\langle x_i,s\rangle^2]+\sum_{i\neq i'}\mathbb{E}[\langle x_i,\theta_k-\theta_1^*\rangle^2]\mathbb{E}[\langle x_{i'},s\rangle^2]\}^{1/2}\\ &\overset{\text{(i)}}{\leq} 2\exp(-2n)\{n\|\theta_k^*-\theta_1^*\|^2+n(n-1)\|\theta_k^*-\theta_1^*\|^2\}^{1/2}\\ &=\mathcal{O}(\Delta_{\max}n\exp(-n))\end{aligned}$$

where the inequality (i) follows from Lemma 4. Similarly,

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_k^*,\varepsilon_i}[|\Delta_1 \sum_{i=1}^n \varepsilon_i \langle x_i, s\rangle| \mathbb{1}_{G_k}] \\
&\leq 2\exp(-2n)\mathbb{E}_{x_i,\varepsilon_i}[(\sum_{i=1}^n \varepsilon_i\langle x_i, s\rangle)^2]^{1/2} \\
&\leq 2\exp(-2n)\{\mathbb{E}[\sum_{i=1}^n \varepsilon_i^2]\mathbb{E}[\sum_{i=1}^n \langle x_i, s\rangle^2]\}^{1/2} \text{ by Cauchy-Schwarz and the independence of } x_i \text{ and } \varepsilon_i \\
&= \mathcal{O}(\sigma n \exp(-n))
\end{aligned}
$$

- Next, on event $G_{k,1}^c$, we bound $\mathbb{P}(G_{k,1}^c)$.

$$
\mathbb{P}(G_{k,1}^c) = \mathbb{P}(\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^*\rangle^2 \leq \frac{320\sigma^2 n}{3}) = \mathbb{P}(\sum_{i=1}^n \frac{\langle x_i, \theta_k^* - \theta_1^*\rangle^2}{\|\theta_k^* - \theta_1^*\|^2} \leq \frac{320\sigma^2 n}{3\|\theta_k^* - \theta_1^*\|^2})
$$

where $\sum_{i=1}^n \frac{\langle x_i, \theta_k^* - \theta_1^*\rangle^2}{\|\theta_k^* - \theta_1^*\|^2} \sim \chi_n^2$. Then by Lemma 5, set $n - 2\sqrt{ns} = \frac{320n}{3\,\mathrm{SNR}^2}$. Since we assume $\mathrm{SNR} \geq (\frac{320}{3}K)^{1/2} = \Omega(\sqrt{K})$, then $n - \frac{320n}{3\,\mathrm{SNR}^2} \geq n - \frac{1}{K}n > 0$ and $s = n(\frac{1}{2} - \frac{160}{3\,\mathrm{SNR}^2})^2$. Therefore,

$$
\mathbb{P}(G_{k,1}^c) \leq \exp(-n(\frac{1}{2} - \frac{160}{3\,\mathrm{SNR}^2})^2) \leq \exp(-n(\frac{1}{2} - \frac{1}{2K})^2) \leq \exp(-\frac{1}{16}n)
$$

Then,

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_k^*}[|\Delta_1 \sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^*\rangle\langle x_i, s\rangle| \mathbb{1}_{G_{k,1}^c}] \\
&\leq \mathbb{E}_{\mathcal{D}_k^*}[|\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^*\rangle\langle x_i, s\rangle||G_{k,1}^c]\mathbb{P}(G_{k,1}^c) \\
&\leq \sqrt{\mathbb{E}_{x_i}[\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^*\rangle^2 | G_{k,1}^c]}\sqrt{\mathbb{E}_{x_i}[\sum_{i=1}^n \langle x_i, s\rangle^2|G_{k,1}^c]}\mathbb{P}(G_{k,1}^c) \\
&\leq \mathcal{O}(\Delta_{\max}\sqrt{n}\exp(-n)) \text{ by Lemma 2}
\end{aligned}
$$

and similarly,

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}_k^*,\varepsilon_i}[|\Delta_1 \sum_{i=1}^n \varepsilon_i\langle x_i, s\rangle| \mathbb{1}_{G_{k,1}^c}] &\leq \sqrt{\mathbb{E}_{\varepsilon_i}[\sum_{i=1}^n \varepsilon_i^2 | G_{k,1}^c]}\sqrt{\mathbb{E}_{x_i}[\sum_{i=1}^n \langle x_i, s\rangle^2|G_{k,1}^c]}\mathbb{P}(G_{k,1}^c) \\
&\overset{\text{(i)}}{\leq} \mathcal{O}(n^{3/4}\sigma\exp(-n))
\end{aligned}
$$

where the inequality (i) follows from the independent of ε_i on $G_{k,1}^c$ and Lemma 2.

- Next, for $G_{k,2}^c$,

$$
\mathbb{P}(G_{k,2}^c) \leq \mathbb{P}(\sum_{i=1}^n \langle x_i, \theta_k - \theta_k^*\rangle^2 \geq \frac{1}{16}\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^*\rangle^2) + \mathbb{P}(\sum_{i=1}^n \langle x_i, \theta_1 - \theta_1^*\rangle^2 \geq \frac{1}{16}\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^*\rangle^2)
$$

Note that $\forall t > 0$,

$$
\begin{aligned}
\mathbb{P}(\sum_{i=1}^{n}\langle x_i, \theta_k - \theta_k^*\rangle^2 \geq \frac{1}{16}\sum_{i=1}^{n}\langle x_i, \theta_k^* - \theta_1^*\rangle^2) &\leq \mathbb{P}(\sum_{i=1}^{n}\langle x_i, \theta_k - \theta_k^*\rangle^2 \geq t) + \mathbb{P}(\frac{1}{16}\sum_{i=1}^{n}\langle x_i, \theta_k^* - \theta_1^*\rangle^2 \leq t) \\
&= \mathbb{P}(\sum_{i=1}^{n}\frac{\langle x_i, \theta_k - \theta_k^*\rangle^2}{\|\theta_k - \theta_k^*\|^2} \geq \frac{t}{\|\theta_k - \theta_k^*\|^2}) \\
&\quad + \mathbb{P}(\frac{1}{16}\sum_{i=1}^{n}\frac{\langle x_i, \theta_k^* - \theta_1^*\rangle^2}{\|\theta_k^* - \theta_1^*\|^2} \leq \frac{t}{\|\theta_k^* - \theta_1^*\|^2})
\end{aligned}
$$

Again, we can use χ^2 tail bounds from Lemma 5. First, we set $\frac{t}{\|\theta_k - \theta_k^*\|^2} = n + 2s + 2\sqrt{ns}$. Then

$$
\sqrt{s} = \frac{-2\sqrt{n} + \sqrt{4n - 8(n - \frac{t}{\|\theta_k - \theta_k^*\|^2})}}{4}.
$$

In order to make s a valid real number, we need $t > n\|\theta_k - \theta_k^*\|^2$.
On the other hand, set $\frac{16t}{\|\theta_k^* - \theta_1^*\|^2} = n - 2\sqrt{ns'}$. Then

$$
\sqrt{s'} = \frac{1}{2}\sqrt{n} - \frac{8t}{\sqrt{n}\|\theta_k^* - \theta_1^*\|^2}.
$$

In order to make s' a valid real number, we need $t < \frac{1}{16}n\|\theta_k^* - \theta_1^*\|^2$.
Since we assume $\max_{k\in[K]}\|\theta_k - \theta_k^*\| \leq \frac{1}{4}\Delta_{\min}$, then the requirement $n\|\theta_k - \theta_k^*\|^2 < t < \frac{1}{16}n\|\theta_k^* - \theta_1^*\|^2$ can be satisfied by taking $t = \frac{1}{2}n(\|\theta_k - \theta_k^*\|^2 + \frac{1}{16}\|\theta_k^* - \theta_1^*\|^2)$. Moreover, recall we assume $\max_{k\in[K]}\|\theta_k - \theta_k^*\| = \alpha\Delta_{\min}$ for some $\alpha \in (0, \frac{1}{4})$. Then,

$$
\mathbb{P}(\sum_{i=1}^{n}\frac{\langle x_i, \theta_k - \theta_k^*\rangle^2}{\|\theta_k - \theta_k^*\|^2} \geq \frac{t}{\|\theta_k - \theta_k^*\|^2}) \leq \exp(-\frac{n}{16}(\frac{1}{2\alpha} - 2)^2)
$$

and

$$
\mathbb{P}(\frac{1}{16}\sum_{i=1}^{n}\frac{\langle x_i, \theta_k^* - \theta_1^*\rangle^2}{\|\theta_k^* - \theta_1^*\|^2} \leq \frac{t}{\|\theta_k^* - \theta_1^*\|^2}) \leq \exp(-n(\frac{1}{4} - 4\alpha^2)^2)
$$

Therefore,

$$
\mathbb{P}(\sum_{i=1}^{n}\langle x_i, \theta_k - \theta_k^*\rangle^2 \geq \frac{1}{16}\sum_{i=1}^{n}\langle x_i, \theta_k^* - \theta_1^*\rangle^2) \leq \exp(-\frac{n}{16}(\frac{1}{2\alpha} - 2)^2) + \exp(-n(\frac{1}{4} - 4\alpha^2)^2).
$$

Note that the analysis also holds true for $\mathbb{P}(\sum_{i=1}^{n}\langle x_i, \theta_1 - \theta_1^*\rangle^2 \geq \frac{1}{16}\sum_{i=1}^{n}\langle x_i, \theta_k^* - \theta_1^*\rangle^2)$.
Therefore,

$$
\begin{aligned}
\mathbb{P}(G_{k,2}^c) &\leq 2\exp(-\frac{n}{16}(\frac{1}{2\alpha} - 2)^2) + 2\exp(-n(\frac{1}{4} - 4\alpha^2)^2) \\
&\leq 4\exp(-n(\frac{1}{4} - 4\alpha^2)^2)
\end{aligned}
$$

Notice that on $G_{k,2}^c$, $\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle^2 \leq 16 \sum_{i=1}^n \langle x_i, \theta_k - \theta_k^* \rangle^2 + 16 \sum_{i=1}^n \langle x_i, \theta_1 - \theta_1^* \rangle^2$. Then,

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_k^*}[|\Delta_1 \sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle \langle x_i, s \rangle | \mathbb{1}_{G_{k,2}^c}] \\
&\leq \sqrt{\mathbb{E}_{x_i}[\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle^2 | G_{k,2}^c]} \sqrt{\mathbb{E}_{x_i}[\sum_{i=1}^n \langle x_i, s \rangle^2 | G_{k,2}^c] \mathbb{P}(G_{k,2}^c)} \\
&\leq \sqrt{16 \mathbb{E}[\sum_{i=1}^n \langle x_i, \theta_k - \theta_k^* \rangle^2 | G_{k,2}^c] + 16 \mathbb{E}[\sum_{i=1}^n \langle x_i, \theta_1 - \theta_1^* \rangle^2 | G_{k,2}^c] \mathbb{P}(G_{k,2}^c)} \\
&\leq \mathcal{O}(D_M n \exp(-C_\alpha n)) \text{ by Lemma 3}
\end{aligned}
$$

where C_α is a constant depending on α and $D_M := \max_{k \in [K]} \|\theta_k - \theta_k^*\|$. Similarly,

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}_k^*, \varepsilon_i}[|\Delta_1 \sum_{i=1}^n \varepsilon_i \langle x_i, s \rangle | \mathbb{1}_{G_{k,2}^c}] &\leq \sqrt{\mathbb{E}_{\varepsilon_i}[\sum_{i=1}^n \varepsilon_i^2 | G_{k,2}^c]} \sqrt{\mathbb{E}_{x_i}[\sum_{i=1}^n \langle x_i, s \rangle^2 | G_{k,2}^c] \mathbb{P}(G_{k,2}^c)} \\
&\leq \mathcal{O}(n\sigma \exp(-C_\alpha n))
\end{aligned}
$$

- Next, for G_3^c,

$$
\begin{aligned}
\mathbb{P}(G_3^c) = \mathbb{P}(\sum_{i=1}^n (\frac{\varepsilon_i}{\sigma})^2 \geq 2n) &\leq \exp(-n(\frac{\sqrt{1}2 - 2}{4})^2) \quad \text{by Lemma 5} \\
&\leq \exp(-\frac{n}{16})
\end{aligned}
$$

Then,

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_k^*}[|\Delta_1 \sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle \langle x_i, s \rangle | \mathbb{1}_{G_3^c}] \\
&\leq \sqrt{\mathbb{E}_{x_i}[\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle^2]} \sqrt{\mathbb{E}_{x_i}[\sum_{i=1}^n \langle x_i, s \rangle^2] \mathbb{P}(G_3^c)} \text{ by the independence of } x_i \text{ on } G_3^c \\
&\leq \mathcal{O}(\Delta_{\max} n \exp(-n))
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}_k^*, \varepsilon_i}[|\Delta_1 \sum_{i=1}^n \varepsilon_i \langle x_i, s \rangle | \mathbb{1}_{G_3^c}] &\leq \sqrt{\mathbb{E}_{x_i, \varepsilon_i}[(\sum_{i=1}^n \varepsilon_i \langle x_i, s \rangle)^2]} \sqrt{\mathbb{P}(G_3^c)} \text{ by Cauchy-Schwarz} \\
&\leq \sqrt{\mathbb{E}_{\varepsilon_i}[\sum_{i=1}^n \varepsilon_i^2]} \sqrt{\mathbb{E}_{x_i}[\sum_{i=1}^n \langle x_i, s \rangle^2]} \sqrt{\mathbb{P}(G_3^c)} \\
&\leq \mathcal{O}(n\sigma \exp(-n))
\end{aligned}
$$

Therefore,

$$
T_k \leq \mathcal{O}((\sigma + \Delta_{\max}) n \exp(-n) + D_M \exp(-C_\alpha n))
$$

B.1.2 Bound on T_1

The analysis is similar as the analysis from the last section.

$$
\begin{aligned}
T_1 &= |\mathbb{E}_{\mathcal{D}_1^*}[\Delta_1 \sum_{i=1}^n (y_i - \langle x_i, \theta_1^* \rangle)\langle x_i, s\rangle]| \\
&\leq \mathbb{E}_{\mathcal{D}_1^*}[\Delta_1^2]^{1/2} \mathbb{E}_{x_i,\varepsilon_i}[(\sum_{i=1}^n \varepsilon_i \langle x_i, s\rangle)^2]^{1/2}
\end{aligned}
$$

Note that

$$
\begin{aligned}
\mathbb{E}_{x_i,\varepsilon_i}[(\sum_{i=1}^n \varepsilon_i \langle x_i, s\rangle)^2] &= \mathbb{E}[\sum_{i=1}^n \varepsilon_i^2 \langle x_i, s\rangle^2] + \mathbb{E}[\sum_{i\neq i'} \varepsilon_i \varepsilon_{i'} \langle x_i, s\rangle \langle x_{i'}, s\rangle] \\
&= \sum_{i=1}^n \mathbb{E}[\varepsilon_i^2] = \sigma^2 n
\end{aligned}
$$

Therefore, $\mathbb{E}_{x_i,\varepsilon_i}[(\sum_{i=1}^n \varepsilon_i \langle x_i, s\rangle)^2]^{1/2} = \sqrt{n}\sigma$.
Next, We define the following good events

$$
G_1 = \{\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle^2 \geq \frac{320\sigma^2 n}{3} \ \forall k \neq 1\},
$$

$$
G_2 = \{\sum_{i=1}^n \langle x_i, \theta_1 - \theta_1^* \rangle^2 \leq \frac{1}{16} \sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle^2 \, \forall k \neq 1\},
$$

$$
G_3 = \{\sum_{i=1}^m \varepsilon_i^2 \leq 2\sigma^2 n\}
$$

and $G = G_1 \cap G_2 \cap G_3$.
Then, we bound $\mathbb{E}_{\mathcal{D}_1^*}[\Delta_1^2]$ by $\mathbb{E}_{\mathcal{D}_1^*}[\Delta_1^2] \leq \mathbb{E}_{\mathcal{D}_1^*}[\Delta_1^2 | G] + \mathbb{P}(G^c)$.
• First, note that

$$
w(\theta)_1 = \frac{\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \langle x_i, \theta_1\rangle)^2\}}{\sum_{l=1}^K \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \langle x_i, \theta_l\rangle)^2\}} = 1 - \sum_{k\neq 1} w(\theta)_k
$$

where we bound $w(\theta)_k$ on G by the similar method as before.

$$
w(\theta)_k \leq \exp\{\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \langle x_i, \theta_1\rangle)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \langle x_i, \theta_k\rangle)^2\} \leq \exp(-2n)
$$

Then $w(\theta)_1 \geq 1 - K\exp(-2n)$. Similarly,

$$
w(\theta^*)_1 = 1 - \sum_{k\neq 1} w(\theta^*)_k
$$

and

$$
w(\theta^*)_k \leq \exp\{\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \langle x_i, \theta_1^*\rangle)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \langle x_i, \theta_k^*\rangle)^2\} \leq \exp(-2n).
$$

Therefore, $w(\theta^*)_1 \geq 1 - K\exp(-2n)$. Since $1 - K\exp(-2n) \leq w(\theta^*)_1, w(\theta)_1 \leq 1$, we have $|\Delta_1| \leq K\exp(-2n)$. Then $\mathbb{E}_{\mathcal{D}_1^*}[\Delta_1^2|G] \leq K^2\exp(-4n)$.

• Next, note that $G_1 = \cap_{k\neq 1} G_{k,1}$ and $G_2 = \cap_{k\neq 1}\{\sum_{i=1}^n \langle x_i, \theta_1 - \theta_1^*\rangle^2 \leq \frac{1}{16}\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^*\rangle^2\}$. Then

$$\mathbb{P}(G_1^c) \leq \sum_{k\neq 1}\mathbb{P}(G_{k,1}^c) \leq K\exp(-\frac{n}{16})$$

and

$$\begin{aligned}
\mathbb{P}(G_2^c) &\leq \sum_{k\neq 1}\mathbb{P}(\sum_{i=1}^n \langle x_i, \theta_1 - \theta_1^*\rangle^2) \geq \frac{1}{16}\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^*\rangle^2) \\
&\leq K\exp(-\frac{n}{16}(\frac{1}{2\alpha} - 2)^2) + K\exp(-n(\frac{1}{4} - 4\alpha^2)^2) \\
&\leq 2K\exp(-n(\frac{1}{4} - 4\alpha^2)^2)
\end{aligned}$$

where the inequality $\mathbb{P}(\sum_{i=1}^n \langle x_i, \theta_1 - \theta_1^*\rangle^2) \geq \frac{1}{16}\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^*\rangle^2) \leq \exp(-\frac{n}{16}(\frac{1}{2\alpha} - 2)^2) + \exp(-n(\frac{1}{4} - 4\alpha^2)^2)$ has been proved in the last part.
Moreover, we have also proved $\mathbb{P}(G_3^c) \leq \exp(-\frac{n}{16})$ in the last part.
Therefore,

$$\mathbb{P}(G^c) \leq (K+1)\exp(-\frac{n}{16}) + 2K\exp(-n(\frac{1}{4} - 4\alpha^2)^2).$$

Putting everything together, we bound T_1 by

$$\begin{aligned}
T_1 &\leq \sqrt{n}\sigma\{K^2\exp(-4n) + (K+1)\exp(-\frac{1}{16}n) + 2K\exp(-n(\frac{1}{4} - 4\alpha^2)^2)\}^{1/2} \\
&= \mathcal{O}(\sqrt{n}\sigma K\exp(-C_\alpha n))
\end{aligned}$$

Therefore, the bound on B is given by

$$\begin{aligned}
\|B\| &= \frac{1}{K}T_1 + \frac{1}{K}\sum_{k\neq 1} T_k \\
&= \mathcal{O}(\sigma\sqrt{n}\exp(-C_\alpha n)) + \mathcal{O}((\sigma + \Delta_{\max})n\exp(-n) + D_M\exp(-C_\alpha n)) \\
&= \mathcal{O}(D_M\exp(-C_\alpha n)) \text{ by the assumption } n \geq \Omega(\log(\Delta_{\max} + \sigma) + \log K)
\end{aligned}$$

B.2 Bound on A

Recall that $A = \mathbb{E}[w(\theta)_1\sum_{i=1}^n x_i x_i^T]^{-1}$. Notice that

$$\mathbb{E}[w(\theta)_1\sum_{i=1}^n x_i x_i^T] = \frac{1}{K}\sum_{k=1}^K \mathbb{E}_{\mathcal{D}_k^*}[w(\theta)_1\sum_{i=1}^n x_i x_i^T] \succcurlyeq \frac{1}{K}\mathbb{E}_{\mathcal{D}_1^*}[w(\theta)_1\sum_{i=1}^n x_i x_i^T]$$

Therefore, it is enough to lower bound $\|\mathbb{E}_{\mathcal{D}_1^*}[w(\theta)_1\sum_{i=1}^n x_i x_i^T]\|$.

$$\begin{aligned}
\|\mathbb{E}_{\mathcal{D}_1^*}[w(\theta)_1\sum_{i=1}^n x_i x_i^T]\| &= \|\mathbb{E}_{\mathcal{D}_1^*}[\sum_{i=1}^n x_i x_i^T] - \mathbb{E}_{\mathcal{D}_1^*}[(1 - w(\theta)_1)\sum_{i=1}^n x_i x_i^T]\| \\
&\geq \sum_{i=1}^n \|\mathbb{E}_{\mathcal{D}_1^*}[x_i x_i^T]\| - \sup_{s\in\mathcal{S}^{d-1}}\|\mathbb{E}_{\mathcal{D}_1^*}[(1 - w(\theta)_1)\sum_{i=1}^n x_i\langle x_i, s\rangle]\|
\end{aligned}$$

$$\geq n - \sup_{s \in \mathcal{S}^{d-1}} \mathbb{E}_{\mathcal{D}_1^*}[\|(1 - w(\theta)_1) \sum_{i=1}^{n} x_i \langle x_i, s \rangle\|]$$

$$= n - \sup_{s \in \mathcal{S}^{d-1}} \mathbb{E}_{\mathcal{D}_1^*}[\sup_{u \in \mathcal{S}^{d-1}} |(1 - w(\theta)_1) \sum_{i=1}^{n} \langle x_i, u \rangle \langle x_i, s \rangle|]$$

Then $\forall u, s \in \mathcal{S}^{d-1}$,

$$\mathbb{E}_{\mathcal{D}_1^*}[|(1 - w(\theta)_1) \sum_{i=1}^{n} \langle x_i, u \rangle \langle x_i, s \rangle|] \leq \mathbb{E}_{\mathcal{D}_1^*}[(1 - w(\theta)_1)^2]^{1/2} \mathbb{E}[(\sum_{i=1}^{n} \langle x_i, s \rangle \langle x_i, u \rangle)^2]^{1/2}$$

Notice that

$$\begin{aligned} \mathbb{E}[(\sum_{i=1}^{n} \langle x_i, s \rangle \langle x_i, u \rangle)^2] &= \mathbb{E}[\sum_{i=1}^{n} \langle x_i, u \rangle \langle x_i, s \rangle]^2 + \mathrm{Var}[\sum_{i=1}^{n} \langle x_i, u \rangle \langle x_i, s \rangle] \\ &= (\sum_{i=1}^{n} \mathbb{E}[\langle x_i, u \rangle \langle x_i, s \rangle])^2 + \sum_{i=1}^{n} \mathrm{Var}[\langle x_i, u \rangle \langle x_i, s \rangle] \end{aligned}$$

where

$$(\sum_{i=1}^{n} \mathbb{E}[\langle x_i, u \rangle \langle x_i, s \rangle])^2 = n^2$$

and

$$\begin{aligned} \mathrm{Var}[\langle x_i, u \rangle \langle x_i, s \rangle] &= \mathbb{E}[\langle x_i, u \rangle^2 \langle x_i, s \rangle^2] - (\mathbb{E}[\langle x_i, u \rangle \langle x_i, s \rangle])^2 \\ &\leq \mathbb{E}[\langle x_i, u \rangle^2 \langle x_i, s \rangle^2] \leq 3 \quad \text{by Lemma 4.} \end{aligned}$$

Therefore,

$$\mathbb{E}[(\sum_{i=1}^{n} \langle x_i, s \rangle \langle x_i, u \rangle)^2]^{1/2} \leq (n^2 + 3n)^{1/2} \leq 2n.$$

Next, we bound $\mathbb{E}_{\mathcal{D}_1^*}[(1 - w(\theta)_1)^2]$. Similarly as before, we use the following good events

$$G_1 = \{\sum_{i=1}^{n} \langle x_i, \theta_k^* - \theta_1^* \rangle^2 \geq \frac{320\sigma^2 n}{3} \ \forall k \neq 1\},$$

$$G_2 = \{\sum_{i=1}^{n} \langle x_i, \theta_1 - \theta_1^* \rangle^2 \leq \frac{1}{16} \sum_{i=1}^{n} \langle x_i, \theta_k^* - \theta_1^* \rangle^2 \, \forall k \neq 1\},$$

$$G_3 = \{\sum_{i=1}^{m} \varepsilon_i^2 \leq 2\sigma^2 n\}$$

and $G = G_1 \cap G_2 \cap G_3$.

On event G, $w(\theta)_1 \geq 1 - K\exp(-2n)$, so $1 - w(\theta)_1 \leq K\exp(-2n)$.

Therefore,

$$\begin{aligned} \mathbb{E}_{\mathcal{D}_1^*}[(1 - w(\theta)_1)^2] &\leq \mathbb{E}_{\mathcal{D}_1^*}[(1 - w(\theta)_1)^2 | G] + \mathbb{P}(G^c) \\ &\leq K^2 \exp(-4n) + (K+1)\exp(-\frac{n}{16}) + 2K\exp(-n(\frac{1}{4} - 4\alpha^2)^2) \end{aligned}$$

Then

$$\|A^{-1}\| \geq \frac{n}{K} - \frac{2n}{K}\{K^2 \exp(-4n) + (K+1)\exp(-\frac{n}{16}) + 2K\exp(-n(\frac{1}{4} - 4\alpha^2)^2)\}^{1/2}$$

which gives the following bound on A

$$\|A\| \leq \frac{K}{n - nK\mathcal{O}(\exp(-C_\alpha n))}$$

Therefore,

$$\begin{aligned}\|\theta_1^+ - \theta_1^*\| \leq \|A\|\|B\| &= \mathcal{O}(D_M K \exp(-C_\alpha n)) \\ &= \mathcal{O}(D_M \exp(-C_\alpha n)) \text{ by the assumption } n \geq \Omega(\log K + \log(\Delta_{\max} + \sigma))\end{aligned}$$

C Proof of Theorem 2

This section proves the convergence rate for the empirical EM. First, we present a more detailed version of Theorem 2.

Theorem 3. *(Detailed Version of Theorem 2) Given the initialization condition* $\max_{k\in[K]} \|\widetilde{\theta}_k^{(0)} - \theta_k^*\| \leq \frac{1}{4}\Delta_{\min}$*, suppose the current estimates* $\{\widetilde{\theta}_k\}_{k=1}^K$ *satisfy* $\max_{k\in[K]} \|\widetilde{\theta}_k - \theta_k^*\| = \alpha\Delta_{\min}$ *for some* $\alpha \in (0, \frac{1}{4})$*. Let* C_α *be a constant depending on* α*. Assume* $n = \Omega(\frac{\log K + \log(\sigma + \Delta_{\max})}{C_\alpha})$ *and* $\mathrm{SNR} \geq \Omega(\sqrt{K})$*. Then with probability at least* $1 - \delta/T$*,* $\{\widetilde{\theta}_k^+\}_{k=1}^K$ *updated by one iteration of the Empirical EM satisfy the following.*

(i) If $\frac{m}{T} = \Omega(K \exp(n) \log(dK^2T/\delta))$ *and* $\frac{mn}{T} = \Omega(dK \log(dK^2T/\delta))$*, then*

$$\begin{aligned}\max_{k\in[K]} \|\widetilde{\theta}_k^+ - \theta_k^*\| \leq &\sqrt{\frac{Kd}{mn}\log(dK^2T/\delta)}(\sigma + \Delta_{\max}\exp(-n)) \\ &+ \max_{k\in[K]} \|\widetilde{\theta}_k - \widetilde{\theta}_k^*\|\mathcal{O}(\exp(-C_\alpha n))\end{aligned}$$

(ii) If $\mathcal{O}(K \exp(n) \log(dK^2T/\delta)) \geq \frac{m}{T} \geq \Omega(K\sqrt{n}\log(dK^2T/\delta))$ *and* $\frac{mn}{T} = \Omega(dK^{3/2}\log(dK^2T/\delta))$*, then*

$$\begin{aligned}\max_{k\in[K]} \|\widetilde{\theta}_k^+ - \theta_k^*\| \leq &\sqrt{\frac{K^{3/2}d}{mn}\log(dK^2T/\delta)}(\sigma + \frac{\Delta_{\max}}{\sqrt{n}}) \\ &+ \max_{k\in[K]} \|\widetilde{\theta}_k - \widetilde{\theta}_k^*\|\mathcal{O}(\exp(-C_\alpha n) + \sqrt{\frac{K^{3/2}d}{mn}\log(dK^2T/\delta)})\end{aligned}$$

(iii) If $\mathcal{O}(K\sqrt{n}\log(dK^2T/\delta)) \geq \frac{m}{T} \geq \Omega(K \log(dK^2T/\delta))$ *and* $\frac{mn}{T} = \Omega(dK^4 \log(dK^2T/\delta))$*, then*

$$\begin{aligned}\max_{k\in[K]} \|\widetilde{\theta}_k^+ - \theta_k^*\| \leq &\sqrt{\frac{dK^4}{mn}\log(dK^2T/\delta)}(\sigma + \Delta_{\max}) \\ &+ \max_{k\in[K]} \|\widetilde{\theta}_k - \widetilde{\theta}_k^*\|\mathcal{O}(\exp(-C_\alpha n) + \sqrt{\frac{dK^3}{m\sqrt{n}}\log(dK^2T/\delta)})\end{aligned}$$

In order to prove the theorem, we perform one step analysis again. Suppose at the current step, we have estimates $\widetilde{\theta}_1, \ldots, \widetilde{\theta}_K$, and one step update of the empirical EM generates new estimates $\widetilde{\theta}_1^+, \ldots, \widetilde{\theta}_K^+$. WLOG, we focus on $\widetilde{\theta}_1^+ - \theta_1^*$.

$$\widetilde{\theta}_1^+ - \theta_1^* = \underbrace{(\frac{1}{mn}\sum_{j=1}^m w(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j x_i^{jT})^{-1}}_{A_m} \underbrace{(\frac{1}{mn}\sum_{j=1}^m w(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j (y_i^j - \langle x_i^j, \theta_1^* \rangle))}_{B_m}$$

C.1 Bound on B_m

$B_m = (B_m - B) + B$ where $B = \mathbb{E}[w(\widetilde{\theta})_1 \sum_{i=1}^n x_i(y_i - \langle x_i, \theta_1^* \rangle)]$ has already been bounded by the analysis before. We only need to bound the following

$$B_m - B = \frac{1}{m}\sum_{j=1}^m \frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j (y_i^j - \langle x_i^j, \theta_1^* \rangle)) - \frac{1}{mn}\sum_{j=1}^m \mathbb{E}[w(\widetilde{\theta})_1 \sum_{i=1}^n x_i(y_i - \langle x_i, \theta_1^* \rangle)].$$

We define the following events

$$E_k = \{\text{data points on the client are generated from the } k\text{-th linear regression}\},$$

$$G_{k,1} = \{\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle^2 \geq \frac{320\sigma^2 n}{3}\},$$

$$G_{k,2} = \{\max\{\sum_{i=1}^n \langle x_i, \theta_k - \theta_k^* \rangle^2, \sum_{i=1}^n \langle x_i, \theta_1 - \theta_1^* \rangle^2\} \leq \frac{1}{16}\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle^2\},$$

$$G_3 = \{\sum_{i=1}^n \varepsilon_i^2 \leq 2\sigma^2 n\}$$

and $G_k = G_{k,1} \cap G_{k,2} \cap G_3$. Then

$$\begin{aligned}
&\frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j (y_i^j - \langle x_i^j, \theta_1^* \rangle) \\
&= \sum_{k \neq 1}\{\frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \langle x_i^j, \theta_k^* - \theta_1^* \rangle \mathbb{1}_{E_k \cap G_k} + \frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \langle x_i^j, \theta_k^* - \theta_1^* \rangle \mathbb{1}_{E_k \cap G_{k,1}^c} \\
&+ \frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \langle x_i^j, \theta_k^* - \theta_1^* \rangle \mathbb{1}_{E_k \cap G_{k,2}^c} + \frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \langle x_i^j, \theta_k^* - \theta_1^* \rangle \mathbb{1}_{E_k \cap G_3^c}\} \\
&+ \sum_{k \neq 1}\{\frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \varepsilon_i^j \mathbb{1}_{E_k \cap G_k} + \frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \varepsilon_i^j \mathbb{1}_{E_k \cap G_{k,1}^c} + \frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \varepsilon_i^j \mathbb{1}_{E_k \cap G_{k,2}^c} \\
&+ \frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \varepsilon_i^j \mathbb{1}_{E_k \cap G_3^c}\} + \frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \varepsilon_i^j \mathbb{1}_{E_1}
\end{aligned}$$

For each term, we will use Lemma 1 to bound the deviation of the empirical mean from the population mean with probability $1 - \frac{3\delta}{K^2 T}$. In order to invoke Lemma 1, the key point is to show that every term conditional on the corresponding event is sub-exponential and compute its sub-exponential norm.

C.1.1 Analysis of $\frac{1}{n}w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \langle x_i^j, \theta_k^* - \theta_1^* \rangle \mathbb{1}_{j \in E_k \cap G_k}$

First of all, we want to show $\forall j$, $W_j := \frac{1}{n}w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \langle x_i^j, \theta_k^* - \theta_1^* \rangle | E_k \cap G_k$ is sub-exponential by showing it has finite sub-exponential norm. This is equivalent to show $\langle W_j, s \rangle$ has finite sub-exponential norm.

$$
\begin{aligned}
&\|\frac{1}{n}w^j(\widetilde{\theta})_1 \sum_{i=1}^n \langle x_i^j, s\rangle \langle x_i^j, \theta_k^* - \theta_1^* \rangle | E_k \cap G_k\|_{\psi_1} \\
&\overset{\text{(i)}}{\leq} \frac{1}{n}\exp(-2n) \| \sum_{i=1}^n \langle x_i^j, s\rangle \langle x_i^j, \theta_k^* - \theta_1^* \rangle | E_k \cap G_k\|_{\psi_1} \\
&= \frac{1}{n}\exp(-2n) \sup_{p \geq 1} p^{-1} \mathbb{E}[| \sum_{i=1}^n \langle x_i^j, s\rangle \langle x_i^j, \theta_k^* - \theta_1^* \rangle |^p | E_k \cap G_k]^{1/p} \\
&\overset{\text{(ii)}}{\leq} \frac{1}{n}\exp(-2n) \sup_{p \geq 1} p^{-1} \mathbb{E}[(\sum_{i=1}^n \langle x_i^j, s\rangle^2)^{p/2} (\sum_{i=1}^n \langle x_i^j, \theta_k^* - \theta_1^* \rangle^2)^{p/2} | E_k \cap G_k]^{1/p} \\
&\overset{\text{(iii)}}{\leq} \frac{1}{n}\exp(-2n) \sup_{p \geq 1} p^{-1} \sqrt{\mathbb{E}[(\sum_{i=1}^n \langle x_i^j, s\rangle^2)^p | E_k \cap G_k]^{1/p}} \sqrt{\mathbb{E}[(\sum_{i=1}^n \langle x_i^j, \theta_k^* - \theta_1^* \rangle^2)^p | E_k \cap G_k]^{1/p}} \\
&\overset{\text{(iiii)}}{=} \frac{1}{n}\exp(-2n) \sup_{p \geq 1} p^{-1} \sqrt{\mathbb{E}[(\sum_{i=1}^n \langle x_i^j, s\rangle^2)^p]^{1/p}} \sqrt{\mathbb{P}(G_k)^{-1/p}\mathbb{E}[(\sum_{i=1}^n \langle x_i^j, \theta_k^* - \theta_1^* \rangle^2)^p \mathbb{1}_{G_k}]^{1/p}} \\
&\overset{\text{(v)}}{\leq} \frac{1}{n}\exp(-2n) \sup_{p \geq 1} p^{-1} \sqrt{\mathcal{O}(\sqrt{n}p)} \sqrt{\mathcal{O}(\|\theta_k^* - \theta_1^*\|^2 \sqrt{n}p)} \\
&= \mathcal{O}(\frac{\|\theta_k^* - \theta_1^*\|}{\sqrt{n}} \exp(-n))
\end{aligned}
$$

Inequality (i) follows from the proof that on event G_k, $w^j(\widetilde{\theta})_1 \leq \exp(-2n)$. Both inequalities (ii) and (iii) follow from Cauchy-Schwarz. Equality (iiii) follows from the fact that all x_i^j's are independent of E_k and $\sum_{i=1}^n \langle x_i^j, s\rangle^2$ is also independent of G_k. Inequality (v) follows from $\mathbb{P}(G_k) > \frac{1}{2}$ with proper conditions and the fact that $\sum_{i=1}^n \langle x_i^j, s\rangle^2 \sim \text{SubE}(4n, 4)$ and $\sum_{i=1}^n \langle x_i^j, \theta_k^* - \theta_1^* \rangle^2 \sim \text{SubE}(4n\|\theta_k - \theta_1^*\|^4, 4\|\theta_k^* - \theta_1^*\|^2)$. Therefore, $\|W\|_{\psi_1} = \mathcal{O}(\frac{\|\theta_k^* - \theta_1^*\|}{\sqrt{n}} \exp(-n))$. By Lemma 1,

$$
\begin{aligned}
t &= \mathcal{O}(\|\theta_k^* - \theta_1^*\| \exp(-n) \sqrt{p \vee \frac{\log(dK^2T/\delta)}{m}} \sqrt{\frac{d\log(dK^2T/\delta)}{mn}}) \\
&= \mathcal{O}(\Delta_{\max} \sqrt{\frac{1}{K}} \exp(-n) \sqrt{\frac{d\log(dK^2T/\delta)}{mn}}) \quad \text{since } p < \frac{1}{K} \text{ and we assume } m > K\log(dK^2T/\delta)
\end{aligned}
$$

Therefore, with probability at least $1 - 3\delta/(K^2T)$,

$$
\begin{aligned}
&\|\frac{1}{m}\sum_{j=1}^m \frac{1}{n}w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \langle x_i^j, \theta_k^* - \theta_1^* \rangle \mathbb{1}_{j \in E_k \cap G_k} - \frac{1}{mn}\mathbb{E}[w(\widetilde{\theta})_1 \sum_{i=1}^n x_i \langle x_i, \theta_k^* - \theta_1^* \rangle \mathbb{1}_{E_k \cap G_k}]\| \\
&= \mathcal{O}(\Delta_{\max} \sqrt{\frac{1}{K}} \exp(-n) \sqrt{\frac{d\log(dK^2T/\delta)}{mn}})
\end{aligned}
$$

C.1.2 Analysis of $\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^n x_i^j\langle x_i^j,\theta_k^*-\theta_1^*\rangle\mathbb{1}_{j\in E_k\cap G_{k,1}^c}$

Recall that $G_{k,1}^c=\{\sum_{i=1}^n\langle x_i,\theta_k^*-\theta_1^*\rangle^2\leq\frac{320\sigma^2 n}{3}\}$.
Similarly as before, define $Z_j:=\mathbb{1}_{j\in E_k\cap G_{k,1}^c}$, $p:=\mathbb{P}(E_k\cap G_{k,1}^c)\leq\mathbb{P}(G_{k,1}^c)\leq\exp(-\frac{n}{16})$ and $W_j:=\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^n x_i^j\langle x_i^j,\theta_k^*-\theta_1^*\rangle|E_k\cap G_{k,1}^c$.
$\forall s\in\mathcal{S}^{d-1}$

$$
\begin{aligned}
&\|\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^n\langle x_i^j,s\rangle\langle x_i^j,\theta_k^*-\theta_1^*\rangle|E_k\cap G_{k,1}^c\|_{\psi_1}\\
&=n^{-1}\sup_{p\geq1}p^{-1}\mathbb{E}_{\mathcal{D}_k^*}[|w^j(\widetilde{\theta})_1\sum_{i=1}^n\langle x_i^j,s\rangle\langle x_i^j,\theta_k^*-\theta_1^*\rangle|^p|G_{k,1}^c]^{1/p}\\
&\leq n^{-1}\sup_{p\geq1}p^{-1}\mathbb{E}_{\mathcal{D}_k^*}[|(\sum_{i=1}^n\langle x_i^j,s\rangle^2)^{1/2}(\sum_{i=1}^n\langle x_i^j,\theta_k^*-\theta_1^*\rangle^2)^{1/2}|^p|G_{k,1}^c]^{1/p}\\
&\leq n^{-1}\sup_{p\geq1}p^{-1}\sqrt{\frac{320\sigma^2 n}{3}}\mathbb{E}_{\mathcal{D}_k^*}[(\sum_{i=1}^n\langle x_i^j,s\rangle^2)^{p/2}|G_{k,1}^c]^{1/p}\\
&=n^{-1}\sup_{p\geq1}p^{-1}\sqrt{\frac{320\sigma^2 n}{3}}\sqrt{\mathbb{E}_{\mathcal{D}_k^*}[(\sum_{i=1}^n\langle x_i^j,s\rangle^2)^{p/2}|G_{k,1}^c]^{2/p}}\\
&=n^{-1}\sup_{p\geq1}p^{-1}\sqrt{\frac{320\sigma^2 n}{3}}\mathcal{O}(n^{1/4}p^{1/2})\text{ by Lemma 2}\\
&\leq\mathcal{O}(n^{-1/4}\sigma)
\end{aligned}
$$

Therefore, $\|W\|_{\psi_1}=\mathcal{O}(n^{-1/4}\sigma)$.
Then

$$
\begin{aligned}
t&=\mathcal{O}(n^{1/4}\sigma\sqrt{p\vee\frac{\log(dK^2T/\delta)}{m}}\sqrt{\frac{d\log(dK^2T/\delta)}{mn}})\\
&=\begin{cases}\mathcal{O}(\sigma\exp(-n)\sqrt{\frac{d}{mn}\log(dK^2T/\delta)}) & \text{if } m\geq\Omega(\exp(n)\log(dK^2T/\delta))\\ \mathcal{O}(\sigma\sqrt{\frac{1}{K}}\sqrt{\frac{d}{mn}\log(dK^2T/\delta)}) & \text{if } \mathcal{O}(\exp(n)\log(dK^2T/\delta))\geq m\geq\Omega(K\sqrt{n}\log(dK^2T/\delta)).\end{cases}
\end{aligned}
$$

If $m\leq\mathcal{O}(K\sqrt{n}\log(dK^2T/\delta)$, we directly apply Lemma 6 to $\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^m x_i^j\langle x_i^j,\theta_k^*-\theta_1^*\rangle\mathbb{1}_{j\in E_k\cap G_{k,1}^c}$.
Notice that $\forall s\in\mathcal{S}^{d-1}$,

$$
\begin{aligned}
&\|\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^m\langle x_i^j,s\rangle\langle x_i^j,\theta_k^*-\theta_1^*\rangle\mathbb{1}_{j\in E_k\cap G_{k,1}^c}\|_{\psi_1}\\
&\leq\frac{1}{n}\|(\sum_{i=1}^m\langle x_i^j,s\rangle^2)^{1/2}(\sum_{i=1}^m\langle x_i^j,\theta_k^*-\theta_1^*\rangle^2)^{1/2}\|_{\psi_1}\\
&\leq\frac{1}{n}\|\sum_{i=1}^m\langle x_i^j,s\rangle^2\|_{\psi_1}^{1/2}\|\sum_{i=1}^n\langle x_i^j,\theta_k^*-\theta_1^*\rangle^2\|_{\psi_1}^{1/2}\text{ by Lemma 7 and Lemma 8}\\
&=\mathcal{O}(\frac{\Delta_{\max}}{\sqrt{n}})
\end{aligned}
$$

Therefore, with probability at least $1 - \frac{\delta}{K^2 T}$, when $m \le \mathcal{O}(K\sqrt{n}\log(dK^2T/\delta))$, the statistical error is bounded by $\mathcal{O}(\Delta_{\max}\sqrt{\frac{d}{mn}\log(dK^2T/\delta)})$.

C.1.3 Analysis of $\frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \langle x_i^j, \theta_k^* - \theta_1^* \rangle \mathbb{1}_{j \in E_k \cap G_{k,2}^c}$

Recall that $G_{k,2}^c = \{\max\{\sum_{i=1}^n \langle x_i, \theta_k - \theta_k^* \rangle^2, \sum_{i=1}^n \langle x_i, \theta_1 - \theta_1^* \rangle^2\} \le \frac{1}{16}\sum_{i=1}^n \langle x_i, \theta_k^* - \theta_1^* \rangle^2\}$.
Define $Z_j := \mathbb{1}_{j \in E_k \cap G_{k,2}^c}$, $p := \mathbb{P}(E_k \cap G_{k,2}^c) \le \mathbb{P}(G_{k,2}^c) = \mathcal{O}(\exp(-C_\alpha n))$ and
$W_j := \frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \langle x_i^j, \theta_k^* - \theta_1^* \rangle | E_k \cap G_{k,2}^c$.
Then $\forall s \in \mathcal{S}^{d-1}$,

$$
\begin{aligned}
&\|\frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n \langle x_i^j, s \rangle \langle x_i^j, \theta_k^* - \theta_1^* \rangle | E_k \cap G_{k,2}^c\|_{\psi_1} \\
&\le \frac{1}{n}\|\sum_{i=1}^n \langle x_i^j, s \rangle \langle x_i^j, \theta_k^* - \theta_1^* \rangle | G_{k,2}^c\|_{\psi_1} \quad \text{by the independence of } x_i^j \text{ on } E_k \\
&\le \frac{1}{n}\|(\sum_{i=1}^n \langle x_i^j, s \rangle^2)^{1/2} (\sum_{i=1}^n \langle x_i^j, \theta_k^* - \theta_1^* \rangle^2)^{1/2} | G_{k,2}^c\|_{\psi_1} \\
&\le \frac{1}{n}\|(\sum_{i=1}^n \langle x_i^j, s \rangle^2)^{1/2} | G_{k,2}^c\|_{\psi_2} \|(\sum_{i=1}^n \langle x_i^j, \theta_k^* - \theta_1^* \rangle^2)^{1/2} | G_{k,2}^c\|_{\psi_2} \quad \text{by Lemma 8}
\end{aligned}
$$

Note that on $G_{k,2}^c$, $\sum_{i=1}^n \langle x_i^j, \theta_k^* - \theta_1^* \rangle^2 \le 16\sum_{i=1}^n \langle x_i^j, \theta_k - \theta_k^* \rangle^2 + 16\sum_{i=1}^n \langle x_i^j, \theta_1 - \theta_1^* \rangle^2$. Then

$$
\begin{aligned}
&\|(\sum_{i=1}^n \langle x_i^j, \theta_k^* - \theta_1^* \rangle^2)^{1/2} | G_{k,2}^c\|_{\psi_2} \\
&\le \|(16\sum_{i=1}^n \langle x_i^j, \theta_k - \theta_k^* \rangle^2 + 16\sum_{i=1}^n \langle x_i^j, \theta_1 - \theta_1^* \rangle^2)^{1/2} | G_{k,2}^c\|_{\psi_2} \\
&= \|16\sum_{i=1}^n \langle x_i^j, \theta_k - \theta_k^* \rangle^2 + 16\sum_{i=1}^n \langle x_i^j, \theta_1 - \theta_1^* \rangle^2\|_{\psi_1}^{1/2} \quad \text{by Lemma 7} \\
&\le (16\|\sum_{i=1}^n \langle x_i^j, \theta_k - \theta_k^* \rangle^2\|_{\psi_1} + 16\|\sum_{i=1}^n \langle x_i^j, \theta_1 - \theta_1^* \rangle^2\|_{\psi_1})^{1/2} \\
&= \mathcal{O}(n^{1/4} D_M) \quad \text{where } D_M = \max_{k \in [K]} \|\theta_k - \theta_k^*\|
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
\|(\sum_{i=1}^n \langle x_i^j, s \rangle^2)^{1/2} | G_{k,2}^c\|_{\psi_2} &= \|\sum_{i=1}^n \langle x_i, s \rangle^2 | G_{k,2}^c\|_{\psi_1}^{1/2} \quad \text{by Lemma 7} \\
&= (\sup_{p \ge 1} p^{-1} \mathbb{E}[(\sum_{i=1}^n \langle x_i, s \rangle^2)^p | G_{k,2}^c]^{1/p})^{1/2} \\
&= (\sup_{p \ge 1} p^{-1} \mathcal{O}(np))^{1/2} \quad \text{by Lemma 3} \\
&= \mathcal{O}(\sqrt{n})
\end{aligned}
$$

Therefore, $\|W\|_{\psi_1} = \mathcal{O}(n^{-1/4} D_M)$. Then by Lemma 1

$$
\begin{aligned}
t &= \mathcal{O}(n^{1/4} D_M \sqrt{p \vee \frac{\log(dK^2T/\delta)}{m}} \sqrt{\frac{d\log(dK^2T/\delta)}{mn}}) \\
&= \begin{cases} \mathcal{O}(D_M \exp(-C_\alpha n)\sqrt{\frac{d}{mn}\log(dK^2T/\delta)}) & \text{if } m \geq \Omega(\exp(C_\alpha n)\log(dK^2T/\delta)) \\ \mathcal{O}(D_M \sqrt{\frac{1}{K}}\sqrt{\frac{d}{mn}\log(dK^2T/\delta)}) & \text{if } \mathcal{O}(\exp(C_\alpha n)\log(dK^2T/\delta)) \geq m \geq \Omega(K\sqrt{n}\log(dK^2T/\delta)). \\ \mathcal{O}(D_M \sqrt{\frac{1}{K}}\sqrt{\frac{d}{m\sqrt{n}}\log(dK^2T/\delta)}) & \text{if } \mathcal{O}(K\sqrt{n}\log(dK^2T/\delta)) \geq m \geq \Omega(K\log(dK^2T/\delta)) \end{cases}
\end{aligned}
$$

C.1.4 Analysis of $\frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \langle x_i^j, \theta_k^* - \theta_1^* \rangle \mathbb{1}_{j \in E_k \cap G_3^c}$

Recall that $G_3^c = \{\sum_{i=1}^n \varepsilon_i^2 \geq 2\sigma^2 n\}$. Define $Z_j := \mathbb{1}_{j \in E_k \cap G_3^c}$, $p := \mathbb{P}(E_k \cap G_3^c) \leq \mathbb{P}(G_3^c) \leq \exp(-\frac{n}{16})$, and $W_j := \frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \langle x_i^j, \theta_k^* - \theta_1^* \rangle | E_k \cap G_3^c$.
Then $\forall s \in \mathcal{S}^{d-1}$,

$$
\begin{aligned}
&\|\frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n \langle x_i^j, s\rangle \langle x_i^j, \theta_k^* - \theta_1^* \rangle | E_k \cap G_3^c\|_{\psi_1} \\
&\leq \frac{1}{n}\|(\sum_{i=1}^n \langle x_i^j, s\rangle^2)^{1/2} (\sum_{i=1}^n \langle x_i^j, \theta_k^* - \theta_1^* \rangle^2)^{1/2} | E_k \cap G_3^c\|_{\psi_1} \quad \text{by Cauchy-Schwarz inequality} \\
&\leq \frac{1}{n}\|\sum_{i=1}^n \langle x_i^j, s\rangle^2 | E_k \cap G_3^c\|_{\psi_1}^{1/2} \|\sum_{i=1}^n \langle x_i^j, \theta_k^* - \theta_1^* \rangle^2 | E_k \cap G_3^c\|_{\psi_1}^{1/2} \quad \text{by Lemma 7 and 8} \\
&= \frac{1}{n}\|\sum_{i=1}^n \langle x_i^j, s\rangle^2\|_{\psi_1}^{1/2} \|\sum_{i=1}^n \langle x_i^j, \theta_k^* - \theta_1^* \rangle^2\|_{\psi_1}^{1/2} \quad \text{by the independence of every term on } E_k \text{ and } G_3^c \\
&= \mathcal{O}(n^{-1/2}\|\theta_k^* - \theta_1^*\|)
\end{aligned}
$$

Therefore, $\|W\|_{\psi_1} = \mathcal{O}(n^{-1/2}\Delta_{\max})$.
Again, by Lemma 1,

$$
\begin{aligned}
t &= \mathcal{O}(\Delta_{\max} \sqrt{p \vee \frac{\log(dK^2T/\delta)}{m}} \sqrt{\frac{d\log(dK^2T/\delta)}{mn}}) \\
&= \begin{cases} \mathcal{O}(\Delta_{\max} \exp(-n)\sqrt{\frac{d\log(dK^2T/\delta)}{mn}}) & \text{if } m \geq \Omega(\exp(n)\log(dK^2T/\delta)) \\ \mathcal{O}(\Delta_{\max} \sqrt{\frac{1}{K\sqrt{n}}}\sqrt{\frac{d\log(dK^2T/\delta)}{mn}}) & \text{if } \mathcal{O}(\exp(n)\log(dK^2T/\delta)) \geq m \geq \Omega(K\sqrt{n}\log(dK^2T/\delta)) \\ \mathcal{O}(\Delta_{\max} \sqrt{\frac{1}{K}}\sqrt{\frac{d\log(dK^2T/\delta)}{mn}}) & \text{if } \mathcal{O}(K\sqrt{n}\log(dK^2T/\delta)) \geq m \geq \Omega(K\log(dK^2T/\delta)). \end{cases}
\end{aligned}
$$

C.1.5 Analysis of $\frac{1}{n} w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \varepsilon_i^j \mathbb{1}_{j \in E_k \cap G_k}$

Define $Z_j := \mathbb{1}_{E_k \cap G_k}, p := \mathbb{P}(E_k \cap G_k) \leq \mathbb{P}(E_k) = \frac{1}{K}$ and $W_j := \frac{1}{n} w(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \varepsilon_i^j | E_k \cap G_k$.
Then $\forall s \in \mathcal{S}^{d-1}$,

$$
\|\frac{1}{n} w(\widetilde{\theta})_1 \sum_{i=1}^n \langle x_i^j, s\rangle \varepsilon_i^j | E_k \cap G_k\|_{\psi_1}
$$

$$
\begin{aligned}
&\leq \frac{1}{n}\exp(-2n)\|\sum_{i=1}^{n}\langle x_i^j, s\rangle\varepsilon_i^j|E_k\cap G_k\|_{\psi_1} \quad \text{since } w(\widetilde{\theta})_1\leq \exp(-2n) \text{ on event } G_k\\
&\overset{\text{(i)}}{\leq} \frac{1}{n}\exp(-2n)\sup_{p\geq 1}p^{-1}\sqrt{\mathbb{E}[(\sum_{i=1}^{n}\langle x_i^j, s\rangle^2)^p]^{1/p}}\sqrt{\mathbb{P}(G_k)^{-1}\mathbb{E}[(\sum_{i=1}^{n}\varepsilon_i^{j2})^p]^{1/p}}\\
&\overset{\text{(ii)}}{=} \mathcal{O}(\exp(-2n)n^{-1/2}\sigma)
\end{aligned}
$$

Notice that (i) follows from the same analysis in Appendix C.1.1, and (ii) follows from $\mathbb{P}(G_k)>\frac{1}{2}$ with proper assumptions and the fact that $\sum_{i=1}^{n}\langle x_i^j, s\rangle^2\sim \mathrm{SubE}(4n,4)$ and $\sum_{i=1}^{n}\varepsilon_i^{j2}\sim\mathrm{SubE}(4n\sigma^4, 4\sigma^2)$. Therefore, $\|W\|_{\psi_1}=\mathcal{O}(\frac{\sigma}{\sqrt{n}}\exp(-n))$.
Then

$$
\begin{aligned}
t &= \mathcal{O}(\sigma\exp(-n)\sqrt{p\vee\frac{\log(dK^2T/\delta)}{m}}\sqrt{\frac{d\log(dK^2T/\delta)}{mn}})\\
&= \mathcal{O}(\sigma\exp(-n)\sqrt{\frac{1}{K}}\sqrt{\frac{d}{mn}\log(dK^2T/\delta)}) \quad \text{by the assumption } m\geq\Omega(K\log(dK^2T/\delta)).
\end{aligned}
$$

C.1.6 Analysis of $\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^{n}x_i^j\varepsilon_i^j\mathbb{1}_{j\in E_k\cap G_{k,1}^c}$

Define $Z_j:=\mathbb{1}_{E_k\cap G_{k,1}^c}, p:=\mathbb{P}(E_k\cap G_{k,1}^c)\leq\mathbb{P}(G_{k,1}^c)\leq\exp(-\frac{n}{16})$ and $W_j:=\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^{n}x_i^j\varepsilon_i^j|E_k\cap G_{k,1}^c$. Then $\forall s\in\mathcal{S}^{d-1}$,

$$
\begin{aligned}
&\|\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^{n}\langle x_i^j, s\rangle\varepsilon_i^j|E_k\cap G_{k,1}^c\|_{\psi_1}\\
&\leq \|(\sum_{i=1}^{n}\langle x_i, s\rangle^2)^{1/2}(\sum_{i=1}^{n}\varepsilon_i^{j2})^{1/2}|E_k\cap G_{k,1}^c\|_{\psi_1}\\
&\leq \|\sum_{i=1}^{n}\langle x_i, s\rangle^2|G_{k,1}^c\|_{\psi_1}^{1/2}\|\sum_{i=1}^{n}\varepsilon_i^{j,2}\|_{\psi_1}^{1/2} \quad \text{by the independence of } x,\varepsilon \text{ on } E_k\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{and the independence of } \varepsilon \text{ on } G_{k,1}^c
\end{aligned}
$$

Note that $\|\sum_{i=1}^{n}\varepsilon_i^{j2}\|_{\psi_1}^{1/2}=\mathcal{O}(\sigma n^{1/4})$ by the analysis before and

$$
\begin{aligned}
\|\sum_{i=1}^{n}\langle x_i, s\rangle^2|G_{k,1}^c\|_{\psi_1}^{1/2} &= \sqrt{\sup_{p\geq 1}p^{-1}\mathbb{E}[(\sum_{i=1}^{n}\langle x_i, s\rangle^2)^p|G_{k,1}^c]^{1/p}}\\
&= \sqrt{\sup_{p\geq 1}p^{-1}\mathcal{O}(\sqrt{n}p)} \quad \text{by Lemma 2}\\
&= \mathcal{O}(n^{1/4})
\end{aligned}
$$

Therefore, $\|W\|_{\psi_1}=\mathcal{O}(\frac{\sigma}{\sqrt{n}})$.
Then by Lemma 1,

$$
t=\mathcal{O}(\sigma\sqrt{p\vee\frac{\log(dK^2T/\delta)}{m}}\sqrt{\frac{d\log(dK^2T/\delta)}{mn}})
$$

$$= \begin{cases} \mathcal{O}(\sigma \exp(-n)\sqrt{\frac{d\log(dK^2T/\delta)}{mn}}) & \text{if } m \geq \Omega(\exp(n)\log(dK^2T/\delta)) \\ \mathcal{O}(\sigma\sqrt{\frac{1}{K\sqrt{n}}}\sqrt{\frac{d\log(dK^2T/\delta)}{mn}}) & \text{if } \mathcal{O}(\exp(n)\log(dK^2T/\delta)) \geq m \geq \Omega(K\sqrt{n}\log(dK^2T/\delta)) \\ \mathcal{O}(\sigma\sqrt{\frac{1}{K}}\sqrt{\frac{d\log(dK^2T/\delta)}{mn}}) & \text{if } \mathcal{O}(K\sqrt{n}\log(dK^2T/\delta)) \geq m \geq \Omega(K\log(dK^2T/\delta)). \end{cases}$$

C.1.7 Analysis of $\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^n x_i^j\varepsilon_i^j\mathbb{1}_{j\in E_k\cap G_{k,2}^c}$

Define $Z_j := \mathbb{1}_{E_k\cap G_{k,2}^c}, p := \mathbb{P}(E_k\cap G_{k,2}^c) \leq \mathbb{P}(G_{k,2}^c) \leq \mathcal{O}(\exp(-C_\alpha n))$ and $W_j := \frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^n x_i^j\varepsilon_i^j|E_k\cap G_{k,2}^c$. Then $\forall s\in\mathcal{S}^{d-1}$,

$$\begin{aligned} &\|\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^n\langle x_i^j,s\rangle\varepsilon_i^j|E_k\cap G_{k,2}^c\|_{\psi_1} \\ &\leq \frac{1}{n}\|\sum_{i=1}^n\langle x_i^j,s\rangle^2|G_{k,2}^c\|_{\psi_1}^{1/2}\|\sum_{i=1}^n\varepsilon_i^{j2}\|_{\psi_1}^{1/2} \quad \text{by the independence of } \varepsilon \text{ on } G_{k,2}^c \\ &= \sqrt{\sup_{p\geq 1}p^{-1}\mathbb{E}[(\sum_{i=1}^n\langle x_i^j,s\rangle^2)^p|G_{k,2}^c]^{1/p}}\mathcal{O}(n^{1/4}\sigma) \\ &= \mathcal{O}(\sqrt{n})\mathcal{O}(n^{1/4}\sigma) \quad \text{by Lemma 3} \\ &= \mathcal{O}(n^{-1/4}\sigma) \end{aligned}$$

Therefore, $\|W\|_{\psi_1} = \mathcal{O}(n^{-1/4}\sigma)$.
Then, by Lemma 1,

$$\begin{aligned} t &= \mathcal{O}(\sigma n^{1/4}\sqrt{p\vee\frac{\log(dK^2T/\delta)}{m}}\sqrt{\frac{d\log(dK^2T/\delta)}{mn}}) \\ &= \begin{cases} \mathcal{O}(\sigma\exp(-C_\alpha n)\sqrt{\frac{d}{mn}\log(dK^2T/\delta)}) & \text{if } m\geq\Omega(\exp(C_\alpha n)\log(dK^2T/\delta)) \\ \mathcal{O}(\sigma\sqrt{\frac{1}{K}}\sqrt{\frac{d}{mn}\log(dK^2T/\delta)}) & \text{if } \mathcal{O}(\exp(C_\alpha n)\log(dK^2T/\delta))\geq m\geq\Omega(K\sqrt{n}\log(dK^2T/\delta)). \end{cases} \end{aligned}$$

If $m\leq\mathcal{O}(K\sqrt{n}\log(dK^2T/\delta))$, similarly as before, we directly apply Lemma 6 to $\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^n x_i^j\varepsilon_i^j\mathbb{1}_{j\in E_k\cap G_{k,2}^c}$. Notice that $\forall s\in\mathcal{S}^{d-1}$,

$$\begin{aligned} &\|\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^n\langle x_i^j,s\rangle\varepsilon_i^j\mathbb{1}_{j\in E_k\cap G_{k,2}^c}\|_{\psi_1} \\ &\leq\frac{1}{n}\|\sum_{i=1}^n\langle x_i^j,s\rangle^2\|_{\psi_1}^{1/2}\|\sum_{i=1}^n\varepsilon_i^{j2}\|_{\psi_1}^{1/2} = \mathcal{O}(\frac{\sigma}{\sqrt{n}}) \end{aligned}$$

Therefore, with probability at least $1-\frac{\delta}{K^2T}$, the statistical error is bounded by $\sigma\sqrt{\frac{d}{mn}\log(dK^2T/\delta)}$.

C.1.8 Analysis of $\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^n x_i^j\varepsilon_i^j\mathbb{1}_{j\in E_k\cap G_3^c}$

Define $Z_j := \mathbb{1}_{E_k\cap G_3^c}, p := \mathbb{P}(E_k\cap G_3^c)\leq\mathbb{P}(G_3^c)\leq\exp(-\frac{n}{16})$ and $W_j := \frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^n x_i^j\varepsilon_i^j|E_k\cap G_3^c$. Then $\forall s\in\mathcal{S}^{d-1}$,

$$\|\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^n\langle x_i^j,s\rangle\varepsilon_i^j|E_k\cap G_3^c\|_{\psi_1}$$

$$\leq \| \sum_{i=1}^{n} \langle x_i^j, s\rangle^2 \|_{\psi_1}^{1/2} \| \sum_{i=1}^{n} \varepsilon_i^{j2} | G_3^c \|_{\psi_1}^{1/2} \quad \text{by the independence of } x \text{ on } G_3^c$$

$$= \mathcal{O}(n^{1/4}) \| \sum_{i=1}^{n} \varepsilon_i^{j2} | G_3^c \|_{\psi_1}^{1/2}$$

Note that $\| \sum_{i=1}^n \varepsilon_i^{j2} | G_3^c \|_{\psi_1} = \sup_{p\geq 1} p^{-1} \mathbb{E}[| \sum_{i=1}^n \varepsilon_i^{j2}|^p \mathbb{1}_{G_3^c}]^{1/p} \mathbb{P}(G_3^c)^{-1/p}$. In order to bound this sub-exponential norm, we decompose G_3^c into $\{12\sigma^2 n \geq \sum_{i=1}^n \varepsilon_i^{j2} \geq 2\sigma^2 n\}$ and $\{\sum_{i=1}^n \varepsilon_i^{j2} \geq 12\sigma^2 n\}$. Then

$$\mathbb{E}[| \sum_{i=1}^{n} \varepsilon_i^{j2}|^p \mathbb{1}_{G_3^c}]^{1/p}$$

$$\leq \mathbb{E}[| \sum_{i=1}^{n} \varepsilon_i^{j2}|^p \mathbb{1}_{12\sigma^2 n \geq \sum_{i=1}^n \varepsilon_i^{j2} \geq 2\sigma^2 n}]^{1/p} + \mathbb{E}[| \sum_{i=1}^{n} \varepsilon_i^{j2}|^p \mathbb{1}_{\sum_{i=1}^n \varepsilon_i^{j2} \geq 12\sigma^2 n}]^{1/p} \quad \text{by Minkowski inequality}$$

$$\leq (12\sigma^2 n)\mathbb{P}(\sum_{i=1}^{n} \varepsilon_i^{j2} \geq 2\sigma^2 n)^{1/p} + \mathbb{E}[| \sum_{i=1}^{n} \varepsilon_i^{j2}|^{2p}]^{\frac{1}{2p}} \mathbb{P}(\sum_{i=1}^{n} \varepsilon_i^{j2} \geq 12\sigma^2 n)^{\frac{1}{2p}} \quad \text{by Cauchy-Schwarz inequality}$$

$$= (12\sigma^2 n)\mathbb{P}(G_3^c)^{1/p} + \mathcal{O}(p\sqrt{n}\sigma^2)\mathbb{P}(\sum_{i=1}^{n} \varepsilon_i^{j2} \geq 12\sigma^2 n)^{\frac{1}{2p}} \quad \text{since } \sum_{i=1}^{n} \varepsilon_i^{j2} \sim \text{SubE}(4n\sigma^4, 4\sigma^2).$$

Note that

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^{n} \varepsilon_i^{j2} \geq 12\sigma^2 n) &\leq \exp(-n(\frac{\sqrt{92}-2}{4})^2) \quad \text{by Lemma 5} \\
&\leq \exp(-3n)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}(G_3^c) &= \mathbb{P}(\sum_{i=1}^{n} (\frac{\varepsilon_i^j}{\sigma})^2 \geq 2n) = \mathbb{P}(\sqrt{\sum_{i=1}^{n} (\frac{\varepsilon_i^j}{\sigma})^2} \geq \sqrt{2n}) \\
&\geq \mathbb{P}(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i^j}{\sigma} \geq \sqrt{2n}) \quad \text{since } \sqrt{\sum_{i=1}^{n} (\frac{\varepsilon_i^j}{\sigma})^2} \geq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\varepsilon_i^j}{\sigma} \\
&\overset{\text{(i)}}{\geq} \frac{1}{\sqrt{2\pi}} \frac{\sqrt{2n}}{2n+1} \exp(-n)
\end{aligned}$$

where (i) follows from the lower bound of complementary cumulative distribution function of standard Gaussian $\Phi^c(t) \geq \frac{1}{\sqrt{2\pi}} \frac{t}{t^2+1} \exp(-t^2/2)$.
Therefore,

$$\mathbb{P}(\sum_{i=1}^{n} \varepsilon_i^{j2} \geq 12\sigma^2 n)^{1/2} \mathbb{P}(G_3^c)^{-1} \leq \exp(-\frac{3n}{2})\sqrt{2\pi}\frac{2n+1}{\sqrt{2n}} \exp(n) = \mathcal{O}(\exp(-n))$$

Then

$$\| \sum_{i=1}^{n} \varepsilon_i^{j2} | G_3^c \|_{\psi_1} \leq \sup_{p\geq 1} p^{-1} \{(12\sigma^2 n)\mathbb{P}(G_3^c)^{1/p} + \mathcal{O}(p\sqrt{n}\sigma^2)\mathbb{P}(\sum_{i=1}^{n} \varepsilon_i^{j2} \geq 12\sigma^2 n)^{\frac{1}{2p}}\} \mathbb{P}(G_3^c)^{-1/p}$$

$$
\begin{aligned}
&= \sup_{p\geq 1} p^{-1}\{12\sigma^2 n + \mathcal{O}(p\sqrt{n}\sigma^2)(\mathbb{P}(\sum_{i=1}^n \varepsilon_i^{j2} \geq 12\sigma^2 n)^{1/2}\mathbb{P}(G_3^c)^{-1})^{1/p}\} \\
&\leq \sup_{p\geq 1} p^{-1}\{12\sigma^2 n + \mathcal{O}(p\sqrt{n}\sigma^2)\mathcal{O}(\exp(-\frac{n}{p}))\} \\
&= \mathcal{O}(n\sigma^2)
\end{aligned}
$$

Therefore, $\|W\|_{\psi_1} = \mathcal{O}(n^{-1/4}\sigma)$.
Then, by Lemma 1,

$$
\begin{aligned}
t &= \mathcal{O}(\sigma n^{1/4}\sqrt{p \vee \frac{\log(dK^2T/\delta)}{m}}\sqrt{\frac{d\log(dK^2T/\delta)}{mn}}) \\
&= \begin{cases} \mathcal{O}(\sigma \exp(-n)\sqrt{\frac{d}{mn}\log(dK^2T/\delta)}) & \text{if } m \geq \Omega(\exp(n)\log(dK^2T/\delta)) \\ \mathcal{O}(\sigma\sqrt{\frac{1}{K}}\sqrt{\frac{d}{mn}\log(dK^2T/\delta)}) & \text{if } \mathcal{O}(\exp(n)\log(dK^2T/\delta)) \geq m \geq \Omega(K\sqrt{n}\log(dK^2T/\delta)). \end{cases}
\end{aligned}
$$

If $m \leq \mathcal{O}(K\sqrt{n}\log(dK^2T/\delta))$, by applying Lemma 6 the same way as before, with probability at least $1 - \frac{\delta}{K^2T}$, the statistical error is bounded by $\mathcal{O}(\sigma\sqrt{\frac{d}{mn}\log(dK^2T/\delta)})$.

C.1.9 Analysis of $\frac{1}{n}w^j(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \varepsilon_i^j \mathbb{1}_{j\in E_1}$

Define $Z_j := \mathbb{1}_{E_1}, p := \mathbb{P}(E_1) = \frac{1}{K}$ and $W_j := \frac{1}{n}w(\widetilde{\theta})_1 \sum_{i=1}^n x_i^j \varepsilon_i^j | E_1$.
Then $\forall s \in \mathcal{S}^{d-1}$,

$$
\begin{aligned}
\|\frac{1}{n}w(\widetilde{\theta})_1 \sum_{i=1}^n \langle x_i^j, s\rangle \varepsilon_i^j | E_1\|_{\psi_1} &\leq \|\sum_{i=1}^n \langle x_i^j, s\rangle \varepsilon_i^j\|_{\psi_1} \quad \text{by the independence of } x, \varepsilon \text{ on } E_1 \\
&= \mathcal{O}(\frac{\sigma}{\sqrt{n}})
\end{aligned}
$$

Therefore, $\|W\|_{\psi_1} = \mathcal{O}(\frac{\sigma}{\sqrt{n}})$. Then, by Lemma 1,

$$
\begin{aligned}
t &= \mathcal{O}(\sigma\sqrt{p \vee \frac{\log(dK^2T/\delta)}{m}}\sqrt{\frac{d\log(dK^2T/\delta)}{mn}}) \\
&= \mathcal{O}(\sigma\sqrt{\frac{1}{K}}\sqrt{\frac{d\log(dK^2T/\delta)}{mn}}) \quad \text{by the assumption } m \geq \Omega(K\log(dK^2T/\delta).
\end{aligned}
$$

Putting all the terms together, taking union over K elements and changing 3δ to δ, we have the following three cases with probability at least $1 - \delta/(KT)$.
If $m \geq \Omega(\exp(n)\log(dK^2T/\delta))$,

$$
\begin{aligned}
\|B_m - B\| \leq \sqrt{\frac{d}{mn}\log(dK^2T/\delta)}\{&\Delta_{\max}\sqrt{K}\exp(-n) + \sigma K\exp(-n+\log n) + \\
&D_M K\exp(-C_\alpha n + \log n) + \Delta_{\max} K\exp(-n) + \sigma\exp(-n)\sqrt{K} \\
&+ \sigma K\exp(-n) + \sigma K\exp(-n+\log n) + \sigma\sqrt{\frac{1}{K}}\} \\
= \sqrt{\frac{d}{mn}\log(dK^2T/\delta)}&\mathcal{O}(\Delta_{\max}\exp(-n) + D_M\exp(-C_\alpha n) + \sigma\sqrt{\frac{1}{K}})
\end{aligned}
$$

If $\mathcal{O}(\exp(n)\log(dK^2T/\delta)) \geq m \geq \Omega(K\sqrt{n}\log(dK^2T/\delta))$,

$$\begin{aligned}\|B_m - B\| &\leq \sqrt{\frac{d}{mn}\log(dK^2T/\delta)}\mathcal{O}(\Delta_{\max}\sqrt{K}\exp(-n) + \sigma\sqrt{K} + D_M\sqrt{K}\\ &\quad + \Delta_{\max}\sqrt{\frac{K}{\sqrt{n}}} + \sigma\exp(-n)\sqrt{K} + \sigma\sqrt{\frac{K}{\sqrt{n}}} + 2\sigma\sqrt{K} + \sigma\sqrt{\frac{1}{K}})\\ &= \sqrt{\frac{d}{mn}\log(dK^2T/\delta)}\mathcal{O}(\sigma\sqrt{K} + \Delta_{\max}\sqrt{\frac{K}{\sqrt{n}}} + D_M\sqrt{K})\end{aligned}$$

If $\mathcal{O}(K\sqrt{n}\log(dK^2T/\delta)) \geq m \geq \Omega(K\log(dK^2T/\delta))$,

$$\begin{aligned}\|B_m - B\| &\leq \sqrt{\frac{d}{mn}\log(dK^2T/\delta)}\mathcal{O}(\Delta_{\max}\sqrt{K}\exp(-n) + \Delta_{\max}K + \Delta_{\max}\sqrt{K} + \sigma\exp(-n)\sqrt{K}\\ &\quad + \sigma\sqrt{\frac{1}{K}} + 2\sigma K) + \sqrt{\frac{dK}{m\sqrt{n}}\log(dK^2T/\delta)}D_M\\ &= \sqrt{\frac{dK^2}{mn}\log(dK^2T/\delta)}\mathcal{O}(\sigma + \Delta_{\max}) + \sqrt{\frac{dK}{m\sqrt{n}}\log(dK^2T/\delta)}D_M\end{aligned}$$

C.2 Bound On A_m

Note that $\frac{1}{mn}\sum_{j=1}^m w^j(\widetilde{\theta})_1\sum_{i=1}^n x_i^j x_i^{jT} \succcurlyeq \frac{1}{mn}\sum_{j=1}^m w(\widetilde{\theta})_1\sum_{i=1}^n x_i^j x_i^{jT}\mathbb{1}_{E_1}$. It's enough to bound the deviation of $\frac{1}{mn}\sum_{j=1}^m w^j(\widetilde{\theta})_1\sum_{i=1}^n x_i^j x_i^{jT}\mathbb{1}_{E_1}$ from $\frac{1}{m}\sum_{j=1}^m \mathbb{E}[w(\widetilde{\theta})_1\sum_{i=1}^n x_i x_i^T\mathbb{1}_{E_1}]$. Let $\mathcal{S}_{1/4}^{d-1}$ be the $\frac{1}{4}$-covering net of $\mathcal{S}^{d-1}$. Then by the standard covering net argument,

$$\begin{aligned}&\|\frac{1}{m}\sum_{j=1}^m\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^n x_i^j x_i^{jT}\mathbb{1}_{E_1} - \frac{1}{mn}\sum_{j=1}^m\mathbb{E}[w(\widetilde{\theta})_1\sum_{i=1}^n x_i x_i^T\mathbb{1}_{E_1}]\|_{op}\\ &\leq 2\sup_{s\in\mathcal{S}_{1/4}^{d-1}}|\frac{1}{m}\sum_{j=1}^m\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^n u^T x_i^j x_i^{jT}u\mathbb{1}_{E_1} - \frac{1}{mn}\sum_{j=1}^m\mathbb{E}[w(\widetilde{\theta})_1\sum_{i=1}^n u^T x_i x_i^T u\mathbb{1}_{E_1}]|\\ &\leq 2\sup_{s\in\mathcal{S}_{1/4}^{d-1}}|\frac{1}{m}\sum_{j=1}^m\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^n\langle x_i^j,u\rangle^2\mathbb{1}_{E_1} - \frac{1}{mn}\sum_{j=1}^m\mathbb{E}[w(\widetilde{\theta})_1\sum_{i=1}^n\langle x_i,u\rangle^2\mathbb{1}_{E_1}]|\end{aligned}$$

Let $Z_j := \mathbb{1}_{E_1}, p := \mathbb{P}(E_k) = \frac{1}{K}$ and $W_j := \frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^n\langle x_i^j,u\rangle^2|E_1$.
Then

$$\|W\|_{\psi_1} \leq \frac{1}{n}\|\sum_{i=1}^n\langle x_i^j,u\rangle^2\|_{\psi_1} = \mathcal{O}(\frac{1}{\sqrt{n}})$$

Again, by Lemma 1,

$$\begin{aligned}t &= \mathcal{O}(\sqrt{p\vee\frac{d\log(K^2T/\delta)}{m}}\sqrt{\frac{d\log(K^2T/\delta)}{mn}})\\ &= \mathcal{O}(\sqrt{\frac{1}{K}}\sqrt{\frac{d}{mn}\log(K^2T/\delta)}) \quad \text{by the assumption } m \geq \Omega(K\log(dK^2T/\delta))\end{aligned}$$

Therefore, with probability at least $1-(K^2T/\delta)^{-d}$,

$$|\frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^{n}\langle x_i^j,u\rangle^2\mathbb{1}_{E_1}-\frac{1}{mn}\sum_{j=1}^{m}\mathbb{E}[w(\widetilde{\theta})_1\sum_{i=1}^{n}\langle x_i,u\rangle^2\mathbb{1}_{E_1}]|\leq\mathcal{O}(\sqrt{\frac{1}{K}}\sqrt{\frac{d}{mn}\log(K^2T/\delta)})$$

Now since $|\mathcal{S}_{1/4}^{d-1}|\leq 9^d$, taking union bounds over 9^d elements, we have

$$\begin{aligned}
&\mathbb{P}(\|\frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^{n}x_i^jx_i^{jT}\mathbb{1}_{E_1}-\frac{1}{mn}\sum_{j=1}^{m}\mathbb{E}[w(\widetilde{\theta})_1\sum_{i=1}^{n}x_ix_i^T\mathbb{1}_{E_1}]\|_{op}\geq\mathcal{O}(\sqrt{\frac{1}{K}}\sqrt{\frac{d}{mn}\log(K^2T/\delta)}))\\
&=\mathbb{P}(\sup_{p\in\mathcal{S}_{1/4}^{d-1}}|\frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^{n}\langle x_i^j,u\rangle^2\mathbb{1}_{E_1}-\frac{1}{mn}\sum_{j=1}^{m}\mathbb{E}[w(\widetilde{\theta})_1\sum_{i=1}^{n}\langle x_i,u\rangle^2\mathbb{1}_{E_1}]|\\
&\qquad\geq\mathcal{O}(\sqrt{\frac{1}{K}}\sqrt{\frac{d}{mn}\log(K^2T/\delta)}))\\
&\leq 9^d(K^2T/\delta)^{-d}=(9\delta/(K^2T))^d
\end{aligned}$$

Notice that in Appendix B.2, we have derived $\mathbb{E}_{\mathcal{D}_1^*}[w(\widetilde{\theta})_1\sum_{i=1}^{n}x_ix_i^T]\geq n(1-K\mathcal{O}(\exp(-C_\alpha n)))$. Then

$$\frac{1}{mn}\sum_{j=1}^{m}\mathbb{E}_{\mathcal{D}_1^*}[w(\widetilde{\theta})_1\sum_{i=1}^{n}x_ix_i^T]\succcurlyeq 1-\mathcal{O}(K\exp(-C_\alpha n)).$$

Therefore, by changing $(9\delta)^d$ to δ, with probability at least $1-(\delta/(K^2T))^d$, assuming good initialization, $n>\Omega(\log K)$ and $mn>\Omega(dK\log(dK^2T/\delta))$,

$$\begin{aligned}
\frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}w^j(\widetilde{\theta})_1\sum_{i=1}^{n}x_i^jx_i^{jT}\mathbb{1}_{E_1}&\succcurlyeq(1-\mathcal{O}(K\exp(-C_\alpha n))-\sqrt{\frac{1}{K}}\sqrt{\frac{d}{mn}\log(K^2T/\delta)})I_d\\
&\succcurlyeq\frac{1}{K}I_d
\end{aligned}$$

Therefore, $\|A_m\|\leq K$.

D Proof of Corollary 1

Corollary 2. *(Detailed Version of Corollary 1) Let T and ε be the number of iterations and precision which will be specified for each of the following three cases. Assume mn is large enough such that $\mathcal{O}(\varepsilon)\leq\frac{\alpha\Delta_{\min}}{2}$. Then after T iterations, either $\max_{k\in[K]}\|\widetilde{\theta}_k^{(t)}-\theta_k^*\|\leq\varepsilon$ for some $t=0,1,\ldots,T-1$, or $\max_{k\in[K]}\|\widetilde{\theta}_k^{(T)}-\theta_k^*\|\leq\mathcal{O}(\varepsilon)$ with probability $1-\delta$.*

(i) If $\frac{m}{T}=\Omega(K\exp(n)\log(dK^2T/\delta))$ and $\frac{mn}{T}=\Omega(dK\log(dK^2T/\delta))$, then

$$\varepsilon=\sqrt{\frac{Kd}{mn}\log(dK^2T/\delta)}(\sigma+\Delta_{\max}\exp(-n))$$

and

$$T=\frac{1}{C_\alpha n}\mathcal{O}(\log(\frac{\max_{k\in[K]}\|\widetilde{\theta}_k^{(0)}-\theta_k^*\|}{\varepsilon})).$$

(ii) If $\mathcal{O}(K\exp(n)\log(dK^2T/\delta)) \geq \frac{m}{T} \geq \Omega(K\sqrt{n}\log(dK^2T/\delta))$ and $\frac{mn}{T} = \Omega(dK^{3/2}\log(dK^2T/\delta))$, then

$$\varepsilon = \sqrt{\frac{K^{3/2}d}{mn}\log(dK^2T/\delta)}(\sigma + \Delta_{\max})$$

and

$$T = \frac{1}{C_\alpha n}\mathcal{O}(\log(\frac{\max_{k\in[K]}\|\widetilde{\theta}_k^{(0)} - \theta_k^*\|}{\varepsilon})) = \mathcal{O}(1).$$

Moreover, if $\alpha \leq \frac{\Delta_{\max}}{\Delta_{\min}\sqrt{n}}$, then we take

$$\varepsilon = \sqrt{\frac{K^{3/2}d}{mn}\log(dK^2T/\delta)}(\sigma + \frac{\Delta_{\max}}{\sqrt{n}})$$

and still have

$$T = \frac{1}{C_\alpha n}\mathcal{O}(\log(\frac{\max_{k\in[K]}\|\widetilde{\theta}_k^{(0)} - \theta_k^*\|}{\varepsilon})) = \mathcal{O}(1).$$

(iii) If $\mathcal{O}(K\sqrt{n}\log(dK^2T/\delta)) \geq \frac{m}{T} \geq \Omega(K\log(dK^2T/\delta))$ and $\frac{mn}{T} = \Omega(dK^4\log(dK^2T/\delta))$, then

$$\varepsilon = \sqrt{\frac{dK^4}{mn}\log(dK^2T/\delta)}(\sigma + \Delta_{\max})$$

and

$$T = \mathcal{O}(\log(\frac{\max_{k\in[K]}\|\widetilde{\theta}_k^{(0)} - \theta_k^*\|}{\varepsilon})).$$

Proof. For simplicity, denote $D_M^{(t)} := \max_{k\in[K]}\|\widetilde{\theta}_k^{(t)} - \theta_k^*\|$. For any $t = 0, 1, \ldots, T-1$, if $D_M^{(t)} \leq \varepsilon$, then the corollary automatically holds true. Therefore, we consider the case when $D_M^{(t)} \geq \varepsilon\, \forall t = 0, 1, \ldots, T-1$. For each of the three cases, we will first use induction to show $D_M^{(t)} \leq \alpha\Delta_{\min}$ for each iteration, and then we will show $D_M^{(T)} \leq C\varepsilon$ where C will be specified for each case.
(i) Assume $D_M^{(s)} \leq \alpha\Delta_{\min}\, \forall s \leq t$. Suppose n is large enough such that $\exp(-C_\alpha n) < \frac{1}{2}$. Then

$$\begin{aligned}
D_M^{(t+1)} &\leq \varepsilon + \exp(-C_\alpha n)D_M^{(t)} \\
&\leq \frac{\alpha\Delta_{\min}}{2} + \frac{\alpha\Delta_{\min}}{2} \\
&= \alpha\Delta_{\min}
\end{aligned}$$

Then, by Theorem 3, iterating over T iterations, we get

$$\begin{aligned}
D_M^{(T)} &\leq \sum_{t=0}^{T-1}\exp(-tC_\alpha n)\varepsilon + D_M^{(0)}\exp(-TC_\alpha n) \\
&\leq 2\varepsilon + D_M^{(0)}\exp(-TC_\alpha n)
\end{aligned}$$

Therefore, after $T = \frac{1}{C_\alpha n}\log(\frac{D_M^{(0)}}{\varepsilon})$ iterations, $D_M^{(T)} \leq \mathcal{O}(\varepsilon)$.
(ii) Similarly as before, we again assume $D_M^{(s)} \leq \alpha\Delta_{\min}\, \forall s \leq t$ and n is large enough such that

$\exp(-C_\alpha n) < \frac{1}{2}$. If $\varepsilon = \sqrt{\frac{K^{3/2}d}{mn}\log(dK^2T/\delta)}(\sigma + \Delta_{\max})$, then since $\alpha\Delta_{\min} < \Delta_{\max}$ and $\frac{\Delta_{\max}}{\sqrt{n}} \le \Delta_{\max}$, the term $\sqrt{\frac{K^{3/2}d}{mn}\log(dK^2T/\delta)}(\sigma + D_M^{(s)} + \frac{\Delta_{\max}}{\sqrt{n}})$ already has statistical error $\mathcal{O}(\varepsilon)$. Therefore,

$$\begin{aligned} D_M^{(t+1)} &\le \mathcal{O}(\varepsilon) + \mathcal{O}(D_M^{(t)}\exp(-C_\alpha n)) \\ &\le \frac{\alpha\Delta_{\min}}{2} + \frac{\alpha\Delta_{\min}}{2} \\ &= \alpha\Delta_{\min} \end{aligned}$$

By Theorem 3, iterating over T iterations, we have

$$D_M^{(T)} \le \mathcal{O}(\varepsilon) + \exp(-C_\alpha T n) D_M^{(0)}.$$

Therefore,

$$\begin{aligned} T &= \frac{1}{C_\alpha n}\log(\frac{D_M^{(0)}}{\varepsilon}) \\ &= \frac{1}{C_\alpha n}\log(\frac{D_M^{(0)}}{\sigma + \Delta_{\max}}\sqrt{\frac{mn}{K^{3/2}d\log(dK^2T/\delta)}}) \\ &= \mathcal{O}(1) \quad \text{since } m \le \mathcal{O}(\exp(n)TK\log(dK^2T/\delta)). \end{aligned}$$

If $\alpha \le \frac{\Delta_{\max}}{\Delta_{\min}\sqrt{n}}$, then $\alpha\Delta_{\min} \le \frac{\Delta_{\max}}{\sqrt{n}}$, which means the precision ε can be further improved to $\sqrt{\frac{K^{3/2}d}{mn}\log(dK^2T/\delta)}(\sigma + \frac{\Delta_{\max}}{\sqrt{n}})$. The rest of the proof remains the same.

(iii) Again, assume $D_M^{(s)} \le \alpha\Delta_{\min}\,\forall s \le t$. Moreover, assume n and m are large enough such that $\exp(-C_\alpha n) + \sqrt{\frac{dK^3}{m\sqrt{n}}\log(dK^2T/\delta)} \le \frac{1}{2}$. Then

$$D_M^{(t+1)} \le \varepsilon + \frac{1}{2}D_M^{(t)} \le \frac{\alpha\Delta_{\min}}{2} + \frac{\alpha\Delta_{\min}}{2} = \alpha\Delta_{\min}.$$

Then by Theorem 3, iterating over T iterations, we get

$$D_M^{(T)} \le \sum_{t=0}^{T-1}(\frac{1}{2})^t\varepsilon + (\frac{1}{2})^T D_M^{(0)} \le 2\varepsilon + 2^{-T}D_M^{(0)}$$

Therefore, $T = \mathcal{O}(\log(\frac{D_M^{(0)}}{\varepsilon}))$. □

E Auxiliary Lemma

Lemma 1. *Let T be the number of iterations and K be the number of components in FMLR. Let U be a d-dimensional random variable and A be an event defined on the same probability space with $p = \mathbb{P}(U \in A) > 0$. Define random variables $W = U|A$ and $Z = \mathbb{1}_A$. Suppose W is sub-exponential with sub-exponential norm $\|W\|_{\psi_1}$. Let U_j, W_j, Z_j be the i.i.d samples from the corresponding distributions. Then, setting*

$$t = \|W\|_{\psi_1}\mathcal{O}(\sqrt{p \vee \frac{\log(dK^2T/\delta)}{m}}\sqrt{\frac{d\log(dK^2T/\delta)}{m}}),$$

with probability at least $1 - 3\delta/(K^2T)$*, we have*

$$\mathbb{P}(\|\frac{1}{m}\sum_{j=1}^{m} U_j \mathbb{1}_{U_j\in A} - \mathbb{E}[U\mathbb{1}_A]\| \le t).$$

Proof. The key idea of the proof lies in the application of Proposition 5.3 from [KC20b]. The proposition is as follows.

Proposition 3. *(Proposition 5.3 in [KC20b]) Let* U *be a* d*-dimensional random variable and* A *be an event defined on the same probability space with* $p = \mathbb{P}(U \in A) > 0$*. Define random variables* $W = U|A$ *and* $Z = \mathbb{1}_A$*. Let* U_j, W_j, Z_j *be the i.i.d samples from the corresponding distributions. Then for any* $0 \le m_e \le m$ *and* $t_1 + t_2 = t$*,*

$$\begin{aligned}\mathbb{P}(\|\frac{1}{m}\sum_{j=1}^{m} U_j \mathbb{1}_{U_j\in A} - \mathbb{E}[U\mathbb{1}_A]\| \ge t) \le \max_{\widetilde{m}\le m_e} \mathbb{P}(\frac{1}{m}\|\sum_{j=1}^{\widetilde{m}}(W_j - \mathbb{E}[W])\| \ge t_1)\\ + \mathbb{P}(\|\mathbb{E}[W]\||\frac{1}{m}\sum_{j=1}^{m} Z_j - p| \ge t_2) + \mathbb{P}(|\sum_{j=1}^{m} Z_j| \ge m_e + 1)\end{aligned}$$

According to the proposition, we define $Z_j = \mathbb{1}_{U_j\in A}$ and $p = \mathbb{P}(A)$. Then Z_j is a Bernoulli random variable with p. Recall Bernstein's inequality for Bernoulli random variable

$$\mathbb{P}(|\frac{1}{m}\sum_{j=1}^{m} Z_j - p| \ge s) \le \exp(-\frac{ms^2}{2p + \frac{2}{3}s}).$$

To choose m_e for Proposition 3, we want to guarantee

$$\mathbb{P}(\sum_{j=1}^{m} Z_j \ge m_e + 1) \le \mathbb{P}(\frac{1}{m} Z_j - \mathbb{E}[z] \ge s) \le \frac{\delta}{K^2T}.$$

Therefore, we should choose

$$s = \frac{1}{m}(\frac{1}{3}\log(\frac{K^2T}{\delta}) + \sqrt{\frac{1}{9}\log^2(\frac{K^2T}{\delta}) + 2pm\log(\frac{K^2T}{\delta})})$$

and

$$\begin{aligned} m_e &= mp + ms = mp + \frac{1}{3}\log(\frac{K^2T}{\delta}) + \sqrt{\frac{1}{9}\log^2(\frac{K^2T}{\delta}) + 2pm\log(\frac{K^2T}{\delta})} \\ &= mp + \mathcal{O}(\log(K^2T/\delta) \vee \sqrt{pm\log(K^2T/\delta)}) \end{aligned}$$

Next, by Proposition 3, we need

$$\begin{aligned} &\mathbb{P}(|\frac{1}{m}\sum_{j=1}^{m} Z_j - p| \ge \frac{t_2}{\mathbb{E}[\|W\|]}) \le \frac{\delta}{K^2T} \\ \Rightarrow &\mathbb{P}(|\frac{1}{m}\sum_{j=1}^{m} Z_j - p| \ge \frac{t_2}{\|W\|_{\psi_1}}) \le \frac{\delta}{K^2T} \quad \text{since } \mathbb{E}[\|W\|] \le \|W\|_{\psi_1}. \end{aligned}$$

Again, by Bernstein's inequality, we should choose

$$\begin{aligned} t_2 &= \frac{1}{m}\|W\|_{\psi_1}(\frac{1}{3}\log(\frac{K^2T}{\delta}) + \sqrt{\frac{1}{9}\log^2(\frac{K^2T}{\delta}) + 2pm\log(\frac{K^2T}{\delta})}) \\ &\leq \|W\|_{\psi_1}\mathcal{O}(\sqrt{p \vee \frac{\log(K^2T/\delta)}{m}}\sqrt{\frac{\log(K^2T/\delta)}{m}}) \end{aligned}$$

Next, since W is sub-exponential, by Lemma 6, $\forall \widetilde{m} \leq m_e$,

$$\begin{aligned} \mathbb{P}(\frac{1}{m}|\sum_{j=1}^{\widetilde{m}} W_j - \mathbb{E}[W]| \geq t_1) &= \mathbb{P}(\frac{1}{\widetilde{m}}|\sum_{j=1}^{\widetilde{m}} W_j - \mathbb{E}[W]| \geq \frac{mt_1}{\widetilde{m}}) \\ &\leq \exp(-C\min\{\frac{mt_1}{\|W\|_{\psi_1}\sqrt{d}}, \frac{m^2t_1^2}{m_e d\|W\|_{\psi_1}^2}\} + C'\log d) \end{aligned}$$

Therefore, we should take

$$\begin{aligned} t_1 &= \|W\|_{\psi_1}\sqrt{d}\max\{\frac{C'\log d + \log(K^2T/\delta)}{Cm}, \sqrt{\frac{C'\log d + \log(K^2T/\delta)}{Cm}}\sqrt{\frac{m_e}{m}}\} \\ &= \|W\|_{\psi_1}\sqrt{d}\mathcal{O}(\sqrt{\frac{m_e}{m} \vee \frac{\log(dK^2T/\delta)}{m}}\sqrt{\frac{\log(dK^2T/\delta)}{m}}) \end{aligned}$$

Plug m_e into t_1 and we will get

$$t_1 = \|W\|_{\psi_1}\sqrt{d}\mathcal{O}(\sqrt{p \vee \frac{\log(dK^2T/\delta)}{m}}\sqrt{\frac{\log(dK^2T/\delta)}{m}}).$$

By Proposition 3, set

$$t = t_1 + t_2 = \|W\|_{\psi_1}\mathcal{O}(\sqrt{p \vee \frac{\log(dK^2T/\delta)}{m}}\sqrt{\frac{d\log(dK^2T/\delta)}{m}})$$

which concludes the proof of the lemma. □

The following two lemmas are used in bounding sub-exponential norms of random variables conditioning on some events. These are analogous to Lemma A.1 and Lemma A.2 in [KC20b].

Lemma 2. *Let $X_1, \ldots, X_n \sim_{i.i.d} \mathcal{N}(0, I_d)$. For any fixed vector u and constant α, define $G = \{\sum_{i=1}^n \langle X_i, u\rangle^2 \geq \alpha^2\}$. Then for any unit vector $s \in \mathcal{S}^{d-1}$ and $p \geq 1$,*

$$\mathbb{E}[(\sum_{i=1}^n \langle X_i, s\rangle^2)^p | G^c] = \mathcal{O}((\sqrt{n}p)^p).$$

Proof. WLOG, we can assume $u = e_1$ due to the rotational invariance property of Gaussian. Denote $Y_i = \langle X_{i,2:d}, s_{2:d}\rangle$ as the inner product between the second to the last coordinate of X_i and s. Then we have

$$\mathbb{E}[(\sum_{i=1}^n \langle X_i, s\rangle^2)^p | G^c] = \frac{\mathbb{E}[(\sum_{i=1}^n (s_1 X_{i,1} + Y_i)^2)^p \mathbb{1}_{\sum_{i=1}^n X_{i,1}^2 \leq \alpha^2}]}{\mathbb{P}(\sum_{i=1}^n X_{i,1}^2 \leq \alpha^2)}$$

$$
\begin{aligned}
&\leq \frac{\mathbb{E}[(\sum_{i=1^n} 2s_1^2 X_{1,i}^2 + 2Y_i^2)^p \mathbb{1}_{\sum_{i=1}^n X_{i,1}^2 \leq \alpha^2}]}{\mathbb{P}(\sum_{i=1}^n X_{i,1}^2 \leq \alpha^2)} \\
&= \frac{\mathbb{E}[(\mathbb{E}[(\sum_{i=1^n} 2s_1^2 X_{1,i}^2 + 2Y_i^2)^p | \{X_i, 1\}_{i=1}^n]^{1/p})^p \mathbb{1}_{\sum_{i=1}^n X_{i,1}^2 \leq \alpha^2}]}{\mathbb{P}(\sum_{i=1}^n X_{i,1}^2 \leq \alpha^2)} \\
&\overset{\text{(i)}}{\leq} \frac{\mathbb{E}[(\mathbb{E}[(\sum_{i=1^n} 2s_1^2 X_{1,i}^2)^p | \{X_{i,1}\}_{i=1}^n]^{1/p} + \mathbb{E}[(\sum_{i=1}^n 2Y_i^2)^p | \{X_i, 1\}_{i=1}^n]^{1/p})^p \mathbb{1}_{\sum_{i=1}^n X_{i,1}^2 \leq \alpha^2}]}{\mathbb{P}(\sum_{i=1}^n X_{i,1}^2 \leq \alpha^2)} \\
&\overset{\text{(ii)}}{=} \frac{\mathbb{E}[(\sum_{i=1}^n 2s_1^2 X_{i,1}^2 + \mathbb{E}[(\sum_{i=1}^n 2Y_i^2)^p]^{1/p})^p \mathbb{1}_{\sum_{i=1}^n X_{i,1}^2 \leq \alpha^2}]}{\mathbb{P}(\sum_{i=1}^n X_{i,1}^2 \leq \alpha^2)} \\
&\overset{\text{(iii)}}{\leq} \frac{(2s_1^2 \alpha^2 + \mathbb{E}[(\sum_{i=1}^n 2Y_i^2)^p]^{1/p})^p \mathbb{E}[\mathbb{1}_{\sum_{i=1}^n X_{i,1}^2 \leq \alpha^2}]}{\mathbb{P}(\sum_{i=1}^n X_{i,1}^2 \leq \alpha^2)} \\
&\overset{\text{(iv)}}{=} (2s_1^2 \alpha^2 + C\sqrt{n} \|s_{2:d}\|^2 p)^p \\
&= \mathcal{O}((\sqrt{n} p)^p)
\end{aligned}
$$

Note that (i) follows from Minkowski inequality, both (ii) and (iii) follow from the independence of $\{X_{i,1}\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, and (iv) follows as $\sum_{i=1}^n 2Y_i^2 \sim \text{SubExp}(16n\|s_{2:d}\|^4, 8\|s_{2:d}\|^2)$ whose L_p norm is $C\sqrt{n}\|s_{2:d}\|^2 p$ for some constant C. $\square$

Lemma 3. *Let $X_1, \ldots, X_n \sim_{i.i.d} \mathcal{N}(0, I_d)$. For any fixed vector $u \in \mathbb{R}^d$ and a set of vectors $\{v_1, \ldots, v_H\} \subset \mathbb{R}^d$ such that $\|u\| \geq \|v_l\|$ $\forall l = 1, \ldots, H$, define $G := \cap_{l=1}^H \{\sum_{i=1}^n \langle X_i, u\rangle^2 \geq \sum_{i=1}^n \langle X_i, v_l\rangle^2\}$. Then for any unit vector $s \in \mathcal{S}^{d-1}$ and $p \geq 1$,*

$$
\mathbb{E}[(\sum_{i=1}^n \langle X_i, s\rangle^2)^p | G^c] = \mathcal{O}(H(np)^p).
$$

Proof. Let $G_l = \{\sum_{i=1}^n \langle X_i, u\rangle^2 \geq \sum_{i=1}^n \langle X_i, v_l\rangle^2\}$. Then $G = \cap_{l=1}^H G_l$. We first focus on G_1^c. Again, by the rotational invariance property of Gaussian, we can assume $\text{span}\{u, v_1\} = \text{span}\{e_1, e_2\}$. We use the following change of coordinates $X_{i,1} = r_i \cos\theta_i$ and $X_{i,2} = r_i \sin\theta_i$ where $r_i \sim_{i.i.d}$ Rayleigh(1) and $\theta_i \sim_{i.i.d}$ Uniform$[0, 2\pi]$. Also, we denote $Y_i = \langle X_{i,3:d}, s_{3:d}\rangle$.

$$
\begin{aligned}
\mathbb{E}[(\sum_{i=1}^n \langle X_i, s\rangle^2)^p | G_1^c] &= \frac{\mathbb{E}[(\sum_{i=1}^n (s_1 r_i \cos\theta_i + s_2 r_i \sin\theta_i + Y_i)^2)^p \mathbb{1}_{G_1^c}]}{\mathbb{P}(G_1^c)} \\
&= \frac{\mathbb{E}_\theta[(\mathbb{E}_{r,Y}[(\sum_{i=1}^n (s_1 r_i \cos\theta_i + s_2 r_i \sin\theta_i + Y_i)^2)^p | \theta]^{1/p})^p \mathbb{1}_{G_1^c}]}{\mathbb{P}(G_1^c)} \\
&\overset{\text{(i)}}{\leq} \frac{\mathbb{E}_\theta[(\mathbb{E}_{r,Y}[(\sum_{i=1}^n 4r_i^2 (s_1^2 \cos^2\theta_i + s_2^2 \sin^2\theta_i) + \sum_{i=1}^n 2Y_i^2)^p | \theta]^{1/p})^p \mathbb{1}_{G_1^c}]}{\mathbb{P}(G_1^c)} \\
&\overset{\text{(ii)}}{\leq} \frac{\mathbb{E}_\theta[(\mathbb{E}_r[(\sum_{i=1}^n 4r_i^2 (s_1^2 \cos^2\theta_i + s_2^2 \sin^2\theta_1))^p | \theta]^{1/p} + \mathbb{E}_Y[(\sum_{i=1}^n 2Y_i^2)^p]^{1/p})^p \mathbb{1}_{G_1^c}]}{\mathbb{P}(G_1^c)}
\end{aligned}
$$

where (i) follows from the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ and (ii) follows from Minkowski inequality. Note that $\sum_{i=1}^n 2Y_i^2 \sim \text{SubE}(16n\|s_{3:d}\|^4, 8\|s_{3:d}\|^2)$ whose L_p norm is $C\sqrt{n}\|s_{3:d}\|^2 p$ for some constant C. Moreover,

$$
\mathbb{E}_r[(\sum_{i=1}^n 4r_i^2 (s_1^2 \cos^2\theta_i + s_2^2 \sin^2\theta_1))^p | \theta]^{1/p}
$$

$$
\begin{aligned}
&\leq \mathbb{E}_r[(\sum_{i=1}^n 16r_i^4)^{p/2}(\sum_{i=1}^n (s_1^2\cos^2\theta_i + s_2^2\sin^2\theta_i)^2)^{p/2}|\theta]^{1/p} \text{ by Cauchy Schwarz inequality} \\
&= \mathbb{E}_r[(\sum_{i=1}^n 16r_i^4)^{p/2}]^{1/p}(\sum_{i=1}^n (s_1^2\cos^2\theta_i + s_2^2\sin^2\theta_i)^2)^{1/2} \\
&\leq \mathbb{E}_r[(4\sqrt{n}r^2)^p]^{1/p}\sqrt{n}\|s_{1:2}\|^2 \\
&= 4n\|s_{1:2}\|^2\mathbb{E}_r[r^{2p}]^{1/p}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}[(\sum_{i=1}^n \langle X_i, s\rangle^2)^p|G_1^c] &\leq \frac{(4n\|s_{1:2}\|^2\mathbb{E}_r[r^{2p}]^{1/p} + C\sqrt{n}\|s_{3:d}\|^2 p)^p\mathbb{E}_\theta[\mathbb{1}_{\theta\in G_1^c}]}{\mathbb{P}(G_1^c)} \\
&= (4n\|s_{1:2}\|^2\mathbb{E}_r[r^{2p}]^{1/p} + C\sqrt{n}\|s_{3:d}\|^2 p)^p
\end{aligned}
$$

Since $r \sim \text{Rayleigh}(1)$, its raw moments are given by $2^{p/2}\Gamma(1+\frac{p}{2})$ where Γ is the Gamma function. Then

$$
\mathbb{E}_r[r^{2p}]^{1/p} = (\mathbb{E}_r[r^{2p}]^{\frac{1}{2p}})^2 = 2\Gamma^{1/p}(1+p).
$$

Note that by Lanczos approximation, $\Gamma^{1/p}(1+p) = \mathcal{O}(p)$. This gives us

$$
\begin{aligned}
\mathbb{E}[(\sum_{i=1}^n \langle X_i, s\rangle^2)^p|G_1^c] &\leq (8n\|s_{1:2}\|^2\Gamma^{1/p}(1+p) + C\sqrt{n}\|s_{3:d}\|^2 p)^p \\
&= \mathcal{O}((np)^p)
\end{aligned}
$$

This is the analysis for the first event G_1, and the analysis hold true for all the other events.

$$
\begin{aligned}
\mathbb{E}[(\sum_{i=1}^n \langle X_i, s\rangle^2)^p|G^c] &\leq \frac{\mathbb{E}[(\sum_{i=1}^n \langle X_i, s\rangle^2)^p \sum_{l=1}^H \mathbb{1}_{G_l^c}]}{\mathbb{P}(G^c)} \\
&= \sum_{l=1}^H \frac{\mathbb{E}[(\sum_{i=1}^n \langle X_i, s\rangle^2)^p \mathbb{1}_{G_l^c}]}{\mathbb{P}(G^c)} \\
&\leq \sum_{l=1}^H \frac{\mathbb{E}[(\sum_{i=1}^n \langle X_i, s\rangle^2)^p \mathbb{1}_{G_l^c}]}{\mathbb{P}(G_l^c)} \\
&= \mathcal{O}(H(np)^p)
\end{aligned}
$$

□

The following lemma is proved in [BWY17]. We state and prove it for completeness.

Lemma 4. *Suppose $X \sim \mathcal{N}(0, I_d)$. Then for any fixed vectors $u, v \in \mathbb{R}^d$, we have*

$$
\mathbb{E}[\langle X, u\rangle^2\langle X, v\rangle^2] \leq 3\|u\|^2\|v\|^2.
$$

Proof. Similarly as before, by the rotational invariance property of Gaussian, we can choose an orthogonal matrix R such that $Ru = \|u\|e_1$ and $R^T X$ still follows $\mathcal{N}(0, I_d)$. Denote $z = Rv$. Then we have

$$
\mathbb{E}[\langle X, Ru\rangle^2\langle X, Rv\rangle^2] = \mathbb{E}[\|u\|^2 X_1^2 \sum_{i=1}^d\sum_{j=1}^d X_i X_j z_i z_j]
$$

$$= \|u\|^2(3z_1^2 + (\|z\|^2 - z_1^2)) \leq 3\|u\|^2\|z\|^2 = 3\|u\|^2\|v\|^2$$

□

The following lemma from [LM00] gives well-known tail bounds for χ^2 random variables.

Lemma 5. *Let U be a χ^2 statistics with D degree of freedom. For any positive s,*

$$\mathbb{P}(U - D \geq 2\sqrt{Ds} + 2s) \leq \exp(-s)$$

and

$$\mathbb{P}(D - U \geq 2\sqrt{Ds}) \leq \exp(-s).$$

The following lemma derived from [Ver18] provides a standard tail bound for sub-exponential d-dimensional random vector.

Lemma 6. *[Ver18] Let W be a random vector in $\mathbb{R}^d$ with all elements being sub-exponential with the same sub-exponential norm K. Then $\forall t > 0$,*

$$\mathbb{P}(\|\frac{1}{m}\sum_{j=1}^{m} W_j - \mathbb{E}[W]\| \geq t) \leq \exp(-Cm\min\{\frac{t}{K\sqrt{d}}, (\frac{t}{K\sqrt{d}})^2\} + C'\log d)$$

where C and C' are some constants.

The following two lemmas from [Ver18] provide relations between sub-exponential norm and sub-gaussian norm, which we will use when computing sub-exponential norm.

Lemma 7. *[Ver18] A random variable X is sub-gaussian if and only if X^2 is sub-exponential and $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$.*

Lemma 8. *[Ver18] Let X and Y be sub-gaussian random variables. Then XY is sub-exponential and $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$.*