Towards the Theory of Unsupervised Federated Learning: Non-asymptotic Analysis of Federated EM Algorithms

Ye Tian [1] **Haolei Weng** [2] **Yang Feng** [3]

Abstract

While supervised federated learning approaches have enjoyed significant success, the domain of unsupervised federated learning remains relatively underexplored. Several federated EM algorithms have gained popularity in practice, however, their theoretical foundations are often lacking. In this paper, we first introduce a federated gradient EM algorithm (FedGrEM) designed for the unsupervised learning of mixture models, which supplements the existing federated EM algorithms by considering task heterogeneity and potential adversarial attacks. We present a comprehensive finite-sample theory that holds for general mixture models, then apply this general theory on specific statistical models to characterize the explicit estimation error of model parameters and mixture proportions. Our theory elucidates when and how FedGrEM outperforms local single-task learning with insights extending to existing federated EM algorithms. This bridges the gap between their practical success and theoretical understanding. Our numerical results validate our theory, and demonstrate FedGrEM's superiority over existing unsupervised federated learning benchmarks.

1. Introduction

Federated learning (FDL) is a machine learning paradigm that allows the training of statistical models by leveraging data from various local tasks, while ensuring the data remains decentralized to protect privacy (Li et al., 2020a). Introduced a few years ago, notably by Google (Konečnỳ et al., 2016; McMahan et al., 2017), FDL has witnessed remarkable success in a diverse range of applications, including smartphones (Hard et al., 2018), healthcare (Antunes et al., 2022), and the internet of things (Nguyen et al., 2021). However, it is important to note that a large portion of current FDL research is centered around supervised learning problems. In this paper, we delve into the realm of unsupervised FDL, a scenario in which each task involves a mixture of distributions.

Before proceeding further, we summarize the mathematical notations used in this paper here. $\mathbb{P}$ and $\mathbb{E}$ denote the probability and expectation, respectively. For two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \ll b_n$ or $a_n = o(b_n)$ means $a_n/b_n \to 0$, $a_n \lesssim b_n$ or $a_n = \mathcal{O}(b_n)$ means $a_n/b_n \leq C < \infty$, and $a_n \asymp b_n$ means $a_n/b_n, b_n/a_n \leq C < \infty$. $\widetilde{\mathcal{O}}(b_n)$ is the same as $a_n = \mathcal{O}(b_n)$ up to logarithmic factors. For a random variable x_n and a positive sequence a_n, $x_n = \mathcal{O}_{\mathbb{P}}(a_n)$ means that for any $\epsilon > 0$, there exists $M > 0$ such that $\sup_n \mathbb{P}(|x_n/a_n| > M) \leq \epsilon$. $\widetilde{\mathcal{O}}_{\mathbb{P}}(a_n)$ has a similar meaning up to logarithmic factors in a_n. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, $\|x\|_2$ represents its Euclidean norm. For two numbers a and b, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For any positive integer K, $1 : K$ and $[K]$ stand for the set $\{1, 2, \ldots, K\}$. And for any set S, $|S|$ denotes its cardinality and S^c denotes its complement. "w.p." stands for "with probability". The absolute constants c and C may vary from line to line.

2. Federated Learning on Mixture of Distributions

2.1. Problem Setting

Consider K tasks, where for the k-th task, we observe data $\{\boldsymbol{x}_i^{(k)}\}_{i=1}^n \subseteq \mathbb{R}^d$ [1]. There exists an *unknown* subset $S \subseteq [K]$, such that each observation in task $k \in S$ comes from a *mixture model* with R components ($R \geq 2$):

$$\boldsymbol{x}_i^{(k)} \overset{i.i.d.}{\sim} \sum_{r=1}^{R} w_r^{(k)*} \cdot p_r^{(k)}(\,\cdot\,; \boldsymbol{\theta}_r^{(k)*}), \tag{1}$$

[1]Department of Statistics, Columbia University, New York, USA [2]Department of Statistics and Probability, Michigan State University, East Lansing, USA [3]Department of Biostatistics, School of Global Public Health, New York University, New York, USA. Correspondence to: Yang Feng <yang.feng@nyu.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]For simplicity, we assume all tasks share the same sample size n. We can easily extend our analysis to the case of heterogeneous task sample sizes with similar theoretical results.

where the *mixture proportion* $\{w_r^{(k)*}\}_{r=1}^R \subseteq (0,1)$ with $\sum_{r=1}^R w_r^{(k)*} = 1$ and $p_r^{(k)}(\cdot\,;\boldsymbol{\theta}_r^{(k)*})$ is a Radon-Nikodym density w.r.t. a base measure σ. $\boldsymbol{\theta}_r^{(k)*} \in \mathbb{R}^d$ [2] are the parameters that index the distribution $p_r^{(k)}$. This is equivalent to

$$z_i^{(k)} \overset{i.i.d.}{\sim} \sum_{r=1}^R w_r^{(k)*}\cdot\delta_r, \quad \boldsymbol{x}_i^{(k)}|z_i^{(k)} = r \overset{i.i.d.}{\sim} p_r^{(k)}(\cdot\,;\boldsymbol{\theta}_r^{(k)*}), \tag{2}$$

for $k \in S$, where $z_i^{(k)}$ is the *unobserved* latent cluster label and δ_r is the point mass at r. Here S is the index of similar tasks (unknown), where the parameters $\{\boldsymbol{\theta}_r^{(k)*}\}_{k\in S}$ of different tasks are "similar" to each other, in the sense that

$$\min_{\overline{\boldsymbol{\theta}}\in\mathbb{R}^d} \max_{k\in S} \|\boldsymbol{\theta}_r^{(k)*} - \overline{\boldsymbol{\theta}}\|_2 \le h, \quad \forall r \in [R],$$

where h is an *unknown* parameter controlling the task similarity level. A small h implies that the tasks are more similar. The data from tasks in set $S^c = [K]\backslash S$ can be *arbitrarily distributed*, i.e., $\{\boldsymbol{x}_i^{(k)}\}_{i\in[n],k\in S^c}$ follow an *arbitrary* joint distribution $\mathbb{Q}_{S^c}$, and we denote the proportion $\epsilon := |S^c|/K \in [0,1)$. Note that h, K, R and d can change with single-task sample size n.

There are two different interpretations of this setting. The first one is from the perspective of *adversarial attacks/contaminations*, where there is an *adversarial attacker* who can arbitrarily contaminate the data of tasks in an index set S^c. The index set S^c and the distribution $\mathbb{Q}_{S^c}$ are picked by the attacker *after* we pick the estimator (hence S^c and $\mathbb{Q}_{S^c}$ are *unknown* to us). A similar setting can be found in Qiao (2018), Konstantinov & Lampert (2019), Konstantinov et al. (2020), and Tian et al. (2023).

Alternatively, aside from adversarial attacks, we can interpret the presence of contaminated data sets as a result of *outlier tasks*. In the era of big data, certain collected data sets may exhibit distributions significantly different from others, particularly when dealing with numerous tasks (Zhang & Yang, 2021). These data sets within S^c can be viewed as outlier tasks. In practice, detecting outlier tasks is challenging.

In the rest of this paper, we may take the views of "adversarial attacks" and "outlier tasks" interchangeably.

Note that in our unsupervised learning setting, similar to Marfoq et al. (2021) and Wu et al. (2023), we avoid assuming that the mixture proportions $\{w_r^{(k)*}\}_{k\in S}$ are similar across tasks, which offers more flexibility in practice.

The goal is to develop an algorithm to estimate the mixture proportions $\{w_r^{(k)*}\}_{k\in S, r\in[R]}$ and the parameters $\{\boldsymbol{\theta}_r^{(k)*}\}_{k\in S, r\in[R]}$ simultaneously, which satisfies the following five desired properties:

(i) **Adaptability** to unknown similarity level h: The algorithm should utilize the data from different sources in an "optimal" way. When h is small, the output estimator should achieve a better convergence rate than the local estimator (or single-task estimator of each task). When h is large, the output estimator should perform no worse than the local estimator.

(ii) **Robustness** against the adversarial attack on a small fraction of sources: The output estimator should maintain a good performance when the contaminated proportion ϵ is small.

(iii) **Privacy** for local data: The algorithm should avoid transferring raw data out of each task.

(iv) **Computation efficiency** on local servers: The local computational cost should be low.

(v) **Communication efficiency** between local and global servers: The communicational cost should be low.

2.2. Related Works

Federated learning (FDL): While there exists an extensive body of literature on FDL, the majority of it centers around the supervised learning paradigm. Notable frameworks within supervised FDL include COCOA (Jaggi et al., 2014), MOCHA (Smith et al., 2017), and FedAvg (McMahan et al., 2017). To accommodate varying task characteristics, FedProx was introduced by Li et al. (2020b). See Yang et al. (2019) and Li et al. (2020a) for a comprehensive overview of supervised FDL. In contrast to supervised FDL, much less attention has been given to unsupervised FDL with mixture models. Marfoq et al. (2021) examined a similar FDL problem presented in this paper and introduced a federated EM algorithm without exploring the estimation error of the parameter estimators. Dieuleveut et al. (2021) also proposed a federated EM algorithm that supports communication compression and partial participation. Wu et al. (2023) adapted the EM algorithm to the scenario where predictors are from Gaussian Mixture Models (GMMs) and the regression models can encompass general mixture models. Notably, none of these papers provided finite-sample results for their EM algorithm (which is the key to interpreting their practical successes) nor discussed the adversarial attacks and outlier tasks. Our work complements this line of research by explicitly accommodating outlier tasks, providing comprehensive finite-sample results, and applying the developed theory to illustrative model examples. Our theory illustrates when and how the aforementioned federated EM algorithms (Dieuleveut

[2] For simplicity, we assume parameters and observations are of the same dimension, but our results can be generalized to the case where the two dimensions are different.

et al., 2021; Marfoq et al., 2021; Wu et al., 2023) outperform local single-task learning. Note that there exist works on clustered FDL, wherein tasks are organized into several groups with tasks within each group being identical (task-level mixture) (Ghosh et al., 2020; Kong et al., 2020; Su et al., 2022). This setting differs from ours, where each task's data originates from a mixture model at the sample level.

Multi-task learning (MTL) and transfer learning (TL): Problems related to federated learning but permitting raw data sharing across tasks include multi-task learning and transfer learning. Analogous to federated learning, a substantial proportion of research in MTL and TL centers on supervised learning. In unsupervised MTL and TL, there have been diverse approaches, including the kernel k-means clustering (Gu et al., 2011), the spectral method (Yang et al., 2014), and the penalized optimization (Dai et al., 2008; Zhang & Zhang, 2011; Zhang et al., 2015; 2018). Specific mixture models such as Gaussian Mixture Models (GMMs) have also been explored (Wang et al., 2021; Tian et al., 2022). Discussions on outlier tasks, adversarial attacks, and negative transfer in MTL and TL have emerged in various model settings, for example, Qiao (2018); Konstantinov & Lampert (2019); Hanneke & Kpotufe (2020); Konstantinov et al. (2020); Li et al. (2021); Duan & Wang (2022); Tian et al. (2023).

EM algorithm: The EM algorithm was formalized by Dempster et al. (1977), and there have been intensive studies on the local convergence of the likelihood and the estimator to some stationary point (Wu, 1983; Redner & Walker, 1984; Meng & Rubin, 1994; McLachlan & Krishnan, 2007). More recently, Xu et al. (2016) established the global convergence of EM algorithm on binary GMMs. Additionally, Balakrishnan et al. (2017), Yan et al. (2017), Cai et al. (2019), Kwon & Caramanis (2020a), Kwon & Caramanis (2020b), and Zhao et al. (2020) provided finite-sample convergence results for EM and its variants under certain initialization conditions.

2.3. A Federated Gradient EM: FedGrEM

Before delving into our main algorithm, we first introduce some key notations and definitions. The posterior, i.e. the probability of $z^{(k)} = r$ conditioned on the observation $\boldsymbol{x}^{(k)}$, given that $(\boldsymbol{x}^{(k)}, z^{(k)})$ is from the mixture model (2) with parameters $\boldsymbol{w}^{(k)} = \{w_r^{(k)}\}_{r=1}^R$ and $\boldsymbol{\theta}^{(k)} = \{\boldsymbol{\theta}_r^{(k)}\}_{r=1}^R$, is defined as

$$\mathbb{P}(z^{(k)} = r|\boldsymbol{x}^{(k)}; \boldsymbol{w}^{(k)}, \boldsymbol{\theta}^{(k)}) = \frac{w_r^{(k)} \times p_r^{(k)}(\boldsymbol{x}^{(k)}; \boldsymbol{\theta}_r^{(k)})}{\sum_{r=1}^R w_r^{(k)} \times p_r^{(k)}(\boldsymbol{x}^{(k)}; \boldsymbol{\theta}_r^{(k)})}.$$

Based on this posterior, we define

$$Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}', \boldsymbol{\theta}') = \mathbb{E}\Bigg[\sum_{r=1}^R \mathbb{P}(z^{(k)} = r|\boldsymbol{x}^{(k)}; \boldsymbol{w}', \boldsymbol{\theta}') \times \log p_r^{(k)}(\boldsymbol{x}^{(k)}; \boldsymbol{\theta}_r)\Bigg],$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_r\}_{r=1}^R$, $\boldsymbol{w}' = \{w_r'\}_{r=1}^R$, and $\boldsymbol{\theta}' = \{\boldsymbol{\theta}_r'\}_{r=1}^R$. By Jensen's inequality, it can be shown that $Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}', \boldsymbol{\theta}')$ is a lower bound of the complete population-level log-likelihood $\mathbb{E}\log\big[\sum_{r=1}^R w_r p_r^{(k)}(\boldsymbol{x}^{(k)}; \boldsymbol{\theta}_r)\big]$. The latter one is difficult to manage as the summation is within the logarithm. The EM algorithm endeavors to maximize the surrogate $Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}', \boldsymbol{\theta}')$ by iteratively alternating with the E-step and M-step. Gradient EM does a one-step gradient ascent in M-step instead of finding the exact optimizer, which can speed up the computation. In practice, we work on a sample-based variant of $Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}', \boldsymbol{\theta}')$ as

$$\widehat{Q}^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}', \boldsymbol{\theta}') = \frac{1}{n}\sum_{i=1}^n \sum_{r=1}^R \mathbb{P}(z^{(k)} = r|\boldsymbol{x}_i^{(k)}; \boldsymbol{w}', \boldsymbol{\theta}') \times \log p_r^{(k)}(\boldsymbol{x}_i^{(k)}; \boldsymbol{\theta}_r).$$

Now, we are ready to introduce our core algorithm *FedGrEM* in Algorithm 1. It executes the E-step and M-step locally on each task, pooling the estimators of $\{\boldsymbol{\theta}_r^{(k)*}\}_{k=1}^K$ obtained in M-step by penalizing the ℓ_2-distance between each estimator and a common center. This approach mirrors the penalization strategy used in other MTL and TL literature, such as Evgeniou & Pontil (2004); Li & Bilmes (2007); Lounici et al. (2011); Solnon et al. (2012); Jalali et al. (2013); Kuzborskij & Orabona (2013; 2017); Denevi et al. (2018); T Dinh et al. (2020); Li et al. (2021); Tian & Feng (2023); He et al. (2024); Lin & Reimherr (2024a;b). After several iterations of local and central updates, FedGrEM yields the final estimators. Figure 1 provides an intuitive illustration of the workflow of FedGrEM.

While the core idea of FedGrEM is akin to FedEM in Marfoq et al. (2021) and FedGMM in Wu et al. (2023), two crucial distinctions set them apart. First, we employ gradient EM, while FedEM and FedGMM use the full EM. Second, the central update of FedGrEM can adapt to the *heterogeneity* of $\boldsymbol{\theta}_r^{(k)*}$'s across tasks and remain *robust* when a small proportion of tasks is contaminated, which are not present in FedEM or FedGMM. On the other hand, by setting $\lambda^{[t]} = +\infty$, FedGrEM can be viewed as a simplified gradient version of FedEM and FedGMM, therefore our non-asymptotic theory in Section 3 can illustrate the empirical success achieved by FedEM and FedGMM.

FedGrEM is *computationally efficient* on local servers as it computes the gradient instead of explicitly solving the

Algorithm 1 FedGrEM: A Federated Gradient EM Algorithm

Input: Initializations $\{\widehat{\boldsymbol{w}}^{(k)[0]}\}_{k\in[K]}$ and $\{\widehat{\boldsymbol{\theta}}^{(k)[0]}\}_{k\in[K]}$ ($\widehat{\boldsymbol{w}}^{(k)[0]} = \{\widehat{w}_r^{(k)[0]}\}_{r=1}^R$, $\widehat{\boldsymbol{\theta}}^{(k)[0]} = \{\widehat{\boldsymbol{\theta}}_r^{(k)[0]}\}_{r=1}^R$), data $\{\boldsymbol{x}_i^{(k)}\}_{i\in[n],k\in[K]}$, iteration number T, penalty parameters $\{\lambda^{[t]}\}_{t=1}^T$, step sizes $\{\eta_r^{(k)}\}_{k\in[K],r\in[R]}$
for $t = 1$ **to** T **do**
Local update: For task $k = 1 : K$:

- E-step: $\widehat{Q}^{(k)}(\boldsymbol{\theta}|\widehat{\boldsymbol{w}}^{(k)[t-1]}, \widehat{\boldsymbol{\theta}}^{(k)[t-1]}) := \frac{1}{n}\sum_{i=1}^n \sum_{r=1}^R \mathbb{P}(z^{(k)} = r|\boldsymbol{x}_i^{(k)}; \widehat{\boldsymbol{w}}^{(k)[t-1]}, \widehat{\boldsymbol{\theta}}^{(k)[t-1]}) \log p_r^{(k)}(\boldsymbol{x}_i^{(k)}; \boldsymbol{\theta}_r)$
- M-step: $\diamond\ \widehat{w}_r^{(k)[t]} = \frac{1}{n}\sum_{i=1}^n \mathbb{P}(z^{(k)} = r|\boldsymbol{x}_i^{(k)}; \widehat{\boldsymbol{w}}^{(k)[t-1]}, \widehat{\boldsymbol{\theta}}^{(k)[t-1]})$
  $\diamond\ \widetilde{\boldsymbol{\theta}}_r^{(k)[t]} = \widehat{\boldsymbol{\theta}}_r^{(k)[t-1]} + \eta_r^{(k)} \cdot \frac{\partial}{\partial \boldsymbol{\theta}_r} \widehat{Q}^{(k)}(\boldsymbol{\theta}|\widehat{\boldsymbol{w}}^{(k)[t-1]}, \widehat{\boldsymbol{\theta}}^{(k)[t-1]})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(k)[t-1]}}$

Central update: $\{\widehat{\boldsymbol{\theta}}_r^{(k)[t]}\}_{k=1}^K, \overline{\boldsymbol{\theta}}_r^{[t]} = \underset{\{\boldsymbol{\nu}^{(k)}\}_{k=1}^K \subseteq \mathbb{R}^d, \overline{\boldsymbol{\nu}}\in\mathbb{R}^d}{\arg\min} \left\{ \sum_{k=1}^K \left(\frac{n}{2}\|\boldsymbol{\nu}^{(k)} - \widetilde{\boldsymbol{\theta}}_r^{(k)[t]}\|_2^2 + \sqrt{n}\lambda^{[t]} \cdot \|\boldsymbol{\nu}^{(k)} - \overline{\boldsymbol{\nu}}\|_2 \right) \right\}$, define $\widehat{\boldsymbol{w}}^{(k)[t]} = \{\widehat{w}_r^{(k)[t]}\}_{r=1}^R$, $\widehat{\boldsymbol{\theta}}^{(k)[t]} = \{\widehat{\boldsymbol{\theta}}_r^{(k)[t]}\}_{r=1}^R$
end for
Output: Final estimators $\{\widehat{w}_r^{(k)[T]}\}_{k\in[K],r\in[R]}$ and $\{\widehat{\boldsymbol{\theta}}_r^{(k)[T]}\}_{k\in[K],r\in[R]}$

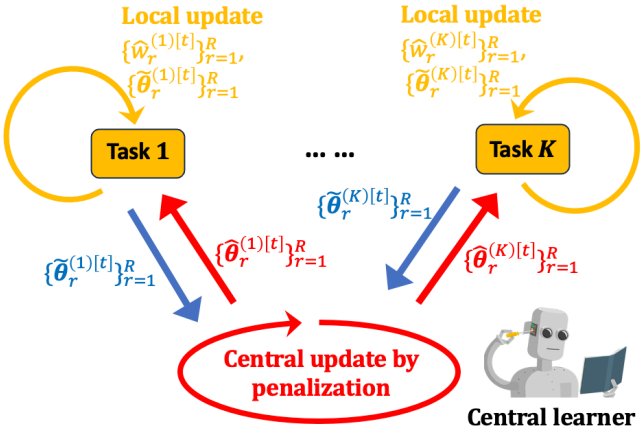

Figure 1. An illustration of Algorithm 1 (the iteration round t).

maximizer of $\widehat{Q}^{(k)}$. It is also *communicationally efficient* because it only necessitates the exchange of parameter estimators between local and central servers. Therefore, FedGrEM fulfills *all five of the desired properties* we outlined in Section 2.1.

2.4. Our Contributions

Firstly, we introduced FedGrEM, a federated EM algorithm that exhibits robustness against a small number of adversarially contaminated tasks while maintaining computational and communication efficiency. FedGrEM supplements the existing federated EM algorithms in literature (Dieuleveut et al., 2021; Marfoq et al., 2021; Wu et al., 2023) by considering the heterogeneity across tasks and adversarial contaminations.

Second, we provided an extensive non-asymptotic theory for FedGrEM on general mixture models. We characterized the estimation error of $w_r^{(k)*}$ and $\boldsymbol{\theta}_r^{(k)*}$ for non-outlier tasks by five main components:

- Iterative error, which vanishes as the number of iterations goes to infinity;
- Aggregation rate, which depends on the combined sample sizes of non-outlier tasks;
- Cost of heterogeneous mixture proportions;
- Cost of task heterogeneity;
- Cost of outlier tasks.

This analysis revealed that when the tasks exhibit sufficient similarity and the proportion of outlier tasks is sufficiently small, the estimation error of FedGrEM surpasses the rate achieved by typical single-task algorithms such as the local single-task EM. Since the setting of existing federated EM papers can be seen as a special case (no model heterogeneity and contaminations) of the scenario we study, our theory helps illustrate the empirical success of existing federated EM algorithms (Dieuleveut et al., 2021; Marfoq et al., 2021; Wu et al., 2023) and offers new theoretical insights for unsupervised federated learning.

Thirdly, we addressed the often overlooked issue of cluster label permutation in federated EMs. While label permutation is not a concern for single-task EM, most federated EM algorithms require all non-outlier tasks to share the same permutation as they take an average over the parameter estimators for the same cluster in the M-step. Failure to align the label permutations across non-outlier tasks can lead to the failure of federated EM algorithms. Due to space limit, we leave this part to Section C of the appendix.

3. Theory

In this section, we introduce a generic non-asymptotic upper bound for the estimation error of FedGrEM and apply this theory to two statistical examples: Gaussian Mixture Models (GMMs) and Mixture of Regressions (MoRs). Our theoretical findings offer a clear interpretation, shedding light on the conditions under which existing federated EM algorithms, including FedGrEM, can outperform local single-task learning. To streamline the presentation, we provide simplified versions of most theoretical results here, with formal details available in Section A of the appendix.

3.1. Generic Analysis

For simplicity, denote $q^{(k)}(\boldsymbol{\theta}) = Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}^{(k)*}, \boldsymbol{\theta}^{(k)*})$. We first state a few assumptions that are necessary for our results on general mixture models.

Assumption 3.1 (Concavity and smoothness, a simplified version of Assumption A.1)**.** For all $k \in S$, there exist non-negative constants $\{\mu_r^{(k)}\}_{r=1}^R$ and $\{L_r^{(k)}\}_{r=1}^R$ such that for all $\boldsymbol{\theta} = \{\boldsymbol{\theta}_r\}_{r=1}^R$, $\boldsymbol{\theta}' = \{\boldsymbol{\theta}_r'\}_{r=1}^R$:

(i) (Strong concavity) $q^{(k)}(\boldsymbol{\theta}') - q^{(k)}(\boldsymbol{\theta}) - \frac{\partial}{\partial \boldsymbol{\theta}} q^{(k)}(\boldsymbol{\theta})^T(\boldsymbol{\theta}' - \boldsymbol{\theta}) \leq -\sum_{r=1}^R \frac{\mu_r^{(k)}}{2}\|\boldsymbol{\theta}_r' - \boldsymbol{\theta}_r\|_2^2$;

(ii) (Smoothness) $q^{(k)}(\boldsymbol{\theta}') - q^{(k)}(\boldsymbol{\theta}) - \frac{\partial}{\partial \boldsymbol{\theta}} q^{(k)}(\boldsymbol{\theta})^T(\boldsymbol{\theta}' - \boldsymbol{\theta}) \geq -\sum_{r=1}^R \frac{L_r^{(k)}}{2}\|\boldsymbol{\theta}_r' - \boldsymbol{\theta}_r\|_2^2$.

Remark 3.2. The same conditions are imposed by Balakrishnan et al. (2017) in the single-task setting. The strong concavity is usually assumed to obtain the parametric convergence rate, and the smoothness is imposed for gradient descent to converge at a geometric rate.

Assumption 3.3 (Contraction and convergence, a simplified version of Assumptions A.3 and A.5)**.** There exist a constant $\kappa \in (0,1)$, and rate functions $\mathcal{R}_w(n)$, $\mathcal{R}_{\boldsymbol{\theta}}(n)$, such that for any $k \in S$, such that for all $\boldsymbol{w}' = \{w_r'\}_{r=1}^R$ and $\boldsymbol{\theta}' = \{\boldsymbol{\theta}_r'\}_{r=1}^R$ close to $\{w_r^{(k)*}\}_{r=1}^R$ and $\{w_r^{(k)*}\}_{r=1}^R$:

(i) (Contraction)

(a) $\big|\mathbb{E}\big[\mathbb{P}(z^{(k)} = r|\boldsymbol{x}^{(k)}; \boldsymbol{w}', \boldsymbol{\theta}')\big] - w_r^{(k)*}\big| \leq \kappa \cdot \sum_{r=1}^R \big(|w_r' - w_r^{(k)*}| + \|\boldsymbol{\theta}_r' - \boldsymbol{\theta}_r^{(k)*}\|_2\big)$;

(b) $\big\|\frac{\partial}{\partial \boldsymbol{\theta}_r} q^{(k)}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} - \frac{\partial}{\partial \boldsymbol{\theta}_r} Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}', \boldsymbol{\theta}')|_{\boldsymbol{\theta}=\boldsymbol{\theta}'}\big\|_2 \leq \kappa \cdot \sum_{r=1}^R \big(|w_r' - w_r^{(k)*}| + \|\boldsymbol{\theta}_r' - \boldsymbol{\theta}_r^{(k)*}\|_2\big)$

(ii) (Uniform convergence) w.p. $1 - \mathcal{O}(1)$,

(a) $\big|\frac{1}{n}\sum_{i=1}^n \mathbb{P}(z^{(k)} = r|\boldsymbol{x}_i^{(k)}; \boldsymbol{w}', \boldsymbol{\theta}') - \mathbb{E}\big[\mathbb{P}(z^{(k)} = r|\boldsymbol{x}^{(k)}; \boldsymbol{w}', \boldsymbol{\theta}')\big]\big| \leq \mathcal{R}_w(n)$;

(b) $\big\|\frac{\partial}{\partial \boldsymbol{\theta}_r}\widehat{Q}^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}', \boldsymbol{\theta}')|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} - \frac{\partial}{\partial \boldsymbol{\theta}_r} Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}', \boldsymbol{\theta}')|_{\boldsymbol{\theta}=\boldsymbol{\theta}'}\big\|_2 \leq \mathcal{R}_{\boldsymbol{\theta}}(n)$;

(c) $\big\|\frac{1}{|S|}\sum_{k\in S}\big[\frac{\partial}{\partial \boldsymbol{\theta}_r}\widehat{Q}^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}', \boldsymbol{\theta}')|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} - \frac{\partial}{\partial \boldsymbol{\theta}_r} Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}', \boldsymbol{\theta}')|_{\boldsymbol{\theta}=\boldsymbol{\theta}'}\big]\big\|_2 \leq \mathcal{R}_{\boldsymbol{\theta}}(nK)$.

Note that $\mathcal{R}_w(n)$ and $\mathcal{R}_{\boldsymbol{\theta}}(n)$ also depend on other parameters such as d and R, which we suppress in the notation for the ease of presentation.

Remark 3.4. Note that by definition $w_r^{(k)*} = \mathbb{E}[\mathbb{P}(z^{(k)} = r|\boldsymbol{x}^{(k)}; \boldsymbol{w}^{(k)*}, \boldsymbol{\theta}^{(k)*})]$. Therefore, condition (i) describes the behavior of $\mathbb{E}\big[\mathbb{P}(z^{(k)} = r|\boldsymbol{x}^{(k)}; \boldsymbol{w}', \boldsymbol{\theta}')\big]$ and $\frac{\partial}{\partial \boldsymbol{\theta}_r} Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}', \boldsymbol{\theta}')$, and condition (ii) is a uniform convergence assumption on the same quantities, when $\boldsymbol{w}'$ and $\boldsymbol{\theta}'$ are close to the true values $\boldsymbol{w}^{(k)*}$ and $\boldsymbol{\theta}^{(k)*}$. Condition (i) has been used by Balakrishnan et al. (2017) in the single-task setting. Condition (ii).(b) and condition (ii).(c) are uniform convergence assumptions on the gradient around the true parameter values, which are often needed when analyzing the EM without data splitting (Yan et al., 2017; Cai et al., 2019). Condition (ii).(c) is a generalization of (ii).(b) when aggregating the data from multiple tasks. As we will see in later examples, we usually have $R_w(n) = \widetilde{\mathcal{O}}(R^2\sqrt{1/n})$ and $R_{\boldsymbol{\theta}}(n) = \widetilde{\mathcal{O}}(R^2\sqrt{d/n})$, and $\mathcal{R}_{\boldsymbol{\theta}}(n)$ is typically the estimation error of local single-task methods.

Assumption 3.5 (Good initialization and step size, a simplified version of Assumption A.7)**.** $|\widehat{w}_r^{(k)[0]} - w_r^{(k)*}| \vee \|\widehat{\boldsymbol{\theta}}_r^{(k)[0]} - \boldsymbol{\theta}_r^{(k)*}\|_2 \leq C$, $\eta_r^{(k)} \leq 1/L_r^{(k)}$, for all $k \in S$ and $r \in [R]$, where $C > 0$ is a constant whose explicit form can be found in the appendix.

We set the penalty parameters in Algorithm 1 by induction as

$$\lambda^{[0]} = C_1\sqrt{n},$$
$$\lambda^{[t]} = \kappa' \cdot \lambda^{[t-1]} + C_2\sqrt{n}[\mathcal{R}_w(n) + \mathcal{R}_{\boldsymbol{\theta}}(n)],$$

where $t \geq 1$ and the explicit forms of $\kappa' \in (0,1)$, C_1, C_2 can be found in the appendix.

Next, we present our primary result for the estimation error of FedGrEM in Theorem 3.6.

Theorem 3.6 (Main result, a simplified version of Theorem A.8)**.** *Suppose Assumptions 3.1, 3.3, and 3.5 hold. Then for any contaminated set S^c with $\epsilon = |S^c|/K < 1/3$ and any contamination distribution $\mathbb{Q}_{S^c}$, w.p. $1 - \mathcal{O}(1)$, for all $T \geq 1$, FedGrEM satisfies*

$$\max_{k\in S, r\in[R]} \big(|\widehat{w}_r^{(k)[T]} - w_r^{(k)*}| \vee \|\widehat{\boldsymbol{\theta}}_r^{(k)[T]} - \boldsymbol{\theta}_r^{(k)*}\|_2\big) \lesssim \underbrace{\kappa_0^T}_{\text{iterative error}} + \underbrace{\mathcal{R}_{\boldsymbol{\theta}}(nK)}_{\text{aggregation rate}} + \underbrace{\mathcal{R}_w(n)}_{\text{cost of heterogeneous mixture proportions}} + \underbrace{\min\big\{h, \mathcal{R}_w(n) + \mathcal{R}_{\boldsymbol{\theta}}(n)\big\}}_{\text{cost of task heterogeneity}} + \underbrace{\epsilon\big[\mathcal{R}_w(n) + \mathcal{R}_{\boldsymbol{\theta}}(n)\big]}_{\text{cost of outlier tasks}}, \quad (3)$$

where $\kappa_0 \in (0,1)$.

The upper bound of convergence rate in Theorem 3.6 comprises multiple terms, each with a clear interpretation. The first term corresponds to the geometric iterative error which vanishes as $T \to +\infty$, and the second term accounts for the aggregation error which arises from combining all the data. The third to fifth terms represent the cost of heterogeneous mixing proportions, task heterogeneity, and outlier tasks, respectively.

As we will observe in later specific examples, $\mathcal{R}_{\boldsymbol{\theta}}(nK)$ scales as $R^2\sqrt{d/(nK)}$ which depends on the total sample size nK of all K tasks, $\mathcal{R}_{\boldsymbol{\theta}}(n) = \widetilde{\mathcal{O}}(R^2\sqrt{d/n})$, and $\mathcal{R}_w(n) = \widetilde{\mathcal{O}}(R^2\sqrt{1/n})$. Note that $\mathcal{R}_{\boldsymbol{\theta}}(n)$ typically represents the estimation error of local single-task algorithms. Comparing (3) with $\mathcal{R}_{\boldsymbol{\theta}}(n) = \widetilde{\mathcal{O}}(R^2\sqrt{d/n})$, we can see that when both h and ϵ are small — indicating sufficient similarity shared across tasks and few contaminated tasks — FedGrEM can achieve a better estimation error than the local single-task methods. When $h = \epsilon = 0$, implying that all tasks share the same parameters, we revert to the setting of Dieuleveut et al. (2021), Marfoq et al. (2021), and Wu et al. (2023). In this context, our finite-sample upper bound demonstrates that federated EM can indeed outperform the local single-task methods, aligning with the empirical success of federated EM algorithms observed in these works.

In subsequent sections, we will substitute specific rate expressions for each term in concrete examples, by showing that $\mathcal{R}_w = \widetilde{\mathcal{O}}(R^2\sqrt{1/n})$ and $\mathcal{R}_{\boldsymbol{\theta}}(n) = \widetilde{\mathcal{O}}(R^2\sqrt{d/n})$.

3.2. Proof Sketch of Theorem 3.6

We briefly describe the proof of Theorem 3.6 here. The proof follows an iterative fashion, where we show a connection between the estimation error rates in two consecutive iteration rounds and then iterate the analysis to obtain the final result. More specifically, by utilizing the contraction and uniform convergence conditions assumed in Assumption 3.3, if we define the estimation error of round t as $\mathrm{Er}(t) = \max_{k\in S, r\in[R]} \big(|\widehat{w}_r^{(k)[t]} - w_r^{(k)*}| \vee \|\widehat{\boldsymbol{\theta}}_r^{(k)[t]} - \boldsymbol{\theta}_r^{(k)*}\|_2\big)$, we can prove that with high probability,

$$\mathrm{Er}(t) \le \kappa_0 \mathrm{Er}(t-1) + \text{other terms},$$

where other terms include the sum of the last four terms on the RHS of (3), and $\kappa_0 \in (0,1)$.

We want to highlight that the proof is much harder and more complicated than the proofs in standard EM theory. The reason is that the mixture proportions $\{w_r^{(k)*}\}_{r=1}^R$ can be heterogenous across tasks, where we can still benefit from federated learning because the similarity between d-dimensional parameters $\{\boldsymbol{\theta}_r^{(k)*}\}_{r=1}^R$ is more important than the heterogeneity of 1-dimensional scalers $\{w_r^{(k)*}\}_{r=1}^R$. However, in standard EM theory (Balakrishnan et al., 2017; Yan et al., 2017; Cai et al., 2019), the estimation errors of $\{w_r^{(k)*}\}_{r=1}^R$ and $\{\boldsymbol{\theta}_r^{(k)*}\}_{r=1}^R$ are entangled and it is challenging to separate them by the existing theory. We creatively used a *localization* technique to address the issue by adaptively shrinking the radius of the ball within which uniform convergence in Assumption 3.3 must hold. This adaptive radius shrinking trick during iterations finally leads to a "fast rate", effectively replacing $\mathcal{R}_{\boldsymbol{\theta}}(n)$ ("the slow rate") with a much smaller $\mathcal{R}_w(n)$ for the term "cost of heterogeneous mixing proportions" in (3). The intuition is visually interpreted in Figure 2, and more details can be found in the full proof of Theorem 3.6 in the appendix.

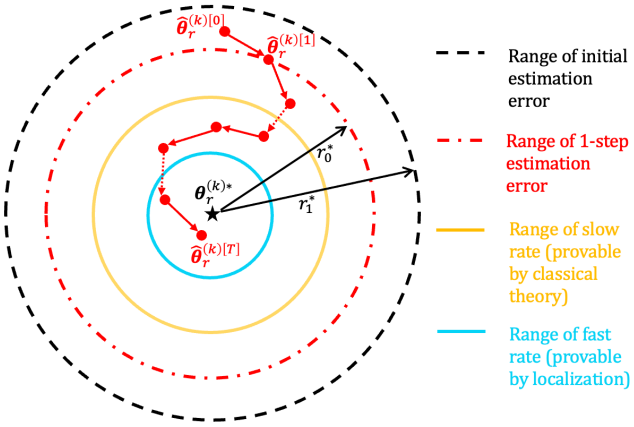

Figure 2. Schematic of the geometric convergence and the localization trick, where we shrink the radius of uniform convergence ball from r_1^* to r_0^* after the first iteration.

3.3. Example 1: Gaussian Mixture Models (GMMs)

In this section, we examine Gaussian Mixture Models (GMMs) as an example of (1). Each observation is from a mixture of R Gaussian distributions ($R \ge 2$):

$$\boldsymbol{x}_i^{(k)} \overset{i.i.d.}{\sim} \sum_{r=1}^{R} w_r^{(k)*} \cdot N(\boldsymbol{\theta}_r^{(k)*}, \boldsymbol{I}_{d\times d}). \tag{4}$$

Hence in (1), $p_r^{(k)}(\,\cdot\,;\boldsymbol{\theta}_r^{(k)*})$ represents the Lebesgue density of Gaussian distribution $N(\boldsymbol{\theta}_r^{(k)*}, \boldsymbol{I}_{d\times d})$. We define $\Delta := \min_{k\in S}\min_{r\neq r'}\|\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_{r'}^{(k)*}\|_2$, which characterizes the signal-to-noise ratio of GMMs among S. We impose the following assumption.

Assumption 3.7. Suppose the following conditions hold:

(i) (Bounded parameters) $w_r^{(k)*} \gtrsim 1/R$, $\|\boldsymbol{\theta}_r^{(k)*}\|_2 \le C$

for all $k \in S$ and $r \in [R]$, where C is a constant;

(ii) (Good initialization) $\max_{k\in S, r\in[R]} |\widehat{w}_r^{(k)[0]} - w_r^{(k)*}| \lesssim \frac{1}{R}$, $\max_{k\in S, r\in[R]} \|\widehat{\boldsymbol{\theta}}_r^{(k)[0]} - \boldsymbol{\theta}_r^{(k)*}\|_2 \lesssim \Delta$;

(iii) (Large signal strength) $\Delta \gtrsim \log(R)$;

(iv) (Sample size) $n \gtrsim R^4[d + \log(RK)]\Delta^{-2}$;

(v) (Step size) $1 - \eta_r^{(k)} w_r^{(k)*} < c$, and $0 < \eta_r^{(k)} \leq 1/w_r^{(k)*}$ for all $k \in S$ and $r \in [R]$, where $c > 0$ is a small constant.

Proposition 3.8. *Under Assumption 3.7, GMMs defined in (4) satisfies Assumptions 3.1, 3.3, and 3.5 with*

$$\mu_r^{(k)} = L_r^{(k)} = w_r^{(k)*}, \quad \kappa \asymp R^2 \exp\{-C\Delta^2\},$$
$$\mathcal{R}_w(n) = \widetilde{\mathcal{O}}\left(R^2\sqrt{\frac{1}{n}}\right), \quad \mathcal{R}_{\boldsymbol{\theta}}(n) = \widetilde{\mathcal{O}}\left(R^2\sqrt{\frac{d}{n}}\right),$$

where $C > 0$ is some constant.

By plugging the rates in Propositions 3.8 into Theorem 3.6, we obtain the following estimation error for GMMs.

Corollary 3.9. *Under Assumption 3.7, for the GMMs defined in (4), for any contaminated set S^c with $\epsilon = |S^c|/K \leq 1/3$ and contaminated distribution $\mathbb{Q}_{S^c}$, w.p. $1 - o(1)$, for all $T \geq 1$, FedGrEM satisfies*

$$\max_{k\in S, r\in[R]} \left(|\widehat{w}_r^{(k)[T]} - w_r^{(k)*}| \vee \|\widehat{\boldsymbol{\theta}}_r^{(k)[T]} - \boldsymbol{\theta}_r^{(k)*}\|_2\right)$$
$$= \widetilde{\mathcal{O}}\left(\kappa_0^T + R^2\sqrt{\frac{d}{nK}} + R^2\sqrt{\frac{1}{n}} + \min\left\{h, R^2\sqrt{\frac{d}{n}}\right\} + \epsilon R^2\sqrt{\frac{d}{n}}\right).$$

where $\kappa_0 \in (0,1)$ is a constant.

Our theoretical analysis is also applicable to local single-task EM and gradient EM, enabling us to establish an upper bound of estimation error $\widetilde{\mathcal{O}}_{\mathbb{P}}\left(R^2\sqrt{\frac{d}{n}}\right)$ on S. Consequently, when $d \to \infty$ (diverging dimension), $K \to \infty$ (many similar tasks), $h \ll R^2\sqrt{\frac{d}{n}}$ (sufficient similarity), and $\epsilon \to 0$ (small proportion of outlier tasks), FedGrEM exhibits a better estimation error rate than single-task EM and gradient EM (up to logarithmic factors). Notably, FedGrEM always achieves an error rate at least as good as the single-task rate $\widetilde{\mathcal{O}}_{\mathbb{P}}\left(R^2\sqrt{\frac{d}{n}}\right)$.

3.4. Example 2: Mixture of Regressions (MoRs)

As a second example, we consider a mixture of linear regressions (MoRs), where each observation comes from a mixture of R linear regression models ($R \geq 2$):

$$z_i^{(k)} \overset{i.i.d.}{\sim} \sum_{r=1}^{R} w_r^{(k)*} \cdot \delta_r,$$
$$\text{Given } z_i^{(k)} = r: \quad y_i^{(k)} = (\widetilde{\boldsymbol{x}}_i^{(k)})^T \boldsymbol{\theta}_r^{(k)*} + \epsilon_i^{(k)}, \tag{5}$$
$$\epsilon_i^{(k)} \overset{i.i.d.}{\sim} N(0,1), \quad \widetilde{\boldsymbol{x}}_i^{(k)} \overset{i.i.d.}{\sim} N(\mathbf{0}_d, \boldsymbol{I}_{d\times d}), \quad \epsilon_i^{(k)} \perp\!\!\!\perp \widetilde{\boldsymbol{x}}_i^{(k)}.$$

Hence in (1) and (2), $\boldsymbol{x}_i^{(k)}$ is the pair $(\widetilde{\boldsymbol{x}}_i^{(k)}, y_i^{(k)})$ and $p_r^{(k)}(\cdot\,; \boldsymbol{\theta}_r^{(k)*})$ is the Lebesgue density of joint distribution of $(\widetilde{\boldsymbol{x}}_i^{(k)}, y_i^{(k)})$. We impose the following assumption set.

Assumption 3.10. Suppose the same conditions in Assumption 3.7 hold by replacing (iii) with:

(iii) (Large signal strength) $\Delta \gtrsim R^3 + R^2(\log\Delta)^{3/2}$.

Proposition 3.11. *Under Assumption 3.10, the MoRs defined in (5) satisfies Assumptions 3.1, 3.3, and 3.5 with*

$$\mu_r^{(k)} = w_r^{(k)*} - CR\frac{\sqrt{\log\Delta}}{\Delta}, L_r^{(k)} = w_r^{(k)*} + CR\frac{\sqrt{\log\Delta}}{\Delta},$$
$$\kappa = \widetilde{\mathcal{O}}\left(\tfrac{R^2}{\Delta}\right), \mathcal{R}_w(n) = \widetilde{\mathcal{O}}\left(R^2\sqrt{\tfrac{1}{n}}\right), \mathcal{R}_{\boldsymbol{\theta}}(n) = \widetilde{\mathcal{O}}\left(R^2\sqrt{\tfrac{d}{n}}\right),$$

where $C > 0$ is some constant.

By plugging the rates in Propositions 3.11 into Theorem 3.6, we have the following estimation error for MoRs.

Corollary 3.12. *Under Assumption 3.10, for the MoRs defined in (5), for any contaminated set S^c with $\epsilon = |S^c|/K \leq 1/3$ and contaminated distribution $\mathbb{Q}_{S^c}$, with probability $1 - o(1)$, for all $T \geq 1$, FedGrEM satisfies*

$$\max_{k\in S, r\in[R]} \left(|\widehat{w}_r^{(k)[T]} - w_r^{(k)*}| \vee \|\widehat{\boldsymbol{\theta}}_r^{(k)[T]} - \boldsymbol{\theta}_r^{(k)*}\|_2\right) = \widetilde{\mathcal{O}}\Bigg(\kappa_0^T + R^2\sqrt{\tfrac{d}{nK}} + R^2\sqrt{\tfrac{1}{n}} + \min\left\{h, R^2\sqrt{\tfrac{d}{n}}\right\} + \epsilon R^2\sqrt{\tfrac{d}{n}}\Bigg).$$

where $\kappa_0 \in (0,1)$ is a constant.

We can similarly discuss when the rate of FedGrEM is better than the estimation error of the local single-task algorithms for GMMs, which we do not repeat here.

4. Numerical Results

4.1. Simulations

In this subsection, we present simulation results to empirically validate our theoretical insights. We consider two examples in the last section: Gaussian Mixture Models (GMMs) and Mixture of Regressions (MoRs). For both examples, we set the number of tasks $K = 10$, the number of

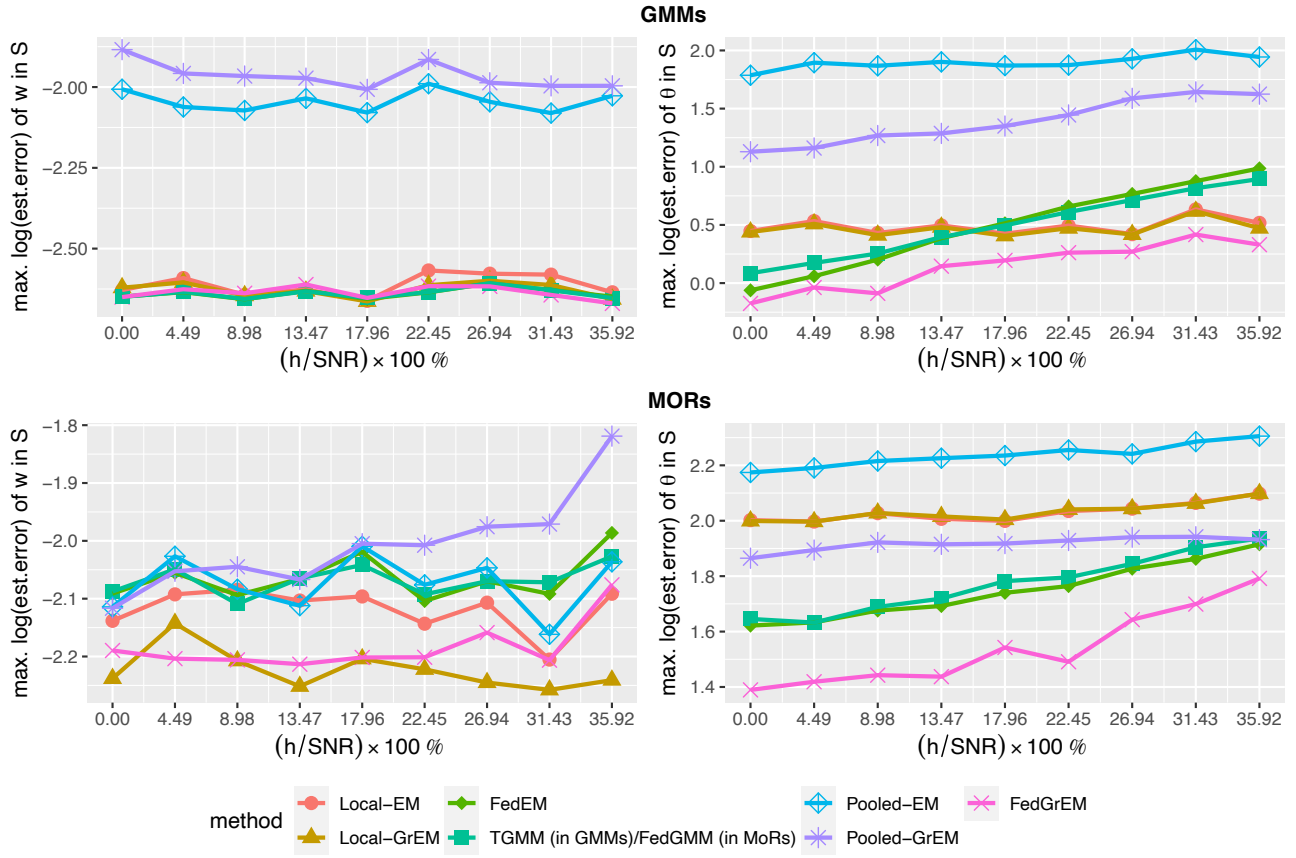

Figure 3. The average estimation errors of different methods in 100 replications of the GMM and MoR simulations (in $\log_e$ scale). The left two figures show the estimation error $\max_{k\in S}\max_{r\in[R]}\log(|\widehat{w}_r^{(k)[T]}-w_r^{(k)*}|)$ and the right two figures show the estimation error $\max_{k\in S}\max_{r\in[R]}\log(\|\widehat{\boldsymbol{\theta}}_r^{(k)[T]}-\boldsymbol{\theta}_r^{(k)*}\|_2)$. x-axis represents the ratio between model heterogeneity h and SNR (signal-to-noise ratio), where the definition of SNR is in Section B.1 of the appendix.

clusters $R=5$, the sample size of each task $n=150$ (per-cluster sample size is around 30), the dimension $p=10$, and introduce one outlier task ($\epsilon=0.1$). The simulations are conducted over 100 repetitions, and we generate $w^{(k)*}$ in dependently from Dirichlet$(5,5,5,5,5)$. Details regarding the values of $\boldsymbol{\theta}^{(k)*}$'s and the generation mechanism of the outlier task can be found in Section B of the appendix. We vary the value of h from 0 to 2 with an increment of 0.25, calculating the average mean estimation error of $w^{(k)*}$'s and $\boldsymbol{\theta}^{(k)*}$'s in $S=1:9$ for different approaches.

The considered methods include several unsupervised federated learning or multi-task learning benchmarks: Local-EM (single-task EM), Local-GrEM (single-task gradient EM), FedEM (Marfoq et al., 2021), TGMM (Wang et al., 2021), FedGMM (Wu et al., 2023), Pooled-EM (EM on pooled data), Pooled-GrEM (gradient EM on pooled data), and FedGrEM (ours).

The results are presented in Figure 3. In the GMM simulation, the estimation errors of $w^{(k)*}$'s for different methods are similar, except for EM methods on the pooled data. FedGrEM, FedEM, TGMM, and FedGMM outperform the others in estimating $\boldsymbol{\theta}^{(k)*}$'s with FedGrEM surpassing the other three due to its robustness to outlier tasks. As h increases, the performance of FedEM and TGMM degrades, becoming inferior to EM and GrEM. Conversely, FedGrEM's performance becomes comparable to EM and GrEM as h/SNR approaches 35.92%, highlighting its adaptability to unknown h. Similar trends are observed in the MoR simulation. These numerical results align with our theoretical analyses and confirm FedGrEM's advantage in handling unknown similarity levels h and a few outlier tasks.

4.2. Real-data Studies

We also conduct experiments on three real datasets:

ϵ/Method	Local-EM	Local-GrEM	FedEM	FedGrEM	TGMM	Pooled-EM	Pooled-GrEM
0%	13.30 (1.17)	13.20 (1.77)	15.75 (2.09)	**9.62 (1.03)**	25.14 (2.72)	26.84 (5.54)	34.88 (11.70)
6.8%	12.88 (1.11)	12.96 (1.70)	15.87 (2.20)	**9.31 (1.10)**	25.00 (3.16)	26.60 (5.80)	34.61 (11.92)
13.6%	12.62 (1.26)	12.99 (1.99)	17.56 (2.44)	**9.04 (1.06)**	25.68 (2.85)	26.79 (5.53)	35.61 (12.52)
20.5%	12.94 (1.37)	13.31 (1.93)	18.52 (2.38)	**9.39 (1.25)**	26.14 (3.08)	27.05 (5.46)	37.94 (13.65)

Table 1. Average mis-clustering error rates (standard deviations) in percentages for Pen-Based Recognition of Handwritten Digits dataset

ϵ/Method	Local-EM	Local-GrEM	FedEM	FedGrEM	TGMM	Pooled-EM	Pooled-GrEM
0%	12.47 (1.19)	12.06 (0.95)	**9.97 (2.09)**	10.63 (2.72)	12.18 (4.08)	12.84 (2.92)	10.67 (2.41)
8%	12.30 (1.10)	11.98 (0.88)	11.97 (2.80)	**10.64 (1.98)**	13.66 (4.13)	15.07 (5.25)	14.16 (4.23)
16%	12.47 (1.19)	12.06 (0.93)	14.46 (2.95)	**10.96 (1.96)**	14.47 (4.24)	16.16 (5.46)	14.97 (4.36)
24%	12.43 (1.11)	12.10 (0.89)	21.67 (4.33)	**10.97 (1.59)**	15.70 (5.01)	15.50 (5.09)	14.35 (3.96)

Table 2. Average mis-clustering error rates (standard deviations) in percentages for MNIST dataset

ϵ/Method	Local-EM	Local-GrEM	FedEM	FedGrEM	TGMM	Pooled-EM	Pooled-GrEM
0%	36.40 (0.67)	35.66 (0.64)	36.06 (1.42)	**34.44 (1.98)**	36.14 (2.89)	36.31 (1.25)	35.72 (1.07)
8%	36.40 (0.72)	35.69 (0.65)	36.82 (1.15)	**33.27 (1.96)**	36.71 (3.32)	39.75 (1.41)	39.71 (1.54)
16%	36.34 (0.66)	35.65 (0.70)	36.80 (1.28)	**33.29 (1.86)**	37.50 (3.16)	39.65 (1.50)	39.34 (1.91)
24%	36.31 (0.69)	35.61 (0.73)	37.93 (1.81)	**33.35 (1.69)**	38.20 (3.28)	39.51 (1.24)	39.44 (1.57)

Table 3. Average mis-clustering error rates (standard deviations) in percentages for Fashion-MNIST dataset

- Pen-Based Recognition of Handwritten Digits [3]: This dataset collects 0-9 digits written by 44 writers on the tablet with 16 features related to each digit such as the pressure level at certain coordinates. The ID of the writer for each handwritten digit is provided. Hence, it is a federated multi-task learning dataset in nature by viewing each writer as a client.
- MNIST [4]: 70000 grayscale images of 28×28 pixels for the handwritten digits 0-9 from different writers. There is no information about the writers, hence we manually created a federated learning dataset by randomly assigning each image to one of 100 clients.
- Fashion-MNIST [5]: 70000 of Zalando's article grayscale images in 28×28 pixels, each associated with a label from 10 classes. We manually created a federated learning dataset by randomly assigning each image to one of 100 clients.

In each replication, 80% data for each task is used as training data, and the remaining 20% is used as test data to calculate the mis-clustering error. We also contaminate different proportions (ϵ) of tasks to showcase the robustness of FedGrEM against adversarial attacks. We run different benchmark methods with GMMs (without the knowledge of the true labels) for 200 replications and compare their performances in terms of mis-clustering error rates, as shown in Tables 1-3, where FedGrEM performs the best in most settings. More details about the datasets, preprocessing steps, and results can be found in Section B.2 of the appendix.

[3] https://archive.ics.uci.edu/dataset/81/pen+based+recognition+of+handwritten+digits

[4] https://www.kaggle.com/datasets/hojjatk/mnist-dataset

[5] https://www.kaggle.com/datasets/zalando-research/fashionmnist

5. Discussions

In this work, we introduced a federated gradient EM algorithm (FedGrEM) to enhance the existing federated EM methods, by considering the task heterogeneity and adversarial attacks. We studied the non-asymptotic theory on general mixture models, and applied the theory to GMMs and MoRs to obtain the explicit estimation error of the model parameters and mixture proportions. Our theory helps illustrate the empirical success of existing federated EM algorithms in literature and offers new theoretical insights on unsupervised federated learning. The proposed FedGrEM was shown to be adaptive to unknown task similarity, robust against the adversarial attack on a small proportion of tasks, protective for the local data, computationally and communicationally efficient. It serves as a valuable supplement to existing federated EM algorithms.

Some additional discussions on the limitations and future extensions are available in Section D of the appendix.

Acknowledgements

All the experiments were conducted on Ginsburg HPC cluster of Columbia University. Haolei Weng was partially supported by NSF-DMS 2210505. Yang Feng was partially supported by NSF Grant DMS-2324489 and NIH Grant 1R21AG074205-01.

Impact Statement

This paper aims to study the non-asymptotic theory underlying the federated EM algorithms, accounting for task heterogeneity and adversarial contaminations. It seeks to illustrate when and how federated learning can improve local performance and offers novel insights into unsupervised federated learning, a field with potential societal impacts, particularly in data privacy and security. While this study primarily advances Machine Learning, we acknowledge its indirect implications on ethical considerations in data handling and algorithm deployment. We believe our findings will contribute to the development of more secure and ethically-aware federated learning algorithms.

References

Abbe, E., Fan, J., Wang, K., and Zhong, Y. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452, 2020.

Antunes, R. S., André da Costa, C., Küderle, A., Yari, I. A., and Eskofier, B. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13 (4):1–23, 2022.

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

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

Dai, W., Yang, Q., Xue, G.-R., and Yu, Y. Self-taught clustering. In *Proceedings of the 25th international conference on Machine learning*, pp. 200–207, 2008.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. Learning to learn around a common mean. *Advances in neural information processing systems*, 31, 2018.

Dieuleveut, A., Fort, G., Moulines, E., and Robin, G. Federated-em with heterogeneity mitigation and variance reduction. *Advances in Neural Information Processing Systems*, 34:29553–29566, 2021.

Duan, Y. and Wang, K. Adaptive and robust multi-task learning. *arXiv preprint arXiv:2202.05250*, 2022.

Evgeniou, T. and Pontil, M. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 109–117, 2004.

Ghosh, A., Chung, J., Yin, D., and Ramchandran, K. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33: 19586–19597, 2020.

Gu, Q., Li, Z., and Han, J. Learning a kernel for multi-task clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pp. 368–373, 2011.

Guminov, S., Dvurechensky, P., Tupitsa, N., and Gasnikov, A. On a combination of alternating minimization and nesterov's momentum. In *International conference on machine learning*, pp. 3886–3898. PMLR, 2021.

Hanneke, S. and Kpotufe, S. On the value of target data in transfer learning. *arXiv preprint arXiv:2002.04747*, 2020.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

He, Z., Sun, Y., and Li, R. Transfusion: Covariate-shift robust transfer learning for high-dimensional regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 703–711. PMLR, 2024.

Hinton, G. E. and Roweis, S. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002.

Jaggi, M., Smith, V., Takác, M., Terhorst, J., Krishnan, S., Hofmann, T., and Jordan, M. I. Communication-efficient distributed dual coordinate ascent. *Advances in neural information processing systems*, 27, 2014.

Jalali, A., Ravikumar, P., and Sanghavi, S. A dirty model for multiple sparse regression. *IEEE Transactions on Information Theory*, 59(12):7947–7968, 2013.

Konečnỳ, J., McMahan, H. B., Ramage, D., and Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

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

Konstantinov, N. and Lampert, C. Robust learning from untrusted sources. In *International conference on machine learning*, pp. 3488–3498. PMLR, 2019.

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

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

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

Kwon, J. and Caramanis, C. The em algorithm gives sample-optimality for learning mixtures of well-separated gaussians. In *Conference on Learning Theory*, pp. 2425–2487. PMLR, 2020a.

Kwon, J. and Caramanis, C. Em converges for a mixture of many linear regressions. In *International Conference on Artificial Intelligence and Statistics*, pp. 1727–1736. PMLR, 2020b.

Lan, G. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

Li, Q., Zhu, Z., and Tang, G. Alternating minimizations converge to second-order optimal solutions. In *International Conference on Machine Learning*, pp. 3935–3943. PMLR, 2019.

Li, S., Cai, T. T., and Li, H. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pp. 1–25, 2021.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020a.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020b.

Li, X. and Bilmes, J. A bayesian divergence prior for classiffier adaptation. In *Artificial Intelligence and Statistics*, pp. 275–282. PMLR, 2007.

Lin, H. and Reimherr, M. On hypothesis transfer learning of functional linear models. *stat*, 1050:22, 2024a.

Lin, H. and Reimherr, M. Smoothness adaptive hypothesis transfer learning. *arXiv preprint arXiv:2402.14966*, 2024b.

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

Marfoq, O., Neglia, G., Bellet, A., Kameni, L., and Vidal, R. Federated multi-task learning under a mixture of distributions. *Advances in Neural Information Processing Systems*, 34:15434–15447, 2021.

Maurer, A. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pp. 3–17. Springer, 2016.

Maurer, A. and Pontil, M. Concentration inequalities under sub-gaussian and sub-exponential conditions. *Advances in Neural Information Processing Systems*, 34: 7588–7597, 2021.

McLachlan, G. J. and Krishnan, T. *The EM algorithm and extensions*. John Wiley & Sons, 2007.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Meng, X.-L. and Rubin, D. B. On the global and componentwise rates of convergence of the em algorithm. *Linear Algebra and its Applications*, 199:413–425, 1994.

Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., and Poor, H. V. Federated learning for internet of things: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 23(3):1622–1658, 2021.

Polson, N., Scott, J. G., and Willard, B. T. Proximal algorithms in statistics and machine learning. *Statistical science*, 30(4):559–581, 2015.

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

Redner, R. A. and Walker, H. F. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26 (2):195–239, 1984.

Rohe, K., Chatterjee, S., and Yu, B. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):596, 2011.

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. S. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

Solnon, M., Arlot, S., and Bach, F. Multi-task regression using minimal penalties. *The Journal of Machine Learning Research*, 13(1):2773–2812, 2012.

Su, L., Xu, J., and Yang, P. Global convergence of federated learning for mixed regression. *Advances in Neural Information Processing Systems*, 35:29889–29902, 2022.

T Dinh, C., Tran, N., and Nguyen, J. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

Tang, M. and Priebe, C. E. Limit theorems for eigenvectors of the normalized laplacian for random graphs. *The Annals of Statistics*, 46(5):2360–2415, 2018.

Tian, Y. and Feng, Y. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697, 2023.

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

Tian, Y., Gu, Y., and Feng, Y. Learning from similar linear representations: Adaptivity, minimaxity, and robustness. *arXiv preprint arXiv:2303.17765*, 2023.

Tupitsa, N., Dvurechensky, P., Gasnikov, A., and Guminov, S. Alternating minimization methods for strongly convex optimization. *Journal of Inverse and Ill-posed Problems*, 29(5):721–739, 2021.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Varshney, P., Thakurta, A., and Jain, P. (nearly) optimal private linear regression for sub-gaussian data via adaptive clipping. In *Conference on Learning Theory*, pp. 1126–1166. PMLR, 2022.

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

Wang, R., Zhou, J., Jiang, H., Han, S., Wang, L., Wang, D., and Chen, Y. A general transfer learning-based gaussian mixture model for clustering. *International Journal of Fuzzy Systems*, 23(3):776–793, 2021.

Wu, C. J. On the convergence properties of the em algorithm. *The Annals of statistics*, pp. 95–103, 1983.

Wu, Y., Zhang, S., Yu, W., Liu, Y., Gu, Q., Zhou, D., Chen, H., and Cheng, W. Personalized federated learning under mixture of distributions. *arXiv preprint arXiv:2305.01068*, 2023.

Xu, J., Hsu, D. J., and Maleki, A. Global analysis of expectation maximization for mixtures of two gaussians. *Advances in Neural Information Processing Systems*, 29, 2016.

Yan, B., Yin, M., and Sarkar, P. Convergence of gradient em on multi-component mixture of gaussians. *Advances in Neural Information Processing Systems*, 30, 2017.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10 (2):1–19, 2019.

Yang, Y., Ma, Z., Yang, Y., Nie, F., and Shen, H. T. Multitask spectral clustering by exploring intertask correlation. *IEEE transactions on cybernetics*, 45(5):1083–1094, 2014.

Zhang, J. and Zhang, C. Multitask bregman clustering. *Neurocomputing*, 74(10):1720–1734, 2011.

Zhang, X., Zhang, X., and Liu, H. Smart multi-task bregman clustering and multitask kernel clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1):1–29, 2015.

Zhang, X., Zhang, X., Liu, H., and Luo, J. Multi-task clustering with model relation learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 3132–3140, 2018.

Zhang, Y. and Yang, Q. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

Zhao, R., Li, Y., and Sun, Y. Statistical convergence of the em algorithm on gaussian mixture models. *Electronic Journal of Statistics*, 14:632–660, 2020.

Appendix

This appendix collects the additional theoretical and numerical details as well as all the technical proofs of the theory. We present the formal theoretical results in Section A which correspond to the simplified versions in Section 3 of the main text. Section B contains more details of the numerical studies presented in Section 4 of the main text. Section C discusses the label permutation issue we mentioned in Section 2.4. Section D includes additional discussions on limitations and potential extensions of the current work in the future. All the technical proofs are summarized in Section E.

We recall our mathematical notations here. $\mathbb{P}$ and $\mathbb{E}$ denote the probability and expectation, respectively. $\mathbb{P}$ and $\mathbb{E}$ denote the probability and expectation, respectively. For two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \ll b_n$ means $a_n/b_n \to 0$, $a_n \lesssim b_n$ or $a_n = \mathcal{O}(b_n)$ means $a_n/b_n \leq C < \infty$, and $a_n \asymp b_n$ means $a_n/b_n, b_n/a_n \leq C < \infty$. For a random variable x_n and a positive sequence a_n, $x_n = \mathcal{O}_p(a_n)$ means that for any $\epsilon > 0$, there exists $M > 0$ such that $\sup_n \mathbb{P}(|x_n/a_n| > M) \leq \epsilon$. $\widetilde{\mathcal{O}}_p(a_n)$ has a similar meaning up to logarithmic factors in a_n. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, $\|x\|_2$ represents its Euclidean norm. For two numbers a and b, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For any positive integer K, $1:K$ and $[K]$ stand for the set $\{1, 2, \ldots, K\}$. Denote $\mathcal{B}_\xi(\boldsymbol{\theta})$ as an Euclidean ball centered at $\boldsymbol{\theta}$ with radius $\xi > 0$. And for any set S, $|S|$ denotes its cardinality and S^c denotes its complement. "w.p." stands for "with probability". "WLOG" stands for "Without loss of generality". The absolute constants c and C may vary from line to line.

A. Formal Theoretical Results

Denote $\bar{\eta} = \max_{k\in S, r\in[R]} \eta_r^{(k)}$ as the maximum step size used in the local M-step on tasks in S. For simplicity, denote $q^{(k)}(\boldsymbol{\theta}) = Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}^{(k)*}, \boldsymbol{\theta}^{(k)*})$. We first state a few assumptions which are necessary for our result on general mixture models.

Assumption A.1. For any $k \in S$, there exist non-negative sets $\{\mu_r^{(k)}\}_{r\in[R]}$, $\{L_r^{(k)}\}_{r\in[R]}$, and a positive constant r_1^* such that for all $\boldsymbol{\theta} = \{\boldsymbol{\theta}_r\}_{r=1}^R$, $\boldsymbol{\theta}' = \{\boldsymbol{\theta}'_r\}_{r=1}^R$ with $\boldsymbol{\theta}_r, \boldsymbol{\theta}'_r \in \mathcal{B}_{r_1^*}(\boldsymbol{\theta}_r^{(k)*})$:

(i) (Strong concavity) $q^{(k)}(\boldsymbol{\theta}') - q^{(k)}(\boldsymbol{\theta}) - \frac{\partial}{\partial\boldsymbol{\theta}} q^{(k)}(\boldsymbol{\theta})^T(\boldsymbol{\theta}' - \boldsymbol{\theta}) \leq -\sum_{r=1}^R \frac{\mu_r^{(k)}}{2}\|\boldsymbol{\theta}'_r - \boldsymbol{\theta}_r\|_2^2$;

(ii) (Smoothness) $q^{(k)}(\boldsymbol{\theta}') - q^{(k)}(\boldsymbol{\theta}) - \frac{\partial}{\partial\boldsymbol{\theta}} q^{(k)}(\boldsymbol{\theta})^T(\boldsymbol{\theta}' - \boldsymbol{\theta}) \geq -\sum_{r=1}^R \frac{L_r^{(k)}}{2}\|\boldsymbol{\theta}'_r - \boldsymbol{\theta}_r\|_2^2$.

Remark A.2. The same conditions are imposed by Balakrishnan et al. (2017) in the single-task setting. The strong concavity is usually assumed to obtain the parametric convergence rate, and the smoothness is imposed for gradient descent to converge at a geometric rate.

Assumption A.3. There exist positive constants r_w^*, r_2^*, and $\kappa \in (0, 1)$, and a function $\mathcal{W}$, such that for any $k \in S$:

(i) For all $\boldsymbol{w}' = \{w'_r\}_{r=1}^R$ and $\boldsymbol{\theta}' = \{\boldsymbol{\theta}'_r\}_{r=1}^R$ with $w'_r \in \mathcal{B}_{r_w^*}(w_r^{(k)*})$ and $\boldsymbol{\theta}'_r \in \mathcal{B}_{r_2^*}(\boldsymbol{\theta}_r^{(k)*})$, we have $\big|\mathbb{E}\big[\mathbb{P}(z^{(k)} = r|\boldsymbol{x}^{(k)}; \boldsymbol{w}', \boldsymbol{\theta}')\big] - w_r^{(k)*}\big| \leq \kappa \cdot \sum_{r=1}^R \big(|w'_r - w_r^{(k)*}| + \|\boldsymbol{\theta}'_r - \boldsymbol{\theta}_r^{(k)*}\|_2\big)$;

(ii) w.p. at least $1-\delta$, for all $\boldsymbol{w}' = \{w'_r\}_{r=1}^R$, $\xi > 0$, and $\boldsymbol{\theta}' = \{\boldsymbol{\theta}'_r\}_{r=1}^R$ with $w'_r \in \mathcal{B}_{r_w^*}(w_r^{(k)*})$ and $\boldsymbol{\theta}'_r \in \mathcal{B}_\xi(\boldsymbol{\theta}_r^{(k)*})$, we have $\big|\frac{1}{n}\sum_{i=1}^n \mathbb{P}(z^{(k)} = r|\boldsymbol{x}_i^{(k)}; \boldsymbol{w}', \boldsymbol{\theta}') - \mathbb{E}\big[\mathbb{P}(z^{(k)} = r|\boldsymbol{x}^{(k)}; \boldsymbol{w}', \boldsymbol{\theta}')\big]\big| \leq \mathcal{W}(n, \delta, \xi)$.

Remark A.4. Note that by definition $w_r^{(k)*} = \mathbb{E}[\mathbb{P}(z^{(k)} = r|\boldsymbol{x}^{(k)}; \boldsymbol{w}^{(k)*}, \boldsymbol{\theta}^{(k)*})]$. Therefore, condition (i) describes the behavior of $\mathbb{E}\big[\mathbb{P}(z^{(k)} = r|\boldsymbol{x}^{(k)}; \boldsymbol{w}', \boldsymbol{\theta}')\big]$, and condition (ii) is a uniform convergence assumption on $\mathbb{P}(z^{(k)} = r|\boldsymbol{x}_i^{(k)}; \boldsymbol{w}', \boldsymbol{\theta}')$, when $\boldsymbol{w}'$ and $\boldsymbol{\theta}'$ are close to the true values $\boldsymbol{w}^{(k)*}$ and $\boldsymbol{\theta}^{(k)*}$.

Assumption A.5. With the same constants r_w^* and r_2^* in Assumption A.3, there exists a constant $\gamma \in (0, 1)$ and functions $\mathcal{E}_1$, $\mathcal{E}_2$ such that for any $k \in S$:

(i) For all $\boldsymbol{w}' = \{w'_r\}_{r=1}^R$ and $\boldsymbol{\theta}' = \{\boldsymbol{\theta}'_r\}_{r=1}^R$ with $w'_r \in \mathcal{B}_{r_w^*}(w_r^{(k)*})$ and $\boldsymbol{\theta}'_r \in \mathcal{B}_{r_2^*}(\boldsymbol{\theta}_r^{(k)*})$, we have $\big\|\frac{\partial}{\partial\boldsymbol{\theta}_r} q^{(k)}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} - \frac{\partial}{\partial\boldsymbol{\theta}_r} Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}', \boldsymbol{\theta}')|_{\boldsymbol{\theta}=\boldsymbol{\theta}'}\big\| \leq \gamma \cdot \sum_{r=1}^R \big(|w'_r - w_r^{(k)*}| + \|\boldsymbol{\theta}'_r - \boldsymbol{\theta}_r^{(k)*}\|_2\big)$;

(ii) w.p. at least $1-\delta$, for all $\boldsymbol{w}' = \{w'_r\}_{r=1}^R$ and $\boldsymbol{\theta}' = \{\boldsymbol{\theta}'_r\}_{r=1}^R$ with $w'_r \in \mathcal{B}_{r_w^*}(w_r^{(k)*})$ and $\boldsymbol{\theta}'_r \in \mathcal{B}_{r_2^*}(\boldsymbol{\theta}_r^{(k)*})$, we have $\big\|\frac{\partial}{\partial\boldsymbol{\theta}_r}\widehat{Q}^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}', \boldsymbol{\theta}')|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} - \frac{\partial}{\partial\boldsymbol{\theta}_r} Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}', \boldsymbol{\theta}')|_{\boldsymbol{\theta}=\boldsymbol{\theta}'}\big\|_2 \leq \mathcal{E}_1(n, \delta)$;

(iii) w.p. at least $1-\delta$, for $\xi > 0$ and all $\boldsymbol{w}^{(k)\prime} = \{w_r^{(k)\prime}\}_{r=1}^R$, $\boldsymbol{\theta}' = \{\boldsymbol{\theta}_r'\}_{r=1}^R$, and $\{\eta_r^{(k)}\}_{k\in S, r\in[R]}$ with $w_r^{(k)\prime} \in \mathcal{B}_{r_w^*}(w_r^{(k)*})$ and $\boldsymbol{\theta}_r' \in \mathcal{B}_{r_2^*}(\boldsymbol{\theta}_r^{(k)*})$, we have $\big\| \frac{1}{|S|}\sum_{k\in S}\eta_r^{(k)} \cdot \big[\frac{\partial}{\partial \boldsymbol{\theta}_r}\widehat{Q}^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}',\boldsymbol{\theta}')|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} - \frac{\partial}{\partial \boldsymbol{\theta}_r}Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{w}',\boldsymbol{\theta}')|_{\boldsymbol{\theta}=\boldsymbol{\theta}'}\big]\big\|_2 \le \bar{\eta}\mathcal{E}_2(n,|S|,\delta)$.

Remark A.6. Conditions (i) and (ii) have been used by Balakrishnan et al. (2017) in the single-task setting, while condition (iii) is a generalization of (ii) when aggregating the data from multiple tasks. Similar to condition (ii) in Assumption A.3, condition (ii) and condition (iii) are uniform convergence assumptions on the gradient around the true parameter values, which are often needed when analyzing the EM without data splitting (Yan et al., 2017; Cai et al., 2019).

Assumption A.7. Denote $r_{\boldsymbol{\theta}}^* = r_1^* \wedge r_2^*$ and $\tilde{\kappa}_0 = 119\Big(\sqrt{1-\min_{r,k\in S}(\mu_r^{(k)}\eta_r^{(k)})} + \bar{\eta}\gamma R + \kappa R\Big)$. Suppose

$$\max_{k\in S, r\in[R]} |\widehat{w}_r^{(k)[0]} - w_r^{(k)*}| \le r_w^*, \max_{k\in S, r\in[R]} \|\widehat{\boldsymbol{\theta}}_r^{(k)[0]} - \boldsymbol{\theta}_r^{(k)*}\|_2 \le r_{\boldsymbol{\theta}}^*.$$

In addition, $\tilde{\kappa}_0$, κ, r_w^*, $r_{\boldsymbol{\theta}}^*$, and functions $\mathcal{W}$, $\mathcal{E}_1$, and $\mathcal{E}_2$ defined in Assumptions A.1-A.5 satisfy

(i) $\tilde{\kappa}_0 \le \frac{r_{\boldsymbol{\theta}}^*}{18(r_w^*+r_{\boldsymbol{\theta}}^*)}$, $\kappa R \le \frac{9}{1199}\cdot\frac{r_w^*}{r_w^*+r_{\boldsymbol{\theta}}^*}$;

(ii) $\bar{\eta}\cdot\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big) \le \Big[\frac{(1-\tilde{\kappa}_0/119)(1-\tilde{\kappa}_0)}{4320} r_{\boldsymbol{\theta}}^*\Big] \wedge \Big(\frac{1}{3}r_w^*\Big)$;

(iii) $\mathcal{W}\Big(n, \frac{\delta}{3RK}, r_{\boldsymbol{\theta}}^*\Big) \le \frac{(1-\tilde{\kappa}_0/119)(1-\tilde{\kappa}_0)}{2160} r_{\boldsymbol{\theta}}^*$;

(iv) $\bar{\eta}\cdot\mathcal{E}_2\Big(n, |S|, \frac{\delta}{3R}\Big) \le \frac{1-\tilde{\kappa}_0/119}{20} r_{\boldsymbol{\theta}}^*$;

(v) $\eta_r^{(k)} \le 1/L_r^{(k)}$ for all $k\in S$ and $r\in[R]$.

We set the penalty parameters in Algorithm 1 by induction as

$$\lambda^{[0]} = \frac{15}{119}\sqrt{n}(r_w^* + r_{\boldsymbol{\theta}}^*),$$
$$\lambda^{[t]} = \tilde{\kappa}_0\lambda^{[t-1]} + 15\sqrt{n}\Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r_{\boldsymbol{\theta}}^*\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big],$$

Theorem A.8. *Suppose Assumptions A.1-A.7 hold. Then for any contaminated set S^c with $\epsilon = |S^c|/K < 1/3$ and any contaminated distribution $\mathbb{Q}_{S^c}$, with probability at least $1-\delta$, for all $T \ge 1$, FedGrEM satisfies*

$$\max_{k\in S, r\in[R]} \big(|\widehat{w}_r^{(k)[T]} - w_r^{(k)*}| \vee \|\widehat{\boldsymbol{\theta}}_r^{(k)[T]} - \boldsymbol{\theta}_r^{(k)*}\|_2\big) \le \underbrace{20T\tilde{\kappa}_0^{T-1}\times(r_w^*\vee r_{\boldsymbol{\theta}}^*)}_{\text{iterative error}} + \underbrace{\frac{1}{1-\tilde{\kappa}_0/119}\bar{\eta}\mathcal{E}_2\Big(n,|S|,\frac{\delta}{3R}\Big)}_{\text{aggregation rate}}$$
$$+ \underbrace{\frac{1}{1-\tilde{\kappa}_0/119}\mathcal{W}\Big(n,\frac{\delta}{3RK},r_{\boldsymbol{\theta},T}^*\Big)}_{\text{cost of heterogeneous mixing proportions}} + \underbrace{\frac{18}{1-\tilde{\kappa}_0/119}\min\Big\{3h, \frac{6}{1-\tilde{\kappa}_0}\Big[\mathcal{W}\Big(n,\frac{\delta}{3RK},r_{\boldsymbol{\theta}}^*\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n,\frac{\delta}{3RK}\Big)\Big]\Big\}}_{\text{cost of task heterogeneity}}$$
$$+ \underbrace{\frac{30}{(1-\tilde{\kappa}_0)(1-\tilde{\kappa}_0/119)}\epsilon\Big[\mathcal{W}\Big(n,\frac{\delta}{3RK},r_{\boldsymbol{\theta}}^*\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n,\frac{\delta}{3RK}\Big)\Big]}_{\text{cost of outlier tasks}},$$

where $r_{\boldsymbol{\theta},T}^*$ is defined in an iterative fashion by

$$A_t = \Big[9\tilde{\kappa}_0\Big(\frac{\tilde{\kappa}_0}{119}\Big)^{t-1} + \frac{118}{119}(t-1)\tilde{\kappa}_0^{t-1}\Big](r_w^*+r_{\boldsymbol{\theta}}^*) + \frac{1}{1-\tilde{\kappa}_0/119}\bar{\eta}\mathcal{E}_2\Big(n,|S|,\frac{\delta}{3R}\Big)$$
$$+ \frac{18}{1-\tilde{\kappa}_0/119}\min\Big\{3h, \frac{6}{1-\tilde{\kappa}_0}\Big[\mathcal{W}\Big(n,\frac{\delta}{3RK},r_{\boldsymbol{\theta}}^*\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n,\frac{\delta}{3R}\Big)\Big]\Big\}$$
$$+ \frac{30}{(1-\tilde{\kappa}_0)(1-\tilde{\kappa}_0/119)}\epsilon\Big[\mathcal{W}\Big(n,\frac{\delta}{3RK},r_{\boldsymbol{\theta}}^*\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n,\frac{\delta}{3R}\Big)\Big];$$

$$A_t + \frac{18}{1-\tilde{\kappa}_0/119}\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta},t}\Big) = r^*_{\boldsymbol{\theta},t+1},$$

for $t \geq 1$ *with* $r^*_{\boldsymbol{\theta},1} \coloneqq r^*_{\boldsymbol{\theta}}$.

Before jumping into two specific examples, we want to point out that the term $r^*_{\boldsymbol{\theta},T}$ is introduced to address a specific challenge in the analysis of gradient EM mentioned in Section 3.2. In the classical EM theory, the estimates of similar $\{\boldsymbol{\theta}_r^{(k)*}\}_{k\in S}$ and heterogeneous $\{w_r^{(k)*}\}_{k\in S}$ are entangled in the iterations, which will make the heterogeneous scalars $\{w_r^{(k)*}\}_{k\in S}$ contribute a large dimension-dependent error $\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big)$ to the estimation and finally lead to a "slow rate" of convergence for $\{\boldsymbol{\theta}_r^{(k)*}\}_{k\in S, r\in[R]}$. To mitigate this issue, we reduced the radius of the ball within which uniform convergence must hold during iterations. This localization trick finally leads to a "fast rate", effectively replacing $\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big)$ in the slow rate with the current $\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta},T}\Big)$, where $r^*_{\boldsymbol{\theta},T} \ll r^*_{\boldsymbol{\theta}}$. The intuition has been visually interpreted in Figure 2, and more details can be found in the proof of Theorem 3.6.

A.1. Example 1: Gaussian Mixture Models (GMMs)

Assumption A.9. Suppose the following conditions hold:

(i) (Bounded parameters) $w_r^{(k)*} \geq c_w/R$, $\|\boldsymbol{\theta}_r^{(k)*}\|_2 \leq M$ with some $c_w \in (0,1]$ for all $k \in S$ and $r \in [R]$ and $M \geq C > 0$, where C is a constant;

(ii) (Good initialization) $\max_{k\in S, r\in[R]} |\widehat{w}_r^{(k)[0]} - w_r^{(k)*}| \leq C_b \frac{c_w}{R}$, $\max_{k\in S, r\in[R]} \|\widehat{\boldsymbol{\theta}}_r^{(k)[0]} - \boldsymbol{\theta}_r^{(k)*}\|_2 \leq C_b\Delta$, with C_b a small constant;

(iii) (Large signal strength) $\Delta \gtrsim \log(MRc_w^{-1})$;

(iv) (Sample size) $n \gtrsim [R^2M^6d + R^2\log^2(Rc_w^{-1})M^2 + M^2\log(RK/\delta)]C_b^{-2}\Delta^{-2}\bar{\eta}^2$;

(v) (Step size) $1 - \min_{k\in S, r\in[R]}(\eta_r^{(k)} w_r^{(k)*}) <$ a small constant c, and $0 < \eta_r^{(k)} \leq 1/w_r^{(k)*}$ for all $k \in S$ and $r \in [R]$.

Remark A.10. Note that we allow c_w, M, C_b, Δ, T, R, K, and d to change with sample size n.

Proposition A.11. *Under Assumption A.9, GMMs defined in* (4) *satisfies Assumptions A.1-A.7 with*

$$\begin{aligned}
\mu_r^{(k)} = L_r^{(k)} &= w_r^{(k)*},\\
r_1^* = +\infty, r_w^* &= C_b\frac{c_w}{R}, r_2^* = C_b\Delta,\\
\kappa \asymp c_w^{-2}R^2\exp\{-C\Delta^2\}&,\ \gamma \asymp M^2c_w^{-2}R^2\exp\{-C\Delta^2\},\\
\mathcal{W}(n,\delta,\xi) \asymp RM\xi\sqrt{\frac{d}{n}} + [RM^2 + R\log(Rc_w^{-1})]&\sqrt{\frac{1}{n}} + \sqrt{\frac{\log(1/\delta)}{n}},\\
\mathcal{E}_1(n,\delta) \asymp RM^3\sqrt{\frac{d}{n}} + RM\log(Rc_w^{-1})&\sqrt{\frac{1}{n}} + M\sqrt{\frac{\log(1/\delta)}{n}},\\
\mathcal{E}_2(n,|S|,\delta) \asymp RM^3\sqrt{\frac{d}{n|S|}} + [RM^3 + RM\log(Rc_w^{-1})]&\sqrt{\frac{1}{n}}\\
+ M\sqrt{\frac{\log(1/\delta)}{n|S|}}&,
\end{aligned}$$

where $C > 0$ *is some constant.*

By plugging the rates in Propositions 3.8 into Theorem 3.6, we obtain the following result for GMMs.

Corollary A.12. *Set* $\eta_r^{(k)} = (1+C_b)^{-1}(\widehat{w}_r^{(k)[0]})^{-1}$. *Under Assumption 3.7, for the GMMs defined in* (4)*, for any contaminated set* S^c *with* $\epsilon = |S^c|/K \leq 1/3$ *and contaminated distribution* $\mathbb{Q}_{S^c}$*, with probability at least* $1-\delta$*, for all* $T \geq 1$*, FedGrEM satisfies*

$$\max_{k\in S, r\in[R]} \big(|\widehat{w}_r^{(k)[T]} - w_r^{(k)*}| \vee \|\widehat{\boldsymbol{\theta}}_r^{(k)[T]} - \boldsymbol{\theta}_r^{(k)*}\|_2\big)$$

$$
\begin{aligned}
&\lesssim T^2\kappa_0^{T-1}M + R^2c_w^{-1}M^3\sqrt{\frac{d}{n|S|}} \\
&\quad + Rc_w^{-1}M\sqrt{\frac{\log(RK/\delta)}{n}} \\
&\quad + R^2c_w^{-1}M[M^2+\log(Rc_w^{-1})]\sqrt{\frac{1}{n}} \\
&\quad + \min\left\{h, R^2c_w^{-1}M^3\sqrt{\frac{d}{n}}\right\} + \epsilon R^2c_w^{-1}M^3\sqrt{\frac{d}{n}},
\end{aligned}
$$

where $\kappa_0 = 119\sqrt{\frac{2C_b}{1+C_b}} + CM^2c_w^{-2}R^3\exp\{-C'\Delta^2\} + Cc_w^{-2}R^3\exp\{-C'\Delta^2\} + \tilde{\kappa}_0' \in (0,1)$*, and* $\tilde{\kappa}_0'$ *satisfies* $1 > \tilde{\kappa}_0' > CMR\sqrt{\frac{d}{n}}$ *for some* $C > 0$.*.*

Remark A.13. If M and c_w are bounded, when $T \gtrsim \log n$, we will have

$$
\begin{aligned}
&\max_{k\in S, r\in[R]}\left(|\widehat{w}_r^{(k)[T]} - w_r^{(k)*}| \vee \|\widehat{\boldsymbol{\theta}}_r^{(k)[T]} - \boldsymbol{\theta}_r^{(k)*}\|_2\right) \\
&= \widetilde{\mathcal{O}}_{\mathbb{P}}\left(R^2\sqrt{\frac{d}{n|S|}} + R^2\sqrt{\frac{1}{n}} + \min\left\{h, R^2\sqrt{\frac{d}{n}}\right\} + \epsilon R^2\sqrt{\frac{d}{n}}\right).
\end{aligned}
$$

Next, we want to illustrate the choice of step size $\eta_r^{(k)}$. In Corollary A.12, we set $\eta_r^{(k)} = (1+C_b)^{-1}(\widehat{w}_r^{(k)[0]})^{-1}$, then under Assumption A.9, it can be shown that Assumptions A.7.(i) and A.7.(v) hold, $\sqrt{1-\min_{k\in S, r\in[R]}(\eta_r^{(k)}\mu_r^{(k)})} \le \sqrt{\frac{2C_b}{1+C_b}}$, and we can replace $\bar{\eta}$ with CRc_w^{-1} for some constant C in Assumptions A.7.(ii) and A.7.(v).

Second, we have the following upper bound for $r^*_{\boldsymbol{\theta},T}$, which can be plugged into Theorem A.8 with the rates in Proposition A.11 to obtain the upper bound of estimation error in Corollary A.12.

Proposition A.14. *Under Assumption A.9, for the GMMs defined in (4), we have*

$$
\begin{aligned}
r^*_{\boldsymbol{\theta},T} &\lesssim T^2\kappa_0^{T-1}M + \bar{\eta}RM^3\sqrt{\frac{d}{n|S|}} + [(\bar{\eta}M)\vee 1][RM^2 + R\log(Rc_w^{-1})]\sqrt{\frac{1}{n}} + [(\bar{\eta}M)\vee 1]\sqrt{\frac{\log(RK/\delta)}{n}} \\
&\quad + \min\left\{h, \bar{\eta}RM^3\sqrt{\frac{d}{n}}\right\} + \epsilon\bar{\eta}RM^2[(\bar{\eta}M)\vee 1]\sqrt{\frac{d}{n}},
\end{aligned}
$$

where $\kappa_0 = 119\sqrt{\frac{2C_b}{1+C_b}} + CM^2c_w^{-2}R^3\exp\{-C'\Delta^2\} + Cc_w^{-2}R^3\exp\{-C'\Delta^2\} + \tilde{\kappa}_0'$*, and* $\tilde{\kappa}_0'$ *satisfies* $1 > \tilde{\kappa}_0' > CMR\sqrt{\frac{d}{n}}$ *for some* $C > 0$.

A.2. Example 2: Mixture of Regressions (MoRs)

Assumption A.15. Suppose the same conditions in Assumption A.9 hold by replacing (iii) with:

(iii) (Strong signal strength) $\Delta \gtrsim R^3c_w^{-1} + R^2c_w^{-1}(\log\Delta)^{3/2}$;

Similar to our previous comments in Remark A.10 for GMMs, we also allow c_w, M, C_b, Δ, T, R, K, and d to change with sample size n in MoRs.

Proposition A.16. *Under Assumption A.15, the MoRs defined in (5) satisfies Assumptions A.1-A.7 with* r_1^*, r_2^*, $\mathcal{W}(n,\delta,\xi)$, $\mathcal{E}_1(n,\delta)$, $\mathcal{E}_2(n,|S|,\delta)$ *the same as in Proposition A.11, and*

$$
\begin{aligned}
\mu_r^{(k)} &= w_r^{(k)*} - CR\frac{\sqrt{\log\Delta}}{\Delta}, \\
L_r^{(k)} &= w_r^{(k)*} + CR\frac{\sqrt{\log\Delta}}{\Delta},
\end{aligned}
$$

$$\kappa \asymp c_w^{-1} R \frac{\sqrt{\log \Delta}}{\Delta} + c_w^{-1} R C_b + R^2 c_w^{-1} \frac{1}{\Delta},$$
$$\gamma \asymp c_w^{-1} R \frac{(\log \Delta)^{3/2}}{\Delta} + c_w^{-1} R C_b + R^2 c_w^{-1} \frac{1}{\Delta},$$

where $C > 0$ is some constant.

Corollary A.17. *Set $\eta_r^{(k)} = (1 + C_b)^{-1} (\widehat{w}_r^{(k)[0]})^{-1}$. Under Assumption A.15, for the MoRs defined in (5), for any contaminated set S^c with $\epsilon = |S^c|/K \leq 1/3$ and contaminated distribution $\mathbb{Q}_{S^c}$, with probability at least $1 - \delta$, for all $T \geq 1$, FedGrEM satisfies*

$$\begin{aligned}
&\max_{k \in S, r \in [R]} \left(|\widehat{w}_r^{(k)[T]} - w_r^{(k)*}| \vee \|\widehat{\boldsymbol{\theta}}_r^{(k)[T]} - \boldsymbol{\theta}_r^{(k)*}\|_2 \right) \\
&\lesssim T^2 \kappa_0^{T-1} M + R^2 c_w^{-1} M^3 \sqrt{\frac{d}{n|S|}} \\
&\quad + R c_w^{-1} M \sqrt{\frac{\log(RK/\delta)}{n}} \\
&\quad + R^2 c_w^{-1} M [M^2 + \log(R c_w^{-1})] \sqrt{\frac{1}{n}} \\
&\quad + \min \left\{ h, R^2 c_w^{-1} M^3 \sqrt{\frac{d}{n}} \right\} + \epsilon R^2 c_w^{-1} M^3 \sqrt{\frac{d}{n}},
\end{aligned}$$

where $\kappa_0 = 119 \sqrt{\frac{3C_b}{1+2C_b} + CR^3 c_w^{-2} \frac{(\log \Delta)^{3/2}}{\Delta}} + CR^3 c_w^{-2} C_b + CR^4 c_w^{-2} \frac{1}{\Delta} + \tilde{\kappa}_0' \in (0, 1)$, and $\tilde{\kappa}_0'$ satisfies $1 > \tilde{\kappa}_0' > CMR\sqrt{\frac{d}{n}}$ for some $C > 0$.

Remark A.18. If M and c_w are bounded, when $T \gtrsim \log n$, we will have the same rate for MoRs as in Remark A.13. We can also compare the rate of FedGrEM and the local single-task rates as in GMMs, which we do not repeat here.

In Corollary A.17, we set $\eta_r^{(k)} = (1 + 2C_b)^{-1} (\widehat{w}_r^{(k)[0]})^{-1}$ and let $C_b \gtrsim R c_w^{-1} \frac{\sqrt{\log \Delta}}{\Delta}$, then under Assumption A.15, it can be shown that Assumptions A.7.(i) and (v) hold, $\sqrt{1 - \min_{k \in S, r \in [R]} (\eta_r^{(k)} \mu_r^{(k)})} \leq \sqrt{\frac{3C_b}{1+2C_b} + C \frac{\sqrt{\log \Delta}}{\Delta}}$, and we can replace $\bar{\eta}$ with CRc_w^{-1} for some constant C in Assumptions A.7.(ii) and A.7.(iv).

In addition, we have the following upper bound for $r^*_{\boldsymbol{\theta}, T}$, which can be plugged into Theorem A.8 with the rates in Proposition A.16 to obtain the upper bound of estimation error in Corollary A.17.

Proposition A.19. *Under Assumption A.15, for the MoRs defined in (5), we have*

$$\begin{aligned}
r^*_{\boldsymbol{\theta}, T} &\lesssim T^2 \kappa_0^{T-1} M + \bar{\eta} R M^3 \sqrt{\frac{d}{n|S|}} + [(\bar{\eta} M) \vee 1][RM^2 + R \log(R c_w^{-1})] \sqrt{\frac{1}{n}} + [(\bar{\eta} M) \vee 1] \sqrt{\frac{\log(RK/\delta)}{n}} \\
&\quad + \min \left\{ h, \bar{\eta} R M^3 \sqrt{\frac{d}{n}} \right\} + \epsilon \bar{\eta} R M^2 [(\bar{\eta} M) \vee 1] \sqrt{\frac{d}{n}},
\end{aligned}$$

where $\kappa_0 = 119 \sqrt{\frac{3C_b}{1+2C_b} + CR^3 c_w^{-2} \frac{(\log \Delta)^{3/2}}{\Delta}} + CR^3 c_w^{-2} C_b + CR^4 c_w^{-2} \frac{1}{\Delta} + \tilde{\kappa}_0'$, and $\tilde{\kappa}_0'$ satisfies $1 > \tilde{\kappa}_0' > CMR\sqrt{\frac{d}{n}}$ for some $C > 0$.

B. Additional Details of Numerical Results

In this section, we provide more details of the numerical studies in Section 4. All experiments are implemented in R, where EM is executed using the `mclust` package for GMMs and the `mixreg` package for MoRs. GrEM, FedEM (Marfoq et al., 2021), TGMM (Wang et al., 2021), FedGMM (Wu et al., 2023), and FedGrEM are initialized with the estimates from local EM. As the empirical results in Wang et al. (2021) suggested, we set the tuning parameter $\lambda = 0.4$ in TGMM, which controls how much information to borrow from the other tasks. TGMM was originally designed for transfer learning in Wang et al. (2021) and we ran it on each task with all the other tasks as sources. For FedGMM, we set the number of mixtures $M_1 = 3$.

B.1. Additional Details of Simulations

In both examples, we generate centers $\bar{\boldsymbol{\theta}}_r$ of parameters $\{\boldsymbol{\theta}_r^{(k)*}\}_{k\in S}$ as

$$\begin{aligned}
\bar{\boldsymbol{\theta}}_1 &= (1,0,3,-1,1,-1,0,1,1,-1)^T,\\
\bar{\boldsymbol{\theta}}_2 &= (0,1,-1,-3,2,-1,2,-1,1,-)^T,\\
\bar{\boldsymbol{\theta}}_3 &= (-3,-1,2,-1,2,-1,1,-3,-1,-2)^T,\\
\bar{\boldsymbol{\theta}}_4 &= (1,-2,0,-1,-2,2,1,3,1,-1)^T,\\
\bar{\boldsymbol{\theta}}_5 &= (3,1,2,-1,-2,1,2,-1,-1,2)^T.
\end{aligned}$$

And we generate the parameter $\boldsymbol{\theta}_r^{(k)*}$ by

$$\boldsymbol{\theta}_r^{(k)*} = \bar{\boldsymbol{\theta}}_r + h \times \frac{\boldsymbol{z}_r^{(k)}}{\|\boldsymbol{z}_r^{(k)}\|_2}, \quad r \in [R],$$

where $\boldsymbol{z}_r \sim N(\mathbf{0}_p, \boldsymbol{I}_{p\times p})$ and they are independent for different k and r. In the GMM simulation, the observations of the outlier task are generated i.i.d. from $N(2\cdot\mathbf{1}_p, 3\boldsymbol{I}_{p\times p})$. In the MoR simulation, all $\boldsymbol{x}_i^{(k)}$'s are generated from $N(\mathbf{0}_p, \boldsymbol{I}_{p\times p})$, and the responses $\{y_i^{(k)}\}_{i=1}^n$ of the outlier task are generated i.i.d. from the regression model $y_i^{(k)} = (\boldsymbol{x}_i^{(k)})^T\boldsymbol{\theta}_r^{(k)*} + \epsilon_i^{(k)}$ with $\boldsymbol{\theta}_r^{(k)*} = 3\cdot\mathbf{1}_p$ and $\epsilon_i^{(k)} \sim N(0,1)$. In Figure 3, the SNR is defined to be $\min_{k\in S}\max_{r\neq r'\in[R]}\|\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_{r'}^{(k)*}\|_2$.

In FedGrEM, we set the number of iterations $T = 1000$, and the penalty parameters $\lambda^{[t]}$ are updated in iterations as follows:

$$\lambda^{[0]} = 1, \quad \lambda^{[t]} = \kappa_0\lambda^{[t-1]} + C\sqrt{p+\log K},$$

where $\kappa = 0.1$ and $C = 2$.

The optimization in the central update is solved by an alternating optimization procedure. We first reparameterize the problem by $\boldsymbol{\nu}^{(k)} = \bar{\boldsymbol{\nu}} + \boldsymbol{\Delta}^{(k)}$, and we solve

$$\underset{\{\boldsymbol{\Delta}^{(k)}\}_{k=1}^K\subseteq\mathbb{R}^d,\bar{\boldsymbol{\nu}}\in\mathbb{R}^d}{\arg\min}\left\{\sum_{k=1}^K\left(\frac{n}{2}\|\bar{\boldsymbol{\nu}} + \boldsymbol{\Delta}^{(k)} - \widetilde{\boldsymbol{\theta}}_r^{(k)[t]}\|_2^2 + \sqrt{n}\lambda^{[t]}\cdot\|\boldsymbol{\Delta}^{(k)}\|_2\right)\right\}, \tag{6}$$

and assign $\widehat{\boldsymbol{\theta}}_r^{(k)[t]} = \bar{\boldsymbol{\nu}} + \boldsymbol{\Delta}^{(k)}$. We solve (6) by the alternating optimization as follows:

(i) Set $\boldsymbol{\Delta}^{(k)} = \mathbf{0}$ for all $k\in[K]$;

(ii) Fix $\{\boldsymbol{\Delta}^{(k)}\}_{k=1}^K$, and update $\bar{\boldsymbol{\nu}} = \frac{1}{K}\sum_{k=1}^K(\widetilde{\boldsymbol{\theta}}_r^{(k)[t]} - \boldsymbol{\Delta}^{(k)})$;

(iii) Fix $\bar{\boldsymbol{\nu}}$, and update $\boldsymbol{\Delta}^{(k)}$ as

$$\begin{aligned}
\boldsymbol{\Delta}^{(k)} &= \arg\min_{\boldsymbol{\Delta}}\left\{\frac{n}{2}\|\bar{\boldsymbol{\nu}} + \boldsymbol{\Delta} - \widetilde{\boldsymbol{\theta}}_r^{(k)[t]}\|_2^2 + \sqrt{n}\lambda^{[t]}\cdot\|\boldsymbol{\Delta}\|_2\right\}\\
&= \begin{cases}\left(1 - \frac{\lambda^{[t]}/\sqrt{n}}{\|\widetilde{\boldsymbol{\theta}}_r^{(k)[t]} - \bar{\boldsymbol{\nu}}\|_2}\right)(\widetilde{\boldsymbol{\theta}}_r^{(k)[t]} - \bar{\boldsymbol{\nu}}), & \text{if } \|\widetilde{\boldsymbol{\theta}}_r^{(k)[t]} - \bar{\boldsymbol{\nu}}\|_2 \geq \frac{\lambda^{[t]}}{\sqrt{n}},\\ \mathbf{0}, & \text{else.}\end{cases}
\end{aligned}$$

We iterate (ii) and (iii) for a few times until convergence.

We also run our simulation example with different C_η values ranging from 0.55 to 1.35. The results are summarized in the following Tables 4 and 5. It can be seen that the estimation error of mixture proportions $w_r^{(k)*}$'s is quite stable with the choice of C_η, and the estimation error of parameters $\boldsymbol{\theta}_r^{(k)*}$ first decreases then becomes stable when C_η changes from 0.55 to 1.35. In fact, according to our observation in the experiments, if we continue increasing the C_η, the algorithm would fail to converge and output useless estimates. Overall the performance is robust to the choice of C_η within this range $[0.55, 1.35]$, and we can further tune it in practice, although the theoretically-guided choice already leads to a decent performance.

C_η / h	$h=0$	$h=0.25$	$h=0.5$	$h=0.75$	$h=1$	$h=1.25$	$h=1.5$
0.55	0.072 (0.017)	0.071 (0.015)	0.068 (0.015)	0.071 (0.016)	0.072 (0.017)	0.071 (0.017)	0.071 (0.016)
0.65	0.072 (0.017)	0.071 (0.015)	0.068 (0.015)	0.071 (0.016)	0.072 (0.017)	0.071 (0.017)	0.071 (0.016)
0.75	0.072 (0.017)	0.071 (0.018)	0.068 (0.015)	0.071 (0.016)	0.072 (0.017)	0.071 (0.019)	0.071 (0.015)
0.85	0.072 (0.017)	0.071 (0.015)	0.068 (0.015)	0.071 (0.016)	0.072 (0.019)	0.071 (0.018)	0.071 (0.015)
0.95	0.072 (0.017)	0.071 (0.016)	0.068 (0.015)	0.071 (0.016)	0.072 (0.02)	0.071 (0.018)	0.071 (0.015)
1.05	0.072 (0.017)	0.071 (0.016)	0.068 (0.015)	0.071 (0.016)	0.071 (0.017)	0.071 (0.017)	0.071 (0.015)
1.15	0.072 (0.017)	0.071 (0.016)	0.069 (0.017)	0.071 (0.016)	0.071 (0.017)	0.071 (0.017)	0.071 (0.016)
1.25	0.072 (0.017)	0.071 (0.016)	0.068 (0.015)	0.071 (0.017)	0.071 (0.017)	0.072 (0.019)	0.071 (0.017)
1.35	0.073 (0.021)	0.071 (0.016)	0.069 (0.017)	0.071 (0.016)	0.071 (0.017)	0.072 (0.018)	0.071 (0.016)

Table 4. Average of the maximum estimation error of $\{w_r^{(k)*}\}_{k\in S,r}$ (standard deviations) in $\log_e$ scale in the GMM simulation, with different constants C_η in learning rate $\eta_r^{(k)} = C_\eta/\hat{w}_r^{(k)[0]}$ and heterogeneity parameter h.

C_η/ h	$h=0$	$h=0.25$	$h=0.5$	$h=0.75$	$h=1$	$h=1.25$	$h=1.5$	$h=1.75$
0.55	1.36 (0.62)	1.47 (0.63)	1.62 (0.63)	1.71 (0.56)	1.89 (0.70)	2.07 (0.72)	2.23 (0.84)	2.44 (0.99)
0.65	1.28 (0.59)	1.38 (0.61)	1.52 (0.62)	1.62 (0.56)	1.79 (0.71)	1.99 (0.75)	2.14 (0.90)	2.33 (0.99)
0.75	1.22 (0.55)	1.34 (0.63)	1.44 (0.59)	1.54 (0.55)	1.71 (0.69)	1.93 (0.78)	2.02 (0.81)	2.25 (1.01)
0.85	1.17 (0.51)	1.30 (0.68)	1.38 (0.57)	1.49 (0.54)	1.66 (0.68)	1.88 (0.77)	1.95 (0.79)	2.22 (1.05)
0.95	1.13 (0.47)	1.30 (0.80)	1.34 (0.56)	1.45 (0.52)	1.60 (0.61)	1.82 (0.74)	1.90 (0.77)	2.14 (1.05)
1.05	1.12 (0.44)	1.27 (0.75)	1.30 (0.51)	1.42 (0.48)	1.54 (0.52)	1.79 (0.73)	1.85 (0.74)	2.07 (0.97)
1.15	1.10 (0.41)	1.25 (0.77)	1.32 (0.66)	1.39 (0.46)	1.52 (0.50)	1.79 (0.79)	1.82 (0.72)	2.04 (0.98)
1.25	1.09 (0.38)	1.23 (0.76)	1.34 (0.71)	1.40 (0.59)	1.50 (0.48)	1.78 (0.81)	1.81 (0.75)	1.99 (0.96)
1.35	1.21 (1.42)	1.24 (0.81)	1.31 (0.57)	1.36 (0.41)	1.49 (0.48)	1.81 (0.91)	1.79 (0.74)	1.99 (1.00)

Table 5. Average of the maximum estimation error of $\{\boldsymbol{\theta}_r^{(k)*}\}_{k\in S,r}$ (standard deviations) in $\log_e$ scale in the GMM simulation, with different constants C_η in learning rate $\eta_r^{(k)} = C_\eta/\hat{w}_r^{(k)[0]}$ and heterogeneity parameter h.

B.2. Additional Details of Real Studies

Due to the high dimensionality of the MNIST and Fashion-MNIST datasets, we first applied tSNE (Hinton & Roweis, 2002; Van der Maaten & Hinton, 2008) to reduce the dimension to 10 then performed all methods on the transformed datasets. In each replication, 80% data for each task is used as training data and the remaining 20% is used as test data to calculate the mis-clustering error. We also contaminate different proportions of tasks to showcase the robustness of FedGrEM against adversarial attacks.

We record the average computational time for each method and the results are summarized below. The experiments were conducted with Dual Intel Xeon Gold 6226R processors (2.9 GHz) and a single core. The results are summarized in Tables 6, 7, and 8. From the performance and computational time, we can see that FedGrEM can be adapted to a federated learning environment with hundreds of nodes, even without parallel computing. By parallelizing the local update steps in each iteration, we can further speed things up.

The reason why gradient EM-based methods are slower than full EM-based methods is that the M-step of full EM has an explicit expression and does not require calculating the matrix inverse since all covariances are identities. Full EM-based methods do not apply to the problems with non-explicit M-steps and would be more time-consuming than gradient EM-based methods if the explicit M-step expression is very complicated.

C. Label Permutation

Denote the family of permutation functions on $[R]$ as $\mathcal{P}^R$.

ϵ / Method	Local-EM	Local-GrEM	FedEM	FedGrEM	TGMM	Pooled-EM	Pooled-GrEM
0%	26.01 (1.22)	36.55 (1.58)	29.86 (1.48)	47.07 (2.38)	86.79 (3.69)	18.85 (2.10)	30.24 (2.97)
6.8%	25.92 (1.04)	36.19 (1.41)	29.74 (1.27)	46.76 (1.92)	85.07 (3.57)	18.28 (1.66)	29.71 (2.23)
13.6%	25.94 (1.46)	36.07 (1.46)	29.78 (1.54)	46.90 (2.31)	84.85 (3.73)	17.81 (1.47)	29.39 (2.00)
20.5%	25.62 (1.05)	35.80 (1.48)	29.49 (1.21)	46.49 (2.12)	83.84 (3.65)	17.36 (1.35)	28.56 (2.14)

Table 6. Average computational time (standard deviations) in seconds for Pen-Based Recognition of Handwritten Digits dataset.

ϵ / Method	Local-EM	Local-GrEM	FedEM	FedGrEM	TGMM	Pooled-EM	Pooled-GrEM
0%	83.99 (4.07)	113.81 (6.59)	93.00 (4.85)	134.69 (6.41)	299.56 (19.18)	94.87 (12.26)	166.12 (27.84)
8%	80.63 (3.92)	110.90 (5.31)	89.36 (4.57)	132.49 (6.63)	291.81 (17.09)	93.51 (13.56)	159.38 (28.76)
16%	80.06 (4.08)	110.27 (5.51)	88.52 (4.82)	133.32 (7.12)	290.71 (17.95)	89.39 (14.91)	151.79 (27.51)
24%	79.32 (3.82)	109.31 (5.20)	90.16 (4.55)	133.74 (6.47)	287.80 (15.12)	86.54 (15.00)	145.39 (25.01)

Table 7. Average computational time (standard deviations) in seconds for MNIST dataset.

ϵ / Method	Local-EM	Local-GrEM	FedEM	FedGrEM	TGMM	Pooled-EM	Pooled-GrEM
0%	83.89 (3.58)	113.21 (4.94)	92.82 (4.10)	134.92 (5.98)	297.01 (13.73)	96.46 (14.15)	169.07 (32.25)
8%	81.15 (3.54)	111.78 (4.89)	89.41 (4.08)	133.29 (6.01)	293.78 (14.59)	94.52 (15.28)	162.50 (30.77)
16%	80.70 (3.56)	111.28 (4.78)	89.80 (4.16)	134.38 (6.17)	293.31 (15.13)	90.28 (15.88)	154.20 (29.21)
24%	80.39 (3.46)	110.75 (4.59)	90.91 (4.17)	135.52 (5.95)	292.91 (13.93)	87.20 (15.59)	147.38 (27.98)

Table 8. Average computational time (standard deviations) in seconds for Fashion-MNIST dataset.

C.1. Label Permutation in Initialization

One challenge that hinders the practical application of FedGrEM (and other federated EM algorithms) and our theoretical framework relates to the initialization condition outlined in Assumption 3.5 (more rigorously, Assumption A.7). As is common in many unsupervised learning problems, the parameters are estimated up to a label permutation. Our results still hold if the initialization condition holds up to a permutation, i.e. there exists a permutation $\pi \in \mathcal{P}^R$ such that $\max_{k\in S, r\in[R]} |\widehat{w}^{(k)[0]}_{\pi(r)} - w^{(k)*}_r| \leq r^*_w$, $\max_{k\in S, r\in[R]} \|\widehat{\boldsymbol{\theta}}^{(k)[0]}_{\pi(r)} - \boldsymbol{\theta}^{(k)*}_r\|_2 \leq r^*_{\boldsymbol{\theta}}$, however, it necessitates the presence of a *shared* permutation $\pi \in \mathcal{P}^R$ among the tasks in S. However, in all clustering methods typically employed for initialization in individual tasks, the estimations are inherently invariant to permutations, and different tasks may yield distinct permutations. To the best of our knowledge, there has been limited discussion in the existing literature on unsupervised FDL regarding this issue, with the exception of Tian et al. (2022). We generalized the solutions proposed in Tian et al. (2022) to address the permutation issue in initialization. By ensuring that different tasks in S share the same permutation, we make FedGrEM and our accompanying theory applicable in practice.

C.2. Alignment Algorithms

We define the following score function of the permutations $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K \in (\mathcal{P}^R)^{\otimes K}$:

$$\text{score}(\boldsymbol{\pi}, K) = \sum_{r=1}^{R} \sum_{k\neq k'\in[K]} \|\widehat{\boldsymbol{\theta}}^{(k)[0]}_{\pi_k(r)} - \widehat{\boldsymbol{\theta}}^{(k)[0]}_{\pi_{k'}(r)}\|_2.$$

Intuitively, the score is smaller if the permutations $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K \in (\mathcal{P}^R)^{\otimes K}$ are more aligned, serving as the basis for adjusting the permutations of each task. We also define the best permutation for task $k \in S$ as

$$\pi^*_k = \mathop{\arg\min}_{\pi_k\in\mathcal{P}^R} \sum_{r=1}^{R} \|\widehat{\boldsymbol{\theta}}^{(k)[0]}_{\pi_k(r)} - \boldsymbol{\theta}^{(k)*}_r\|_2.$$

Next, we introduce an exhaustive search algorithm for permutation alignment, which seeks the permutations minimizing the score.

Permutation Alignment Algorithm 1 (Exhaustive search): Let $\widehat{\boldsymbol{\pi}} = \underset{\boldsymbol{\pi}\in(\mathcal{P}^R)^{\otimes K}}{\arg\min}\ \text{score}(\boldsymbol{\pi}, K)$.

Under the assumption detailed below, we demonstrate that the output permutations in tasks of S from the exhaustive search algorithm are well-aligned.

Assumption C.1. The following conditions hold:

(i) $\Delta > \frac{2+2\epsilon}{1-\epsilon}h + \frac{4+4\epsilon}{1-\epsilon}\max_{k\in S}\min_{\pi_k\in\mathcal{P}^R}\max_{r\in[R]}\|\widehat{\boldsymbol{\theta}}^{(k)[0]}_{\pi_k(r)} - \boldsymbol{\theta}^{(k)*}_r\|_2$;

(ii) $\epsilon < 1/2$.

Theorem C.2. *Under Assumption C.1, for Alignment Algorithm 1 (Exhaustive search), there exists a permutation $\iota \in \mathcal{P}^R$ such that $\widehat{\pi}_k = \iota \circ \pi^*_k$ for all $k \in S$.*

The computational cost of the exhaustive search algorithm is $\mathcal{O}((R!)^K \cdot K^2R)$ as it explores all the possible permutations on $[R]$ for all K tasks and takes $\mathcal{O}(K^2R)$ time to calculate the score for each permutation. We introduce the following stepwise search algorithm which can reduce the computational cost to $\mathcal{O}(R!K \cdot K^2R)$.

Permutation Alignment Algorithm 2 (Stepwise search): For $k = 1 : K$, with $\{\widehat{\pi}_{k'}\}_{k'=1}^{k-1}$ fixed, set $\widehat{\pi}_k = \underset{\pi_k\in\mathcal{P}^R}{\arg\min}$ $\text{score}(\{\widehat{\pi}_{k'}\}_{k'=1}^{k-1} \cup \pi_k, k)$. Finally, let $\widehat{\boldsymbol{\pi}} = \{\widehat{\pi}_k\}_{k=1}^K$.

Under the following assumption, we show that the output permutations in tasks of S from the stepwise search algorithm are well-aligned.

Assumption C.3. Suppose there are no outlier tasks in the first K_0 tasks and

(i) $\Delta > 2\frac{K_0+K\epsilon}{K_0-K\epsilon}h + 6\frac{K_0+K\epsilon}{K_0-K\epsilon}\max_{k\in S}\min_{\pi_k\in\mathcal{P}^R}\max_{r\in[R]}\|\widehat{\boldsymbol{\theta}}^{(k)[0]}_{\pi_k(r)} - \boldsymbol{\theta}^{(k)*}_r\|_2$;

(ii) $K_0 > K\epsilon$;

(iii) $\epsilon < 1/2$.

Theorem C.4. *Under Assumption C.3, there exists a permutation $\iota \in \mathcal{P}^R$ such that $\widehat{\pi}_k = \iota \circ \pi^*_k$ for all $k \in S$.*

The limitation of the stepwise search algorithm is that it requires the first K_0 tasks to be non-outlier tasks. One potential solution is running the algorithm multiple times with a random shuffling of tasks and then picking the permutations based on recurring patterns observed across multiple experiment runs. We leave a full investigation of this random algorithm for future study.

D. Additional Discussions

We want to comment a bit more about our FedGrEM algorithm.

- When mixture proportions $\{w^{(k)*}_r\}_{k=1}^K$ are also similar across tasks, we can apply the same aggregation by regularization in the central update to further improve the performance of FedGrEM. The same analysis tools can be applied to obtain stronger results.
- When there exist various computational capabilities across different tasks, we can replace the current local gradient descent with full local data by local stochastic gradient descent (SGD) with small batches of local data. This would decrease the computational cost for each task. Moreover, instead of including all users in each iteration round, we can randomly sample a few users in each round to update their local estimates and run the central update with only these active users. These two approaches could further decrease the computational cost. And the size of SGD batches and the proportion of active users in each round could depend on the computational and communicational budget for each user, which could differ from user to user.

- The current communicational cost for FedGrEM or FedEM (Marfoq et al., 2021) is already low because they only pass the gradients across tasks in each iteration. The sampling of active users could help further decrease the communication cost.
- Our current analysis and the method FedGrEM can be extended to a fully decentralized version with the same idea used by Algorithm 4 in (Marfoq et al., 2021). In each iteration, instead of sending all local estimates to the central server, each node can send their local estimates only to their neighbors (the nodes that are closest to them in the geometric sense or the communication cost sense) and perform the aggregation locally with the estimates received from the neighbors. We can use the current theoretical framework to analyze the estimation error of this fully decentralized algorithm and derive similar results.
- We can consider other types of attacks and contaminations. For example, instead of the corruption of the entire dataset from some users, we can assume partial observations are contaminated. In this case, we can create a robust version of local estimates by using truncated gradients, similar to the gradient clipping used in differential privacy (Varshney et al., 2022). Aggregating these robust local estimates in the central update can make the whole procedure robust to both observation-level and user-level attacks.
- We assume that the central update can be exactly solved and we use alternating optimization to solve it in practice. We did not directly analyze the alternating optimization itself because the EM procedure is already very complicated to study due to its iterative nature. We believe that this is an important question. One nice characteristic about our central update in FedGrEM (Algorithm 1) is that it is a convex problem. There exist convergence results about alternating optimization, such as (Li et al., 2019; Guminov et al., 2021; Tupitsa et al., 2021), which can be helpful. We can also consider other optimization methods such as proximal gradient descent (Polson et al., 2015). We will work on this problem in the future.

We also want to point out the possibility of generalizing our theory and method to high-dimensional or non-i.i.d. data.

- For i.i.d. high-dimensional data, which is very common in healthcare and biomedical studies, we can add an additional regularization term (e.g., ℓ_1-penalty) for the global estimator $\bar{\boldsymbol{\nu}}$ in the central update of FedGrEM (Algorithm 1). Another solution is to truncate the current local estimator (one-step gradient descent) by a coordinate-wise soft-thresholding function and keep the central update as it is. The challenges of analyzing such a problem are mainly aligned with the challenges in other high-dimensional problems. For example, the strong concavity would fail for the empirical surrogate risk function $\widehat{Q}^{(k)}$. Instead, as one of the standard techniques used in the high-dimensional analysis, we need to first prove that the estimators belong to a small subset (usually a cone in $\mathbb{R}^d$), and within this subset, the so-called restricted strong convexity or concavity (RSC) holds. In the context of federated EM algorithms, the analysis would be more complicated due to the nature of the iterative procedure. Some analysis has been done for the single-task EM in (Cai et al., 2019), and some techniques therein might be helpful.
- For non-i.i.d. data, for example, the data of social networks, we can first apply some embedding methods to transform the original data into a standard unsupervised learning problem. For instance, for adjacency or Laplacian matrices in social networks, we can compute its spectral embedding and use the embedding as the input for the federated EM algorithms. The challenge here is that the embedded data is not independent. But in many situations, the dependence within the embedded data can be shown to be somewhat weak, which is sufficient to derive some theoretical guarantees (Rohe et al., 2011; Tang & Priebe, 2018; Abbe et al., 2020).

E. Proofs

E.1. Proof of Theorem A.8

Let us first fix an $S \subseteq [K]$ and introduce the following key lemma.

Lemma E.1. *(Duan & Wang, 2022) The following results hold:*

(i) If $\lambda^{[t]} \geq \frac{5\sqrt{n}\max_{k\in S}\|\widetilde{\boldsymbol{\theta}}_r^{(k)[t]}-\boldsymbol{\theta}_r^{(k)*}\|_2}{1-2\epsilon}$, *then*

$$\max_{k\in S}\|\widehat{\boldsymbol{\theta}}_r^{(k)[t]}-\boldsymbol{\theta}_r^{(k)*}\|_2 \leq \frac{1}{|S|}\left\|\sum_{k\in S}(\widetilde{\boldsymbol{\theta}}_r^{(k)[t]}-\boldsymbol{\theta}_r^{(k)*})\right\|_2 + \frac{6}{1-2\epsilon}\min\left\{3h, \frac{2\lambda^{[t]}}{5\sqrt{n}}\right\} + \frac{2\lambda^{[t]}}{\sqrt{n}}\epsilon.$$

(ii) If we further have $\lambda^{[t]} \geq \frac{15\sqrt{n}}{1-2\epsilon}h$*, then*

$$\max_{k\in S}\|\widehat{\boldsymbol{\theta}}_r^{(k)[t]} - \boldsymbol{\theta}_r^{(k)*}\|_2 \leq \frac{1}{|S|}\left\|\sum_{k\in S}(\widetilde{\boldsymbol{\theta}}_r^{(k)[t]} - \boldsymbol{\theta}_r^{(k)*})\right\|_2 + 2h + \frac{2\lambda^{[t]}}{\sqrt{n}}\epsilon.$$

Define a random event $\mathcal{V}$ which is the intersection of the following three events:

- The event in Assumption A.3.(ii) holds for all $k \in S$ with failure probability $\frac{\delta}{3RK}$;
- The event in Assumption A.3.(ii) holds for all $k \in S$ with failure probability $\frac{\delta}{3RK}$;
- The event in Assumption A.5.(iii) holds with failure probability $\frac{\delta}{3}$.

Then by the union bound, $\mathbb{P}(\mathcal{V}) \geq 1-\delta$. In the following analysis, we condition on $\mathcal{V}$. Hence all arguments hold with probability at least $1-\delta$.

(I) Part 1: Iteration round $t = 1$.

By Lemma E.1, when $\lambda^{[t]} \geq \frac{5\sqrt{n}}{1-2\epsilon}\max_{k\in S}\max_{r\in[R]}\|\widetilde{\boldsymbol{\theta}}_r^{(k)[1]} - \boldsymbol{\theta}_r^{(k)*}\|_2$:

$$\max_{k\in S}\|\widehat{\boldsymbol{\theta}}_r^{(k)[1]} - \boldsymbol{\theta}_r^{(k)*}\|_2 \leq \frac{1}{|S|}\left\|\sum_{k\in S}(\widetilde{\boldsymbol{\theta}}_r^{(k)[1]} - \boldsymbol{\theta}_r^{(k)*})\right\|_2 + \frac{6}{1-2\epsilon}\min\left\{3h, \frac{2\lambda^{[1]}}{5\sqrt{n}}\right\} + \frac{2\lambda^{[1]}}{\sqrt{n}}\epsilon.$$

Note that

$$\begin{aligned}\frac{1}{|S|}\left\|\sum_{k\in S}(\widetilde{\boldsymbol{\theta}}_r^{(k)[1]} - \boldsymbol{\theta}_r^{(k)*})\right\|_2 \leq& \underbrace{\frac{1}{|S|}\left\|\sum_{k\in S}\left[\widehat{\boldsymbol{\theta}}_r^{(k)[0]} - \boldsymbol{\theta}_r^{(k)*} + \eta_r^{(k)}\frac{\partial}{\partial\boldsymbol{\theta}_r}Q^{(k)}(\widehat{\boldsymbol{\theta}}^{(k)[0]}|\widehat{w}^{(k)[0]}, \widehat{\boldsymbol{\theta}}^{(k)[0]})\right]\right\|_2}_{[1]} \\ &+ \underbrace{\frac{1}{|S|}\left\|\sum_{k\in S}\eta_r^{(k)}\left[\frac{\partial}{\partial\boldsymbol{\theta}_r}Q^{(k)}(\widehat{\boldsymbol{\theta}}^{(k)[0]}|\widehat{w}^{(k)[0]}, \widehat{\boldsymbol{\theta}}^{(k)[0]}) - \frac{\partial}{\partial\boldsymbol{\theta}_r}\widehat{Q}^{(k)}(\widehat{\boldsymbol{\theta}}^{(k)[0]}|\widehat{w}^{(k)[0]}, \widehat{\boldsymbol{\theta}}^{(k)[0]})\right]\right\|_2}_{[2]}.\end{aligned}$$

For [1], we have

$$\begin{aligned}[1] \leq& \frac{1}{|S|}\left\|\sum_{k\in S}\left[\widehat{\boldsymbol{\theta}}_r^{(k)[0]} - \boldsymbol{\theta}_r^{(k)*} + \eta_r^{(k)}\frac{\partial}{\partial\boldsymbol{\theta}_r}q^{(k)}(\widehat{\boldsymbol{\theta}}^{(k)[0]})\right]\right\|_2 \\ &+ \frac{1}{|S|}\left\|\sum_{k\in S}\eta_r^{(k)}\left[\frac{\partial}{\partial\boldsymbol{\theta}_r}Q^{(k)}(\widehat{\boldsymbol{\theta}}^{(k)[0]}|\widehat{w}^{(k)[0]}, \widehat{\boldsymbol{\theta}}^{(k)[0]}) - \frac{\partial}{\partial\boldsymbol{\theta}_r}q^{(k)}(\widehat{\boldsymbol{\theta}}^{(k)[0]})\right]\right\|_2 \\ \leq& \frac{1}{|S|}\sum_{k\in S}\sqrt{1-\eta_r^{(k)}\mu_r^{(k)}}\|\widehat{\boldsymbol{\theta}}^{(k)[0]} - \boldsymbol{\theta}_r^{(k)*}\|_2 + \frac{1}{|S|}\sum_{k\in S}\gamma\cdot\sum_{r=1}^{R}\eta_r^{(k)}(|\widehat{w}_r^{(k)[0]} - w_r^{(k)*}| + \|\widehat{\boldsymbol{\theta}}_r^{(k)[0]} - \boldsymbol{\theta}_r^{(k)*}\|_2) \\ \leq& \left(\sqrt{1-\min_{k\in S, r\in[R]}(\eta_r^{(k)}\mu_r^{(k)})} + \gamma\bar{\eta}R\right)\max_{k\in S}\max_{r\in[R]}\|\widehat{\boldsymbol{\theta}}_r^{(k)[0]} - \boldsymbol{\theta}_r^{(k)*}\|_2 + \gamma\bar{\eta}R\cdot\max_{k\in S}\max_{r\in[R]}|\widehat{w}_r^{(k)[0]} - w_r^{(k)*}|,\end{aligned}$$

where the first part of the second inequality comes from the classical result of gradient descent (e.g., see Theorem 3.4 in Lan (2020)).

Similarly, we can show that

$$\max_{k\in S}\|\widetilde{\boldsymbol{\theta}}_r^{(k)[1]} - \boldsymbol{\theta}_r^{(k)*}\|_2 \leq \kappa_0 G^{[0]} + \bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big).$$

Therefore $\lambda^{[1]} = \widetilde{\kappa}_0\lambda^{[0]} + 15\sqrt{n}[\mathcal{W}(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}) + 2\bar{\eta}\mathcal{E}_1(n, \frac{\delta}{3RK})] = \frac{15}{119}\widetilde{\kappa}_0\sqrt{n}(r^*_w + r^*_{\boldsymbol{\theta}}) + 15\sqrt{n}[\mathcal{W}(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}) + 2\bar{\eta}\mathcal{E}_1(n, \frac{\delta}{3RK})] \geq \frac{5\sqrt{n}}{1-2\epsilon}\max_{k\in S}\|\widetilde{\boldsymbol{\theta}}^{(k)[1]}_r - \boldsymbol{\theta}^{(k)*}_r\|_2$ indeed holds, where $\widetilde{\kappa}_0 = 119\Big[\sqrt{1-\min_{k\in S, r\in[R]}(\eta^{(k)}_r\mu^{(k)}_r)} + \gamma\bar{\eta}R + \kappa R\Big]$.

For [2], we have

$$[2] \leq \bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big).$$

Combine the bounds of [1] and [2]:

$$\max_{k\in S}\max_{r\in[R]}\|\widehat{\boldsymbol{\theta}}^{(k)[1]}_r - \boldsymbol{\theta}^{(k)*}_r\|_2 \leq \left(\sqrt{1-\min_{k\in S, r\in[R]}(\eta^{(k)}_r\mu^{(k)}_r)} + \gamma\bar{\eta}R\right)\max_{k\in S}\max_{r\in[R]}(\|\widehat{\boldsymbol{\theta}}^{(k)[0]}_r - \boldsymbol{\theta}^{(k)*}_r\|_2 + |\widehat{w}^{(k)[0]}_r - w^{(k)*}_r|)$$

$$+ \frac{6}{1-2\epsilon}\min\Big\{3h, \frac{2\lambda^{[1]}}{5\sqrt{n}}\Big\} + \frac{2\lambda^{[1]}}{\sqrt{n}}\epsilon + \bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big).$$

On the other hand,

$$\begin{aligned}
\max_{k\in S}\max_{r\in[R]}|\widehat{w}^{(k)[0]}_r - w^{(k)*}_r| &\leq \max_{k\in S}\max_{r\in[R]}\left|\frac{1}{n}\sum_{i=1}^n\mathbb{P}(z^{(k)} = r|\boldsymbol{x}^{(k)}_i, \widehat{w}^{(k)[0]}, \widehat{\boldsymbol{\theta}}^{(k)[0]}) - \mathbb{E}\big[\mathbb{P}(z^{(k)} = r|\boldsymbol{x}^{(k)}_i, \widehat{w}^{(k)[0]}, \widehat{\boldsymbol{\theta}}^{(k)[0]})\big]\right| \\
&\quad + \max_{k\in S}\max_{r\in[R]}\Big|\mathbb{E}\big[\mathbb{P}(z^{(k)} = r|\boldsymbol{x}^{(k)}_i, \widehat{w}^{(k)[0]}, \widehat{\boldsymbol{\theta}}^{(k)[0]})\big] - w^{(k)*}_r\Big| \\
&\leq \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + \kappa\max_{k\in S}\sum_{r=1}^R(\|\widehat{\boldsymbol{\theta}}^{(k)[0]}_r - \boldsymbol{\theta}^{(k)*}_r\|_2 + |\widehat{w}^{(k)[0]}_r - w^{(k)*}_r|) \\
&\leq \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + \kappa R\max_{k\in S}\max_{r\in[R]}(\|\widehat{\boldsymbol{\theta}}^{(k)[0]}_r - \boldsymbol{\theta}^{(k)*}_r\|_2 + |\widehat{w}^{(k)[0]}_r - w^{(k)*}_r|). \qquad (7)
\end{aligned}$$

As a result,

$$\begin{aligned}
&\max_{k\in S}\max_{r\in[R]}(\|\widehat{\boldsymbol{\theta}}^{(k)[1]}_r - \boldsymbol{\theta}^{(k)*}_r\|_2 + |\widehat{w}^{(k)[1]}_r - w^{(k)*}_r|) \\
&\leq \left(\sqrt{1-\min_{k\in S, r\in[R]}(\eta^{(k)}_r\mu^{(k)}_r)} + \gamma\bar{\eta}R + \kappa R\right)\max_{k\in S}\max_{r\in[R]}(\|\widehat{\boldsymbol{\theta}}^{(k)[0]}_r - \boldsymbol{\theta}^{(k)*}_r\|_2 + |\widehat{w}^{(k)[0]}_r - w^{(k)*}_r|) \\
&\quad + \frac{6}{1-2\epsilon}\min\Big\{3h, \frac{2\lambda^{[1]}}{5\sqrt{n}}\Big\} + \frac{2\lambda^{[1]}}{\sqrt{n}}\epsilon + \bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big) + \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big).
\end{aligned}$$

Denote $G^{[t]} = \max_{k\in S}\max_{r\in[R]}(\|\widehat{\boldsymbol{\theta}}^{(k)[t]}_r - \boldsymbol{\theta}^{(k)*}_r\|_2 + |\widehat{w}^{(k)[t]}_r - w^{(k)*}_r|)$ and $\kappa_0 = \sqrt{1-\min_{k\in S, r\in[R]}(\eta^{(k)}_r\mu^{(k)}_r)} + \gamma\bar{\eta}R + \kappa R$. Then

$$\begin{aligned}
G^{[1]} &\leq \kappa_0 G^{[0]} + \frac{6}{1-2\epsilon}\min\Big\{3h, \frac{2\lambda^{[1]}}{5\sqrt{n}}\Big\} + \frac{2\lambda^{[1]}}{\sqrt{n}}\epsilon + \bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big) + \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) \\
&\leq \kappa_0 G^{[0]} + \left[\frac{12}{5(1-2\epsilon)} + 2\epsilon\right]\frac{\lambda^{[1]}}{\sqrt{n}} + \bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big) + \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) \\
&\leq \kappa_0 G^{[0]} + \frac{118}{15}\cdot\frac{\lambda^{[1]}}{\sqrt{n}} + \bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big) + \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big).
\end{aligned}$$

(II) Part 2: Iteration round $t \geq 2$.

Repeating the analysis in (I), we can see that when $\lambda^{[t]} \geq 15\sqrt{n}\kappa_0 G^{[t-1]} + 15\sqrt{n}\bar{\eta}\mathcal{E}_1(n, \frac{\delta}{3RK}) \geq 15\sqrt{n}\Big(\sqrt{1-\min_{k\in S, r\in[R]}(\eta^{(k)}_r\mu^{(k)}_r)} + \gamma\bar{\eta}R\Big)G^{[t-1]} + 15\sqrt{n}\bar{\eta}\mathcal{E}_1(n, \frac{\delta}{3RK})$,

$$G^{[t]} \leq \kappa_0 G^{[t-1]} + \frac{118}{15}\cdot\frac{\lambda^{[t]}}{\sqrt{n}} + \bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big) + \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big)$$

$$\leq \frac{119}{15}\cdot\frac{\lambda^{[t]}}{\sqrt{n}} + \bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big) + \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big). \tag{8}$$

Recall our setting of $\{\lambda^{[t]}\}_{t=1}^T$:

$$\lambda^{[0]} = \frac{15}{119}\sqrt{n}(r^*_w + r^*_{\boldsymbol{\theta}}),$$
$$\lambda^{[t]} = \tilde{\kappa}_0\lambda^{[t-1]} + 15\sqrt{n}\Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big].$$

Hence $\lambda^{[t]} \geq 15\sqrt{n}\kappa_0 G^{[t-1]} + 15\sqrt{n}\bar{\eta}\mathcal{E}_1(n, \frac{\delta}{3RK})$ indeed holds and

$$\lambda^{[t]} = (\widetilde{\kappa}_0)^t\lambda^{[0]} + \frac{1-(\widetilde{\kappa}_0)^t}{1-\widetilde{\kappa}_0}15\sqrt{n}\Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big], \tag{9}$$

which together with (8) implies

$$G^{[t]} \leq (\widetilde{\kappa}_0)^t(r^*_w + r^*_{\boldsymbol{\theta}}) + \left(\frac{119}{1-\widetilde{\kappa}_0} + 1\right)\Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big] \leq r^*_{\boldsymbol{\theta}}, \tag{10}$$

when $t \geq 1$. The last inequality holds due to Assumption A.7. Similar to (7), we have

$$\begin{aligned}\max_{k\in S}\max_{r\in[R]}|\widehat{w}_r^{(k)[t]} - w_r^{(k)*}| &\leq \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + \kappa R\cdot G^{[t-1]}\\ &\leq \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + \frac{119}{15}(\widetilde{\kappa}_0)^{t-1}\kappa R(r^*_w + r^*_{\boldsymbol{\theta}})\\ &\quad + \kappa R\left(\frac{119}{1-\widetilde{\kappa}_0} + 1\right)\Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big]\\ &\leq r^*_w,\end{aligned}$$

where the last inequality is due to Assumption A.7.

(III) Part 3: The case when $h \leq \frac{1}{3}[\mathcal{W}(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}) + 2\bar{\eta}\mathcal{E}_1(n, \frac{\delta}{3RK})]$.

In this case, we have $\lambda^{[t]} \geq 15\sqrt{n}\big[\mathcal{W}(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}) + 2\bar{\eta}\mathcal{E}_1(n, \frac{\delta}{3RK})\big] \geq \frac{15\sqrt{n}}{1-2\epsilon}h$ for $t \geq 1$. Then by Lemma E.1.(ii), $\widehat{\boldsymbol{\theta}}^{(k)[t]}$'s are equal for $k \in S$ when $t \geq 1$. Thus

$$G^{[1]} \leq \frac{119}{15}\tilde{\kappa}_0\lambda^{[0]} + \left(\frac{119}{1-\widetilde{\kappa}_0} + 1\right)\Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big],$$

$$\begin{aligned}\max_{k\in S}\max_{r\in[R]}\|\widehat{\boldsymbol{\theta}}_r^{(k)[t]} - \boldsymbol{\theta}_r^{(k)*}\|_2 &\leq \left(\sqrt{1 - \min_{k\in S, r\in[R]}(\eta_r^{(k)}\mu_r^{(k)})} + \gamma\bar{\eta}R\right)G^{[t-1]} + \bar{\eta}\mathcal{E}_2\Big(n, |S|, \frac{\delta}{3R}\Big)\\ &\quad + \frac{6}{1-2\epsilon}\min\Big\{3h, \frac{2\lambda^{[t]}}{5\sqrt{n}}\Big\} + \frac{2\lambda^{[t]}}{\sqrt{n}}\epsilon,\\ \max_{k\in S}\max_{r\in[R]}|\widehat{w}_r^{(k)[t]} - w_r^{(k)*}| &\leq \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + \kappa R G^{[t-1]},\end{aligned}$$

which implies that

$$G^{[t]} \leq \kappa_0 G^{[t-1]} + \bar{\eta}\mathcal{E}_2\Big(n, |S|, \frac{\delta}{3R}\Big) + \frac{6}{1-2\epsilon}\min\Big\{3h, \frac{2\lambda^{[t]}}{5\sqrt{n}}\Big\} + \frac{2\lambda^{[t]}}{\sqrt{n}}\epsilon + \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big),$$

when $t \geq 2$. By induction,

$$G^{[t]} \leq \kappa_0^{t-1}G^{[1]} + \frac{1-\kappa_0^{t-1}}{1-\kappa_0}\bar{\eta}\mathcal{E}_2\Big(n, |S|, \frac{\delta}{3R}\Big) + \frac{6}{1-2\epsilon}\underbrace{\sum_{t'=2}^{t}\kappa_0^{t-t'}\cdot\min\Big\{3h, \frac{2\lambda^{[t']}}{5\sqrt{n}}\Big\}}_{[3]} + \frac{2\epsilon}{\sqrt{n}}\underbrace{\sum_{t'=2}^{t}\kappa_0^{t-t'}\cdot\lambda^{[t']}}_{[4]}$$

$$+ \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big).$$

By (9),

$$[3] \le \min\Big\{3\frac{1-\kappa_0^{t-1}}{1-\kappa_0}h, \frac{2}{5}(t-1)(\tilde{\kappa}_0)^t \cdot \frac{\lambda^{[0]}}{\sqrt{n}} + \frac{6}{1-\tilde{\kappa}_0} \cdot \frac{1-\kappa_0^{t-1}}{1-\kappa_0}\Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big]\Big\},$$
$$[4] \le (t-1)(\tilde{\kappa}_0)^t\lambda^{[0]} + \frac{15}{1-\tilde{\kappa}_0} \cdot \frac{1-\kappa_0^{t-1}}{1-\kappa_0}\Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big]\sqrt{n}.$$

Therefore,

$$\begin{aligned}
G^{[t]} &\le (\tilde{\kappa}_0/119)^{t-1}G^{[1]} + \frac{1}{1-\kappa_0}\Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + \bar{\eta}\mathcal{E}_2\Big(n, |S|, \frac{\delta}{3R}\Big)\Big] \\
&\quad + \frac{18}{1-\tilde{\kappa}_0/119} \cdot \min\Big\{3h, \frac{6}{1-\tilde{\kappa}_0}\Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big]\Big\} \\
&\quad + \frac{30}{(1-\tilde{\kappa}_0)(1-\tilde{\kappa}_0/119)}\epsilon \cdot \Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big] + \Big(\frac{2}{3} + \frac{2}{5} \cdot 18\Big) \cdot (t-1)(\tilde{\kappa}_0)^t\frac{\lambda^{[0]}}{\sqrt{n}} \\
&\le \frac{119}{15}\tilde{\kappa}_0(\tilde{\kappa}_0/119)^{t-1}\Big\{\frac{\lambda^{[0]}}{\sqrt{n}} + \Big(\frac{119}{1-\tilde{\kappa}_0} + 1\Big)\Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big]\Big\} \\
&\quad + \frac{1}{1-\kappa_0}\Big[\bar{\eta}\mathcal{E}_2\Big(n, |S|, \frac{\delta}{3R}\Big) + \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big)\Big] \\
&\quad + \frac{18}{1-\tilde{\kappa}_0/119} \cdot \min\Big\{3h, \frac{6}{1-\tilde{\kappa}_0}\Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big]\Big\} \\
&\quad + \frac{30}{(1-\tilde{\kappa}_0)(1-\tilde{\kappa}_0/119)}\epsilon \cdot \Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big] + \frac{118}{15} \cdot (t-1)(\tilde{\kappa}_0)^t\frac{\lambda^{[0]}}{\sqrt{n}} \\
&\le \tilde{\kappa}_0(\tilde{\kappa}_0/119)^{t-1}(r^*_w + r^*_{\boldsymbol{\theta}}) + \Big[\frac{119}{15}\tilde{\kappa}_0(\tilde{\kappa}_0/119)^{t-1} + \frac{118}{119}(t-1)(\tilde{\kappa}_0)^t\Big](r^*_w + r^*_{\boldsymbol{\theta}}) \\
&\quad + \frac{1}{1-\kappa_0}\Big[\bar{\eta}\mathcal{E}_2\Big(n, |S|, \frac{\delta}{3R}\Big) + \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big)\Big] \\
&\quad + \frac{18}{1-\tilde{\kappa}_0/119} \cdot \min\Big\{3h, \frac{6}{1-\tilde{\kappa}_0}\Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big]\Big\} \\
&\quad + \frac{30}{(1-\tilde{\kappa}_0)(1-\tilde{\kappa}_0/119)}\epsilon \cdot \Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big].
\end{aligned}$$

Note that $\tilde{\kappa}_0(\tilde{\kappa}_0/119)^{t-1} + \frac{119}{15}\tilde{\kappa}_0(\tilde{\kappa}_0/119)^{t-1} + \frac{118}{119}(t-1)(\tilde{\kappa}_0)^t \le 9\tilde{\kappa}_0(\tilde{\kappa}_0/119)^{t-1} + \frac{118}{119}(t-1)(\tilde{\kappa}_0)^t \le 10t(\tilde{\kappa}_0)^t$, hence

$$\begin{aligned}
G^{[t]} &\le 20t(\tilde{\kappa}_0)^{t-1}(r^*_w \vee r^*_{\boldsymbol{\theta}}) + \Big[\frac{119}{15}\tilde{\kappa}_0(\tilde{\kappa}_0/119)^{t-1} + \frac{118}{119}(t-1)(\tilde{\kappa}_0)^t\Big](r^*_w + r^*_{\boldsymbol{\theta}}) \\
&\quad + \frac{1}{1-\kappa_0}\Big[\bar{\eta}\mathcal{E}_2\Big(n, |S|, \frac{\delta}{3R}\Big) + \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big)\Big] \\
&\quad + \frac{18}{1-\tilde{\kappa}_0/119} \cdot \min\Big\{3h, \frac{6}{1-\tilde{\kappa}_0}\Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big]\Big\} \\
&\quad + \frac{30}{(1-\tilde{\kappa}_0)(1-\tilde{\kappa}_0/119)}\epsilon \cdot \Big[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta}}\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\Big]. \qquad (11)
\end{aligned}$$

By Assumption 4, we obtain $\max_{k\in S}\max_{r\in[R]}\|\widehat{\boldsymbol{\theta}}_r^{(k)[t]} - \boldsymbol{\theta}_r^{(k)*}\|_2 \le G^{[t]} \le r^*_{\boldsymbol{\theta}}$.

Next, let us shrink the contraction radius to obtain the desired rate. Recall that

$$A_t = \Big[9\tilde{\kappa}_0\Big(\frac{\tilde{\kappa}_0}{119}\Big)^{t-1} + \frac{118}{119}(t-1)\tilde{\kappa}_0^{t-1}\Big](r^*_w + r^*_{\boldsymbol{\theta}}) + \frac{1}{1-\tilde{\kappa}_0/119}\bar{\eta}\mathcal{E}_2\Big(n, |S|, \frac{\delta}{3R}\Big)$$

$$+\frac{18}{1-\tilde{\kappa}_0/119}\min\left\{3h,\frac{6}{1-\tilde{\kappa}_0}\Big[\mathcal{W}\Big(n,\frac{\delta}{3RK},r^*_{\boldsymbol{\theta}}\Big)+2\bar{\eta}\mathcal{E}_1\Big(n,\frac{\delta}{3R}\Big)\Big]\right\}$$
$$+\frac{30}{(1-\tilde{\kappa}_0)(1-\tilde{\kappa}_0/119)}\epsilon\Big[\mathcal{W}\Big(n,\frac{\delta}{3RK},r^*_{\boldsymbol{\theta}}\Big)+2\bar{\eta}\mathcal{E}_1\Big(n,\frac{\delta}{3R}\Big)\Big],$$

and

$$A_t+\frac{18}{1-\tilde{\kappa}_0/119}\mathcal{W}\Big(n,\frac{\delta}{3RK},r^*_{\boldsymbol{\theta},t}\Big)=r^*_{\boldsymbol{\theta},t+1},$$

with $r^*_{\boldsymbol{\theta},1}:=r^*_{\boldsymbol{\theta}}$. Then we repeat the previous analysis in part (III) for $t=1:T$, then we will get the same rate as in (11) but replace $r^*_{\boldsymbol{\theta}}$ with $r^*_{\boldsymbol{\theta},T}$ in the term $\frac{1}{1-\kappa_0}\big[\bar{\eta}\mathcal{E}_2(n,|S|,\frac{\delta}{3R})+\mathcal{W}(n,\frac{\delta}{3RK},r^*_{\boldsymbol{\theta}})\big]$.

(IV) Part 4: Combining this rate (which holds when $h\le\frac{1}{3}[\mathcal{W}(n,\frac{\delta}{3RK},r^*_{\boldsymbol{\theta}})+2\bar{\eta}\mathcal{E}_1(n,\frac{\delta}{3RK})]$) with (10) (which holds for any $h\ge 0$ but is only used when $h>\frac{1}{3}[\mathcal{W}(n,\frac{\delta}{3RK},r^*_{\boldsymbol{\theta}})+2\bar{\eta}\mathcal{E}_1(n,\frac{\delta}{3RK})]$) completes our proof.

E.2. Proof of Proposition A.11

We first introduce two useful lemmas.

Lemma E.2 (Theorem 3 in Maurer & Pontil (2021)). *Let $f:\mathcal{X}^n\to\mathbb{R}$ and $X=(X_1,\ldots,X_n)$ be a vector of independent random variables with values in a space $\mathcal{X}$. Then for any $t>0$ we have*

$$\mathbb{P}(f(X)-\mathbb{E}f(X)>t)\le\exp\left\{-\frac{t^2}{32e\left\|\sum_{i=1}^n\|f_i(X)\|^2_{\psi_2}\right\|_\infty}\right\},$$

where $f_i(X)$ as a random function of x is defined to be $(f_i(X))(x):=f(x_1,\ldots,x_{i-1},X_i,x_{i+1},\ldots,X_n)-\mathbb{E}_{X_i}[f(x_1,\ldots,x_{i-1},X_i,x_{i+1},\ldots,X_n)]$, the sub-Gaussian norm $\|Z\|_{\psi_2}:=\sup_{d\ge1}\{\|Z\|_d/\sqrt{d}\}$, and $\|Z\|_d=(\mathbb{E}|Z|^d)^{1/d}$.

Lemma E.3 (Vectorized contraction of Rademacher complexity, Corollary 1 in Maurer (2016)). *Suppose $\{\epsilon_{ir}\}_{i\in[n],r\in[R]}$ and $\{\epsilon_i\}_{i=1}^n$ are independent Rademacher variables. Let $\mathcal{F}$ be a class of functions $f:\mathbb{R}^d\to\mathcal{S}\subseteq\mathbb{R}^R$ and $h:\mathcal{S}\to\mathbb{R}$ is L-Lipschitz under ℓ_2-norm, i.e., $|h(\boldsymbol{y})-h(\boldsymbol{y}')|\le L\|\boldsymbol{y}-\boldsymbol{y}'\|_2$, where $\boldsymbol{y}=(y_1,\ldots,y_R)^T$, $\boldsymbol{y}'=(y'_1,\ldots,y'_R)^T\in\mathcal{S}$. Then*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n\epsilon_i h(f(x_i))\le\sqrt{2}L\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n\sum_{r=1}^R\epsilon_{ir}f_r(x_i),$$

where $f_r(x_i)$ is the r-th component of $f(x_i)\in\mathcal{S}\subseteq\mathbb{R}^R$.

Define the posterior

$$\gamma^{(r)}_{\boldsymbol{\theta}^{(k)},\boldsymbol{w}^{(k)}}(\boldsymbol{x}^{(k)})=\frac{w^{(k)}_r\exp\{(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}^{(k)}_r-\boldsymbol{\theta}^{(k)}_1)-\frac{1}{2}(\|\boldsymbol{\theta}^{(k)}_r\|_2^2-\|\boldsymbol{\theta}^{(k)}_1\|_2^2)\}}{w^{(k)}_1+\sum_{r=2}^R w^{(k)}_r\exp\{(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}^{(k)}_r-\boldsymbol{\theta}^{(k)}_1)-\frac{1}{2}(\|\boldsymbol{\theta}^{(k)}_r\|_2^2-\|\boldsymbol{\theta}^{(k)}_1\|_2^2)\}}$$
$$=\mathbb{P}(z^{(k)}=r|\boldsymbol{x}^{(k)};\boldsymbol{\theta}^{(k)},\boldsymbol{w}^{(k)}),\ r\in[R],$$

where $\boldsymbol{w}^{(k)}=\{w^{(k)}\}_{k=1}^K$ and $\boldsymbol{\theta}^{(k)}=\{\boldsymbol{\theta}^{(k)}_r\}_{k=1}^K$.

By definition, $q^{(k)}(\boldsymbol{\theta})=Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)*},\boldsymbol{w}^{(k)*})=-\frac{1}{2}\mathbb{E}\big[\sum_{r=1}^R\gamma^{(r)}_{\boldsymbol{\theta}^{(k)*},\boldsymbol{w}^{(k)*}}(\boldsymbol{x}^{(k)})\|\boldsymbol{x}^{(k)}-\boldsymbol{\theta}\|_2^2\big]$, hence $\mu^{(k)}_r=L^{(k)}_r=w^{(k)*}_r$ with $r^*_1=+\infty$. And

$$\widehat{Q}^{(k)}(\boldsymbol{\theta}|\boldsymbol{\theta}',\boldsymbol{w}')=-\frac{1}{2n_k}\sum_{i=1}^n\sum_{r=1}^R\gamma^{(r)}_{\boldsymbol{\theta}',\boldsymbol{w}'}(\boldsymbol{x}^{(k)}_i)\|\boldsymbol{x}^{(k)}_i-\boldsymbol{\theta}\|_2^2,$$
$$\frac{\partial}{\partial\boldsymbol{\theta}_r}Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{\theta}',\boldsymbol{w}')=-\mathbb{E}_{\boldsymbol{x}^{(k)}}\big[\gamma^{(r)}_{\boldsymbol{\theta}',\boldsymbol{w}'}(\boldsymbol{x}^{(k)})(\boldsymbol{\theta}-\boldsymbol{x}^{(k)})\big],$$
$$\frac{\partial}{\partial\boldsymbol{\theta}_r}\widehat{Q}^{(k)}(\boldsymbol{\theta}|\boldsymbol{\theta}',\boldsymbol{w}')=-\frac{1}{n_k}\sum_{i=1}^n\gamma^{(r)}_{\boldsymbol{\theta}',\boldsymbol{w}'}(\boldsymbol{x}^{(k)}_i)(\boldsymbol{\theta}-\boldsymbol{x}^{(k)}_i).$$

From the proof of Theorem 1 in Tian et al. (2022), we have $\kappa \asymp c_w^{-2}R^2\exp\{-C\Delta^2\}$, $\gamma \asymp M^2c_w^{-2}R^2\exp\{-C\Delta^2\}$ with $r_2 = C_b\Delta$. Consider $r_{\boldsymbol{\theta}}^* = r_1^* \wedge r_2^* = C_b\Delta \le M$. In the remaining proof of Proposition 1, we will derive the expressions of $\mathcal{W}, \mathcal{E}_1$ and $\mathcal{E}_2$. Let

$$
\begin{aligned}
V &= \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le \xi}} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{P}(z^{(k)} = r | \boldsymbol{x}_i^{(k)}; \boldsymbol{w}, \boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{x}^{(k)}}\left[\mathbb{P}(z^{(k)} = r | \boldsymbol{x}^{(k)}; \boldsymbol{w}, \boldsymbol{\theta})\right] \right| \\
&= \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le \xi}} \left| \frac{1}{n}\sum_{i=1}^n \gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}_i^{(k)}) - \mathbb{E}\left[\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}^{(k)})\right] \right|.
\end{aligned}
$$

By bounded difference inequality (Corollary 2.21 in Wainwright (2019)), w.p. at least $1-\delta$,

$$
V \le \mathbb{E}V + \sqrt{\frac{\log(1/\delta)}{n}}.
$$

And by classical symmetrization arguments (e.g., see Proposition 4.11 in Wainwright (2019)),

$$
\mathbb{E}V \le \frac{2}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le \xi}} \left| \sum_{i=1}^n \epsilon_i^{(k)} \gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(k)}(\boldsymbol{x}_i^{(k)}) \right|.
$$

Let $g_{ir}^{(k)} = (\boldsymbol{\theta}_r - \boldsymbol{\theta}_1)^T\boldsymbol{x}_i^{(k)} - \frac{1}{2}(\|\boldsymbol{\theta}_r\|_2^2 - \|\boldsymbol{\theta}_1\|_2^2) + \log w_r - \log w_1$, $\varphi(\boldsymbol{x}) = \frac{\exp\{x_r\}}{1+\sum_{r=2}^R \exp\{x_r\}}$, where φ is 1-Lipschitz (w.r.t. ℓ_2-norm) and $\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}) = \varphi(\{g_{ir}^{(k)}\}_{r=2}^R)$. Then by Lemma E.3,

$$
\begin{aligned}
&\frac{2}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le \xi}} \left| \sum_{i=1}^n \epsilon_i^{(k)} \gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(k)}(\boldsymbol{x}_i^{(k)}) \right| \\
&\lesssim \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le \xi}} \left| \sum_{i=1}^n \sum_{r=2}^R \epsilon_{ir}^{(k)} g_{ir}^{(k)} \right| \\
&\lesssim \frac{1}{n}\sum_{r=2}^R \mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le \xi}} \left| \sum_{i=1}^n \epsilon_{ir}^{(k)} g_{ir}^{(k)} \right| \\
&\lesssim \sum_{r=2}^R \left\{ \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le \xi}} \left| \sum_{i=1}^n \epsilon_{ir}^{(k)} (\boldsymbol{\theta}_r - \boldsymbol{\theta}_1)^T\boldsymbol{x}_i^{(k)} \right| + \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le \xi}} \left| \sum_{i=1}^n \epsilon_{ir}^{(k)} (\|\boldsymbol{\theta}_r\|_2^2 - \|\boldsymbol{\theta}_1\|_2^2) \right| \right. \\
&\qquad \left. + \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le \xi}} \left| \sum_{i=1}^n \epsilon_{ir}^{(k)} (\log w_r - \log w_1) \right| \right\} \\
&\lesssim \sum_{r=2}^R \left\{ \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le \xi}} \left| \sum_{i=1}^n \epsilon_{ir}^{(k)} (\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*})^T\boldsymbol{x}_i^{(k)} \right| + \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le \xi}} \left| \sum_{i=1}^n \epsilon_{ir}^{(k)} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_1^{(k)*})^T\boldsymbol{x}_i^{(k)} \right| \right. \\
&\qquad + \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \left| \sum_{i=1}^n \epsilon_{ir}^{(k)} (\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_1^{(k)*})^T\boldsymbol{x}_i^{(k)} \right| + \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le \xi}} \left| \sum_{i=1}^n \epsilon_{ir}^{(k)} (\boldsymbol{\theta}_r + \boldsymbol{\theta}_1)^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1) \right| \\
&\qquad \left. + \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le \xi}} \left| \sum_{i=1}^n \epsilon_{ir}^{(k)} (\log w_r - \log w_1) \right| \right\}
\end{aligned}
$$

$$\lesssim RM\xi\sqrt{\frac{d}{n}}+[RM^2+R\log(Rc_w^{-1})]\sqrt{\frac{1}{n}},$$

which implies

$$V\lesssim RM\xi\sqrt{\frac{d}{n}}+[RM^2+R\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+\sqrt{\frac{\log(1/\delta)}{n}}\asymp\mathcal{W}(n,\delta,\xi).$$

w.p. at least $1-\delta$.

Next, let

$$\begin{aligned}U&=\sup_{\substack{|w_r-w_r^{(k)*}|\le\frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le r_{\boldsymbol{\theta}}^*}}\left\|\frac{1}{n}\sum_{i=1}^n\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}_i^{(k)})\boldsymbol{x}_i^{(k)}-\mathbb{E}\big[\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}^{(k)})\boldsymbol{x}^{(k)}\big]\right\|_2\\&=\sup_{\|\boldsymbol{u}\|_2\le1}\sup_{\substack{|w_r-w_r^{(k)*}|\le\frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le r_{\boldsymbol{\theta}}^*}}\left|\frac{1}{n}\sum_{i=1}^n\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}_i^{(k)})(\boldsymbol{x}_i^{(k)})^T\boldsymbol{u}-\mathbb{E}\big[\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k)})^T\boldsymbol{u}\big]\right|\\&\le2\max_{j=1:N}\underbrace{\sup_{\substack{|w_r-w_r^{(k)*}|\le\frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le r_{\boldsymbol{\theta}}^*}}\left|\frac{1}{n}\sum_{i=1}^n\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}_i^{(k)})(\boldsymbol{x}_i^{(k)})^T\boldsymbol{u}_j-\mathbb{E}\big[\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k)})^T\boldsymbol{u}_j\big]\right|}_{U_j},\end{aligned}$$

where $\{\boldsymbol{u}_j\}_{j=1}^N$ is a $1/2$-cover of the unit ball $\mathcal{B}(\mathbf{0},1)$ in $\mathbb{R}^d$ w.r.t. ℓ_2-norm, with $N\le5^d$ (by Example 5.8 in (Wainwright, 2019)). We first bound $U_j-\mathbb{E}U_j$ as follows. Fix $\boldsymbol{x}_1^{(k)},\dots,\boldsymbol{x}_{i-1}^{(k)},\boldsymbol{x}_{i+1}^{(k)},\dots,\boldsymbol{x}_n^{(k)}$ and define $s_{ir}^{(k)}(\boldsymbol{x}_i^{(k)})=V_j-\mathbb{E}[V_j|\boldsymbol{x}_1^{(k)},\dots,\boldsymbol{x}_{i-1}^{(k)},\boldsymbol{x}_{i+1}^{(k)},\dots,\boldsymbol{x}_n^{(k)}]$. Then

$$|s_{ir}^{(k)}(\boldsymbol{x}_i^{(k)})|\le\underbrace{\frac{1}{n}\sup_{\substack{|w_r-w_r^{(k)*}|\le\frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le r_{\boldsymbol{\theta}}^*}}\left|\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}_i^{(k)})(\boldsymbol{x}_i^{(k)})^T\boldsymbol{u}_j\right|}_{W_1}+\underbrace{\frac{2}{n}\mathbb{E}\left|\sup_{\substack{|w_r-w_r^{(k)*}|\le\frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le r_{\boldsymbol{\theta}}^*}}\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}_i^{(k)})(\boldsymbol{x}_i^{(k)})^T\boldsymbol{u}_j\right|}_{W_2},$$

where $[\mathbb{E}(W_1+W_2)^d]^{1/d}\le(\mathbb{E}W_1^d)^{1/d}+(\mathbb{E}W_2^d)^{1/d}$, and $(\mathbb{E}W_1^d)^{1/d},(\mathbb{E}W_2^d)^{1/d}\le CM\sqrt{d}/n$ with some constant$C>0$. Then by Lemma E.2,

$$\mathbb{P}(U_j-\mathbb{E}U_j\ge t)\lesssim\exp\left\{-\frac{Cnt^2}{M^2}\right\}.$$

By a similar procedure used in deriving $\mathcal{W}(n,\delta,\xi)$, we can show that

$$\mathbb{E}U_j\lesssim RM^2r_{\boldsymbol{\theta}}^*\sqrt{\frac{d}{n}}+[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}.$$

As a consequence,

$$\mathbb{P}\left(U_j\ge CRM^2r_{\boldsymbol{\theta}}^*\sqrt{\frac{d}{n}}+C[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+t\right)\lesssim\exp\left\{-\frac{Cnt^2}{M^2}\right\}.$$

Therefore

$$\mathbb{P}\left(\max_{j=1:N}U_j\ge CRM^2r_{\boldsymbol{\theta}}^*\sqrt{\frac{d}{n}}+C[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+t\right)\lesssim N\exp\left\{-\frac{Cnt^2}{M^2}\right\},$$

which implies that

$$U\lesssim(RM^2r_{\boldsymbol{\theta}}^*+M)\sqrt{\frac{d}{n}}+[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+M\sqrt{\frac{\log(1/\delta)}{n}},$$

w.p. at least $1-\delta$. On the other hand, by $\mathcal{W}(n,\delta,r^*_{\boldsymbol{\theta}})$, we have

$$\sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le r^*_{\boldsymbol{\theta}}}}\left\|\frac{1}{n}\sum_{i=1}^n \gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}_i^{(k)})\boldsymbol{\theta}-\mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}^{(k)})\boldsymbol{\theta}\big]\right\|_2 \lesssim \left[RM\xi\sqrt{\frac{d}{n}}+[RM^2+R\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+\sqrt{\frac{\log(1/\delta)}{n}}\right]\cdot M,$$

hence

$$\begin{aligned}\mathcal{E}_1(n,\delta) &\asymp (RM^2r^*_{\boldsymbol{\theta}}+M)\sqrt{\frac{d}{n}}+[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+M\sqrt{\frac{\log(1/\delta)}{n}}\\ &\asymp RM^3\sqrt{\frac{d}{n}}+[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+M\sqrt{\frac{\log(1/\delta)}{n}},\end{aligned}$$

where the last inequality is due to $r^*_{\boldsymbol{\theta}}=r_1^*\wedge r_2^*=C_b\Delta\le M$.

Let

$$\begin{aligned}Z &= \sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le r^*_{\boldsymbol{\theta}}\\ 0<\le\eta_r^{(k)}\le\bar\eta}}\frac{1}{n|S|}\left\|\sum_{k\in S}\eta_r^{(k)}\cdot\sum_{i=1}^n\big(\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}_i^{(k)})\boldsymbol{x}_i^{(k)}-\mathbb{E}[\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}^{(k)})\boldsymbol{x}^{(k)}]\big)\right\|_2\\ &= \sup_{\|\boldsymbol{u}\|_2\le1}\sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le r^*_{\boldsymbol{\theta}}\\ 0<\le\eta_r^{(k)}\le\bar\eta}}\frac{1}{n|S|}\left|\sum_{k\in S}\eta_r^{(k)}\cdot\sum_{i=1}^n\big(\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}_i^{(k)})(\boldsymbol{x}_i^{(k)})^T\boldsymbol{u}-\mathbb{E}[\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k)})^T\boldsymbol{u}]\big)\right|\\ &\le \sup_{j_1',\ldots,j_k'=1:N'}\sup_{j=1:N}\underbrace{\frac{2}{n|S|}\sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le r^*_{\boldsymbol{\theta}}}}\left|\sum_{k\in S}\eta_{j_k'}\cdot\sum_{i=1}^n\big(\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}_i^{(k)})(\boldsymbol{x}_i^{(k)})^T\boldsymbol{u}_j-\mathbb{E}[\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k)})^T\boldsymbol{u}_j]\big)\right|}_{Z(j,j_1',\ldots,j_k')},\end{aligned}$$

where $\{\boldsymbol{u}_j\}_{j=1}^N$ is a $1/2$-cover of the unit ball $\mathcal{B}(\mathbf{0},1)$ in $\mathbb{R}^d$ w.r.t. ℓ_2-norm with $N\le 5^d$ and $\{\eta_{j'}\}_{j'=1}^{N'}$ is a $1/2$-cover of $[0,1]$ with $N'\le 2$. We first bound $Z(j,j_1',\ldots,j_k')-\mathbb{E}Z(j,j_1',\ldots,j_k')$ as follows. Fix $\boldsymbol{x}_1^{(k)},\ldots,\boldsymbol{x}_{i-1}^{(k)},\boldsymbol{x}_{i+1}^{(k)},\ldots,\boldsymbol{x}_n^{(k)}$ and define $v_{ir}^{(k)}(\boldsymbol{x}_i^{(k)})=Z(j,j_1',\ldots,j_k')-\mathbb{E}[Z(j,j_1',\ldots,j_k')|\{\boldsymbol{x}_i^{(k)}\}_{k\in S,i\in[n]}\backslash\{\boldsymbol{x}_i^{(k)}\}]$. Then

$$|v_{ir}^{(k)}(\boldsymbol{x}_i^{(k)})|\le \underbrace{\frac{\eta_{j_k}}{n|S|}\sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le r^*_{\boldsymbol{\theta}}}}\left|\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}_i^{(k)})(\boldsymbol{x}_i^{(k)})^T\boldsymbol{u}_j\right|}_{W_1}+\underbrace{\frac{2\eta_{j_k}}{n|S|}\mathbb{E}\left|\sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le r^*_{\boldsymbol{\theta}}}}\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}_i^{(k)})(\boldsymbol{x}_i^{(k)})^T\boldsymbol{u}_j\right|}_{W_2}.$$

Via the same procedure used to bound U_j, it can be shown that

$$\begin{aligned}\mathbb{P}(Z(j,j_1',\ldots,j_k')-\mathbb{E}Z(j,j_1',\ldots,j_k')\ge t)&\lesssim\exp\left\{-\frac{Cn|S|t^2}{M^2\bar\eta^2}\right\},\\ \mathbb{E}Z(j,j_1',\ldots,j_k')&\lesssim\bar\eta RM^2r^*_{\boldsymbol{\theta}}\sqrt{\frac{d}{n|S|}}+\bar\eta[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}},\end{aligned}$$

leading to

$$\mathbb{P}\left(Z(j,j_1',\ldots,j_k')\ge\bar\eta RM^2r^*_{\boldsymbol{\theta}}\sqrt{\frac{d}{n|S|}}+\bar\eta[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+t\right)\lesssim\exp\left\{-\frac{Cn|S|t^2}{M^2\bar\eta^2}\right\}.$$

Therefore

$$\mathbb{P}\left(\max_{j_1',\ldots,j_k'=1:N'}\max_{j=1:N}Z(j,j_1',\ldots,j_k')\ge C\bar\eta RM^2r^*_{\boldsymbol{\theta}}\sqrt{\frac{d}{n|S|}}+C\bar\eta[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+t\right)$$

$$\lesssim N(N')^K \exp\Big\{-\frac{Cn|S|t^2}{M^2\bar{\eta}^2}\Big\},$$

which implies that

$$Z \le \max_{j'_1,\ldots,j'_k=1:N'} \max_{j=1:N} Z(j,j'_1,\ldots,j'_k) \lesssim \bar{\eta}(RM^2 r^*_{\boldsymbol{\theta}}+M)\sqrt{\frac{d}{n|S|}}+\bar{\eta}[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+\bar{\eta}M\sqrt{\frac{\log(1/\delta)}{n|S|}},$$

w.p. at least $1-\delta$. Similarly,

$$\begin{aligned}&\sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le r^*_{\boldsymbol{\theta}}\\ 0<\le\eta_r^{(k)}\le\bar{\eta}}} \frac{1}{n|S|}\Big\|\sum_{k\in S}\eta_r^{(k)}\cdot\sum_{i=1}^n\big(\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}_i^{(k)})\boldsymbol{\theta}_r-\mathbb{E}[\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}^{(k)})\boldsymbol{\theta}_r]\big)\Big\|_2\\ &\lesssim \bar{\eta}(RM^2 r^*_{\boldsymbol{\theta}}+M)\sqrt{\frac{d}{n|S|}}+\bar{\eta}[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+\bar{\eta}M\sqrt{\frac{\log(1/\delta)}{n|S|}},\end{aligned}$$

w.p. at least $1-\delta$. Considering that $r^*_{\boldsymbol{\theta}}=r_1^*\wedge r_2^*=C_b\Delta\le M$, we have

$$\mathcal{E}_2(n,|S|,\delta)\asymp RM^3\sqrt{\frac{d}{n|S|}}+[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+M\sqrt{\frac{\log(1/\delta)}{n|S|}}.$$

E.3. Proof of Proposition A.14

Recall that

$$\begin{aligned}A_t=&\Big[9\tilde{\kappa}_0\Big(\frac{\tilde{\kappa}_0}{119}\Big)^{t-1}+\frac{118}{119}(t-1)\tilde{\kappa}_0^{t-1}\Big](r_w^*+r^*_{\boldsymbol{\theta}})+\frac{1}{1-\tilde{\kappa}_0/119}\bar{\eta}\mathcal{E}_2\Big(n,|S|,\frac{\delta}{3R}\Big)\\&+\frac{18}{1-\tilde{\kappa}_0/119}\min\Big\{3h,\frac{6}{1-\tilde{\kappa}_0}\Big[\mathcal{W}\Big(n,\frac{\delta}{3RK},r^*_{\boldsymbol{\theta}}\Big)+2\bar{\eta}\mathcal{E}_1\Big(n,\frac{\delta}{3R}\Big)\Big]\Big\}\\&+\frac{30}{(1-\tilde{\kappa}_0)(1-\tilde{\kappa}_0/119)}\epsilon\Big[\mathcal{W}\Big(n,\frac{\delta}{3RK},r^*_{\boldsymbol{\theta}}\Big)+2\bar{\eta}\mathcal{E}_1\Big(n,\frac{\delta}{3R}\Big)\Big],\end{aligned}$$

and

$$A_t+\frac{18}{1-\tilde{\kappa}_0/119}\mathcal{W}\Big(n,\frac{\delta}{3RK},r^*_{\boldsymbol{\theta},t}\Big)=r^*_{\boldsymbol{\theta},t+1},$$

for $t\ge 1$ with $r^*_{\boldsymbol{\theta},1}:=r^*_{\boldsymbol{\theta}}$.

By Assumption A.9.(iv), there exists $\tilde{\kappa}_0'\in(0,1)$ such that $CRM\sqrt{\frac{p}{n}}\le\tilde{\kappa}_0'$ with a large C. Hence by plugging in the explicit rates obtained in Proposition A.11,

$$\begin{aligned}r^*_{\boldsymbol{\theta},t+1}\le&\tilde{\kappa}_0' r^*_{\boldsymbol{\theta},t}+Ct(\tilde{\kappa}_0)^{t-1}(r_w^*\vee r^*_{\boldsymbol{\theta}})+C\bar{\eta}RM^3\sqrt{\frac{d}{n|S|}}+C[(\bar{\eta}M)\vee 1][RM^2+R\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}\\&+C[(\bar{\eta}M)\vee 1]\sqrt{\frac{\log(RK/\delta)}{n}}+C\min\Big\{h,\bar{\eta}RM^3\sqrt{\frac{d}{n}}\Big\}+\epsilon\bar{\eta}RM^2[(\bar{\eta}M)\vee 1]\sqrt{\frac{d}{n}},\end{aligned}$$

implying that

$$\begin{aligned}r^*_{\boldsymbol{\theta},T}\lesssim&(\tilde{\kappa}_0')^{T-1}r^*_{\boldsymbol{\theta}}+T^2(\tilde{\kappa}_0')^{T-1}(r_w^*\vee r^*_{\boldsymbol{\theta}})+\bar{\eta}RM^3\sqrt{\frac{d}{n|S|}}+[(\bar{\eta}M)\vee 1][RM^2+R\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}\\&+[(\bar{\eta}M)\vee 1]\sqrt{\frac{\log(RK/\delta)}{n}}+\min\Big\{h,\bar{\eta}RM^3\sqrt{\frac{d}{n}}\Big\}+\epsilon\bar{\eta}RM^2[(\bar{\eta}M)\vee 1]\sqrt{\frac{d}{n}}\end{aligned}$$

$$\lesssim T^2(\tilde{\kappa}_0 \vee \tilde{\kappa}_0')^{T-1}(r_w^* \vee r_{\boldsymbol{\theta}}^*) + \bar{\eta}RM^3\sqrt{\frac{d}{n|S|}} + [(\bar{\eta}M) \vee 1][RM^2 + R\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}$$
$$+ [(\bar{\eta}M) \vee 1]\sqrt{\frac{\log(RK/\delta)}{n}} + \min\left\{h, \bar{\eta}RM^3\sqrt{\frac{d}{n}}\right\} + \epsilon\bar{\eta}RM^2[(\bar{\eta}M) \vee 1]\sqrt{\frac{d}{n}}$$
$$\lesssim T^2(\tilde{\kappa}_0 \vee \tilde{\kappa}_0')^{T-1}(r_w^* \vee r_{\boldsymbol{\theta}}^*) + R^2M^3c_w^{-1}\sqrt{\frac{d}{n|S|}} + R^2Mc_w^{-1}[M^2 + \log(Rc_w^{-1})]\sqrt{\frac{1}{n}}$$
$$+ MRc_w^{-1}\sqrt{\frac{\log(RK/\delta)}{n}} + \min\left\{h, R^2M^3c_w^{-1}\sqrt{\frac{d}{n}}\right\} + \epsilon RM^3c_w^{-1}\sqrt{\frac{d}{n}},$$

where $\kappa_0 = 119\sqrt{\frac{2C_b}{1+C_b}} + CM^2c_w^{-2}R^3\exp\{-C'\Delta^2\} + Cc_w^{-2}R^3\exp\{-C'\Delta^2\} + \tilde{\kappa}_0'$, and $\tilde{\kappa}_0'$ satisfies $1 > \tilde{\kappa}_0' > CMR\sqrt{\frac{d}{n}}$ for some $C > 0$.

E.4. Proof of Corollary A.12

By the rate of $\mathcal{W}(n, \frac{\delta}{3RK}, r_{\boldsymbol{\theta},T}^*)$ in Proposition A.11 and the upper bound of $r_{\boldsymbol{\theta},T}^*$ in Proposition A.14,

$$\mathcal{W}\Big(n, \frac{\delta}{3RK}, r_{\boldsymbol{\theta},T}^*\Big) \asymp RMr_{\boldsymbol{\theta},T}^*\sqrt{\frac{d}{n}} + [RM^2 + R\log(Rc_w^{-1})]\sqrt{\frac{1}{n}} + \sqrt{\frac{\log(RK/\delta)}{n}}$$
$$\lesssim T^2(\tilde{\kappa}_0 \vee \tilde{\kappa}_0')^{T-1}(r_w^* \vee r_{\boldsymbol{\theta}}^*) + R^2M^3c_w^{-1}\sqrt{\frac{d}{n|S|}} + R^2Mc_w^{-1}[M^2 + \log(Rc_w^{-1})]\sqrt{\frac{1}{n}}$$
$$+ MRc_w^{-1}\sqrt{\frac{\log(RK/\delta)}{n}} + \min\left\{h, R^2M^3c_w^{-1}\sqrt{\frac{d}{n}}\right\} + \epsilon RM^3c_w^{-1}\sqrt{\frac{d}{n}}.$$

Applying Theorem A.8, we have

$$\max_{k\in S}\max_{r\in[R]}(\|\widehat{\boldsymbol{\theta}}_r^{(k)[T]} - \boldsymbol{\theta}_r^{(k)*}\|_2 + |\widehat{w}_r^{(k)[T]} - w_r^{(k)*}|)$$
$$\leq 20T(\tilde{\kappa}_0)^{T-1}(r_w^* \vee r_{\boldsymbol{\theta}}^*) + \left[\frac{119}{15}\tilde{\kappa}_0(\tilde{\kappa}_0/119)^{T-1} + \frac{118}{119}(T-1)(\tilde{\kappa}_0)^T\right](r_w^* + r_{\boldsymbol{\theta}}^*)$$
$$+ \frac{1}{1-\kappa_0}\left[\bar{\eta}\mathcal{E}_2\Big(n, |S|, \frac{\delta}{3R}\Big) + \mathcal{W}\Big(n, \frac{\delta}{3RK}, r_{\boldsymbol{\theta},J}^*\Big)\right]$$
$$+ \frac{18}{1-\tilde{\kappa}_0/119}\cdot\min\left\{3h, \frac{6}{1-\tilde{\kappa}_0}\left[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r_{\boldsymbol{\theta}}^*\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\right]\right\}$$
$$+ \frac{30}{(1-\tilde{\kappa}_0)(1-\tilde{\kappa}_0/119)}\epsilon\cdot\left[\mathcal{W}\Big(n, \frac{\delta}{3RK}, r_{\boldsymbol{\theta}}^*\Big) + 2\bar{\eta}\mathcal{E}_1\Big(n, \frac{\delta}{3RK}\Big)\right]$$
$$\leq T^2(\tilde{\kappa}_0 \vee \tilde{\kappa}_0')^{T-1}(r_w^* \vee r_{\boldsymbol{\theta}}^*) + R^2M^3c_w^{-1}\sqrt{\frac{d}{n|S|}} + R^2Mc_w^{-1}[M^2 + \log(Rc_w^{-1})]\sqrt{\frac{1}{n}}$$
$$+ MRc_w^{-1}\sqrt{\frac{\log(RK/\delta)}{n}} + \min\left\{h, R^2M^3c_w^{-1}\sqrt{\frac{d}{n}}\right\} + \epsilon RM^3c_w^{-1}\sqrt{\frac{d}{n}}. \tag{12}$$

Note that conditioned on the event $\mathcal{V}$ defined in the proof of Theorem A.8,

$$\eta_r^{(k)} = (1+C_b)^{-1}(\widehat{w}_r^{(k)[0]})^{-1} \lesssim Rc_w^{-1},$$

for all $k \in S$ and $r \in [R]$. Plugging it in equation (12) implies the desired upper bound in Corollary A.12.

E.5. Proof of Proposition A.16

Since this proof is very long, we divide it into several parts.

(I) Part 1: Deriving the expressions of $\mu_r^{(k)}$ and $L_r^{(k)}$.

First, note that

$$q^{(k)}(\boldsymbol{\theta}) = Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)*}, \boldsymbol{w}^{(k)*}) = -\frac{1}{2}\mathbb{E}\left[\sum_{r=1}^{R}\gamma^{(r)}_{\boldsymbol{\theta}^{(k)*},\boldsymbol{w}^{(k)*}}(\boldsymbol{x}^{(k)}, y^{(k)})(y^{(k)} - (\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_r)^2\right].$$

and

$$\begin{aligned}
\widehat{Q}^{(k)}(\boldsymbol{\theta}|\boldsymbol{\theta}', \boldsymbol{w}') &= -\frac{1}{2n}\sum_{i=1}^{n}\sum_{r=1}^{R}\gamma^{(r)}_{\boldsymbol{\theta}',\boldsymbol{w}'}(\boldsymbol{x}_i^{(k)})(y_i^{(k)} - (\boldsymbol{x}_i^{(k)})^T\boldsymbol{\theta}_r)^2,\\
\frac{\partial}{\partial\boldsymbol{\theta}_r}Q^{(k)}(\boldsymbol{\theta}|\boldsymbol{\theta}', \boldsymbol{w}') &= -\mathbb{E}_{\boldsymbol{x}^{(k)}}\left[\gamma^{(r)}_{\boldsymbol{\theta}',\boldsymbol{w}'}(\boldsymbol{x}^{(k)}, y^{(k)})\boldsymbol{x}^{(k)}((\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_r - y^{(k)})\right],\\
\frac{\partial}{\partial\boldsymbol{\theta}_r}\widehat{Q}^{(k)}(\boldsymbol{\theta}|\boldsymbol{\theta}', \boldsymbol{w}') &= -\frac{1}{n}\sum_{i=1}^{n}\gamma^{(r)}_{\boldsymbol{\theta}',\boldsymbol{w}'}(\boldsymbol{x}_i^{(k)}, y_i^{(k)})\boldsymbol{x}_i^{(k)}((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\theta}_r - y_i^{(k)}),\\
\nabla^2_{\boldsymbol{\theta}}q^{(k)}(\boldsymbol{\theta}) &= \mathrm{diag}\Big(\Big\{\mathbb{E}\Big[\gamma^{(r)}_{\boldsymbol{\theta}_r^{(k)*},\boldsymbol{w}_r^{(k)*}}(\boldsymbol{x}^{(k)}, y^{(k)})\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\Big]\Big\}_{r=1}^{R}\Big).
\end{aligned}$$

We have the following lemma.

Lemma E.4. *Under Assumption A.15:*

(i) $\lambda_{\max}\big(\mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{\theta}_r^{(k)*},\boldsymbol{w}_r^{(k)*}}(\boldsymbol{x}^{(k)}, y^{(k)})\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\big]\big) \le w_r^{(k)*} + C\frac{\sqrt{\log\Delta}}{\Delta} := L_r^{(k)}$;

(ii) $\lambda_{\min}\big(\mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{\theta}_r^{(k)*},\boldsymbol{w}_r^{(k)*}}(\boldsymbol{x}^{(k)}, y^{(k)})\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\big]\big) \ge w_r^{(k)*} - CR\frac{\sqrt{\log\Delta}}{\Delta} := \mu_r^{(k)}$.

Now let us prove the lemma.

(i) Note that

$$\gamma^{(r)}_{\boldsymbol{w}^{(k)*},\boldsymbol{\theta}^{(k)*}}(\boldsymbol{x}^{(k)}, y^{(k)}) = \frac{w_r^{(k)*}\exp\{y^{(k)}(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_1^{(k)*}) - \frac{1}{2}[((\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_r^{(k)*})^2 - ((\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_1^{(k)*})^2]\}}{w_1^{(k)*} + \sum_{r=2}^{R}w_r^{(k)*}\exp\{y^{(k)}(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_1^{(k)*}) - \frac{1}{2}[((\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_r^{(k)*})^2 - ((\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_1^{(k)*})^2]\}},$$

where

$$\begin{aligned}
&y^{(k)}(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_1^{(k)*}) - \frac{1}{2}[((\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_r^{(k)*})^2 - ((\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_1^{(k)*})^2]\\
&= [y^{(k)} - (\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_1^{(k)*}]\cdot(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_1^{(k)*}) + [(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_1^{(k)*})]\Big[(\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_1^{(k)*} - \frac{1}{2}((\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_r^{(k)*}\\
&\qquad + (\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_1^{(k)*})\Big]\\
&= [y^{(k)} - (\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_1^{(k)*}]\cdot(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_1^{(k)*}) - \frac{1}{2}[(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_1^{(k)*})]^2.
\end{aligned}$$

Conditioned on the event $\{z^{(k)} = 1\}$, we have $[y^{(k)} - (\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_1^{(k)*}]\cdot(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_1^{(k)*}) \sim N(\mathbf{0}, [(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_1^{(k)*})]^2)$. Define events

$$\begin{aligned}
\mathcal{V}_1 &= \big\{|[y^{(k)} - (\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_1^{(k)*}]\cdot(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_1^{(k)*})| \le \tau_1|(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_1^{(k)*})|\big\},\\
\mathcal{V}_2 &= \big\{|(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_1^{(k)*})| \ge \tau_2\|\boldsymbol{\theta}_r^{(k)*} - \boldsymbol{\theta}_1^{(k)*}\|_2\big\},
\end{aligned}$$

then by the tail bounds for Gaussian variables and the boundedness of Gaussian density, we have $\mathbb{P}(\mathcal{V}_1^c) \lesssim \exp\{-C\tau_1^2\}$, $\mathbb{P}(\mathcal{V}_2^c) \lesssim \tau_2$. Therefore,

$$\lambda_{\max}\Big(\mathbb{E}\Big[\gamma^{(r)}_{\boldsymbol{\theta}_r^{(k)*},\boldsymbol{w}_r^{(k)*}}(\boldsymbol{x}^{(k)}, y^{(k)})\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\Big|z^{(k)} = 1\Big]\Big)$$

$$\lesssim \underbrace{\sup_{\|\boldsymbol{u}\|_2=1} \mathbb{E}\Big[\gamma^{(r)}_{\boldsymbol{\theta}_r^{(k)*},\boldsymbol{w}_r^{(k)*}}(\boldsymbol{x}^{(k)},y^{(k)})((\boldsymbol{x}^{(k)})^T\boldsymbol{u})^2\Big|z^{(k)}=1\Big]\mathbb{P}(\mathcal{V}_1\cap\mathcal{V}_2)}_{[1]} + \underbrace{\sup_{\|\boldsymbol{u}\|_2=1}\sqrt{\mathbb{E}[((\boldsymbol{x}^{(k)})^T\boldsymbol{u})^4|z^{(k)}=1]}\sqrt{\mathbb{P}(\mathcal{V}_1^c)}}_{[2]}$$
$$+ \underbrace{\sup_{\|\boldsymbol{u}\|_2=1} \mathbb{E}\Big[\gamma^{(r)}_{\boldsymbol{\theta}_r^{(k)*},\boldsymbol{w}_r^{(k)*}}(\boldsymbol{x}^{(k)},y^{(k)})((\boldsymbol{x}^{(k)})^T\boldsymbol{u})^2\mathbb{1}(\mathcal{V}_2^c)\Big|z^{(k)}=1\Big]}_{[3]},$$

where

$$\begin{aligned}
[1] &\lesssim \frac{w_r^{(k)*}}{w_1^{(k)*}}\cdot\mathbb{E}\bigg[\exp\Big\{\tau_1|(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_r^{(k)*}-\boldsymbol{\theta}_1^{(k)*})| - \frac{1}{2}|(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_r^{(k)*}-\boldsymbol{\theta}_1^{(k)*})|^2\Big|z^{(k)}=1,\mathcal{V}_1\cap\mathcal{V}_2\Big\}\bigg]\\
&\lesssim \frac{w_r^{(k)*}}{w_1^{(k)*}}\cdot\exp\Big\{\tau_1\tau_2\|\boldsymbol{\theta}_r^{(k)*}-\boldsymbol{\theta}_1^{(k)*}\|_2 - \frac{1}{2}\tau_2^2\|\boldsymbol{\theta}_r^{(k)*}-\boldsymbol{\theta}_1^{(k)*}\|_2^2\Big\},\\
[2] &\lesssim \exp\{-C\tau_1^2\},\\
[3] &\lesssim \sup_{\|\boldsymbol{u}\|_2=1}\mathbb{E}[((\boldsymbol{x}^{(k)})^T\boldsymbol{u})^2|\mathcal{V}_2]\cdot\mathbb{P}(\mathcal{V}_2)\lesssim\mathbb{P}(\mathcal{V}_2)\lesssim\tau_2,.
\end{aligned}\tag{13}$$

The second inequality in (13) holds due to Lemma A.1 in Kwon & Caramanis (2020b). Let $\tau_1 = c\sqrt{\log\|\boldsymbol{\theta}_r^{(k)*}-\boldsymbol{\theta}_1^{(k)*}\|_2}$, $\tau_2 = 3C\frac{\sqrt{\log\|\boldsymbol{\theta}_r^{(k)*}-\boldsymbol{\theta}_1^{(k)*}\|_2}}{\|\boldsymbol{\theta}_r^{(k)*}-\boldsymbol{\theta}_1^{(k)*}\|_2}$ with some constant $c>0$. Note that $\|\boldsymbol{\theta}_r^{(k)*}-\boldsymbol{\theta}_1^{(k)*}\|_2\le\tau_1/\tau_2 = \|\boldsymbol{\theta}_r^{(k)*}-\boldsymbol{\theta}_1^{(k)*}\|_2/3$. Then

$$\lambda_{\max}\Big(\mathbb{E}\Big[\gamma^{(r)}_{\boldsymbol{\theta}_r^{(k)*},\boldsymbol{w}_r^{(k)*}}(\boldsymbol{x}^{(k)},y^{(k)})\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\Big|z^{(k)}=1\Big]\Big)\le[1]+[2]+[3]\lesssim\frac{w_r^{(k)*}}{w_1^{(k)*}}\frac{1}{\Delta}+\frac{\sqrt{\log\Delta}}{\Delta}.$$

Similarly, the same bound holds for $\lambda_{\max}(\mathbb{E}[\gamma^{(r)}_{\boldsymbol{\theta}_r^{(k)*},\boldsymbol{w}_r^{(k)*}}(\boldsymbol{x}^{(k)},y^{(k)})\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T|z^{(k)}=r'])$ with all $r'\neq r$. In addition, we can rewrite $\gamma^{(r)}_{\boldsymbol{w}^{(k)*},\boldsymbol{\theta}^{(k)*}}(\boldsymbol{x}^{(k)},y^{(k)})$ as

$$\gamma^{(r)}_{\boldsymbol{w}^{(k)*},\boldsymbol{\theta}^{(k)*}}(\boldsymbol{x}^{(k)},y^{(k)}) = \frac{w_r^{(k)*}}{w_r^{(k)*}+\sum_{r'\neq r}w_{r'}^{(k)*}\exp\{y^{(k)}(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_{r'}^{(k)*}-\boldsymbol{\theta}_1^{(k)*})-\frac{1}{2}[((\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_{r'}^{(k)*})^2-((\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_1^{(k)*})^2]\}}$$
$$\le 1,$$

which implies that

$$\lambda_{\max}\Big(\mathbb{E}\Big[\gamma^{(r)}_{\boldsymbol{\theta}_r^{(k)*},\boldsymbol{w}_r^{(k)*}}(\boldsymbol{x}^{(k)},y^{(k)})\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\Big|z^{(k)}=r\Big]\Big)\le 1.$$

Hence by the convexity of maximum eigenvalues,

$$\begin{aligned}
\lambda_{\max}\Big(\mathbb{E}\Big[\gamma^{(r)}_{\boldsymbol{\theta}_r^{(k)*},\boldsymbol{w}_r^{(k)*}}(\boldsymbol{x}^{(k)},y^{(k)})\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\Big]\Big) &\le \sum_{r'=1}^R w_{r'}^{(k)*}\cdot\lambda_{\max}\Big(\mathbb{E}\Big[\gamma^{(r)}_{\boldsymbol{\theta}_r^{(k)*},\boldsymbol{w}_r^{(k)*}}(\boldsymbol{x}^{(k)},y^{(k)})\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\Big|z^{(k)}=r'\Big]\Big)\\
&\le w_r^{(k)*}\Big(1+\frac{C}{\Delta}\Big)+C\frac{\sqrt{\log\Delta}}{\Delta}\\
&\le w_r^{(k)*}+C'\frac{\sqrt{\log\Delta}}{\Delta}.
\end{aligned}$$

(ii) We have

$$\begin{aligned}
\lambda_{\min}\Big(\mathbb{E}\Big[\gamma^{(r)}_{\boldsymbol{\theta}_r^{(k)*},\boldsymbol{w}_r^{(k)*}}(\boldsymbol{x}^{(k)},y^{(k)})\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\Big]\Big) &\ge w_r^{(k)*}\lambda_{\min}\Big(\mathbb{E}\Big[\gamma^{(r)}_{\boldsymbol{\theta}_r^{(k)*},\boldsymbol{w}_r^{(k)*}}(\boldsymbol{x}^{(k)},y^{(k)})\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\Big|z^{(k)}=r\Big]\Big)\\
&\ge w_r^{(k)*}\lambda_{\min}\big(\mathbb{E}\big[\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\big|z^{(k)}=r\big]\big)
\end{aligned}$$

$$- w_r^{(k)*}\lambda_{\min}\Big(\mathbb{E}\Big[(1-\gamma^{(r)}_{\boldsymbol{\theta}_r^{(k)*},\boldsymbol{w}_r^{(k)*}}(\boldsymbol{x}^{(k)},y^{(k)}))\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\Big|z^{(k)}=r\Big]\Big).$$

Similar to (i), it is straightforward to show that

$$\lambda_{\max}\Bigg(\mathbb{E}\Bigg[\frac{w_{r'}^{(k)*}\exp\{y^{(k)}(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_{r'}^{(k)*}-\boldsymbol{\theta}_1^{(k)*})-\frac{1}{2}[((\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_{r'}^{(k)*})^2-((\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_1^{(k)*})^2]\}}{w_r^{(k)*}+\sum_{r'\neq r}w_{r'}^{(k)*}\exp\{y^{(k)}(\boldsymbol{x}^{(k)})^T(\boldsymbol{\theta}_{r'}^{(k)*}-\boldsymbol{\theta}_1^{(k)*})-\frac{1}{2}[((\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_{r'}^{(k)*})^2-((\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_1^{(k)*})^2]\}}\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T$$

$$\Big|z^{(k)}=r\Bigg]\Bigg)\lesssim\frac{w_{r'}^{(k)*}}{w_r^{(k)*}}\frac{1}{\Delta}+\frac{\sqrt{\log\Delta}}{\Delta},$$

for any $r'\neq r$. Hence

$$\lambda_{\min}\Big(\mathbb{E}\Big[\gamma^{(r)}_{\boldsymbol{\theta}_r^{(k)*},\boldsymbol{w}_r^{(k)*}}(\boldsymbol{x}^{(k)},y^{(k)})\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\Big]\Big)\geq w_r^{(k)*}-C\sum_{r'\neq r}\left(w_{r'}^{(k)*}\frac{1}{\Delta}+w_r^{(k)*}\frac{\sqrt{\log\Delta}}{\Delta}\right)\geq w_r^{(k)*}-CR\frac{\sqrt{\log\Delta}}{\Delta},$$

which completes the proof of Lemma E.4.

(II) Part 2: Deriving the rate of κ in Assumption 2.(i).

Since Assumption A.3 is assumed to hold for all $k\in[K]$, in this part, for notation simplicity, we drop the task index k in the superscript and write $\boldsymbol{w}^{(k)}=\{w_r^{(k)}\}_{r=1}^R$, $\boldsymbol{\theta}^{(k)}=\{\boldsymbol{\theta}_r^{(k)}\}_{r=1}^R$, $\boldsymbol{w}^{(k)*}=\{w_r^{(k)*}\}_{r=1}^R$, $\boldsymbol{\theta}^{(k)*}=\{\boldsymbol{\theta}_r^{(k)*}\}_{r=1}^R$, $\boldsymbol{x}^{(k)}$, and $y^{(k)}$ simply as $\boldsymbol{w}=\{w_r\}_{r=1}^R$, $\boldsymbol{\theta}=\{\boldsymbol{\theta}_r\}_{r=1}^R$, $\boldsymbol{w}^*=\{w_r^*\}_{r=1}^R$, $\boldsymbol{\theta}^{(k)}=\{\boldsymbol{\theta}_r^*\}_{r=1}^R$, $\boldsymbol{x}$, and y.

By Taylor expansion:

$$\mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x},y)-\gamma^{(r)}_{\boldsymbol{\theta}^*,\boldsymbol{w}^*}(\boldsymbol{x},y)\big]=\sum_{r'=1}^R\mathbb{E}\left[\frac{\partial\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x},y)}{\partial w_{r'}}\Big|_{w_{r'}=\widetilde{w}_{r'}}(w_{r'}-w_{r'}^*)\right]+\sum_{r'=1}^R\mathbb{E}\left[\left(\frac{\partial\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x},y)}{\partial\boldsymbol{\theta}_{r'}}\Big|_{\boldsymbol{\theta}_{r'}=\widetilde{\boldsymbol{\theta}}_{r'}}\right)^T(\boldsymbol{\theta}_{r'}-\boldsymbol{\theta}_{r'}^*)\right],$$

where $\widetilde{w}_{r'}$ is at the line segment between $w_{r'}$ and w_r^*, and $\widetilde{\boldsymbol{\theta}}_{r'}$ is at the line segment between $\boldsymbol{\theta}_{r'}$ and $\boldsymbol{\theta}_r^*$. And

$$\frac{\partial\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x},y)}{\partial w_{r'}}=\begin{cases}\frac{\exp\{y\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)-\frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_r)^2-(\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]\}(w_1+\sum_{r'\neq r}w_{r'}\exp\{y\boldsymbol{x}^T(\boldsymbol{\theta}_{r'}-\boldsymbol{\theta}_1)-\frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_{r'})^2-(\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]\})}{(w_1+\sum_{r'=1}^R w_{r'}\exp\{y\boldsymbol{x}^T(\boldsymbol{\theta}_{r'}-\boldsymbol{\theta}_1)-\frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_{r'})^2-(\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]\})^2}, & r'=r;\\ -\frac{w_r\exp\{y\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)-\frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_r)^2-(\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]\}\cdot\exp\{y\boldsymbol{x}^T(\boldsymbol{\theta}_{r'}-\boldsymbol{\theta}_1)-\frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_{r'})^2-(\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]\}}{(w_1+\sum_{r'=1}^R w_{r'}\exp\{y\boldsymbol{x}^T(\boldsymbol{\theta}_{r'}-\boldsymbol{\theta}_1)-\frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_{r'})^2-(\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]\})^2}, & r'\neq r.\end{cases}$$

We want to upper bound $\mathbb{E}\Big[\frac{\partial\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x},y)}{\partial w_r}\Big|z=\tilde{r}\Big]$ for all $\tilde{r}$. Since in the expression of $\mathbb{E}\Big[\frac{\partial\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x},y)}{\partial w_r}\Big|z=\tilde{r}\Big]$, $\{w_{r'}\}_{r'\neq r}$ and $\{\boldsymbol{\theta}_{r'}\}_{r'\neq r}$ are symmetric, i.e., any class can be the reference class. WLOG, we only show how to bound $\mathbb{E}\Big[\frac{\partial\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x},y)}{\partial w_r}\Big|z=1\Big]$ and the same arguments can be used to bound $\mathbb{E}\Big[\frac{\partial\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x},y)}{\partial w_r}\Big|z=\tilde{r}\Big]$ for $\tilde{r}\neq 1$.

Denote $(*)=y^{(1)}\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)-\frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_r)^2-(\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]$, $\widetilde{z}=(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_1^*)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)|\boldsymbol{x}\sim N(0,[\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)]^2)$, where $y^{(1)}\overset{d}{=}(y|z=1)$. Then

$$(*)=\widetilde{z}+(\boldsymbol{x}^T\boldsymbol{\theta}_1^*)\cdot x^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)-\frac{1}{2}x^T(\boldsymbol{\theta}_r+\boldsymbol{\theta}_1)\cdot x^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1),$$

where

$$\begin{aligned}&(\boldsymbol{x}^T\boldsymbol{\theta}_1^*)\cdot x^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)-\frac{1}{2}x^T(\boldsymbol{\theta}_r+\boldsymbol{\theta}_1)\cdot x^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)\\&=\big[x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)+x^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1+\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^*)\big]\left[-\frac{1}{2}x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)+-\frac{1}{2}x^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)+\frac{1}{2}x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)\right]\\&=-\frac{1}{2}[x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)]^2+\frac{1}{2}\{x^T[(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)+(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)]\}^2.\end{aligned}$$

Define events

$$\mathcal{V}_1 = \left\{ |x^T(\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_1)| \vee |x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_r)| \leq \frac{1}{4}|x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)| \right\},$$
$$\mathcal{V}_2 = \{|\widetilde{z}| \leq \tau_1 |x^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1)|\},$$
$$\mathcal{V}_3 = \{|\boldsymbol{x}^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)| > \tau_2 |\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*|\}.$$

We know that

$$\begin{aligned}
\mathbb{P}(\mathcal{V}_1^c) &\leq \mathbb{P}\bigg(|x^T(\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_1)| \vee |x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_r)| > \frac{1}{4}|x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)| \bigg) \\
&\leq \mathbb{P}\bigg(|x^T(\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_1)| > \frac{1}{4}|x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)| \bigg) + \mathbb{P}\bigg(|x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_r)| > \frac{1}{4}|x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)| \bigg) \\
&\leq 4\bigg(\frac{\|\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_1\|_2}{\|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2} + \frac{\|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_r\|_2}{\|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2} \bigg) \\
&\leq 8C_b,
\end{aligned}$$

where we applied Lemma A.1 in Kwon & Caramanis (2020b) to get the second last inequality. And

$$\mathbb{P}(\mathcal{V}_2^c) \leq C\exp\{-C'\tau_1^2\}, \quad \mathbb{P}(\mathcal{V}_3^c) \leq C\tau_2.$$

Given $\mathcal{V}_1 \cap \mathcal{V}_2 \cap \mathcal{V}_3$, we have

$$-\frac{1}{2}[x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)]^2 + \frac{1}{2}\{x^T[(\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_1) + (\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_r)]\}^2 \leq -\frac{3}{8}[x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)]^2,$$

leading to

$$\widetilde{z} \leq \tau_1 |x^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1)| \leq \tau_1(|x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)| + |x^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^*)| + |x^T(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_1^*)|) \leq \tau_1 \cdot \frac{3}{2}|x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)|.$$

Hence

$$\begin{aligned}
\left| \mathbb{E}\left[\frac{\partial \gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}, y)}{\partial w_r} \middle| z = 1 \right] \right| &\leq \left| \mathbb{E}\left[\frac{\partial \gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}, y)}{\partial w_r} \middle| z = 1, \mathcal{V}_1 \cap \mathcal{V}_2 \cap \mathcal{V}_3 \right] \mathbb{P}(\mathcal{V}_1 \cap \mathcal{V}_2 \cap \mathcal{V}_3) \right| + \frac{R}{c_w}[\mathbb{P}(\mathcal{V}_1^c) + \mathbb{P}(\mathcal{V}_2^c) + \mathbb{P}(\mathcal{V}_3^c)] \\
&\leq \frac{R}{c_w} \exp\left\{ \frac{3}{2}\tau_1 |x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)| - \frac{3}{8}[x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)]^2 \right\} + C\frac{R}{c_w}\big(C_b + \exp\{-C'\tau_1^2\} + \tau_2\big) \\
&\leq \frac{R}{c_w} \exp\left\{ \frac{3}{2}\tau_1\tau_2 \|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2 - \frac{3}{8}\tau_2^2 \|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2^2 \right\} + C\frac{R}{c_w}\big(C_b + \exp\{-C'\tau_1^2\} + \tau_2\big),
\end{aligned}$$

where the last inequality requires $|x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)| \geq \tau_2 \|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2 \geq 2\tau_1$. Let $\tau_1 = c\sqrt{\log \|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2}$ and $\tau_2 = 10c\frac{\sqrt{\log \|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2}}{\|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2}$ with some constant $c > 0$, then

$$\left| \mathbb{E}\left[\frac{\partial \gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}, y)}{\partial w_r} \middle| z = 1 \right] \right| \lesssim \frac{R}{c_w}\left(C_b + \frac{\sqrt{\log \|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2}}{\|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2} \right) \lesssim \frac{R}{c_w}\left(C_b + \frac{\sqrt{\log \Delta}}{\Delta} \right).$$

Similarly,

$$\left| \mathbb{E}\left[\frac{\partial \gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}, y)}{\partial w_{r'}} \middle| z = \widetilde{r} \right] \right| \lesssim \frac{R}{c_w}\left(C_b + \frac{\sqrt{\log \Delta}}{\Delta} \right),$$

for any r, r', and $\widetilde{r}$. Therefore,

$$\sum_{r'=1}^{R} \mathbb{E}\left[\frac{\partial \gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}, y)}{\partial w_{r'}}\bigg|_{w_{r'} = \widetilde{w}_{r'}} (w_{r'} - w_{r'}^*) \right] \leq C\frac{R}{c_w}\left(C_b + \frac{\sqrt{\log \Delta}}{\Delta} \right) \sum_{r=1}^{R} |w_r - w_r^*|.$$

On the other hand,

$$\frac{\partial \gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y)}{\partial \boldsymbol{\theta}_{r'}} = \begin{cases} \gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y)\boldsymbol{x}(y-\boldsymbol{x}^T\boldsymbol{\theta}_r) - \big(\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y)\big)^2\boldsymbol{x}(y-\boldsymbol{x}^T\boldsymbol{\theta}_r), & r'=r; \\ -\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y)\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r')}(\boldsymbol{x},y)\boldsymbol{x}(y-\boldsymbol{x}^T\boldsymbol{\theta}_{r'}), & r'\neq r, \end{cases} \tag{14}$$

where

$$\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y) = \frac{w_r \exp\{y\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1) - \frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_r)^2 - (\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]\}}{w_1 + \sum_{r'=2}^R w_{r'}\exp\{y\boldsymbol{x}^T(\boldsymbol{\theta}_{r'}-\boldsymbol{\theta}_1) - \frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_{r'})^2 - (\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]\}},$$

We will show how to upper bound $\mathbb{E}[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y)(y-\boldsymbol{x}^T\boldsymbol{\theta}_r)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}'_r)]$, and the same arguments can be used to bound $\mathbb{E}[(\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y))^2(y-\boldsymbol{x}^T\boldsymbol{\theta}_r)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}'_r)]$. Then we will have an upper bound for $\mathbb{E}\Big[\big(\frac{\partial\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x},y)}{\partial\boldsymbol{\theta}_r}\big|_{\boldsymbol{\theta}_r=\widetilde{\boldsymbol{\theta}}_r}\big)^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}'_r)\Big]$, and the analysis is the same for $\mathbb{E}\Big[\big(\frac{\partial\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x},y)}{\partial\boldsymbol{\theta}_{r'}}\big|_{\boldsymbol{\theta}_{r'}=\widetilde{\boldsymbol{\theta}}_{r'}}\big)^T(\boldsymbol{\theta}_{r'}-\boldsymbol{\theta}^*_{r'})\Big]$ with $r'\neq r$.

Let us start from $\mathbb{E}[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y)(y-\boldsymbol{x}^T\boldsymbol{\theta}_r)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}'_r)|z=1]$. Consider $y^{(1)}\overset{d}{=}(y|z=1)$ and

$$y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_r = (y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}^*_1) + \boldsymbol{x}^T(\boldsymbol{\theta}^*_1-\boldsymbol{\theta}^*_r) + \boldsymbol{x}^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}_r).$$

Define events

$$\begin{aligned}
\mathcal{V}_1 &= \left\{|x^T(\boldsymbol{\theta}^*_1-\boldsymbol{\theta}_1)| \vee |x^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}_r)| \le \frac{1}{4}|x^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)|\right\},\\
\mathcal{V}_2 &= \{|y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}^*_1|\le\tau_1\},\\
\mathcal{V}_3 &= \{|\boldsymbol{x}^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)| > \tau_2|\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1|\}.
\end{aligned}$$

We know that

$$\begin{aligned}
\mathbb{P}(\mathcal{V}_1^c) &\le \mathbb{P}\left(|x^T(\boldsymbol{\theta}^*_1-\boldsymbol{\theta}_1)| \vee |x^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}_r)| > \frac{1}{4}|x^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)|\right)\\
&\le \mathbb{P}\left(|x^T(\boldsymbol{\theta}^*_1-\boldsymbol{\theta}_1)| > \frac{1}{4}|x^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)|\right) + \mathbb{P}\left(|x^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}_r)| > \frac{1}{4}|x^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)|\right)\\
&\le 4\left(\frac{\|\boldsymbol{\theta}^*_1-\boldsymbol{\theta}_1\|_2}{\|\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1\|_2} + \frac{\|\boldsymbol{\theta}^*_r-\boldsymbol{\theta}_r\|_2}{\|\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1\|_2}\right)\\
&\le 8C_b,
\end{aligned}$$

where we applied Lemma A.1 in Kwon & Caramanis (2020b) to get the second last inequality. And

$$\mathbb{P}(\mathcal{V}_2^c)\le C\exp\{-C'\tau_1^2\},\quad \mathbb{P}(\mathcal{V}_3^c)\le C\tau_2.$$

Note that

$$\begin{aligned}
&\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y)(y-\boldsymbol{x}^T\boldsymbol{\theta}_r)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}'_r)|z=1\big]\\
&\le \underbrace{\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y)(y-\boldsymbol{x}^T\boldsymbol{\theta}_r)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}'_r)|z=1,\mathcal{V}_1\cap\mathcal{V}_2\cap\mathcal{V}_3\big]\mathbb{P}(\mathcal{V}_1\cap\mathcal{V}_2\cap\mathcal{V}_3)}_{[1]}\\
&\quad + \underbrace{\mathbb{E}\big[\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}'_r)(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_r)\mathbb{1}(\mathcal{V}_1^c\cup\mathcal{V}_2^c\cup\mathcal{V}_3^c)\big]}_{[2]}.
\end{aligned}$$

(a) Case 1: $\max_r\|\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r\|_2\le 1$.

Note that for term [2], by Lemma A.1 in Kwon & Caramanis (2020b),

$$\mathbb{E}\big[\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}'_r)(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}^*_1)\mathbb{1}(\mathcal{V}_1^c)\big]$$

$$
\begin{aligned}
&\le \mathbb{E}\bigg[\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r')(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_1^*)\mathbb{1}(\mathcal{V}_1^c)\bigg||x^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)|>\frac{1}{4}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|\bigg]\mathbb{P}\bigg(|x^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)|>\frac{1}{4}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|\bigg)\\
&\quad+\mathbb{E}\bigg[\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r')(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_1^*)\mathbb{1}(\mathcal{V}_1^c)\bigg||x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)|>\frac{1}{4}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|\bigg]\mathbb{P}\bigg(|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)|>\frac{1}{4}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|\bigg)\\
&\le \sqrt{\mathbb{E}\bigg[(\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'))^2\bigg||x^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)|>\frac{1}{4}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|\bigg]}\cdot\sqrt{\mathbb{E}(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_1^*)^2}\\
&\quad\cdot\mathbb{P}\bigg(|x^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)|>\frac{1}{4}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|\bigg)\\
&\quad+\sqrt{\mathbb{E}\bigg[(\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'))^2\bigg||x^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)|>\frac{1}{4}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|\bigg]}\cdot\sqrt{\mathbb{E}(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_1^*)^2}\\
&\quad\cdot\mathbb{P}\bigg(|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)|>\frac{1}{4}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|\bigg)\\
&\lesssim C_b\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2.
\end{aligned}
$$

And by Cauchy-Schwarz inequality,

$$
\begin{aligned}
&\mathbb{E}\big[\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r')(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_1^*)\mathbb{1}(\mathcal{V}_2^c)\big]\le\sqrt{\mathbb{E}\big[(\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'))^2(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_1^*)^2\big]}\sqrt{\mathbb{P}(\mathcal{V}_2^c)}\lesssim\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2\cdot\exp\{-C\tau_1^2\},\\
&\mathbb{E}\big[\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r')(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_1^*)\mathbb{1}(\mathcal{V}_3^c)\big]\le\sqrt{\mathbb{E}\big[(\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'))^2\big]\cdot\mathbb{P}(\mathcal{V}_2^c)}\sqrt{\mathbb{E}\big[(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_1^*)^2\big]\cdot\mathbb{P}(\mathcal{V}_2^c)}\lesssim\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2\cdot\tau_2.
\end{aligned}
$$

Combine them together:

$$
\mathbb{E}\big[\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r')(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_1^*)\mathbb{1}(\mathcal{V}_1^c\cup\mathcal{V}_2^c\cup\mathcal{V}_3^c)\big]\lesssim(C_b+\exp\{-C\tau_1^2\}+\tau_2)\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2.
$$

Furthermore,

$$
\begin{aligned}
\mathbb{E}\big[\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r')\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_r^*)\mathbb{1}(\mathcal{V}_1^c)\big]&\le\sqrt{\mathbb{E}\big[(\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'))^2|\mathcal{V}_1^c\big]\cdot\mathbb{P}(\mathcal{V}_1^c)}\cdot\sqrt{\mathbb{E}\big[(\boldsymbol{x}^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_r^*))^2|\mathcal{V}_1^c\big]\cdot\mathbb{P}(\mathcal{V}_1^c)}\\
&\lesssim\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2\cdot C_b\cdot\sqrt{\mathbb{E}\big[|x^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)|^2\vee|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)|^2|\mathcal{V}_1^c\big]}\\
&\lesssim C_b\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2,
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}\big[\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r')\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_r^*)\mathbb{1}(\mathcal{V}_2^c)\big]&\lesssim\exp\{-C\tau_1^2\}\cdot\|\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_r^*\|_2\cdot\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2,\\
\mathbb{E}\big[\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r')\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_r^*)\mathbb{1}(\mathcal{V}_3^c)\big]&\lesssim\sqrt{\mathbb{E}\big[(\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'))^2|\mathcal{V}_3^c\big]\cdot\mathbb{P}(\mathcal{V}_3^c)}\cdot\sqrt{\mathbb{E}\big[(\boldsymbol{x}^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_r^*))^2|\mathcal{V}_3^c\big]\cdot\mathbb{P}(\mathcal{V}_3^c)}\\
&\lesssim\tau_2^2\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2\cdot\|\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_r^*\|_2.
\end{aligned}
$$

Therefore

$$
\mathbb{E}\big[\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r')\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_r^*)\mathbb{1}(\mathcal{V}_1^c\cup\mathcal{V}_2^c\cup\mathcal{V}_3^c)\big]\lesssim(C_b+\exp\{-C\tau_1^2\}+\tau_2^2\|\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_r^*\|_2)\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2.
$$

Similarly,

$$
\begin{aligned}
\mathbb{E}\big[\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r')\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)\mathbb{1}(\mathcal{V}_1^c\cup\mathcal{V}_2^c\cup\mathcal{V}_3^c)\big]&\lesssim(C_b+\exp\{-C\tau_1^2\}+\tau_2^2\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r\|_2)\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2\\
&\lesssim(C_b+\exp\{-C\tau_1^2\}+\tau_2^2)\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2.
\end{aligned}
$$

Therefore we have

$$
[2]\lesssim(C_b+\exp\{-C\tau_1^2\}+\tau_2+\tau_2^2\|\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_r^*\|_2)\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2.
$$

For [1], given $\mathcal{V}_1\cap\mathcal{V}_2\cap\mathcal{V}_3$, we know that

$$
-\frac{1}{2}[x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)]^2+\frac{1}{2}\{x^T[(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)+(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)]\}^2\le-\frac{3}{8}[x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)]^2,
$$

leading to $(y^{(1)} - \boldsymbol{x}^T\boldsymbol{\theta}_1^*) + \boldsymbol{x}^T(\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_r^*)$

$$(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_1^*)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1) \le \tau_1|x^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)| \le \tau_1(|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|+|x^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^*)|+|x^T(\boldsymbol{\theta}_1-\boldsymbol{\theta}_1^*)|) \le \tau_1\cdot\frac{3}{2}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|.$$

Similar to the previous analysis, we can show that

$$\begin{aligned}[1] &\lesssim \frac{R}{c_w}\exp\left\{\frac{3}{2}\tau_1|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)| - \frac{3}{8}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|^2\right\}(\tau_1+1)\cdot\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2 \\ &\lesssim \frac{R}{c_w}\exp\left\{\frac{3}{2}\tau_1\tau_2\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2 - \frac{3}{8}\tau_2^2\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2^2\right\}(\tau_1+1)\cdot\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2.\end{aligned}$$

This implies that

$$\begin{aligned}&\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y)(y-\boldsymbol{x}^T\boldsymbol{\theta}_r)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r')|z=1\big] \\ &\le [1]+[2] \\ &\lesssim \left[\frac{R}{c_w}\exp\left\{\frac{3}{2}\tau_1\tau_2\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2 - \frac{3}{8}\tau_2^2\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2^2\right\}(\tau_1+1) + C_b + \exp\{-C\tau_1^2\} + \tau_2 + \tau_2^2\|\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_r^*\|_2\right]\cdot\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2.\end{aligned} \tag{15}$$

Let $\tau_1 = c\sqrt{\log\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2}$ and $\tau_2 = 10c\frac{\sqrt{\log\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2}}{\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2}$ with some constant $c>0$. Then

$$\text{RHS of (15)} \lesssim \left(\frac{R}{c_w}\frac{1}{\Delta}\sqrt{\log\Delta} + \frac{R}{c_w}C_b\right)\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2.$$

Similarly, we can show that $\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y)(y-\boldsymbol{x}^T\boldsymbol{\theta}_r)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r')|z=r'\big]$ has the same upper bound. Therefore

$$\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y)(y-\boldsymbol{x}^T\boldsymbol{\theta}_r)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r')\big] \lesssim \left(\frac{R}{c_w}\frac{1}{\Delta}\sqrt{\log\Delta} + \frac{R}{c_w}C_b\right)\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2.$$

Similarly, following the same arguments, it can be shown that

$$\mathbb{E}\big[(\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y))^2(y-\boldsymbol{x}^T\boldsymbol{\theta}_r)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r')\big] \lesssim \left(\frac{R}{c_w}\frac{1}{\Delta}\sqrt{\log\Delta} + \frac{R}{c_w}C_b\right)\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2.$$

Hence by (14),

$$\left|\mathbb{E}\left[\left(\frac{\partial\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x},y)}{\partial\boldsymbol{\theta}_r}\Big|_{\boldsymbol{\theta}_r=\widetilde{\boldsymbol{\theta}}_r}\right)^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^*)\right]\right| \lesssim \left(\frac{R}{c_w}\frac{1}{\Delta}\sqrt{\log\Delta} + \frac{R}{c_w}C_b\right)\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r'\|_2.$$

(b) Case 2: $\max_r\|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^*\|_2 > 1$.

Suppose $r_0\in[R]$ satisfies $\|\boldsymbol{\theta}_{r_0}-\boldsymbol{\theta}_{r_0}^*\|_2>1$. For $r'\neq r$, we have

$$\big|\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y)-\gamma_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}^{(r)}(\boldsymbol{x},y)\big|z=r'\big]\big| \le \big|\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y)\big|z=r'\big]\big| + \big|\mathbb{E}\big[\gamma_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}^{(r)}(\boldsymbol{x},y)\big|z=r'\big]\big|.$$

WLOG, let us consider the case $r'=1\neq r$, and the other cases can be similarly discussed. We have

$$\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x},y)\big|z=1\big] = \mathbb{E}\left[\frac{w_r\exp\{y\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)-\frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_r)^2-(\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]\}}{w_1+\sum_{r'=2}^R w_{r'}\exp\{y\boldsymbol{x}^T(\boldsymbol{\theta}_{r'}-\boldsymbol{\theta}_1)-\frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_{r'})^2-(\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]\}}\right].$$

Recall that

$$\mathcal{V}_1 = \left\{|x^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)| \vee |x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)| \le \frac{1}{4}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|\right\},$$

$$\mathcal{V}_2 = \{|y^{(1)} - \boldsymbol{x}^T\boldsymbol{\theta}_1^*| \le \tau_1\},$$
$$\mathcal{V}_3 = \{|\boldsymbol{x}^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)| > \tau_2|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*|\}.$$

and

$$\mathbb{P}(\mathcal{V}_1^c) \le 8C_b, \quad \mathbb{P}(\mathcal{V}_2^c) \le C\exp\{-C'\tau_1^2\}, \quad \mathbb{P}(\mathcal{V}_3^c) \le C\tau_2.$$

Similar to the previous analysis,

$$y^{(1)}\boldsymbol{x}^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1) - \frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_r)^2 - (\boldsymbol{x}^T\boldsymbol{\theta}_1)^2] = (y^{(1)} - \boldsymbol{x}^T\boldsymbol{\theta}_1^*)\boldsymbol{x}^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1) + (\boldsymbol{x}^T\boldsymbol{\theta}_1^*)\cdot x^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1) - \frac{1}{2}x^T(\boldsymbol{\theta}_r + \boldsymbol{\theta}_1)\cdot x^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1).$$

Conditioned on $\mathcal{V}_1 \cap \mathcal{V}_2 \cap \mathcal{V}_3$,

$$\begin{aligned}
&(\boldsymbol{x}^T\boldsymbol{\theta}_1^*)\cdot x^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1) - \frac{1}{2}x^T(\boldsymbol{\theta}_r + \boldsymbol{\theta}_1)\cdot x^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1)\\
&= \big[x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*) + x^T(\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_1 + \boldsymbol{\theta}_r - \boldsymbol{\theta}_r^*)\big]\bigg[-\frac{1}{2}x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*) + -\frac{1}{2}x^T(\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_1) + \frac{1}{2}x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_r)\bigg]\\
&= -\frac{1}{2}[x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)]^2 + \frac{1}{2}\{x^T[(\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_1) + (\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_r)]\}^2\\
&\le -\frac{3}{8}[x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)]^2,
\end{aligned}$$

and

$$|(y^{(1)} - \boldsymbol{x}^T\boldsymbol{\theta}_1^*)\boldsymbol{x}^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1)| \le \frac{1}{4}\tau_1|x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)|,$$

hence

$$\begin{aligned}
\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x}, y)\big|z = 1, \mathcal{V}_1 \cap \mathcal{V}_2 \cap \mathcal{V}_3\big] &\le \frac{w_r}{w_1}\mathbb{E}\bigg[\exp\bigg\{\frac{1}{4}\tau_1|x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)| - \frac{3}{8}|x^T(\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*)|^2\bigg\}\bigg]\\
\lesssim \frac{R}{c_w}\exp\bigg\{\frac{1}{4}\tau_1\tau_2\|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2 - \frac{3}{8}\tau_2\|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2^2\bigg\}&.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x}, y)\big|z = 1\big] &\lesssim \frac{R}{c_w}\exp\bigg\{\frac{1}{4}\tau_1\tau_2\|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2 - \frac{3}{8}\tau_2\|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2^2\bigg\} + \mathbb{P}(\mathcal{V}_1^c) + \mathbb{P}(\mathcal{V}_2^c) + \mathbb{P}(\mathcal{V}_3^c)\\
&\lesssim \frac{R}{c_w}\exp\bigg\{\frac{1}{4}\tau_1\tau_2\|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2 - \frac{3}{8}\tau_2\|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2^2\bigg\} + C_b + \exp\{-C\tau_1^2\} + \tau_2.
\end{aligned}$$

Let $\tau_1 = c\sqrt{\log\|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2}$ and $\tau_2 = 10c\frac{\sqrt{\log\|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2}}{\|\boldsymbol{\theta}_r^* - \boldsymbol{\theta}_1^*\|_2}$ with some constant $c > 0$. Then

$$\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x}, y)\big|z = 1\big] \lesssim \frac{R}{c_w}\frac{1}{\Delta} + C_b + \frac{1}{\Delta} + \frac{\sqrt{\log\Delta}}{\Delta} \lesssim \frac{R}{c_w}\frac{1}{\Delta} + C_b + \frac{\sqrt{\log\Delta}}{\Delta}.$$

Similarly, we can show the same bound for $\mathbb{E}[\gamma_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}^{(r)}(\boldsymbol{x}, y)|z = 1]$. Then

$$\big|\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x}, y) - \gamma_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}^{(r)}(\boldsymbol{x}, y)\big|z = 1\big]\big| \lesssim \frac{R}{c_w}\frac{1}{\Delta} + C_b + \frac{\sqrt{\log\Delta}}{\Delta},$$

and the same bound holds for $\big|\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x}, y) - \gamma_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}^{(r)}(\boldsymbol{x}, y)\big|z = r'\big]\big|$ for any $r' \ne r$. On the other hand,

$$\big|\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x}, y) - \gamma_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}^{(r)}(\boldsymbol{x}, y)\big|z = r\big]\big| \le \sum_{r'\ne r}\big|\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r')}(\boldsymbol{x}, y) - \gamma_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}^{(r')}(\boldsymbol{x}, y)\big|z = r\big]\big| \lesssim \frac{R^2}{c_w}\frac{1}{\Delta} + RC_b + R\frac{\sqrt{\log\Delta}}{\Delta}.$$

Therefore,

$$\big|\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x}, y) - \gamma_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}^{(r)}(\boldsymbol{x}, y)\big]\big| \le \sum_{r'=1}^{R} w_{r'}^*\big|\mathbb{E}\big[\gamma_{\boldsymbol{w},\boldsymbol{\theta}}^{(r)}(\boldsymbol{x}, y) - \gamma_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}^{(r)}(\boldsymbol{x}, y)\big|z = r'\big]\big|$$

$$\lesssim \frac{R^2}{c_w}\frac{1}{\Delta} + RC_b + R\frac{\sqrt{\log\Delta}}{\Delta}$$
$$\lesssim \left(\frac{R^2}{c_w}\frac{1}{\Delta} + RC_b + R\frac{\sqrt{\log\Delta}}{\Delta}\right)\cdot\|\boldsymbol{\theta}_{r_0} - \boldsymbol{\theta}^*_{r_0}\|_2,$$

where the last inequality comes from the fact that $\|\boldsymbol{\theta}_{r_0} - \boldsymbol{\theta}^*_{r_0}\|_2 > 1$.

Combining two cases entails Assumption A.3.(i) with $\kappa \asymp \frac{R}{c_w}\frac{\sqrt{\log\Delta}}{\Delta} + \frac{R}{c_w}C_b + \frac{R^2}{c_w}\frac{1}{\Delta}$.

(III) Part 3: Deriving the rate of γ in Assumption A.5.(i).

Similar to Part 2, for notation simplicity, we drop the task index k in the superscript and write $\boldsymbol{w}^{(k)} = \{w_r^{(k)}\}_{r=1}^R$, $\boldsymbol{\theta}^{(k)} = \{\boldsymbol{\theta}_r^{(k)}\}_{r=1}^R$, $\boldsymbol{w}^{(k)*} = \{w_r^{(k)*}\}_{r=1}^R$, $\boldsymbol{\theta}^{(k)*} = \{\boldsymbol{\theta}_r^{(k)*}\}_{r=1}^R$, $\boldsymbol{x}^{(k)}$, $y^{(k)}$, $Q^{(k)}$, and $q^{(k)}$ simply as $\boldsymbol{w}^{(k)} = \{w_r^{(k)}\}_{r=1}^R$, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_r\}_{r=1}^R$, $\boldsymbol{w} = \{w_r^*\}_{r=1}^R$, $\boldsymbol{\theta}^{(k)} = \{\boldsymbol{\theta}_r^*\}_{r=1}^R$, $\boldsymbol{x}$, y, Q, and q.

Note that $\frac{\partial Q^{(k)}}{\partial \boldsymbol{\theta}_r}(\boldsymbol{\theta}|\boldsymbol{w}', \boldsymbol{\theta}') = -\mathbb{E}[\gamma^{(r)}_{\boldsymbol{w}',\boldsymbol{\theta}'}(\boldsymbol{x}, y)\boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{\theta}_r - y)]$, which implies that

$$\left\|\frac{\partial Q}{\partial \boldsymbol{\theta}_r}(\boldsymbol{\theta}|\boldsymbol{w}, \boldsymbol{\theta}) - \frac{\partial q}{\partial \boldsymbol{\theta}_r}(\boldsymbol{\theta})\right\|_2 = \left\|\mathbb{E}\big[\big(\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x}, y) - \gamma^{(r)}_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}(\boldsymbol{x}, y)\big)\boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{\theta}_r - y)\big]\right\|_2$$

(a) Case 1: $\max_r \|\boldsymbol{\theta}_r - \boldsymbol{\theta}^*_r\|_2 \le 1$.

WLOG, let us consider $\mathbb{E}\big[\big(\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x}, y) - \gamma^{(r)}_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}(\boldsymbol{x}, y)\big)\boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{\theta}_r - y)\big|z = 1\big]$ with $r \neq 1$. For any $\boldsymbol{u} \in \mathcal{S}^{d-1}$,

$$\begin{aligned}
&\big|\mathbb{E}\big[\big(\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x}, y) - \gamma^{(r)}_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}(\boldsymbol{x}, y)\big)\boldsymbol{x}^T\boldsymbol{u}\cdot(\boldsymbol{x}^T\boldsymbol{\theta}_r - y)\big|z = 1\big]\big| \\
&\le \underbrace{\big|\mathbb{E}\big[\big(\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x}, y) - \gamma^{(r)}_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}(\boldsymbol{x}, y)\big)\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r - \boldsymbol{\theta}^*_1)\big|z = 1\big]\big|}_{[1]} \\
&\quad + \underbrace{\big|\mathbb{E}\big[\big(\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x}, y) - \gamma^{(r)}_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}(\boldsymbol{x}, y)\big)\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}^*_r)\big|z = 1\big]\big|}_{[2]} \\
&\quad + \underbrace{\big|\mathbb{E}\big[\big(\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x}, y) - \gamma^{(r)}_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}(\boldsymbol{x}, y)\big)\boldsymbol{x}^T\boldsymbol{u}\cdot(y - \boldsymbol{x}^T\boldsymbol{\theta}^*_1)\big|z = 1\big]\big|}_{[3]}.
\end{aligned} \tag{16}$$

First,

$$\begin{aligned}
[1] \le &\sum_{r'=1}^R \left|\mathbb{E}\left[\frac{\partial\gamma^{(r)}(\boldsymbol{x}, y)}{\partial w_{r'}}(w_{r'} - w^*_{r'})\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r - \boldsymbol{\theta}^*_1)\middle|z = 1\right]\right| \\
&+ \sum_{r'=1}^R \left|\mathbb{E}\left[\left(\frac{\partial\gamma^{(r)}(\boldsymbol{x}, y)}{\partial \boldsymbol{\theta}_{r'}}\right)^T(\boldsymbol{\theta}_{r'} - \boldsymbol{\theta}^*_{r'})\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r - \boldsymbol{\theta}^*_1)\middle|z = 1\right]\right|.
\end{aligned} \tag{17}$$

Recall that

$$\begin{aligned}
&\frac{\partial\gamma^{(r)}(\boldsymbol{x}, y)}{\partial w_r} \\
&= \frac{\exp\{y\boldsymbol{x}^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1) - \frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_r)^2 - (\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]\}(w_1 + \sum_{r'\neq r} w_{r'}\exp\{y\boldsymbol{x}^T(\boldsymbol{\theta}_{r'} - \boldsymbol{\theta}_1) - \frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_{r'})^2 - (\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]\})}{(w_1 + \sum_{r'=1}^R w_{r'}\exp\{y\boldsymbol{x}^T(\boldsymbol{\theta}_{r'} - \boldsymbol{\theta}_1) - \frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_{r'})^2 - (\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]\})^2}.
\end{aligned}$$

Denote $(*) = y^{(1)}\boldsymbol{x}^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1) - \frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_r)^2 - (\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]$, $\widetilde{z} = (y^{(1)} - \boldsymbol{x}^T\boldsymbol{\theta}^*_1)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1)|\boldsymbol{x} \sim N(0, [\boldsymbol{x}^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1)]^2)$, where $y^{(1)} \overset{d}{=} (y|z = 1)$. Then

$$(*) = \widetilde{z} + (\boldsymbol{x}^T\boldsymbol{\theta}^*_1)\cdot x^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1) - \frac{1}{2}x^T(\boldsymbol{\theta}_r + \boldsymbol{\theta}_1)\cdot x^T(\boldsymbol{\theta}_r - \boldsymbol{\theta}_1),$$

where

$$
\begin{aligned}
&(\boldsymbol{x}^T\boldsymbol{\theta}_1^*)\cdot x^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)-\frac{1}{2}x^T(\boldsymbol{\theta}_r+\boldsymbol{\theta}_1)\cdot x^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)\\
&=\big[x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)+x^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1+\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^*)\big]\bigg[-\frac{1}{2}x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)+-\frac{1}{2}x^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)+\frac{1}{2}x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)\bigg]\\
&=-\frac{1}{2}[x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)]^2+\frac{1}{2}\{x^T[(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)+(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)]\}^2.
\end{aligned}
$$

Define events

$$
\begin{aligned}
\mathcal{V}_1&=\bigg\{|x^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)|\vee|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)|\le\frac{1}{4}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|\bigg\},\\
\mathcal{V}_2&=\{|\widetilde{z}|\le\tau_1|x^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)|\},\\
\mathcal{V}_3&=\{|\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|>\tau_2|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*|\}.
\end{aligned}
$$

We know that

$$
\begin{aligned}
\mathbb{P}(\mathcal{V}_1^c)&\le\mathbb{P}\bigg(|x^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)|\vee|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)|>\frac{1}{4}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|\bigg)\\
&\le\mathbb{P}\bigg(|x^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)|>\frac{1}{4}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|\bigg)+\mathbb{P}\bigg(|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)|>\frac{1}{4}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|\bigg)\\
&\le 4\bigg(\frac{\|\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1\|_2}{\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2}+\frac{\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r\|_2}{\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2}\bigg)\\
&\le 8C_b,
\end{aligned}
$$

where we applied Lemma A.1 in Kwon & Caramanis (2020b) to get the second last inequality. And

$$
\mathbb{P}(\mathcal{V}_2^c)\le C\exp\{-C'\tau_1^2\},\quad \mathbb{P}(\mathcal{V}_3^c)\le C\tau_2.
$$

Given $\mathcal{V}_1\cap\mathcal{V}_2\cap\mathcal{V}_3$, we have

$$
-\frac{1}{2}[x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)]^2+\frac{1}{2}\{x^T[(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_1)+(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)]\}^2\le-\frac{3}{8}[x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)]^2,
$$

leading to

$$
\widetilde{z}\le\tau_1|x^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)|\le\tau_1(|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|+|x^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^*)|+|x^T(\boldsymbol{\theta}_1-\boldsymbol{\theta}_1^*)|)\le\tau_1\cdot\frac{3}{2}|x^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)|.
$$

Hence

$$
\begin{aligned}
&\bigg|\mathbb{E}\bigg[\frac{\partial\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x},y)}{\partial w_r}(w_r-w_r^*)\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)\bigg|z=1\bigg]\bigg|\\
&\le\bigg|\mathbb{E}\bigg[\frac{\partial\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x},y)}{\partial w_r}(w_r-w_r^*)\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)\bigg|z=1,\mathcal{V}_1\cap\mathcal{V}_2\cap\mathcal{V}_3\bigg]\mathbb{P}(\mathcal{V}_1\cap\mathcal{V}_2\cap\mathcal{V}_3)\bigg|\\
&\quad+\sum_{j=1}^3\bigg|\mathbb{E}\bigg[\frac{\partial\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x},y)}{\partial w_r}(w_r-w_r^*)\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)\cdot\mathbb{1}(\mathcal{V}_j^c)\bigg|z=1\bigg]\bigg|.
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\bigg|\mathbb{E}\bigg[\frac{\partial\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x},y)}{\partial w_r}(w_r-w_r^*)\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)\bigg|z=1,\mathcal{V}_1\cap\mathcal{V}_2\cap\mathcal{V}_3\bigg]\bigg|\\
&\lesssim\frac{R}{c_w}\exp\bigg\{\frac{3}{2}\tau_1\tau_2\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2-\frac{3}{8}\tau_2^2\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2^2\bigg\}\cdot\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2\cdot|w_r-w_r^*|.
\end{aligned}
$$

Also,

$$\begin{aligned}
&\left|\mathbb{E}\Big[\frac{\partial\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x},y)}{\partial w_r}(w_r-w_r^*)\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)\cdot\mathbb{1}(\mathcal{V}_1^c)\Big|z=1\Big]\right|\\
&\le\frac{R}{c_w}\sqrt{\mathbb{E}[(\boldsymbol{x}^T\boldsymbol{u})^2|\mathcal{V}_1^c]\mathbb{P}(\mathcal{V}_1^c)}\sqrt{\mathbb{E}[(\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*))^2|\mathcal{V}_1^c]\mathbb{P}(\mathcal{V}_1^c)}\cdot|w_r-w_r^*|\\
&\lesssim\frac{R}{c_w}\sqrt{\mathbb{E}[(\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^*))^2|\mathcal{V}_1^c]+\mathbb{E}[(\boldsymbol{x}^T(\boldsymbol{\theta}_1-\boldsymbol{\theta}_1^*))^2|\mathcal{V}_1^c]}\cdot\mathbb{P}(\mathcal{V}_1^c)\cdot|w_r-w_r^*|\\
&\lesssim\frac{R}{c_w}C_b|w_r-w_r^*|.
\end{aligned}$$

$$\left|\mathbb{E}\Big[\frac{\partial\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x},y)}{\partial w_r}(w_r-w_r^*)\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)\cdot\mathbb{1}(\mathcal{V}_2^c)\Big|z=1\Big]\right|\lesssim\exp\{-C\tau_1^2\}\cdot\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2\cdot|w_r-w_r^*|.$$

$$\begin{aligned}
&\left|\mathbb{E}\Big[\frac{\partial\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x},y)}{\partial w_r}(w_r-w_r^*)\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)\cdot\mathbb{1}(\mathcal{V}_3^c)\Big|z=1\Big]\right|\\
&\le\frac{R}{c_w}\sqrt{\mathbb{E}[(\boldsymbol{x}^T\boldsymbol{u})^2|\mathcal{V}_3^c]\mathbb{P}(\mathcal{V}_3^c)}\sqrt{\mathbb{E}[(\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*))^2|\mathcal{V}_3^c]\mathbb{P}(\mathcal{V}_3^c)}\cdot|w_r-w_r^*|\\
&\lesssim\frac{R}{c_w}\tau_2^2\cdot\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2\cdot|w_r-w_r^*|.
\end{aligned}$$

Let $\tau_1=c\sqrt{\log\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2}$ and $\tau_2=10c\frac{\sqrt{\log\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2}}{\|\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*\|_2}$ with some constant $c>0$, then

$$\begin{aligned}
\left|\mathbb{E}\Big[\frac{\partial\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x},y)}{\partial w_r}(w_r-w_r^*)\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)\Big|z=1\Big]\right|&\lesssim\left(\frac{R}{c_w}\frac{1}{\Delta}+\frac{R}{c_w}C_b+\frac{1}{\Delta}+\frac{R}{C_w}\frac{\log\Delta}{\Delta}\right)\cdot|w_r-w_r^*|\\
&\lesssim\left(\frac{R}{c_w}C_b+\frac{R}{C_w}\frac{\log\Delta}{\Delta}\right)\cdot|w_r-w_r^*|.
\end{aligned}$$

Similarly, we can show that

$$\left|\mathbb{E}\Big[\frac{\partial\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x},y)}{\partial w_{r'}}(w_{r'}-w_{r'}^*)\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_{r'}^*-\boldsymbol{\theta}_1^*)\Big|z=1\Big]\right|\lesssim\left(\frac{R}{c_w}C_b+\frac{R}{C_w}\frac{\log\Delta}{\Delta}\right)\cdot|w_{r'}-w_{r'}^*|,$$

for $r'\neq r$.

On the other hand, recall that

$$\frac{\partial\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)}{\partial\boldsymbol{\theta}_r}=\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)\boldsymbol{x}(y-\boldsymbol{x}^T\boldsymbol{\theta}_r)-\big(\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)\big)^2\boldsymbol{x}(y-\boldsymbol{x}^T\boldsymbol{\theta}_r),$$

where

$$\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)=\frac{w_r\exp\{y\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)-\frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_r)^2-(\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]\}}{w_1+\sum_{r'=2}^R w_{r'}\exp\{y\boldsymbol{x}^T(\boldsymbol{\theta}_{r'}-\boldsymbol{\theta}_1)-\frac{1}{2}[(\boldsymbol{x}^T\boldsymbol{\theta}_{r'})^2-(\boldsymbol{x}^T\boldsymbol{\theta}_1)^2]\}}.$$

We have

$$\begin{aligned}
&\big|\mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^*)\cdot(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_r)\cdot\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)\big|z=1\big]\big|\\
&\le\big|\mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^*)\cdot(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_1^*)\cdot\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)\big|z=1\big]\big|\\
&\quad+\big|\mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^*)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_1^*-\boldsymbol{\theta}_r^*)\cdot\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)\big|z=1\big]\big|\\
&\quad+\big|\mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^*)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_r)\cdot\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}_r^*-\boldsymbol{\theta}_1^*)\big|z=1\big]\big|.
\end{aligned}$$

Similar to the previous analysis,

$$\begin{aligned}
&\big|\mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r)\cdot(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}^*_1)\cdot\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)\big|z=1\big]\big|\\
&\le\big|\mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r)\cdot(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}^*_1)\cdot\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)\big|z=1,\mathcal{V}_1\cap\mathcal{V}_2\cap\mathcal{V}_3\big]\big|\\
&\quad+\sum_{j=1}^3\big|\mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r)\cdot(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}^*_1)\cdot\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)\big|z=1,\mathcal{V}_j^c\big]\big|\cdot\mathbb{P}(\mathcal{V}_j^c)\\
&\lesssim\frac{R}{c_w}\exp\left\{\frac{3}{2}\tau_1\tau_2\|\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1\|_2-\frac{3}{8}\tau_2\|\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1\|_2^2\right\}\|\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1\|_2\cdot\|\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r\|_2\\
&\quad+\frac{R}{c_w}C_b+\exp\{-C\tau_1^2\}\cdot\|\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1\|_2\cdot\|\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r\|_2+\frac{R}{c_w}\tau_2^2\|\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1\|_2\cdot\|\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r\|_2.
\end{aligned}$$

Let $\tau_1=c\sqrt{\log\|\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1\|_2}$ and $\tau_2=10c\frac{\sqrt{\log\|\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1\|_2}}{\|\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1\|_2}$ with some constant $c>0$, then

$$\big|\mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r)\cdot(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}^*_1)\cdot\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)\big|z=1\big]\big|\lesssim\left(\frac{R}{c_w}\frac{\log\Delta}{\Delta}+\frac{R}{c_w}C_b\right)\cdot\|\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r\|_2.$$

Similarly,

$$\begin{aligned}
&\big|\mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_1-\boldsymbol{\theta}^*_r)\cdot\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)\big|z=1\big]\big|\lesssim\left[\frac{R}{c_w}\frac{(\log\Delta)^{3/2}}{\Delta}+\frac{R}{c_w}C_b\right]\cdot\|\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r\|_2,\\
&\big|\mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r)\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}_r)\cdot\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)\big|z=1\big]\big|\lesssim\left(\frac{R}{c_w}\frac{\log\Delta}{\Delta}+\frac{R}{c_w}C_b\right)\cdot\|\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r\|_2.
\end{aligned}$$

Hence,

$$\big|\mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r)\cdot(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_r)\cdot\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)\big|z=1\big]\big|\lesssim\left[\frac{R}{c_w}\frac{(\log\Delta)^{3/2}}{\Delta}+\frac{R}{c_w}C_b\right]\cdot\|\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r\|_2.$$

Similarly,

$$\big|\mathbb{E}\big[(\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y))^2\boldsymbol{x}^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r)\cdot(y^{(1)}-\boldsymbol{x}^T\boldsymbol{\theta}_r)\cdot\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)\big|z=1\big]\big|\lesssim\left[\frac{R}{c_w}\frac{(\log\Delta)^{3/2}}{\Delta}+\frac{R}{c_w}C_b\right]\cdot\|\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r\|_2.$$

Therefore,

$$\left|\mathbb{E}\left[\left(\frac{\partial\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)}{\partial\boldsymbol{\theta}_r}\right)^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r)\cdot\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)\big|z=1\right]\right|\lesssim\left[\frac{R}{c_w}\frac{(\log\Delta)^{3/2}}{\Delta}+\frac{R}{c_w}C_b\right]\cdot\|\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r\|_2.$$

Similarly,

$$\left|\mathbb{E}\left[\left(\frac{\partial\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)}{\partial\boldsymbol{\theta}_{r'}}\right)^T(\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r)\cdot\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)\big|z=1\right]\right|\lesssim\left[\frac{R}{c_w}\frac{(\log\Delta)^{3/2}}{\Delta}+\frac{R}{c_w}C_b\right]\cdot\|\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r\|_2,$$

for $r'\neq r$. Recall (16) and (17),

$$\begin{aligned}
[1]&\le\sum_{r'=1}^R\left|\mathbb{E}\left[\frac{\partial\gamma^{(r)}(\boldsymbol{x},y)}{\partial w_{r'}}(w_{r'}-w^*_{r'})\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)\Big|z=1\right]\right|\\
&\quad+\sum_{r'=1}^R\left|\mathbb{E}\left[\left(\frac{\partial\gamma^{(r)}(\boldsymbol{x},y)}{\partial\boldsymbol{\theta}_{r'}}\right)^T(\boldsymbol{\theta}_{r'}-\boldsymbol{\theta}^*_{r'})\boldsymbol{x}^T\boldsymbol{u}\cdot\boldsymbol{x}^T(\boldsymbol{\theta}^*_r-\boldsymbol{\theta}^*_1)\Big|z=1\right]\right|\\
&\lesssim\left[\frac{R}{c_w}\frac{(\log\Delta)^{3/2}}{\Delta}+\frac{R}{c_w}C_b\right]\cdot\sum_{r'=1}^R(|w_{r'}-w^*_{r'}|+\|\boldsymbol{\theta}_{r'}-\boldsymbol{\theta}^*_{r'}\|_2).
\end{aligned}$$

Similarly, the same bound holds for terms [2] and [3] in (16) as well, therefore

$$\left|\mathbb{E}\left[\left(\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)-\gamma^{(r)}_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}(\boldsymbol{x},y)\right)\boldsymbol{x}^T\boldsymbol{u}\cdot(\boldsymbol{x}^T\boldsymbol{\theta}_r-y)\middle|z=1\right]\right| \lesssim \left[\frac{R}{c_w}\frac{(\log\Delta)^{3/2}}{\Delta}+\frac{R}{c_w}C_b\right]\cdot\sum_{r'=1}^{R}(|w_{r'}-w^*_{r'}|+\|\boldsymbol{\theta}_{r'}-\boldsymbol{\theta}^*_{r'}\|_2).$$

With similar arguments, we have

$$\left|\mathbb{E}\left[\left(\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)-\gamma^{(r)}_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}(\boldsymbol{x},y)\right)\boldsymbol{x}^T\boldsymbol{u}\cdot(\boldsymbol{x}^T\boldsymbol{\theta}_r-y)\middle|z=\tilde{r}\right]\right| \lesssim \left[\frac{R}{c_w}\frac{(\log\Delta)^{3/2}}{\Delta}+\frac{R}{c_w}C_b\right]\cdot\sum_{r'=1}^{R}(|w_{r'}-w^*_{r'}|+\|\boldsymbol{\theta}_{r'}-\boldsymbol{\theta}^*_{r'}\|_2),$$

for $\tilde{r}\neq 1$. Therefore,

$$\begin{aligned}\left\|\frac{\partial Q}{\partial\boldsymbol{\theta}_r}(\boldsymbol{\theta}|\boldsymbol{w},\boldsymbol{\theta})-\frac{\partial q}{\partial\boldsymbol{\theta}_r}(\boldsymbol{\theta})\right\|_2 &= \left\|\mathbb{E}\left[\left(\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)-\gamma^{(r)}_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}(\boldsymbol{x},y)\right)\boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{\theta}_r-y)\right]\right\|_2\\ &= \sup_{\|\boldsymbol{u}\|_2\le 1}\sum_{r'=1}^{R}\mathbb{E}\left[\left(\gamma^{(r)}_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x},y)-\gamma^{(r)}_{\boldsymbol{w}^*,\boldsymbol{\theta}^*}(\boldsymbol{x},y)\right)(\boldsymbol{x}^T\boldsymbol{u})(\boldsymbol{x}^T\boldsymbol{\theta}_r-y)\middle|z=r'\right]\cdot\mathbb{P}(z=r')\\ &\le \left[C\frac{R}{c_w}\frac{(\log\Delta)^{3/2}}{\Delta}+C\frac{R}{c_w}C_b\right]\cdot\sum_{r'=1}^{R}(|w_{r'}-w^*_{r'}|+\|\boldsymbol{\theta}_{r'}-\boldsymbol{\theta}^*_{r'}\|_2).\end{aligned}$$

(b) Case 2: $\max_r\|\boldsymbol{\theta}_r-\boldsymbol{\theta}^*_r\|_2>1$.

Similar to case 2 of Part 2, it can be shown that

$$\left\|\frac{\partial Q}{\partial\boldsymbol{\theta}_r}(\boldsymbol{\theta}|\boldsymbol{w},\boldsymbol{\theta})-\frac{\partial q}{\partial\boldsymbol{\theta}_r}(\boldsymbol{\theta})\right\|_2 \le \left[C\frac{R^2}{c_w}+C\frac{R}{c_w}\frac{(\log\Delta)^{3/2}}{\Delta}+C\frac{R}{c_w}C_b\right]\cdot\sum_{r'=1}^{R}(|w_{r'}-w^*_{r'}|+\|\boldsymbol{\theta}_{r'}-\boldsymbol{\theta}^*_{r'}\|_2).$$

Therefore $\gamma\asymp\frac{R^2}{c_w}+\frac{R}{c_w}\frac{(\log\Delta)^{3/2}}{\Delta}+\frac{R}{c_w}C_b$ in Assumption A.5.(i).

(IV) Part 4: Deriving the rate of $\mathcal{W}$ in Assumption A.3.(ii). Let

$$\begin{aligned}V &= \sup_{\substack{|w_r-w^{(k)*}_r|\le\frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}^{(k)*}_r\|_2\le\xi}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(z^{(k)}=r|\boldsymbol{x}^{(k)}_i,y^{(k)}_i;\boldsymbol{w},\boldsymbol{\theta})-\mathbb{E}_{\boldsymbol{x}^{(k)}}\left[\mathbb{P}(z^{(k)}=r|\boldsymbol{x}^{(k)},y^{(k)};\boldsymbol{w},\boldsymbol{\theta})\right]\right|\\ &= \sup_{\substack{|w_r-w^{(k)*}_r|\le\frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}^{(k)*}_r\|_2\le\xi}}\left|\frac{1}{n}\sum_{i=1}^{n}\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}^{(k)}_i,y^{(k)}_i)-\mathbb{E}\left[\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}^{(k)},y^{(k)})\right]\right|.\end{aligned}$$

By bounded difference inequality (Corollary 2.21 in Wainwright (2019)), w.p. at least $1-\delta$,

$$V\le\mathbb{E}V+\sqrt{\frac{\log(1/\delta)}{n}}.$$

And by classical symmetrization arguments (e.g., see Proposition 4.11 in Wainwright (2019)),

$$\mathbb{E}V\le\frac{2}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{\substack{|w_r-w^{(k)*}_r|\le\frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}^{(k)*}_r\|_2\le\xi}}\left|\sum_{i=1}^{n}\epsilon^{(k)}_i\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}^{(k)}_i,y^{(k)}_i)\right|.$$

Let $g^{(k)}_{ir}=(\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)^T\boldsymbol{x}^{(k)}_i\cdot y^{(k)}_i-\frac{1}{2}[((\boldsymbol{x}^{(k)}_i)^T\boldsymbol{\theta}_r)^2-((\boldsymbol{x}^{(k)}_i)^T\boldsymbol{\theta}_1)^2]+\log w_r-\log w_1$, $\varphi(\boldsymbol{x})=\frac{\exp\{x_r\}}{1+\sum_{r=2}^{R}\exp\{x_r\}}$, where φ is 1-Lipschitz (w.r.t. ℓ_2-norm) and $\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x},y)=\varphi(\{g^{(k)}_{ir}\}_{r=2}^{R})$. Then by Lemma E.3,

$$\frac{2}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{\substack{|w_r-w^{(k)*}_r|\le\frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}^{(k)*}_r\|_2\le\xi}}\left|\sum_{i=1}^{n}\epsilon^{(k)}_i\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}^{(k)}_i,y^{(k)}_i))\right|$$

$$
\begin{aligned}
&\lesssim \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le\xi}} \left|\sum_{i=1}^n\sum_{r=2}^R \epsilon_{ir}^{(k)} g_{ir}^{(k)}\right| \\
&\lesssim \frac{1}{n}\sum_{r=2}^R \mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le\xi}} \left|\sum_{i=1}^n \epsilon_{ir}^{(k)} g_{ir}^{(k)}\right| \\
&\lesssim \sum_{r=2}^R \Bigg\{ \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le\xi}} \left|\sum_{i=1}^n \epsilon_{ir}^{(k)} (\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)^T \boldsymbol{x}_i^{(k)}\cdot y_i^{(k)}\right| \\
&\qquad + \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le\xi}} \left|\sum_{i=1}^n \epsilon_{ir}^{(k)} [((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\theta}_r)^2 - ((\boldsymbol{x}_i^{(k)})^T\boldsymbol{\theta}_1)^2]\right| \\
&\qquad + \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le\xi}} \left|\sum_{i=1}^n \epsilon_{ir}^{(k)} (\log w_r - \log w_1)\right| \Bigg\} \\
&\lesssim \sum_{r=2}^R \Bigg\{ \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le\xi}} \left|\sum_{i=1}^n \epsilon_{ir}^{(k)} (\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*})^T \boldsymbol{x}_i^{(k)}\cdot y_i^{(k)}\right| \\
&\qquad + \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le\xi}} \left|\sum_{i=1}^n \epsilon_{ir}^{(k)} (\boldsymbol{\theta}_1-\boldsymbol{\theta}_1^{(k)*})^T \boldsymbol{x}_i^{(k)}\cdot y_i^{(k)}\right| \\
&\qquad + \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \left|\sum_{i=1}^n \epsilon_{ir}^{(k)} (\boldsymbol{\theta}_r^{(k)*}-\boldsymbol{\theta}_1^{(k)*})^T \boldsymbol{x}_i^{(k)}\cdot y_i^{(k)}\right| \\
&\qquad + \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le\xi}} \left|\sum_{i=1}^n \epsilon_{ir}^{(k)} (\boldsymbol{\theta}_r+\boldsymbol{\theta}_1)^T \boldsymbol{x}_i^{(k)}\cdot (\boldsymbol{\theta}_r-\boldsymbol{\theta}_1)^T \boldsymbol{x}_i^{(k)}\right| \\
&\qquad + \frac{1}{n}\mathbb{E}_{\boldsymbol{x}^{(k)}}\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r-w_r^{(k)*}|\le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le\xi}} \left|\sum_{i=1}^n \epsilon_{ir}^{(k)} (\log w_r - \log w_1)\right| \Bigg\} \\
&\lesssim RM\xi\sqrt{\frac{d}{n}} + [RM^2 + R\log(Rc_w^{-1})]\sqrt{\frac{1}{n}},
\end{aligned}
$$

which implies

$$
V \lesssim RM\xi\sqrt{\frac{d}{n}} + [RM^2 + R\log(Rc_w^{-1})]\sqrt{\frac{1}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \asymp \mathcal{W}(n,\delta,\xi).
$$

w.p. at least $1-\delta$.

(V) Part 5: Deriving the rate of $\mathcal{E}_1$ in Assumption A.5.(ii).

We first introduce the following useful lemma.

Lemma E.5 (Theorem 4 in Maurer & Pontil (2021)). *Let $f : \mathcal{X}^n \to \mathbb{R}$ and $X = (X_1, \ldots, X_n)$ be a vector of independent random variables with values in a space $\mathcal{X}$. Then for any $t > 0$ we have*

$$
\mathbb{P}(f(X) - \mathbb{E}f(X) > t) \le \exp\left\{ -\frac{t^2}{4e^2 \left\| \sum_{i=1}^n \|f_i(X)\|_{\psi_1}^2 \right\|_\infty + 2e \max_i \|\|f_i(X)\|_{\psi_1}\|_\infty t} \right\},
$$

where $f_i(X)$ as a random function of x is defined to be $(f_i(X))(x) := f(x_1, \ldots, x_{i-1}, X_i, x_{i+1}, \ldots, X_n) - \mathbb{E}_{X_i}[f(x_1, \ldots, x_{i-1}, X_i, x_{i+1}, \ldots, X_n)]$, the sub-Gaussian norm $\|Z\|_{\psi_1} := \sup_{d\ge 1}\{\|Z\|_d/d\}$, and $\|Z\|_d = (\mathbb{E}|Z|^d)^{1/d}$.

Let

$$
\begin{aligned}
U &= \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le r_{\boldsymbol{\theta}}^*}} \left\| \frac{1}{n} \sum_{i=1}^n \gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}_i^{(k)}) \boldsymbol{x}_i^{(k)} y_i^{(k)} - \mathbb{E}\big[\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}^{(k)}) \boldsymbol{x}^{(k)} y^{(k)}\big] \right\|_2 \\
&= \sup_{\|\boldsymbol{u}\|_2 \le 1} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le r_{\boldsymbol{\theta}}^*}} \left| \frac{1}{n} \sum_{i=1}^n \gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}_i^{(k)}) (\boldsymbol{x}_i^{(k)})^T \boldsymbol{u} \cdot y_i^{(k)} - \mathbb{E}\big[\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}^{(k)}) (\boldsymbol{x}^{(k)})^T \boldsymbol{u} \cdot y^{(k)}\big] \right| \\
&\le 2 \max_{j=1:N} \underbrace{\sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le r_{\boldsymbol{\theta}}^*}} \left| \frac{1}{n} \sum_{i=1}^n \gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}_i^{(k)}) (\boldsymbol{x}_i^{(k)})^T \boldsymbol{u}_j \cdot y_i^{(k)} - \mathbb{E}\big[\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}^{(k)}) (\boldsymbol{x}^{(k)})^T \boldsymbol{u}_j \cdot y_i^{(k)}\big] \right|}_{U_j},
\end{aligned}
$$

where $\{\boldsymbol{u}_j\}_{j=1}^N$ is a $1/2$-cover of the unit ball $\mathcal{B}(\mathbf{0}, 1)$ in $\mathbb{R}^d$ w.r.t. ℓ_2-norm, with $N \le 5^d$ (by Example 5.8 in Wainwright (2019)). We first bound $U_j - \mathbb{E}U_j$ as below. Fix $(\boldsymbol{x}_1^{(k)}, y_i^{(k)}), \ldots, (\boldsymbol{x}_{i-1}^{(k)}, y_{i-1}^{(k)}), (\boldsymbol{x}_{i+1}^{(k)}, y_{i+1}^{(k)}), \ldots, (\boldsymbol{x}_n^{(k)}, y_n^{(k)})$ and define $s_{ir}^{(k)}(\boldsymbol{x}_i^{(k)}, y_i^{(k)}) = V_j - \mathbb{E}\big[V_j | (\boldsymbol{x}_1^{(k)}, y_i^{(k)}), \ldots, (\boldsymbol{x}_{i-1}^{(k)}, y_{i-1}^{(k)}), (\boldsymbol{x}_{i+1}^{(k)}, y_{i+1}^{(k)}), \ldots, (\boldsymbol{x}_n^{(k)}, y_n^{(k)})\big]$. Then

$$
\begin{aligned}
|s_{ir}^{(k)}(\boldsymbol{x}_i^{(k)}, y_i^{(k)})| \le{}& \underbrace{\frac{1}{n} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le r_{\boldsymbol{\theta}}^*}} \left| \gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}_i^{(k)}, y_i^{(k)}) (\boldsymbol{x}_i^{(k)})^T \boldsymbol{u}_j \cdot y_i^{(k)} \right|}_{W_1} \\
&+ \underbrace{\frac{2}{n} \mathbb{E} \left| \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le r_{\boldsymbol{\theta}}^*}} \gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}^{(k)}, y^{(k)}) (\boldsymbol{x}^{(k)})^T \boldsymbol{u}_j \cdot y^{(k)} \right|}_{W_2},
\end{aligned}
$$

where $[\mathbb{E}(W_1 + W_2)^d]^{1/d} \le (\mathbb{E}W_1^d)^{1/d} + (\mathbb{E}W_2^d)^{1/d}$, and $(\mathbb{E}W_1^d)^{1/d}, (\mathbb{E}W_2^d)^{1/d} \le CMd/n$ with some constant$C > 0$. Then by Lemma E.2,

$$
\mathbb{P}(U_j - \mathbb{E}U_j \ge t) \lesssim \exp\left\{ - \frac{Cnt^2}{M^2} \right\}.
$$

By a similar procedure used in deriving $\mathcal{W}(n, \delta, \xi)$, we can show that

$$
\begin{aligned}
\mathbb{E}U_j &\lesssim \frac{1}{n} \mathbb{E}_{\boldsymbol{x}^{(k)}} \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le r_{\boldsymbol{\theta}}^*}} \left| \sum_{i=1}^n \gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}_i^{(k)}, y_i^{(k)}) (\boldsymbol{x}_i^{(k)})^T \boldsymbol{u}_j \cdot y_i^{(k)} \cdot \epsilon_i^{(k)} \right| \\
&\lesssim \frac{1}{n} \mathbb{E}_{\boldsymbol{x}^{(k)}} \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le r_{\boldsymbol{\theta}}^*}} \left| \sum_{i=1}^n (\boldsymbol{x}_i^{(k)})^T \boldsymbol{u}_j \cdot y_i^{(k)} \cdot \epsilon_{i1}^{(k)} \right| \\
&\quad + \sum_{r=2}^R \left\{ \frac{1}{n} \mathbb{E}_{\boldsymbol{x}^{(k)}} \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le r_{\boldsymbol{\theta}}^*}} \left| \sum_{i=1}^n \epsilon_{ir}^{(k)} (\boldsymbol{\theta}_r - \boldsymbol{\theta}_1)^T \boldsymbol{x}_i^{(k)} \cdot y_i^{(k)} \right| \right. \\
&\qquad + \frac{1}{n} \mathbb{E}_{\boldsymbol{x}^{(k)}} \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le r_{\boldsymbol{\theta}}^*}} \left| \sum_{i=1}^n \epsilon_{ir}^{(k)} y_i^{(k)} \big[(\boldsymbol{\theta}_r^T \boldsymbol{x}_i^{(k)})^2 - (\boldsymbol{\theta}_1^T \boldsymbol{x}_i^{(k)})^2 \big] \right| \\
&\qquad \left. + \frac{1}{n} \mathbb{E}_{\boldsymbol{x}^{(k)}} \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{\substack{|w_r - w_r^{(k)*}| \le \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \le r_{\boldsymbol{\theta}}^*}} \left| \sum_{i=1}^n \epsilon_{ir}^{(k)} y_i^{(k)} (\log w_r - \log w_1) \right| \right\}
\end{aligned}
$$

$$\lesssim [RM^2 r^*_{\boldsymbol{\theta}} + RM(M + r^*_{\boldsymbol{\theta}})^2]\sqrt{\frac{d}{n}} + MR\log\Big(\frac{R}{c_w}\Big)\sqrt{\frac{1}{n}}$$
$$\lesssim RM^3\sqrt{\frac{d}{n}} + MR\log\Big(\frac{R}{c_w}\Big)\sqrt{\frac{1}{n}},$$

which implies that

$$\mathbb{P}\bigg(U_j \geq RM^3\sqrt{\frac{d}{n}} + MR\log\Big(\frac{R}{c_w}\Big)\sqrt{\frac{1}{n}} + t\bigg) \lesssim \exp\bigg\{-\frac{Cn^2t^2}{nM^2 + Mtn}\bigg\} = \exp\bigg\{-\frac{Cnt^2}{M^2 + Mt}\bigg\}.$$

Therefore

$$\mathbb{P}\bigg(\max_{j=1:N} U_j \geq CRM^3\sqrt{\frac{d}{n}} + CMR\log\Big(\frac{R}{c_w}\Big)\sqrt{\frac{1}{n}} + t\bigg) \lesssim N\exp\bigg\{-\frac{Cnt^2}{M^2 + M^2t}\bigg\},$$

which implies that

$$U \lesssim RM^3\sqrt{\frac{d}{n}} + MR\log(Rc_w^{-1})\sqrt{\frac{1}{n}} + M\sqrt{\frac{\log(1/\delta)}{n}},$$

w.p. at least $1-\delta$. On the other hand, similarly, we have

$$\sup_{\substack{|w_r - w_r^{(k)*}| \leq \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \leq r^*_{\boldsymbol{\theta}}}} \bigg\|\frac{1}{n}\sum_{i=1}^n \gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}_i^{(k)})\boldsymbol{x}_i^{(k)}(\boldsymbol{x}_i^{(k)})^T\boldsymbol{\theta}_r - \mathbb{E}\big[\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}^{(k)})\boldsymbol{x}^{(k)}(\boldsymbol{x}^{(k)})^T\boldsymbol{\theta}_r\big]\bigg\|_2$$
$$\lesssim RM^3\sqrt{\frac{d}{n}} + MR\log(Rc_w^{-1})\sqrt{\frac{1}{n}} + M\sqrt{\frac{\log(1/\delta)}{n}},$$

hence

$$\mathcal{E}_1(n,\delta) \asymp RM^3\sqrt{\frac{d}{n}} + MR\log(Rc_w^{-1})\sqrt{\frac{1}{n}} + M\sqrt{\frac{\log(1/\delta)}{n}}.$$

(VI) Part 6: Deriving the rate of $\mathcal{E}_2$ in Assumption A.5.(iii). Let

$$Z = \sup_{\substack{|w_r - w_r^{(k)*}| \leq \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \leq r^*_{\boldsymbol{\theta}} \\ 0 < \eta_r^{(k)} \leq \bar{\eta}}} \frac{1}{n|S|}\bigg\|\sum_{k\in S}\eta_r^{(k)}\cdot\sum_{i=1}^n\big(\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}_i^{(k)})\boldsymbol{x}_i^{(k)}y_i^{(k)} - \mathbb{E}[\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}^{(k)})\boldsymbol{x}^{(k)}y^{(k)}]\big)\bigg\|_2$$
$$= \sup_{\|\boldsymbol{u}\|_2\leq 1}\sup_{\substack{|w_r - w_r^{(k)*}| \leq \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \leq r^*_{\boldsymbol{\theta}} \\ 0 < \eta_r^{(k)} \leq \bar{\eta}}} \frac{1}{n|S|}\bigg|\sum_{k\in S}\eta_r^{(k)}\cdot\sum_{i=1}^n\big(\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}_i^{(k)}, y_i^{(k)})(\boldsymbol{x}_i^{(k)})^T\boldsymbol{u}\cdot y_i^{(k)} - \mathbb{E}[\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k)})^T\boldsymbol{u}\cdot y^{(k)}]\big)\bigg|$$
$$\leq \sup_{j'_1,\ldots,j'_k=1:N'}\sup_{j=1:N}$$
$$\underbrace{\frac{2}{n|S|}\sup_{\substack{|w_r - w_r^{(k)*}| \leq \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \leq r^*_{\boldsymbol{\theta}}}}\bigg|\sum_{k\in S}\eta_{j'_k}\cdot\sum_{i=1}^n\big(\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}_i^{(k)}, y_i^{(k)})(\boldsymbol{x}_i^{(k)})^T\boldsymbol{u}_j\cdot y_i^{(k)} - \mathbb{E}[\gamma^{(k)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}^{(k)}, y^{(k)})(\boldsymbol{x}^{(k)})^T\boldsymbol{u}_j\cdot y^{(k)}]\big)\bigg|}_{Z(j,j'_1,\ldots,j'_k)},$$

where $\{\boldsymbol{u}_j\}_{j=1}^N$ is a $1/2$-cover of the unit ball $\mathcal{B}(\mathbf{0},1)$ in $\mathbb{R}^d$ w.r.t. ℓ_2-norm with $N \leq 5^d$ and $\{\eta_{j'}\}_{j'=1}^{N'}$ is a $1/2$-cover of $[0,1]$ with $N' \leq 2$. We first bound $Z(j,j'_1,\ldots,j'_k) - \mathbb{E}Z(j,j'_1,\ldots,j'_k)$ as follows. Fix $(\boldsymbol{x}_1^{(k)}, y_i^{(k)}),\ldots,(\boldsymbol{x}_{i-1}^{(k)}, y_{i-1}^{(k)}),(\boldsymbol{x}_{i+1}^{(k)}, y_{i+1}^{(k)}),\ldots,(\boldsymbol{x}_n^{(k)}, y_n^{(k)})$ and define $v_{ir}^{(k)}(\boldsymbol{x}_i^{(k)}, y_i^{(k)}) = Z(j,j'_1,\ldots,j'_k) - \mathbb{E}[Z(j,j'_1,\ldots,j'_k)|\{(\boldsymbol{x}_i^{(k)}, y_i^{(k)})\}_{k\in S, i\in[n]}\backslash\{(\boldsymbol{x}_i^{(k)}, y_i^{(k)})\}]$. Then

$$|v_{ir}^{(k)}(\boldsymbol{x}_i^{(k)})| \leq \underbrace{\frac{\eta_{j_k}}{n|S|}\sup_{\substack{|w_r - w_r^{(k)*}| \leq \frac{c_w}{2R} \\ \|\boldsymbol{\theta}_r - \boldsymbol{\theta}_r^{(k)*}\|_2 \leq r^*_{\boldsymbol{\theta}}}}\Big|\gamma^{(r)}_{\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{x}_i^{(k)}, y_i^{(k)})(\boldsymbol{x}_i^{(k)})^T\boldsymbol{u}_j\cdot y_i^{(k)}\Big|}_{W_1}$$

$$+\underbrace{\frac{2\eta_{j_k}}{n|S|}\mathbb{E}\left|\sup_{\substack{|w_r-w_r^{(k)*}|\le\frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le r_{\boldsymbol{\theta}}^*}}\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(r)}(\boldsymbol{x}^{(k)},y^{(k)})(\boldsymbol{x}^{(k)})^T\boldsymbol{u}_j\cdot y^{(k)}\right|}_{W_2}.$$

Via the same procedure used to bound U_j, it can be shown that

$$\mathbb{P}(Z(j,j_1',\dots,j_k')-\mathbb{E}Z(j,j_1',\dots,j_k')\ge t)\lesssim\exp\Big\{-\frac{C(n|S|)^2t^2}{n|S|M^2\bar\eta^2+\bar\eta Mtn|S|}\Big\}=\exp\Big\{-\frac{Cn|S|t^2}{M^2\bar\eta^2+\bar\eta Mt}\Big\},$$
$$\mathbb{E}Z(j,j_1',\dots,j_k')\lesssim\bar\eta RM^3\sqrt{\frac{d}{n|S|}}+\bar\eta[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}},$$

leading to

$$\mathbb{P}\Big(Z(j,j_1',\dots,j_k')\ge\bar\eta RM^3\sqrt{\frac{d}{n}}+\bar\eta[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+t\Big)\lesssim\exp\Big\{-\frac{Cn|S|t^2}{M^2\bar\eta^2+\bar\eta Mt}\Big\}.$$

Therefore

$$\mathbb{P}\Big(\max_{j_1',\dots,j_k'=1:N'}\max_{j=1:N}Z(j,j_1',\dots,j_k')\ge C\bar\eta RM^3\sqrt{\frac{d}{n}}+C[RM^3+RM\log(Rc_w^{-1})]\log(Rc_w^{-1})\sqrt{\frac{1}{n}}+t\Big)$$
$$\lesssim N(N')^K\exp\Big\{-\frac{Cn|S|t^2}{M^2\bar\eta^2+\bar\eta Mt}\Big\},$$

which implies that

$$Z\le\max_{j_1',\dots,j_k'=1:N'}\max_{j=1:N}Z(j,j_1',\dots,j_k')\lesssim\bar\eta RM^3\sqrt{\frac{d}{n}}+\bar\eta[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+\bar\eta M\sqrt{\frac{\log(1/\delta)}{n|S|}},$$

w.p. at least $1-\delta$. Similarly,

$$\sup_{\substack{|w_r-w_r^{(k)*}|\le\frac{c_w}{2R}\\ \|\boldsymbol{\theta}_r-\boldsymbol{\theta}_r^{(k)*}\|_2\le r_{\boldsymbol{\theta}}^*\\ 0<\eta_r^{(k)}\le\bar\eta}}\frac{1}{n|S|}\Big\|\sum_{k\in S}\eta_r^{(k)}\cdot\sum_{i=1}^n\big(\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(k)}(\boldsymbol{x}_i^{(k)})\boldsymbol{\theta}_r-\mathbb{E}[\gamma_{\boldsymbol{\theta},\boldsymbol{w}}^{(k)}(\boldsymbol{x}^{(k)})\boldsymbol{\theta}_r]\big)\Big\|_2$$
$$\lesssim\bar\eta RM^3\sqrt{\frac{d}{n}}+\bar\eta[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+\bar\eta M\sqrt{\frac{\log(1/\delta)}{n|S|}},$$

w.p. at least $1-\delta$. Hence

$$\mathcal{E}_2(n,|S|,\delta)\asymp RM^3\sqrt{\frac{d}{n}}+[RM^3+RM\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}+M\sqrt{\frac{\log(1/\delta)}{n|S|}}.$$

E.6. Proof of Proposition A.19

Recall that

$$A_t=\Big[9\tilde\kappa_0\Big(\frac{\tilde\kappa_0}{119}\Big)^{t-1}+\frac{118}{119}(t-1)\tilde\kappa_0^{t-1}\Big](r_w^*+r_{\boldsymbol{\theta}}^*)+\frac{1}{1-\tilde\kappa_0/119}\bar\eta\mathcal{E}_2\Big(n,|S|,\frac{\delta}{3R}\Big)$$
$$+\frac{18}{1-\tilde\kappa_0/119}\min\Big\{3h,\frac{6}{1-\tilde\kappa_0}\Big[\mathcal{W}\Big(n,\frac{\delta}{3RK},r_{\boldsymbol{\theta}}^*\Big)+2\bar\eta\mathcal{E}_1\Big(n,\frac{\delta}{3R}\Big)\Big]\Big\}$$
$$+\frac{30}{(1-\tilde\kappa_0)(1-\tilde\kappa_0/119)}\epsilon\Big[\mathcal{W}\Big(n,\frac{\delta}{3RK},r_{\boldsymbol{\theta}}^*\Big)+2\bar\eta\mathcal{E}_1\Big(n,\frac{\delta}{3R}\Big)\Big],$$

and

$$A_t + \frac{18}{1-\tilde{\kappa}_0/119}\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta},t}\Big) = r^*_{\boldsymbol{\theta},t+1},$$

for $t \geq 1$ with $r^*_{\boldsymbol{\theta},1} \coloneqq r^*_{\boldsymbol{\theta}}$.

By Assumption A.15.(iv), there exists $\tilde{\kappa}'_0 \in (0,1)$ such that $CRM\sqrt{\frac{p}{n}} \leq \tilde{\kappa}'_0$ with a large C. Hence by plugging in the explicit rates obtained in Proposition A.16,

$$\begin{aligned}
r^*_{\boldsymbol{\theta},t+1} &\leq \tilde{\kappa}'_0 r^*_{\boldsymbol{\theta},t} + Ct(\tilde{\kappa}_0)^{t-1}(r^*_w \vee r^*_{\boldsymbol{\theta}}) + C\bar{\eta}RM^3\sqrt{\frac{d}{n|S|}} + C[(\bar{\eta}M)\vee 1][RM^2 + R\log(Rc_w^{-1})]\sqrt{\frac{1}{n}} \\
&\quad + C[(\bar{\eta}M)\vee 1]\sqrt{\frac{\log(RK/\delta)}{n}} + C\min\left\{h, \bar{\eta}RM^3\sqrt{\frac{d}{n}}\right\} + \epsilon\bar{\eta}RM^2[(\bar{\eta}M)\vee 1]\sqrt{\frac{d}{n}},
\end{aligned}$$

implying that

$$\begin{aligned}
r^*_{\boldsymbol{\theta},T} &\lesssim (\tilde{\kappa}'_0)^{T-1} r^*_{\boldsymbol{\theta}} + T^2(\tilde{\kappa}'_0)^{T-1}(r^*_w \vee r^*_{\boldsymbol{\theta}}) + \bar{\eta}RM^3\sqrt{\frac{d}{n|S|}} + [(\bar{\eta}M)\vee 1][RM^2 + R\log(Rc_w^{-1})]\sqrt{\frac{1}{n}} \\
&\quad + [(\bar{\eta}M)\vee 1]\sqrt{\frac{\log(RK/\delta)}{n}} + \min\left\{h, \bar{\eta}RM^3\sqrt{\frac{d}{n}}\right\} + \epsilon\bar{\eta}RM^2[(\bar{\eta}M)\vee 1]\sqrt{\frac{d}{n}} \\
&\lesssim T^2(\tilde{\kappa}_0 \vee \tilde{\kappa}'_0)^{T-1}(r^*_w \vee r^*_{\boldsymbol{\theta}}) + \bar{\eta}RM^3\sqrt{\frac{d}{n|S|}} + [(\bar{\eta}M)\vee 1][RM^2 + R\log(Rc_w^{-1})]\sqrt{\frac{1}{n}} \\
&\quad + [(\bar{\eta}M)\vee 1]\sqrt{\frac{\log(RK/\delta)}{n}} + \min\left\{h, \bar{\eta}RM^3\sqrt{\frac{d}{n}}\right\} + \epsilon\bar{\eta}RM^2[(\bar{\eta}M)\vee 1]\sqrt{\frac{d}{n}} \\
&\lesssim T^2(\tilde{\kappa}_0 \vee \tilde{\kappa}'_0)^{T-1}(r^*_w \vee r^*_{\boldsymbol{\theta}}) + R^2M^3c_w^{-1}\sqrt{\frac{d}{n|S|}} + R^2Mc_w^{-1}[M^2 + \log(Rc_w^{-1})]\sqrt{\frac{1}{n}} \\
&\quad + MRc_w^{-1}\sqrt{\frac{\log(RK/\delta)}{n}} + \min\left\{h, R^2M^3c_w^{-1}\sqrt{\frac{d}{n}}\right\} + \epsilon RM^3c_w^{-1}\sqrt{\frac{d}{n}},
\end{aligned}$$

where $\kappa_0 = 119\sqrt{\frac{3C_b}{1+2C_b} + CR^3c_w^{-2}\frac{(\log\Delta)^{3/2}}{\Delta}} + CR^3c_w^{-2}C_b + CR^4c_w^{-2}\frac{1}{\Delta} + \tilde{\kappa}'_0$, and $\tilde{\kappa}'_0$ satisfies $1 > \tilde{\kappa}'_0 > CMR\sqrt{\frac{d}{n}}$ for some $C > 0$.

E.7. Proof of Corollary A.17

By the rate of $\mathcal{W}(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta},T})$ in Proposition A.16 and the upper bound of $r^*_{\boldsymbol{\theta},T}$ in Proposition A.19,

$$\begin{aligned}
\mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta},T}\Big) &\asymp RMr^*_{\boldsymbol{\theta},T}\sqrt{\frac{d}{n}} + [RM^2 + R\log(Rc_w^{-1})]\sqrt{\frac{1}{n}} + \sqrt{\frac{\log(RK/\delta)}{n}} \\
&\lesssim T^2(\tilde{\kappa}_0 \vee \tilde{\kappa}'_0)^{T-1}(r^*_w \vee r^*_{\boldsymbol{\theta}}) + R^2M^3c_w^{-1}\sqrt{\frac{d}{n|S|}} + R^2Mc_w^{-1}[M^2 + \log(Rc_w^{-1})]\sqrt{\frac{1}{n}} \\
&\quad + MRc_w^{-1}\sqrt{\frac{\log(RK/\delta)}{n}} + \min\left\{h, R^2M^3c_w^{-1}\sqrt{\frac{d}{n}}\right\} + \epsilon RM^3c_w^{-1}\sqrt{\frac{d}{n}}.
\end{aligned}$$

Applying Theorem A.8, we have

$$\begin{aligned}
&\max_{k\in S}\max_{r\in[R]}(\|\widehat{\boldsymbol{\theta}}_r^{(k)[T]} - \boldsymbol{\theta}_r^{(k)*}\|_2 + |\widehat{w}_r^{(k)[T]} - w_r^{(k)*}|) \\
&\leq 20T(\tilde{\kappa}_0)^{T-1}(r^*_w \vee r^*_{\boldsymbol{\theta}}) + \left[\frac{119}{15}\tilde{\kappa}_0(\tilde{\kappa}_0/119)^{T-1} + \frac{118}{119}(T-1)(\tilde{\kappa}_0)^T\right](r^*_w + r^*_{\boldsymbol{\theta}}) \\
&\quad + \frac{1}{1-\kappa_0}\Big[\bar{\eta}\mathcal{E}_2\Big(n, |S|, \frac{\delta}{3R}\Big) + \mathcal{W}\Big(n, \frac{\delta}{3RK}, r^*_{\boldsymbol{\theta},J}\Big)\Big]
\end{aligned}$$

$$
\begin{aligned}
&+\frac{18}{1-\tilde{\kappa}_0/119}\cdot\min\left\{3h,\frac{6}{1-\tilde{\kappa}_0}\Big[\mathcal{W}\Big(n,\frac{\delta}{3RK},r_{\boldsymbol{\theta}}^*\Big)+2\bar{\eta}\mathcal{E}_1\Big(n,\frac{\delta}{3RK}\Big)\Big]\right\}\\
&+\frac{30}{(1-\tilde{\kappa}_0)(1-\tilde{\kappa}_0/119)}\epsilon\cdot\Big[\mathcal{W}\Big(n,\frac{\delta}{3RK},r_{\boldsymbol{\theta}}^*\Big)+2\bar{\eta}\mathcal{E}_1\Big(n,\frac{\delta}{3RK}\Big)\Big]\\
&\le T^2(\tilde{\kappa}_0\vee\tilde{\kappa}_0')^{T-1}(r_w^*\vee r_{\boldsymbol{\theta}}^*)+R^2M^3c_w^{-1}\sqrt{\frac{d}{n|S|}}+R^2Mc_w^{-1}[M^2+\log(Rc_w^{-1})]\sqrt{\frac{1}{n}}\\
&+MRc_w^{-1}\sqrt{\frac{\log(RK/\delta)}{n}}+\min\left\{h,R^2M^3c_w^{-1}\sqrt{\frac{d}{n}}\right\}+\epsilon RM^3c_w^{-1}\sqrt{\frac{d}{n}}. \qquad (18)
\end{aligned}
$$

Note that conditioned on the event $\mathcal{V}$ defined in the proof of Theorem A.8,

$$\eta_r^{(k)}=(1+2C_b)^{-1}(\widehat{w}_r^{(k)[0]})^{-1}\lesssim Rc_w^{-1},$$

for all $k\in S$ and $r\in[R]$. Plugging it in equation (18) implies the desired upper bound in Corollary 2.

E.8. Proof of Theorem C.2

Recall that our best permutation $\pi_k^*\in\mathcal{P}^R$ on task k can be up to a permutation on $[K]$. WLOG, consider π_k^* satisfying that $\pi_k^*(r)=$ "the majority class" $\tilde{r}$ if $\#\{k\in S:\pi_k(r)=\tilde{r}\}>\frac{1}{2}|S|$, for all $k\in S$. Define "the majority class" of $\{\pi_k(r)\}_{k\in S}$ as the $\tilde{r}\in[R]$ which satisfies $\#\{k\in S:\pi_k(r)=\tilde{r}\}\ge\max_{r'\neq\tilde{r}}\#\{k\in S:\pi_k(r)=r'\}$. Denote the majority class of $\{\pi_k(r)\}_{k\in S}$ as $m_r\in[R]$ and $S_r=\{k\in S:\pi_k(r)=m_r\}$. WLOG, suppose $\boldsymbol{\pi}^*=\{\pi_k^*\}_{k=1}^K$ satisties $\pi_k^*(r)=r$ for any r and $k\in S$. Consider any $\boldsymbol{\pi}=\{\pi_k\}_{k=1}^K$ with $\pi_k(r)=\pi_k^*(r)$ for all $k\in S^c$ and $\boldsymbol{\pi}\neq\boldsymbol{\pi}^*$. It suffices to show that $\text{score}(\boldsymbol{\pi},K)>\text{score}(\boldsymbol{\pi}^*,K)$.

For convenience, denote $\xi=\max_{k\in S}\min_{\pi_k}\max_{r\in[R]}\|\widehat{\boldsymbol{\theta}}_{\pi_k(r)}^{(k)[0]}-\boldsymbol{\theta}_r^{(k)*}\|_2$. We have

$$
\begin{aligned}
\text{score}(\boldsymbol{\pi},K)-\text{score}(\boldsymbol{\pi}^*,K)=&\underbrace{\sum_{k\neq k'\in S}\sum_{r=1}^R\|\widehat{\boldsymbol{\theta}}_{\pi_k(r)}^{(k)[0]}-\widehat{\boldsymbol{\theta}}_{\pi_{k'}(r)}^{(k')[0]}\|_2}_{[1]}+2\underbrace{\sum_{k\in S,k'\in S^c}\sum_{r=1}^R\|\widehat{\boldsymbol{\theta}}_{\pi_k(r)}^{(k)[0]}-\widehat{\boldsymbol{\theta}}_{\pi_{k'}(r)}^{(k')[0]}\|_2}_{[2]}\\
&-\underbrace{\sum_{k\neq k'\in S}\sum_{r=1}^R\|\widehat{\boldsymbol{\theta}}_r^{(k)[0]}-\widehat{\boldsymbol{\theta}}_r^{(k')[0]}\|_2}_{[1]'}-2\underbrace{\sum_{k\neq k'\in S}\sum_{r=1}^R\|\widehat{\boldsymbol{\theta}}_r^{(k)[0]}-\widehat{\boldsymbol{\theta}}_{\pi_{k'}(r)}^{(k')[0]}\|_2}_{[2]'}.
\end{aligned}
$$

Note that

$$
\begin{aligned}
[1]-[1]'&=\sum_{k\neq k'\in S}\sum_{r:\pi_k(r)\neq\pi_{k'}(r)}\Big(\|\widehat{\boldsymbol{\theta}}_{\pi_k(r)}^{(k)[0]}-\widehat{\boldsymbol{\theta}}_{\pi_{k'}(r)}^{(k')[0]}\|_2-\|\widehat{\boldsymbol{\theta}}_r^{(k)[0]}-\widehat{\boldsymbol{\theta}}_r^{(k')[0]}\|_2\Big)\\
&\ge\sum_{k\neq k'\in S}\sum_{r:\pi_k(r)\neq\pi_{k'}(r)}\Big(\|\boldsymbol{\theta}_{\pi_k(r)}^{(k)*}-\boldsymbol{\theta}_{\pi_{k'}(r)}^{(k')*}\|_2-\|\boldsymbol{\theta}_r^{(k)*}-\boldsymbol{\theta}_r^{(k')*}\|_2-4\xi\Big)\\
&\ge\sum_{k\neq k'\in S}\sum_{r:\pi_k(r)\neq\pi_{k'}(r)}\Big(\|\boldsymbol{\theta}_{\pi_k(r)}^{(k)*}-\boldsymbol{\theta}_{\pi_{k'}(r)}^{(k)*}\|_2-\|\boldsymbol{\theta}_{\pi_{k'}(r)}^{(k)*}-\boldsymbol{\theta}_{\pi_{k'}(r)}^{(k')*}\|_2-\|\boldsymbol{\theta}_r^{(k)*}-\boldsymbol{\theta}_r^{(k')*}\|_2-4\xi\Big)\\
&\ge\sum_{k\neq k'\in S}\sum_{r:\pi_k(r)\neq\pi_{k'}(r)}(\Delta-2h-4\xi)\\
&=\sum_{r=1}^R\sum_{k\neq k'\in S,\pi_k(r)\neq\pi_{k'}(r)}(\Delta-2h-4\xi).
\end{aligned}
$$

For r with $|S_r|>\frac{1}{2}|S|$:

$$\sum_{k\neq k':\pi_k(r)\neq\pi_{k'}(r)}(\Delta-2h-4\xi)\ge|S_r|(|S|-|S_r|)(\Delta-2h-4\xi).$$

For r with $|S_r| \le \frac{1}{2}|S|$: denote $\#\{K \in S : \pi_k(r) = r'\}$ as $v_{r'}$, where $\sum_{r'=1}^{R} v_{r'} = |S|$ and $|v_{r'}| \le \frac{1}{2}|S|$ for all r'. Then

$$\begin{aligned}
\sum_{k \neq k' : \pi_k(r) \neq \pi_{k'}(r)} (\Delta - 2h - 4\xi) &\ge (\Delta - 2h - 4\xi)\left[|S|(|S|-1) - \sum_{r'=1}^{R} v_{r'}(v_{r'}-1)\right] \\
&= (\Delta - 2h - 4\xi)\left[|S|^2 - \sum_{r'=1}^{R} v_{r'}^2\right] \\
&\ge (\Delta - 2h - 4\xi)\left[|S|^2 - 2 \cdot \left(\frac{1}{2}|S|\right)^2\right] \\
&\ge (\Delta - 2h - 4\xi) \cdot \frac{1}{2}|S|^2.
\end{aligned}$$

Hence

$$[1] - [1]' \ge (\Delta - 2h - 4\xi) \cdot \left[\sum_{r:|S_r|>|S|/2} |S_r|(|S| - |S_r|) + \sum_{r:|S_r|\le|S|/2} \frac{1}{2}|S|^2\right].$$

And

$$\begin{aligned}
[2] - [2]' &\ge -2 \sum_{k \in S, k' \in S^c} \sum_{r=1}^{R} \|\widehat{\boldsymbol{\theta}}_{\pi_k(r)}^{(k)[0]} - \widehat{\boldsymbol{\theta}}_{r}^{(k)[0]}\|_2 \\
&\ge -2|S^c| \sum_{k \in S} \sum_{r:\pi_k(r)\neq r} (h + 2\xi) \\
&\ge -2|S^c| \left[\sum_{r:|S_r|>|S|/2} \sum_{k \in S, \pi_k(r)\neq r} (h+2\xi) + \sum_{r:|S_r|\le|S|/2} \sum_{k \in S, \pi_k(r)\neq r} (h+2\xi)\right] \\
&\ge -2|S^c|(h+2\xi)\left[\sum_{r:|S_r|>|S|/2} (|S| - |S_r|) + \sum_{r:|S_r|\le|S|/2} |S|\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\text{score}(\boldsymbol{\pi}) - \text{score}(\boldsymbol{\pi}^*) &\ge \sum_{r:|S_r|>|S|/2} (|S| - |S_r|)[|S_r|(\Delta - 2h - 4\xi) - 2|S^c|(h+2\xi)] \\
&\quad + \sum_{r:|S_r|\le|S|/2} \left[\frac{1}{2}|S|^2(\Delta - 2h - 4\xi) - 2|S^c||S|(h+2\xi)\right] \\
&\ge \sum_{r:|S_r|>|S|/2} (|S| - |S_r|)\left[\frac{1}{2}|S|\Delta - h(|S| + 2|S^c|) - \xi(2|S| + 4|S^c|)\right] \\
&\quad + \sum_{r:|S_r|\le|S|/2} |S|\left[\frac{1}{2}|S|(\Delta - 2h - 4\xi) - 2|S^c|(h+2\xi)\right] \\
&\ge \sum_{r:|S_r|>|S|/2} (|S| - |S_r|) \cdot \frac{1}{2}|S|\left[\Delta - h\left(2 + 4 \cdot \frac{|S^c|}{|S|}\right) - \xi\left(4 + 8 \cdot \frac{|S^c|}{|S|}\right)\right] \\
&\quad + \sum_{r:|S_r|\le|S|/2} \frac{1}{2}|S|^2\left[\Delta - h\left(2 + 4 \cdot \frac{|S^c|}{|S|}\right) - \xi\left(4 + 8 \cdot \frac{|S^c|}{|S|}\right)\right] \\
&> 0,
\end{aligned}$$

which completes the proof, where in the last inequality we used the fact that $|S^c|/|S| \le \frac{\epsilon}{1-\epsilon}$.

E.9. Proof of Theorem C.4

Consider the $\widetilde{K}$-th round where $\widetilde{K} \in S$. WLOG, suppose ι is the identity permutation on $[R]$. Denote $\widetilde{S} = [\widetilde{K}] \cap S$, $\widetilde{S}^c = [\widetilde{K}] \cap S^c$, hence $[K] = \widetilde{S} \cup \widetilde{S}^c$. WLOG, suppose $\pi_1 = \pi_2 = \ldots = \pi_{\widetilde{K}-1}$ are the identity permutations on $[R]$.

Denote $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^{\widetilde{K}-1} \cup \pi_{\widetilde{K}}$ and $\widetilde{\boldsymbol{\pi}} = \{\pi_k\}_{k=1}^{\widetilde{K}-1} \cup \widetilde{\pi}_{\widetilde{K}}$, where $\widetilde{\pi}_{\widetilde{K}}$ is the identity permutation on $[R]$ and $\pi_{\widetilde{K}}$ can be any non-identity permutation. We claim that it suffices to show that $\text{score}(\boldsymbol{\pi}) > \text{score}(\widetilde{\boldsymbol{\pi}})$ for any $\widetilde{K} \geq K_0$, because if this is the case, then $\widetilde{\pi}_{\widetilde{K}}$ will be chosen in the $\widetilde{K}$-th round of the "for" loop. Hence all chosen permutations in $\widetilde{S}$ have the same alignment. By induction, the output permutations from "Permutation Alignment Algorithm 2 (Stepwise search)" are identity permutations on $[R]$ among tasks in S, which completes our proof. In the remaining part of this proof, we show $\text{score}(\boldsymbol{\pi}) > \text{score}(\widetilde{\boldsymbol{\pi}})$ for any $\widetilde{K} \geq K_0$.

In fact,

$$\begin{aligned}\text{score}(\boldsymbol{\pi}) - \text{score}(\widetilde{\boldsymbol{\pi}}) = &\underbrace{\sum_{k\in\widetilde{S}}\sum_{r:\pi_{\widetilde{K}}(r)\neq r} \|\widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_{\pi_{\widetilde{K}}(r)} - \widehat{\boldsymbol{\theta}}^{(k)[0]}_r\|_2}_{[1]} + \underbrace{\sum_{k\in\widetilde{S}^c}\sum_{r:\pi_{\widetilde{K}}(r)\neq r} \|\widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_{\pi_{\widetilde{K}}(r)} - \widehat{\boldsymbol{\theta}}^{(k)[0]}_r\|_2}_{[2]} \\ &- \underbrace{\sum_{k\in\widetilde{S}}\sum_{r:\pi_{\widetilde{K}}(r)\neq r} \|\widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_r - \widehat{\boldsymbol{\theta}}^{(k)[0]}_r\|_2}_{[1]'} - \underbrace{\sum_{k\in\widetilde{S}^c}\sum_{r:\pi_{\widetilde{K}}(r)\neq r} \|\widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_r - \widehat{\boldsymbol{\theta}}^{(k)[0]}_r\|_2}_{[2]'}.\end{aligned}$$

And

$$\begin{aligned}[1]-[1]' &= \sum_{k\in\widetilde{S}}\sum_{r:\pi_{\widetilde{K}}(r)\neq r}\Big(\|\widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_{\pi_{\widetilde{K}}(r)} - \widehat{\boldsymbol{\theta}}^{(k)[0]}_r\|_2 - \|\widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_r - \widehat{\boldsymbol{\theta}}^{(k)[0]}_r\|_2\Big) \\ &\geq \sum_{k\in\widetilde{S}}\sum_{r:\pi_{\widetilde{K}}(r)\neq r}\Big(\|\widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_{\pi_{\widetilde{K}}(r)} - \widehat{\boldsymbol{\theta}}^{(k)[0]}_r\|_2 - \|\boldsymbol{\theta}^{(\widetilde{K})*}_r - \boldsymbol{\theta}^{(k)*}_r\|_2 - 2\xi\Big) \\ &\geq \sum_{k\in\widetilde{S}}\sum_{r:\pi_{\widetilde{K}}(r)\neq r}\Big(\|\widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_{\pi_{\widetilde{K}}(r)} - \widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_r\|_2 - 2h - 4\xi\Big) \\ &\geq \sum_{r:\pi_{\widetilde{K}}(r)\neq r}|\widetilde{S}|\Big(\|\widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_{\pi_{\widetilde{K}}(r)} - \widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_r\|_2 - 2h - 4\xi\Big),\end{aligned}$$

and

$$\begin{aligned}[2]-[2]' &\geq \sum_{k\in\widetilde{S}^c}\sum_{r:\pi_{\widetilde{K}}(r)\neq r}\Big(\|\widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_{\pi_{\widetilde{K}}(r)} - \widehat{\boldsymbol{\theta}}^{(k)[0]}_r\|_2 - \|\widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_r - \widehat{\boldsymbol{\theta}}^{(k)[0]}_r\|_2\Big) \\ &\geq -\sum_{k\in\widetilde{S}^c}\sum_{r:\pi_{\widetilde{K}}(r)\neq r}\|\widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_{\pi_{\widetilde{K}}(r)} - \widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_r\|_2 \\ &= -\sum_{r:\pi_{\widetilde{K}}(r)\neq r}|\widetilde{S}^c|\cdot\|\widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_{\pi_{\widetilde{K}}(r)} - \widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_r\|_2.\end{aligned}$$

Therefore,

$$\begin{aligned}\text{score}(\boldsymbol{\pi}) - \text{score}(\widetilde{\boldsymbol{\pi}}) &= [1]-[1]'+[2]-[2]' \\ &\geq \sum_{r:\pi_{\widetilde{K}}(r)\neq r}\Big[(|\widetilde{S}|-|\widetilde{S}^c|)\|\widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_{\pi_{\widetilde{K}}(r)} - \widehat{\boldsymbol{\theta}}^{(\widetilde{K})[0]}_r\|_2 - 2|\widetilde{S}|h - 4|\widetilde{S}|\xi\Big] \\ &\geq \sum_{r:\pi_{\widetilde{K}}(r)\neq r}\Big[(|\widetilde{S}|-|\widetilde{S}^c|)\Delta - 2|\widetilde{S}|h - (6|\widetilde{S}|-2|\widetilde{S}^c|)\xi\Big] \\ &= \sum_{r:\pi_{\widetilde{K}}(r)\neq r}\widetilde{K}\Big[\Big(2\frac{|\widetilde{S}|}{\widetilde{K}}-1\Big)\Delta - 2\frac{|\widetilde{S}|}{\widetilde{K}}\cdot h - \Big(8\frac{|\widetilde{S}|}{\widetilde{K}}-2\Big)\xi\Big] \\ &\geq \sum_{r:\pi_{\widetilde{K}}(r)\neq r}\widetilde{K}\cdot\Big(\frac{K_0-K\epsilon}{K_0+K\epsilon}\cdot\Delta - 2h - 6\xi\Big) \\ &> 0,\end{aligned}$$

where the second last inequality is due to the fact that $\frac{K_0}{K_0+K\epsilon} \leq |\widetilde{S}/\widetilde{K}| \leq 1$.