EM for Mixture of Linear Regression with Clustered Data

Amirhossein Reisizadeh
EECS, MIT

Khashayar Gatmiry
EECS, MIT

Asuman Ozdaglar
EECS, MIT

Abstract

Modern data-driven and distributed learning frameworks deal with diverse massive data generated by clients spread across heterogeneous environments. Indeed, *data heterogeneity* is a major bottleneck in scaling up many distributed learning paradigms. In many settings however, heterogeneous data may be generated in *clusters* with shared structures, as is the case in several applications such as federated learning where a common latent variable governs the distribution of all the samples generated by a client. It is therefore natural to ask how the underlying clustered structures in distributed data can be exploited to improve learning schemes. In this paper, we tackle this question in the special case of estimating d-dimensional parameters of a two-component mixture of linear regressions problem where each of m nodes generates n samples with a *shared* latent variable. We employ the well-known Expectation-Maximization (EM) method to estimate the maximum likelihood parameters from m batches of dependent samples each containing n measurements. Discarding the clustered structure in the mixture model, EM is known to require $\mathcal{O}(\log(mn/d))$ iterations to reach the statistical accuracy of $\mathcal{O}(\sqrt{d/(mn)})$. In contrast, we show that if initialized properly, EM on the structured data requires only $\mathcal{O}(1)$ iterations to reach the same statistical accuracy, as long as m grows up as $e^{o(n)}$. Our analysis establishes and combines novel asymptotic optimization and generalization guarantees for population and empirical EM with dependent samples, which may be of independent interest.

1 INTRODUCTION

With the ever-growing applications of data-intensive and distributed learning paradigms, it becomes more critical to address new challenges associated with such frameworks. For instance, federated learning is a novel distributed learning architecture consisting a central parameter server and a network of clients (or nodes) each equipped with locally generated data. In general, the main premise of such distributed learning methods is to estimate the underlying ground truth model using the collective data samples across the clients. *Data heterogeneity* (or non-i.i.d. data) is among the most significant challenges in scaling up distributed learning methods. Indeed, naive distributed and federated benchmarks such as FedAvg are known to diverge if deployed on highly heterogeneous settings, unless particularly tailored for non-i.i.d. data (Karimireddy et al., 2020).

In this paper, we consider a *structured* or *clustered* data heterogeneity model which roots in an observation specific to modern data-driven distributed and federated learning applications. Under this structured heterogeneity model, an *identical* and unobserved latent variable governs the distribution of *all* the samples generated at any node (Pei et al., 2017; Hendrycks and Dietterich, 2019; Robey et al., 2020; Diamandis et al., 2021). Particularly in this paper, we zoom in on *mixture of linear regression* model which is a classical approach to capture data heterogeneity (Jordan and Jacobs, 1994; Xu et al., 2016; Viele and Tong, 2002). To be more clear, in our setting each node observes not one but a potentially large number of linear measurements for all of which a common latent variable governs the true parameter. These latent variables are unknown, random, independent and identically distributed across the nodes. Throughout the paper, we refer to this model as *clustered mixture of linear regressions*, or C-MLR in short.

Our goal in this work is to estimate the maximum likelihood parameters of the regression model in the above-described C-MLR heterogeneity model using

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

the collection of *all* the observations across all the devices. However, maximizing likelihood objectives are notoriously intractable in general, due to non-convexity of the likelihood function (Yi et al., 2014). The most popular approach for computationally efficient inference in such models with latent variables is the Expected-Maximization (EM) method (Dempster et al., 1977; Redner and Walker, 1984; Wu, 1983). We therefore aim to study optimization and generalization characteristics of the EM method in estimating the C-MLR models.

To this end, we first characterize and analyse the so-called *population EM* variant for which we establish an asymptotic, local and deterministic convergence guarantee. Next, we move to the empirical counterpart with finite number of observations known as the *empirical EM* method and provide probabilistic generalization bounds on its estimation error. Both results are local and asymptotic. That is, our analysis relies on the assumption that the initial iterate of the EM method is suitable (as opposed to random). Moreover, we let the number of nodes and the number of samples per node grow while all the other parameters assumed to be constants. To be more specific, let us precisely describe the C-MLR model in the following.

1.1 Clustered MLR Model

As discussed above and motivated by distributed learning applications, we consider a collection of m nodes where each node $j = 1, \cdots, m$ observes n pairs of measurements denoted by $\{(x_i^j, y_i^j) | i = 1, \cdots, n\}$. Here, $x_i^j \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y_i^j \in \mathcal{Y} \subseteq \mathbb{R}$ denote the covariate and response variables, respectively. These observations are linear measurements of a *clustered* mixture of linear regressions (C-MLR) model described below

$$y_i^j = \xi^j \langle x_i^j, \theta^* \rangle + \epsilon_i^j, \qquad \text{(C-MLR)} \qquad (1)$$

for all $i = 1, \cdots, n$ and $j = 1, \cdots, m$. In this model, $\xi^j \in \Xi$ denotes the hidden latent variable corresponding to node j. In this paper, we focus on a symmetric and two-component mixture of linear regressions with $\Xi = \{\pm 1\}$, where ξ^j takes on values uniformly at random, denoted by $\xi^j \sim \mathcal{U}\{\pm 1\}$. Note that this latent variable is *identical* for *all* the measurements of a given node, however, we assume that they are *independent* across different nodes. Moreover, we let $\theta^* \in \mathbb{R}^d$ denote the fixed and unknown ground truth regression vector and assume that covariates and noises are independent and Gaussian with $x_i^j \sim \mathcal{N}(0, I_d)$ and $\epsilon_i^j \sim \mathcal{N}(0, \sigma^2)$, respectively. This model clearly implies that the observations of any given node are *not* independent due to the shared latent variable. In the remainder of the paper, we denote the signal-to-noise ratio (SNR) by $\mathsf{snr} = \|\theta^*\|/\sigma$.

Remark 1. C-MLR model in (1) captures the underlying node-dependent data heterogeneity through the latent variable ξ^j which is shared and identical for all the n samples measured by node j. Therefore, C-MLR is a well-motivated abstract model to encapsulate the structured data heterogeneity observed in modern distributed learning application as discussed before (Diamandis et al., 2021).

Remark 2. We further clarify that in the C-MLR model described above, the term "clustered" referrers to the fact that data samples are available in batches of size n where all the n samples in each batch share the same latent variable ξ. Though, it is worth noting that the folklore two-component MLR model with independent latent variables partitions the samples into two clusters, as well. However, we adopt the term "clustered" to particularly underscore the batched structure modeled in (1).

Remark 3. In our asymptotic analysis in this paper, we are interested in the regime that m and n grow while other problem parameters, that are $\|\theta^*\|$, σ, and d remain constant.

Our main goal in this paper is to answer the following:

> *What is iteration complexity of the sample-based EM algorithm to estimate the ground truth θ^* from m batches of samples, each of size n generated by the C-MLR described in (1)?*

We answer this question in this paper as follows. We assume that m batches of in total mn samples generated by the C-MLR model in (1) are available where m grows at most up to $e^{o(n)}$. We prove that if initialized within a constant-size neighbourhood of the ground truth θ^* and after $T = \mathcal{O}(1)$ iterations of the sample-based (or empirical) EM algorithm, either (i) there exists an iterate $0 \leq t \leq T$ of the algorithm for which $\|\theta_t - \theta^*\| \leq \mathcal{O}(\sqrt{d/(mn)})$; or (ii) the $\|\theta_T - \theta^*\| \leq \mathcal{O}(\sqrt{d/(mn)})$ with high probability. Our result is asymptotic, that is, it holds for sufficiently large n. To highlight this result, it is worth noting that the underlying clustered structure in C-MLR is essential for a constant iteration complexity. Indeed, if such a structure is discarded, the EM algorithm requires $\mathcal{O}(\log(mn/d))$ iterates to reach the same statistical accuracy.

Contribution. To summarize the above discussion, we consider a data heterogeneity structure observed in various distributed learning applications such as federated learning where a latent variable governs the distribution of all the samples generated on any node. In particular, we zoom in on a *clustered* two-component mixture of linear regression model described in (1) where all the linear measurements of any node share their binary latent variable. We utilize the EM algorithm to estimate the maximum likelihood regressor and establish asymptotic and local optimization and generalization guarantees for both population and empirical EM updates. Lastly, we employ these two results and asymptotically characterize the iteration complexity of the sample-based EM algorithm to estimate the ground truth parameters of the C-MLR model. All in all, our result demonstrates that employing the clustered nature of the distributed data improves the iteration complexity of EM method.

As it will become more clear in the paper, the EM update for the C-MLR model in (1) contains nonlinear terms which make the existing approaches such as Balakrishnan et al. (2017) inapplicable. To circumvent this challenge, we resort to the concentration of sub-exponential random variables to analyze both optimization and generalization errors.

Related Work. Studying convergence characteristics of Expectation-Maximization (EM) dates back to the seminal work of Wu (1983) in which asymptotic and local convergence of EM is established for general latent variable models. Balakrishnan et al. (2017) provides a general framework to analyze local onvergence of the EM algorithm in several settings such as mixture of linear regressions (MLR) and Gaussian mixture model (GMM). Several follow up works study GMM, MLR and Missing Covariate Regression (MCR) models including Yi and Caramanis (2015); Daskalakis et al. (2017); Li and Liang (2018); Klusowski et al. (2019); Ghosh and Kannan (2020); Yan et al. (2017).

Although it is not the main focus of this paper, global convergence of the EM method (with random initialization) has been extensively studied for Gaussian mixture model (Chen et al., 2019) and mixture of linear regressions (Kwon et al., 2019; Wu and Zhou, 2019). Another interesting direction is establishing statistical lower bounds on the accuracy of the EM method for the MLR model Kwon et al. (2021). Going beyond two-component MLR model, Kwon and Caramanis (2020) proves that well-initialized EM converges to the true regression parameters of k-component MLR in certain SNR regimes. In the same setting, Chen et al. (2020) proposes an algorithm that is sub-exponential in k. For noiseless MLR model, Yi et al. (2014, 2016) were among the first works to establish convergence guarantees for EM. To tackle the computational complexity of EM in learning MLR models, Li and Liang (2018); Zhong et al. (2016) propose gradient descent-type methods with nearly optimal sample complexity. From practical point of view, EM has demonstrated empirical success in MLR models (Jordan and Jacobs, 1994; De Veaux, 1989) and its simple implementation has made it a suitable choice in several applications (Chen and Li, 2009; Li et al., 2009).

2 PRELIMINARIES

In this section, we first review backgrounds on MLE and EM and then characterize the population and empirical EM updates for our C-MLR model followed by an insightful benchmark.

2.1 Maximum Likelihood Estimator and EM Algorithm

Population EM. Let us focus on one node observing n samples $\{(x_i, y_i)|i = 1, \cdots, n\}$ where we adopt the shorthand notations $x^n = (x_1, \cdots, x_n)$ and $y^n = (y_1, \cdots, y_n)$. Furthermore, let ξ denote the latent variables in the C-MLR model described in (1), respectively. To reiterate the underlying C-MLR model, we have that

$$y_i = \xi\langle x_i, \theta^*\rangle + \epsilon_i, \quad i = 1, \cdots, n. \tag{2}$$

As discussed before, in our setting, only the variables (x^n, y^n) are observed and the latent variable $\xi \in \Xi$ remains hidden. Suppose that the tuple (x^n, y^n, ξ) is generated by the joint distribution f_{θ^*} where $\{f_\theta|\theta \in \Omega\}$ and Ω is a non-empty compact convex set.

As our main goal in this paper, we aim to estimate the ground-truth model θ^* by maximizing the likelihood function, that is, finding $\hat{\theta} \in \Omega$ that maximizes the following likelihood

$$g_\theta(x^n, y^n) = \int_\Xi f_\theta(x^n, y^n, \xi)\mathrm{d}\xi.$$

In many settings, it is computationally expensive to compute the likelihood function $g_\theta(x^n, y^n)$, while computing log-likelihood $\log f_\theta(x^n, y^n, \xi)$ is relatively easier. The EM method is an iterative algorithm that aims to maximize a lower bound on the log-likelihood $\log g_\theta(\cdot, \cdot)$. This lower bound which is known as the Q-function can be written as follows

$$Q(\theta'|\theta) = \int_{\mathcal{X}^n \times \mathcal{Y}^n} \left(\int_\Xi f_\theta(\xi|x^n, y^n) \cdot \log f_{\theta'}(x^n, y^n, \xi)\mathrm{d}\xi \right) f_{\theta^*}(x^n, y^n)\mathrm{d}x^n\mathrm{d}y^n. \tag{3}$$

At each iteration of the empirical EM (Algorithm 1) and given the current estimate of the true model θ, the next model is obtained by maximizing the above Q-function, that is, $\theta \leftarrow M(\theta)$ where

$$M(\theta) := \arg\max_{\theta' \in \Omega} Q(\theta'|\theta). \tag{4}$$

Note that computing $M(\cdot)$ requires having access to the joint distribution f_{θ^*}, or to put it differently, observed data from infinitely many nodes ($m \to \infty$) is required. We call such variant of the EM algorithm *population EM* and discuss the *empirical* variant with finite clients (finite m) in the following section. Next proposition characterizes the M-function and the population EM update.

Proposition 2.1 (Population EM)**.** *Consider n linear measurements from the C-MLR model in (2) with Gaussian features $X_i \sim \mathcal{N}(0, I_d)$ and noises $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ with shared latent variable $\xi \sim \mathcal{U}\{\pm 1\}$. Then, the $M(\cdot)$ function of the population EM defined in (4) is as follows*

$$M(\theta) = \mathbb{E}\left[X_1 Y_1 \tanh\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\langle X_i, \theta\rangle Y_i\right)\right]. \tag{5}$$

Proof. We defer the proof to Appendix D.1. □

Note that equally likely $\xi \in \{\pm 1\}$ makes the distribution of Y^n symmetric given X^n. Moreover, $\tanh(\cdot)$ is an odd function and therefore, the expectation in (5) can also be taken with respect to $X_i \sim \mathcal{N}(0, I_d)$ and $Y_i|X_i \sim \mathcal{N}(\langle X_i, \theta^*\rangle, \sigma^2)$, *i.e.* no randomness in the the latent variable ξ.

Empirical EM. For a finite number of nodes m, the empirical EM algorithm updates the estimate of the true model using the empirical Q_m-function defined below

$$Q_m(\theta'|\theta) = \frac{1}{m}\sum_{j=1}^{m}\int_{\Xi} f_\theta(\xi|x_j^n, y_j^n) \log f_{\theta'}(x_j^n, y_j^n, \xi) \mathrm{d}\xi, \tag{6}$$

where samples are independent across different nodes. Similarly, in each iteration of the empirical EM algorithm (Algorithm 2), the current model estimate θ is updated to $\theta \leftarrow M_m(\theta)$ where

$$M_m(\theta) := \arg\max_{\theta' \in \Omega} Q_m(\theta'|\theta). \tag{7}$$

Next proposition characterises the empirical M_m-function defined in (7).

Algorithm 1 Population EM

```
Require: initialization $\theta_0$
  for $t = 0, 1, \cdots$ do
      Update $\theta_{t+1} = M(\theta_t)$ as defined in (4)
  end for
```

Algorithm 2 Empirical EM

```
Require: initialization $\theta_0$
  for $t = 0, 1, \cdots$ do
      Update $\theta_{t+1} = M_m(\theta_t)$ as defined in (7)
  end for
```

Proposition 2.2 (Empirical EM)**.** *Consider m nodes each observing n linear measurements generated by the C-MLR model in (1) denoted by $\{(x_i^j, y_i^j)|i = 1, \cdots, n,\ j = 1, \cdots, m\}$. Then, the $M_m(\cdot)$ function of the empirical EM defined in (7) can be computed as follows*

$$M_m(\theta) = \widehat{\Sigma}^{-1}\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n} x_i^j y_i^j \tanh\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\langle x_i^j, \theta\rangle y_i^j\right), \tag{8}$$

where $\widehat{\Sigma} := 1/(mn)\sum_{j=1}^{m}\sum_{i=1}^{n} x_i^j {x_i^j}^\top$ denotes the sample covariance matrix of the total mn observations.

Proof. We defer the proof to Appendix D.2. □

Our goal in the remainder of the paper is to rigorously study the optimization and generalization performance of the two population and empirical EM algorithms described above. Before that, let us elaborate on a simple and intuitive benchmark.

2.2 A Benchmark: EM with Independent Samples

As we described in our C-MLR model in (1), the measurements observed on a given node share the same latent variable, making them dependent. In contrast, the well-established literature on EM is centered around the i.i.d. setting where each sample is generated through a latent variable independent of the ones for any other sample. To be more precise, consider the setting where N i.i.d. linear measurements $\{(x_i, y_i)|i = 1, \cdots, N\}$ generated by a mixture of two component linear regression model are available. That is, $y_i = \xi_i\langle x_i, \theta^*\rangle + \epsilon_i$ for all $i = 1, \cdots, N$ where $\xi_i \sim \mathcal{U}\{\pm 1\}$, $x_i \sim \mathcal{N}(0, I_d)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. and mutually independent. In this setting, the population and empirical EM update rules are as follows

$$M(\theta) = \mathbb{E}\left[XY \tanh\left(\frac{1}{\sigma^2}\langle X, \theta\rangle Y\right)\right], \quad \text{and}$$

Table 1: Iteration complexity vs. accuracy for EM employed on clustered and independent samples. †For $m \leq e^{o(n)}$.

Method	Reference	Iteration complexity	Accuracy
EM with independent data	Balakrishnan et al. (2017)	$\mathcal{O}\left(\log(mn/d)\right)$	$\mathcal{O}\left(\sqrt{d\log(1/\delta)/(mn)}\right)$
EM with clustered data	**This paper**	$\mathcal{O}(1)^\dagger$	$\mathcal{O}\left(\sqrt{d\log(1/\delta)/(mn)}\right)$

$$M_N(\theta) = \widehat{\Sigma}^{-1}\frac{1}{N}\sum_{i=1}^{N} x_i y_i \tanh\left(\frac{1}{\sigma^2}\langle x_i, \theta\rangle y_i\right),$$

where the expectation is over $X \sim \mathcal{N}(0, I_d)$, $\xi \sim \mathcal{U}\{\pm 1\}$ and $Y|X, \xi \sim \mathcal{N}(\xi\langle X, \theta^*\rangle, \sigma^2)$. In above, $\widehat{\Sigma} = 1/N\sum_{i=1}^{N} x_i x_i^\top$ denotes the sample covariance matrix (Balakrishnan et al., 2017; Kwon et al., 2019). In particular, it was shown in Balakrishnan et al. (2017) that for any suitable initialization with $\|\theta_0 - \theta^*\| \leq \|\theta^*\|/32$, after $T = \log(N/d \cdot \|\theta^*\|^2/(\|\theta^*\|^2 + \sigma^2)) \cdot \mathcal{O}(1)$ iterations of empirical EM with update rule $M_N(\cdot)$ as above, the following sub-optimality is guaranteed with probability at least $1-\delta$,

$$\|\theta_T - \theta^*\| \leq \sqrt{\|\theta^*\|^2 + \sigma^2}\sqrt{\frac{d + \log(1/\delta)}{N}}\,\log\left(\frac{N}{d}\,\frac{\|\theta^*\|^2}{\|\theta^*\|^2 + \sigma^2}\right)\cdot\mathcal{O}(1).$$

Now, consider $N = mn$ linear measurements generated by the C-MLR model in (1) which we also denote by the same notation $\{(x_i, y_i) | i = 1, \cdots, N\}$. Clearly, the EM update rules in (2.2) may not be employed in this setting as samples are not independent due to the shared latent variables. However, one could make such N samples independent by the following simple trick. For each sample $i = 1, \cdots, N$, let us denote $\tilde{y}_i = \tilde{\xi}_i \cdot y_i$ where $\tilde{\xi}_i$s are independent Rademacher variables. In words, $\tilde{y}_i = y_i$ or $\tilde{y}_i = -y_i$ equally likely. It is straightforward to check that the new N samples $\{(x_i, \tilde{y}_i) | i = 1, \cdots, N\}$ are indeed independent. Therefore, one may employ the guarantee above and conclude that with a suitable initialization and after T iterations of EM (on the new samples), the final sub-optimality is with probability $1-\delta$ bounded by

$$\|\theta_T - \theta^*\| \leq \sqrt{\|\theta^*\|^2 + \sigma^2}\sqrt{\frac{d + \log(1/\delta)}{mn}} \cdot \tilde{\mathcal{O}}(1),$$
$$\text{where } T = \log\left(\frac{mn}{d} \cdot \frac{\|\theta^*\|^2}{\|\theta^*\|^2 + \sigma^2}\right) \cdot \mathcal{O}(1).$$

As mentioned before, we aim to characterize the complexity of the EM algorithm deployed on clustered samples per the C-MLR model described in (1). Before laying out our formal analysis, it is worth to highlight our main result here and compare it to the simple benchmark described above.

Theorem (Main, informal)**.** *Consider the empirical EM in Algorithm 2 with a constant* $\mathsf{snr} \geq 4$ *and any tolerance probability* $\delta \in (0,1)$. *Moreover, assume that* $mn \geq \mathcal{O}(d + \log(1/\delta))$ *and* $n \geq \mathcal{O}(\log(m) + d + \log(1/\delta))$. *Then, for a suitable initialization and sufficiently large* n*, after* $T = \mathcal{O}(1)$ *iterations of Algorithm 2, either*

(i) there exists an iterate $0 \leq t \leq T$ *such that*

$$\|\theta_t - \theta^*\| \leq \sqrt{\|\theta^*\|^2 + \sigma^2}\sqrt{\frac{d + \log(1/\delta)}{mn}},$$

(ii) or with probability at least $1-\delta$,

$$\|\theta_T - \theta^*\| \leq \sqrt{\|\theta^*\|^2 + \sigma^2}\sqrt{\frac{d + \log(1/\delta)}{mn}} \cdot \mathcal{O}(1).$$

Our result above demonstrates that incorporating the underlying clustered structure in the C-MLR model, EM requires only $\mathcal{O}(1)$ iterations to reach the statistical accuracy $\mathcal{O}(\sqrt{d/(mn)})$ under proper scaling assumptions. In contrast and as illustrated above, discarding such structure makes EM algorithm to run for $\mathcal{O}(\log(mn/d))$ iterations to reach the same accuracy. Table 1 summarizes the above discussion.

In the following sections, we prove this result by laying out optimization and generalization guarantees for the EM algorithm on samples generated by the C-MLR model.

3 ANALYSIS OF POPULATION AND EMPIRICAL EM UPDATES

3.1 Population EM Update

In this section, we consider the population EM updates in Algorithm 1 with the M operator characterized in (5) and establish optimization guarantees for it. Let us recall the population EM scenario and the underlying C-MLR model. Denoted by $\{(x_i, y_i) | i = 1, \cdots, n\}$ are n pairs of linear measurements generated according to the mixture model (2), that is, $y_i = \xi\langle x_i, \theta^*\rangle + \epsilon_i$ for all $i = 1, \cdots, n$. In Proposition 2.1, we characterised the population M-function and in the following theorem, we establish its contraction property.

Here and throughout the paper, we denote a Euclidean ball of radius r around the fixed point θ^* by $\mathbb{B}(r;\theta^*) \coloneqq \{\theta \in \Omega \,|\, \|\theta - \theta^*\| \leq r\}$.

Theorem 3.1. *Consider the population EM update rule M in (5) and assume that $\theta \in \mathbb{B}(\alpha\|\theta^*\|;\theta^*)$ for some constant $0 \leq \alpha < 1$. If $\|\theta - \theta^*\| \geq \varepsilon$, then there exist constants $N_0(\alpha, \mathsf{snr})$ and $C(\alpha, \mathsf{snr})$ depending on α and $\mathsf{snr} = \|\theta^*\|/\sigma$ such that for any $n \geq N_0(\alpha, \mathsf{snr})$ we have*

$$\|M(\theta) - \theta^*\| \leq \kappa \|\theta - \theta^*\|,$$
$$\text{for}\, \kappa = (\sqrt{d}\|\theta^*\| + \sigma)\Big(\mathsf{snr} + \frac{1}{n\varepsilon}\Big) \exp\left(-n \cdot C(\alpha, \mathsf{snr})\right).$$

Proof. We defer the proof to Appendix A. □

The result of this theorem reveals a number of insightful remarks as follows.

Remark 4. First, for any constant accuracy lower bound ε, as the number of samples per node n grows, the factor κ decreases and there exists a constant N_0 depending on the problem parameters such that for any $n \geq N_0$, the M-operator is a contraction, that is, $\kappa < 1$. Secondly and more importantly, it shows that if initialized within a ball around the ground truth model θ^*, iterates of the population EM in Algorithm 1 converge *linearly* in n till reaching the accuracy ε. The following corollary provides an informal but insightful implication of this theorem.

Corollary 3.1.1 (Informal)**.** *Suppose that the population EM in Algorithm 1 is initialized with θ_0 where $\|\theta_0 - \theta^*\| = \mathcal{O}(\|\theta^*\|)$. Then, for sufficiently large n and after $T = \mathcal{O}(1 + \log(n/d)/n) = \mathcal{O}(1)$ iterations, either there exists an iterate $0 \leq t \leq T$ for which $\|\theta_t - \theta^*\| = \mathcal{O}(\sqrt{d/n}\,\|\theta^*\|)$.*

While we provide the proof of Theorem 3.1 in Section A, it is worth elaborating on the proof technique as follows.

3.2 Proof Sketch

To establish optimization guarantees for the population EM iterates and Algorithm 1, we first adopt the *First-Order Stability* (FOS) notion (Balakrishnan et al., 2017) as defined below.

Definition 3.1 (First-Order Stability (FOS))**.** *The functions $\{Q(\cdot|\theta) | \theta \in \Omega\}$ satisfy condition FOS(γ) over $\mathbb{B}(r;\theta^*)$ if for all $\theta \in \mathbb{B}(r;\theta^*)$,*

$$\|\nabla Q(M(\theta)|\theta^*) - \nabla Q(M(\theta)|\theta)\| \leq \gamma \|\theta - \theta^*\|.$$

This property of the Q-function helps showing the contraction of the population EM operator M. The following general theorem from Balakrishnan et al. (2017) characterizes the conditions under which the population EM operator M is contractive.

Theorem 3.2 (Balakrishnan et al. (2017))**.** *For some radius $r > 0$ and pair (γ, λ) such that $0 \leq \gamma < \lambda$, suppose that the function $Q(\cdot|\theta^*)$ is λ-strongly concave, and that the FOS(γ) condition holds on the ball $\mathbb{B}(r;\theta^*)$. Then, the population EM operator M is contractive over $\mathbb{B}(r;\theta^*)$, in particular,*

$$\|M(\theta) - \theta^*\| \leq \frac{\gamma}{\lambda}\|\theta - \theta^*\|, \quad \textit{for all } \theta \in \mathbb{B}(r;\theta^*).$$

For the EM function in (5), we prove the first-order stability property in Definition 3.1 for a fixed θ. More precisely, for any $\theta \in \mathbb{B}(\alpha\|\theta^*\|;\theta^*)$, we show that for the population Q-function (3) the FOS(γ) property holds true with

$$\gamma = \frac{1}{\sigma^2}(\sqrt{d}\|\theta^*\| + \sigma)\Big(n \cdot \mathsf{snr} + \frac{1}{\varepsilon}\Big) \exp\left(-n \cdot C(\alpha, \mathsf{snr})\right),$$

as long as $\|\theta - \theta^*\| \geq \varepsilon$. On the other hand, it is straightforward to check that population Q-function is λ-strongly concave with $\lambda = n/\sigma^2$. This, together with the first-order stability and Theorem 3.2 yields the contractive property of the population M-function in Theorem 3.1.

3.3 Empirical EM Update

Having set up the optimization guarantees for the population EM (Algorithm 1) in the previous section, we move to the sample-based setting and establish generalization characteristics the empirical EM. Coupling these two results, we provide convergence guarantees of the (empirical) EM algorithm later in this section.

Let us recall the empirical setting of our interest where each node $j = 1, \cdots, m$ nodes observes n linear measurements denoted by $\{(x_i^j, y_i^j) | i = 1, \cdots, n\}$ and generated by the C-MLR model in (1), that is, $y_i^j = \xi^j \langle x_i^j, \theta^* \rangle + \epsilon_i^j$. In the following, we establish a uniform generalization error bound for the empirical EM update with finitely many nodes m and samples per node n.

Theorem 3.3 (Generalization gap)**.** *Consider the C-MLR model in (1) with $\mathsf{snr} \geq 4$, any tolerance probability $\delta \in (0, 1)$ and the empirical and population EM operators in (8) and (5) with $mn \geq 192^2(d + \log(8/\delta))$ and $n - 64\log m \geq 104(2d + \log(4/\delta))$. Then, with* *probability at least* $1-\delta$,

$$\sup_{\theta\in\mathbb{Sh}(\varepsilon,r;\theta^*)}\|M_m(\theta)-M(\theta)\|\leq \sqrt{\|\theta^*\|^2+\sigma^2}\sqrt{\frac{d+\log(1/\delta)}{mn}}\cdot\mathcal{O}(1+\kappa(\varepsilon)).$$

Here, the supermom is over the spherical shell

$$\mathbb{Sh}(\varepsilon,r;\theta^*)\coloneqq\{\theta\in\mathbb{R}^d:\varepsilon\leq\|\theta-\theta^*\|\leq r\},$$

with $r=\|\theta^*\|/14$ *and* $\kappa(\varepsilon)$ *is the contraction factor of the expected EM update characterized in Theorem 3.1, i.e.,*

$$\kappa(\varepsilon)=(\sqrt{d}\|\theta^*\|+\sigma)\left(\mathsf{snr}+\frac{1}{n\varepsilon}\right)\exp\left(-n\cdot C(\mathsf{snr})\right).$$

Proof. We defer the proof to Appendix B. □

Let us provide a useful implication of Theorem 3.3. Assume the signal-to-noise ratio is a constant larger than 1 and the total number of samples are at least $mn=\mathcal{O}(d+\log(1/\delta))$. Moreover, suppose that the number of nodes is at most $m=\exp(o(n))$, for instance, it grows at a rate polynomial in n. Now take the accuracy

$$\varepsilon_\ell=\sqrt{\|\theta^*\|^2+\sigma^2}\sqrt{\frac{d+\log(1/\delta)}{mn}},$$

which is particularly of our interest in this paper. This pick of the accuracy lower bound yields that for sufficiently large n, the expected EM update is contractive, i.e. $\kappa(\varepsilon_\ell)<1$. Now, we denote by $\varepsilon_\ell^{\mathrm{unif}}$ the smallest scalar for which

$$\sup_{\theta\in\mathbb{Sh}(\varepsilon_\ell,\frac{1}{14}\|\theta^*\|;\theta^*)}\|M_m(\theta)-M(\theta)\|\leq\varepsilon_\ell^{\mathrm{unif}}$$

with probability at least $1-\delta$. As a result of Theorem 3.3, we have with high probability that the supermom generalization gap $\|M_m(\theta)-M(\theta)\|$ over the spherical shell $\theta\in\mathbb{Sh}(\varepsilon_\ell,\|\theta^*\|/14;\theta^*)$ is at most $\varepsilon_\ell^{\mathrm{unif}}\leq C_\varepsilon\varepsilon_\ell$ for a constant $C_\varepsilon\geq 1$. To put it differently, for any parameter θ in a ball around θ^* with $\|\theta-\theta^*\|\leq\|\theta^*\|/14$, if $\|\theta-\theta^*\|\leq\varepsilon_\ell$, then θ is already a fairly accurate estimate of θ^*. Otherwise, Theorem 3.3 guarantees that the generalization error of the empirical EM update is with high probability bounded by a constant multiplicative factor of ε_ℓ.

4 MAIN RESULTS ON SAMPLE-BASED EM ALGORITHM

Having laid out the main two components of our analysis in Theorems 3.1 and 3.3, we are ready to formally state the main result of the paper.

Theorem 4.1 (Main)**.** *Consider the empirical EM update (8) with* $\mathsf{snr}\geq 4$ *and any tolerance probability* $\delta\in(0,1)$ *and suppose that the initialization* θ_0 *is in* $\mathbb{B}(r;\theta^*)$ *for* $r=\|\theta^*\|/14$. *Moreover, assume that* $mn\geq 192^2(d+\log(8/\delta))$ *and* $n\geq 64\log(m)+104(2d+\log(4/\delta))$ *while* n *is large enough that* $\kappa(\varepsilon_\ell)\leq 1/2$, $\kappa(\varepsilon_\ell)\leq\exp(-C_\kappa n)$ *for a constant* C_κ *and* $4C_\varepsilon\varepsilon_\ell\leq r/2$. *Then, after*

$$T=1+\frac{1}{2C_\kappa n}\log\left(mn\cdot\frac{1}{28C_\varepsilon}\cdot\frac{\|\theta^*\|^2}{\|\theta^*\|^2+\sigma^2}\cdot\frac{1}{d+\log(1/\delta)}\right)$$

iterations of Algorithm 2, either

(i) $\|\theta_t-\theta^*\|\leq\varepsilon_\ell$ *for some iteration* $t=0,1,\cdots,T$, *or*

(ii) $\|\theta_T-\theta^*\|\leq 4C_\varepsilon\varepsilon_\ell$ *with probability at least* $1-\delta$.

Remark 5. The result of Theorem 4.1 implies the following remarks. Let the empirical EM (Algorithm 2) be initialized with θ_0 where $\|\theta_0-\theta^*\|\leq\|\theta^*\|/14$. In addition, consider the C-MLR model in (1) with a constant SNR larger than 4 where m and n are such that $mn\geq\mathcal{O}(d+\log(1/\delta))$ and $n\geq\mathcal{O}(\log(m)+d+\log(1/\delta))$, that is, m grows at a rate no greater than $e^{o(n)}$. Then, Theorem 4.1 implies that for sufficiently large n and after

$$T=\mathcal{O}(1)+\frac{1}{n}\log\left(\frac{mn}{d}\cdot\frac{\|\theta^*\|^2}{\|\theta^*\|^2+\sigma^2}\right)\cdot\mathcal{O}(1)=\mathcal{O}(1)$$

iterations, either $\|\theta_t-\theta^*\|\leq\varepsilon_\ell$ for some iteration $t=0,1,\cdots,T$; or otherwise,

$$\|\theta_T-\theta^*\|\leq\mathcal{O}(\varepsilon_\ell)=\sqrt{\|\theta^*\|^2+\sigma^2}\sqrt{\frac{d+\log(1/\delta)}{mn}}\cdot\mathcal{O}(1),$$

with probability at least $1-\delta$. Note that since $m\leq e^{o(n)}$, then the iteration complexity is indeed bounded by a constant, that is,

$$T=\mathcal{O}(1+1/n\cdot\log(mn/d))=\mathcal{O}(1).$$

Remark 6. We would like to particularly highlight the fact that implications of the above theorem are two-folded. Theorem 4.1 shows that if the EM method in Algorithm 2 is applied to the mn samples generated by the C-MLR while honoring the underlying structure (i.e. shared latent variables for samples of any node), after only a constant number of iterations independent of the number of samples, the statistical accuracy $\mathcal{O}(\sqrt{d/(mn)})$ is attained with high probability. On the one hand and regarding the iteration complexity, this is a significant improvement over the benchmark described in Section 2.2 where the iteration complexity grows logarithmically with the number of samples. On the other hand, Theorem 4.1 guarantees that the statistical accuracy $\mathcal{O}(\sqrt{d/(mn)})$ is indeed achievable by the same EM algorithm.

4.1 Proof of Theorem 4.1

As mentioned in the theorem's statement, suppose that Algorithm 2 is initialized with θ_0 such that $\|\theta_0 - \theta^*\| \leq r = \|\theta^*\|/14$ and consider any iteration $t = 0, 1, \cdots$. We can write that

$$\begin{aligned}\|\theta_{t+1} - \theta^*\| &= \|M_m(\theta_t) - \theta^*\| \\ &\leq \|M(\theta_t) - \theta^*\| + \|M_m(\theta_t) - M(\theta_t)\|. \quad (9)\end{aligned}$$

Assume that for all iterates $0 \leq k \leq t$ we have $\|\theta_k - \theta^*\| > \varepsilon_\ell$, otherwise the theorem's first claim is concluded. Then from Theorem 3.1, for large enough n, we have $\|M(\theta_t) - \theta^*\| \leq \kappa(\varepsilon_\ell) \cdot \|\theta_t - \theta^*\|$ for

$$\kappa(\varepsilon_\ell) = (\sqrt{d}\|\theta^*\| + \sigma)\left(\mathsf{snr} + \frac{1}{n\varepsilon_\ell}\right)\exp\left(-n \cdot C(\mathsf{snr})\right).$$

In particular, note that

$$\begin{aligned}\frac{1}{n\varepsilon_\ell} &= \frac{1}{n}\left(\sqrt{\|\theta^*\|^2 + \sigma^2}\sqrt{\frac{d + \log(1/\delta)}{mn}}\right)^{-1} \\ &= \mathcal{O}\left(\sqrt{\frac{m}{n}}\right),\end{aligned}$$

and since m grows at a rate at most $m = \exp(o(n))$, there exists a constant C_κ that for large enough n, we have $\kappa(\varepsilon_\ell) \leq \exp(-C_\kappa n)$ and $\kappa(\varepsilon_\ell) \leq 1/2$.

In the course of the proof, we show by induction that the iterates remain in the r-neighbourhood of θ^*. Assume that for all iterates $0 \leq k \leq t$ we have $\|\theta_k - \theta^*\| \leq r$ and therefore, $\|M_m(\theta_t) - M(\theta_t)\| \leq \varepsilon_\ell^{\text{unif}}$ with probability at least $1 - \delta$. Plugging in (9) we have that with probability at least $1 - \delta$

$$\|\theta_{t+1} - \theta^*\| \leq e^{-C_\kappa n}\|\theta_t - \theta^*\| + \varepsilon_\ell^{\text{unif}} \quad (10)$$

Note that the above inequality also implies that $\|\theta_{t+1} - \theta^*\| \leq r/2 + r/2 = r$, where we used the fact that for large enough n, we have $\kappa(\varepsilon_\ell) \leq 1/2$. This concludes the induction argument described before, that is for any t, if $\|\theta_k - \theta^*\| > \varepsilon_\ell$ for all $0 \leq k \leq t$, then with probability at least $1 - \delta$, we have that $\|\theta_k - \theta^*\| \leq r$ for all $0 \leq k \leq t$. Now, consider the last iterate T and assume that $\|\theta_t - \theta^*\| > \varepsilon_\ell$ for all $0 \leq t \leq T$. We condition the rest of the analysis on the event

$$\{\|M_m(\theta_t) - M(\theta_t)\| \leq \varepsilon_\ell^{\text{unif}} \text{ for all } t = 0, \cdots, T-1\}$$

which happens with probability at least $1 - \delta$. Repeating the argument yielding to (10) implies that

$$\begin{aligned}\|\theta_T - \theta^*\| &\leq e^{-C_\kappa n T}\|\theta_0 - \theta^*\| + \sum_{t=0}^{T}\left(\frac{1}{2}\right)^t \varepsilon_\ell^{\text{unif}} \\ &\leq e^{-C_\kappa n T}\frac{\|\theta^*\|}{14} + 2C_\varepsilon\varepsilon_\ell.\end{aligned}$$

Balancing the two terms above yields that after T iterations for

$$\begin{aligned}T &= \frac{1}{C_\kappa n}\log\left(\frac{\|\theta^*\|}{28C_\varepsilon\varepsilon_\ell}\right) \\ &= \frac{1}{2C_\kappa n}\log\left(mn \cdot \frac{1}{28C_\varepsilon} \cdot \frac{\|\theta^*\|^2}{\|\theta^*\|^2 + \sigma^2} \cdot \frac{1}{d + \log(1/\delta)}\right),\end{aligned}$$

we have with probability at least $1 - \delta$ that

$$\begin{aligned}\|\theta_T - \theta^*\| &\leq 4C_\varepsilon\varepsilon_\ell \\ &= 4C_\varepsilon\sqrt{\|\theta^*\|^2 + \sigma^2}\sqrt{\frac{d + \log(1/\delta)}{mn}} \\ &= \sqrt{\|\theta^*\|^2 + \sigma^2}\sqrt{\frac{d + \log(1/\delta)}{mn}} \cdot \mathcal{O}(1).\end{aligned}$$

Note that Algorithm 2 has to iterate at least for one iteration and since $m = e^{o(n)}$, therefore we can write that $T = \mathcal{O}(1 + 1/n \cdot \log(mn/d)) = \mathcal{O}(1)$.

5 CONCLUSION

Data heterogeneity is a major challenge in scaling up distributed learning frameworks such as federated learning. However, there exist underlying structures in the data generation model of such paradigms that can be employed. In this paper, we focus on a particular model of two-component mixture of linear regressions where m batches of samples each containing n samples with identical latent variable are available. Expectation-Maximization is a popular method to estimate parameters of models with latent variables, while its theoretical analysis is typically complicated. We provide optimization and generalization guarantees for EM algorithm on clustered samples which enables us to characterize its iteration complexity to estimate he true parameters. An interesting follow-up of our work is to implement the EM algorithm in a distributed fashion which is aligned with modern applications such as federated learning. While new challenges such as consensus of local estimates arise, we believe that our techniques and analysis in this paper will be highly applicable.

6 ACKNOWLEDGMENT

This work was supported, in part, by MIT-DSTA grant 031017-00016 and the MIT-IBM Watson AI Lab.

References

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

Jiahua Chen and Pengfei Li. Hypothesis test for normal mixture models: The em approach. 2009.

Sitan Chen, Jerry Li, and Zhao Song. Learning mixtures of linear regressions in subexponential time via fourier moments. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 587–600, 2020.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176:5–37, 2019.

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

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

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Theo Diamandis, Yonina Eldar, Alireza Fallah, Farzan Farnia, and Asuman Ozdaglar. A wasserstein minimax framework for mixed linear regression. In *International Conference on Machine Learning*, pages 2697–2706. PMLR, 2021.

Avishek Ghosh and Ramchandran Kannan. Alternating minimization converges super-linearly for mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1093–1103. PMLR, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

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

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

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

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

P Li, J Chen, and P Marriott. Non-finite fisher information and homogeneity: an em approach. *Biometrika*, 96(2):411–426, 2009.

Yuanzhi Li and Yingyu Liang. Learning mixtures of linear regressions with nearly optimal complexity. In *Conference On Learning Theory*, pages 1125–1144. PMLR, 2018.

Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*, pages 1–18, 2017.

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

Alexander Robey, Hamed Hassani, and George J Pappas. Model-based robust deep learning: Generalizing to natural, out-of-distribution data. *arXiv preprint arXiv:2005.10247*, 2020.

Kert Viele and Barbara Tong. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12: 315–330, 2002.

CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

Yihong Wu and Harrison H Zhou. Randomly initialized em algorithm for two-component gaussian mixture achieves near optimality in $O(\sqrt{n})$ iterations. *arXiv preprint arXiv:1908.10935*, 2019.

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

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

Xinyang Yi and Constantine Caramanis. Regularized em algorithms: A unified framework and statistical guarantees. *Advances in Neural Information Processing Systems*, 28, 2015.

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

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *arXiv preprint arXiv:1608.05749*, 2016.

Kai Zhong, Prateek Jain, and Inderjit S Dhillon. Mixed linear regression with multiple components. *Advances in neural information processing systems*, 29, 2016.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials for

EM for Mixture of Linear Regression with Clustered Data

A Proof of Theorem 3.1

We first show the first-order stability of the population Q-function (3) and then employ the result of Theorem 3.2 to conclude the contractive property of the operator M. As we will show in the proof of Proposition 2.1, the gradient of the function $Q(\theta'|\theta)$ (with respect to θ') is as follows,

$$\nabla Q(\theta'|\theta) = -\frac{n}{\sigma^2}\theta' + \mathbb{E}\Bigg[\frac{1}{\sigma^2}\sum_{i=1}^{n} X_iY_i \tanh\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\langle X_i,\theta\rangle Y_i\right)\Bigg], \tag{11}$$

where the expectation is over i.i.d. feature vectors $X_i \sim \mathcal{N}(0, I_d)$ and response variables $Y_i = \langle X_i, \theta^*\rangle + \epsilon_i$ with i.i.d. Gaussian noises $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$ for $i \in \{1,\cdots,n\}$. To ease the presentation, we use the following short-hand notation throughout the paper,

$$Z \coloneqq \sum_{i=1}^{n} X_iY_i.$$

Therefore, we can rewrite the gradient of the Q-functions as follows

$$\begin{aligned}\nabla Q(\theta'|\theta) &= -\frac{n}{\sigma^2}\theta' + \mathbb{E}\Bigg[\frac{1}{\sigma^2}\sum_{i=1}^{n} X_iY_i \tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right)\Bigg] \\ &= -\frac{n}{\sigma^2}\theta' + \frac{n}{\sigma^2}\mathbb{E}\Bigg[X_1Y_1 \tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right)\Bigg],\end{aligned} \tag{12}$$

where we used the fact that the expectation in (11) is symmetric with respect to indices $i \in [n]$. Now, we plug in $\theta, M(\theta)$ and θ^* in (12) and write

$$\begin{aligned}&\big\|\nabla Q(M(\theta)|\theta^*) - \nabla Q(M(\theta)|\theta)\big\| \\ &= \frac{n}{\sigma^2}\Bigg\|\mathbb{E}\Bigg[X_1Y_1\Bigg(\tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right) - \tanh\left(\frac{1}{\sigma^2}\langle Z,\theta^*\rangle\right)\Bigg)\Bigg]\Bigg\| \\ &= \frac{n}{\sigma^2}\max_{\beta:\|\beta\|=1}\mathbb{E}\Bigg[\langle X_1,\beta\rangle Y_1\Bigg(\tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right) - \tanh\left(\frac{1}{\sigma^2}\langle Z,\theta^*\rangle\right)\Bigg)\Bigg] \\ &\le \frac{n}{\sigma^2}\max_{\beta:\|\beta\|=1}\Bigg|\mathbb{E}\Bigg[\langle X_1,\beta\rangle Y_1\Bigg(\tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right) - \tanh\left(\frac{1}{\sigma^2}\langle Z,\theta^*\rangle\right)\Bigg)\Bigg]\Bigg| \\ &\le \frac{n}{\sigma^2}\max_{\beta:\|\beta\|=1}\sqrt{\mathbb{E}[\langle X_1,\beta\rangle^2 Y_1^2]}\sqrt{T_1},\end{aligned} \tag{13}$$

where in the last step above, we used Cauchy–Schwarz inequality and the following short-hand notation,

$$T_1 \coloneqq \mathbb{E}\Bigg[\Bigg(\tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right) - \tanh\left(\frac{1}{\sigma^2}\langle Z,\theta^*\rangle\right)\Bigg)^2\Bigg].$$

In the following, we bound both terms in (13), starting with the first term. According to the regression model $Y_1 = \langle X_1,\theta^*\rangle + \epsilon_1$, we can write for any unit-norm β that

$$\begin{aligned}\mathbb{E}[\langle X_1,\beta\rangle^2 Y_1^2] &= \mathbb{E}[\langle X_1,\beta\rangle^2\langle X_1,\theta^*\rangle^2] + \mathbb{E}[\langle X_1,\beta\rangle^2\epsilon_1^2] \\ &\overset{(a)}{\le} 3\|\beta\|^2\|\theta^*\|^2 + \sigma^2\|\beta\|^2 \\ &= 3\|\theta^*\|^2 + \sigma^2,\end{aligned}$$

where in (a) we used Lemma 5 from Balakrishnan et al. (2017) which shows that for Gaussian vector $X_1 \sim \mathcal{N}(0, I_d)$ and any two fixed vectors β, θ, we have $\mathbb{E}[\langle X_1, \beta\rangle^2 \langle X_1, \theta\rangle^2] \leq 3\|\beta\|^2\|\theta\|^2$. Therefore,

$$\max_{\beta:\|\beta\|=1} \sqrt{\mathbb{E}[\langle X_1, \beta\rangle^2 Y_1^2]} \leq \sqrt{3}\|\theta^*\|+\sigma.$$

Next, we upper bound the second terms in (13), that is T_1. We begin by defining the following three good events for a given $\theta \in \mathbb{B}(\alpha\|\theta^*\|; \theta^*)$

$$\begin{aligned}
\mathcal{E}_1 &= \left\{ \langle Z, \theta\rangle \geq \frac{n}{4}(1-\alpha)\|\theta^*\|^2 \right\}, \\
\mathcal{E}_2 &= \left\{ \langle Z, \theta^*\rangle \geq \frac{n}{4}\|\theta^*\|^2 \right\}, \\
\mathcal{E}_3 &= \left\{ |\langle Z, \theta\rangle - \langle Z, \theta^*\rangle| \leq 3n\|\theta^*\|\|\theta - \theta^*\| \right\},
\end{aligned}$$

and letting $\mathcal{E}$ denote their intersection, that is, $\mathcal{E} \coloneqq \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$. Now, we can write that

$$\begin{aligned}
T_1 &= \mathbb{E}\left[\left(\tanh\left(\frac{1}{\sigma^2}\langle Z, \theta\rangle\right) - \tanh\left(\frac{1}{\sigma^2}\langle Z, \theta^*\rangle\right)\right)^2\right] \\
&\leq \mathbb{E}\left[\left(\tanh\left(\frac{1}{\sigma^2}\langle Z, \theta\rangle\right) - \tanh\left(\frac{1}{\sigma^2}\langle Z, \theta^*\rangle\right)\right)^2 \middle| \mathcal{E}\right] \\
&\quad + \mathbb{E}\left[\left(\tanh\left(\frac{1}{\sigma^2}\langle Z, \theta\rangle\right) - \tanh\left(\frac{1}{\sigma^2}\langle Z, \theta^*\rangle\right)\right)^2 \middle| \mathcal{E}^c\right] \cdot \mathbb{P}(\mathcal{E}^c). \qquad (14)
\end{aligned}$$

The first term above can be bounded as follows,

$$\begin{aligned}
&\mathbb{E}\left[\left(\tanh\left(\frac{1}{\sigma^2}\langle Z, \theta\rangle\right) - \tanh\left(\frac{1}{\sigma^2}\langle Z, \theta^*\rangle\right)\right)^2 \middle| \mathcal{E}\right] \\
&= \mathbb{E}\left[|\langle Z, \theta\rangle - \langle Z, \theta^*\rangle|^2 \left(\frac{\tanh\left(\frac{1}{\sigma^2}\langle Z, \theta\rangle\right) - \tanh\left(\frac{1}{\sigma^2}\langle Z, \theta^*\rangle\right)}{\langle Z, \theta\rangle - \langle Z, \theta^*\rangle}\right)^2 \middle| \mathcal{E}\right] \\
&\overset{(a)}{\leq} \frac{1}{\sigma^2}\mathbb{E}\left[|\langle Z, \theta\rangle - \langle Z, \theta^*\rangle|^2 \left(1 - \tanh^2\left(\frac{1}{\sigma^2}\min\left\{\langle Z, \theta\rangle, \langle Z, \theta^*\rangle\right\}\right)\right)^2 \middle| \mathcal{E}\right] \\
&\leq 9n^2 \frac{\|\theta^*\|^2}{\sigma^2}\|\theta - \theta^*\|^2 \left(1 - \tanh^2\left(\frac{n}{4}(1-\alpha)\frac{\|\theta^*\|^2}{\sigma^2}\right)\right)^2 \\
&\leq 144n^2 \mathsf{snr}^2 \|\theta - \theta^*\|^2 \exp\left(-n(1-\alpha)\mathsf{snr}^2\right), \qquad (15)
\end{aligned}$$

where in (a) we used the following inequality (stated and proved in Lemma D.2),

$$\frac{\tanh(x_2) - \tanh(x_1)}{x_2 - x_1} \leq \max\{1 - \tanh^2(x_1), 1 - \tanh^2(x_2)\}, \quad \text{for all } x_1, x_2 \geq 0.$$

The second term in the RHS of (14) can be bounded as follows,

$$\begin{aligned}
&\mathbb{E}\left[\left(\tanh\left(\frac{1}{\sigma^2}\langle Z, \theta\rangle\right) - \tanh\left(\frac{1}{\sigma^2}\langle Z, \theta^*\rangle\right)\right)^2 \middle| \mathcal{E}^c\right] \cdot \mathbb{P}(\mathcal{E}^c) \\
&\leq 4\mathbb{P}(\mathcal{E}^c) \\
&\overset{(b)}{\leq} 8\exp\left(-\frac{n}{32}\left(\frac{1-\alpha}{1+\alpha}\right)^2\right) + 8\exp\left(-n\min\left\{\frac{1}{16}\left(\frac{1-\alpha}{1+\alpha}\right)^2\mathsf{snr}^2, \frac{1}{8}\left(\frac{1-\alpha}{1+\alpha}\right)\mathsf{snr}\right\}\right) \\
&\quad + 8\exp\left(-\frac{n}{32}\right) + 8\exp\left(-n\min\left\{\frac{1}{16}\mathsf{snr}^2, \frac{1}{8}\mathsf{snr}\right\}\right) \\
&\quad + 8\exp\left(-\frac{n}{8}\right) + 8\exp\left(-n\min\left\{\mathsf{snr}^2, \frac{\mathsf{snr}}{2}\right\}\right) \\
&\leq \exp(-2c_1 n), \qquad (16)
\end{aligned}$$

for a constant c_1 depending on α and snr. In inequality (b) above, we used the high probability of the good event $\mathcal{E}$ stated and proved in Lemma A.1. Putting (15) and (16) in (14) yields that

$$\begin{aligned}\frac{T_1}{\|\theta-\theta^*\|^2} &= \frac{1}{\|\theta-\theta^*\|^2}\mathbb{E}\left[\left(\tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right)-\tanh\left(\frac{1}{\sigma^2}\langle Z,\theta^*\rangle\right)\right)^2\right]\\ &\leq 144n^2\mathsf{snr}^2\exp\left(-n(1-\alpha)\mathsf{snr}^2\right)+\frac{\exp(-2c_1n)}{\|\theta-\theta^*\|^2}\\ &\leq 144n^2\mathsf{snr}^2\exp\left(-n(1-\alpha)\mathsf{snr}^2\right)+\varepsilon^{-2}\exp(-2c_1n),\end{aligned}$$

where we assume that $\|\theta-\theta^*\|\geq\varepsilon$. Therefore,

$$\frac{\sqrt{T_1}}{\|\theta-\theta^*\|}\leq 12\mathsf{snr}\cdot n\exp\left(-\frac{n}{2}(1-\alpha)\mathsf{snr}^2\right)+\varepsilon^{-1}\exp(-c_1n)$$

Putting all together in (13), we have the following FOS satisfied

$$\|\nabla Q(M(\theta)|\theta^*)-\nabla Q(M(\theta)|\theta)\|\leq\gamma\|\theta-\theta^*\|,$$

for every θ in $\mathbb{B}(\alpha\|\theta^*\|;\theta^*)$ such that $\|\theta-\theta^*\|\geq\varepsilon$. Here, the γ parameter is

$$\begin{aligned}\gamma &= \frac{1}{\sigma^2}\left(12n\frac{\|\theta^*\|}{\sigma}\exp\left(-\frac{n}{2}(1-\alpha)\mathsf{snr}^2\right)+\varepsilon^{-1}\exp(-c_1n)\right)\cdot(\sqrt{3}\|\theta^*\|+\sigma)\\ &\leq\frac{1}{\sigma^2}(\|\theta^*\|+\sigma)\left(n\cdot\mathsf{snr}+\frac{1}{\varepsilon}\right)\exp(-n\cdot C(\alpha,\mathsf{snr})),\end{aligned}$$

and for $n\geq N_0(\alpha,\mathsf{snr})$ where both $N_0(\alpha,\mathsf{snr})$ and $C(\alpha,\mathsf{snr})$ are constants depending on c_1, α and snr (and therefore depending on α and snr). Moreover, $Q(\cdot|\theta^*)$ is λ–strongly concave with $\lambda=n/\sigma^2$. Following the proof of Theorem 1 in Balakrishnan et al. (2017), it can be shown that for every θ in $\mathbb{B}(\alpha\|\theta^*\|;\theta^*)$ with $\|\theta-\theta^*\|\geq\varepsilon$, we have

$$\|M(\theta)-\theta^*\|\leq\frac{\gamma}{\lambda}\|\theta-\theta^*\|,$$

where in our case

$$\frac{\gamma}{\lambda}\leq\kappa:=(\|\theta^*\|+\sigma)\left(\mathsf{snr}+\frac{1}{n\varepsilon}\right)\exp(-n\cdot C(\alpha,\mathsf{snr})),$$

which concludes Theorem 3.1's claim. It is worth noting that the constraint $\|\theta-\theta^*\|\geq\varepsilon$ in our case does not affect the conclusion of Theorem 1 in Balakrishnan et al. (2017).

A.1 Useful lemmas and proofs

Lemma A.1. *Assume that $\|\theta-\theta^*\|\leq\alpha\|\theta^*\|$ for some $0\leq\alpha<1$ and let $\mathsf{snr}=\|\theta^*\|/\sigma$ denote the SNR. Then, the following three events are high probability,*

$$\begin{aligned}\mathcal{E}_1 &= \left\{\langle Z,\theta\rangle\geq\frac{n}{4}(1-\alpha)\|\theta^*\|^2\right\},\\ \mathcal{E}_2 &= \left\{\langle Z,\theta^*\rangle\geq\frac{n}{4}\|\theta^*\|^2\right\},\\ \mathcal{E}_3 &= \left\{|\langle Z,\theta\rangle-\langle Z,\theta^*\rangle|\leq 3n\|\theta^*\|\|\theta-\theta^*\|\right\}.\end{aligned}$$

In particular, we have that

$$\begin{aligned}\mathbb{P}(\mathcal{E}_1) &\geq 1-2\exp\left(-\frac{n}{32}\left(\frac{1-\alpha}{1+\alpha}\right)^2\right)-2\exp\left(-n\min\left\{\frac{1}{16}\left(\frac{1-\alpha}{1+\alpha}\right)^2\mathsf{snr}^2,\frac{1}{8}\left(\frac{1-\alpha}{1+\alpha}\right)\mathsf{snr}\right\}\right),\\ \mathbb{P}(\mathcal{E}_2) &\geq 1-2\exp\left(-\frac{n}{32}\right)-2\exp\left(-n\min\left\{\frac{1}{16}\mathsf{snr}^2,\frac{1}{8}\mathsf{snr}\right\}\right),\\ \mathbb{P}(\mathcal{E}_3) &\geq 1-2\exp\left(-\frac{n}{8}\right)-2\exp\left(-n\min\left\{\mathsf{snr}^2,\frac{\mathsf{snr}}{2}\right\}\right).\end{aligned}$$

A.2 Proof of Lemma A.1

To prove the first two parts of the lemma, we employ the following result on the concentration of sub-exponential random variables.

Lemma A.2. *Consider the linear regression model $Y_i = \langle \theta^*, X_i \rangle + \epsilon_i$ where $X_i \sim \mathcal{N}(0, I)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ are independent and $i \in \{1, \cdots, n\}$. Also, assume that $\|\theta - \theta^*\| \leq \alpha \|\theta^*\|$ for some $0 \leq \alpha < 1$. Then, for any n,*

$$\left| \sum_{i=1}^{n} \langle X_i, \theta \rangle \langle X_i, \theta^* \rangle - n\langle \theta, \theta^* \rangle \right| \leq \frac{n}{2} \langle \theta, \theta^* \rangle, \ \text{w.p. at least } 1 - 2\exp\left(-\frac{n}{32} \left(\frac{1-\alpha}{1+\alpha} \right)^2 \right),$$

and

$$\left| \sum_{i=1}^{n} \langle X_i, \theta \rangle \epsilon_i \right| \leq \frac{n}{4} \langle \theta, \theta^* \rangle, \ \text{w.p. at least } 1 - 2\exp\left(-n \min\left\{ \frac{1}{16} \left(\frac{1-\alpha}{1+\alpha} \right)^2 \mathsf{snr}^2, \frac{1}{8} \left(\frac{1-\alpha}{1+\alpha} \right) \mathsf{snr} \right\} \right),$$

where $\mathsf{snr} \coloneqq \|\theta^*\| / \sigma$.

A.3 Proof of Lemma A.2

Let us denote signal random variables $S_i \coloneqq \langle X_i, \theta \rangle \langle X_i, \theta^* \rangle$ where $\mathbb{E}[S_i] = \langle \theta, \theta^* \rangle \geq (1-\alpha) \|\theta^*\|^2 > 0$ for each $i \in \{1, \cdots, n\}$. As shown in Lemma A.3, S_is are i.i.d. sub-exponential, in particular, $S_i \sim \mathsf{SubE}(4\|\theta\|^2 \|\theta^*\|^2, 4\|\theta\|\|\theta^*\|)$. This yields that

$$\sum_{i=1}^{n} \langle X_i, \theta \rangle \langle X_i, \theta^* \rangle = \sum_{i=1}^{n} S_i \sim \mathsf{SubE}\left(4n\|\theta\|^2 \|\theta^*\|^2, 4\|\theta\|\|\theta^*\| \right),$$

and therefore for any $t \geq 0$, we have the following concentration of sum of S_is around its mean value, that is,

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} S_i - n\langle \theta, \theta^* \rangle \right| \geq nt \right) \leq 2\exp\left(-\min\left\{ \frac{nt^2}{8\|\theta\|^2 \|\theta^*\|^2}, \frac{nt}{8\|\theta\|\|\theta^*\|} \right\} \right).$$

We pick $t = \langle \theta, \theta^* \rangle / 2$ which yields that

$$\begin{aligned} \min\left\{ \frac{nt^2}{8\|\theta\|^2 \|\theta^*\|^2}, \frac{nt}{8\|\theta\|\|\theta^*\|} \right\} &= \min\left\{ \frac{n}{32} \left(\frac{\langle \theta, \theta^* \rangle}{\|\theta\|\|\theta^*\|} \right)^2, \frac{n}{16} \left(\frac{\langle \theta, \theta^* \rangle}{\|\theta\|\|\theta^*\|} \right) \right\} \\ &= \frac{n}{32} \left(\frac{\langle \theta, \theta^* \rangle}{\|\theta\|\|\theta^*\|} \right)^2 \\ &\geq \frac{n}{32} \left(\frac{1-\alpha}{1+\alpha} \right)^2. \end{aligned}$$

Therefore, for any n we have that

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} S_i - n\langle \theta, \theta^* \rangle \right| \leq \frac{n}{2} \langle \theta, \theta^* \rangle \right) \geq 1 - 2\exp\left(-\frac{n}{32} \left(\frac{1-\alpha}{1+\alpha} \right)^2 \right). \tag{17}$$

Next, we define i.i.d. noise signals $N_i \coloneqq \langle X_i, \theta \rangle \epsilon_i$ for $i \in \{1, \cdots, n\}$ where we have $\mathbb{E}[N_i] = 0$. As we show in Lemma A.3, N_is are sub-exponential random variables with parameters $N_i \sim \mathsf{SubE}(\|\theta\|^2 \sigma^2 / 2, \|\theta\| \sigma)$. Now, we can write concentration for sum of N_is as follows. We have that

$$\sum_{i=1}^{n} \langle X_i, \theta \rangle \epsilon_i = \sum_{i=1}^{n} N_i \sim \mathsf{SubE}\left(\frac{1}{2} n\|\theta\|^2 \sigma^2, \|\theta\| \sigma \right)$$

and therefore for any $t \geq 0$, we have

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} N_i \right| \geq nt \right) \leq 2\exp\left(-\min\left\{ \frac{nt^2}{\|\theta\|^2 \sigma^2}, \frac{nt}{2\|\theta\| \sigma} \right\} \right).$$

In particular, it yields for $t = \langle \theta, \theta^* \rangle / 4$ that

$$\begin{aligned}\min\left\{\frac{nt^2}{\|\theta\|^2\sigma^2}, \frac{nt}{2\|\theta\|\sigma}\right\} &= \min\left\{\frac{n}{16}\left(\frac{\langle\theta,\theta^*\rangle}{\|\theta\|\sigma}\right)^2, \frac{n}{8}\left(\frac{\langle\theta,\theta^*\rangle}{\|\theta\|\sigma}\right)\right\} \\ &\geq n\min\left\{\frac{1}{16}\left(\frac{1-\alpha}{1+\alpha}\right)^2\mathsf{snr}^2, \frac{1}{8}\left(\frac{1-\alpha}{1+\alpha}\right)\mathsf{snr}\right\},\end{aligned}$$

and consequently,

$$\mathbb{P}\left(\left|\sum_{i=1}^n N_i\right| \leq \frac{n}{4}\langle\theta,\theta^*\rangle\right) \geq 1 - 2\exp\left(-n\min\left\{\frac{1}{16}\left(\frac{1-\alpha}{1+\alpha}\right)^2\mathsf{snr}^2, \frac{1}{8}\left(\frac{1-\alpha}{1+\alpha}\right)\mathsf{snr}\right\}\right). \tag{18}$$

Putting the two high probability events in (17) and (18) implies that event $\mathcal{E}_1$ holds, that is,

$$\begin{aligned}\langle Z,\theta\rangle &= \sum_{i=1}^n \langle X_i,\theta\rangle Y_i \\ &= \sum_{i=1}^n \langle X_i,\theta\rangle\langle X_i,\theta^*\rangle + \sum_{i=1}^n \langle X_i,\theta\rangle\epsilon_i \\ &\geq \frac{n}{4}\langle\theta,\theta^*\rangle \\ &\geq \frac{n}{4}(1-\alpha)\|\theta^*\|^2,\end{aligned}$$

with probability $\mathbb{P}(\mathcal{E}_1)$ stated in the lemma. As a particular case, we set $\theta = \theta^*$ (and thus $\alpha = 0$) in above and conclude the high probability of event $\mathcal{E}_2$. Next, we move to show the high probability of event $\mathcal{E}_3$. We have that

$$\langle Z,\theta\rangle - \langle Z,\theta^*\rangle = \sum_{i=1}^n \langle X_i,\theta-\theta^*\rangle Y_i = \sum_{i=1}^n \langle X_i,\theta-\theta^*\rangle\langle X_i,\theta^*\rangle + \sum_{i=1}^n \langle X_i,\theta-\theta^*\rangle\epsilon_i.$$

From Lemma A.3, we have

$$\sum_{i=1}^n \langle X_i,\theta-\theta^*\rangle\langle X_i,\theta^*\rangle \sim \mathsf{SubE}(4\|\theta^*\|^2\|\theta-\theta^*\|^2, 4\|\theta^*\|\|\theta-\theta^*\|).$$

Therefore, with probability at least $1 - 2\exp(-n/8)$,

$$\left|\sum_{i=1}^n \langle X_i,\theta-\theta^*\rangle\langle X_i,\theta^*\rangle - n\langle\theta-\theta^*,\theta^*\rangle\right| \leq n\|\theta^*\|\|\theta-\theta^*\|,$$

implying that

$$\left|\sum_{i=1}^n \langle X_i,\theta-\theta^*\rangle\langle X_i,\theta^*\rangle\right| \leq n|\langle\theta-\theta^*,\theta^*\rangle| + n\|\theta^*\|\|\theta-\theta^*\| \leq 2n\|\theta^*\|\|\theta-\theta^*\|.$$

On the other hand,

$$\sum_{i=1}^n \langle X_i,\theta-\theta^*\rangle\epsilon_i \sim \mathsf{SubE}\left(\frac{1}{2}n\|\theta-\theta^*\|^2\sigma^2, \|\theta-\theta^*\|\sigma\right)$$

Therefore,

$$\left|\sum_{i=1}^n \langle X_i,\theta-\theta^*\rangle\epsilon_i\right| \leq n\|\theta-\theta^*\|\|\theta^*\|,$$

with probability at least

$$1 - 2\exp\left(-n\min\left\{\mathsf{snr}^2, \frac{\mathsf{snr}}{2}\right\}\right).$$

Therefore,

$$\begin{aligned}\mathbb{P}(\mathcal{E}_3) &= \mathbb{P}\left(\left\{|\langle Z,\theta\rangle - \langle Z,\theta^*\rangle| \leq 3n\|\theta^*\|\|\theta-\theta^*\|\right\}\right)\\ &\geq 1-2\exp\left(-\frac{n}{8}\right) - 2\exp\left(-n\min\left\{\mathsf{snr}^2, \frac{\mathsf{snr}}{2}\right\}\right).\end{aligned}$$

Lemma A.3. *Let $X \sim \mathcal{N}(0, I_d)$ be Gaussian. Then, for any $u, v \in \mathbb{R}^d$, $\langle X,u\rangle\langle X,v\rangle$ is sub-exponential* $\mathsf{SubE}(\tau^2, b)$ *with*

$$\tau^2 = 4\|u\|^2\|v\|^2, \quad \textit{and} \quad b = 4\|u\|\|v\|.$$

Also, assume that $\epsilon \sim \mathcal{N}(0,\sigma^2)$ is independent of X. Then, $\langle X,u\rangle\epsilon \sim \mathsf{SubE}(\|u\|^2\sigma^2/2, \|u\|\sigma)$.

A.4 Proof of Lemma A.3

Define $S = \langle X,u\rangle\langle X,v\rangle$. We can write

$$\frac{S}{\|u\|\|v\|} = \langle X, u/\|u\|\rangle\langle X, v/\|v\|\rangle = \langle X,\overline{u}\rangle\langle X,\overline{v}\rangle = \frac{1}{4}\langle X,\overline{u}+\overline{v}\rangle^2 - \frac{1}{4}\langle X,\overline{u}-\overline{v}\rangle^2,$$

where we denote $\overline{u} := u/\|u\|$ and $\overline{v} := v/\|v\|$. Moreover, $S_1 := \langle X,\overline{u}+\overline{v}\rangle$ and $S_2 := \langle X,\overline{u}-\overline{v}\rangle$ are zero-mean Gaussian RVs and

$$\mathbb{E}[S_1S_2] = \mathbb{E}[(\overline{u}+\overline{v})^\top XX^\top(\overline{u}-\overline{v})] = \langle \overline{u}+\overline{v}, \overline{u}-\overline{v}\rangle = 0,$$

implying that S_1 and S_2 are independent. Therefore, we can write that

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\left(S-\mu_S\right)\right)\right] &= \mathbb{E}\left[\exp\left(\lambda\|u\|\|v\|\left(\frac{S}{\|u\|\|v\|} - \frac{\mu_S}{\|u\|\|v\|}\right)\right)\right]\\ &= \mathbb{E}\left[\exp\left(\frac{\lambda}{4}\|u\|\|v\|\left(S_1^2 - \mu_{S_1^2}\right)\right)\right]\\ &\quad\cdot\mathbb{E}\left[\exp\left(-\frac{\lambda}{4}\|u\|\|v\|\left(S_2^2 - \mu_{S_2^2}\right)\right)\right],\end{aligned} \tag{19}$$

where we used the independence of S_1 and S_2 in above. Next, we employ sub-exponential property of S_1^2 and S_2^2. More precisely, $S_1^2 \sim \mathsf{SubE}(4\sigma_1^4, 4\sigma_1^2)$ and $S_1^2 \sim \mathsf{SubE}(4\sigma_2^4, 4\sigma_2^2)$ where

$$\sigma_1^2 := \mathbb{E}[S_1^2] = 2\left(1+\langle\overline{u},\overline{v}\rangle\right), \quad \text{and} \quad \sigma_2^2 := \mathbb{E}[S_2^2] = 2\left(1-\langle\overline{u},\overline{v}\rangle\right).$$

In other words,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda'\left(S_1^2 - \mu_{S_1^2}\right)\right)\right] &\leq \exp\left(2\lambda'^2\sigma_1^4\right), \quad \forall|\lambda'|\leq \frac{1}{4\sigma_1^2},\\ \mathbb{E}\left[\exp\left(\lambda'\left(S_2^2 - \mu_{S_2^2}\right)\right)\right] &\leq \exp\left(2\lambda'^2\sigma_2^4\right), \quad \forall|\lambda'|\leq \frac{1}{4\sigma_2^2}.\end{aligned}$$

Putting $\lambda' = \pm\frac{\lambda}{4}\|u\|\|v\|$ and using (19), we have that

$$\mathbb{E}\left[\exp\left(\lambda\left(S-\mu_S\right)\right)\right] \leq \exp\left(\frac{\lambda^2}{8}\|u\|^2\|v\|^2\left(\sigma_1^4+\sigma_2^4\right)\right), \quad \forall|\lambda|\leq \frac{1}{\|u\|\|v\|\max\{\sigma_1^2,\sigma_2^2\}}.$$

Finally,

$$\sigma_1^4+\sigma_2^4 = 4\left(1+\langle\overline{u},\overline{v}\rangle\right)^2 + 4\left(1-\langle\overline{u},\overline{v}\rangle\right)^2 = 8\left(1+\langle\overline{u},\overline{v}\rangle^2\right) \leq 16,$$

yielding that

$$\|u\|^2\|v\|^2\left(\sigma_1^4+\sigma_2^4\right) \leq 16\|u\|^2\|v\|^2,$$

and

$$\|u\|\|v\|\max\{\sigma_1^2,\sigma_2^2\} = 2\|u\|\|v\|\max\{1+\langle\overline{u},\overline{v}\rangle, 1-\langle\overline{u},\overline{v}\rangle\} = 2\|u\|\|v\|(1+|\langle\overline{u},\overline{v}\rangle|)\leq 4\|u\|\|v\|.$$

Next, consider $N = \langle X, u\rangle\epsilon$. We know that N is sub-exponential with parameters $\mathsf{SubE}(\tau_N^2, b_N)$. Denote $Z_x \coloneqq \frac{\langle X,u\rangle}{\|u\|}$ and $Z_\epsilon \coloneqq \frac{\epsilon}{\sigma}$. Clearly, Z_x and Z_ϵ are independent standard Gaussian. Moreover,

$$N = \|u\|\sigma Z_x Z_\epsilon = \frac{\|u\|\sigma}{4}\left(Z_x+Z_\epsilon\right)^2 - \frac{\|u\|\sigma}{4}\left(Z_x-Z_\epsilon\right)^2.$$

Furthermore, $Z_x + Z_\epsilon$ and $Z_x - Z_\epsilon$ are zero-mean Gaussian RVs and $\mathbb{E}[(Z_x+Z_\epsilon)(Z_x-Z_\epsilon)] = 0$, implying that $Z_x + Z_\epsilon$ and $Z_x - Z_\epsilon$ are independent. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda N\right)\right] &= \mathbb{E}\left[\exp\left(\lambda\|u\|\sigma Z_x Z_\epsilon\right)\right]\\
&= \mathbb{E}\left[\exp\left(\frac{\lambda\|u\|\sigma}{4}\left(Z_x+Z_\epsilon\right)^2 - \frac{\lambda\|u\|\sigma}{4}\left(Z_x-Z_\epsilon\right)^2\right)\right]\\
&= \mathbb{E}\left[\exp\left(\frac{\lambda\|u\|\sigma}{4}\left(Z_x+Z_\epsilon\right)^2\right)\right]\cdot\mathbb{E}\left[\exp\left(-\frac{\lambda\|u\|\sigma}{4}\left(Z_x-Z_\epsilon\right)^2\right)\right]\\
&\leq \exp\left(\frac{\lambda^2}{2}\frac{\|u\|^2\sigma^2}{2}\right),
\end{aligned}$$

for any $|\lambda|\leq\frac{1}{\|u\|\sigma}$. This concludes that $N\sim\mathsf{SubE}(\|u\|^2\sigma^2/2, \|u\|\sigma)$. In above, we use the fact that if $Z\sim\mathcal{N}(0,1)$, then $Z^2\sim\mathsf{SubE}(4,4)$.

B Proof of Theorem 3.3

We begin by setting up a few shorthand notations as follows

$$\widehat{\Sigma} = \frac{1}{mn}\sum_{j=1}^m\sum_{i=1}^n x_i^j x_i^{j\top}, \qquad Z^j = \sum_{i=1}^n x_i^j y_i^j,$$

$$\widehat{v} = \frac{1}{mn}\sum_{j=1}^m\sum_{i=1}^n x_i^j y_i^j\tanh\left(\frac{1}{\sigma^2}\sum_{i=1}^n\langle x_i^j,\theta\rangle y_i^j\right) = \frac{1}{mn}\sum_{j=1}^m Z^j\tanh\left(\frac{1}{\sigma^2}\langle Z^j,\theta\rangle\right),$$

$$v = \mathbb{E}\left[X_1Y_1\tanh\left(\frac{1}{\sigma^2}\sum_{i=1}^n\langle X_i,\theta\rangle Y_i\right)\right]. \tag{20}$$

Therefore, we can write that

$$\|M_m(\theta)-M(\theta)\| = \|\widehat{\Sigma}^{-1}\widehat{v}-v\|\leq\underbrace{\|\widehat{\Sigma}^{-1}\|_{\mathrm{op}}\|v-\widehat{v}\|}_{T_1}+\underbrace{\|\widehat{\Sigma}^{-1}-I\|_{\mathrm{op}}\|v\|}_{T_2}. \tag{21}$$

In the following, we bound each of the two terms above.

B.0.1 Bounding T_1:

We use the following concentration bounds in bounding both T_1 and T_2.

Lemma B.1. *Let standard Gaussian random variables $X_i\sim\mathcal{N}(0,I_d)$ be independent for $i=1,\cdots,N$. For any $\delta\in(0,1)$, if $N\geq 192^2(d+\log(2/\delta))$, then we have that*

$$\|\widehat{\Sigma}_N - I_d\|_{\mathrm{op}}\leq 96\sqrt{\frac{d+\log(2/\delta)}{N}},\quad \|\widehat{\Sigma}_N^{-1}-I_d\|_{\mathrm{op}}\leq 192\sqrt{\frac{d+\log(2/\delta)}{N}},\quad \|\widehat{\Sigma}_N^{-1}\|_{\mathrm{op}}\leq 2,$$

each with probability at least $1-\delta$. Here, we denote the sample covariance matrix of the N samples by

$$\widehat{\Sigma}_N \coloneqq \frac{1}{N}\sum_{i=1}^N x_i x_i^\top.$$

As a result of Lemma B.1, if $mn \geq 192^2(d+\log(2/\delta))$, then $\|\widehat{\Sigma}^{-1}\|_{\mathrm{op}} \leq 2$ and therefore $T_1 \leq 2\|v-\widehat{v}\|$ with probability $1-\delta$. Next, we upper bound $\|v-\widehat{v}\|$ by first decomposing it to the following three terms

$$\|\widehat{v}-v\| \leq \|\widehat{v}-\widehat{v}_0\| + \|v-v_0\| + \|\widehat{v}_0-v_0\|, \tag{22}$$

where we use the following notations

$$\widehat{v}_0 = \frac{1}{mn}\sum_{j=1}^{m} Z^j, \quad v_0 = \frac{1}{n}\mathbb{E}\left[Z\right], \quad Z = \sum_{i=1}^{n} X_i Y_i. \tag{23}$$

Here, since Z^js are i.i.d. across different nodes (i.e. different js), we denote by Z the generic random variable with the same distribution as Z^j for any $j = 1, \cdots, m$. It is also worth noting that v_0 and $\widehat{v}_0$ defined above approach v and $\widehat{v}$ defined in (20) respectively, as the $\tanh(\cdot)$ terms therein approach 1. In the following three lemmas, we upper bound each of the terms in (22).

Lemma B.2. *Assuming that $n \geq d$, there exist $C_4(\alpha, \mathsf{snr})$ and $N_1(\alpha, \mathsf{snr})$, constants depending on $0 \leq \alpha < 1$ and snr, such that for any $\theta \in \mathbb{B}(\alpha\|\theta^*\|; \theta^*)$ and $n \geq N_1(\alpha, \mathsf{snr})$, we have that*

$$\|v-v_0\| \leq (1+\|\theta^*\|+\sigma)\exp\left(-n\cdot C_4(\alpha,\mathsf{snr})\right).$$

Proof. We defer the proof to Section B.1. □

Lemma B.3. *Fix $\delta \in (0,1)$ and assume that $mn \geq 32^2(2d+\log(1/\delta))$. Then, with probability at least $1-\delta$, we have*

$$\|\widehat{v}_0 - v_0\| \leq 8\sqrt{\|\theta^*\|^2+\sigma^2}\sqrt{\frac{2d+\log(1/\delta)}{mn}}.$$

Proof. We defer the proof to Section B.2. □

Lemma B.4. *For $r = \frac{1}{14}\|\theta^*\|$ and $\mathsf{snr} \geq 4$, with probability at least $1-(m+2)\cdot 5^d \cdot \exp(-n/64)$,*

$$\sup_{\theta\in\mathbb{B}(r;\theta^*)} \|\widehat{v}-\widehat{v}_0\| \leq 2(3\|\theta^*\|+\sigma)\exp(-4n).$$

Proof. We defer the proof to Section B.3. □

Putting the results of the above three lemmas back in the decomposition of $\|v-\widehat{v}\|$ in (22) yields that

$$\begin{aligned}
\sup_{\theta\in\mathbb{B}(r;\theta^*)} \|\widehat{v}-v\| &\leq \sup_{\theta\in\mathbb{B}(r;\theta^*)} \|\widehat{v}-\widehat{v}_0\| + \sup_{\theta\in\mathbb{B}(r;\theta^*)} \|v-v_0\| + \|\widehat{v}_0-v_0\| \\
&\leq 8\sqrt{\|\theta^*\|^2+\sigma^2}\sqrt{\frac{2d+\log(1/\delta)}{mn}} + 2(3\|\theta^*\|+\sigma)\exp(-4n) \\
&\quad + (1+\|\theta^*\|+\sigma)\exp\left(-n\cdot C_4(\alpha,\mathsf{snr})\right) \\
&= \sqrt{\|\theta^*\|^2+\sigma^2}\sqrt{\frac{d+\log(1/\delta)}{mn}}\,\mathcal{O}(1)
\end{aligned} \tag{24}$$

with probability at least $1-\delta-(m+2)\cdot 5^d\cdot\exp(-n/64) \geq 1-2\delta$. Here, $\mathcal{O}(1)$ hides constants depending on $\|\theta^*\|$, σ and α. Also, we used the assumption that $n - 64\log m \geq 104(2d+\log(1/\delta))$.

Moreover, we showed in Lemma B.1 that if the total number of samples in at least $mn \geq 192^2(d+\log(2/\delta))$, then with probability $1-\delta$, we have $\|\widehat{\Sigma}^{-1}\|_{\mathrm{op}} \leq 2$. This together with (24) yields that with probability $1-3\delta$, it holds that

$$\sup_{\theta\in\mathbb{B}(r;\theta^*)} T_1 = \sup_{\theta\in\mathbb{B}(r;\theta^*)} \|\widehat{\Sigma}^{-1}\|_{\mathrm{op}}\|v-\widehat{v}\| \leq \sqrt{\|\theta^*\|^2+\sigma^2}\sqrt{\frac{d+\log(1/\delta)}{mn}}\,\mathcal{O}(1).$$

B.0.2 Bounding T_2:

As we showed in Lemma B.1, for a fixed $\delta \in (0,1)$ and $mn \geq 192^2(d+\log(2/\delta))$, we have

$$\|\widehat{\Sigma}^{-1} - I\|_{\text{op}} \leq 192\sqrt{\frac{d+\log(2/\delta)}{mn}}$$

with probability $1-\delta$. To bound $\|v\|$, we can write that $\|v\| = \|M(\theta)\| \leq \|M(\theta)-\theta^*\| + \|\theta^*\|$. The term $\|M(\theta)-\theta^*\|$ denotes the distance of the regression parameter θ to the optimal one θ^* after an iteration of updates by the population operator $M(\cdot)$. In Theorem 3.1, we proved that this operator is contractive. More precisely, for any θ such that $\varepsilon \leq \|\theta-\theta^*\| \leq \alpha\|\theta^*\|$, we have

$$\|M(\theta)-\theta^*\| \leq \kappa(\varepsilon)\cdot\|\theta-\theta^*\|, \text{ where } \kappa(\varepsilon) = (\|\theta^*\|+\sigma)\left(\mathsf{snr} + \frac{1}{n\varepsilon}\right)\exp\left(-n\cdot C(\alpha,\mathsf{snr})\right).$$

From the assumption of the theorem, we know that $\kappa(\varepsilon) \leq 1$ for large enough n. Therefore

$$\|v\| = \|M(\theta)\| \leq \|M(\theta)-\theta^*\| + \|\theta^*\| \leq \|\theta-\theta^*\| + \|\theta^*\| \leq 2\|\theta^*\|.$$

All in all, we have with probability at least $1-\delta$ that

$$T_2 \leq \|\theta^*\|\sqrt{\frac{d+\log(1/\delta)}{mn}}\,\mathcal{O}(1).$$

Now having bounded both terms T_1 and T_2, we can write from (21) that with probability at least $1-4\delta$,

$$\sup_{\theta\in\mathbb{B}(r;\theta^*)} \|M_m(\theta)-M(\theta)\| \leq \sqrt{\|\theta^*\|^2+\sigma^2}\sqrt{\frac{d+\log(1/\delta)}{mn}}\cdot\mathcal{O}(1).$$

We can further change the probability $1-4\delta$ to $1-\delta$ by replacing the $\log(1/\delta)$ to $\log(4/\delta)$ which implies slightly tighter bounds on the sample sizes m and n. It is also worth noting that all the assumptions made in the auxiliary lemmas above can be implied by the ones made in Theorem 3.3. Here, we conclude the proof of Theorem 3.3 and move to prove the auxiliary lemmas used above.

B.1 Proof of Lemma B.2

Using the definition of v and v_0 in (20) and (23), we have that

$$\|v-v_0\| = \left\|\frac{1}{n}\mathbb{E}\left[Z\right] - \frac{1}{n}\mathbb{E}\left[Z\tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right)\right]\right\| \leq \frac{1}{n}\mathbb{E}\left[\|Z\|\left(1-\tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right)\right)\right].$$

Next, consider a fixed θ with $\|\theta-\theta^*\| \leq \alpha\|\theta^*\|$ and define a good event $\mathcal{E}_1$ as follows

$$\mathcal{E}_1 = \left\{\langle Z,\theta\rangle \geq \frac{n}{4}(1-\alpha)\|\theta^*\|^2\right\}.$$

Therefore, we can write that

$$\begin{aligned}\|v-v_0\| \leq \frac{1}{n}&\mathbb{E}\left[\|Z\|\left(1-\tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right)\right)\cdot\mathbb{1}\{\mathcal{E}_1\}\right]\\ &+\frac{1}{n}\mathbb{E}\left[\|Z\|\left(1-\tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right)\right)\cdot\mathbb{1}\{\mathcal{E}_1^c\}\right]\end{aligned}$$

Let us denote each of the two terms above as T_3 and T_4, that is,

$$\begin{aligned}T_3 &= \mathbb{E}\left[\|Z\|\left(1-\tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right)\right)\cdot\mathbb{1}\{\mathcal{E}_1\}\right],\\ T_4 &= \mathbb{E}\left[\|Z\|\left(1-\tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right)\right)\cdot\mathbb{1}\{\mathcal{E}_1^c\}\right].\end{aligned} \tag{25}$$

We bound T_3 by first noting that

$$\|Z\|\left(1-\tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right)\right)\cdot\mathbb{1}\{\mathcal{E}_1\}\leq 2\|Z\|\cdot\exp\left(-\frac{n}{2}(1-\alpha)\mathsf{snr}^2\right),$$

where we used the fact that under $\mathcal{E}_g$, we have $\langle Z,\theta\rangle\geq\frac{n}{4}(1-\alpha)\|\theta^*\|^2$ which from the monotonicity of $\tanh(\cdot)$ implies that

$$\tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right)\geq\tanh\left(\frac{n}{4}(1-\alpha)\frac{\|\theta^*\|^2}{\sigma^2}\right)\geq 1-2\exp\left(-\frac{n}{2}(1-\alpha)\mathsf{snr}^2\right).$$

In the last inequality above, we used the fact that $\tanh(x)\geq 1-2\exp(-2x)$ for all x. Consequently, we have that

$$T_3\leq 2\mathbb{E}[\|Z\|]\cdot\exp\left(-\frac{n}{2}(1-\alpha)\mathsf{snr}^2\right).$$

In the following, we upper bound $\mathbb{E}[\|Z\|]$. We can write that

$$\begin{aligned}\mathbb{E}[\|Z\|^2]&=\mathbb{E}\left[\left(\sum_{i=1}^{n}\langle\theta^*,X_i\rangle X_i+\epsilon_i X_i\right)^\top\left(\sum_{j=1}^{n}\langle\theta^*,X_j\rangle X_j+\epsilon_j X_j\right)\right]\\&=\sum_{i=1}^{n}\mathbb{E}[Y_i^2X_i^\top X_i]+\sum_{1\leq i\neq j\leq n}\mathbb{E}[Y_iX_i^\top X_jY_j]\\&=(n^2-n+d+2)\|\theta^*\|^2+d\sigma^2,\end{aligned}$$

which implies that $\mathbb{E}[\|Z\|]\leq\sqrt{n^2-n+d+2}\|\theta^*\|+\sqrt{d}\sigma$ and consequently,

$$T_3\leq 2\left(n\|\theta^*\|+\sqrt{d}\sigma\right)\cdot\exp\left(-\frac{n}{2}(1-\alpha)\mathsf{snr}^2\right).$$

Next, we upper bound the term T_4 in (25) as follows,

$$\begin{aligned}T_4&=\mathbb{E}\left[\|Z\|\left(1-\tanh\left(\frac{1}{\sigma^2}\langle Z,\theta\rangle\right)\right)\cdot\mathbb{1}\{\mathcal{E}_1^c\}\right]\\&\leq 2\mathbb{E}\left[\|Z\|\cdot\mathbb{1}\{\mathcal{E}_1^c\}\right]\\&=2\int_0^\infty\mathbb{P}\left(\|Z\|\cdot\mathbb{1}\{\mathcal{E}_1^c\}\geq\gamma\right)\mathrm{d}\gamma.\end{aligned}\tag{26}$$

For any $\gamma\geq 0$, we use ϵ-net argument and write that

$$\begin{aligned}\mathbb{P}\left(\|Z\|\cdot\mathbb{1}\{\mathcal{E}_1^c\}\geq\gamma\right)&\leq\mathbb{P}\left(\|Z\|\geq\gamma\right)\\&=\mathbb{P}\left(\max_{\|u\|=1}\langle Z,u\rangle\geq\gamma\right)\\&\leq\mathbb{P}\left(\max_{u'\in\mathcal{N}_{1/2}}\langle Z,u'\rangle\geq\gamma/2\right)\\&\leq 5^d\cdot\mathbb{P}\left(\langle Z,u'\rangle\geq\gamma/2\right),\end{aligned}\tag{27}$$

where $\mathcal{N}_{1/2}$ denotes a 1/2-covering of the unit sphere $\mathbb{S}^d=\{u\in\mathbb{R}^d\,|\,\|u\|=1\}$, known to have cardinality of at most $|\mathcal{N}_{1/2}|\leq 5^d$. Moreover, in deriving the above inequalities, we used the fact that for any unit vector u, there exists $u'\in\mathcal{N}_{1/2}$ such that $\|u-u'\|\leq 1/2$ and therefore,

$$\langle Z,u\rangle=\langle Z,u-u'\rangle+\langle Z,u'\rangle\leq\max_{\|w\|=\frac{1}{2}}\langle Z,w\rangle+\max_{u'\in\mathcal{N}_{1/2}}\langle Z,u'\rangle,$$

which yields that $\max_{\|u\|=1}\langle Z,u\rangle \le 2\max_{u'\in\mathcal{N}_{1/2}}\langle Z,u'\rangle$. Now, consider a fixed unit vector $u'\in\mathbb{S}^d$. We know that $\langle Z,u'\rangle$ is $\mathsf{SubE}(8n\|\theta^*\|^2{+}n\sigma^2, 4\|\theta^*\|{+}\sigma)$ and $\mathbb{E}[\langle Z,u'\rangle]=n\langle\theta^*,u'\rangle$. Therefore,

$$\begin{aligned}\mathbb{P}\left(\langle Z,u'\rangle\ge\gamma/2\right)&=\mathbb{P}\left(\langle Z,u'\rangle-n\langle\theta^*,u'\rangle\ge\gamma/2-n\langle\theta^*,u'\rangle\right)\\&\le\mathbb{P}\left(\langle Z,u'\rangle-n\langle\theta^*,u'\rangle\ge\gamma/2-n\|\theta^*\|\right)\\&\le\mathbb{P}\left(|\langle Z,u'\rangle-n\langle\theta^*,u'\rangle|\ge\gamma/2-n\|\theta^*\|\right)\\&\le\exp\left(-\frac{1}{2}\min\left\{\frac{(\gamma/2-n\|\theta^*\|)^2}{8n\|\theta^*\|^2{+}n\sigma^2},\frac{\gamma/2-n\|\theta^*\|}{4\|\theta^*\|{+}\sigma}\right\}\right).\end{aligned}$$

for any $\gamma\ge 2n\|\theta^*\|$. Now pick

$$\gamma_0=2n\|\theta^*\|+2\frac{8n\|\theta^*\|^2{+}n\sigma^2}{4\|\theta^*\|{+}\sigma}+2n^2(\|\theta^*\|{+}\sigma).$$

It yields that

$$\begin{aligned}\int_{\gamma_0}^{\infty}\mathbb{P}\left(\|Z\|\cdot\mathbb{1}\{\mathcal{E}_1^c\}\ge\gamma\right)\mathrm{d}\gamma&\le 5^d\int_{\gamma_0}^{\infty}\exp\left(-\frac{\gamma/2-n\|\theta^*\|}{4\|\theta^*\|{+}\sigma}\right)\mathrm{d}\gamma\\&=5^d\cdot 2(4\|\theta^*\|{+}\sigma)\cdot\exp\left(-\frac{\gamma_0/2-n\|\theta^*\|}{4\|\theta^*\|{+}\sigma}\right)\\&\le 8\cdot 5^d(\|\theta^*\|{+}\sigma)\exp(-n^2/4).\end{aligned}\tag{28}$$

Moreover,

$$\begin{aligned}\int_{0}^{\gamma_0}\mathbb{P}\left(\|Z\|\cdot\mathbb{1}\{\mathcal{E}_1^c\}\ge\gamma\right)\mathrm{d}\gamma&\le\int_{e^{-c_3n}}^{\gamma_0}\mathbb{P}\left(\|Z\|\cdot\mathbb{1}\{\mathcal{E}_1^c\}\ge\gamma\right)\mathrm{d}\gamma+e^{-c_3n}\\&\le\gamma_0\mathbb{P}(\mathcal{E}_1^c)+e^{-c_3n}\\&\le(\gamma_0+1)\exp(-c_3n)\\&\le\Big(1+(2n^2+10n)(\|\theta^*\|{+}\sigma)\Big)\exp(-c_3n),\end{aligned}\tag{29}$$

where we used Lemma A.1 to conclude that $\mathbb{P}(\mathcal{E}_1^c)\le\exp(-c_3n)$ for a constant c_3 depending on α and snr. Putting (28) and (29) in (26) yields that

$$T_4\le 2\Big(1+(2n^2+10n)(\|\theta^*\|{+}\sigma)\Big)\exp(-c_3n)+16\cdot 5^d(\|\theta^*\|{+}\sigma)\exp(-n^2/4).$$

Finally, we put everything together and conclude the lemma as follows,

$$\begin{aligned}\|v-v_0\|&\le\frac{1}{n}T_3+\frac{1}{n}T_4\\&\le 2\left(\|\theta^*\|{+}\frac{\sqrt{d}}{n}\sigma\right)\cdot\exp\left(-\frac{n}{2}(1-\alpha)\mathsf{snr}^2\right)\\&\quad+\frac{2}{n}\left(\|\theta^*\|{+}\frac{\sqrt{d}}{n}\sigma\right)\cdot\exp\left(-\frac{n}{2}(1-\alpha)\mathsf{snr}^2\right)+\frac{16}{n}\cdot 5^d(\|\theta^*\|{+}\sigma)\exp(-n^2/4)\\&\le(1+\|\theta^*\|{+}\sigma)\exp\left(-C_4(\alpha,\mathsf{snr})n\right),\end{aligned}$$

for any $n\ge N_1(\alpha,\mathsf{snr})$. Here, we used the assumption that $d\le n$ which is also implied by the assumptions in Theorem 3.3, particularly from $n\ge 64\log m+104(2d+\log(4/\delta))$.

B.2 Proof of Lemma B.3

Recall from the notations that

$$\|\widehat{v}_0-v_0\|=\left\|\frac{1}{mn}\sum_{j=1}^{m}Z^j-\frac{1}{n}\mathbb{E}\left[Z\right]\right\|,$$

For a fixed $j \in \{1, \cdots, m\}$ and unit-norm vector u, the inner product $\langle Z^u, u\rangle$ is sub-exponential. More precisely,

$$\frac{1}{n}\langle Z^j - \mathbb{E}[Z^j], u\rangle \sim \mathsf{SubE}\Big(\frac{1}{n}(8\|\theta^*\|^2 + \sigma^2), \frac{1}{n}(4\|\theta^*\| + \sigma)\Big).$$

Therefore,

$$\langle \widehat{v}_0 - v_0, u\rangle = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\langle Z^j - \mathbb{E}[Z^j], u\rangle \sim \mathsf{SubE}\Big(\frac{1}{mn}(8\|\theta^*\|^2 + \sigma^2), \frac{1}{mn}(4\|\theta^*\| + \sigma)\Big).$$

From concentration of sub-exponential random variables in Theorem D.1, we can write for every $t \geq 0$ that

$$\begin{aligned}
\mathbb{P}\left(\|\widehat{v}_0 - v_0\| \geq t\right) &= \mathbb{P}\bigg(\max_{\|u\|=1}\langle \widehat{v}_0 - v_0, u\rangle \geq t\bigg)\\
&\leq \mathbb{P}\bigg(\max_{u'\in\mathcal{N}_{1/2}}\langle \widehat{v}_0 - v_0, u'\rangle \geq t/2\bigg)\\
&\leq 5^d \cdot \exp\left(-\frac{1}{2}\min\left\{\frac{mn(t/2)^2}{8\|\theta^*\|^2 + \sigma^2}, \frac{mnt/2}{4\|\theta^*\| + \sigma}\right\}\right)
\end{aligned}$$

where we used a 1/2-covering argument similar to (27). Now, assuming $mn \geq 32^2(2d + \log(1/\delta))$, we pick

$$t = 8\sqrt{\frac{2d + \log(1/\delta)}{mn}}\sqrt{\|\theta^*\|^2 + \sigma^2},$$

which implies that $\mathbb{P}(\|\widehat{v}_0 - v_0\| \geq t) \leq \delta$ as desired.

B.3 Proof of Lemma B.4

We begin the proof by using the definitions of $\widehat{v}$ and $\widehat{v}_0$ and write

$$\begin{aligned}
\sup_{\theta\in\mathbb{B}(r;\theta^*)}\|\widehat{v} - \widehat{v}_0\| &= \sup_{\theta\in\mathbb{B}(r;\theta^*)}\left\|\frac{1}{mn}\sum_{j=1}^{m}Z^j - \frac{1}{mn}\sum_{j=1}^{m}Z^j\tanh\left(\frac{1}{\sigma^2}\langle Z^j, \theta\rangle\right)\right\|\\
&\leq \sup_{\theta\in\mathbb{B}(r;\theta^*)}\frac{1}{mn}\sum_{j=1}^{m}\|Z^j\|\left(1 - \tanh\left(\frac{1}{\sigma^2}\langle Z^j, \theta\rangle\right)\right)\\
&\leq \frac{1}{mn}\sum_{j=1}^{m}\|Z^j\| \cdot \sup_{\theta\in\mathbb{B}(r;\theta^*)}\left(1 - \tanh\left(\frac{1}{\sigma^2}\langle Z^j, \theta\rangle\right)\right).
\end{aligned}$$

Next, for each $j = 1, \cdots, m$ we have that

$$\inf_{\theta\in\mathbb{B}(r;\theta^*)}\langle Z^j, \theta\rangle = \langle Z^j, \theta^*\rangle + \inf_{\theta\in\mathbb{B}(r;\theta^*)}\langle Z^j, \theta - \theta^*\rangle = \langle Z^j, \theta^*\rangle + r \cdot \inf_{\|u\|=1}\langle Z^j, u\rangle. \tag{30}$$

Moreover, from a 1/2-covering argument similar to (27) we know that

$$\sup_{\|u\|=1}\langle Z^j, u\rangle \leq 2\sup_{u'\in\mathcal{N}_{1/2}}\langle Z^j, u'\rangle,$$

which yields that

$$\mathbb{P}\bigg(\sup_{\|u\|=1}\langle Z^j, u\rangle \geq t\bigg) \leq \mathbb{P}\bigg(\sup_{u'\in\mathcal{N}_{1/2}}\langle Z^j, u'\rangle \geq t/2\bigg) \leq 5^d \cdot \mathbb{P}\left(\langle Z^j, u'\rangle \geq t/2\right).$$

For a fixed unit-norm u', $\langle Z^j, u'\rangle$ is $\mathsf{SubE}(8n\|\theta^*\|^2 + n\sigma^2, 4\|\theta^*\| + \sigma)$ with $\mathbb{E}[\langle Z^j, u'\rangle] = n\langle\theta^*, u'\rangle$. Therefore, for any $t \geq 0$, we have that

$$|\langle Z^j, u'\rangle - n\langle\theta^*, u'\rangle| \leq t/2,$$

with probability at least

$$1-2\exp\left(-\min\left\{\frac{t^2/8}{8n\|\theta^*\|^2+n\sigma^2},\frac{t/4}{4\|\theta^*\|+\sigma}\right\}\right).$$

Taking $t=n(\|\theta^*\|+\sigma)$ in above yields that with probability at least $1-\exp(-n/64)$ we have

$$\langle Z^j,u'\rangle\leq n\langle\theta^*,u'\rangle+\frac{1}{2}n(\|\theta^*\|+\sigma)\leq\frac{3}{2}n\|\theta^*\|+\frac{1}{2}n\sigma.$$

Together with the 1/2-covering argument, it holds that

$$\sup_{\|u\|=1}\langle Z^j,u\rangle\leq 3n\|\theta^*\|+n\sigma, \tag{31}$$

with probability at least $1-5^d\cdot\exp(-n/64)$. Moreover, following the above logic for a specific case of $u'=\theta^*/\|\theta^*\|$, we have with probability $1-\exp(-n/64)$ that

$$\Big|\langle Z^j,\theta^*/\|\theta^*\|\rangle-n\|\theta^*\|\Big|\leq\frac{1}{2}n(\|\theta^*\|+\sigma),$$

which implies that

$$\langle Z^j,\theta^*\rangle\geq\frac{1}{2}n\|\theta^*\|^2-\frac{1}{2}n\|\theta^*\|\sigma, \tag{32}$$

with the same probability. Putting (31) and (32) back in (30) yields that

$$\inf_{\theta\in\mathbb{B}(r;\theta^*)}\langle Z^j,\theta\rangle=\langle Z^j,\theta^*\rangle+r\cdot\inf_{\|u\|=1}\langle Z^j,u\rangle\geq\frac{1}{2}n\|\theta^*\|^2-\frac{1}{2}n\|\theta^*\|\sigma-3rn\|\theta^*\|-rn\sigma\geq\frac{1}{7}n\|\theta^*\|^2,$$

for $r=\frac{1}{14}\|\theta^*\|$ and $\mathsf{snr}\geq 4$. Therefore, with probability at least $1-(5^d+1)\exp(-n/64)$, we have

$$\sup_{\theta\in\mathbb{B}(r;\theta^*)}\left(1-\tanh\left(\frac{1}{\sigma^2}\langle Z^j,\theta\rangle\right)\right)\leq 2\exp\left(-\frac{2}{7}n\cdot\mathsf{snr}^2\right)\leq 2\exp(-4n), \tag{33}$$

where we used $\tanh(x)\geq 1-2\exp(-2x)$ for all x. Furthermore, we showed above that with probability at least $1-5^d\cdot\exp(-n/64)$ and for each $1\leq j\leq m$, we have

$$\|Z^j\|=\sup_{\|u\|=1}\langle Z^j,u\rangle\leq 3n\|\theta^*\|+n\sigma. \tag{34}$$

Putting (33) and (34) together, we have

$$\sup_{\theta\in\mathbb{B}(r;\theta^*)}\|\widehat{v}-\widehat{v}_0\|\leq\frac{1}{mn}\sum_{j=1}^m\|Z^j\|\cdot\sup_{\theta\in\mathbb{B}(r;\theta^*)}\left(1-\tanh\left(\frac{1}{\sigma^2}\langle Z^j,\theta\rangle\right)\right)\leq 2(3\|\theta^*\|+\sigma)\exp(-4n),$$

with probability at least $1-(m+2)\cdot 5^d\cdot\exp(-n/64)$.

B.4 Proof of Lemma B.1

The proof follows from basic standard Gaussian concentration. We provide the proof here for completeness. Using concentration of sub-exponential RVs and ϵ-net arguments we have that

$$\mathbb{P}\left(\|\widehat{\Sigma}-I_d\|_{\mathrm{op}}\geq t\right)\leq 2\cdot 9^d\exp\left(-\frac{n}{2}\min\left\{\left(\frac{t}{32}\right)^2,\frac{t}{32}\right\}\right).$$

Picking $t=96\sqrt{(d+\log(2/\delta))/N}$ yields the desired concentration bound. For the second inequality, we note that

$$\begin{aligned}\|\widehat{\Sigma}^{-1}-I_d\|_{\mathrm{op}}&\leq\|\widehat{\Sigma}^{-1}\|_{\mathrm{op}}\|\widehat{\Sigma}-I_d\|_{\mathrm{op}}\\&\leq\left(1+\|\widehat{\Sigma}^{-1}-I_d\|_{\mathrm{op}}\right)\|\widehat{\Sigma}-I_d\|_{\mathrm{op}}\\&\leq 96\sqrt{\frac{d+\log(2/\delta)}{N}}+\frac{1}{2}\|\widehat{\Sigma}^{-1}-I_d\|_{\mathrm{op}},\end{aligned}$$

with probability at least $1-\delta$. In above, we used the concentration proved in the first part, as well as the assumption $N\geq 192^2(d+\log(2/\delta))$ to conclude that $\|\widehat{\Sigma}-I_d\|_{\mathrm{op}}\leq 1/2$.

C Useful Lemmas

C.1 Proof of Proposition 2.1

Let us denote $x_{[n]} \coloneqq (x_1 \cdots, x_n)$ and $y_{[n]} \coloneqq (y_1 \cdots, y_n)$. Then, according to the C-MLR model in (2) with true regression parameters θ, we have $y_i|\xi, x_i \sim \mathcal{N}(\xi\langle x_i, \theta\rangle, \sigma^2)$. Therefore,

$$\begin{aligned} f_\theta(x_{[n]}, y_{[n]}) &= \int f_\theta(x_{[n]}, y_{[n]}, \xi)\mathrm{d}\xi \\ &= \mathbb{P}(\xi = -1) f_\theta(x_{[n]}, y_{[n]}|\xi = -1) + \mathbb{P}(\xi = +1) f_\theta(x_{[n]}, y_{[n]}|\xi = +1) \\ &= \frac{1}{2}\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} f(x_{[n]})\left\{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \langle x_i, \theta\rangle)^2\right) + \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i + \langle x_i, \theta\rangle)^2\right)\right\} \\ &= \frac{1}{2}\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} f(x_{[n]})\left(\exp(s_{-1}(\theta)) + \exp(s_{+1}(\theta))\right), \end{aligned}$$

where we denote for any θ and $\xi \in \{-1, +1\}$

$$s_\xi(\theta) = -\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \xi\langle x_i, \theta\rangle)^2.$$

Moreover,

$$f_\theta(\xi|x_{[n]}, y_{[n]}) = \frac{f_\theta(x_{[n]}, y_{[n]}, \xi)}{f_\theta(x_{[n]}, y_{[n]})} = 2p(\xi)\frac{\exp(s_\xi(\theta)))}{\exp(s_{-1}(\theta)) + \exp(s_{+1}(\theta))},$$

and

$$\log f_{\theta'}(x_{[n]}, y_{[n]}, \xi) = s_\xi(\theta') + \log p(\xi) + \log f(x_{[n]}) - \frac{n}{2}\log(2\pi). \tag{35}$$

Let us define

$$\hat{Q}(\theta'|\theta) = \int_\xi f_\theta(\xi|x_{[n]}, y_{[n]}) \log f_{\theta'}(x_{[n]}, y_{[n]}, \xi)\mathrm{d}\xi.$$

According to (38), only the first term in the RHS of (38) depends on θ' and therefore, we only keep the term $s_\xi(\theta')$ in computing $\hat{Q}(\theta'|\theta)$ as follows

$$\begin{aligned} \int_\xi f_\theta(\xi|x_{[n]}, y_{[n]}) s_\xi(\theta')\mathrm{d}\xi &= \int_\xi 2p(\xi)\frac{\exp(s_\xi(\theta)))}{\exp(s_{-1}(\theta)) + \exp(s_{+1}(\theta))} s_\xi(\theta')\mathrm{d}\xi \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^n y_i^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n \langle x_i, \theta'\rangle^2 \\ &\quad + \frac{1}{\sigma^2}\frac{\exp(s_{+1}(\theta)) - \exp(s_{-1}(\theta))}{\exp(s_{+1}(\theta)) + \exp(s_{-1}(\theta))}\sum_{i=1}^n y_i\langle x_i, \theta'\rangle. \end{aligned} \tag{36}$$

Now, the Q-function defined in (4) can be written as $Q(\theta'|\theta) = \mathbb{E}[\hat{Q}(\theta'|\theta)]$ where the expectation is w.r.t. randomness in $(X_{[n]}, Y_{[n]})$ generated by the ground truth distribution governed by θ^*. From (39) we have that

$$\begin{aligned} \mathbb{E}\left[\int_\xi f_\theta(\xi|x_{[n]}, y_{[n]}) s_\xi(\theta')\mathrm{d}\xi\right] &= -\frac{1}{2\sigma^2}\sum_{i=1}^n \mathbb{E}[y_i^2] - \frac{n}{2\sigma^2}\|\theta'\|^2 \\ &\quad + \frac{1}{\sigma^2}\mathbb{E}\left[\tanh\left(\frac{1}{\sigma^2}\sum_{i=1}^n y_i\langle x_i, \theta\rangle\right)\sum_{i=1}^n y_i\langle x_i, \theta'\rangle\right], \end{aligned} \tag{37}$$

where we used the fact that

$$\frac{\exp(s_{+1}(\theta)) - \exp(s_{-1}(\theta))}{\exp(s_{+1}(\theta)) + \exp(s_{-1}(\theta))} = \tanh\left(\frac{1}{\sigma^2}\sum_{i=1}^n y_i\langle x_i, \theta\rangle\right).$$

Note that the first term in RHS of (40) dose not depend on θ', therefore,

$$\nabla Q(\theta'|\theta) = -\frac{n}{\sigma^2}\theta' + \frac{1}{\sigma^2}\mathbb{E}\bigg[\sum_{i=1}^n X_i Y_i \tanh\left(\frac{1}{\sigma^2}\sum_{i=1}^n \langle X_i, \theta\rangle Y_i\right)\bigg].$$

Putting $\nabla Q(\theta'|\theta) = 0$ yields that

$$\begin{aligned} M(\theta) &:= \arg\max_{\theta'} Q(\theta'|\theta) \\ &= \frac{1}{n}\mathbb{E}\bigg[\sum_{i=1}^n X_i Y_i \tanh\left(\frac{1}{\sigma^2}\sum_{i=1}^n \langle X_i, \theta\rangle Y_i\right)\bigg] \\ &= \mathbb{E}\bigg[X_1 Y_1 \tanh\left(\frac{1}{\sigma^2}\sum_{i=1}^n \langle X_i, \theta\rangle Y_i\right)\bigg]. \end{aligned}$$

C.2 Proof of Proposition 2.2

Using (39) and definition of Q_m-function in (6) we have that

$$\begin{aligned} \frac{1}{m}\sum_{j=1}^m \int_\xi f_\theta(\xi|x^j_{[n]}, y^j_{[n]}) s_\xi(\theta')\mathrm{d}\xi = &-\frac{1}{2m\sigma^2}\sum_{j=1}^m\sum_{i=1}^n {y_i^j}^2 - \frac{1}{2m\sigma^2}\sum_{j=1}^m\sum_{i=1}^n \langle x_i^j, \theta'\rangle^2 \\ &+ \frac{1}{m\sigma^2}\sum_{j=1}^m \tanh\left(\frac{1}{\sigma^2}\sum_{i=1}^n y_i^j\langle x_i^j, \theta\rangle\right)\sum_{i=1}^n y_i^j\langle x_i^j, \theta'\rangle. \end{aligned}$$

Therefore,

$$\nabla Q_m(\theta'|\theta) = -\frac{1}{m\sigma^2}\sum_{j=1}^m\sum_{i=1}^n x_i^j {x_i^j}^\top\theta' + \frac{1}{m\sigma^2}\sum_{j=1}^m \tanh\left(\frac{1}{\sigma^2}\sum_{i=1}^n y_i^j\langle x_i^j, \theta\rangle\right)\sum_{i=1}^n y_i^j x_i^j,$$

and finally putting $\nabla Q_m(\theta'|\theta) = 0$ yields the desired result.

Definition C.1 (Sub-exponential RV)**.** *A random variable X is said to be sub-exponential with parameter (τ^2, b) if*

$$\mathbb{E}\left[\exp\left(\lambda(X - \mu_X)\right)\right] \le \exp\left(\frac{\lambda^2\tau^2}{2}\right), \quad \forall|\lambda|\le \frac{1}{b}.$$

We denote such RV by $\mathsf{SubE}(\tau^2, b)$.

Theorem C.1. *Let $X_i \sim \mathsf{SubE}(\tau^2, b)$ be iid sub-exponentials. Then,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu_X\right| \ge t\right) \le 2\exp\left(-\min\left\{\frac{nt^2}{2\tau^2}, \frac{nt}{2b}\right\}\right)$$

Lemma C.2. *For any $x_1, x_2 \ge 0$, we have*

$$\frac{\tanh(x_2) - \tanh(x_1)}{x_2 - x_1} \le \max\{1 - \tanh^2(x_1), 1 - \tanh^2(x_2)\}.$$

Proof. Assume that $x_2 \ge x_1 \ge 0$. Since the function $f(x) := \tanh(x)$ is concave in $[0, +\infty)$, we can write

$$f(x_2) \le f(x_1) + (x_2 - x_1)f'(x_1),$$

which yields that

$$\frac{\tanh(x_2) - \tanh(x_1)}{x_2 - x_1} \leq 1 - \tanh^2(x_1).$$

Similar argument holds for $x_1 \geq x_2 \geq 0$, i.e.,

$$\frac{\tanh(x_1) - \tanh(x_2)}{x_1 - x_2} \leq 1 - \tanh^2(x_2).$$

□