Regularized EM Algorithms: A Unified Framework and Statistical Guarantees

Xinyang Yi
Dept. of Electrical and Computer Engineering
The University of Texas at Austin
yixy@utexas.edu

Constantine Caramanis
Dept. of Electrical and Computer Engineering
The University of Texas at Austin
constantine@utexas.edu

Abstract

Latent models are a fundamental modeling tool in machine learning applications, but they present significant computational and analytical challenges. The popular EM algorithm and its variants, is a much used algorithmic tool; yet our rigorous understanding of its performance is highly incomplete. Recently, work in [1] has demonstrated that for an important class of problems, EM exhibits linear local convergence. In the high-dimensional setting, however, the M-step may not be well defined. We address precisely this setting through a unified treatment using regularization. While regularization for high-dimensional problems is by now well understood, the iterative EM algorithm requires a careful balancing of making progress towards the solution while identifying the right structure (e.g., sparsity or low-rank). In particular, regularizing the M-step using the state-of-the-art high-dimensional prescriptions (e.g., à la [19]) is not guaranteed to provide this balance. Our algorithm and analysis are linked in a way that reveals the balance between optimization and statistical errors. We specialize our general framework to sparse gaussian mixture models, high-dimensional mixed regression, and regression with missing variables, obtaining statistical guarantees for each of these examples.

1 Introduction

We give general conditions for the convergence of the EM method for high-dimensional estimation. We specialize these conditions to several problems of interest, including high-dimensional sparse and low-rank mixed regression, sparse gaussian mixture models, and regression with missing covariates. As we explain below, the key problem in the high-dimensional setting is the M-step. A natural idea is to modify this step via appropriate regularization, yet choosing the appropriate sequence of regularizers is a critical problem. As we know from the theory of regularized M-estimators (e.g., [19]) the regularizer should be chosen proportional to the target estimation error. For EM, however, the target estimation error changes at each step.

The main contribution of our work is technical: we show how to perform this iterative regularization. We show that the regularization sequence must be chosen so that it converges to a quantity controlled by the ultimate estimation error. In existing work, the estimation error is given by the relationship between the population and empirical M-step operators, but this too is not well defined in the high-dimensional setting. Thus a key step, related both to our algorithm and its convergence analysis, is obtaining a different characterization of statistical error for the high-dimensional setting.

Background and Related Work

EM (e.g., [8, 12]) is a general algorithmic approach for handling latent variable models (including mixtures), popular largely because it is typically computationally highly scalable, and easy to implement. On the flip side, despite a fairly long history of studying EM in theory (e.g., [12, 17, 21]),

very little has been understood about general statistical guarantees until recently. Very recent work in [1] establishes a general local convergence theorem (i.e., assuming initialization lies in a local region around true parameter) and statistical guarantees for EM, which is then specialized to obtain near-optimal rates for several specific *low-dimensional* problems – low-dimensional in the sense of the classical statistical setting where the samples outnumber the dimension. A central challenge in extending EM (and as a corollary, the analysis in [1]) to the high-dimensional regime is the M-step. On the algorithm side, the M-step will not be stable (or even well-defined in some cases) in the high-dimensional setting. To make matters worse, any analysis that relies on showing that the finite-sample M-step is somehow "close" to the M-step performed with infinite data (the population-level M-step) simply cannot apply in the high-dimensional regime. Recent work in [20] treats high-dimensional EM using a truncated M-step. This works in some settings, but also requires specialized treatment for every different setting, precisely because of the difficulty with the M-step.

In contrast to work in [20], we pursue a high-dimensional extension via regularization. The central challenge, as mentioned above, is in picking the sequence of regularization coefficients, as this must control the optimization error (related to the special structure of $\boldsymbol{\beta}^*$), as well as the statistical error. Finally, we note that for finite mixture regression, Städler et al.[16] consider an ℓ_1 regularized EM algorithm for which they develop some asymptotic analysis and oracle inequality. However, this work doesn't establish the theoretical properties of local optima arising from regularized EM. Our work addresses this issue from a local convergence perspective by using a novel choice of regularization.

2 Classical EM and Challenges in High Dimensions

The EM algorithm is an iterative algorithm designed to combat the non-convexity of max likelihood due to latent variables. For space concerns we omit the standard derivation, and only give the definitions we need in the sequel. Let $\boldsymbol{Y}, \boldsymbol{Z}$ be random variables taking values in $\mathcal{Y}$,$\mathcal{Z}$, with joint distribution $f_{\boldsymbol{\beta}}(\mathbf{y}, \mathbf{z})$ depending on model parameter $\boldsymbol{\beta} \subseteq \Omega \subseteq \mathbb{R}^p$. We observe samples of Y but not of the latent variable Z. EM seeks to maximize a lower bound on the maximum likelihood function for $\boldsymbol{\beta}$. Letting $\kappa_{\boldsymbol{\beta}}(\mathbf{z}|\mathbf{y})$ denote the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{Y} = \mathbf{y}$, and defining the function

$$Q_n(\boldsymbol{\beta}'|\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{Z}} \kappa_{\boldsymbol{\beta}}(\mathbf{z}|\mathbf{y}_i)\log f_{\boldsymbol{\beta}'}(\mathbf{y}_i, \mathbf{z})d\mathbf{z}, \tag{2.1}$$

one iteration of the EM algorithm, mapping $\boldsymbol{\beta}^{(t)}$ to $\boldsymbol{\beta}^{(t+1)}$, consists of the following two steps:

- E-step: Compute function $Q_n(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)})$ given $\boldsymbol{\beta}^{(t)}$.
- M-step: $\boldsymbol{\beta}^{(t+1)} \leftarrow \mathcal{M}_n(\boldsymbol{\beta}) := \arg\max_{\boldsymbol{\beta}' \in \Omega} Q_n(\boldsymbol{\beta}'|\boldsymbol{\beta}^{(t)})$.

We can define the population (infinite sample) versions of Q_n and $\mathcal{M}_n$ in a natural manner:

$$Q(\boldsymbol{\beta}'|\boldsymbol{\beta}) := \int_{\mathcal{Y}} y_{\boldsymbol{\beta}^*}(\mathbf{y})\int_{\mathcal{Z}} \kappa_{\boldsymbol{\beta}}(\mathbf{z}|\mathbf{y})\log_{\boldsymbol{\beta}'}(\mathbf{y}, \mathbf{z})d\mathbf{z}d\mathbf{y} \tag{2.2}$$

$$\mathcal{M}(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}' \in \Omega} Q(\boldsymbol{\beta}'|\boldsymbol{\beta}). \tag{2.3}$$

This paper is about the high-dimensional setting where the number of samples n may be far less than the dimensionality p of the parameter $\boldsymbol{\beta}$, but where $\boldsymbol{\beta}$ exhibits some special structure, e.g., it may be a sparse vector or a low-rank matrix. In such a setting, the M-step of the EM algorithm may be highly problematic. In many settings, for example sparse mixed regression, the M-step may not even be well defined. More generally, when $n \ll p$, $\mathcal{M}_n(\boldsymbol{\beta})$ may be far from the population version, $\mathcal{M}(\boldsymbol{\beta})$, and in particular, the minimum estimation error $\|\mathcal{M}_n(\boldsymbol{\beta}^*) - \mathcal{M}(\boldsymbol{\beta}^*)\|$ can be much larger than the signal strength $\|\boldsymbol{\beta}^*\|$. This quantity is used in [1] as well as in follow-up work in [20], as a measure of statistical error. In the high dimensional setting, something else is needed.

3 Algorithm

The basis of our algorithm is the by-now well understood concept of regularized high dimensional estimators, where the regularization is tuned to the underlying structure of $\boldsymbol{\beta}^*$, thus defining a regu-

larized M-step via

$$\mathcal{M}_n^r(\boldsymbol{\beta}) := \arg\max_{\boldsymbol{\beta}' \in \Omega} Q_n(\boldsymbol{\beta}'|\boldsymbol{\beta}) - \lambda_n \mathcal{R}(\boldsymbol{\beta}'), \tag{3.1}$$

where $\mathcal{R}(\cdot)$ denotes an appropriate regularizer chosen to match the structure of $\boldsymbol{\beta}^*$. The key challenge is how to choose the sequence of regularizers $\{\lambda_n^{(t)}\}$ in the iterative process, so as to control optimization and statistical error. As detailed in Algorithm 1, our sequence of regularizers attempts to match the target estimation error at each step of the EM iteration. For an intuition of what this might look like, consider the estimation error at step t: $\|\mathcal{M}_n^r(\boldsymbol{\beta}^{(t)}) - \boldsymbol{\beta}^*\|_2$. By the triangle inequality, we can bound this by a sum of two terms: the optimization error and the final estimation error:

$$\|\mathcal{M}_n^r(\boldsymbol{\beta}^{(t)}) - \boldsymbol{\beta}^*\|_2 \leq \|\mathcal{M}_n^r(\boldsymbol{\beta}^{(t)}) - \mathcal{M}_n^r(\boldsymbol{\beta}^*)\|_2 + \|\mathcal{M}_n^r(\boldsymbol{\beta}^*) - \boldsymbol{\beta}^*\|_2. \tag{3.2}$$

Since we expect (and show) linear convergence of the optimization, it is natural to update $\lambda_n^{(t)}$ via a recursion of the form $\lambda_n^{(t)} = \kappa \lambda_n^{(t-1)} + \Delta$ as in (3.3), where the first term represents the optimization error, and Δ represents the final statistical error, i.e., the last term above in (3.2). A key part of our analysis shows that this error (and hence Δ) is controlled by $\|\nabla Q_n(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - \nabla Q(\boldsymbol{\beta}^*|\boldsymbol{\beta})\|_{\mathcal{R}^*}$, which in turn can be bounded uniformly for a variety of important applications of EM, including the three discussed in this paper (see Section 5). While a technical point, it is this key insight that enables the right choice of algorithm and its analysis. In the cases we consider, we obtain min-max optimal rates of convergence, demonstrating that no algorithm, let alone another variant of EM, can perform better.

Algorithm 1 Regularized EM Algorithm

Input Samples $\{\mathbf{y}_i\}_{i=1}^n$, regularizer $\mathcal{R}$, number of iterations T, initial parameter $\boldsymbol{\beta}^{(0)}$, initial regularization parameter $\lambda_n^{(0)}$, estimated statistical error Δ, contractive factor $\kappa < 1$.

1: **For** $t = 1, 2, \ldots, T$ **do**

2: **Regularization parameter update:**

$$\lambda_n^{(t)} \leftarrow \kappa \lambda_n^{(t-1)} + \Delta. \tag{3.3}$$

3: **E-step**: Compute function $Q_n(\cdot|\boldsymbol{\beta}^{(t-1)})$ according to (2.1).

4: **Regularized M-step:**

$$\boldsymbol{\beta}^{(t)} \leftarrow \mathcal{M}_n^r(\boldsymbol{\beta}^{(t-1)}) := \arg\max_{\boldsymbol{\beta} \in \Omega} Q_n(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t-1)}) - \lambda_n^{(t)} \cdot \mathcal{R}(\boldsymbol{\beta}).$$

5: **End For**

Output $\boldsymbol{\beta}^{(T)}$.

4 Statistical Guarantees

We now turn to the theoretical analysis of regularized EM algorithm. We first set up a general analytical framework for regularized EM where the key ingredients are decomposable regularizer and several technical conditions on the population based $Q(\cdot|\cdot)$ and the sample based $Q_n(\cdot|\cdot)$. In Section 4.3, we provide our main result (Theorem 1) that characterizes both computational and statistical performance of the proposed variant of regularized EM algorithm.

4.1 Decomposable Regularizers

Decomposable regularizers (e.g., [3, 6, 14, 19]), have been shown to be useful both empirically and theoretically for high dimensional structural estimation, and they also play an important role in our analytical framework. Recall that for $\mathcal{R} : \mathbb{R}^p \to \mathbb{R}^+$ a norm, and a pair of subspaces $(\mathcal{S}, \overline{\mathcal{S}})$ in $\mathbb{R}^p$ such that $\mathcal{S} \subseteq \overline{\mathcal{S}}$, we have the following definition:

Definition 1 (Decomposability)**.** *Regularizer $\mathcal{R} : \mathbb{R}^p \to \mathbb{R}^+$ is decomposable with respect to $(\mathcal{S}, \overline{\mathcal{S}})$ if*

$$\mathcal{R}(\mathbf{u} + \mathbf{v}) = \mathcal{R}(\mathbf{u}) + \mathcal{R}(\mathbf{v}), \text{ for any } \mathbf{u} \in \mathcal{S}, \mathbf{v} \in \overline{\mathcal{S}}^{\perp}.$$

Typically, the structure of model parameter $\boldsymbol{\beta}^*$ can be characterized by specifying a subspace $\mathcal{S}$ such that $\boldsymbol{\beta}^* \in \mathcal{S}$. The common use of a regularizer is thus to penalize the compositions of solution that

live outside $\mathcal{S}$. We are interested in bounding the estimation error in some norm $\|\cdot\|$. The following quantity is critical in connecting $\mathcal{R}$ to $\|\cdot\|$.

Definition 2 (Subspace Compatibility Constant)**.** *For any subspace $\mathcal{S} \subseteq \mathbb{R}^p$, a given regularizer $\mathcal{R}$ and some norm $\|\cdot\|$, the subspace compatibility constant of $\mathcal{S}$ with respect to $\mathcal{R}, \|\cdot\|$ is given by*

$$\Psi(\mathcal{S}) := \sup_{\mathbf{u} \in \mathcal{S} \setminus \{\mathbf{0}\}} \frac{\mathcal{R}(\mathbf{u})}{\|\mathbf{u}\|}.$$

As is standard, the dual norm of $\mathcal{R}$ is defined as $\mathcal{R}^*(\mathbf{v}) := \sup_{\mathcal{R}(\mathbf{u}) \leq 1} \langle \mathbf{u}, \mathbf{v} \rangle$. To simplify notation, we let $\|\mathbf{u}\|_{\mathcal{R}} := \mathcal{R}(\mathbf{u})$ and $\|\mathbf{u}\|_{\mathcal{R}^*} := \mathcal{R}^*(\mathbf{u})$.

4.2 Conditions on $Q(\cdot|\cdot)$ and $Q_n(\cdot|\cdot)$

Next, we review three technical conditions, originally proposed by [1], on the population level $Q(\cdot|\cdot)$ function, and then we give two important conditions that the empirial function $Q_n(\cdot|\cdot)$ must satisfy, including one that characterizes the statistical error.

It is well known that performance of EM algorithm is sensitive to initialization. Following the low-dimensional development in [1], our results are local, and apply to an r-neighborhood region around $\boldsymbol{\beta}^*$: $\mathcal{B}(r; \boldsymbol{\beta}^*) := \{\mathbf{u} \in \Omega, \|\mathbf{u} - \boldsymbol{\beta}^*\| \leq r\}$.

We first require that $Q(\cdot|\boldsymbol{\beta}^*)$ is *self consistent* as stated below. This is satisfied, in particular, when $\boldsymbol{\beta}^*$ maximizes the population log likelihood function, as happens in most settings of interest [12].

Condition 1 (Self Consistency)**.** *Function $Q(\cdot|\boldsymbol{\beta}^*)$ is self consistent, namely*

$$\boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\beta} \in \Omega} Q(\boldsymbol{\beta}|\boldsymbol{\beta}^*).$$

We also require that the function $Q(\cdot|\cdot)$ satisfies a certain strong concavity condition and is smooth over Ω.

Condition 2 (Strong Concavity and Smoothness (γ, μ, r))**.** *$Q(\cdot|\boldsymbol{\beta}^*)$ is γ-strongly concave over Ω, i.e.,*

$$Q(\boldsymbol{\beta}_2|\boldsymbol{\beta}^*) - Q(\boldsymbol{\beta}_1|\boldsymbol{\beta}^*) - \langle \nabla Q(\boldsymbol{\beta}_1|\boldsymbol{\beta}^*), \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1 \rangle \leq -\frac{\gamma}{2}\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|^2, \ \forall\, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \Omega. \tag{4.1}$$

For any $\boldsymbol{\beta} \in \mathcal{B}(r; \boldsymbol{\beta}^*)$, $Q(\cdot|\boldsymbol{\beta})$ is μ-smooth over Ω, i.e.,

$$Q(\boldsymbol{\beta}_2|\boldsymbol{\beta}) - Q(\boldsymbol{\beta}_1|\boldsymbol{\beta}) - \langle \nabla Q(\boldsymbol{\beta}_1|\boldsymbol{\beta}), \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1 \rangle \geq -\frac{\mu}{2}\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|^2, \ \forall\, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \Omega. \tag{4.2}$$

The next condition is key in guaranteeing the curvature of $Q(\cdot|\boldsymbol{\beta})$ is similar to that of $Q(\cdot|\boldsymbol{\beta}^*)$ when $\boldsymbol{\beta}$ is close to $\boldsymbol{\beta}^*$. It has also been called *First Order Stability* in [1].

Condition 3 (Gradient Stability (τ, r))**.** *For any $\boldsymbol{\beta} \in \mathcal{B}(r; \boldsymbol{\beta}^*)$, we have*

$$\big\|\nabla Q(\mathcal{M}(\boldsymbol{\beta})|\boldsymbol{\beta}) - \nabla Q(\mathcal{M}(\boldsymbol{\beta})|\boldsymbol{\beta}^*)\big\| \leq \tau\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|.$$

The above condition only requires that the gradient be stable at one point $\mathcal{M}(\boldsymbol{\beta})$. This is sufficient for our analysis. In fact, for many concrete examples, one can verify a stronger version of Condition 3 that is $\big\|\nabla Q(\boldsymbol{\beta}'|\boldsymbol{\beta}) - \nabla Q(\boldsymbol{\beta}'|\boldsymbol{\beta}^*)\big\| \leq \tau\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|, \ \forall\, \boldsymbol{\beta}' \in \mathcal{B}(r; \boldsymbol{\beta}^*)$.

Next we require two conditions on the empirical function $Q_n(\cdot|\cdot)$, which is computed from finite number of samples according to (2.1). Our first condition, parallel to Condition 2, imposes a curvature constraint on $Q_n(\cdot|\cdot)$. In order to guarantee that the estimation error $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|$ in step t of the EM algorithm is well controlled, we would like $Q_n(\cdot|\boldsymbol{\beta}^{(t-1)})$ to be strongly concave at $\boldsymbol{\beta}^*$. However, in the setting where $n \ll p$, there might exist directions along which $Q_n(\cdot|\boldsymbol{\beta}^{(t-1)})$ is flat, e.g., as in mixed linear regression and missing covariate regression. In contrast with Condition 2, we only require $Q_n(\cdot|\cdot)$ to be strongly concave over a particular set $\mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \mathcal{R})$ that is defined in terms of the subspace pair $(\mathcal{S}, \overline{\mathcal{S}})$ and regularizer $\mathcal{R}$. This set is defined as follows:

$$\mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \mathcal{R}) := \left\{\mathbf{u} \in \mathbb{R}^p : \big\|\Pi_{\overline{\mathcal{S}}^\perp}(\mathbf{u})\big\|_{\mathcal{R}} \leq 2 \cdot \big\|\Pi_{\overline{\mathcal{S}}}(\mathbf{u})\big\|_{\mathcal{R}} + 2 \cdot \Psi(\overline{\mathcal{S}}) \cdot \big\|\mathbf{u}\big\|\right\}, \tag{4.3}$$

where the projection operator $\Pi_{\mathcal{S}} : \mathbb{R}^p \to \mathbb{R}^p$ is defined as $\Pi_{\mathcal{S}}(\mathbf{u}) := \arg\min_{\mathbf{v} \in \mathcal{S}} \|\mathbf{v} - \mathbf{u}\|$. The restricted strong concavity (RSC) condition is as follows.

Condition 4 (RSC $(\gamma_n, \mathcal{S}, \overline{\mathcal{S}}, r, \delta)$)**.** *For any fixed $\boldsymbol{\beta} \in \mathcal{B}(r; \boldsymbol{\beta}^*)$, with probability at least $1-\delta$, we have that for all $\boldsymbol{\beta}' - \boldsymbol{\beta}^* \in \Omega \bigcap \mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \mathcal{R})$,*

$$Q_n(\boldsymbol{\beta}'|\boldsymbol{\beta}) - Q_n(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - \left\langle \nabla Q_n(\boldsymbol{\beta}^*|\boldsymbol{\beta}), \boldsymbol{\beta}' - \boldsymbol{\beta}^* \right\rangle \leq -\frac{\gamma_n}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}^*\|^2.$$

The above condition states that $Q_n(\cdot|\boldsymbol{\beta})$ is strongly concave in directions $\boldsymbol{\beta}' - \boldsymbol{\beta}^*$ that belong to $\mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \mathcal{R})$. It is instructive to compare Condition 4 with a related condition proposed by [14] for analyzing high dimensional M-estimators. They require the loss function to be strongly convex over the cone $\{\mathbf{u} \in \mathbb{R}^p : \|\Pi_{\overline{\mathcal{S}}^\perp}(\mathbf{u})\|_{\mathcal{R}} \lesssim \|\Pi_{\overline{\mathcal{S}}}(\mathbf{u})\|_{\mathcal{R}}\}$. Therefore our restrictive set (4.3) is similar to the cone but has the additional term $2\Psi(\overline{\mathcal{S}})\|\mathbf{u}\|$. The main purpose of the term $2\Psi(\overline{\mathcal{S}})\|\mathbf{u}\|$ is to allow the regularization parameter λ_n to jointly control optimization and statistical error. We note that while Condition 4 is stronger than the usual RSC condition in M-estimator, in typical settings the difference is immaterial. This is because $\left\|\Pi_{\overline{\mathcal{S}}}(\mathbf{u})\right\|_{\mathcal{R}}$ is within a constant factor of $\Psi(\overline{\mathcal{S}}) \cdot \left\|\mathbf{u}\right\|$, and hence checking RSC over $\mathcal{C}$ amounts to checking it over $\|\Pi_{\overline{\mathcal{S}}^\perp}(\mathbf{u})\|_{\mathcal{R}} \lesssim \Psi(\overline{\mathcal{S}})\|\mathbf{u}\|$, which is indeed what is typically also done in the M-estimator setting.

Finally, we establish the condition that characterizes the achievable statistical error.

Condition 5 (Statistical Error (Δ_n, r, δ))**.** *For any fixed $\boldsymbol{\beta} \in \mathcal{B}(r; \boldsymbol{\beta}^*)$, with probability at least $1-\delta$, we have*

$$\left\|\nabla Q_n(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - \nabla Q(\boldsymbol{\beta}^*|\boldsymbol{\beta})\right\|_{\mathcal{R}^*} \leq \Delta_n. \tag{4.4}$$

This quantity replaces the term $\|\mathcal{M}_n(\boldsymbol{\beta}) - \mathcal{M}(\boldsymbol{\beta})\|$ which appears in [1] and [20], and which presents problems in the high dimensional regime.

4.3 Main Results

In this section, we provide the theoretical guarantees for a *resampled version* of our regularized EM algorithm: we split the whole dataset into T pieces and use a fresh piece of data in each iteration of regularized EM. As in [1], resampling makes it possible to check that Conditions 4-5 are satisfied without requiring them to hold uniformly for all $\boldsymbol{\beta} \in \mathcal{B}(r; \boldsymbol{\beta}^*)$ with high probability. Our empirical results indicate that it is not in fact required and is an artifact of the analysis. We refer to this resampled version as **Algorithm 2**. In the sequel, we let $m := n/T$ to denote the sample complexity in each iteration. We let $\alpha := \sup_{\mathbf{u} \in \mathbb{R}^p \setminus \{0\}} \|\mathbf{u}\|_*/\|\mathbf{u}\|$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

For Algorithm 2, our main result is as follows. The proof is deferred to the Supplemental Material.

Theorem 1. *Assume the model parameter $\boldsymbol{\beta}^* \in \mathcal{S}$ and regularizer $\mathcal{R}$ is decomposable with respect to $(\mathcal{S}, \overline{\mathcal{S}})$ where $\mathcal{S} \subseteq \overline{\mathcal{S}} \subseteq \mathbb{R}^p$. Assume $r > 0$ is such that $\mathcal{B}(r; \boldsymbol{\beta}^*) \subseteq \Omega$. Further, assume function $Q(\cdot|\cdot)$, defined in* (2.2)*, is self consistent and satisfies Conditions 2-3 with parameters (γ, μ, r) and (τ, r). Given n samples and T iterations, let $m := n/T$. Assume $Q_m(\cdot|\cdot)$, computed from any m i.i.d. samples according to* (2.1)*, satisfies Conditions 4-5 with parameters $(\gamma_m, \mathcal{S}, \overline{\mathcal{S}}, r, 0.5\delta/T)$ and $(\Delta_m, r, 0.5\delta/T)$. Let $\kappa^* := 5\frac{\alpha\mu\tau}{\gamma\gamma_m}$, and assume $0 < \tau < \gamma$ and $0 < \kappa^* \leq 3/4$. Define $\overline{\Delta} := r\gamma_m/[60\Psi(\overline{\mathcal{S}})]$ and assume Δ_m is such that $\Delta_m \leq \overline{\Delta}$.*

Consider Algorithm 2 with initialization $\boldsymbol{\beta}^{(0)} \in \mathcal{B}(r; \boldsymbol{\beta}^*)$ and with regularization parameters given by

$$\lambda_m^{(t)} = \kappa^t \frac{\gamma_m}{5\Psi(\overline{\mathcal{S}})}\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\| + \frac{1-\kappa^t}{1-\kappa}\Delta,\ t = 1, 2, \ldots, T \tag{4.5}$$

for any $\Delta \in [3\Delta_m, 3\overline{\Delta}]$, $\kappa \in [\kappa^*, 3/4]$. Then with probability at least $1-\delta$, we have that for any $t \in [T]$,

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\| \leq \kappa^t\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\| + \frac{5}{\gamma_m}\frac{1-\kappa^t}{1-\kappa}\Psi(\overline{\mathcal{S}})\Delta. \tag{4.6}$$

The estimation error is bounded by a term decaying linearly with number of iterations t, which we can think of as the *optimization error* and a second term that characterizes the ultimate *estimation error* of our algorithm. With $T = O(\log n)$ and suitable choice of Δ such that $\Delta = O(\Delta_{n/T})$, we bound the ultimate estimation error as

$$\|\boldsymbol{\beta}^{(T)} - \boldsymbol{\beta}^*\| \lesssim \frac{1}{(1-\kappa)\gamma_{n/T}}\Psi(\overline{\mathcal{S}})\Delta_{n/T}. \tag{4.7}$$

We note that overestimating the initial error, $\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|$ is not important, as it may slightly increase the overall number of iterations, but will not impact the ultimate estimation error.

The constraint $\Delta_m \lesssim r\gamma_m/\Psi(\overline{\mathcal{S}})$ ensures that $\boldsymbol{\beta}^{(t)}$ is contained in $\mathcal{B}(r; \boldsymbol{\beta}^*)$ for all $t \in [T]$. This constraint is quite mild in the sense that if $\Delta_m = \Omega(r\gamma_m/\Psi(\overline{\mathcal{S}}))$, $\boldsymbol{\beta}^{(0)}$ is a decent estimator with estimation error $O(\Psi(\overline{\mathcal{S}})\Delta_m/\gamma_m)$ that already matches our expectation.

5 Examples: Applying the Theory

Now we introduce three well known latent variable models. For each model, we first review the standard EM algorithm formulations, and discuss the extensions to the high dimensional setting. Then we apply Theorem 1 to obtain the statistical guarantee of the regularized EM with data splitting (Algorithm 2). The key ingredient underlying these results is to check the technical conditions in Section 4 hold for each model. We postpone these tedious details to the Supplemental Material.

5.1 Gaussian Mixture Model

We consider the balanced isotropic Gaussian mixture model (GMM) with two components where the distribution of random variables $(Y, Z) \in \mathbb{R}^p \times \{-1, 1\}$ is characterized as

$$\Pr(Y = \mathbf{y}|Z = z) = \phi(\mathbf{y}; z \cdot \boldsymbol{\beta}^*, \sigma^2\mathbf{I}_p),\ \Pr(Z = 1) = \Pr(Z = -1) = 1/2.$$

Here we use $\phi(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the probability density function of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In this example, Z is the latent variable that indicates the cluster id of each sample. Given n i.i.d. samples $\{\mathbf{y}_i\}_{i=1}^n$, function $Q_n(\cdot|\cdot)$ defined in (2.1) corresponds to

$$Q_n^{GMM}(\boldsymbol{\beta}'|\boldsymbol{\beta}) = -\frac{1}{2n}\sum_{i=1}^n \left[w(\mathbf{y}_i; \boldsymbol{\beta})\|\mathbf{y}_i - \boldsymbol{\beta}'\|_2^2 + (1 - w(\mathbf{y}_i; \boldsymbol{\beta}))\|\mathbf{y}_i + \boldsymbol{\beta}'\|_2^2\right], \tag{5.1}$$

where $w(\mathbf{y}; \boldsymbol{\beta}) := \exp(-\frac{\|\mathbf{y}-\boldsymbol{\beta}\|_2^2}{2\sigma^2})[\exp(-\frac{\|\mathbf{y}-\boldsymbol{\beta}\|_2^2}{2\sigma^2}) + \exp(-\frac{\|\mathbf{y}+\boldsymbol{\beta}\|_2^2}{2\sigma^2})]^{-1}$. We assume $\boldsymbol{\beta}^* \in \mathcal{B}_0(s; p) := \{\mathbf{u} \in \mathbb{R}^p : |\operatorname{supp}(\mathbf{u})| \le s\}$. Naturally, we choose the regularizer $\mathcal{R}(\cdot)$ to be the ℓ_1 norm. We define the signal-to-noise ratio $\mathrm{SNR} := \|\boldsymbol{\beta}^*\|_2/\sigma$.

Corollary 1 (Sparse Recovery in GMM). *There exist constants ρ, C such that if $\mathrm{SNR} \ge \rho$, $n/T \ge [80C(\|\boldsymbol{\beta}^*\|_\infty + \sigma)/\|\boldsymbol{\beta}^*\|_2]^2\, s\log p$, $\boldsymbol{\beta}^{(0)} \in \mathcal{B}(\|\boldsymbol{\beta}^*\|_2/4; \boldsymbol{\beta}^*)$; then with probability at least $1 - T/p$ Algorithm 2 with parameters $\Delta = C(\|\boldsymbol{\beta}^*\|_\infty + \sigma)\sqrt{T\log p/n}$, $\lambda_{n/T}^{(0)} = 0.2\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2/\sqrt{s}$, any $\kappa \in [1/2, 3/4]$ and ℓ_1 regularization generates $\boldsymbol{\beta}^{(t)}$ that has estimation error*

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2 \le \kappa^t\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2 + \frac{5C(\|\boldsymbol{\beta}^*\|_\infty + \sigma)}{1 - \kappa}\sqrt{\frac{s\log p}{n}}T,\ \textit{for all } t \in [T]. \tag{5.2}$$

Note that by setting $T \asymp \log(n/\log p)$, the order of final estimation error turns out to be $(\|\boldsymbol{\beta}^*\|_\infty + \delta)\sqrt{(s\log p)/n)\log(n/\log p)}$. The minimax rate for estimating s-sparse vector in a single Gaussian cluster is $\sqrt{s\log p/n}$, thereby the rate is optimal on (n, p, s) up to a log factor.

5.2 Mixed Linear Regression

Mixed linear regression (MLR), as considered in some recent work [5, 7, 22], is the problem of recovering two or more linear vectors from mixed linear measurements. In the case of mixed linear regression with two symmetric and balanced components, the response-covariate pair $(Y, X) \in \mathbb{R} \times \mathbb{R}^p$ is linked through

$$Y = \langle X,\ Z \cdot \boldsymbol{\beta}^* \rangle + W,$$

where W is the noise term and Z is the latent variable that has Rademacher distribution over $\{-1, 1\}$. We assume $X \sim \mathcal{N}(0, \mathbf{I}_p)$, $W \sim \mathcal{N}(0, \sigma^2)$. In this setting, with n i.i.d. samples $\{y_i, \mathbf{x}_i\}_{i=1}^n$ of pair (Y, X), function $Q_n(\cdot|\cdot)$ then corresponds to

$$Q_n^{MLR}(\boldsymbol{\beta}'|\boldsymbol{\beta}) = -\frac{1}{2n}\sum_{i=1}^n \left[w(y_i, \mathbf{x}_i; \boldsymbol{\beta})(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta}' \rangle)^2 + (1 - w(y_i, \mathbf{x}_i; \boldsymbol{\beta}))(y_i + \langle \mathbf{x}_i, \boldsymbol{\beta}' \rangle)^2\right], \tag{5.3}$$

where $w(y, \mathbf{x}; \boldsymbol{\beta}) := \exp\left(-\frac{(y-\langle \mathbf{x}, \boldsymbol{\beta}\rangle)^2}{2\sigma^2}\right)[\exp\left(-\frac{(y-\langle \mathbf{x}, \boldsymbol{\beta}\rangle)^2}{2\sigma^2}\right) + \exp\left(-\frac{(y+\langle \mathbf{x}, \boldsymbol{\beta}\rangle)^2}{2\sigma^2}\right)]^{-1}$.

We consider two kinds of structure on $\boldsymbol{\beta}^*$:

Sparse Recovery. Assume $\boldsymbol{\beta}^* \in \mathcal{B}_0(s; p)$. Then let $\mathcal{R}$ be the ℓ_1 norm, as in the previous section. We define $\mathrm{SNR} := \|\boldsymbol{\beta}^*\|_2/\sigma$.

Corollary 2 (Sparse recovery in MLR)**.** *There exist constant ρ, C, C' such that if* $\mathrm{SNR} \geq \rho$, $n/T \geq C'\left[(\|\boldsymbol{\beta}^*\|_2 + \delta)/\|\boldsymbol{\beta}^*\|_2\right]^2 s \log p$, $\boldsymbol{\beta}^{(0)} \in \mathcal{B}(\|\boldsymbol{\beta}^*\|_2/240, \boldsymbol{\beta}^*)$*; then with probability at least* $1 - T/p$ *Algorithm 2 with parameters* $\Delta = C(\|\boldsymbol{\beta}^*\|_2 + \delta)\sqrt{T \log p/n}$, $\lambda^{(0)}_{n/T} = \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2/(15\sqrt{s})$, *any* $\kappa \in [1/2, 3/4]$ *and* ℓ_1 *regularization generates* $\boldsymbol{\beta}^{(t)}$ *that has estimation error*

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2 \leq \kappa^t \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2 + \frac{15C(\|\boldsymbol{\beta}^*\|_2 + \delta)}{1-\kappa}\sqrt{\frac{s \log p}{n}T}, \text{ for all } t \in [T].$$

Performing $T \asymp \log(n/(s \log p))$ iterations gives us estimation rate $(\|\boldsymbol{\beta}^*\|_2 + \delta)\sqrt{(s \log p/n) \log(n/(s \log p))}$ which is near-optimal on (s, p, n). The dependence on $\|\boldsymbol{\beta}^*\|_2$, which also appears in the analysis of EM in the classical (low dimensional) setting [1], arises from fundamental limits of EM. Removing such dependence for MLR is possible by convex relaxation [7]. It is interesting to study how to remove it in the high dimensional setting.

Low Rank Recovery. Second we consider the setting where the model parameter is a matrix $\boldsymbol{\Gamma}^* \in \mathbb{R}^{p_1 \times p_2}$ with $\mathrm{rank}(\boldsymbol{\Gamma}^*) = \theta \ll \min(p_1, p_2)$. We further assume $X \in \mathbb{R}^{p_1 \times p_2}$ is an i.i.d. Gaussian matrix, i.e., entries of X are independent random variables with distribution $\mathcal{N}(0, 1)$. Note that in the low dimensional case $n \gg p_1 \times p_2$, there is no essential difference between assuming the parameter is a vector or matrix since we can always treat X and $\boldsymbol{\Gamma}^*$ as $(p_1 \times p_2)$-dimensional vectors. In the high dimensional regime, we apply nuclear norm regularization, i.e, $\mathcal{R}(\boldsymbol{\Gamma}) = \sum_{i=1}^{p_1, p_2} |s_i(\boldsymbol{\Gamma})|$, where $s_i(\boldsymbol{\Gamma})$ is the ith singular value of $\boldsymbol{\Gamma}$. Similarly, $\mathrm{SNR} := \|\boldsymbol{\Gamma}^*\|_F/\sigma$.

Corollary 3 (Low rank recovery in MLR)**.** *There exist constant ρ, C, C' such that if* $\mathrm{SNR} \geq \rho$, $n/T \geq C'\left[(\|\boldsymbol{\Gamma}^*\|_F + \sigma)/\|\boldsymbol{\Gamma}^*\|_F\right]^2 \theta(p_1 + p_2)$, $\boldsymbol{\Gamma}^{(0)} \in \mathcal{B}(\|\boldsymbol{\Gamma}^*\|_F/1600, \boldsymbol{\Gamma}^*)$*; then with probability at least* $1 - T\exp(-p_1 - p_2)$ *Algorithm 2 with parameters* $\Delta = C(\|\boldsymbol{\Gamma}^*\|_F + \sigma)\sqrt{T(p_1 + p_2)/n}$, $\lambda^{(0)}_{n/T} = 0.01\|\boldsymbol{\Gamma}^{(0)} - \boldsymbol{\Gamma}^*\|_F/\sqrt{2\theta}$, *any* $\kappa \in [1/2, 3/4]$ *and nuclear norm regularization generates* $\boldsymbol{\Gamma}^{(t)}$ *that has estimation error*

$$\|\boldsymbol{\Gamma}^{(t)} - \boldsymbol{\Gamma}^*\|_F \leq \kappa^t \|\boldsymbol{\Gamma}^{(0)} - \boldsymbol{\Gamma}^*\|_F + \frac{100C'(\|\boldsymbol{\Gamma}^*\|_F + \sigma)}{1-\kappa}\sqrt{\frac{2\theta(p_1 + p_2)}{n}T}, \text{ for all } t \in [T].$$

For suitable choice of T, one can show the final estimation error is near-optimal by following similar analysis as for sparse recovery. The standard low rank matrix recovery with a single component, including other sensing matrix designs beyond the Gaussian matrix, has been studied extensively (e.g., [2, 4, 13, 15]). To the best of our knowledge, the theoretical study of the mixed low rank matrix recovery has not been considered.

5.3 Missing Covariate Regression

As our last example, we consider the missing covariate regression (MCR) problem. To parallel standard linear regression, $\{y_i, \mathbf{x}_i\}_{i=1}^n$ are samples of (Y, X) linked through $Y = \langle X, \boldsymbol{\beta}^*\rangle + W$. However, we assume each entry of $\mathbf{x}_i$ is missing independently with probability $\epsilon \in (0, 1)$. Therefore, the observed covariate vector $\widetilde{\mathbf{x}}_i$ takes the form

$$\widetilde{x}_{i,j} = \begin{cases} x_{i,j} & \text{with probability } 1 - \epsilon \\ * & \text{otherwise} \end{cases}.$$

We assume the model is under Gaussian design $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p), W \sim \mathcal{N}(0, \sigma^2)$. We refer the reader to our Supplementary Material for the specific $Q_n(\cdot|\cdot)$ function. In high dimensional case, we assume $\boldsymbol{\beta}^* \in \mathcal{B}_0(s; p)$. We define $\rho := \|\boldsymbol{\beta}^*\|_2/\sigma$ to be the SNR and $\omega := r/\|\boldsymbol{\beta}^*\|_2$ to be the *relative contractivity radius*. In particular, let $\zeta := (1 + \omega)\rho$.

Corollary 4 (Sparse Recovery in MCR)**.** *There exist constants C, C', C_0, C_1 such that if* $(1+\omega)\rho \leq C_0 < 1$, $\epsilon < C_1$, $n/T \geq C' \max\{\sigma^2(\omega\rho)^{-1}, 1\} s \log p$, $\boldsymbol{\beta}^{(0)} \in \mathcal{B}(\omega\|\boldsymbol{\beta}^*\|_2, \boldsymbol{\beta}^*)$*; then with probability at least* $1 - T/p$ *Algorithm 2 with parameters* $\Delta = C\sigma\sqrt{T \log p/n}$, $\lambda^{(0)}_{n/T} = \|\boldsymbol{\beta}^{(0)} -$

$\boldsymbol{\beta}^*\|_2/(45\sqrt{s})$, *any* $\kappa \in [1/2, 3/4]$ *and* ℓ_1 *regularization generates* $\boldsymbol{\beta}^{(t)}$ *that has estimation error*

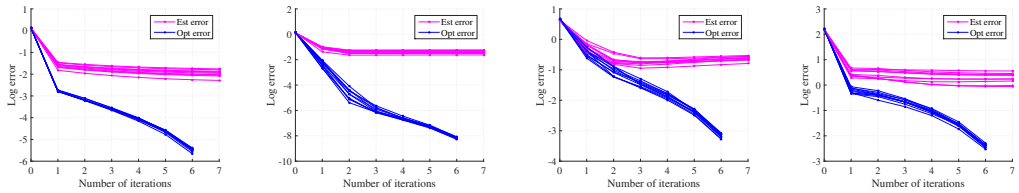
$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2 \leq \kappa^t \|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2 + \frac{45C\sigma}{1-\kappa}\sqrt{\frac{s\log p}{n}}T, \text{ for all } t \in [T],$$

Unlike the previous two models, we require an upper bound on the signal to noise ratio. This unusual constraint is in fact unavoidable [10]. By optimizing T, the order of final estimation error turns out to be $\sigma\sqrt{s\log p/n\log(n/(s\log p))}$.

6 Simulations

We now provide some simulation results to back up our theory. Note that while Theorem 1 requires resampling, we believe in practice this is unnecessary. This is validated by our results, where we apply Algorithm 1 to the four latent variable models discussed in Section 5.

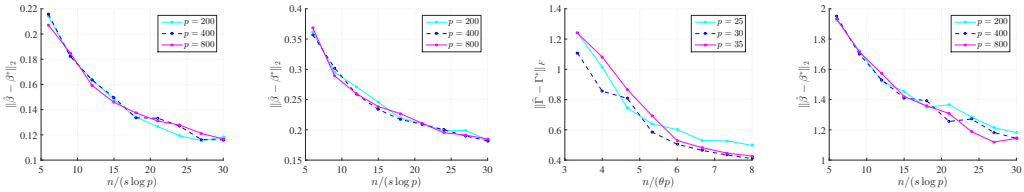
Convergence Rate. We first evaluate the convergence of Algorithm 1 assuming only that the initialization is a bounded distance from $\boldsymbol{\beta}^*$. For a given error $\omega\|\boldsymbol{\beta}^*\|_2$, the initial parameter $\boldsymbol{\beta}^{(0)}$ is picked randomly from the sphere centered around $\boldsymbol{\beta}^*$ with radius $\omega\|\boldsymbol{\beta}^*\|_2$. We use Algorithm 1 with $T = 7$, $\kappa = 0.7$, $\lambda_n^{(0)}$ in Theorem 1. The choice of the critical parameter Δ is given in the Supplementary Material. For every single trial, we report *estimation error* $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|_2$ and *optimization error* $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(T)}\|_2$ in every iteration. We plot the log of errors over iteration t in Figure 1.



(a) GMM (b) MLR(sparse) (c) MLR(low rank) (d) MCR

Figure 1: Convergence of regularized EM algorithm. In each panel, one curve is plotted from single independent trial. **Settings**: (a,b,d) $(n, p, s) = (500, 800, 5)$; (d) $(n, p, \theta) = (600, 30, 3)$; (a-c) SNR $= 5$; (d) $(\text{SNR}, \epsilon) = (0.5, 0.2)$; (a-d) $\omega = 0.5$.

Statistical Rate. We now evaluate the statistical rate. We set $T = 7$ and compute estimation error on $\widehat{\boldsymbol{\beta}} := \boldsymbol{\beta}^{(T)}$. In Figure 2, we plot $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ over normalized sample complexity, i.e., $n/(s\log p)$ for s-sparse parameter and $n/(\theta p)$ for rank θ p-by-p parameter. We refer the reader to Figure 1 for other settings. We observe that the same normalized sample complexity leads to almost identical estimation error in practice, which thus supports the corresponding statistical rate established in Section 5.



(a) GMM (b) MLR(sparse) (c) MLR(low rank) (d) MCR

Figure 2: Statistical rates. Each point is an average of 20 independent trials. **Settings**: (a,b,d) $s = 5$; (c) $\theta = 3$.

References

[1] Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *arXiv preprint arXiv:1408.2156*, 2014.

[2] T Tony Cai and Anru Zhang. Rop: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

[3] Emmanuel Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

[4] Emmanuel J Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57(4):2342–2359, 2011.

[5] Arun Tejasvi Chaganty and Percy Liang. Spectral experts for estimating mixtures of linear regressions. *arXiv preprint arXiv:1306.3729*, 2013.

[6] Yudong Chen, Sujay Sanghavi, and Huan Xu. Improved graph clustering. *Information Theory, IEEE Transactions on*, 60(10):6440–6455, Oct 2014.

[7] Yudong Chen, Xinyang Yi, and Constantine Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. In *Conf. on Learning Theory*, 2014.

[8] 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)*, pages 1–38, 1977.

[9] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.

[10] Po-Ling Loh and Martin J Wainwright. Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, pages 2601–2605. IEEE, 2012.

[11] Jinwen Ma and Lei Xu. Asymptotic convergence properties of the em algorithm with respect to the overlap in the mixture. *Neurocomputing*, 68:105–129, 2005.

[12] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[13] Sahand Negahban, Martin J Wainwright, et al. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

[14] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

[15] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[16] Nicolas Städler, Peter Bühlmann, and Sara Van De Geer. L1-penalization for mixture regression models. *Test*, 19(2):209–256, 2010.

[17] Paul Tseng. An analysis of the em algorithm and entropy-like proximal point methods. *Mathematics of Operations Research*, 29(1):27–44, 2004.

[18] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[19] Martin J Wainwright. Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and Its Application*, 1:233–253, 2014.

[20] Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729*, 2014.

[21] C.F.Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

[22] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. *arXiv preprint arXiv:1310.3745*, 2013.

Regularized EM: Supplemental Material

We give the proof of the main result, as well as the proofs showing the specialization of our results to the examples discussed in Section B. We collect several technical lemmas in Section C, to which we forward-reference in the results in the next sections.

A Proof of Main Result

In this section, we provide the proof of Theorem 1 that characterizes the computational and statistical performance of the regularized EM algorithm with resampling. We first present a result which shows that the population EM operator $\mathcal{M} : \Omega \to \Omega$ is contractive when $\tau < \gamma$.

Lemma 1. *Suppose $Q(\cdot|\cdot)$ satisfies all the corresponding conditions stated in Theorem 1. Mapping $\mathcal{M}$ is contractive over $\mathcal{B}(r; \boldsymbol{\beta}^*)$, namely*

$$\|\mathcal{M}(\boldsymbol{\beta}) - \boldsymbol{\beta}^*\| \leq \frac{\tau}{\gamma}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|, \ \forall \ \boldsymbol{\beta} \in \mathcal{B}(r; \boldsymbol{\beta}^*).$$

Proof. A similar result is proved in [1]. The slight difference is that [1] shows Lemma 1 with ℓ_2 norm. Extending ℓ_2 norm to arbitrary norm is trivial, so we omit the details. □

Now we are ready to prove Theorem 1.

Proof of Theorem 1. We first consider one iteration of Algorithm 1 and show the relationship between $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|$ and $\|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|$. Recall that

$$\boldsymbol{\beta}^{(t)} = \arg\max_{\boldsymbol{\beta}' \in \Omega} Q_m(\boldsymbol{\beta}'|\boldsymbol{\beta}^{(t-1)}) - \lambda_m^{(t)} \cdot \mathcal{R}(\boldsymbol{\beta}'),$$

where $m = n/T$ is the number of samples in each step. We assume $\boldsymbol{\beta}^{(t-1)} \in \mathcal{B}(r; \boldsymbol{\beta}^*)$. To simplify the notation, we drop the superscripts of $\boldsymbol{\beta}^{(t-1)}$, $\lambda_m^{(t)}$ and denote $\boldsymbol{\beta}^{(t)}$ as $\boldsymbol{\beta}^+$. From the optimality of $\boldsymbol{\beta}^+$, we have

$$Q_m(\boldsymbol{\beta}^+|\boldsymbol{\beta}) - \lambda_m \cdot \mathcal{R}(\boldsymbol{\beta}^+) \geq Q_m(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - \lambda_m \cdot \mathcal{R}(\boldsymbol{\beta}^*). \tag{A.1}$$

Equivalently,

$$\lambda_m \cdot \mathcal{R}(\boldsymbol{\beta}^+) - \lambda_m \cdot \mathcal{R}(\boldsymbol{\beta}^*) \leq Q_m(\boldsymbol{\beta}^+|\boldsymbol{\beta}) - Q_m(\boldsymbol{\beta}^*|\boldsymbol{\beta}). \tag{A.2}$$

Using the fact that $Q_m(\cdot|\boldsymbol{\beta})$ is a concave function, the right hand side of the above inequality can be bounded as

$$Q_m(\boldsymbol{\beta}^+|\boldsymbol{\beta}) - Q_m(\boldsymbol{\beta}^*|\boldsymbol{\beta}) \leq \big\langle \nabla Q_m(\boldsymbol{\beta}^*|\boldsymbol{\beta}), \boldsymbol{\beta}^+ - \boldsymbol{\beta} \big\rangle \leq \underbrace{\big|\big\langle \nabla Q_m(\boldsymbol{\beta}^*|\boldsymbol{\beta}), \boldsymbol{\beta}^+ - \boldsymbol{\beta} \big\rangle\big|}_{A}. \tag{A.3}$$

A key ingredient of our proof is to bound the term A. Letting $\Theta := \boldsymbol{\beta}^+ - \boldsymbol{\beta}^*$, we have

$$\begin{aligned}
\big|\big\langle \nabla Q_m(\boldsymbol{\beta}^*|\boldsymbol{\beta}), \boldsymbol{\beta}^+ - \boldsymbol{\beta} \big\rangle\big| &= \big|\big\langle \nabla Q_m(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - \nabla Q(\boldsymbol{\beta}^*|\boldsymbol{\beta}) + \nabla Q(\boldsymbol{\beta}^*|\boldsymbol{\beta}), \Theta \big\rangle\big| \\
&\leq \big|\big\langle \nabla Q_m(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - \nabla Q(\boldsymbol{\beta}^*|\boldsymbol{\beta}), \Theta \big\rangle\big| + \big|\big\langle \nabla Q(\boldsymbol{\beta}^*|\boldsymbol{\beta}), \Theta \big\rangle\big| \\
&\overset{(a)}{\leq} \big\|Q_m(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - \nabla Q(\boldsymbol{\beta}^*|\boldsymbol{\beta})\|_{\mathcal{R}^*} \cdot \mathcal{R}(\Theta) + \big\|\nabla Q(\boldsymbol{\beta}^*|\boldsymbol{\beta})\big\|_* \times \|\Theta\| \\
&\overset{(b)}{\leq} \Delta_m \mathcal{R}(\Theta) + \alpha\big\|\nabla Q(\boldsymbol{\beta}^*|\boldsymbol{\beta})\big\| \times \|\Theta\| \\
&\overset{(c)}{\leq} \Delta_m \mathcal{R}(\Theta) + \alpha\big\|\nabla Q(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - \nabla Q(\mathcal{M}(\boldsymbol{\beta})|\boldsymbol{\beta})\big\| \times \|\Theta\| \\
&\overset{(d)}{\leq} \Delta_m \mathcal{R}(\Theta) + \alpha\mu\big\|\mathcal{M}(\boldsymbol{\beta}) - \boldsymbol{\beta}^*\big\| \times \|\Theta\| \\
&\overset{(e)}{\leq} \Delta_m \mathcal{R}(\Theta) + \frac{\alpha\mu\tau}{\gamma}\big\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\big\| \times \|\Theta\|
\end{aligned} \tag{A.4}$$

where (a) follows from the Cauchy-Schwarz inequality, (b) follows from the statistical error Condition 5 and the definition of α, (c) follows from the fact that $\mathcal{M}(\boldsymbol{\beta})$ maximizes $Q(\cdot|\boldsymbol{\beta})$, (d) follows from the smoothness Condition 2, and (e) follows from Lemma 1. For inequality (c), note that we

assume that $\mathcal{B}(r;\boldsymbol{\beta}^*) \subseteq \Omega$. From Lemma 1, we know that if $\boldsymbol{\beta} \in \mathcal{B}(r;\boldsymbol{\beta}^*)$, under condition $\tau < \gamma$, we must have $\mathcal{M}(\boldsymbol{\beta}) \in \mathcal{B}(r\tau/\gamma;\boldsymbol{\beta}^*) \subseteq \mathcal{B}(r;\boldsymbol{\beta}^*)$. Therefore $\mathcal{M}(\boldsymbol{\beta})$ lies in the interior of Ω thus the optimality condition corresponds to $\nabla Q(\mathcal{M}(\boldsymbol{\beta})|\boldsymbol{\beta}) = \mathbf{0}$.

Plugging (A.4) back into (A.3), we obtain

$$Q_m(\boldsymbol{\beta}^+|\boldsymbol{\beta}) - Q_m(\boldsymbol{\beta}^*|\boldsymbol{\beta}) \leq \Delta_m \mathcal{R}(\Theta) + \frac{\alpha\mu\tau}{\gamma}\big\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\big\| \times \|\Theta\|.$$

Using the above result and (A.2), we have

$$\lambda_m \mathcal{R}(\boldsymbol{\beta}^* + \Theta) - \lambda_m \mathcal{R}(\boldsymbol{\beta}^*) \leq \Delta_m \mathcal{R}(\Theta) + \frac{\alpha\mu\tau}{\gamma}\big\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\big\| \times \|\Theta\|. \tag{A.5}$$

To ease notation, we use $\mathbf{u}_{\mathcal{S}}$ to denote the projection operator $\Pi_{\mathcal{S}}(\mathbf{u})$ defined in (**??**). From the decomposability of $\mathcal{R}$, we have

$$\begin{aligned}\mathcal{R}(\boldsymbol{\beta}^* + \Theta) - \mathcal{R}(\boldsymbol{\beta}^*) &\geq \mathcal{R}(\boldsymbol{\beta}^* + \Theta_{\overline{\mathcal{S}}^\perp}) - \mathcal{R}(\Theta_{\overline{\mathcal{S}}}) - \mathcal{R}(\boldsymbol{\beta}^*) \\ &= \mathcal{R}(\Theta_{\overline{\mathcal{S}}^\perp}) - \mathcal{R}(\Theta_{\overline{\mathcal{S}}}),\end{aligned}$$

where the inequality is from the triangle inequality and the equality is from decomposability of $\mathcal{R}$. Plugging the above result back into (A.5) yields that

$$\lambda_m \cdot \big(\mathcal{R}(\Theta_{\overline{\mathcal{S}}^\perp}) - \mathcal{R}(\Theta_{\overline{\mathcal{S}}})\big) \leq \Delta_m \mathcal{R}(\Theta) + \frac{\alpha\mu\tau}{\gamma}\big\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\big\| \times \|\Theta\|.$$

By choosing λ_m so that it satisfies the following condition

$$\lambda_m \geq 3\Delta_m + \frac{\alpha\mu\tau}{\gamma\Psi(\overline{\mathcal{S}})}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|, \tag{A.6}$$

we have that

$$\mathcal{R}(\Theta_{\overline{\mathcal{S}}^\perp}) - \mathcal{R}(\Theta_{\overline{\mathcal{S}}}) \leq \frac{\Delta_m}{\lambda_m}\mathcal{R}(\Theta) + \frac{\alpha\mu\tau\big\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\big\|}{\gamma\lambda_m}\|\Theta\| \leq \frac{1}{3}\mathcal{R}(\Theta) + \Psi(\overline{\mathcal{S}})\|\Theta\|.$$

Plugging $\mathcal{R}(\Theta) \leq \mathcal{R}(\Theta_{\overline{\mathcal{S}}}) + \mathcal{R}(\Theta_{\overline{\mathcal{S}}^\perp})$ into the above inequality, we obtain

$$2\mathcal{R}(\Theta_{\overline{\mathcal{S}}^\perp}) \leq 4\mathcal{R}(\Theta_{\overline{\mathcal{S}}}) + 3\Psi(\overline{\mathcal{S}}) \cdot \|\Theta\|. \tag{A.7}$$

Therefore, we have shown that Θ lies in the quasi cone $\mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \mathcal{R})$ defined in (4.3). Recall that Condition 4 states that for any fixed $\boldsymbol{\beta} \in \mathcal{B}(r;\boldsymbol{\beta}^*)$, $Q_m(\cdot|\boldsymbol{\beta})$ is strongly concave over the set $\Omega \bigcap \big(\{\boldsymbol{\beta}^*\} + \mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \mathcal{R})\big)$. Using this condition yields that

$$\begin{aligned}Q_m(\boldsymbol{\beta}^* + \Theta|\boldsymbol{\beta}) - Q_m(\boldsymbol{\beta}^*|\boldsymbol{\beta}) &\leq \big\langle \nabla Q_m(\boldsymbol{\beta}^*|\boldsymbol{\beta}), \Theta\big\rangle - \frac{\gamma_m}{2}\|\Theta\|^2 \\ &\leq \Delta_m \mathcal{R}(\Theta) + \frac{\alpha\mu\tau}{\gamma}\big\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\big\| \times \|\Theta\| - \frac{\gamma_m}{2}\|\Theta\|^2,\end{aligned} \tag{A.8}$$

where the second inequality follows from (A.4).

Now we turn back to optimality condition (A.2), following which we have

$$Q_m(\boldsymbol{\beta}^* + \Theta|\boldsymbol{\beta}) - Q_m(\boldsymbol{\beta}^*|\boldsymbol{\beta}) \geq \lambda_m \cdot \mathcal{R}(\boldsymbol{\beta}^* + \Theta) - \lambda_m \cdot \mathcal{R}(\boldsymbol{\beta}^*) \geq -\lambda_m \mathcal{R}(\Theta_{\overline{\mathcal{S}}}). \tag{A.9}$$

Putting (A.8) and (A.9) together gives us

$$\frac{\gamma_m}{2}\|\Theta\|^2 \leq \lambda_m \mathcal{R}(\Theta_{\overline{\mathcal{S}}}) + \Delta_m \mathcal{R}(\Theta) + \frac{\alpha\mu\tau}{\gamma}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \times \|\Theta\|.$$

Using $\mathcal{R}(\Theta) \leq \mathcal{R}(\Theta_{\overline{\mathcal{S}}^\perp}) + \mathcal{R}(\Theta_{\overline{\mathcal{S}}}) \leq (9/2)\Psi(\overline{\mathcal{S}})\|\Theta\|$, we further have

$$\frac{\gamma_m}{2}\|\Theta\|^2 \leq \lambda_m \Psi(\overline{\mathcal{S}})\|\Theta\| + \frac{9}{2}\Delta_m \Psi(\overline{\mathcal{S}})\|\Theta\| + \frac{\alpha\mu\tau}{\gamma}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \times \|\Theta\|.$$

Canceling the term $\|\Theta\|$ on both sides of the above inequality yields that

$$\|\Theta\| \leq 2\Psi(\overline{\mathcal{S}})\frac{\lambda_m}{\gamma_m} + \frac{\Psi(\overline{\mathcal{S}})}{\gamma_m}\left(9\Delta_m + 2\frac{\alpha\mu\tau}{\gamma\Psi(\overline{\mathcal{S}})}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|\right) \leq 5\Psi(\overline{\mathcal{S}})\frac{\lambda_m}{\gamma_m}. \tag{A.10}$$

The last inequality follows from our assumption (A.6). Putting (A.6) and (A.10) together, we reach the conclusion that if $\boldsymbol{\beta}^{(t-1)} \in \mathcal{B}(r; \boldsymbol{\beta}^*)$ and

$$\lambda_m^{(t)} \geq 3\Delta_m + \frac{\alpha\mu\tau}{\gamma\Psi(\overline{\mathcal{S}})}\|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|, \tag{A.11}$$

then we have

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\| \leq 5\Psi(\overline{\mathcal{S}})\frac{\lambda_m^{(t)}}{\gamma_m}. \tag{A.12}$$

As in the statement of the theorem, let $\kappa^* := 5\frac{\alpha\mu\tau}{\gamma\gamma_m}$ and assume $\kappa^* \leq 3/4$. Then for any $\kappa \in [\kappa^*, 3/4]$, $\Delta \geq 3\Delta_m$, we can set

$$\lambda_m^{(t)} = \frac{1-\kappa^t}{1-\kappa}\Delta + \kappa^t\frac{\gamma_m}{5\Psi(\overline{\mathcal{S}})}\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\| \tag{A.13}$$

for all $t \in [T]$. When $t = 1$, we have $\boldsymbol{\beta}^{(0)} \in \mathcal{B}(r; \boldsymbol{\beta}^*)$ and one can check inequality (A.11) holds by setting $t = 1$ in (A.13), thereby applying (A.12) yields that

$$\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^*\| \leq 5\Psi(\overline{\mathcal{S}})\frac{\lambda_m^{(1)}}{\gamma_m} = \frac{5\Psi(\overline{\mathcal{S}})}{\gamma_m}\frac{1-\kappa}{1-\kappa}\Delta + \kappa\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|.$$

Now we prove Theorem 1 by induction. Assume that for some $t \geq 1$,

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\| \leq \frac{5\Psi(\overline{\mathcal{S}})}{\gamma_m}\frac{1-\kappa^t}{1-\kappa}\Delta + \kappa^t\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|. \tag{A.14}$$

Under condition $\Delta \leq 3\overline{\Delta}, \kappa \leq 3/4$, we have

$$\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\| \leq \frac{15\Psi(\overline{\mathcal{S}})}{\gamma_m}\frac{1-(3/4)^t}{1-3/4}\overline{\Delta} + (3/4)^t\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\| \leq \frac{15\Psi(\overline{\mathcal{S}})}{\gamma_m}\frac{1-(3/4)^t}{1-3/4}\overline{\Delta} + (3/4)^t \cdot r$$
$$= (1-(3/4)^t)\cdot r + (3/4)^t \cdot r = r,$$

where the first equality is from our definition of $\overline{\Delta}$. Consequently, we have $\boldsymbol{\beta}^{(t)} \in \mathcal{B}(r; \boldsymbol{\beta}^*)$. Now we check that by our choice of $\lambda_m^{(t+1)}$, inequality (A.11) holds. Note that

$$3\Delta_m + \frac{\alpha\mu\tau}{\gamma\Psi(\overline{\mathcal{S}})}\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\| \leq \Delta + \frac{5\alpha\mu\tau}{\gamma\gamma_m}\frac{1-\kappa^t}{1-\kappa}\Delta + \frac{\alpha\mu\tau}{\gamma\Psi(\overline{\mathcal{S}})}\kappa^t\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|$$
$$\leq \Delta + \kappa\frac{1-\kappa^t}{1-\kappa}\Delta + \kappa^{t+1}\frac{\gamma_m}{5\Psi(\overline{\mathcal{S}})}\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\| = \frac{1-\kappa^{t+1}}{1-\kappa}\Delta + \kappa^{t+1}\frac{\gamma_m}{5\Psi(\overline{\mathcal{S}})}\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\| = \lambda_m^{(t+1)},$$

where the first inequality is from (A.14) and the second inequality is from the fact $\kappa \geq \kappa^* = 5\frac{\alpha\mu\tau}{\gamma\gamma_m}$. Therefore (A.11) holds for $t+1$. Then applying (A.12) with $t+1$ implies that

$$\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^*\| \leq \frac{5\Psi(\overline{\mathcal{S}})}{\gamma_m}\frac{1-\kappa^{t+1}}{1-\kappa}\Delta + \kappa^{t+1}\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|.$$

Putting pieces together we prove that (A.14) holds for all $t \in [T]$ when Conditions 4 and 5 hold in every step. Applying probabilistic union bound, we reach the conclusion. □

B Applications to Example Models

We fill in the details for the example models discussed in Section 5 in the main body: Gaussian mixture models, mixed linear regression (with sparse and low-rank regressors) and missing covariate regression.

B.1 Gaussian Mixture Model

Recall that we consider the isotropic, balanced Gaussian Mixture Model with two components where sample $\mathbf{y}_i$ is generated from either $\mathcal{N}(\boldsymbol{\beta}^*, \sigma^2\mathbf{I}_p)$ or $\mathcal{N}(-\boldsymbol{\beta}^*, \sigma^2\mathbf{I}_p)$.

We focus on the high SNR regime where we assume $\mathrm{SNR} \geq \rho$ for some constant ρ. Note that the work in [11] provides empirical and theoretical evidence that in the low SNR regime, where the overlap density of two Gaussian clusters is small, the standard EM algorithm suffers from sublinear convergence asymptotically. Therefore the high SNR condition is necessary for showing exponential/linear convergence of the EM algorithm and our high dimensional variant. In particular, we are interested in quantizing estimation error using ℓ_2 norm. We thus set the norm $\|\cdot\|$ in our framework to be $\|\cdot\|_2$ in this section. Recall that we set regularizer $\mathcal{R}$ to be the ℓ_1 norm. For any subset $\mathcal{S} \subseteq \{1, \ldots, p\}$, ℓ_1 norm is decomposable with respect to $(\mathcal{S}, \overline{\mathcal{S}})$. For any $\boldsymbol{\beta}^* \in \mathcal{B}_0(s; p)$, by letting $\mathcal{S} = \mathrm{supp}(\boldsymbol{\beta}^*), \overline{\mathcal{S}} = \mathrm{supp}(\boldsymbol{\beta}^*)$, we have $\Psi(\mathcal{S}) = \sqrt{s}$ and $\mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \mathcal{R})$ corresponds to $\{\|\mathbf{u}_{\mathcal{S}^\perp}\|_1 \leq 2\|\mathbf{u}_{\mathcal{S}}\|_1 + 2\sqrt{s}\|\mathbf{u}\|_2\}$.

According to the $Q_n^{GMM}(\cdot|\cdot)$ introduced in (5.1), by taking its expectation, we have

$$Q^{GMM}(\boldsymbol{\beta}'|\boldsymbol{\beta}) = -\frac{1}{2}\mathbb{E}\left[w(Y;\boldsymbol{\beta})\|Y - \boldsymbol{\beta}'\|_2^2 + (1 - w(Y;\boldsymbol{\beta}))\|Y + \boldsymbol{\beta}'\|_2^2\right]. \tag{B.1}$$

We now check that Conditions 1-3 hold for $Q^{GMM}(\cdot|\cdot)$. We begin with proving the following result.

Lemma 2 (Self consistency of GMM)**.** *Consider the Gaussian mixture model with $Q^{GMM}(\cdot|\cdot)$ given in* (B.1)*. For model parameter $\boldsymbol{\beta}^*$ we have*

$$\boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} Q^{GMM}(\boldsymbol{\beta}|\boldsymbol{\beta}^*).$$

Proof. In this example, we have

$$\mathcal{M}(\boldsymbol{\beta}^*) = 2\mathbb{E}\left[w(Y;\boldsymbol{\beta}^*)Y\right] = 2\mathbb{E}\left[\frac{1}{1 + \exp(-\frac{2}{\sigma^2}\langle Z \cdot \boldsymbol{\beta}^* + W, \boldsymbol{\beta}^*\rangle)}(Z \cdot \boldsymbol{\beta}^* + W)\right],$$

where $W \sim \mathcal{N}(0, \sigma^2)$ and Z has Rademacher distribution over $\{-1, 1\}$. Due to the rotation invariance of Gaussianity, without loss of generality, we assume $\boldsymbol{\beta}^* = A\boldsymbol{e}_1$. It is easy to check $\mathrm{supp}(\mathcal{M}(\boldsymbol{\beta}^*)) = \{1\}$. Moreover, the first coordinate of $\mathcal{M}(\boldsymbol{\beta}^*)$ takes form

$$(\mathcal{M}(\boldsymbol{\beta}^*))_1 = 2\mathbb{E}\left[\frac{1}{1 + \exp(-\frac{2}{\sigma^2}(AZ + W_1))}(AZ + W_1)\right] = A,$$

where the last equality follows by the substitution $X = W_1, Z = Z, \gamma = 0, a = A$ in Lemma 25. Therefore, $\mathcal{M}(\boldsymbol{\beta}^*) = \boldsymbol{\beta}^*$. □

The above result shows that $Q^{GMM}(\cdot|\cdot)$ satisfies Condition 1. It is easy to see $\nabla^2 Q^{GMM}(\boldsymbol{\beta}'|\boldsymbol{\beta}) = -\mathbf{I}_p$, which implies that $Q^{GMM}(\cdot|\cdot)$ satisfies Condition 2 with parameters $(\gamma, \mu, r) = (1, 1, r)$ for any $r > 0$. Next we present a result showing that $Q^{GMM}(\cdot|\cdot)$ satisfies Condition 3 with arbitrarily small stability factor τ when SNR is sufficiently large.

Lemma 3 (Gradient stability of GMM)**.** *Consider the Gaussian Mixture Model with $Q^{GMM}(\cdot|\cdot)$ given in* (B.1)*. Suppose $\mathrm{SNR} > \rho$. Function $Q^{GMM}(\cdot|\cdot)$ satisfies Condition 3 with parameters $(\tau, \|\boldsymbol{\beta}^*\|_2/4)$, where $\tau \leq \exp(-C\rho^2)$ for some absolute constant C.*

Proof. See the proof of Lemma 3 in [1]. □

Now we turn to the conditions on $Q_n^{GMM}(\cdot|\cdot)$.

Lemma 4 (RSC of GMM)**.** *Consider the Gaussian mixture model with any $\boldsymbol{\beta}^* \in \mathcal{B}_0(s; p)$ and $Q_n^{GMM}(\cdot|\cdot)$ given in* (5.1)*. For any $r > 0$, we have $Q_n^{GMM}(\cdot|\cdot)$ satisfies Condition 4 with parameters $(\gamma_n, \mathcal{S}, \overline{\mathcal{S}}, r, \delta)$, where*

$$\gamma_n = 1,\ \delta = 0,\ (\mathcal{S}, \overline{\mathcal{S}}) = (\mathit{supp}(\boldsymbol{\beta}^*), \mathit{supp}(\boldsymbol{\beta}^*)).$$

Proof. Although Condition 4 is a stochastic condition, for Gaussian mixture model in particular, it is satisfied deterministically. Note that

$$Q_n^{GMM}(\boldsymbol{\beta}'|\boldsymbol{\beta}) = -\frac{1}{2n}\sum_{i=1}^{n}\left[w(\mathbf{y}_i;\boldsymbol{\beta})\|\mathbf{y}_i-\boldsymbol{\beta}'\|_2^2 + (1-w(\mathbf{y}_i;\boldsymbol{\beta}))\|\mathbf{y}_i+\boldsymbol{\beta}'\|_2^2\right].$$

We have that for any $\boldsymbol{\beta}', \boldsymbol{\beta} \in \mathbb{R}^p$, $\nabla^2 Q_n^{GMM}(\boldsymbol{\beta}'|\boldsymbol{\beta}) = -\mathbf{I}_p$, which implies that $Q_n^{GMM}(\boldsymbol{\beta}'|\boldsymbol{\beta})$ is strongly concave with parameter 1. Consequently, Condition 4 holds with $\gamma_n = 1$. □

This above result indicates that the restricted strong concavity condition holds deterministically in this example. The next lemma validates the statistical error condition and provides the corresponding parameters.

Lemma 5 (Statistical error of GMM)**.** *Consider the Gaussian mixture model with $Q_n^{GMM}(\cdot|\cdot)$ and $Q^{GMM}(\cdot|\cdot)$ given in* (5.1) *and* (B.1) *respectively. For any $r > 0$, $\delta \in (0,1)$ and some absolute constant C, Condition 5 holds with parameters (Δ_n, r, δ) where*

$$\Delta_n = C(\|\boldsymbol{\beta}^*\|_\infty + \sigma)\sqrt{\frac{\log p + \log(2e/\delta)}{n}}.$$

Proof. Note that $\mathcal{R}^*$ is $\|\cdot\|_\infty$ in this example. Following the specific formulations of $Q_n^{GMM}(\cdot|\cdot)$ and $Q^{GMM}(\cdot|\cdot)$ in (5.1) and (B.1), we have

$$\nabla Q_n^{GMM}(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - \nabla Q^{GMM}(\boldsymbol{\beta}^*|\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i + \frac{2}{n}\sum_{i=1}^{n} w(\mathbf{y}_i;\boldsymbol{\beta})\mathbf{y}_i - 2\mathbb{E}\left[w(Y;\boldsymbol{\beta})Y\right].$$

Therefore,

$$\left\|\nabla Q_n^{GMM}(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - \nabla Q^{GMM}(\boldsymbol{\beta}^*|\boldsymbol{\beta})\right\|_\infty \le \underbrace{\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i\right\|_\infty}_{(a)} + \underbrace{\left\|\frac{2}{n}\sum_{i=1}^{n} w(\mathbf{y}_i;\boldsymbol{\beta})\mathbf{y}_i - 2\mathbb{E}\left[w(Y;\boldsymbol{\beta})Y\right]\right\|_\infty}_{(b)}$$

Next we bound the two terms (a) and (b) respectively.

Term (a)**.** Let $\boldsymbol{\zeta} := \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i$. Let $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,p})^\top$ for all $i \in [n]$. Consider the j-th coordinate ζ_j of $\boldsymbol{\zeta}$, we have

$$\zeta_j = \frac{1}{n}\sum_{i=1}^{n} y_{i,j}.$$

Note that $\{y_{i,j}\}_{i=1}^n$ are independent copies of random variable Y_j that is

$$Y_j = Z \cdot \beta_j^* + V, \tag{B.2}$$

where Z is Rademacher random variable taking values in $\{-1, 1\}$ and V has distribution $\mathcal{N}(0, \sigma^2)$. Since $Z \cdot \beta_j^*$ and V are both sub-Gaussian random variables with norm $\|Z \cdot \beta_j^*\|_{\psi_2} \le |\beta_j^*|$ and $\|V\|_{\psi_2} \lesssim \delta$. Following the rotation invariance sub-Gaussian random variables (e.g., Lemma 5.9 in [18]), we have that

$$\|Y_j\|_{\psi_2} \lesssim \sqrt{\|Z \cdot \beta_j^*\|_{\psi_2}^2 + \|V\|_{\psi_2}^2} \lesssim \sqrt{\|\boldsymbol{\beta}^*\|_\infty^2 + \sigma^2}.$$

Following the standard sub-Gaussian concentration argument in Lemma 19, there exists some constant C such that for any $j \in [p]$ and all $t \ge 0$,

$$\Pr\left(\left|\zeta_j\right| \ge t\right) \le e \cdot \exp\left(-\frac{Cnt^2}{\|\boldsymbol{\beta}^*\|_\infty^2 + \sigma^2}\right).$$

Then by applying union bound, we have

$$\Pr\left(\sup_{j\in[p]}\left|\zeta_j\right| \ge t\right) \le pe \cdot \exp\left(-\frac{Cnt^2}{\|\boldsymbol{\beta}^*\|_\infty^2 + \sigma^2}\right).$$

Setting the right hand side to be δ, we have that, with probability at least $1-\delta/2$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i\right\|_\infty \lesssim (\|\boldsymbol{\beta}^*\|_\infty+\delta)\sqrt{\frac{\log p+\log(2e/\delta)}{n}}. \tag{B.3}$$

Term (b)**.** Now let $\boldsymbol{\zeta} := \frac{2}{n}\sum_{i=1}^{n} w(\mathbf{y}_i;\boldsymbol{\beta})\mathbf{y}_i - 2\mathbb{E}\left[w(Y;\boldsymbol{\beta})Y\right]$. We also consider the j-th coordinate ζ_j of $\boldsymbol{\zeta}$, which takes form

$$\zeta_j = \frac{2}{n}\sum_{i=1}^{n}\Big\{w(\mathbf{y}_i;\boldsymbol{\beta})y_{i,j} - \mathbb{E}(w(Y;\boldsymbol{\beta})Y_j)\Big\}.$$

Note that $w(\mathbf{y}_i;\boldsymbol{\beta})y_{i,j} - \mathbb{E}(w(Y;\boldsymbol{\beta})Y_j), i=1,\ldots,n$ are independent copies of random variable $w(Y;\boldsymbol{\beta})Y_j - \mathbb{E}(w(Y;\boldsymbol{\beta})Y_j)$ where Y_j is given in (B.2). We have shown that Y_j is sub-Gaussian random variable. Note that $w(Y;\boldsymbol{\beta})$ is random variable taking values in $[0,1]$. We thus always have

$$\Pr\left(|w(Y;\boldsymbol{\beta})Y_j|\geq t\right) \leq \Pr(|Y_j|>t) \leq \exp(1-Ct^2/\|Y_j\|_{\psi_2}^2).$$

Using the equivalent properties of sub-Gaussian (see Lemma 5.5 in [18]) , we conclude that $w(Y;\boldsymbol{\beta})Y_j$ is sub-Gaussian random variable with norm $\|w(Y;\boldsymbol{\beta})Y_j\|_{\psi_2} \leq \|Y_j\|_{\psi_2} \lesssim \sqrt{\|\boldsymbol{\beta}^*\|_\infty^2+\sigma^2}$. Following Lemma 21, we have $\|w(Y;\boldsymbol{\beta})Y_j - \mathbb{E}\left[w(Y;\boldsymbol{\beta})Y_j\right]\|_{\psi_2} \leq 2\|w(Y;\boldsymbol{\beta})Y_j\|_{\psi_2}$. Using the concentration result from Lemma 19 yields that for any $j\in[p]$ and some constant C,

$$\Pr\left(|\zeta_j|\geq t\right) = \Pr\left\{\left|\frac{2}{n}\sum_{i=1}^{n} w(\mathbf{y}_i;\boldsymbol{\beta})y_{i,j} - \mathbb{E}(w(Y;\boldsymbol{\beta})Y)\right| > t\right\} \leq e\cdot\exp\left(-\frac{Cnt^2}{\|\boldsymbol{\beta}^*\|_\infty^2+\sigma^2}\right).$$

Applying union bound over p coordinates, we have

$$\Pr\left(\sup_{j\in[p]}|\zeta_j|>t\right) \leq pe\cdot\exp\left(-\frac{Cnt^2}{\|\boldsymbol{\beta}^*\|_\infty^2+\sigma^2}\right),$$

which implies that, with probability at least $1-\delta/2$,

$$\left\|\frac{2}{n}\sum_{i=1}^{n} w(\mathbf{y}_i;\boldsymbol{\beta})\mathbf{y}_i - 2\mathbb{E}\left[w(Y;\boldsymbol{\beta})Y\right]\right\|_\infty \lesssim (\|\boldsymbol{\beta}^*\|_\infty+\sigma)\sqrt{\frac{\log p+\log(2e/\delta)}{n}}. \tag{B.4}$$

Putting (B.3) and (B.4) together completes the proof. □

Now we give the guarantees of Algorithm 2 for the Gaussian mixture model.

Proof of Corollary 1. This result follows from Theorem 1. First, recall that the minimum contractive factor κ^* is $\kappa^* = 5\frac{\alpha\mu\tau}{\gamma\gamma_{n/T}}$. For the ℓ_2 norm, we have $\alpha=1$. Following the fact that $(\gamma,\mu)=(1,1)$ and Lemma 3-4, we have $\kappa^* \leq 20\exp(-C\rho^2)$ for some constant C. We further have $\kappa^*\leq\frac{1}{2}$ when ρ is sufficiently large. Second, based on Lemma 5, we set $\delta=1/p$ and choose Δ as $\Delta = C(\|\boldsymbol{\beta}^*\|_\infty+\sigma)\sqrt{T\log p/n}$ with sufficiently large C such that $\Delta\geq 3\Delta_{n/T}$. By the assumption on n/T, we have that $\Delta\leq 3\overline{\Delta}$ where $\overline{\Delta} = \|\boldsymbol{\beta}^*\|_2/(240\sqrt{s})$ in this example. Finally, we choose $\lambda_{n/T}^{(0)} = \|\boldsymbol{\beta}^{(0)}-\boldsymbol{\beta}^*\|/(5\sqrt{s})$ by following Theorem 1. Packing up these ingredients and following Theorem 1, we have that by choosing any $\kappa\in[1/2,3/4]$, $\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^*\|_2 \leq \kappa^t\|\boldsymbol{\beta}^{(0)}-\boldsymbol{\beta}^*\|_2 + 5\sqrt{s}\Delta/(1-\kappa)$, which thus completes the proof. □

B.2 Mixed Linear Regression

Recall that for Mixed Linear Regression (MLR) model, we consider two sets of model parameters: $\boldsymbol{\beta}^*\in\mathcal{B}_0(s;p)$ and $\boldsymbol{\Gamma}^*\in\mathbb{R}^{p_1\times p_2}$ with $\mathrm{rank}(\boldsymbol{\Gamma}^*)=\theta$. For the two settings, the population level analysis is identical under i.i.d. Gaussian covariate design. Without loss of generality, we begin with treating the model parameter as a vector $\boldsymbol{\beta}^*\in\mathbb{R}^p$ and validate Conditions 1-3 for $Q^{MLR}(\cdot|\cdot)$ in this example. Given function $Q_n^{MLR}(\cdot|\cdot)$ in (5.3), taking its expectation, yields

$$Q^{MLR}(\boldsymbol{\beta}'|\boldsymbol{\beta}) = -\frac{1}{2}\mathbb{E}\left[w(Y,X;\boldsymbol{\beta})(Y-\langle X,\boldsymbol{\beta}'\rangle)^2 + (1-w(Y,X;\boldsymbol{\beta}))(Y+\langle X,\boldsymbol{\beta}'\rangle)^2\right]. \tag{B.5}$$

For now, we set the norm $\|\cdot\|$ in our framework to $\|\cdot\|_2$. We begin by checking the self consistency condition.

Lemma 6 (Self consistency of MLR)**.** *Consider mixed linear regression with model parameter $\boldsymbol{\beta}^* \in \mathbb{R}^p$ and $Q^{MLR}(\cdot|\cdot)$ given in* (B.5)*. We have*

$$\boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} Q^{MLR}(\boldsymbol{\beta}|\boldsymbol{\beta}^*).$$

Proof. In this example, we have

$$\mathcal{M}(\boldsymbol{\beta}^*) = 2\mathbb{E}\left[w(Y,X;\boldsymbol{\beta}^*)YX\right] = 2\mathbb{E}\left[\frac{1}{1+\exp(-\frac{2(\langle X, Z\cdot\boldsymbol{\beta}^*\rangle + W)\langle X,\boldsymbol{\beta}^*\rangle}{\sigma^2})}(Z\cdot\boldsymbol{\beta}^* + W)X\right],$$

where $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p), W \sim \mathcal{N}(0,\sigma^2)$, Z has Rademacher distribution. Due to the rotation invariance of Gaussianity, without loss of generality, we can assume $\boldsymbol{\beta}^* = A\boldsymbol{e}_1$. It is easy to check $\mathrm{supp}(\mathcal{M}(\boldsymbol{\beta}^*)) = \{1\}$. Moreover,

$$(\mathcal{M}(\boldsymbol{\beta}^*))_1 = 2\mathbb{E}\left[\frac{1}{1+\exp(-\frac{2}{\sigma^2}(AZX_1+W)AX_1)}(AZX_1^2 + X_1W)\right] = \mathbb{E}(AX_1^2) = A,$$

where the second inequality follows by the substitution $X = W, Z = Z, \gamma = 0, a = AX_1$ in Lemma 25. We thus have $\mathcal{M}(\boldsymbol{\beta}^*) = \boldsymbol{\beta}^*$. □

It is easy to check $\nabla^2 Q^{MLR}(\boldsymbol{\beta}'|\boldsymbol{\beta}) = -\mathbf{I}_p$. Therefore, $Q^{MLR}(\cdot|\cdot)$ satisfies Condition 2 with parameters $(\gamma,\mu,r) = (1,1,r)$ for any $r > 0$. Similar to the Gaussian mixture model, we introduce the following SNR quantity to characterize the difficulty of the problem.

$$\mathrm{SNR} := \|\boldsymbol{\beta}^*\|/\sigma.$$

The work in [7] shows that there exists an unavoidable phase transition of statistical rate from high SNR to low SNR. In detail, in low-dimensional setting, the obtainable statistical error is $\Omega(\sqrt{p/n})$ that matches the standard linear regression when $\mathrm{SNR} \geq \rho$ for some constant ρ. Meanwhile, the unavoidable rate becomes $\Omega((p/n)^{1/4})$ when $\mathrm{SNR} \ll \rho$. We conjecture such transition phenomenon still exists in high dimensional setting. For now we focus on the high SNR regime and show our algorithm achieves statistical rate that matches the standard sparse linear regression and low rank matrix recovery (up to logarithmic factor) in the end.

The following result shows Condition 3 holds with arbitrarily small stability factor τ when SNR is sufficiently large and the radius r of ball $\mathcal{B}(r;\boldsymbol{\beta}^*)$ is sufficiently small.

Lemma 7 (Gradient Stability of MLR)**.** *Consider mixed linear regression model with function $Q^{MLR}(\cdot|\cdot)$ given in* (B.5)*. For any $\omega \in [0,1/4]$, let $r = \omega\|\boldsymbol{\beta}^*\|_2$. Suppose $\mathrm{SNR} \geq \rho$ for some constant ρ. Then for any $\boldsymbol{\beta} \in \mathcal{B}(r;\boldsymbol{\beta}^*)$, we have*

$$\|\nabla Q^{MLR}(\mathcal{M}(\boldsymbol{\beta})|\boldsymbol{\beta}) - \nabla Q^{MLR}(\mathcal{M}(\boldsymbol{\beta})|\boldsymbol{\beta}^*)\|_2 \leq \tau\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2$$

with

$$\tau = \frac{17}{\rho} + 7.3\omega.$$

Proof. Recall that we hope to find τ such that for any $\boldsymbol{\beta} \in \mathcal{B}(r;\boldsymbol{\beta}^*)$

$$\|\nabla Q^{MLR}(\mathcal{M}(\boldsymbol{\beta})|\boldsymbol{\beta}) - \nabla Q^{MLR}(\mathcal{M}(\boldsymbol{\beta})|\boldsymbol{\beta}^*)\|_2 \leq \tau\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2.$$

In this example, we have

$$\mathcal{M}(\boldsymbol{\beta}) = 2\mathbb{E}\left[w(Y,X;\boldsymbol{\beta})YX\right],$$

and

$$\nabla Q^{MLR}(\boldsymbol{\beta}'|\boldsymbol{\beta}) = 2\mathbb{E}\left[w(Y,X;\boldsymbol{\beta})YX\right] - \boldsymbol{\beta}'.$$

Therefore,

$$\begin{aligned}&\nabla Q^{MLR}(\mathcal{M}(\boldsymbol{\beta})|\boldsymbol{\beta}) - \nabla Q^{MLR}(\mathcal{M}(\boldsymbol{\beta})|\boldsymbol{\beta}^*)\\ &= 2\mathbb{E}\left[w(Y,X;\boldsymbol{\beta})YX\right] - 2\mathbb{E}\left[w(Y,X;\boldsymbol{\beta}^*)YX\right] = 2\mathbb{E}\left[w(Y,X;\boldsymbol{\beta})YX\right] - \boldsymbol{\beta}^*,\end{aligned}$$

where the last equality is from the self consistent property of $Q^{MLR}(\cdot|\cdot)$. Due to the rotation invariance of Gaussianity, without loss of generality, we assume $\boldsymbol{\beta}^* = A\boldsymbol{e}_1, \boldsymbol{\beta} = (1+\epsilon_1)A\boldsymbol{e}_1 + \epsilon_2 A\boldsymbol{e}_2$, where $A = \|\boldsymbol{\beta}^*\|_2$, $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 = A\sqrt{\epsilon_1^2 + \epsilon_2^2}$. Let random vector T be

$$T := w(Y, X; \boldsymbol{\beta})YX - \frac{1}{2}\boldsymbol{\beta}^*.$$

Note that for any $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$w(Y, X; \boldsymbol{\beta}) = \frac{\exp(-\frac{(Y-\langle X,\boldsymbol{\beta}\rangle)^2}{2\sigma^2})}{\exp(-\frac{(Y-\langle X,\boldsymbol{\beta}\rangle)^2}{2\sigma^2}) + \exp(-\frac{(Y+\langle X,\boldsymbol{\beta}\rangle)^2}{2\sigma^2})} = \frac{1}{1+\exp(-\frac{2Y\langle X,\boldsymbol{\beta}\rangle}{\sigma^2})},$$

thereby

$$\begin{aligned} T &= \frac{1}{1+\exp(-\frac{2Y\langle X,\boldsymbol{\beta}\rangle}{\sigma^2})}YX - \frac{1}{2}\boldsymbol{\beta}^* \\ &= \frac{1}{1+\exp(-\frac{2(ZAX_1+W)(A(1+\epsilon_1)X_1+\epsilon_2X_2)}{\sigma^2})}(ZAX_1+W)X - \frac{1}{2}A\boldsymbol{e}_1, \end{aligned}$$

where Z is Rademacher random variable taking values in $\{-1, 1\}$, W is stochastic noise with distribution $\mathcal{N}(0, \sigma^2)$, X_1 and X_2 are the first two coordinates of X. It is easy to note that $\mathbb{E}[T_i] = 0$ for $i = 3, \ldots, p$. We focus on characterizing the first two coordinates T_1, T_2 of T.

Coordinate T_1.
First, we compute the expectation of T_1. Particularly we let $\gamma = \epsilon_1 + \epsilon_2 X_2/X_1$. Then we have

$$\begin{aligned} \big|\mathbb{E}[T_1]\big| &= \left|\mathbb{E}\left[\frac{X_1(W+ZAX_1)}{1+\exp(-\frac{2AX_1(1+\gamma)}{\sigma^2}(W+ZAX_1))} - \frac{1}{2}AX_1^2\right]\right| \\ &\le \mathbb{E}\left[|X_1| \cdot \left|\frac{(W+ZAX_1)}{1+\exp(-\frac{2AX_1(1+\gamma)}{\sigma^2}(W+ZAX_1))} - \frac{1}{2}AX_1\right|\right] \\ &= \mathbb{E}_{X_1,X_2}\left\{|X_1| \cdot \mathbb{E}_{W,Z}\left[\left|\frac{(W+ZAX_1)}{1+\exp(-\frac{2AX_1(1+\gamma)}{\sigma^2}(W+ZAX_1))} - \frac{1}{2}AX_1\right|\right]\right\} \\ &\le \mathbb{E}_{X_1,X_2}\left[|X_1| \cdot \min\left\{\frac{1}{2}A \cdot |X_1\gamma| \cdot \exp(\frac{\gamma^2(AX_1)^2 - (AX_1)^2}{2\sigma^2}),\ \frac{\sigma}{\sqrt{2\pi}} + A|X_1|\right\}\right], \end{aligned} \tag{B.6}$$

where the last inequality follows from Lemma 25 by replacing the parameters (X, Z, a, γ) in the statement with (W, Z, AX_1, γ). Let event $\mathcal{E}$ be $\mathcal{E} := \{\gamma^2 \le 0.9\}$. Computing the expectation in (B.6) conditioning on $\mathcal{E}$ and $\mathcal{E}^c$ yields that

$$\begin{aligned} \big|\mathbb{E}[T_1]\big| \le& \mathbb{E}\left[\frac{1}{2}|\gamma|AX_1^2\exp(\frac{\gamma^2(AX_1)^2 - (AX_1)^2}{2\sigma^2})\ \middle|\ \mathcal{E}\right] \cdot \Pr(\mathcal{E}) \\ &+ \mathbb{E}\left[\frac{\sigma|X_1|}{\sqrt{2\pi}} + AX_1^2\ \middle|\ \mathcal{E}^c\right] \cdot \Pr(\mathcal{E}^c). \end{aligned} \tag{B.7}$$

We bound the two terms on the right hand side of the above inequality respectively. For the first term we have

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{2}|\gamma|AX_1^2\exp(\frac{\gamma^2(AX_1)^2 - (AX_1)^2}{2\sigma^2})\ \middle|\ \mathcal{E}\right] \cdot \Pr(\mathcal{E}) \le \mathbb{E}\left[\frac{1}{2}|\gamma|AX_1^2\exp(\frac{-(AX_1)^2}{20\sigma^2})\ \middle|\ \mathcal{E}\right] \cdot \Pr(\mathcal{E}) \\ &\le \mathbb{E}\left[\frac{1}{2}|\gamma|AX_1^2\exp(\frac{-(AX_1)^2}{20\sigma^2})\right] \le \mathbb{E}\left[\frac{1}{2}A\left(|\epsilon_1| \cdot X_1^2 + |\epsilon_2X_1X_2|\right)\exp(-\frac{1}{20}\rho^2X_1^2)\right] \\ &= \frac{1}{2}A\frac{|\epsilon_1|}{(1+0.1\rho^2)^{3/2}} + \frac{1}{\pi}A\frac{|\epsilon_2|}{1+0.1\rho^2} \le \frac{1}{2}A\frac{1}{1+0.1\rho^2}(|\epsilon_1| + |\epsilon_2|), \end{aligned} \tag{B.8}$$

where the third inequality is from $\|\boldsymbol{\beta}^*\|_2/\sigma \ge \rho$. For the second term in (B.7), first note that

$$\sqrt{\epsilon_1^2 + \epsilon_2^2} \le \frac{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2}{\|\boldsymbol{\beta}^*\|_2} \le \omega \le 1/4,$$

thereby

$$|\gamma| \le |\epsilon_1| + |\epsilon_2| \cdot |X_2/X_1| \le 1/4 + |\epsilon_2| \cdot |X_2/X_1|.$$

We define event $\mathcal{E}' := \{X_2^2/X_1^2 \ge (2.1\epsilon_2^2)^{-1}\}$. Note that $\mathcal{E}^c = \{\gamma^2 \ge 0.9\}$, we thus have $\mathcal{E}^c \subseteq \mathcal{E}'$, i.e., the occurrence of $\mathcal{E}^c$ must lead to the occurrence of $\mathcal{E}'$. For the second term in (B.7), we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\sigma|X_1|}{\sqrt{2\pi}} + AX_1^2 \,\middle|\, \mathcal{E}^c\right] \cdot \Pr(\mathcal{E}^c) \le \mathbb{E}\left[\frac{\sigma|X_1|}{\sqrt{2\pi}} + AX_1^2 \,\middle|\, \mathcal{E}'\right] \cdot \Pr(\mathcal{E}') \\
&\le \mathbb{E}\left[\frac{\sigma|X_1|}{\sqrt{2\pi}} + \sqrt{2.1\epsilon_2^2}A|X_1X_2| \,\middle|\, \mathcal{E}'\right] \cdot \Pr(\mathcal{E}') \qquad \text{(B.9)} \\
&= \frac{\sigma}{\pi}\left[1 - \sqrt{\frac{1}{1+2.1\epsilon_2^2}}\right] + \sqrt{2.1\epsilon_2^2}A\frac{2}{\pi}\frac{2.1\epsilon_2^2}{1+2.1\epsilon_2^2} \le \frac{\sqrt{2.1}\sigma}{\pi}|\epsilon_2| + \frac{2\sqrt{2.1}^3}{\pi}A|\epsilon_2|^3, \qquad \text{(B.10)}
\end{aligned}$$

where the equality is from Lemma 24 by setting C in the statement to be $\sqrt{2.1\epsilon_2^2}$.

Putting (B.8) and (B.9) together, we have

$$|\mathbb{E}[T_1]| \le \frac{1}{2}A\frac{1}{1+0.1\rho^2}(|\epsilon_1| + |\epsilon_2|) + \frac{\sqrt{2.1}\sigma}{\pi}|\epsilon_2| + \frac{2\sqrt{2.1}^3}{\pi}A|\epsilon_2|^3. \qquad \text{(B.11)}$$

Coordinate T_2.

Now we turn to the second coordinate T_2. Using $\mathbb{E}[X_1X_2] = 0$, we have

$$\begin{aligned}
|\mathbb{E}[T_2]| &= \left|\mathbb{E}\left[\frac{X_2(W + ZAX_1)}{1+\exp(-\frac{2AX_1(1+\gamma)}{\sigma^2}(W+ZAX_1))} - \frac{1}{2}AX_1X_2\right]\right| \\
&\le \mathbb{E}\left[|X_2| \cdot \left|\frac{(W+ZAX_1)}{1+\exp(-\frac{2AX_1(1+\gamma)}{\sigma^2}(W+ZAX_1))} - \frac{1}{2}AX_1\right|\right].
\end{aligned}$$

Similar to (B.6), using Lemma 25 leads to

$$\begin{aligned}
|\mathbb{E}[T_2]| &\le \mathbb{E}\left[|X_2| \cdot \min\left\{\frac{1}{2}A \cdot |X_1\gamma| \cdot \exp(\frac{\gamma^2(AX_1)^2 - (AX_1)^2}{2\sigma^2}),\ \frac{\sigma}{\sqrt{2\pi}} + A|X_1|\right\}\right] \\
&\le \mathbb{E}\left[\frac{1}{2}A|\gamma| \cdot |X_1X_2| \exp(\frac{\gamma^2(AX_1)^2 - (AX_1)^2}{2\sigma^2}) \,\middle|\, \mathcal{E}\right] \cdot \Pr(\mathcal{E}) \\
&\qquad + \mathbb{E}\left[\frac{\sigma|X_2|}{\sqrt{2\pi}} + A|X_1X_2| \,\middle|\, \mathcal{E}^c\right] \cdot \Pr(\mathcal{E}^c).
\end{aligned}$$

We bound the two terms in the right hand side of the above inequality respectively. For the first term, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{2}A|\gamma| \cdot |X_1X_2| \exp(\frac{\gamma^2(AX_1)^2 - (AX_1)^2}{2\sigma^2}) \,\middle|\, \mathcal{E}\right] \cdot \Pr(\mathcal{E}) \\
&\le \mathbb{E}\left[\frac{1}{2}A|\gamma| \cdot |X_1X_2| \exp(\frac{-0.1(AX_1)^2}{2\sigma^2}) \mid \mathcal{E}\right] \cdot \Pr(\mathcal{E}) \le \mathbb{E}\left[\frac{1}{2}A|\gamma| \cdot |X_1X_2| \exp(\frac{-0.1(AX_1)^2}{2\sigma^2})\right] \\
&\le \mathbb{E}\left[\frac{1}{2}A\left(|\epsilon_1X_1X_2| + |\epsilon_2|X_2^2\right)\exp(-\frac{1}{20}\rho^2X_1^2)\right] = \frac{1}{\pi}A\frac{|\epsilon_1|}{1+0.1\rho^2} + \frac{1}{2}A\frac{|\epsilon_2|}{\sqrt{1+0.1\rho^2}} \qquad \text{(B.12)}
\end{aligned}$$

For the second term, recall that event $\mathcal{E}'$ is defined as $\{X_2^2/X_1^2 \ge (2.1\epsilon_2^2)^{-1}\}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\sigma|X_2|}{\sqrt{2\pi}} + A|X_1X_2| \,\middle|\, \mathcal{E}^c\right] \cdot \Pr(\mathcal{E}^c) \le \mathbb{E}\left[\frac{\sigma|X_2|}{\sqrt{2\pi}} + A|X_1X_2| \,\middle|\, \mathcal{E}'\right] \cdot \Pr(\mathcal{E}') \\
&= \frac{\sigma}{\pi}\frac{\sqrt{2.1}\epsilon_2}{\sqrt{1+2.1\epsilon_2^2}} + \frac{2A}{\pi}\frac{2.1\epsilon_2^2}{1+2.1\epsilon_2^2} \le \frac{\sqrt{2.1}\sigma}{\pi}|\epsilon_2| + \frac{4.2A}{\pi}\epsilon_2^2. \qquad \text{(B.13)}
\end{aligned}$$

where the equality follows from Lemma 24 by setting C in the statement to be $\sqrt{2.1\epsilon_2^2}$. Putting (B.12) and (B.13) together, we have

$$|\mathbb{E}[T_2]| \le \frac{1}{\pi}A\frac{|\epsilon_1|}{1+0.1\rho^2} + \frac{1}{2}A\frac{|\epsilon_2|}{\sqrt{1+0.1\rho^2}} + \frac{\sqrt{2.1}\sigma}{\pi}|\epsilon_2| + \frac{4.2A}{\pi}\epsilon_2^2. \qquad \text{(B.14)}$$

Now based on (B.11) and (B.14), we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\|T\|_2\right] = \mathbb{E}\left[\sqrt{T_1^2 + T_2^2}\right] \leq \mathbb{E}\left[|T_1| + |T_2|\right] \\
&\leq A\frac{1}{\sqrt{1+0.1\rho^2}}(|\epsilon_1| + |\epsilon_2|) + \frac{\sqrt{2.1}\sigma}{\pi}|\epsilon_2| + \frac{2\sqrt{2.1}^3}{\pi}A|\epsilon_2|^3 + \frac{\sqrt{2.1}\sigma}{\pi}|\epsilon_2| + \frac{4.2A}{\pi}\epsilon_2^2 \\
&\leq A\left(\frac{1}{\sqrt{1+0.1\rho^2}}(|\epsilon_1| + |\epsilon_2|) + |\epsilon_2|/\rho + 1.83\omega|\epsilon_2|\right) \\
&\leq A(|\epsilon_1| + |\epsilon_2|)\cdot\left(\frac{4.2}{\rho} + 1.83\omega\right) \leq 2A\sqrt{\epsilon_1^2 + \epsilon_2^2}\cdot\left(\frac{4.2}{\rho} + 1.83\omega\right) \\
&= 2\left(\frac{4.2}{\rho} + 1.83\omega\right)\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2.
\end{aligned}$$

Note that $\nabla Q^{MLR}(\mathcal{M}(\boldsymbol{\beta})|\boldsymbol{\beta}) - \nabla Q^{MLR}(\mathcal{M}(\boldsymbol{\beta})|\boldsymbol{\beta}^*) = 2T$, thereby we conclude that for any $\omega \leq 1/4$, $Q^{MLR}(\cdot|\cdot)$ satisfies gradient stability condition over $\mathcal{B}(\omega\|\boldsymbol{\beta}^*\|_2; \boldsymbol{\beta}^*)$ with parameter

$$\tau = \frac{17}{\rho} + 7.3\omega.$$

□

In [1], it is proved that when $r = \frac{1}{32}\|\boldsymbol{\beta}^*\|_2$, there exists $\tau \in [0, 1/2]$ such that $Q^{MLR}(\cdot|\cdot)$ satisfies Condition 3 with parameter τ when ρ is sufficiently large. Note that Lemma 7 recovers this result. Moreover, Lemma 7 provides an explicit function to characterize the relationship between τ and ρ, ω.

Next we turn to validate the two technical conditions of $Q_n^{MLR}(\cdot|\cdot)$ and establish the computational and statistical guarantees of estimating mixed linear parameters in the high dimensional regime. We consider two different structures of linear parameters: (1) model parameter $\boldsymbol{\beta}^*$ is a sparse vector; (2) model parameter $\boldsymbol{\Gamma}^*$ is a low rank matrix. Note that we assume X is a fully random Gaussian vector/matrix, thereby the population level conditions on $Q^{MLR}(\cdot|\cdot)$ hold in both settings.

Sparse Recovery. We assume model parameter $\boldsymbol{\beta}^*$ is s-sparse, i.e., $\boldsymbol{\beta}^* \in \mathcal{B}_0(s; p)$. Recall that, in order to serve sparse structure, we choose $\mathcal{R}$ to be ℓ_1 norm. Setting $\mathcal{S} = \overline{\mathcal{S}} = \text{supp}(\boldsymbol{\beta}^*)$, set $\mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \mathcal{R})$ corresponds to $\{\mathbf{u} : \|\mathbf{u}_{\mathcal{S}^\perp}\|_1 \leq 2\|\mathbf{u}_{\mathcal{S}}\|_1 + 2\sqrt{s}\|\mathbf{u}\|_2\}$. Restricted concavity of $Q^{MLR}(\cdot|\cdot)$ is validated in the following result.

Lemma 8 (RSC of MLR with sparsity)**.** *Consider mixed linear regression with any model parameter $\boldsymbol{\beta}^* \in \mathcal{B}_0(s; p)$ and function $Q_n^{MLR}(\cdot|\cdot)$ defined in* (5.3)*. There exit absolute constants $\{C_i\}_{i=0}^3$ such that, if $n \geq C_0 s\log p$, then for any $r > 0$, $Q_n^{MLR}(\cdot|\cdot)$ satisfies Condition 4 with parameters $(\gamma_n, \mathcal{S}, \overline{\mathcal{S}}, r, \delta)$, where*

$$\gamma_n = \frac{1}{3},\ (\mathcal{S}, \overline{\mathcal{S}}) = (\textit{supp}(\boldsymbol{\beta}^*), \textit{supp}(\boldsymbol{\beta}^*)),\ \delta = C_1\exp(-C_2 n).$$

Proof. Recall that

$$Q_n^{MLR}(\boldsymbol{\beta}'|\boldsymbol{\beta}) = -\frac{1}{2n}\sum_{i=1}^n\left[w(y_i, \mathbf{x}_i; \boldsymbol{\beta})(y_i - \langle\mathbf{x}_i, \boldsymbol{\beta}'\rangle)^2 + (1 - w(y_i, \mathbf{x}_i; \boldsymbol{\beta}))(y_i + \langle\mathbf{x}_i, \boldsymbol{\beta}'\rangle)^2\right].$$

For any $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^p$, we have

$$Q_n^{MLR}(\boldsymbol{\beta}'|\boldsymbol{\beta}) - Q_n^{MLR}(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - \langle\nabla Q_n^{MLR}(\boldsymbol{\beta}^*|\boldsymbol{\beta}), \boldsymbol{\beta}' - \boldsymbol{\beta}^*\rangle = -\frac{1}{2}(\boldsymbol{\beta}' - \boldsymbol{\beta}^*)^\top\left(\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^\top\right)(\boldsymbol{\beta}' - \boldsymbol{\beta}^*). \tag{B.15}$$

Note that we want to find γ_n such that the right hand side of (B.15) is less than $-\frac{\gamma_n}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2$ for any $\boldsymbol{\beta}' - \boldsymbol{\beta}^* \in \mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \mathcal{R})$. In this example, we have $\mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \mathcal{R}) =$

$\{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}_{\mathcal{S}^\perp}\|_1 \leq 2\|\mathbf{u}_{\mathcal{S}}\|_1 + 2\sqrt{s}\|\mathbf{u}\|_2\}$. It is sufficient to prove that the sample covariance matrix has restricted eigenvalues over set $\mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \mathcal{R})$. The following statement is follows by the substitution $\mathbf{\Sigma} = \mathbf{I}_p$ and $X = X$ in Lemma 23: there exist constants $\{C_i\}_{i=0}^2$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{x}_i, \mathbf{u}\rangle^2 \geq \frac{1}{2}\|\mathbf{u}\|_2^2 - C_0\frac{\log p}{n}\|\mathbf{u}\|_1^2, \text{ for all } \mathbf{u} \in \mathbb{R}^p, \tag{B.16}$$

with probability at least $1 - C_1\exp(-C_2 n)$. For any $\mathbf{u} \in \mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \mathcal{R})$, we have

$$\|\mathbf{u}\|_1 = \|\mathbf{u}_{\mathcal{S}}\|_1 + \|\mathbf{u}_{\mathcal{S}^\perp}\|_1 \leq 3\|\mathbf{u}_{\mathcal{S}}\|_1 + 2\sqrt{s}\|\mathbf{u}\|_2 \leq 5\sqrt{s}\|\mathbf{u}\|_2.$$

Applying (B.16) yields that

$$\frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{x}_i, \mathbf{u}\rangle^2 \geq \frac{1}{2}\|\mathbf{u}\|_2^2 - 25C_0\frac{s\log p}{n}\|\mathbf{u}\|_2^2, \text{ for all } \mathbf{u} \in \mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \mathcal{R}).$$

Consequently, when $n \geq C_3 s\log p$ for sufficiently large C_3, $\frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{x}_i, \mathbf{u}\rangle^2 \geq 1/3\|\mathbf{u}\|_2^2$, which implies $\gamma_n = 1/3$. □

Lemma 8 states that using $n = O(s\log p)$ samples makes $Q_n^{MLR}(\cdot|\cdot)$ be strongly concave over $\mathcal{C}$ with high probability.

Lemma 9 (Statistical error of MLR with sparsity)**.** *Consider mixed linear regression model with any $\boldsymbol{\beta}^* \in \mathcal{B}_0(s; p)$ and functions $Q_n^{MLR}(\cdot|\cdot), Q^{MLR}(\cdot|\cdot)$ defined in* (5.3) *and* (B.5) *respectively. There exist constants C and C_1 such that, for any $r > 0$ and $\delta \in (0, 1)$, if $n \geq C_1(\log p + \log(6/\delta))$, then*

$$\|\nabla Q_n^{MLR}(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - \nabla Q^{MLR}(\boldsymbol{\beta}^*|\boldsymbol{\beta})\|_\infty \leq C(\|\boldsymbol{\beta}^*\|_2 + \delta)\sqrt{\frac{\log p + \log(6/\delta)}{n}} \text{ for all } \boldsymbol{\beta} \in \mathcal{B}(r; \boldsymbol{\beta}^*)$$

with probability at least $1 - \delta$.

Proof. According to the formulations of $Q_n^{MLR}(\cdot|\cdot)$ and $Q^{MLR}(\cdot|\cdot)$ in (5.3) and (B.5), we have

$$\begin{aligned}
&\nabla Q_n^{MLR}(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - \nabla Q^{MLR}(\boldsymbol{\beta}^*|\boldsymbol{\beta}) \\
&= \boldsymbol{\beta}^* - \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)\boldsymbol{\beta}^* + \frac{2}{n}\sum_{i=1}^{n} w(y_i, \mathbf{x}_i; \boldsymbol{\beta})y_i\mathbf{x}_i - 2\mathbb{E}\left[w(Y, X; \boldsymbol{\beta})YX\right] - \frac{1}{n}\sum_{i=1}^{n} y_i\mathbf{x}_i.
\end{aligned} \tag{B.17}$$

So

$$\begin{aligned}
&\|\nabla Q_n^{MLR}(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - \nabla Q^{MLR}(\boldsymbol{\beta}^*|\boldsymbol{\beta})\|_\infty \\
&\leq \underbrace{\left\|\frac{1}{n}\sum_{i=1}^{n} y_i\mathbf{x}_i\right\|_\infty}_{(a)} + \underbrace{\left\|\boldsymbol{\beta}^* - \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)\boldsymbol{\beta}^*\right\|_\infty}_{(b)} + \underbrace{\left\|\frac{2}{n}\sum_{i=1}^{n} w(y_i, \mathbf{x}_i; \boldsymbol{\beta})y_i\mathbf{x}_i - 2\mathbb{E}\left[w(Y, X; \boldsymbol{\beta})YX\right]\right\|_\infty}_{(c)}.
\end{aligned}$$

Next we bound the above three terms (a), (b) and (c) respectively.

Term (a)**.** We let vector $\boldsymbol{\zeta} := \frac{1}{n}\sum_{i=1}^{n} y_i\mathbf{x}_i$. Consider jth coordinate of $\boldsymbol{\zeta}$. For any $j \in [p]$, we have

$$\zeta_j = \frac{1}{n}\sum_{i=1}^{n} y_i x_{i,j},$$

where $x_{i,j}$ is the jth coordinate of $\mathbf{x}_i$. Note that $\{y_i x_{ij}\}_{i=1}^n$ are independent copies of random variables $(\langle X, Z\cdot\boldsymbol{\beta}^*\rangle + W)X_j$ where $X \sim \mathcal{N}(0, \mathbf{I}_p)$, $W \sim \mathcal{N}(0, \sigma^2)$ and Z has Rademacher distribution. $\langle X, Z\cdot\boldsymbol{\beta}^*\rangle + W$ is sub-Gaussian random variable that has norm $\|\langle X, Z\cdot\boldsymbol{\beta}^*\rangle + W\|_{\psi_2} \lesssim \sqrt{\|\boldsymbol{\beta}^*\|_2^2 + \sigma^2}$. Also X_j is sub-Gaussian random variable that has norm $\|X_j\|_{\psi_2} \lesssim 1$. Then based on Lemma 22, $(\langle X, Z\cdot\boldsymbol{\beta}^*\rangle + W)X_j$ is sub-exponential with norm $\|(\langle X, Z\cdot\boldsymbol{\beta}^*\rangle + W)X_j\|_{\psi_1} \lesssim$

$\sqrt{\|\boldsymbol{\beta}^*\|_2^2 + \sigma^2}$. Following standard concentration result of sub-exponential random variables (e.g., Lemma 20), there exists some constant C such that the following inequality

$$\Pr\left(|\zeta_j| \geq t\right) \leq 2\exp\left(-C\frac{t^2 n}{\|\boldsymbol{\beta}^*\|_2^2 + \sigma^2}\right)$$

holds for sufficiently small $t > 0$. Therefore,

$$\Pr\left(\sup_{j\in[p]} |\zeta_j| > t\right) \leq 2p\exp\left(-C\frac{t^2 n}{\|\boldsymbol{\beta}^*\|_2^2 + \sigma^2}\right).$$

Setting the right hand side to be $\delta/3$, we have that, when n is sufficiently large (i.e., $n \geq C(\log p + \log(6/\delta))$ for some constant C), with probability at least $1 - \delta/3$.

$$\left\|\frac{1}{n}\sum_{i=1}^n y_i \mathbf{x}_i\right\|_\infty \lesssim (\|\boldsymbol{\beta}^*\|_2 + \sigma)\sqrt{\frac{\log p + \log(6/\delta)}{n}}. \tag{B.18}$$

Term (b)**.** Now we let $\boldsymbol{\zeta} = \boldsymbol{\beta}^* - \frac{1}{n}\mathbf{x}_i\mathbf{x}_i\boldsymbol{\beta}^*$. For any $j \in [p]$,

$$\zeta_j = \frac{1}{n}\sum_{i=1}^n \beta_j^* - x_{i,j}\langle \mathbf{x}_i, \boldsymbol{\beta}^*\rangle.$$

Note that $\{\beta_j^* - x_{i,j}\langle \mathbf{x}_i, \boldsymbol{\beta}^*\rangle\}_{i=1}^n$ are independent copies of random variable $\beta_j^* - X_j\langle X, \boldsymbol{\beta}^*\rangle$. Using similar analysis in bounding term (a), we claim that $\beta_j^* - X_j\langle X, \boldsymbol{\beta}^*\rangle$ is centered sub-exponential random variable with norm $\|\beta_j^* - X_j\langle X, \boldsymbol{\beta}^*\rangle\|_{\psi_1} \lesssim \|\boldsymbol{\beta}^*\|_2$. Therefore, for sufficiently small t and some constant C,

$$\Pr\left(|\zeta_j| \geq t\right) \leq 2\exp\left(-C\frac{t^2 n}{\|\boldsymbol{\beta}^*\|_2^2}\right).$$

Using union bound implies that

$$\Pr\left(\sup_{j\in[p]} |\zeta_j| \geq t\right) \leq 2p\cdot\exp\left(-C\frac{t^2 n}{\|\boldsymbol{\beta}^*\|_2^2}\right).$$

Setting the right hand side to be $\delta/3$, we have that, when n is sufficiently large,

$$\left\|\boldsymbol{\beta}^* - \left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\right)\boldsymbol{\beta}^*\right\|_\infty \lesssim \|\boldsymbol{\beta}^*\|_2\sqrt{\frac{\log p + \log(6/\delta)}{n}} \tag{B.19}$$

holds with probability at least $1 - \delta/3$.

Term (c). The analysis of this term is similar to the previous two terms. We let

$$\boldsymbol{\zeta} := \frac{1}{n}\sum_{i=1}^n w(\boldsymbol{y}_i, \mathbf{x}_i; \boldsymbol{\beta}) y_i \mathbf{x}_i - \mathbb{E}\left[w(Y, X; \boldsymbol{\beta}) Y X\right].$$

For any $j \in [p]$,

$$\zeta_j = \frac{1}{n}\sum_{i=1}^n w(\boldsymbol{y}_i, \mathbf{x}_i; \boldsymbol{\beta}) y_i x_{i,j} - \mathbb{E}\left[w(Y, X; \boldsymbol{\beta}) Y X\right].$$

Note that $\{w(\boldsymbol{y}_i, \mathbf{x}_i; \boldsymbol{\beta}) y_i x_{i,j}\}_{i=1}^n$ are independent copies of random variable $w(Y, X; \boldsymbol{\beta}) Y X_j$. We know that Y is sub-Gaussian with norm $\|Y\|_{\psi_2} \lesssim \sqrt{\|\boldsymbol{\beta}^*\|_2^2 + \sigma^2}$. Since $w(Y, X; \boldsymbol{\beta})$ is bounded, $w(Y, X; \boldsymbol{\beta})Y$ is also sub-Gaussian. Consequently, $w(Y, X; \boldsymbol{\beta}) Y X_j$ is sub-exponential. By standard concentration result, for some constant C and sufficiently small t,

$$\Pr(|\zeta_j| \geq t) \leq 2\exp\left(-C\frac{nt^2}{\|\boldsymbol{\beta}^*\|_2^2 + \sigma^2}\right).$$

Therefore,

$$\Pr(\sup_{j\in[p]} |\zeta_j| \geq t) \leq 2\exp\left(-C\frac{nt^2}{\|\boldsymbol{\beta}^*\|_2^2 + \sigma^2}\right).$$

Setting the right hand side to be $\delta/3$, we have that, when n is sufficiently large,

$$\left\| \frac{2}{n} \sum_{i=1}^{n} w(\boldsymbol{y}_i, \mathbf{x}_i; \boldsymbol{\beta}) y_i \mathbf{x}_i - 2\mathbb{E}\left[w(Y, X; \boldsymbol{\beta}) Y X \right] \right\|_\infty \lesssim (\|\boldsymbol{\beta}^*\|_2 + \delta) \sqrt{\frac{\log p + \log(6/\delta)}{n}} \quad \text{(B.20)}$$

with probability at least $1 - \delta/3$.

Putting (B.18), (B.19) and (B.20) together completes the proof. □

Lemma 9 implies Condition 5 hold with parameters $\Delta_n = O\left((\|\boldsymbol{\beta}^*\|_2 + \delta) \sqrt{\log p / n} \right)$, any $r > 0$ and $\delta = 1/p$. Putting all the ingredients together leads to the following guarantee about sparse recovery in mixed linear regression using regularized EM algorithm.

Proof of Corollary 2. The result follows from Theorem 1. First, we note that the minimum contractive factor $\kappa^* = 5\frac{\alpha\mu\tau}{\gamma\gamma_{n/T}} = 15\tau$ in this example since $\alpha = 1, \mu = \gamma = 1$ and $\gamma_{n/T} = 1/3$ w.h.p when $n \gtrsim s \log p$ (see Lemma 8). Following Lemma 7, $\kappa^* \leq 1/2$ when $w \leq 1/240$ and ρ is sufficiently large. Second, by choosing $n/T \gtrsim s \log p$, we have $\Delta_{n/T} \lesssim (\|\boldsymbol{\beta}^*\|_2 + \delta)\sqrt{\frac{T \log p}{n}}$ w.h.p., as proved in Lemma 9. Lastly, we have $\Delta \leq 3\overline{\Delta}$ by assuming $n/T \gtrsim [(\|\boldsymbol{\beta}^*\|_2 + \delta)/\|\boldsymbol{\beta}^*\|_2]^2 \, s \log p$. Putting these ingredients together and plugging the established parameters into (4.6) complete the proof. □

Low Rank Recovery. In the sequel, we assume model parameter $\boldsymbol{\Gamma}^* \in \mathbb{R}^{p_1 \times p_2}$ is a low rank matrix that has $\mathrm{rank}(\boldsymbol{\Gamma}^*) = \theta \ll \min\{p_1, p_2\}$. We focus on measuring the estimation error in Frobenius norm thus set $\|\cdot\|$ in our framework to be $\|\cdot\|_F$. Note that by treating $\boldsymbol{\Gamma}^*$ as a vector, Frobenius norm is equivalent to ℓ_2 norm, thereby we still have Lemma 6-7 in this setting. Moreover, SNR is similarly defined as

$$\mathrm{SNR} := \|\boldsymbol{\Gamma}^*\|_F / \sigma.$$

In order to serve the low rank structure, we choose $\mathcal{R}$ to be nuclear norm $\|\cdot\|_*$. For any matrix $\mathbf{M}$, we let $\mathrm{row}(\mathbf{M})$ denote the subspace spanned by the rows of $\mathbf{M}$ and $\mathrm{col}(\mathbf{M})$ denote the subspace spanned by the columns of $\mathbf{M}$. Moreover, for subspace represented by the columns of matrix $\mathbf{U}$, we denote the subspace orthogonal to $\mathbf{U}$ as $\mathbf{U}^\perp$. For $\boldsymbol{\Gamma}^*$ with singular value decomposition $\mathbf{U}^* \boldsymbol{\Sigma} \mathbf{V}^{*\top}$, we thus let

$$\mathcal{S} = \left\{ \mathbf{M} \in \mathbb{R}^{p_1 \times p_2} : \mathrm{col}(\mathbf{M}) \subseteq \mathbf{U}^*, \mathrm{row}(\mathbf{M}) \subseteq \mathbf{V}^* \right\} \quad \text{(B.21)}$$

and

$$\overline{\mathcal{S}}^\perp = \left\{ \mathbf{M} \in \mathbb{R}^{p_1 \times p_2} : \mathrm{col}(\mathbf{M}) \subseteq \mathbf{U}^{*\perp}, \mathrm{row}(\mathbf{M}) \subseteq \mathbf{V}^{*\perp} \right\}. \quad \text{(B.22)}$$

So $\mathcal{S}$ contains all matrices with rows (and columns) living in the row (and column) space of $\boldsymbol{\Gamma}^*$. Subspace $\overline{\mathcal{S}}^\perp$ contains all matrices with rows (and columns) orthogonal to the row (and column) space of $\boldsymbol{\Gamma}^*$. Nuclear norm is decomposable with respect to $(\mathcal{S}, \overline{\mathcal{S}})$. We have $\Psi(\overline{\mathcal{S}}) = \sup_{\mathbf{M} \in \overline{\mathcal{S}} \setminus \{\mathbf{0}\}} \|\mathbf{M}\|_* / \|\mathbf{M}\|_F \leq \sqrt{2\theta}$ since matrix in $\overline{\mathcal{S}}$ has rank at most 2θ. Similar to Lemma 8 and 9 for sparse structure, we have the following two results for low rank structure.

Lemma 10 (RSC of MLR with low rank structure)**.** *Consider mixed linear regression with model parameter $\boldsymbol{\Gamma}^* \in \mathbb{R}^{p_1 \times p_2}$ that has $\mathrm{rank}(\boldsymbol{\Gamma}^*) = \theta$. There exists constants $\{C_i\}_{i=0}^2$ such that, if $n \geq C_0 \theta \max\{p_1, p_2\}$, then for any $\theta \in (0, \min\{p_1, p_2\})$, $Q_n^{MLR}(\cdot|\cdot)$ satisfies Condition 4 with parameters $(\gamma_n, \mathcal{S}, \overline{\mathcal{S}}, r, \delta)$, where $(\mathcal{S}, \overline{\mathcal{S}})$ are given in* (B.21) *and* (B.22)*,*

$$\gamma_n = \frac{1}{20}, \ \delta = C_1 \exp(-C_2 n).$$

Proof. Similarly to (B.15), we have that for any $\boldsymbol{\Gamma}', \boldsymbol{\Gamma} \in \mathbb{R}^{p_1 \times p_2}$,

$$Q_n^{MLR}(\boldsymbol{\Gamma}'|\boldsymbol{\Gamma}) - Q_n^{MLR}(\boldsymbol{\Gamma}^*|\boldsymbol{\Gamma}) - \langle \nabla Q_n^{MLR}(\boldsymbol{\Gamma}^*|\boldsymbol{\Gamma}), \boldsymbol{\Gamma}' - \boldsymbol{\Gamma}^* \rangle = -\frac{1}{2n} \sum_{i=1}^{n} \langle \mathbf{X}_i, \boldsymbol{\Gamma}' - \boldsymbol{\Gamma}^* \rangle^2. \quad \text{(B.23)}$$

Note that $\boldsymbol{\Gamma}' - \boldsymbol{\Gamma}^* \in \mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \|\cdot\|_*)$. Let $\Theta := \boldsymbol{\Gamma}' - \boldsymbol{\Gamma}^*$, we thus have

$$\|\Theta_{\overline{\mathcal{S}}^\perp}\|_* \leq 2 \cdot \|\Theta_{\overline{\mathcal{S}}}\|_* + 2 \cdot \sqrt{2\theta} \|\Theta\|_F.$$

We make use of the following result.

Lemma 11. *Let $\{\mathbf{X}_i\}_{i=1}^n$ be n independent samples of random matrix $X \in \mathbb{R}^{p_1 \times p_2}$ where the entries are i.i.d. Gaussian random variable with distribution $\mathcal{N}(0,1)$. There exits constants C_1, C_2 such that*

$$\frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^n \langle \mathbf{X}_i, \Theta\rangle^2} \geq \frac{1}{4}\|\Theta\|_F - 12\left(\sqrt{\frac{p_1}{n}} + \sqrt{\frac{p_2}{n}}\right)\|\Theta\|_*, \text{ for all } \Theta \in \mathbb{R}^{p_1 \times p_2},$$

with probability at least $1 - C_1\exp(-C_2 n)$.

Proof. See Proposition 1 in [13] for detailed proof. □

Then for our Θ, using the above result yields that

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^n \langle \mathbf{X}_i, \Theta\rangle^2} &\geq \frac{1}{4}\|\Theta\|_F - 12\left(\sqrt{\frac{p_1}{n}} + \sqrt{\frac{p_2}{n}}\right)\left(\|\Theta_{\overline{\mathcal{S}}}\|_* + \|\Theta_{\overline{\mathcal{S}}^\perp}\|_*\right) \\
&\geq \frac{1}{4}\|\Theta\|_F - 12\left(\sqrt{\frac{p_1}{n}} + \sqrt{\frac{p_2}{n}}\right)\left(3\|\Theta_{\overline{\mathcal{S}}}\|_* + 2\sqrt{2r}\|\Theta\|_F\right) \\
&\geq \left[\frac{1}{4} - 60\sqrt{2\theta}\left(\sqrt{\frac{p_1}{n}} + \sqrt{\frac{p_2}{n}}\right)\right]\|\Theta\|_F.
\end{aligned}$$

So when $n \geq C\theta\max\{p_1, p_2\}$ for sufficient large C, we have $\frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^n \langle \mathbf{X}_i, \Theta\rangle^2} \geq \|\Theta\|_F/\sqrt{20}$. Plugging this result back into (B.23) gives us $\gamma_n = 1/20$ thus completes the proof. □

Lemma 12 (Statistical error of MLR with low rank structure)**.** *Consider the mixed linear regression with any $\mathbf{\Gamma}^* \in \mathbb{R}^{p_1 \times p_2}$. There exists constants C and C_1 such that, for any fixed $\mathbf{\Gamma} \in \mathbb{R}^{p_1 \times p_2}$ and $\delta \in (0,1)$, if $n \geq C_1(p_1 + p_2 + \log(6/\delta))$, then*

$$\|\nabla Q^{MLR}(\mathbf{\Gamma}^*|\mathbf{\Gamma}) - \nabla Q_n^{MLR}(\mathbf{\Gamma}^*|\mathbf{\Gamma})\|_2 \leq C(\|\mathbf{\Sigma}^*\|_F + \sigma)\sqrt{\frac{p_1 + p_2 + \log(6/\delta)}{n}}$$

with probability at least $1 - \delta$.

Proof. Parallel to (B.17), we have

$$\begin{aligned}
&\nabla Q_n^{MLR}(\mathbf{\Gamma}^*|\mathbf{\Gamma}) - \nabla Q^{MLR}(\mathbf{\Gamma}^*|\mathbf{\Gamma}) \\
&= \mathbf{\Gamma}^* - \frac{1}{n}\sum_{i=1}^n \langle \mathbf{X}_i, \mathbf{\Gamma}^*\rangle\mathbf{\Gamma}^* + \frac{2}{n}\sum_{i=1}^n w(y_i, \mathbf{X}_i; \mathbf{\Gamma})y_i\mathbf{X}_i - 2\mathbb{E}\left[w(Y, X; \mathbf{\Gamma})YX\right] - \frac{1}{n}\sum_{i=1}^n y_i\mathbf{X}_i.
\end{aligned}$$

The dual norm of nuclear norm is spectral norm. So we are interested in bounding the following term for fixed $\mathbf{\Gamma}$:

$$\begin{aligned}
&\left\|\nabla Q_n^{MLR}(\mathbf{\Gamma}^*|\mathbf{\Gamma}) - \nabla Q^{MLR}(\mathbf{\Gamma}^*|\mathbf{\Gamma})\right\|_2 \\
&\leq \underbrace{\left\|\frac{1}{n}\sum_{i=1}^n y_i\mathbf{X}_i\right\|_2}_{U_1} + \underbrace{\left\|\mathbf{\Gamma}^* - \frac{1}{n}\sum_{i=1}^n \langle \mathbf{X}_i, \mathbf{\Gamma}^*\rangle\mathbf{X}_i\right\|_2}_{U_2} + \underbrace{\left\|\frac{2}{n}\sum_{i=1}^n w(y_i, \mathbf{X}_i; \mathbf{\Gamma})y_i\mathbf{X}_i - 2\mathbb{E}\left[w(Y, X; \mathbf{\Gamma})YX\right]\right\|_2}_{U_3}.
\end{aligned}$$

Next we bound the three terms U_1, U_2 and U_3 respectively.

Term U_1. We first note that

$$U_1 = \sup_{\substack{\mathbf{u} \in \mathbb{S}^{p_1 - 1} \\ \mathbf{v} \in \mathbb{S}^{p_2 - 1}}} \frac{1}{n}\sum_{i=1}^n y_i\langle \mathbf{u}\mathbf{v}^\top, \mathbf{X}_i\rangle.$$

In particular, we let

$$Z(a, b) = \sup_{\substack{\mathbf{u} \in a\mathbb{S}^{p_1 - 1} \\ \mathbf{v} \in b\mathbb{S}^{p_2 - 1}}} \frac{1}{n}\sum_{i=1}^n y_i\langle \mathbf{u}\mathbf{v}^\top, \mathbf{X}_i\rangle.$$

We thus have $Z(a,b) = abZ(1,1)$. We construct $1/4$-covering sets of $\mathbb{S}^{p_1-1}$ and $\mathbb{S}^{p_2-1}$, which we denote as $\mathcal{N}_1$ and $\mathcal{N}_2$ respectively. Therefore, for any $\boldsymbol{u} \in \mathbb{S}^{p-1}, \boldsymbol{v} \in \mathbb{S}^{p_2-1}$, we can always find $\mathbf{u}' \in \mathcal{N}_1, \mathbf{v}' \in \mathcal{N}_2$ such that $\|\mathbf{u}-\mathbf{u}'\|_2 \le 1/4$, $\|\mathbf{v}-\mathbf{v}'\|_2 \le 1/4$. Moreover, we have the following decomposition $\mathbf{u}\mathbf{v}^\top = \mathbf{u}'\mathbf{v}'^\top + (\mathbf{u}-\mathbf{u}')\mathbf{v}'^\top + \mathbf{u}'(\mathbf{v}-\mathbf{v}')^\top + (\mathbf{u}-\mathbf{u}')(\mathbf{v}-\mathbf{v}')^\top$. Therefore, we have

$$Z(1,1) \le \max_{\mathbf{u}\in\mathcal{N}_1,\mathbf{v}\in\mathcal{N}_2} \frac{1}{n}\sum_{i=1}^n y_i\langle \mathbf{u}\mathbf{v}^\top, \mathbf{X}_i\rangle + Z(1/4,1) + Z(1/4,1) + Z(1/4,1/4),$$

which implies that

$$Z(1,1) \le \frac{16}{7}\max_{\mathbf{u}\in\mathcal{N}_1,\mathbf{v}\in\mathcal{N}_2} \frac{1}{n}\sum_{i=1}^n y_i\langle \mathbf{u}\mathbf{v}^\top, \mathbf{X}_i\rangle.$$

For any fixed $\mathbf{u}$ and $\mathbf{v}$, $\{y_i\langle \mathbf{u}\mathbf{v}^\top, \mathbf{X}_i\rangle\}_{i=1}^n$ are n independent copies of random variable $Y\langle \mathbf{u}\mathbf{v}^\top, X\rangle$ where Y is sub-Gaussian with norm $\|Y\|_{\psi_2} \lesssim \sqrt{\|\mathbf{\Gamma}^*\|_F^2+\sigma^2}$, $\langle \mathbf{u}\mathbf{v}^\top, X\rangle$ is zero mean Gaussian with variance 1. Following Lemma 22, $Y\langle \mathbf{u}\mathbf{v}^\top, X\rangle$ is sub-exponential with norm $\|Y\langle \mathbf{u}\mathbf{v}^\top, X\rangle\|_{\psi_1} \lesssim \sqrt{\|\mathbf{\Gamma}^*\|_F^2+\sigma^2}$. Using concentration result in Lemma 20, we have

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n y_i\langle \mathbf{u}\mathbf{v}^\top, \mathbf{X}_i\rangle\right| \ge t\right) \le 2\exp\left(-\frac{Ct^2 n}{\|\mathbf{\Gamma}^*\|_F^2+\sigma^2}\right)$$

for sufficiently small $t > 0$. Note that $|\mathcal{N}_1| \le 9^{p_1}, |\mathcal{N}_2| \le 9^{p_2}$. By applying union bounds over $\mathcal{N}_1$ and $\mathcal{N}_2$, we have

$$\Pr\left(\max_{\mathbf{u}\in\mathcal{N}_1,\mathbf{v}\in\mathcal{N}_2} \frac{1}{n}\sum_{i=1}^n y_i\langle \mathbf{u}\mathbf{v}^\top, \mathbf{X}_i\rangle \ge t\right) \le 2\cdot 9^{(p_1+p_2)}\exp\left(-\frac{Ct^2 n}{\|\mathbf{\Gamma}^*\|_F^2+\sigma^2}\right).$$

By setting the right hand side to be $\delta/3$, we have that if $n \ge C(p_1+p_2+\log(6/\delta))$ for sufficiently large C, then

$$U_1 \lesssim (\|\mathbf{\Gamma}^*\|_F + \sigma)\sqrt{\frac{p_1+p_2+\log(6/\delta)}{n}} \tag{B.24}$$

with probability at least $1-\delta/3$.

Term U_2. Parallel to the analysis of term U_1, we have

$$U_2 = \sup_{\substack{\mathbf{u}\in\mathbb{S}^{p_1-1}\\ \mathbf{v}\in\mathbb{S}^{p_2-1}}} \langle \mathbf{u}\mathbf{v}^\top, \mathbf{\Gamma}^*\rangle - \frac{1}{n}\sum_{i=1}^n \langle \mathbf{X}_i, \mathbf{\Gamma}^*\rangle\cdot\langle \mathbf{u}\mathbf{v}^\top, \mathbf{X}_i\rangle.$$

We construct $1/4$-nets $\mathcal{N}_1, \mathcal{N}_2$ of $\mathbb{S}^{p_1-1}$ and $\mathbb{S}^{p_2-1}$ respectively. Then

$$U_2 \le \frac{16}{7}\max_{\mathbf{u}\in\mathcal{N}_1,\mathbf{v}\in\mathcal{N}_2} \langle \mathbf{u}\mathbf{v}^\top, \mathbf{\Gamma}^*\rangle - \frac{1}{n}\sum_{i=1}^n \langle \mathbf{X}_i, \mathbf{\Gamma}^*\rangle\cdot\langle \mathbf{u}\mathbf{v}^\top, \mathbf{X}_i\rangle.$$

For any fixed $\mathbf{u}, \mathbf{v}$, note that $\{\langle \mathbf{X}_i, \mathbf{\Gamma}^*\rangle\cdot\langle \mathbf{u}\mathbf{v}^\top, \mathbf{X}_i\rangle\}_{i=1}^n$ are n independent samples of random variable $\langle X, \mathbf{\Gamma}^*\rangle\cdot\langle \mathbf{u}\mathbf{v}^\top, X\rangle$ where $\langle X, \mathbf{\Gamma}^*\rangle \sim \mathcal{N}(0, \|\mathbf{\Gamma}^*\|_F^2)$ and $\langle \mathbf{u}\mathbf{v}^\top, X\rangle \sim \mathcal{N}(0,1)$. So $\langle X, \mathbf{\Gamma}^*\rangle\cdot\langle \mathbf{u}\mathbf{v}^\top, X\rangle$ is sub-exponential with norm $O(\|\mathbf{\Gamma}^*\|_F)$. Using the centering argument (Lemma 21) and concentration result (Lemma 20), we have

$$\Pr\left(\left|\langle \mathbf{u}\mathbf{v}^\top, \mathbf{\Gamma}^*\rangle - \frac{1}{n}\sum_{i=1}^n \langle \mathbf{X}_i, \mathbf{\Gamma}^*\rangle\cdot\langle \mathbf{u}\mathbf{v}^\top, \mathbf{X}_i\rangle\right| \ge t\right) \le 2\cdot\exp\left(-C\frac{t^2 n}{\|\mathbf{\Gamma}^*\|_F^2}\right)$$

for sufficiently small t. Using the union bound over sets $\mathcal{N}_1, \mathcal{N}_2$, we conclude that when $n \ge C(p_1+p_2+\log(6/\delta))$ for sufficiently large C, we have

$$U_2 \lesssim \|\mathbf{\Gamma}^*\|_F\sqrt{\frac{p_1+p_2+\log(6/\delta)}{n}} \tag{B.25}$$

with probability at least $1-\delta/3$.

Term U_3. We first have

$$U_3 = \sup_{\substack{\mathbf{u}\in\mathbb{S}^{p_1-1}\\ \mathbf{v}\in\mathbb{S}^{p_2-1}}} \frac{2}{n}\sum_{i=1}^{n} w\cdot y_i\langle \mathbf{u}\mathbf{v}^\top, \mathbf{X}_i\rangle - 2\mathbb{E}\left[w\cdot Y\langle \mathbf{u}\mathbf{v}^\top, X\rangle\right].$$

Similar to the analysis of the first two terms, by constructing $\mathcal{N}_1, \mathcal{N}_2$, we have

$$U_3 \le \frac{16}{7}\max_{\mathbf{u}\in\mathcal{N}_1,\mathbf{v}\in\mathcal{N}_2} \frac{2}{n}\sum_{i=1}^{n} w\cdot y_i\langle \mathbf{u}\mathbf{v}^\top, \mathbf{X}_i\rangle - 2\mathbb{E}\left[w\cdot Y\langle \mathbf{u}\mathbf{v}^\top, X\rangle\right].$$

Note that $\{w\boldsymbol{y}_i\langle \mathbf{u}\mathbf{v}^\top, \mathbf{X}_i\rangle\}_{i=1}^n$ are n independent samples of random variable $wY\langle \mathbf{u}\mathbf{v}^\top, X\rangle$ where $\langle \mathbf{u}\mathbf{v}^\top, X\rangle \sim \mathcal{N}(0,1)$ and wY is sub-Gaussian with norm $\|wY\|_{\psi_2} \lesssim \sqrt{\|\mathbf{\Gamma}^*\|_F^2+\sigma^2}$ since $|w|\le 1$. We thus have $wY\langle \mathbf{u}\mathbf{v}^\top, X\rangle$ is sub-exponential with norm $\|wY\langle \mathbf{u}\mathbf{v}^\top, X\rangle\|_{\psi_1} \lesssim \sqrt{\|\mathbf{\Gamma}^*\|_F^2+\sigma^2}$. Then following the similar steps in analyzing the first two terms, we reach the conclusion that

$$U_3 \lesssim (\|\mathbf{\Gamma}^*\|_F+\sigma)\sqrt{\frac{p_1+p_2+\log(6/\delta)}{n}} \tag{B.26}$$

with probability at least $1-\delta/3$ when $n \gtrsim p_1+p_2+\log(6/\delta)$.

Putting (B.24), (B.25) and (B.26) together completes the proof. □

Setting $\delta = 6\exp(-(p_1+p_2))$ in Lemma 12 suggests that Condition 5 holds with parameters (Δ_n, r, δ) where $\Delta_n \lesssim (\|\mathbf{\Gamma}^*\|_F+\delta)\sqrt{(p_1+p_2)/n}$, $\delta = \exp(-(p_1+p_2))$ and r can be any positive number. Putting these pieces together leads to the following guarantee about low rank recovery.

Proof of Corollary 3. This result is parallel to Corollary 2 for sparse recovery thus can be proved similarly. We omit the details. □

B.3 Missing Covariate Regression

We now turn to missing covariate regression. We first reveal function $Q_n^{MCR}(\cdot|\cdot)$ and $Q^{MCR}(\cdot|\cdot)$. To ease notation, we introduce vector $\mathbf{z}_i \in \{0,1\}^p$ to indicate the positions of missing entries, i.e., $z_{i,j}=1$ if $x_{i,j}$ is missing. In this example, the E step involves computing the distribution of missing entries given current parameter guess $\boldsymbol{\beta}$. Under Gaussian design $X\sim\mathcal{N}(\mathbf{0},\mathbf{I}_p), W\sim\mathcal{N}(0,\sigma^2)$, given observed covariate entries $(\mathbf{1}-\mathbf{z}_i)\odot\mathbf{x}_i$ and y_i, the conditional mean vector of $\widetilde{\mathbf{x}}_i$ has form

$$\boldsymbol{\mu}_{\boldsymbol{\beta}}(y_i,\mathbf{z}_i,\mathbf{x}_i) := \mathbb{E}[\widetilde{\mathbf{x}}_i|\boldsymbol{\beta}, y_i, (\mathbf{1}-\mathbf{z}_i)\odot\mathbf{x}_i] = (\mathbf{1}-\mathbf{z}_i)\odot\mathbf{x}_i + \frac{y_i-\langle\boldsymbol{\beta},(\mathbf{1}-\mathbf{z}_i)\odot\mathbf{x}_i\rangle}{\sigma^2+\|\mathbf{z}_i\odot\boldsymbol{\beta}\|_2^2}\mathbf{z}_i\odot\boldsymbol{\beta}, \tag{B.27}$$

and the conditional correlation matrix of $\widetilde{\mathbf{x}}_i$ has form

$$\begin{aligned}\mathbf{\Sigma}_{\boldsymbol{\beta}}(y_i,\mathbf{z}_i,\mathbf{x}_i) &:= \mathbb{E}\left[\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^\top|\boldsymbol{\beta}, y_i, (\mathbf{1}-\mathbf{z}_i)\odot\mathbf{x}_i\right]\\ &= \boldsymbol{\mu}_{\boldsymbol{\beta}}\boldsymbol{\mu}_{\boldsymbol{\beta}}^\top + \mathrm{diag}(\mathbf{z}_i) - \left(\frac{1}{\sigma^2+\|\mathbf{z}_i\odot\boldsymbol{\beta}\|_2^2}\right)(\mathbf{z}_i\odot\boldsymbol{\beta})(\mathbf{z}_i\odot\boldsymbol{\beta})^\top.\end{aligned} \tag{B.28}$$

Consequently, $Q_n(\cdot|\cdot)$ corresponds to

$$Q_n^{MCR}(\boldsymbol{\beta}'|\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\langle y_i\boldsymbol{\mu}_{\boldsymbol{\beta}}(y_i,\mathbf{z}_i,\mathbf{x}_i),\boldsymbol{\beta}'\rangle - \frac{1}{2}\boldsymbol{\beta}^\top\mathbf{\Sigma}_{\boldsymbol{\beta}}(y_i,\mathbf{z}_i,\mathbf{x}_i)\boldsymbol{\beta}. \tag{B.29}$$

We thus have that $Q^{MCR}(\cdot|\cdot)$ takes form

$$Q^{MCR}(\boldsymbol{\beta}'|\boldsymbol{\beta}) = \langle\mathbb{E}\left[Y\boldsymbol{\mu}_{\boldsymbol{\beta}}(Y,Z,X)\right],\boldsymbol{\beta}'\rangle - \frac{1}{2}\left\langle\mathbb{E}\left[\mathbf{\Sigma}_{\boldsymbol{\beta}}(Y,Z,X)\right],\boldsymbol{\beta}\boldsymbol{\beta}^\top\right\rangle. \tag{B.30}$$

In particular, we let $\overline{\mathbf{\Sigma}}_{\boldsymbol{\beta}} := \mathbb{E}\left[\mathbf{\Sigma}_{\boldsymbol{\beta}}(Y,Z,X)\right]$. We first present a key result that characterizes the spectral property of $\overline{\mathbf{\Sigma}}_{\boldsymbol{\beta}}$.

Lemma 13. *For $\overline{\mathbf{\Sigma}}_{\boldsymbol{\beta}}$, we have the following decomposition*

$$\overline{\mathbf{\Sigma}}_{\boldsymbol{\beta}} = \epsilon \mathbf{I}_p + \mathbf{\Sigma}_1 - \mathbf{\Sigma}_2,$$

where

$$\mathbf{\Sigma}_1 = \mathbb{E}\left\{[(\mathbf{1}-Z)\odot X + \nu Z \odot \boldsymbol{\beta}]\cdot[(\mathbf{1}-Z)\odot X + \nu Z\odot\boldsymbol{\beta}]^\top\right\},$$
$$\mathbf{\Sigma}_2 = \mathbb{E}\left[\frac{1}{\sigma^2+\|Z\odot\boldsymbol{\beta}\|_2^2}(Z\odot\boldsymbol{\beta})(Z\odot\boldsymbol{\beta})^\top\right], \quad \nu = \frac{Y-\langle\boldsymbol{\beta},(\mathbf{1}-Z)\odot X\rangle}{\sigma^2+\|Z\odot\boldsymbol{\beta}\|_2^2}.$$

Let $\zeta := (1+\omega)\rho$, we have

$$\lambda_{min}(\mathbf{\Sigma}_1) \geq 1-\epsilon-2\zeta^2\sqrt{\epsilon}, \tag{B.31}$$
$$\lambda_{max}(\mathbf{\Sigma}_2) \leq \zeta^2\epsilon, \tag{B.32}$$
$$\lambda_{max}(\overline{\mathbf{\Sigma}}_{\boldsymbol{\beta}}) \leq 1+2\zeta^2\sqrt{\epsilon}+(1+\zeta^2)\zeta^2\epsilon. \tag{B.33}$$

In particular, let $\boldsymbol{\beta}=\boldsymbol{\beta}^*$, we have $\overline{\mathbf{\Sigma}}_{\boldsymbol{\beta}^*}=\mathbf{I}_p$.

Proof. The decomposition follows by taking expectation of (B.28). For $\mathbf{\Sigma}_1$, expanding the bracket leads to

$$\mathbf{\Sigma}_1 = (1-\epsilon)\mathbf{I}_p + \underbrace{\mathbb{E}\left\{\nu[(\mathbf{1}-Z)\odot X](Z\odot\boldsymbol{\beta})^\top + \nu(Z\odot\boldsymbol{\beta})[(\mathbf{1}-Z)\odot X]^\top\right\}}_{\mathbf{M}} + \underbrace{\mathbb{E}\left[\nu^2(Z\odot\boldsymbol{\beta})(Z\odot\boldsymbol{\beta})^\top\right]}_{\mathbf{N}}.$$

For term $\mathbf{M}$, consider its spectral norm. Since it is symmetric, we have

$$\begin{aligned}
\|\mathbf{M}\|_2 &= \sup_{\mathbf{u}\in\mathbb{S}^{p-1}} 2\left|\mathbb{E}\left[\nu\langle Z\odot\boldsymbol{\beta},\mathbf{u}\rangle\cdot\langle(\mathbf{1}-Z)\odot X,\mathbf{u}\rangle\right]\right| \\
&= 2\sup_{\mathbf{u}\in\mathbb{S}^{p-1}}\left|\mathbb{E}\left[\frac{1}{\sigma^2+\|Z\odot\boldsymbol{\beta}\|_2}\langle(\mathbf{1}-Z)\odot(\boldsymbol{\beta}^*-\boldsymbol{\beta}),\mathbf{u}\rangle\cdot\langle Z\odot\boldsymbol{\beta},\mathbf{u}\rangle\right]\right| \\
&\leq 2\frac{1}{\sigma^2}\mathbb{E}\left[\|(\mathbf{1}-Z)\odot(\boldsymbol{\beta}^*-\boldsymbol{\beta})\|_2\|Z\odot\boldsymbol{\beta}\|_2\right] \leq 2\frac{1}{\sigma^2}\sqrt{\mathbb{E}\left[\|(\mathbf{1}-Z)\odot(\boldsymbol{\beta}^*-\boldsymbol{\beta})\|_2^2\cdot\|Z\odot\boldsymbol{\beta}\|_2^2\right]} \\
&\leq 2\frac{1}{\sigma^2}\sqrt{\epsilon(1-\epsilon)}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\|\boldsymbol{\beta}\|_2 \leq 2\rho^2\omega(1+\omega)\sqrt{\epsilon(1-\epsilon)} \leq 2\zeta^2\sqrt{\epsilon}.
\end{aligned}$$

where the second equality follows by taking expectation of X and Gaussian noise W, the last inequality follows from the definitions of ω,ρ given in Section B.3. Note that $\mathbf{N}\succeq\mathbf{0}$. Then the lower bound of $\lambda_{min}(\mathbf{\Sigma}_1)$ follows by using $\lambda_{min}(\mathbf{\Sigma}_1)\geq 1-\epsilon-\|\mathbf{M}\|_2$. For $\mathbf{\Sigma}_2$, we have

$$\mathbf{\Sigma}_2 = \mathbb{E}\left[\frac{1}{\sigma^2+\|Z\odot\boldsymbol{\beta}\|_2^2}(Z\odot\boldsymbol{\beta})(Z\odot\boldsymbol{\beta})^\top\right] \preceq \frac{1}{\sigma^2}\left((\epsilon-\epsilon^2)\mathrm{diag}(\boldsymbol{\beta}\odot\boldsymbol{\beta})+\epsilon^2\boldsymbol{\beta}\boldsymbol{\beta}^\top\right).$$

Therefore, $\lambda_{max}(\mathbf{\Sigma}_2)\leq\zeta^2\epsilon$. Note that

$$\begin{aligned}
\mathbf{N} &\preceq \frac{1}{\sigma^4}\mathbb{E}\left[(Y-\langle\boldsymbol{\beta},(\mathbf{1}-Z)\odot X\rangle)^2(Z\odot\boldsymbol{\beta})(Z\odot\boldsymbol{\beta})^\top\right] \\
&= \frac{1}{\sigma^4}\mathbb{E}\left[(\sigma^2+\|\boldsymbol{\beta}^*-(\mathbf{1}-Z)\odot\boldsymbol{\beta}\|_2^2)(Z\odot\boldsymbol{\beta})(Z\odot\boldsymbol{\beta})^\top\right] \\
&\preceq \frac{1}{\sigma^4}(\sigma^2+\|\boldsymbol{\beta}^*\|_2^2+\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2)\left((\epsilon-\epsilon^2)\mathrm{diag}(\boldsymbol{\beta}\odot\boldsymbol{\beta})+\epsilon^2\boldsymbol{\beta}\boldsymbol{\beta}^\top\right).
\end{aligned}$$

We thus have $\lambda_{max}(\mathbf{N})\leq\frac{1}{\sigma^4}(\sigma^2+\|\boldsymbol{\beta}^*\|_2^2+\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2)\epsilon\|\boldsymbol{\beta}\|_2^2\leq(1+\zeta^2)\zeta^2\epsilon$. The corresponding bound for $\lambda_{max}(\overline{\mathbf{\Sigma}}_{\boldsymbol{\beta}})$ then follows from $\lambda_{max}(\overline{\mathbf{\Sigma}}_{\boldsymbol{\beta}})\leq 1+\lambda_{max}(\mathbf{M})+\lambda_{max}(\mathbf{N})$.

When $\boldsymbol{\beta}=\boldsymbol{\beta}^*$, we have

$$\mathbb{E}_{X,W}(\nu^2) = \frac{\mathbb{E}_{X,W}\left[(\langle X,\boldsymbol{\beta}^*\rangle+W-\langle X,(1-Z)\odot\boldsymbol{\beta}^*\rangle)^2\right]}{(\sigma^2+\|Z\odot\boldsymbol{\beta}^*\|_2^2)^2} = \frac{1}{\sigma^2+\|Z\odot\boldsymbol{\beta}^*\|_2^2}$$

and

$$\begin{aligned}
\mathbb{E}_{X,W}(\nu(\mathbf{1}-Z)\odot X) &= \frac{\mathbb{E}\left[(\langle X,\boldsymbol{\beta}^*\rangle+W-\langle X,(1-Z)\odot\boldsymbol{\beta}^*\rangle)(\mathbf{1}-Z)\odot X\right]}{\sigma^2+\|Z\odot\boldsymbol{\beta}^*\|_2^2} \\
&= \frac{(\mathbf{1}-Z)\odot Z\odot\boldsymbol{\beta}^*}{\sigma^2+\|Z\odot\boldsymbol{\beta}^*\|_2^2} = \mathbf{0}.
\end{aligned}$$

Therefore, $\mathbf{M}=\mathbf{0}$ and $\mathbf{N}=\mathbf{\Sigma}_2$. We thus have $\overline{\mathbf{\Sigma}}_{\boldsymbol{\beta}^*}=\epsilon\mathbf{I}_p+(1-\epsilon)\mathbf{I}_p=\mathbf{I}_p$. □

We now turn to check technical conditions about $Q^{MCR}(\cdot|\cdot)$. First, $\mathcal{M}(\cdot)$ is self consistent as stated below.

Lemma 14 (Self-consistency of MCR). *Consider missing covariate regression with parameter $\boldsymbol{\beta}^* \in \mathbb{R}^p$ and $Q^{MCR}(\cdot|\cdot)$ given in* (B.30)*. We have*

$$\boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} Q^{MCR}(\boldsymbol{\beta}|\boldsymbol{\beta}^*).$$

Proof. In this example

$$\mathcal{M}(\boldsymbol{\beta}^*) = \left(\mathbb{E}\left[\boldsymbol{\Sigma}_{\boldsymbol{\beta}^*}(Y,Z,X)\right]\right)^{-1}\mathbb{E}\left[Y\boldsymbol{\mu}_{\boldsymbol{\beta}^*}(Y,Z,X)\right].$$

Following Lemma 13, we have $\boldsymbol{\Sigma}_{\boldsymbol{\beta}^*}(Y,Z,X) = \mathbf{I}_p$. Meanwhile, we have

$$\begin{aligned}\mathbb{E}\left[Y\boldsymbol{\mu}_{\boldsymbol{\beta}^*}(Y,Z,X)\right] &= \mathbb{E}\left[(\langle\boldsymbol{\beta}^*, X\rangle + W)\left((\mathbf{1}-Z)\odot X + \frac{\langle Z\odot\boldsymbol{\beta}^*, X\rangle + W}{\sigma^2 + \|Z\odot\boldsymbol{\beta}^*\|_2^2} Z\odot\boldsymbol{\beta}^*\right)\right]\\ &= \mathbb{E}\left[(\mathbf{1}-Z)\odot\boldsymbol{\beta}^* + Z\odot\boldsymbol{\beta}^*\right] = \boldsymbol{\beta}^*.\end{aligned}$$

Thus $\mathcal{M}(\boldsymbol{\beta}^*) = \boldsymbol{\beta}^*$. □

For our analysis, we define $\rho := \|\boldsymbol{\beta}^*\|_2/\sigma$ to be the *signal to noise ratio* and $\omega := r/\|\boldsymbol{\beta}^*\|_2$ to be the *relative contractivity radius*. Let

$$\zeta := (1+\omega)\rho.$$

Recall that ϵ is the missing probability of every entry. The next result characterizes the smoothness and concavity of $Q^{MCR}(\cdot|\cdot)$.

Lemma 15 (Smoothness and concavity of MCR). *Consider missing covariate regression with parameter $\boldsymbol{\beta}^* \in \mathbb{R}^p$ and $Q^{MCR}(\cdot|\cdot)$ given in* (B.30)*. For any $\omega > 0$, we have that $Q^{MCR}(\cdot|\cdot)$ satisfies Condition 2 with parameters $(\gamma, \mu, \omega\|\boldsymbol{\beta}^*\|_2)$, where*

$$\gamma = 1, \;\; \mu = 1 + 2\zeta^2\sqrt{\epsilon} + (1+\zeta^2)\zeta^2\epsilon.$$

Proof. Following Lemma 13, we have $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}^*} = \mathbf{I}_p$. Therefore, $Q^{MCR}(\cdot|\boldsymbol{\beta}^*)$ is 1-strongly concave. For any $\boldsymbol{\beta} \in \mathcal{B}(w\|\boldsymbol{\beta}^*\|; \boldsymbol{\beta}^*)$, following (B.33), we have that $Q^{MCR}(\cdot|\boldsymbol{\beta})$ is μ-smooth with $\mu = 1 + 2\zeta^2\sqrt{\epsilon} + (1+\zeta^2)\zeta^2\epsilon$. □

We revisit the following result about the gradient stability from [1].

Lemma 16 (Gradient stability of MCR). *Consider the missing covariate regression with $\boldsymbol{\beta}^* \in \mathbb{R}^p$ and $Q^{MCR}(\cdot|\cdot)$ given in* (B.30)*. For any $\omega > 0, \rho > 0$, $Q^{MCR}(\cdot|\cdot)$ satisfies Condition 3 with parameter $(\tau, \omega\|\boldsymbol{\beta}^*\|_2)$ where*

$$\tau = \frac{\zeta^2 + 2\epsilon(1+\zeta^2)^2}{1+\zeta^2}.$$

Proof. See the proof of Corollary 6 in [1]. □

Unlike the previous two models, we require an upper bound on the signal to noise ratio. This unusual constraint is in fact unavoidable, as pointed out in [10].

We now turn to validate the conditions on finite sample function $Q_n^{MCR}(\cdot|\cdot)$. In particular, we have the following two guarantees.

Lemma 17 (RSC of MCR). *Consider missing covariate regression with any fixed parameter $\boldsymbol{\beta}^* \in \mathcal{B}_0(s;p)$ and $Q_n^{MCR}(\cdot|\cdot)$ given in* (B.29)*. There exist constants $\{C_i\}_{i=0}^3$ such that if $\epsilon \le C_0\min\{1, \zeta^{-4}\}$ and $n \ge C_1(1+\zeta)^8 s\log p$, then we have $Q_n^{MCR}(\cdot|\cdot)$ satisfies Condition 4 with parameters $(\gamma_n, \mathcal{S}, \overline{\mathcal{S}}, \omega\|\boldsymbol{\beta}^*\|_2, \delta)$, where*

$$\gamma_n = \frac{1}{9}, \;\; (\mathcal{S}, \overline{\mathcal{S}}) = (\mathit{supp}(\boldsymbol{\beta}^*), \mathit{supp}(\boldsymbol{\beta}^*)), \;\; \delta = C_2\exp(-C_3 n(1+\zeta)^{-8}).$$

Proof. In order to show $Q_n^{MCR}(\cdot|\boldsymbol{\beta})$ is γ_n-strongly concave over $\mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \mathcal{R})$, since $Q_n^{MCR}(\cdot|\boldsymbol{\beta})$ is quadratic, it is then equivalent to show

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{u}^\top \boldsymbol{\Sigma}_{\boldsymbol{\beta}}(y_i, \mathbf{z}_i, \mathbf{x}_i)\mathbf{u} \geq \gamma_n \|\mathbf{u}\|_2^2$$

for all $\mathbf{u} \in \mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}, \mathcal{R})$. Expanding $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ gives us

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{u}^\top \boldsymbol{\Sigma}_{\boldsymbol{\beta}}(y_i, \mathbf{z}_i, \mathbf{x}_i)\mathbf{u} \geq \underbrace{\frac{1}{n}\sum_{i=1}^{n} \langle \boldsymbol{\mu}_{\boldsymbol{\beta}}(y_i, \mathbf{z}_i, \mathbf{x}_i), \mathbf{u}\rangle^2}_{L_1} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} \left(\frac{1}{\sigma^2 + \|\mathbf{z}_i \odot \boldsymbol{\beta}\|_2^2}\right) \langle \mathbf{z}_i \odot \boldsymbol{\beta}, \mathbf{u}\rangle^2}_{L_2}.$$

We choose to bound each term using restricted eigenvalue argument in Lemma 23. To ease notation, we let $\nu := \frac{y_i - \langle (\mathbf{1}-\mathbf{z}_i)\odot\boldsymbol{\beta}, \mathbf{x}_i\rangle}{\sigma^2 + \|\mathbf{z}_i \odot \boldsymbol{\beta}\|_2^2}$.

Term L_1. Note that $\boldsymbol{\mu}_{\boldsymbol{\beta}}(y_i, \mathbf{z}_i, \mathbf{x}_i)$ are samples of $\boldsymbol{\mu}_{\boldsymbol{\beta}}(Y, Z, X)$ which is zero mean sub-Gaussian random vector with covariance matrix $\boldsymbol{\Sigma}_1$ given in Lemma 13. Moreover, we have $\lambda_{min}(\boldsymbol{\Sigma}_1) \geq 1 - \epsilon - 2\zeta^2\sqrt{\epsilon}$. By restricting $\epsilon \leq 1/4$ and assuming $\epsilon \leq C\zeta^{-4}$ for sufficiently small C, we have $\lambda_{min}(\boldsymbol{\Sigma}_1) \geq \frac{1}{2}$. Moreover

$$\|\boldsymbol{\mu}_{\boldsymbol{\beta}}(Y, Z, X)\|_{\psi_2} \lesssim \|(\mathbf{1} - Z) \odot X\|_{\psi_2} + \|\nu Z \odot \boldsymbol{\beta}\|_{\psi_2} \lesssim 1 + \|\nu Z \odot \boldsymbol{\beta}\|_{\psi_2}.$$

Note that $\|\nu Z \odot \boldsymbol{\beta}\|_{\psi_2} = \sup_{\mathbf{u}\in\mathbb{S}^{p-1}} \|\nu\langle Z \odot \boldsymbol{\beta}, \mathbf{u}\rangle\|_{\psi_2} \leq \|\boldsymbol{\beta}\|_2 \cdot \big\||\nu|\big\|_{\psi_2} \leq \sigma^{-2}\|\boldsymbol{\beta}\|_2 \cdot \big\||W + \langle X, \boldsymbol{\beta}^* - (\mathbf{1} - Z)\odot\boldsymbol{\beta}\rangle|\big\|_{\psi_2} \lesssim (1+\omega)\rho + (1+\omega)^2\rho^2$. As $\zeta := (1+\omega)\rho$. We thus have $\|\boldsymbol{\mu}_{\boldsymbol{\beta}}(Y, Z, X)\|_{\psi_2} \lesssim (1+\zeta)^2$. Using Lemma 23 with the substitution $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$ and $X = \boldsymbol{\mu}_{\boldsymbol{\beta}}(Y, Z, X)$, we claim that there exist constants C_i such that

$$L_1 \geq \frac{1}{4}\|\mathbf{u}\|_2^2 - C_0(1+\zeta)^8 \frac{\log p}{n}\|\mathbf{u}\|_1^2 \text{ for all } \mathbf{u} \in \mathbb{R}^p. \tag{B.34}$$

with probability at least $1 - C_1\exp(-C_2 n(1+\zeta)^{-8})$.

Term L_2**.** We now turn to term L_2. We introduce n i.i.d. samples $\{p_i\}_{i=1}^n$ of Rademacher random variable P with $\Pr(P = 1) = \Pr(P = -1) = 1/2$. Equivalently, we have

$$L_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sigma^2 + \|\mathbf{z}_i \odot \boldsymbol{\beta}\|_2^2} \langle p_i \mathbf{z}_i \odot \boldsymbol{\beta}, \mathbf{u}\rangle^2.$$

Note that $\sqrt{(\sigma^2 + \|Z \odot \boldsymbol{\beta}\|_2^2)^{-1}} PZ \odot \boldsymbol{\beta}$ is zero mean sub-Gaussian random vector with covariance matrix $\boldsymbol{\Sigma}_2$ given in Lemma 13. Moreover, we have $\lambda_{max}(\boldsymbol{\Sigma}_2) \leq \zeta^2\epsilon \leq 1/12$, where the last inequality follows by letting $\epsilon \leq C\zeta^{-2}$ for sufficiently small C. Also note that

$$\left\|\sqrt{(\sigma^2 + \|Z \odot \boldsymbol{\beta}\|_2^2)^{-1}} PZ \odot \boldsymbol{\beta}\right\|_{\psi_2} \lesssim \sigma^{-1}\|Z \odot \boldsymbol{\beta}\|_{\psi_2} \lesssim \zeta.$$

Using Lemma 23 with substitution $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_2$ and $X = \sqrt{(\sigma^2 + \|Z \odot \boldsymbol{\beta}\|_2^2)^{-1}} PZ \odot \boldsymbol{\beta}$, we claim there exists constants C_i' such that

$$L_2 \leq \frac{1}{8}\|\mathbf{u}\|_2^2 + C_0' \max\{\zeta^4, 1\}\frac{\log p}{n}\|\mathbf{u}\|_1^2, \text{ for all } \mathbf{u} \in \mathbb{R}^p. \tag{B.35}$$

with probability at least $1 - C_1'\exp(-C_2' n \min\{\zeta^{-4}, 1\})$.

Now we put (B.34) and (B.35) together. So we obtain

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{u}^\top \boldsymbol{\Sigma}_{\boldsymbol{\beta}}(y_i, \mathbf{z}_i, \mathbf{x}_i)\mathbf{u} \geq \frac{1}{8}\|\mathbf{u}\|_2^2 - (C_0 + C_0')(1+\zeta)^8 \frac{\log p}{n}\|\mathbf{u}\|_1^2.$$

For any $\mathbf{u} \in \mathcal{C}(\mathcal{S}, \overline{\mathcal{S}}; \mathcal{R})$, we have $\|\mathbf{u}\|_1 \leq 5\sqrt{s}\|\mathbf{u}\|_2$. Consequently, when $n \geq C(1+\zeta)^8 s \log p$ for sufficiently large C, we have that, with high probability, $Q_n^{MCR}(\cdot|\boldsymbol{\beta})$ is γ_n-strongly concave over $\mathcal{C}$ with $\gamma_n = 1/9$. □

Lemma 18 (Statistical error of MCR)**.** *Consider missing covariate regression with any fixed parameter $\boldsymbol{\beta}^* \in \mathcal{B}_0(s;p)$ and $Q_n^{MCR}(\cdot|\cdot)$ given in* (B.29)*. There exist constants C_0, C_1 such that if $n \geq C_0[\log p + \log(24/\delta)]$, then for any $\delta \in (0,1)$ and any fixed $\boldsymbol{\beta} \in \mathcal{B}(\omega\|\boldsymbol{\beta}^*\|_2, \boldsymbol{\beta}^*)$, we have that for*

$$\|\nabla Q_n^{MCR}(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - Q^{MCR}(\boldsymbol{\beta}^*|\boldsymbol{\beta})\|_\infty \leq C_1(1+\zeta)^5\sigma\sqrt{\frac{\log p + \log(24/\delta)}{n}}$$

with probability at least $1-\delta$.

Proof. In this example,

$$\begin{aligned}
&\|\nabla Q_n^{MCR}(\boldsymbol{\beta}^*|\boldsymbol{\beta}) - \nabla Q^{MCR}(\boldsymbol{\beta}^*|\boldsymbol{\beta})\|_{\mathcal{R}^*} \\
&\leq \underbrace{\left\|\frac{1}{n}\sum_{i=1}^n y_i\boldsymbol{\mu}_{\boldsymbol{\beta}}(y_i,\mathbf{z}_i,\mathbf{x}_i) - \mathbb{E}\left[Y\boldsymbol{\mu}_{\boldsymbol{\beta}}(Y,Z,X)\right]\right\|_\infty}_{U_1} + \underbrace{\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\Sigma}_{\boldsymbol{\beta}}(y_i,\mathbf{z}_i,\mathbf{x}_i)\boldsymbol{\beta}^* - \mathbb{E}\left[\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(Y,Z,X)\right]\boldsymbol{\beta}^*\right\|_\infty}_{U_2}.
\end{aligned}$$

To ease notation, we let $\nu := \frac{y_i - \langle(\mathbf{1}-\mathbf{z}_i)\odot\boldsymbol{\beta}, \mathbf{x}_i\rangle}{\sigma^2 + \|\mathbf{z}_i\odot\boldsymbol{\beta}\|_2^2}$. Next we bound the term U_1 and U_2 respectively.

Term U_1. Consider one coordinate of vector $V := Y\boldsymbol{\mu}_{\boldsymbol{\beta}}(Y,Z,X)$. For any $j \in [p]$, we have

$$V_j = Y[(1-Z_j)X_j + \nu Z_j\beta_j].$$

So V_j is sub-exponential random variable since Y and $(1-Z_j)X_j + \nu Z_j\beta_j$ are both sub-Gaussians. Moreover, we have $\|Y\|_{\psi_2} \lesssim \sigma + \|\boldsymbol{\beta}^*\|_2$ and $\|(1-Z_j)X_j + \nu Z_j\beta_j\|_{\psi_2} \lesssim \|(1-Z_j)X_j\|_{\psi_2} + \|\nu Z_j\boldsymbol{\beta}_j\|_{\psi_2} \lesssim 1 + \sigma^{-2}(\sigma + \sqrt{1+\omega^2}\|\boldsymbol{\beta}^*\|_2)\|\boldsymbol{\beta}\|_2$. The last inequality follows from the fact that ν is sub-Gaussian with $\|\nu\|_{\psi_2} \lesssim \sigma^{-2}(\sigma + \sqrt{1+\omega^2}\|\boldsymbol{\beta}^*\|_2)$. We have $\|V_i\|_{\psi_1} \lesssim \|Y\|_{\psi_2}\cdot\|(1-Z_j)X_j + \nu Z_j\beta_j\|_{\psi_2} \lesssim (1+\zeta)^3\sigma$, where $\zeta := (1+\omega)\rho$. By concentration result of sub-exponentials (Lemma 20) and applying union bound, we have that there exists constant C such that for $t \lesssim (1+\zeta)^3\sigma$,

$$\Pr(U_1 \geq t) \leq pe\cdot\exp(-\frac{Cnt^2}{(1+\zeta)^6\sigma^2}).$$

Setting the right hand side to be $\delta/2$ implies that for $n \gtrsim \log p + \log(2e/\delta)$,

$$U_1 \lesssim (1+\zeta)^3\sigma\sqrt{\frac{\log p + \log(2e/\delta)}{n}} \tag{B.36}$$

with probability at least $1-\delta/2$.

Term U_2**.** Term U_2 can be further decomposed into several terms as follows

$$U_2 \leq \|\mathbf{a}_1\|_\infty + \|\mathbf{a}_2\|_\infty + \|\mathbf{a}_3\|_\infty + \|\mathbf{a}_4\|_\infty + \sigma^{-2}\|\mathbf{a}_5\|_\infty + \|\mathbf{a}_6\|_\infty,$$

where

$$\begin{aligned}
\mathbf{a}_1 &= \frac{1}{n}\sum_{i=1}^n \left\langle(\mathbf{1}-\mathbf{z}_i)\odot\mathbf{x}_i, \boldsymbol{\beta}^*\right\rangle(\mathbf{1}-\mathbf{z}_i)\odot\mathbf{x}_i - \mathbb{E}\left[\left\langle(\mathbf{1}-Z)\odot X, \boldsymbol{\beta}^*\right\rangle(\mathbf{1}-Z)\odot X\right], \\
\mathbf{a}_2 &= \frac{1}{n}\sum_{i=1}^n \left\langle\nu\mathbf{z}_i\odot\boldsymbol{\beta}, \boldsymbol{\beta}^*\right\rangle(\mathbf{1}-\mathbf{z}_i)\odot\mathbf{x}_i - \mathbb{E}\left[\left\langle\nu Z\odot\boldsymbol{\beta}, \boldsymbol{\beta}^*\right\rangle(\mathbf{1}-Z)\odot X\right], \\
\mathbf{a}_3 &= \frac{1}{n}\sum_{i=1}^n \left\langle(\mathbf{1}-\mathbf{z}_i)\odot\mathbf{x}_i, \boldsymbol{\beta}^*\right\rangle\nu\mathbf{z}_i\odot\boldsymbol{\beta} - \mathbb{E}\left[\left\langle(\mathbf{1}-Z)\odot X, \boldsymbol{\beta}^*\right\rangle\nu Z\odot\boldsymbol{\beta}\right], \\
\mathbf{a}_4 &= \frac{1}{n}\sum_{i=1}^n \nu^2\langle\mathbf{z}_i\odot\boldsymbol{\beta}, \boldsymbol{\beta}^*\rangle\mathbf{z}_i\odot\boldsymbol{\beta} - \mathbb{E}\left[\nu^2\langle Z\odot\boldsymbol{\beta}, \boldsymbol{\beta}^*\rangle Z\odot\boldsymbol{\beta}\right], \\
\mathbf{a}_5 &= \frac{1}{n}\sum_{i=1}^n \left\langle\mathbf{z}_i\odot\boldsymbol{\beta}, \boldsymbol{\beta}^*\right\rangle\mathbf{z}_i\odot\boldsymbol{\beta} - \mathbb{E}\left[\left\langle Z\odot\boldsymbol{\beta}, \boldsymbol{\beta}^*\right\rangle Z\odot\boldsymbol{\beta}\right],\ \mathbf{a}_6 = \frac{1}{n}\sum_{i=1}^n \mathrm{diag}(\mathbf{z}_i)\boldsymbol{\beta}^* - \epsilon\boldsymbol{\beta}^*.
\end{aligned}$$

The key idea to bound the infinite norm of each term $\mathbf{a}_i$ is the same: showing that each coordinate is finite summation of independent sub-Gaussian (or sub-exponential) random variables and applying

concentration result and probabilistic union bound. For each term $\mathbf{a}_i, i = 1, 2, \ldots, 6$, we have that for any $j \in [p]$,

$\|\langle(\mathbf{1} - Z) \odot X, \boldsymbol{\beta}^*\rangle(1 - Z_j) \odot X_j\|_{\psi_1} \lesssim \|\boldsymbol{\beta}^*\|_2,\ \|\langle \nu Z \odot \boldsymbol{\beta}, \boldsymbol{\beta}^*\rangle(1 - Z_j) \odot X_j\|_{\psi_1} \lesssim \sigma(1 + \zeta)\zeta^2,$
$\|\langle(\mathbf{1} - Z) \odot X, \boldsymbol{\beta}^*\rangle \nu Z_j \beta_j\|_{\psi_1} \lesssim \sigma(1 + \zeta)\zeta^2,\ \|\nu^2\langle Z \odot \boldsymbol{\beta}, \boldsymbol{\beta}^*\rangle Z_j \beta_j\|_{\psi_1} \lesssim \sigma(1 + \zeta^2)\zeta^3,$
$\sigma^{-2}\|\langle Z \odot \boldsymbol{\beta}, \boldsymbol{\beta}^*\rangle Z_j \odot \beta_j\|_{\psi_2} \lesssim \sigma\zeta^3,\ \|\epsilon\beta_j^*\|_{\psi_2} \lesssim \epsilon\|\boldsymbol{\beta}^*\|_\infty$

respectively. For simplicity, we treat coordinates of every $\mathbf{a}_i$ as finite sum of sub-exponentials with ψ_1 norm $O(\sigma(1 + \zeta)^5)$. Consequently, by concentration result in Lemma 20, there exists constant C such that

$$\Pr(U_2 \geq t) \leq 12p \cdot \exp\left(-\frac{Cnt^2}{\sigma^2(1 + \zeta)^{10}}\right)$$

for $t \lesssim \sigma(1 + \zeta)^5$. By setting the right hand side to be $\delta/2$ in the above inequality, we have that when $n \gtrsim \log p + \log(24/\delta)$,

$$U_2 \lesssim \sigma(1 + \zeta)^5\sqrt{\frac{\log p + \log(24/\delta)}{n}}. \tag{B.37}$$

with probability at least $1 - \delta/2$.

Finally, putting (B.36) and (B.37) together completes the proof. □

By setting $\delta = 1/p$ in Lemma 18 immediately implies that Q_n^{MCR} satisfies Condition 5 with parameters $\Delta_n = O\left((1 + \zeta)^5\sigma\sqrt{\log p/n}\right)$, $r = \omega\|\boldsymbol{\beta}^*\|_2$ and $\delta = 1/p$.

Putting together all the pieces leads to the following guarantee about resampling version of regularized EM on missing covariate regression.

Proof of Corollary 4. Following Theorem 1, we have $\kappa^* = 5\frac{\alpha\mu\tau}{\gamma\gamma_{n/T}}$. For ℓ_2 norm, $\alpha = 1$. Based on Lemma 17, we have $\gamma_n = 1/9$. Following Lemma 15 and 16, we have $\gamma = 1$ and can always find sufficiently small constants C_0, C_1 such that $\mu \leq 10/9$ and $\tau \leq 1/100$. We thus obtain $\kappa^* \leq 1/2$. From Lemma 18, one can check $\Delta > 3\Delta_{n/T}$ under suitable C. We choose $n/T \gtrsim \sigma^2(\omega\rho)^{-1}s\log p$ to make sure $\Delta \leq 3\overline{\Delta}$. With these conditions in hand, direct applying Theorem 1 completes the proof. □

C Supporting Lemmas

Lemma 19. *Suppose $X_1, X_2, \ldots, X_n$ are n i.i.d. centered sub-Gaussian random variables with Orlicz norm $\|X_1\|_{\psi_2} \leq K$. Then for every $t \geq 0$, we have*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq t\right) \leq e \cdot \exp\left(-\frac{Cnt^2}{K^2}\right),$$

where C is an absolute constant.

Proof. See the proof of Proposition 5.10 in [18]. □

Lemma 20. *Suppose $X_1, X_2, \ldots, X_n$ are n i.i.d. centered sub-exponential random variables with Orlicz norm $\|X_1\|_{\psi_1} \leq K$. Then for every $t > 0$, we have*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq t\right) \leq 2 \cdot \exp\left(-C\min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\}n\right),$$

where C is an absolute constant.

Proof. See the proof of Corollary 5.7 in [18]. □

Lemma 21. *Let X be sub-Gaussian random variable and Y be sub-exponential random variable. Then $X - \mathbb{E}[X]$ is also sub-Gaussian; $Y - \mathbb{E}[Y]$ is also sub-exponential. Moreover, we have*

$$\|X - \mathbb{E}[X]\|_{\psi_2} \leq 2\|X\|_{\psi_2}, \quad \|Y - \mathbb{E}[Y]\|_{\psi_1} \leq 2\|Y\|_{\psi_1}.$$

Proof. See Remark 5.18 in [18]. □

Lemma 22. *Let X, Y be two sub-Gaussian random variables. Then $Z = X \cdot Y$ is sub-exponential random variable. Moreover, there exits constant C such that*

$$\|Z\|_{\psi_1} \leq C\|X\|_{\psi_2} \cdot \|Y\|_{\psi_2}.$$

Proof. It follows from the basic properties. We omit the details. □

Lemma 23. *Let matrix $\mathbf{X}$ be an n-by-p random matrix with i.i.d. rows drawn from X, which is zero mean sub-Gaussian random vector with $\|X\|_{\psi_2} \leq K$ and covariance matrix $\mathbf{\Sigma}$. We let $\lambda_1 := \lambda_{min}(\mathbf{\Sigma}), \lambda_p := \lambda_{max}(\mathbf{\Sigma})$.*

(1) There exist constants C_i such that

$$\frac{1}{n}\|\mathbf{X}\mathbf{u}\|_2^2 \geq \frac{\lambda_1}{2}\|\mathbf{u}\|_2^2 - C_0\lambda_1 \max\left\{\frac{K^4}{\lambda_1^2}, 1\right\}\frac{\log p}{n}\|\mathbf{u}\|_1^2, \text{ for all } \mathbf{u} \in \mathbb{R}^p,$$

with probability at least $1 - C_1\exp\left(-C_2 n \min\left\{\frac{\lambda_1^2}{K^4}, 1\right\}\right)$.

(2) In Parallel, there exist constants C_i' such that

$$\frac{1}{n}\|\mathbf{X}\mathbf{u}\|_2^2 \leq \frac{3\lambda_p}{2}\|\mathbf{u}\|_2^2 + C_0'\lambda_p \max\left\{\frac{K^4}{\lambda_p^2}, 1\right\}\frac{\log p}{n}\|\mathbf{u}\|_1^2, \text{ for all } \mathbf{u} \in \mathbb{R}^p,$$

with probability at least $1 - C_1'\exp\left(-C_2' n \min\left\{\frac{\lambda_p^2}{K^4}, 1\right\}\right)$.

Proof. It follows by putting Lemma 12 and Lemma 15 in [9] together. □

Lemma 24. *Let X_1 and X_2 be independent random variables with distribution $\mathcal{N}(0,1)$. For any positive constant $C > 0$, let event $\mathcal{E} := \{C \cdot |X_2| \geq |X_1|\}$. Then we have*

(a)

$$\mathbb{E}\left[|X_1| \,\middle|\, \mathcal{E}\right] \cdot \Pr(\mathcal{E}) = \sqrt{\frac{2}{\pi}}\left[1 - \sqrt{\frac{1}{C^2+1}}\right].$$

(b)

$$\mathbb{E}\left[|X_2| \,\middle|\, \mathcal{E}\right] \cdot \Pr(\mathcal{E}) = \sqrt{\frac{2}{\pi}}\frac{C}{\sqrt{1+C^2}}.$$

(c)

$$\mathbb{E}\left[|X_1 X_2| \,\middle|\, \mathcal{E}\right] \cdot \Pr(\mathcal{E}) = \frac{2C^2}{\pi(1+C^2)}.$$

Proof. (a)

$$\mathbb{E}\left[|X_1| \,\middle|\, \mathcal{E}\right] \cdot \Pr(\mathcal{E}) = 4 \cdot \int_0^\infty \int_0^{uC} \frac{1}{2\pi}\exp(-\frac{1}{2}v^2)\exp(-\frac{u^2}{2}) v \mathrm{d}v \mathrm{d}u = \sqrt{\frac{2}{\pi}}\left[1 - \sqrt{\frac{1}{C^2+1}}\right].$$

(b)

$$\mathbb{E}\left[|X_2| \,\middle|\, \mathcal{E}\right] \cdot \Pr(\mathcal{E}) = 4 \cdot \int_0^\infty \int_{v/C}^\infty \frac{1}{2\pi}\exp(-\frac{1}{2}v^2)\exp(-\frac{u^2}{2}) u \mathrm{d}u \mathrm{d}v = \sqrt{\frac{2}{\pi}}\frac{C}{\sqrt{1+C^2}}.$$

(c)

$$\mathbb{E}\left[|X_1X_2| \,\middle|\, \mathcal{E}\right] \cdot \Pr(\mathcal{E}) = 4 \cdot \int_0^\infty \int_{v/C}^\infty \frac{1}{2\pi} \exp(-\frac{u^2}{2}) \exp(-\frac{v^2}{2}) uv \mathrm{d}u\mathrm{d}v$$
$$= \frac{2}{\pi} \int_0^\infty \exp(-\frac{C^2+1}{2}v^2) v dv = \frac{2C^2}{\pi(1+C^2)}.$$

□

Lemma 25. *Let $X \sim \mathcal{N}(0, \sigma^2)$ and Z be Rademacher random variable taking values in $\{-1, 1\}$. Moreover, X and Z are independent. Function $f(x, z; a, \gamma)$ is defined as*

$$f(x, z; a, \gamma) = \frac{x + az}{1 + \exp(-\frac{2(1+\gamma)}{\sigma^2} a(x + az))}.$$

Then for any $a \in \mathbb{R}, \gamma \in \mathbb{R}$, we have

$$\left| \mathbb{E}\left[f(X, Z; a, \gamma)\right] - \frac{a}{2} \right| \le \min\left\{ \frac{1}{2}|a\gamma| \exp(\frac{\gamma^2 a^2 - a^2}{2\sigma^2}),\ \frac{\sigma}{\sqrt{2\pi}} + |a| \right\}.$$

In the special case $\gamma = 0$, we have $\mathbb{E}\left[f(X, Z; a, \gamma)\right] = a/2$.

Proof. First note that

$$\mathbb{E}\left[f(X, Z; a, \gamma)\right] = \frac{1}{2}\mathbb{E}\left[\frac{X + a}{1 + \exp(-\frac{2(1+\gamma)}{\sigma^2} a(X + a))} + \frac{X - a}{1 + \exp(-\frac{2(1+\gamma)}{\sigma^2} a(X - a))} \right]$$
$$= \frac{1}{2}\mathbb{E}\left[\frac{X + a}{1 + \exp(-\frac{2(1+\gamma)}{\sigma^2} a(X + a))} + \frac{-X - a}{1 + \exp(-\frac{2(1+\gamma)}{\sigma^2} a(-X - a))} \right],$$

where the first equality is from taking expectation of Z, the second equality is from the fact that the distribution of X is symmetric around 0. Let $X' = X + a$, then we have

$$\mathbb{E}\left[f(X, Z; a, \gamma)\right] = \frac{1}{2}\mathbb{E}\left[\frac{X'}{1 + \exp(-\frac{2(1+\gamma)}{\sigma^2} aX')} + \frac{-X'}{1 + \exp(\frac{2(1+\gamma)}{\sigma^2} aX')} \right]$$
$$= \frac{1}{2}\mathbb{E}\left[X' - 2\frac{\exp(-\frac{2(1+\gamma)}{\sigma^2} aX')X'}{1 + \exp(-\frac{2(1+\gamma)}{\sigma^2} aX')} \right].$$

Using $\mathbb{E}\left[X'\right] = a$, we have

$$\mathbb{E}\left[f(X, Z; a, \gamma)\right] - a/2 = \mathbb{E}\left[-\frac{\exp(-\frac{2(1+\gamma)}{\sigma^2} aX')X'}{1 + \exp(-\frac{2(1+\gamma)}{\sigma^2} aX')} \right]$$
$$= \int_{-\infty}^\infty \frac{\exp(-\frac{(x-a)^2}{2\sigma^2})}{\sqrt{2\pi}\sigma} \frac{-\exp(-\frac{2(1+\gamma)}{\sigma^2} ax)x}{1 + \exp(-\frac{2(1+\gamma)}{\sigma^2} ax)} \mathrm{d}x = \int_{-\infty}^\infty \frac{\exp(-\frac{x^2+a^2}{2\sigma^2})x}{\sqrt{2\pi}\sigma} \frac{-\exp(-\frac{\gamma ax}{\sigma^2})}{\exp(\frac{a(1+\gamma)x}{\sigma^2}) + \exp(\frac{-a(1+\gamma)x}{\sigma^2})} \mathrm{d}x$$
$$= \int_0^\infty \frac{\exp(-\frac{x^2+a^2}{2\sigma^2})x}{\sqrt{2\pi}\sigma} \frac{\exp(\frac{\gamma ax}{\sigma^2}) - \exp(-\frac{\gamma ax}{\sigma^2})}{\exp(\frac{a(1+\gamma)x}{\sigma^2}) + \exp(\frac{-a(1+\gamma)x}{\sigma^2})} \mathrm{d}x \tag{C.1}$$

When $a\gamma \ge 0$, we have $\mathbb{E}\left[f(X, Z; a, \gamma)\right] - a/2 \ge 0$. Under this setting, (C.1) yields that

$$\mathbb{E}\left[f(X, Z; a, \gamma)\right] - a/2 \le \int_0^\infty \frac{\exp(-\frac{x^2+a^2}{2\sigma^2})x}{2\sqrt{2\pi}\sigma} \left[\exp(\frac{\gamma ax}{\sigma^2}) - \exp(-\frac{\gamma ax}{\sigma^2})\right] \mathrm{d}x$$
$$= \frac{1}{2}\exp(\frac{\gamma^2 a^2 - a^2}{2\sigma^2}) \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} \left[\exp\left(-\frac{(x - \gamma a)^2}{2\sigma^2}\right) - \exp\left(-\frac{(x + \gamma a)^2}{2\sigma^2}\right) \right] x \mathrm{d}x$$
$$= \frac{1}{2}\exp(\frac{\gamma^2 a^2 - a^2}{2\sigma^2}) \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \gamma a)^2}{2\sigma^2}\right) x \mathrm{d}x = \frac{1}{2}\exp(\frac{\gamma^2 a^2 - a^2}{2\sigma^2})\gamma a,$$

	GMM	MLR(sparse)	MLR(low rank)	MCR
Δ	$0.1(\|\boldsymbol{\beta}^*\|_\infty + \sigma)\sqrt{\frac{\log p}{n}}$	$0.1(\|\boldsymbol{\beta}^*\|_2 + \sigma)\sqrt{\frac{\log p}{n}}$	$0.01(\|\boldsymbol{\Gamma}^*\|_F + \sigma)\sqrt{\frac{p_1+p_2}{n}}$	$0.2\sigma\sqrt{\frac{\log p}{n}}$

Table 1: Choice of parameter Δ in Algorithm 1.

where the first inequality follows from the fact that $x + 1/x \geq 2$ for any $x > 0$, the second equality is from

$$-\int_0^\infty \exp\left(-\frac{(x+\gamma a)^2}{2\sigma^2}\right) x\mathrm{d}x = \int_{-\infty}^0 \exp\left(-\frac{(x-\gamma a)^2}{2\sigma^2}\right) x\mathrm{d}x.$$

When $a\gamma \leq 0$, using similar proof, we have $\frac{1}{2}\exp(\frac{\gamma^2 a^2 - a^2}{2\sigma^2})\gamma a \leq \mathbb{E}\left[f(X, Z; a, \gamma)\right] - a/2 \leq 0$. Combining the two cases, we prove that

$$\left|\mathbb{E}\left[f(X, Z; a, \gamma)\right] - a/2\right| \leq \frac{1}{2}|a\gamma| \exp(\frac{\gamma^2 a^2 - a^2}{2\sigma^2}). \tag{C.2}$$

In the special case when $\gamma = 0$, we thus have $\mathbb{E}(f(X, Z; a, \gamma)) = a/2$.

Note that when $a\gamma \geq 0$, (C.1) also implies that

$$\mathbb{E}\left[f(X, Z; a, \gamma)\right] - a/2 \leq \int_0^\infty \frac{\exp(-\frac{x^2+a^2}{2\sigma^2})x}{\sqrt{2\pi}\sigma} \frac{\exp(\frac{\gamma a x}{\sigma^2})}{\exp(\frac{a(1+\gamma)x}{\sigma^2})}\mathrm{d}x = \int_0^\infty \frac{\exp(-\frac{(x+a)^2}{2\sigma^2})x}{\sqrt{2\pi}\sigma}\mathrm{d}x$$

$$= \int_0^\infty \frac{\exp(-\frac{(x+a)^2}{2\sigma^2})(x+a)}{\sqrt{2\pi}\sigma}\mathrm{d}x - \int_0^\infty \frac{\exp(-\frac{(x+a)^2}{2\sigma^2})a}{\sqrt{2\pi}\sigma}\mathrm{d}x \leq \frac{\sigma}{\sqrt{2\pi}} + |a|.$$

Similarly, when $a\gamma \leq 0$, we have

$$\mathbb{E}\left[f(X, Z; a, \gamma)\right] - a/2 \geq \int_0^\infty \frac{\exp(-\frac{x^2+a^2}{2\sigma^2})x}{\sqrt{2\pi}\sigma} \frac{-\exp(\frac{-\gamma a x}{\sigma^2})}{\exp(\frac{-a(1+\gamma)x}{\sigma^2})}\mathrm{d}x = -\int_0^\infty \frac{\exp(-\frac{(x-a)^2}{2\sigma^2})x}{\sqrt{2\pi}\sigma}\mathrm{d}x$$

$$= -\int_0^\infty \frac{\exp(-\frac{(x-a)^2}{2\sigma^2})(x-a)}{\sqrt{2\pi}\sigma}\mathrm{d}x - \int_0^\infty \frac{\exp(-\frac{(x-a)^2}{2\sigma^2})a}{\sqrt{2\pi}\sigma}\mathrm{d}x \geq -\frac{\sigma}{\sqrt{2\pi}} - |a|.$$

Therefore, we have that

$$\left|\mathbb{E}\left[f(X, Z; a, \gamma)\right] - a/2\right| \leq \frac{\sigma}{\sqrt{2\pi}} + |a|. \tag{C.3}$$

Putting (C.2) and (C.3) together completes the proof. □

D Additional Experiment Setting

In our simulations, parameter Δ for each model is set according to Table 1.