Noname manuscript No.
(will be inserted by the editor)

Robust High Dimensional Expectation Maximization Algorithm via Trimmed Hard Thresholding

Di Wang · Xiangyu Guo ⋆ · Shi Li · Jinhui Xu

Received: date / Accepted: date

Abstract In this paper, we study the problem of estimating latent variable models with arbitrarily corrupted samples in high dimensional space (*i.e.,* $d \gg n$) where the underlying parameter is assumed to be sparse. Specifically, we propose a method called Trimmed (Gradient) Expectation Maximization which adds a trimming gradients step and a hard thresholding step to the Expectation step (E-step) and the Maximization step (M-step), respectively. We show that under some mild assumptions and with an appropriate initialization, the algorithm is corruption-proofing and converges to the (near) optimal statistical rate geometrically when the fraction of the corrupted samples ϵ is bounded by $\tilde{O}(\frac{1}{\sqrt{n}})$. Moreover, we apply our general framework to three canonical models:

⋆ The first two authors contributed equally.

Di Wang
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia
E-mail: dwang45@bufffalo.edu

Xiangyu Guo
Department of Computer Science and Engineering
State University of New York at Buffalo, Buffalo
NY, USA 14260
xiangyug@buffalo.edu

Shi Li
Department of Computer Science and Engineering
State University of New York at Buffalo, Buffalo
NY, USA 14260
E-mail: shil@bufffalo.edu

Jinhui Xu
Department of Computer Science and Engineering
Buffalo, NY, USA 14260
State University of New York at Buffalo
E-mail: jinhui@bufffalo.edu
Corresponding author

mixture of Gaussians, mixture of regressions and linear regression with missing covariates. Our theory is supported by thorough numerical results.

Keywords Robust Statistics · High Dimensional Statistics · Gaussian Mixture Model · Expectation Maximixation · Iterative Hard Thresholding

1 Introduction

As one of the most popular techniques for estimating the maximum likelihood of mixture models or incomplete data problems, Expectation Maximization (EM) algorithm has been widely applied to many areas such as genomics (Laird, 2010), finance (Faria and Gonçalves, 2013), and crowdsourcing (Dawid and Skene, 1979). Although EM algorithm is well-known to converge to an empirically good local estimator (Wu et al., 1983), finite sample statistical guarantees for its performance have not been established until recent studies (Balakrishnan et al., 2017b)(Zhu et al., 2017),(Wang et al., 2015),(Yi and Caramanis, 2015). Specifically, the first local convergence theory and finite sample statistical rate of convergence for the classical EM and its gradient ascent variant (gradient EM) were established in (Balakrishnan et al., 2017b). Later, (Wang et al., 2015) extended the classical EM and gradient EM algorithms to the high dimensional sparse setting, and the key idea in their methods is an additional truncation step after the M-step, which can exploit the intrinsic sparse structure of the high dimensional latent variable models. Later on, (Yi and Caramanis, 2015) also studied the high dimensional sparse EM algorithm and proposed a method which uses a regularized M-estimator in the M-step. Recently, (Zhu et al., 2017) considered the computational issue of the previous methods of the problem in high dimensional sparse case. They proposed a method called VRSGEM which combines the idea of SVRG (Johnson and Zhang, 2013) and the high dimensional gradient EM algorithm. Their method has less gradient complexity while also can achieve almost the same statistical estimation errors as the previous ones.

Although the above methods could achieve (near) optimal minimax rate for some statistical models such as Gaussian mixture model, mixture of regressions and linear regression with missing covariates (see Preliminaries section for details), all of these results need to assume that the data samples have no corruptions and also should satisfy some statistical assumptions, such as sub-Gaussian. This means that some arbitrary corruptions among the data samples may cause the dataset violate these statistical assumptions which are required for convergence of the above methods, or they will even make the above methods achieve unacceptable statistical estimation errors (see Figure 1 for experimental studies). Thus, the classical EM algorithm and its variants are sensitive to these corruptions. Although statistical estimation with arbitrary corruptions has long been a focus in robust statistics (Huber, 2011), it is still unknown that **whether there exist some variant of (gradient) EM algorithm which is robust to arbitrary corruptions while also has finite sample statistical guarantees as in the non-corrupted case**.

To address the aforementioned issue, in this paper, we study the problem of statistical estimation of latent variable models with arbitrarily corrupted samples in high dimensional space[1] (*i.e.,* $d \gg n$) where the underlying parameter is assumed to be sparse. Specifically, we propose a new algorithm called Trimmed (Gradient) Expectation Maximization, which attaches a trimming gradient and hard thresholding step to the E-step and M-step in each iteration, respectively. We show that under certain conditions, our algorithm is robust against corruption and converges with a statistical estimation error which is (near) statistically optimal. Below is a summary of our main contributions.

1. We show that, given an appropriate initialization β^{init}, *i.e.,* $\|\beta^{\text{init}} - \beta^*\| \leq \kappa\|\beta^*\|_2$ for some constant $\kappa \in (0, 1)$, if the model satisfies some additional assumptions, the iterative solution sequence β^t satisfies $\|\beta^t - \beta^*\|_2 \leq \tilde{O}\big(c_1\rho^t + \sqrt{s^*}c_2(\epsilon\log(nd) + \sqrt{\frac{\log d}{n}})\big)$ with high probability, where $\rho \in (0, 1)$, c_1, c_2 are some constants dependent on the model, ϵ is the fraction of the perturbed samples, and s^* is the sparsity parameter of the underlying parameter β^*. Particularly, when c_2 is a constant and $\epsilon \leq O(\frac{1}{\sqrt{n}\log(nd)})$, the above estimation error geometrically converges to $O(\sqrt{\frac{s^*\log d}{n}})$, which is statistically optimal. This means that our algorithm is corruption-proofing for a certain level of corruption that is only dependent on the sample size, which is quite useful in the high dimensional setting.
2. We implement our algorithm on three canonical models: mixture of Gaussians, mixture of regressions and linear regression with missing covariates. Experimental results on these models support our theoretical analysis.

Due to the space limit, some background, lemmas and all the proofs are included in the Appendix.

2 Related Work

There are mainly two perspectives on the study of EM algorithm. The first one focus on its statistical guarantees (Balakrishnan et al., 2017b; Zhu et al., 2017; Wang et al., 2015; Yi and Caramanis, 2015). However, as we mentioned above, although in this paper we study the same statistical setting as these previous work, our method is corruption-proofing. Another direction focus on the practical performance, and there are many robust variants of the EM algorithm such as (Aitkin and Wilson, 1980; Yang et al., 2012). However, we note that these methods are incomparable with ours. Firstly, in this paper we mainly focus on statistical setting and the statistical guarantees while there is no any theoretical guarantees of these methods. Secondly, previous methods can only be used in the low dimension case while we focus on the high dimensional sparse case. Thus, to our best knowledge, there is no previous work on the

[1] Since high dimensional sparse case is much more harder than the low dimension case, our algorithm can be easily extended to the low dimension case by using the results in (Balakrishnan et al., 2017b). Due to the space limit, we omit it in the paper.

variants of the EM algorithm that is both robust to some corruptions and also has statistical guarantees. Thus, in the following we will only compare with some other methods that are close to ours.

(Diakonikolas et al., 2016, 2018, 2017; Chen et al., 2013) studied the problem of robustly estimating the mixture of distributions. However, some of them are not computationally practical as they rely on the rather time-consuming ellipsoid method. Moreover, these methods in general cannot be extended to the distributed or Byzantine setting (Chen et al., 2017), while ours can be easily extended to such scenarios.

(Du et al., 2017; Balakrishnan et al., 2017a; Li, 2017; Suggala et al., 2019; Dalalyan and Thompson, 2019; Thompson and Dalalyan, 2018) studied the robust high dimensional sparse estimation problem for some specified tasks, such as GLM, linear regression, mean and covariance matrix estimation. However, none of them considered estimating the latent variable models and thus is quite different from ours.

Recently, several robust methods have been proposed based on (stochastic) gradient descent, such as (Alistarh et al., 2018; Chen et al., 2017; Yin et al., 2018; Prasad et al., 2018; Holland, 2018). However, none of them studies the latent variable models and all of them consider only the low dimensional case. (Liu et al., 2019) recently investigated the robust high dimensional sparse M-estimation problem by combining hard thresholding with trimming steps. However, their method can only be used in the M-estimation, while ours focuses on the latent variable model and the EM algorithm.

3 Preliminaries

Let Y and Z be two random variables taking values in the sample spaces $\mathcal{Y}$ and $\mathcal{Z}$, respectively. Suppose that the pair (Y, Z) has a joint density function f_{β^*} that belongs to some parameterized family $\{f_{\beta^*}|\beta^* \in \Omega\}$. Rather than considering the whole pair of (Y, Z), we observe only component Y. Thus, component Z can be viewed as the missing or latent structure. We assume that the term $h_\beta(y)$ is the margin distribution over the latent variable Z, *i.e.,* $h_\beta(y) = \int_{\mathcal{Z}} f_\beta(y, z)dz$. Let $k_\beta(z|y)$ be the density of Z conditional on the observed variable $Y = y$, that is, $k_\beta(z|y) = \frac{f_\beta(y,z)}{h_\beta(y)}$.

Given n observations $y_1, y_2, \cdots, y_n$ of Y, the EM algorithm is to maximize the log-likelihood $\max_{\beta\in\Omega} \ell_n(\beta) = \sum_{i=1}^n \log h_\beta(y_i)$. Due to the unobserved latent variable Z, it is often difficult to directly evaluate $\ell_n(\beta)$. Thus, we consider the lower bound of $\ell_n(\beta)$. By Jensen's inequality, we have

$$\frac{1}{n}[\ell_n(\beta) - \ell_n(\beta')] \geq \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{Z}} k_{\beta'}(z|y_i) \log f_\beta(y_i, z)dz$$
$$- \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{Z}} k_{\beta'}(z|y_i) \log f_{\beta'}(y_i, z)dz. \quad (1)$$

Let $Q_n(\beta;\beta') = \frac{1}{n}\sum_{i=1}^n q_i(\beta;\beta')$, where

$$q_i(\beta;\beta') = \int_{\mathcal{Z}} k_{\beta'}(z|y_i)\log f_\beta(y_i,z)dz. \tag{2}$$

Also, it is convenient to let $Q(\beta;\beta')$ denote the expectation of $Q_n(\beta;\beta')$ w.r.t $\{y_i\}_{i=1}^n$, that is,

$$Q(\beta;\beta') = \mathbb{E}_{y\sim h_{\beta^*}}\int_{\mathcal{Z}} k_{\beta'}(z|y)\log f_\beta(y,z)dz. \tag{3}$$

We can see that the second term on the right hand side of (1) is not dependent on β. Thus, given some fixed β', we can maximize the lower bound function $Q_n(\beta;\beta')$ over β to obtain sufficiently large $\ell_n(\beta) - \ell_n(\beta')$. Thus, in the t-th iteration of the standard EM algorithm, we can evaluate $Q_n(\cdot;\beta^t)$ at the E-step and then perform the operation of $\max_{\beta\in\Omega} Q_n(\beta;\beta^t)$ at the M-step. See (McLachlan and Krishnan, 2007) for more details.

In addition to the exact maximization implementation of the M-step, we add a gradient ascent implementation of the M-step, which performs an approximate maximization via a gradient descent step.

Gradient EM Procedure (Balakrishnan et al., 2017b) When $Q_n(\cdot;\beta^t)$ is differentiable, the update of β^t to β^{t+1} consists of the following two steps.

- E-step: Evaluate the functions in (2) to compute $Q_n(\cdot;\beta^t)$.
- M-step: Update $\beta^{t+1} = \beta^t + \eta\nabla Q_n(\beta^t;\beta^t)$, where ∇ is the derivative of Q_n w.r.t the first component and η is the step size.

Next, we give some examples that use the gradient EM algorithm. Note that they are the typical examples for studying the statistical property of EM algorithm (Wang et al., 2015; Balakrishnan et al., 2017b; Yi and Caramanis, 2015; Zhu et al., 2017).

Gaussian Mixture Model Let $y_1,\cdots,y_n$ be n i.i.d samples from $Y\in\mathbb{R}^d$ with

$$Y = Z\cdot\beta^* + V, \tag{4}$$

where Z is a Rademacher random variable (*i.e.,* $\mathbb{P}(Z=+1) = \mathbb{P}(Z=-1) = \frac{1}{2}$), and $V\sim\mathcal{N}(0,\sigma^2 I_d)$ is independent of Z for some known standard deviation σ. In our high dimensional setting, we assume that $\|\beta^*\|_0 = s^*$ is sparse.

For Gaussian Mixture Model, we have

$$\nabla q_i(\beta;\beta) = [2w_\beta(y_i) - 1]\cdot y_i - \beta, \tag{5}$$

where $w_\beta(y) = \frac{1}{1+\exp(-\langle\beta,y\rangle/\sigma^2)}$.

Mixture of (Linear) Regressions Model Let n samples (x_1,y_1), (x_2,y_2), $\cdots,(x_n,y_n)$ i.i.d sampled from $Y\in\mathbb{R}$ and $X\in\mathbb{R}^d$ with

$$Y = Z\langle\beta^*,X\rangle + V, \tag{6}$$

where $X \sim \mathcal{N}(0, I_d)$, $V \sim \mathcal{N}(0, \sigma^2)$, Z is a Rademacher random variable, and X, V, Z are independent. In the high dimensional case, we assume that $\|\beta^*\|_0 = s^*$ is sparse.

In this case, we have

$$\nabla q_i(\beta; \beta) = (2w_\beta(x_i, y_i) - 1) \cdot y_i \cdot x_i - x_i x_i^T \cdot \beta, \tag{7}$$

where $w_\beta(x_i, y_i) = \frac{1}{1+\exp(-y\langle\beta, x\rangle/\sigma^2)}$.

Linear Regression with Missing Covariates We assume that $Y \in \mathbb{R}$ and $X \in \mathbb{R}^d$ satisfy

$$Y = \langle X, \beta^* \rangle + V, \tag{8}$$

where $X \sim \mathcal{N}(0, I_d)$ and $V \sim \mathcal{N}(0, \sigma^2)$ are independent. In our high dimensional setting, we assume that $\|\beta^*\|_0 = s^*$ is sparse. Let $x_1, x_2, \cdots, x_n$ be n observations of X with each coordinate of x_i missing (unobserved) independently with probability $p_m \in [0, 1)$.

In this case, we have

$$\nabla q_i(\beta; \beta) = y_i \cdot m_\beta(x_i^{\text{obs}}, y_i) - K_\beta(x_i^{\text{obs}}, y_i)\beta, \tag{9}$$

where the functions $m_\beta(x_i^{\text{obs}}, y_i) \in \mathbb{R}^d$ and $K_\beta(x_i^{\text{obs}}, y_i) \in \mathbb{R}^{d\times d}$ are defined as:

$$m_\beta(x_i^{\text{obs}}, y_i) = z_i \odot x_i + \frac{y_i - \langle \beta, z_i \odot x_i \rangle}{\sigma^2 + \|(1 - z_i) \odot \beta\|_2^2}(1 - z_i) \odot \beta \tag{10}$$

and

$$\begin{aligned} K_\beta(x_i^{\text{obs}}, y_i) = \text{diag}(1 - z_i) + m_\beta(x_i^{\text{obs}}, y_i) \cdot [m_\beta(x_i^{\text{obs}}, y_i)]^T \\ - [(1 - z_i) \odot m_\beta(x_i^{\text{obs}}, y_i)] \cdot [(1 - z_i) \odot m_\beta(x_i^{\text{obs}}, y_i)]^T, \end{aligned} \tag{11}$$

where vector $z_i \in \mathbb{R}^d$ is defined as $z_{i,j} = 1$ if $x_{i,j}$ is observed and $z_{i,j} = 0$ is $x_{i,j}$ is missing, and $\odot$ denotes the Hadamard product of matrices.

Next, we provide several definitions on the required properties of functions $Q_n(\cdot; \cdot)$ and $Q(\cdot; \cdot)$. Note that some of them have been used in the previous studies on EM (Balakrishnan et al., 2017b; Wang et al., 2015; Zhu et al., 2017).

Definition 1 Function $Q(\cdot; \beta^*)$ is self-consistent if $\beta^* = \arg\max_{\beta\in\Omega} Q(\beta; \beta^*)$. That is, β^* maximizes the lower bound of the log likelihood function.

Definition 2 (Lipschitz-Gradient-2$(\gamma, \mathcal{B})$) $Q(\cdot; \cdot)$ is called Lipschitz-Gradient-2$(\gamma, \mathcal{B})$, if for the underlying parameter β^* and any $\beta \in \mathcal{B}$ for some set $\mathcal{B}$, the following holds

$$\|\nabla Q(\beta; \beta^*) - \nabla Q(\beta; \beta)\|_2 \leq \gamma\|\beta - \beta^*\|_2. \tag{12}$$

We note that there are some differences between the definition of Lipschitz-Gradient-2 and the Lipschitz continuity condition in the convex optimization literature (Nesterov, 2013). Firstly, in (12), the gradient is w.r.t the second component, while the Lipschitz continuity is w.r.t the first component. Secondly, the property holds only for fixed β^* and any β, while the Lipschitz continuity is for all $\beta, \beta' \in \mathcal{B}$.

Definition 3 (μ-smooth) $Q(\cdot;\beta^*)$ is μ-smooth, that is if for any $\beta, \beta' \in \mathcal{B}$, $Q(\beta;\beta^*) \geq Q(\beta';\beta^*) + (\beta-\beta')^T\nabla Q(\beta';\beta^*) - \frac{\mu}{2}\|\beta'-\beta\|_2^2$.

Definition 4 (υ-strongly concave) $Q(\cdot;\beta^*)$ is υ-strongly concave, that is if for any $\beta, \beta' \in \mathcal{B}$, $Q(\beta;\beta^*) \leq Q(\beta';\beta^*) + (\beta-\beta')^T\nabla Q(\beta';\beta^*) - \frac{\upsilon}{2}\|\beta'-\beta\|_2^2$.

Next, we assume that each coordinate of $\nabla q(\beta;\beta)$ in (2) is sub-exponential for every $\beta \in \mathcal{B}$, where ∇ is the derivative of q w.r.t the first component.

Definition 5 (ξ-sub-exponential) A random variable X with mean $\mathbb{E}(X)$ is ξ-sub-exponential for $\xi > 0$ if for all $|t| < \frac{1}{\xi}$, $\mathbb{E}\{\exp(t[X-\mathbb{E}(X)])\} \leq \exp(\frac{\xi^2 t^2}{2})$.

Assumption 1. We assume that $Q(\cdot;\cdot)$ in (3) is self-consistent, Lipschitz-Gradient-2$(\gamma,\mathcal{B})$, μ-smooth and υ-strongly convex for some $\mathcal{B}$. Moreover, we assume that for any fixed $\beta \in \mathcal{B}$ with $\|\beta\|_0 \leq s$ (where the value of s will be specified later) and $\forall j \in [d]$, the j-th coordinate of $\nabla q(\beta;\beta)$ (*i.e.,* $[\nabla q(\beta;\beta)]_j$) is ξ-sub-exponential and for each $i \in [n]$, $[\nabla q_i(\beta,\beta)]_j$ is independent with others.

We note that the sub-exponential assumption on each coordinate is stronger than the assumption of Statistical-Error in (Wang et al., 2015; Balakrishnan et al., 2017b). However, since the model considered in this paper could have arbitrarily corrupted samples, we will see later that this assumption is necessary.

Finally, we give the definition of the corruption model studied in the paper.

Definition 6 (ϵ-corrupted samples) Let $\{y_1, y_2, \cdots, y_n\}$ be n i.i.d observations with distribution P. We say that a collection of samples $\{z_1, z_2, \cdots, z_n\}$ is ϵ-corrupted if an adversary chooses an arbitrary ϵ-fraction of the samples in $\{y_i\}_{i=1}^n$ and modifies them with arbitrary values.

4 Trimmed Expectation Maximization Algorithm

To obtain a robust estimator for the high dimensional model with ϵ-corrupted samples, we propose a trimmed EM algorithm, which is based on the gradient EM algorithm. See Algorithm 1 for details.

Note that compared with the previous gradient EM algorithm, Trimmed EM algorithm has two additional steps in each iteration, *i.e.,* the trimming gradient and hard thresholding step. For the trimming gradient step 4 in Algorithm 1, we use the dimensional α-trimmed estimator on the gradients $\{\nabla q_i(\beta^t;\beta^t)\}_{i=1}^n$. We note that while this operator has also been studied in (Liu et al., 2019; Yin et al., 2018) for the M-estimators, we use it for the EM algorithm.

Definition 7 (Dimensional α-trimmed estimator) Given a set of ϵ-corrupted samples in the form of d-dimensional vectors $\{z_i\}_{i=1}^n$, the D-Trim operator $\text{D-Trim}_\alpha(\{z_i\}_{i=1}^n) \in \mathbb{R}^d$ performs as follows. For each dimension $j \in [d]$, it first

Algorithm 1 Trimmed (Gradient) Expectation Maximization

Input: T is the iteration number, β^{init} is the initial parameter, η is the flexed step-size and s is the sparsity parameter to be specified later. $\{z_i\}_{i=1}^n$ are the ϵ corrupted samples of $\{y_i\}_{i=1}^n$.

1: Let $\hat{\mathcal{S}}^{\text{init}} = \text{supp}(\beta^{\text{init}}, s)$, $\beta^0 = \text{trunc}(\beta^{\text{init}}, \hat{\mathcal{S}}^{\text{init}})$.
2: **for** $t = 0, 1, \cdots, T-1$ **do**
3: **E-step:** Evaluate $\{\nabla q_i(\beta^t; \beta^t)\}_{i=1}^n$ for $\{z_i\}_{i=1}^n$.
4: **Trimming step:** Use a dimensional α-trimmed gradient estimator to get a vector $\nabla \tilde{Q}_n(\beta^t; \beta^t) = \text{D-Trim}_\alpha(\{\nabla q_i(\beta^t; \beta^t)\}_{i=1}^n)$.
5: **M-step:** Update $\beta^{t+0.5} = \beta^t + \eta \nabla \tilde{Q}_n(\beta^t; \beta^t)$.
6: **Thresholding step:** Let $\hat{\mathcal{S}}^{t+0.5} = \text{supp}(\beta^{t+0.5}, s)$ and $\beta^{t+1} = \text{trunc}(\beta^{t+0.5}, \hat{\mathcal{S}}^{t+0.5})$
7: **end for**
8: Return β^T.

removes the largest and the smallest α fraction of elements in the j-the coordinate of $\{z_i\}_{i=1}^n$, *i.e.,* $\{z_{i,j}\}_{i=1}^n$, and then calculates the mean of the remaining terms, where $\alpha = c_0\epsilon$ and $\alpha \leq \frac{1}{2} - c_1$ for some constant $c_0 \geq 1$ and a small constant c_1.

The rationale behind the use of the dimensional trimmed estimator is that due to the existence of ϵ fraction of corrupted samples, directly calculating the the mean of the gradient could introduce a large error to the population gradient $\nabla Q(\beta^t; \beta^t)$ in (3). Also, it can be shown that if each coordinate of $\nabla q_i(\beta^t; \beta^t)$ is sub-exponential, it will be robust against the ϵ-corruption for some small ϵ. This motivates us to use the dimensional trimmed operation.

To ensure the sparsity of our estimator, after getting $\beta^{t+0.5}$, we need to use the hard thresholding operation (Blumensath and Davies, 2009). More specifically, we first find the set $\hat{\mathcal{S}}^{t+0.5} \subseteq [d]$ of indices j corresponding to the top s largest $|\beta_j^{t+0.5}|$ (we denote $\hat{\mathcal{S}}^{t+0.5} = \text{supp}(\beta^{t+0.5}, s)$), and make the value of the remaining entries $\beta_j^{t+0.5}$ for $j \in [d] \backslash \hat{\mathcal{S}}^{t+0.5}$ be 0 (we denote $\beta^{t+1} = \text{trunc}(\beta^{t+0.5}, \hat{\mathcal{S}}^{t+0.5})$). The sparsity level s controls the sparsity of the estimator and the estimation error.

The following main theorem shows that under Assumption 1 and with some proper initial vector β^{init}, the estimator β^T converges to the underlying β^* at a geometric rate with high probability.

Theorem 1 *Let $\mathcal{B} = \{\beta : \|\beta - \beta^*\|_2 \leq R\}$ be a set with $R = k\|\beta^*\|_2$ for some $k \in (0, 1)$. Assume that Assumption 1 holds for parameters $\mathcal{B}, \gamma, \mu, \upsilon, \xi$ satisfying the condition of $1 - 2\frac{\upsilon - \gamma}{\upsilon + \mu} \in (0, 1)$ and the sparsity parameter s is chosen to be*

$$s = \lceil C \max\{\frac{16}{\{1/[1 - 2(\upsilon - \gamma)/(\upsilon + \mu)] - 1\}^2}, \frac{4(1+k)^2}{(1-k)^2}\} s^* \rceil, \qquad (13)$$

where C is some absolute constant. Also, assume that $\|\beta^{init} - \beta^*\|_2 \leq \frac{R}{2}$ and there exist some absolute constants C_1 and C_2 satisfying the condition of

$$\frac{1}{\upsilon+\mu} C_2(\sqrt{s} + \frac{C_1\sqrt{s^*}}{\sqrt{1-k}})\xi(\epsilon\log(nd) + \sqrt{\frac{\log d}{n}}) \leq \min\big\{\big(1 - \sqrt{1 - \frac{2(\upsilon-\gamma)}{\upsilon+\mu}}\big)^2 R, \frac{(1-k)^2}{2(1+k)}\|\beta^*\|_2\big\}. \quad (14)$$

Then, if taking $\eta = \frac{2}{\upsilon+\mu}$ in Algorithm 1, the following holds for $t = 1, \ldots, T$ with probability at least $1 - Td^{-3}$

$$\|\beta^t - \beta^*\|_2 \leq \underbrace{(1 - 2\frac{\upsilon-\gamma}{\upsilon+\mu})^{\frac{t}{2}} R}_{\textit{Optimization Error}} + \underbrace{\frac{2C_2\xi(\epsilon\log(nd) + \sqrt{\frac{\log d}{n}})}{\upsilon+\mu} \frac{\sqrt{s} + \frac{C_1}{\sqrt{1-k}}\sqrt{s^*}}{1 - \sqrt{1 - 2\frac{\upsilon-\gamma}{\upsilon+\mu}}}}_{\textit{Statistical and Corruption Error}}. \quad (15)$$

In the above theorem, assumption (13) indicates that the sparsity level s in Algorithm 1 should be sufficiently large but still in the same order as the underlying sparsity s^*. Although s seems quite complex, in the experiments, we can see that it is suffcient to set $s = s^*$. Assumption (14) suggests that in order to ensure an upper bound in the hard thresholding step, we need $\sqrt{s^*}\xi(\epsilon\log(nd) + \sqrt{\frac{\log d}{n}} \leq O(\|\beta^*\|_2)$, which means that n should be sufficiently large and the fraction of corruption ϵ cannot be too large. In the error bound of (15), there are three types of errors. The first one is caused by optimization, which decreases to zero at a geometric rate of convergence. The second one is the term related to ϵ (*i.e.,* $O(\xi\sqrt{s^*}\epsilon\log(nd))$), which is caused by estimating the population gradient via the trimming step due the ϵ-corrupted samples. In the special case of no corrupted samples (*i.e.,* $\epsilon = 0$), the bound will be zero. The third one is the term $O(\xi\sqrt{\frac{s^*\log d}{n}})$, which corresponds to the statistical error. It is independent of both ϵ and t and only dependent on the model itself. Even though Theorem 1 requires that the initial estimator be close enough to the optimal one, our experiments show that the algorithm actually performs quite well for any random initialization.

From Theorem 1, we can also see that when the fraction of corruption ϵ is sufficiently small such that $\epsilon \leq O(\frac{1}{\sqrt{n\log(nd)}})$ and the iteration number is sufficiently large, the error bound in (15) becomes $O(\xi\sqrt{\frac{s^*\log d}{n}})$, which is the same as the optimal rate of estimating a high dimensional sparse vector when ξ is some constant. This means that our method has the same rate as the non-corrupted ones in (Wang et al., 2015). This rate of corruption also has been appeared in the corrupted sparse linear regression (Dalalyan and Thompson, 2019; Liu et al., 2019). Also, we can see that when $\alpha = 0$, our algorithm will be reduced to the high dimensional gradient EM algorithm in (Wang et al., 2015).

5 Implications for Some Specific Models

In this section, we apply our framework (*i.e.,* Algorithm 1) to the models mentioned in Section 3. To obtain results for these models, we only need to find the corresponding $\mathcal{B}, \gamma, k, R, \upsilon, \mu, \xi$ to ensure that Assumption 1 and assumptions in Theorem 1 hold.

5.1 Corrupted Gaussian Mixture Model

The following lemma, which was given in (Balakrishnan et al., 2017b), ensures the properties of Lipschitz-Gradient-2$(\gamma, \mathcal{B})$, smoothness and strongly concave for model (4). It is easy to show that the model is self-consistent (Yi and Caramanis, 2015).

Lemma 1 ((Balakrishnan et al., 2017b; Yi and Caramanis, 2015)) *If $\frac{\|\beta^*\|_2}{\sigma} \geq r$, where r is a sufficiently large constant denoting the minimum signal-to-noise ratio (SNR), then there exists an absolute constant $C > 0$ such that the properties of self-consistent, Lipschitz-Gradient-2$(\gamma, \mathcal{B})$, μ-smoothness and υ-strongly concave hold for function $Q(\cdot;\cdot)$ with $\gamma = \exp(-Cr^2), \mu = \upsilon = 1, R = k\|\beta^*\|_2, k = \frac{1}{4}$, and $\mathcal{B} = \{\beta : \|\beta - \beta^*\|_2 \leq R\}$.*

Lemma 2 *With the same notations as in Lemma 1, for each $\beta \in \mathcal{B}$ with $\|\beta\|_0 \leq s$, the j-the coordinate of $\nabla q_i(\beta;\beta)$ is ξ-sub-exponential with*

$$\xi = C_1\sqrt{\|\beta^*\|_\infty^2 + \sigma^2}, \tag{16}$$

where C_1 is some absolute constant. Also, each $[\nabla q_i(\beta;\beta)]_j$, where $i \in [n]$, is independent of others for any fixed $j \in [d]$.

Theorem 2 *In an ϵ-corrupted high dimensional Gaussian Mixture Model with ϵ satisfying the condition of*

$$\sqrt{(\|\beta^*\|_\infty^2 + \sigma^2)}\sqrt{s^*}(\epsilon\log(nd) + \sqrt{\frac{\log d}{n}}) \leq O(\|\beta\|_2^*), \tag{17}$$

if $\frac{\|\beta^*\|_2}{\sigma} \geq r$ for some sufficiently large constant r denoting the minimum SNR and the initial estimator β^{init} satisfies the inequality of $\|\beta^{init} - \beta^*\|_2 \leq \frac{1}{8}\|\beta^*\|_2$, then the output β^T of Algorithm 1 after choosing $s = O(s^*)$ and $\eta = O(1)$ satisfies the following with probability at least $1 - Td^{-3}$

$$\begin{aligned}\|\beta^T - \beta^*\|_2 \leq{}& \exp(-CTr^2)\|\beta^*\|_2 \\ &+ O\big(\sqrt{(\|\beta^*\|_\infty^2 + \sigma^2)}\sqrt{s^*}(\epsilon\log(nd) + \sqrt{\frac{\log d}{n}})\big),\end{aligned} \tag{18}$$

where C is some absolute constant.

From Theorem 2, we can see that when $\epsilon \leq \tilde{O}(\frac{1}{\sqrt{n}})$ and $T = O(\log \frac{n}{s^* \log d})$, the output achieves an estimation error of $O(\sqrt{\frac{s^* \log d}{n}})$, which matches the best-known error bound of the no-outlier case (Yi and Caramanis, 2015; Wang et al., 2015). Also, we assume that the SNR ratio is large, which is reasonable since it has been shown that for Gaussian Mixture Model with low SNR, the variance of noise makes it harder for the algorithm to converge (Ma et al., 2000).

5.2 Corrupted Mixture of Regressions Model

The following lemma, which was given in (Balakrishnan et al., 2017b; Yi and Caramanis, 2015), shows the properties of Lipschitz-Gradient-2$(\gamma, \mathcal{B})$, smoothness and strongly concave for model (6).

Lemma 3 ((Balakrishnan et al., 2017b; Yi and Caramanis, 2015)) *If $\frac{\|\beta^*\|_2}{\sigma} \geq r$, where r is a sufficiently large constant denoting the required minimal signal-to-noise ratio (SNR), then function $Q(\cdot;\cdot)$ of the Mixture of Regressions Model has the properties of self-consistent, Lipschitz-Gradient-2$(\gamma, \mathcal{B})$, μ-smoothness, and υ-strongly with $\gamma \in (0, \frac{1}{4}), \mu = \upsilon = 1, \mathcal{B} = \{\beta : \|\beta - \beta^*\|_2 \leq R\}, R = k\|\beta^*\|_2$, and $k = \frac{1}{32}$.*

Lemma 4 *With the same notations as in Lemma 3, for each $\beta \in \mathcal{B}$ and $\|\beta\|_0 = s$, the j-the coordinate of $\nabla q_i(\beta;\beta)$ is ξ-sub-exponential with*

$$\xi = C \max\{\|\beta^*\|_2^2 + \sigma^2, 1, \sqrt{s}\|\beta^*\|_2\}, \tag{19}$$

where $C > 0$ is some absolute constant. Also, each $[\nabla q_i(\beta;\beta)]_j$, where $i \in [n]$, is independent of others for any fixed $j \in [d]$.

Theorem 3 *In an ϵ-corrupted high dimensional Mixture of Regressions Model with ϵ satisfying the condition of*

$$\max\{\|\beta^*\|_2 + \sigma^2, 1, \sqrt{s^*}\|\beta^*\|_2\}\sqrt{s^*}(\epsilon \log(nd) + \sqrt{\frac{\log d}{n}}) \leq O(\|\beta\|_2^*), \tag{20}$$

if $\frac{\|\beta^*\|_2}{\sigma} \geq r$ for some sufficiently large constant r denoting the minimum SNR and the initial estimator β^{init} satisfies the inequality of $\|\beta^{init} - \beta^*\|_2 \leq \frac{1}{64}\|\beta^*\|_2$, then the output β^T of Algorithm 1 after choosing $s = O(s^*)$ and $\eta = O(1)$ satisfies the following with probability at least $1 - Td^{-3}$

$$\begin{aligned}\|\beta^T - \beta^*\|_2 \leq \gamma^{\frac{T}{2}}\|\beta^*\|_2 + O\big(&\max\{\|\beta^*\|_2 + \sigma^2, 1, \sqrt{s^*}\|\beta^*\|_2\} \\ &\times \sqrt{s^*}(\epsilon \log(nd) + \sqrt{\frac{\log d}{n}})\big),\end{aligned} \tag{21}$$

where $\gamma \in (0, \frac{1}{4})$ is a constant.

Note that in the above theorem, when $\epsilon \leq \tilde{O}(\frac{1}{\sqrt{n}})$ and $T = O(\log \frac{\sqrt{n}}{\sqrt{\log d} s^*})$, the estimation error becomes $O(s^* \sqrt{\frac{\log d}{n}})$, which differs from the $O(\sqrt{\frac{s^* \log d}{n}})$ minimax lower bound by only a factor of $\sqrt{s^*}$. We leave it as an open problem for further improvement. Recently, (Chen et al., 2018) shows that in the no-outlier and low dimensional setting, an assumption of $SNR \geq \rho$ for some constant ρ is necessary for achieving the optimal rate $\Theta(\sqrt{\frac{d}{n}})$.

5.3 Corrupted Linear Regression with Missing Covariates

Lemma 5 ((Balakrishnan et al., 2017b; Yi and Caramanis, 2015)) *If* $\frac{\|\beta^*\|_2}{\sigma} \leq r$ *and* $p_m < \frac{1}{1+2b+2b^2}$, *where* r *is a constant denoting the required maximum signal-to-noise ratio (SNR) and* $b = r^2(1+k)^2$ *for some constant* $k \in (0,1)$, *then function* $Q(\cdot;\cdot)$ *of the linear regression with missing covariates has the properties of self-consistent, Lipschitz-Gradient-2*$(\gamma, \mathcal{B})$, μ*-smoothness and* υ*-strongly with*

$$\gamma = \frac{b + p_m(1+2b+2b^2)}{1+b} < 1, \mu = \upsilon = 1,$$
$$\mathcal{B} = \{\beta : \|\beta - \beta^*\|_2 \leq R\}, \text{ where } R = k\|\beta^*\|_2. \tag{22}$$

Lemma 6 *With the same assumptions as in Lemma 5, for each* $\beta \in \mathcal{B}$ *with* $\|\beta\|_0 = s$, $[\nabla q_i(\beta;\beta)]_j$ *is* ξ*-sub-exponential with*

$$\xi = C[(1+k)(1+kr)^2\sqrt{s}\|\beta^*\|_2 + \max\{(1+kr)^2, \sigma^2 + \|\beta^*\|_2^2\}] \tag{23}$$

for some constant $C > 0$. *Also, each* $[\nabla q_i(\beta;\beta)]_j$, *where* $i \in [n]$, *is independent of others for any fixed* $j \in [d]$.

Theorem 4 *In an* ϵ*-corrupted high dimensional linear regression with missing covariates model with* ϵ *satisfying the condition of*

$$[(1+k)(1+kr)^2\sqrt{s}\|\beta^*\|_2 + \max\{(1+kr)^2, \sigma^2 + \|\beta^*\|_2^2\}]\sqrt{s^*}(\epsilon \log(nd) + \sqrt{\frac{\log d}{n}}) \leq O(\|\beta^*\|_2)$$

for some $k \in (0,1)$, *if* $\|\beta^{init} - \beta^*\|_2 \leq \frac{k\|\beta^*\|_2^2}{2}$ *and the assumptions in Lemma 5 hold, then, the output* β^T *of Algorithm 1 after taking* $s = O(s^*)$ *and* $\eta = O(1)$ *satisfies the following with probability at least* $1 - Td^{-3}$

$$\|\beta^T - \beta^*\|_2 \leq \gamma^{\frac{t}{2}}\|\beta^*\|_2 + O\big(\max\{\|\beta^*\|_2^2 + \sigma^2, 1, \sqrt{s^*}\|\beta^*\|_2\} \times \sqrt{s^*}(\epsilon \log(nd) + \sqrt{\frac{\log d}{n}})\big), \tag{24}$$

where the Big-O term hides the terms of k *and* r.

Note that similar to the mixture of regressions model, when $\epsilon \leq \tilde{O}(\frac{1}{\sqrt{n}})$, the estimation error is $O(s^*\sqrt{\frac{\log d}{n}})$, which is only a factor of $\sqrt{s^*}$ away from the optimal. However, unlike the previous two models, we assume here that SNR is upper bounded by some constant which is unavoidable as pointed out in (Loh and Wainwright, 2011).

6 Experiments

In this section, we empirically study the performance of Algorithm 1 on the three models mentioned in the previous section. Since in the paper we mainly focus on the statistical setting and its theoretical behaviors, thus, we will only perform our algorithm on the synthetic data. It is notable that previous papers on the statistical guarantees of EM algorithm all perform their algorithms on synthetic data only such as (Balakrishnan et al., 2017b; Wang et al., 2015; Yi and Caramanis, 2015). Thus, performing experiments on synthetic data only is enough for the paper. For the real world datasets, since most of them do not follow the assumptions (model assumption and high dimensional sparse assumption) in our paper. Thus, we will left designing practical robust EM algorithm with statistical guarantees as future research.

For each of these models, we generate synthesized datasets according to the underlying distribution. We will use $\|\beta - \beta^*\|_2$ to measure the estimation error, and test how it is affected by different parameter settings from two aspects. Firstly, we examine how the *underlying sparsity* parameter s^* of the model affects the estimation error and whether it is consistent with our theoretical results. Secondly, we test how the corruption fraction ϵ of the data and the dimensionality d affect the convergence rate, as well as the estimation error. All experiments are repeated for 20 runs and the average results are reported.

Parameter setting Throughout the experiments we will follow the setting of the previous related works on high dimensional EM algorithms which have statistical guarantees but are not corruption-proofing (Zhu et al., 2017; Wang et al., 2015; Yi and Caramanis, 2015). We fix the dataset size n to be 2000, because using a larger n does not exhibit significant difference. For each model, the experiment is divided into three parts as mentioned previously: The first one (Figure 2) measures $\|\beta - \beta^*\|_2$ v.s. $\sqrt{n/(s^* \log d)}$ by varying s^* from 3 to 15, with d fixed to be 100, which follows the previous works (Wang et al., 2015; Zhu et al., 2017); The second one (Figure 3) examines the convergence behavior under different corruption rate ϵ which varies from 0 to 0.2; The last one (Figure 4) shows the convergence behavior under different data dimensionality d which ranges from 80 to 240, with fixed $\epsilon = 0.2$.

For each experiment, instead of choosing the initial vectors which are close to the optimal ones, we use random initialization. We will set $s = s^*$ in our algorithm, which is also used in the previous methods. Besides the parameter s, there are also two other parameters of the algorithm that need to be specified: the D-Trim parameter α and the step size η. We are also required to set the

"noise level" for each of the three models, which is quantified by σ in their definitions. It is notable that the choices of these parameters are quite flexible.

GMM : Corrupted Gaussian Mixture Model (4). We fix σ to 0.5, α to 0.2 and η to 0.1.

MRM Corrupted Mixture of Regressions Model (6). We fix σ to 0.2, α to 0.2 and η to 0.1.

RMC Corrupted Linear Regression with Missing Covariates Model (8). We set $\sigma = 0.1$, $\alpha = 0.3$, and the missing probability $p_m = 0.1$, but use three different step sizes $\eta = 0.05, 0.1, 0.08$ for the three parts of the experiment, respectively.

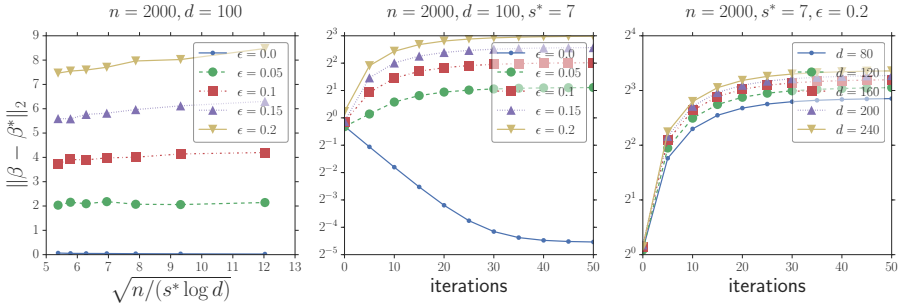
Results Firstly, we will mainly show that the classical high dimensional gradient EM algorithm in (Wang et al., 2015) is not robust against to the corruptions. Here we conduct the algorithm on the three models. For each experiment, we tune the parameters to be optimal as showed in (Wang et al., 2015). We test the algorithm w.r.t to $\sqrt{n/(s^* \log d)}$, iteration and different dimensions d.

As we can see from Figure 1. In all the three models, the algorithm performs quite well if there is no corruptions ($\epsilon = 0$) which also has been showed in the previous papers (Wang et al., 2015; Zhu et al., 2017). However, when there are $\epsilon = 0.05$ fraction of the samples are corrupted, the classical high dimensional EM algorithm will achieve a large estimation error. These results motivate us to design some robust high dimensional EM algorithms while also have provable statistical guarantees.

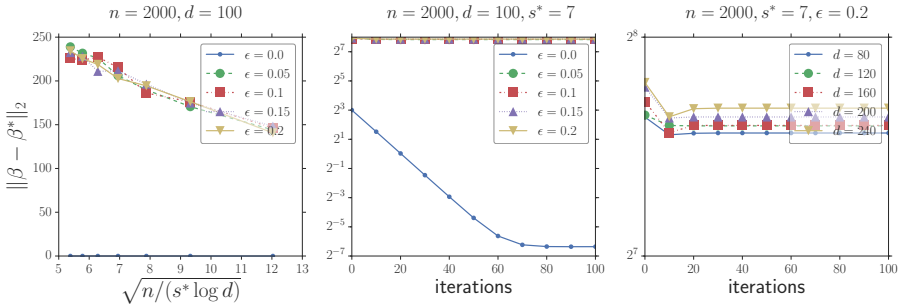
Next, we show the performance of our Algorithm 1. For the first part (Figure 2), we can see that when ϵ is small, the final estimation error in each of the three models decreases when the term $\sqrt{n/(s^* \log d)}$ increases, as predicted by Theorem 2. But when ϵ is relatively large, the trend becomes less obvious for the Gaussian Mixture Model and the Mixture of Regressions model, because now the factor $\epsilon \log(nd)$ comes into play.

Figure 3 shows that our algorithm achieves linear convergence on all three models and all values of ϵ, but the final converged error is heavily affected by ϵ, and especially for the Gaussian Mixture and Linear Regression with Missing Covariates Models. Moreover, when ϵ is small, the estimation errors are comparable to or even the same as the non-corrupted ones, this is actually reasonable since it is corruption-proofing when ϵ is small theoretically. In the third part of the experiments (Figure 4), varying d seems not affect the convergence behavior much, which is reasonable as the error bound depends on d only logarithmically and changes fairly slow. Thus, these results support Theorem 1.

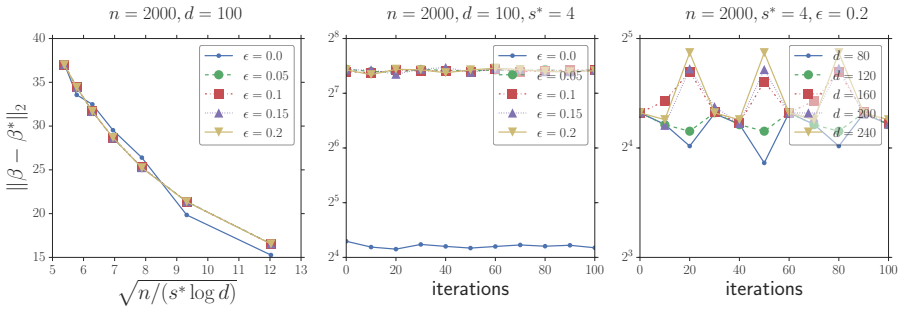
All the results show that our algorithm is robust against to some level of corruption while also could achieve an estimation error that is comparable to the non-corrupted ones.



(a) GMM



(b) MRM



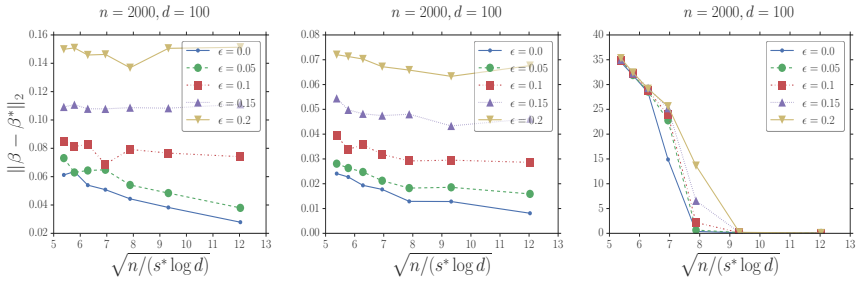
(c) RMC

Fig. 1: Estimation error of classical high dimensional gradient EM algorithm in (Wang et al., 2015) w.r.t sample size, iteration and dimension.

7 Conclusion

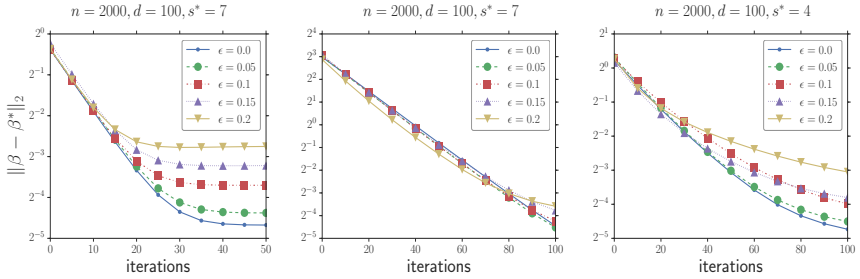
In this paper we study the problem of estimating latent variable models with arbitrarily corrupted samples in the high dimensional sparse case and propose a method called Trimmed Gradient Expectation Maximization. Specifically, we show that our algorithm is corruption-proofing and could achieve the (near) optimal statistical rate for some statistical models under some levels of corruption. Experimental results support our theoretical analysis and also show that our algorithm is indeed robust against to some corrupted samples.

There are still many open problems. Firstly, in this paper, all of our theoretical guarantees need the initial parameter be close enough to the underlying parameter, which is quite strong. So how do we relax this assumption? Second, the three specific models we considered in the paper are quite simple, can we generalize to more models such as multi-component Gaussian Mixture Model or Mixture of Linear Regressions Model? Thirdly, in this paper we assume that the sparsity of the underlying parameter is known, how to deal with the case where it is unknown?



(a) GMM (b) MRM (c) RMC

Fig. 2: Estimation error v.s. $\sqrt{n/(s^* \log d)}$



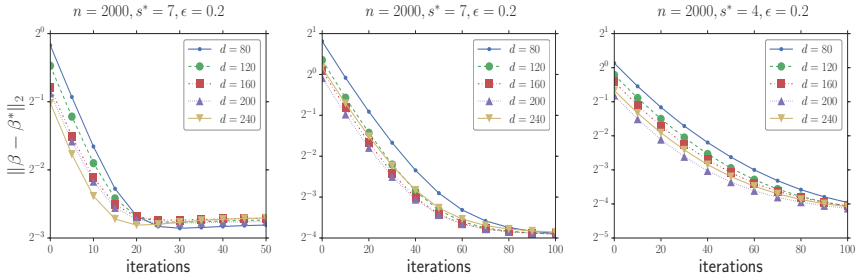
(a) GMM (b) MRM (c) RMC

Fig. 3: Estimation error v.s. iterations t under different corruption rate ϵ

References

Aitkin M, Wilson GT (1980) Mixture models, outliers, and the em algorithm. Technometrics 22(3):325–331

Alistarh D, Allen-Zhu Z, Li J (2018) Byzantine stochastic gradient descent. In: Advances in Neural Information Processing Systems, pp 4613–4623



(a) GMM (b) MRM (c) RMC

Fig. 4: Estimation error v.s. iterations t under different dimensionality d

Balakrishnan S, Du SS, Li J, Singh A (2017a) Computationally efficient robust sparse estimation in high dimensions. In: Conference on Learning Theory, pp 169–212

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

Blumensath T, Davies ME (2009) Iterative hard thresholding for compressed sensing. Applied and computational harmonic analysis 27(3):265–274

Boucheron S, Lugosi G, Massart P (2013) Concentration inequalities: A nonasymptotic theory of independence. Oxford university press

Chen Y, Caramanis C, Mannor S (2013) Robust sparse regression under adversarial corruption. In: International Conference on Machine Learning, pp 774–782

Chen Y, Su L, Xu J (2017) Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. Proceedings of the ACM on Measurement and Analysis of Computing Systems 1(2):44

Chen Y, Yi X, Caramanis C (2018) Convex and nonconvex formulations for mixed regression with two components: Minimax optimal rates. IEEE Transactions on Information Theory 64(3):1738–1766

Dalalyan AS, Thompson P (2019) Outlier-robust estimation of a sparse linear model using ℓ_1-penalized huber's m-estimator. arXiv preprint arXiv:190406288

Dawid AP, Skene AM (1979) Maximum likelihood estimation of observer error-rates using the em algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics) 28(1):20–28

Diakonikolas I, Kamath G, Kane DM, Li J, Moitra A, Stewart A (2016) Robust estimators in high dimensions without the computational intractability. In: 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), IEEE, pp 655–664

Diakonikolas I, Kane DM, Stewart A (2017) Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In: 2017 IEEE 58th Annual Symposium on Foundations of Computer Science

(FOCS), IEEE, pp 73–84
Diakonikolas I, Kane DM, Stewart A (2018) List-decodable robust mean estimation and learning mixtures of spherical gaussians. In: Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, ACM, pp 1047–1060
Du SS, Balakrishnan S, Singh A (2017) Computationally efficient robust estimation of sparse functionals. arXiv preprint arXiv:170207709
Faria S, Gonçalves F (2013) Financial data modeling by poisson mixture regression. Journal of Applied Statistics 40(10):2150–2162
Holland MJ (2018) Robust descent using smoothed multiplicative noise. arXiv preprint arXiv:181006207
Huber PJ (2011) Robust statistics. Springer
Johnson R, Zhang T (2013) Accelerating stochastic gradient descent using predictive variance reduction. In: Advances in neural information processing systems, pp 315–323
Laird NM (2010) The em algorithm in genetics, genomics and public health. Statistical Science pp 450–457
Li J (2017) Robust sparse estimation tasks in high dimensions. arXiv preprint arXiv:170205860
Liu L, Li T, Caramanis C (2019) High dimensional robust estimation of sparse models via trimmed hard thresholding. arXiv preprint arXiv:190108237
Loh PL, Wainwright MJ (2011) High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In: Advances in Neural Information Processing Systems, pp 2726–2734
Ma J, Xu L, Jordan MI (2000) Asymptotic convergence rate of the em algorithm for gaussian mixtures. Neural Computation 12(12):2881–2907
McLachlan G, Krishnan T (2007) The EM algorithm and extensions, vol 382. John Wiley & Sons
Nesterov Y (2013) Introductory lectures on convex optimization: A basic course, vol 87. Springer Science & Business Media
Prasad A, Suggala AS, Balakrishnan S, Ravikumar P (2018) Robust estimation via robust gradient estimation. arXiv preprint arXiv:180206485
Suggala AS, Bhatia K, Ravikumar P, Jain P (2019) Adaptive hard thresholding for near-optimal consistent robust regression. arXiv preprint arXiv:190308192
Thompson P, Dalalyan AS (2018) Restricted eigenvalue property for corrupted gaussian designs. arXiv preprint arXiv:180508020
Vershynin R (2010) Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv:10113027
Wang Z, Gu Q, Ning Y, Liu H (2015) High dimensional em algorithm: Statistical optimization and asymptotic normality. In: Advances in neural information processing systems, pp 2521–2529
Wu CJ, et al. (1983) On the convergence properties of the em algorithm. The Annals of statistics 11(1):95–103
Yang MS, Lai CY, Lin CY (2012) A robust em clustering algorithm for gaussian mixture models. Pattern Recognition 45(11):3950–3961

Yi X, Caramanis C (2015) Regularized em algorithms: A unified framework and statistical guarantees. In: Advances in Neural Information Processing Systems, pp 1567–1575

Yin D, Chen Y, Ramchandran K, Bartlett P (2018) Byzantine-robust distributed learning: Towards optimal statistical rates. arXiv preprint arXiv:180301498

Zhu R, Wang L, Zhai C, Gu Q (2017) High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, pp 4180–4188

A Auxiliary Lemmas

In this section, we introduce prerequisite knowledge and technical lemmas in order to prove the main results.

In order to analyze the Dimensional α-trimmed estimator, we first give some results for 1-dimensional samples and denote it as $\text{trmean}_\alpha(\cdot)$.

Definition 8 Given a set of ϵ-corrupted samples $\{z_i\}_{i=1}^n \subseteq \mathbb{R}$, the trimmed mean estimator $\text{trmean}_\alpha(\{z_i\}_{i=1}^n) \in \mathbb{R}$ removes the largest and smallest α fraction of elements in $\{z_i\}_{i=1}^n$ and calculate the mean of the remaining terms. We choose $\alpha = c_0\epsilon$, for some constant $c_0 \geq 1$. We also require that $\alpha \leq \frac{1}{2} - c_1$ for some small constant $c_1 > 0$.

For the 1-dimensional trimmed mean estimator, we have the following upper bound on the error w.r.t the population mean.

Lemma 7 (Lemma A.2 in (Liu et al., 2019)) *Let $\{z_i\}_{i=1}^n \subset \mathbb{R}^d$ be $n = \Omega(\log d)$ ϵ-corrupted samples. If the j-th coordinate, for each $j \in [d]$, of the samples $\{z_{i,j}\}_{i=1}^n$ are i.i.d ξ-exponential with mean μ^j, then after using the dimensional α-trimmed mean estimator, the following upper bound of error holds with probability at least $1 - d^{-3}$, for every $j \in [d]$*

$$|trmean_\alpha(\{z_{i,j}\}_{i=1}^n) - \mu^j| \leq C_2\xi(\epsilon \log(nd) + \sqrt{\frac{\log d}{n}}), \tag{25}$$

where C_2 is some constant dependent on c_1.

Next, we provide some symmetrization results of random variables, which will be used in our proofs. See (Boucheron et al., 2013) for details.

Lemma 8 *Let $y_1, y_2, \cdots, y_n$ be the n independent realizations of the random vector $Y \in \mathcal{Y}$, and $\mathcal{F}$ be a function class defined on $\mathcal{Y}$. For any increasing convex function $\phi(\cdot)$, the following holds*

$$\mathbb{E}\{\phi[\sup_{f\in\mathcal{F}} |\sum_{i=1}^n f(y_i) - \mathbb{E}(f(Y))|]\} \leq \mathbb{E}\{\phi[\sup_{f\in\mathcal{F}} |\sum_{i=1}^n \epsilon_i f(y_i)|]\},$$

where $\epsilon_1, \cdots, \epsilon_n$ are i.i.d Rademacher random variables that are independent of $y_1, \cdots, y_n$.

Lemma 9 *Let $y_1, \cdots, y_n$ be n independent realization of the random vector $Z \in \mathcal{Z}$ and $\mathcal{F}$ be a function class defined on $\mathcal{Z}$. If Lipschitz functions $\{\phi_i(\cdot)\}_{i=1}^n$ satisfy the following for all $v, v' \in \mathbb{R}$*

$$|\phi_i(v) - \phi_i(v')| \leq L|v - v'|$$

and $\phi_i(0) = 0$, then for any increasing convex function $\phi(\cdot)$, the following holds

$$\mathbb{E}\{\phi[|\sup_{f\in\mathcal{F}} \sum_{i=1}^n \epsilon_i\phi_i(f(y_i))|]\} \leq \mathbb{E}\{\phi[2|L \sup_{f\in\mathcal{F}} \sum_{i=1}^n \epsilon_i f(y_i)|]\},$$

where $\epsilon_1, \cdots, \epsilon_n$ are i.i.d Rademacher random variables that are independent of $y_1, \cdots, y_n$.

Finally we recall some definitions and lemmas on the sub-exponential and sub-Gaussian random variables. See (Vershynin, 2010) for details.

Definition 9 For a sub-exponential random vector X, its sub-exponential norm $\|X\|_{\psi_1}$ is defined as

$$\|X\|_{\psi_1} = \sup_{p\geq 1} p^{-1}(\mathbb{E}|X|^p)^{\frac{1}{p}}.$$

Lemma 10 *Let X be a zero-mean sub-exponential random variable, then there are absolute constants $C, c > 0$, such that when $|t| \leq \frac{c}{\|X\|_{\psi_1}}$,*

$$\mathbb{E}[\exp(tX)] \leq \exp(Ct^2\|X\|_{\psi_1}^2).$$

Lemma 11 (Bernstein's inequality) *Let $X_1, \cdots, X_n$ be n i.i.d realizations of υ-sub-exponential random variable X with mean μ. Then,*

$$Pr(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \geq t) \leq 2\exp(-n\min(-\frac{t^2}{\upsilon^2}, \frac{t}{2\upsilon})).$$

Definition 10 A random variable X is sub-Gaussian with variance σ^2 if for all $t > 0$, the following holds

$$\Pr(|X - \mathbb{E}X| \geq t) \leq 2\exp(-\frac{t^2}{2\sigma^2}).$$

Definition 11 For a sub-Gaussian random variable X, its sub-Gaussian norm $\|X\|_{\psi_2}$ is defined as

$$\|X\|_{\psi_2} = \sum_{p\geq 1} p^{-\frac{1}{2}}(\mathbb{E}|X|^p)^{\frac{1}{p}}.$$

Lemma 12 *If X is sub-Gaussian or sub-exponential, then $\|X - \mathbb{E}X\|_{\psi_2} \leq 2\|X\|_{\psi_2}$ or $\|X - \mathbb{E}X\|_{\psi_1} \leq 2\|X\|_{\psi_1}$ holds, respectively.*

Lemma 13 *For two sub-Gaussian random variables X_1, X_2, $X_1 \cdot X_2$ is a sub-exponential random variable with*

$$\|X_1 \cdot X_2\|_{\psi_1} \leq C\max\{\|X_1\|_{\psi_2}^2, \|X_2\|_{\psi_2}^2\}.$$

Lemma 14 *Let $X_1, X_2, \cdots, X_k$ be k independent zero-mean sub-Gaussian random variables, and $X = \sum_{j=1}^{k} X_j$. Then, X is sub-Gaussian with $\|X\|_{\psi_2}^2 \leq C\sum_{j=1}^{k}\|X_j\|_{\psi_2}^2$ for some absolute constant $C > 0$.*

B Omitted Proofs

B.1 Proof of Theorem 1

By Lemma 7 and our assumption on the ξ-sub-exponential property of each coordinate, we have the following in the t-th iteration with probability at least $1 - d^{-3}$ for some constant $C_2 > 0$

$$\|\nabla\tilde{Q}_n(\beta^t; \beta^t) - \nabla Q(\beta^t; \beta^t)\|_\infty \leq C_2\xi(\epsilon\log(nd) + \sqrt{\frac{\log d}{n}}). \tag{26}$$

For convenience, we let $\alpha = C_2\xi(\epsilon\log(nd) + \sqrt{\frac{\log d}{n}})$, and assume that for all iterations $t \in [T-1]$, event (26) holds (then all events hold with probability at least $1 - Tp^{-3}$).

In the t-th iteration, we define

$$\bar{\beta}^{t+0.5} = \beta^t + \eta\nabla Q(\beta^t; \beta^t) \tag{27}$$

and

$$\bar{\beta}^{t+1} = \text{trunc}(\bar{\beta}^{t+0.5}, \hat{\mathcal{S}}^{t+0.5}). \tag{28}$$

That is, $\bar{\beta}^{t+0.5}$ is the gradient update of β^t w.r.t the non-corrupted population gradient of $Q_n(\beta^t;\beta^t)$, and $\bar{\beta}^{t+1}$ is the estimation after truncating $\bar{\beta}^{t+0.5}$ w.r.t set $\hat{\mathcal{S}}^{t+0.5}$, which is the set of the s-largest coordinates of $\beta^{t+0.5}$.

By the definition, we have the following inequalities

$$\begin{aligned}
&\|\beta^{t+1} - \beta^*\|_2 = \|\text{trunc}(\beta^{t+0.5}, \hat{\mathcal{S}}^{t+0.5}) - \beta^*\|_2 \\
&\leq \|\text{trunc}(\beta^{t+0.5}, \hat{\mathcal{S}}^{t+0.5}) - \text{trunc}(\bar{\beta}^{t+0.5}, \hat{\mathcal{S}}^{t+0.5})\|_2 + \|\text{trunc}(\bar{\beta}^{t+0.5}, \hat{\mathcal{S}}^{t+0.5}) - \beta^*\|_2 \\
&= \|\text{trunc}(\beta^{t+0.5}, \hat{\mathcal{S}}^{t+0.5}) - \text{trunc}(\bar{\beta}^{t+0.5}, \hat{\mathcal{S}}^{t+0.5})\|_2 + \|\bar{\beta}^{t+1} - \beta^*\|_2 \\
&\leq \underbrace{\|(\beta^{t+0.5} - \bar{\beta}^{t+0.5})_{\hat{\mathcal{S}}^{t+0.5}}\|_2}_{A} + \underbrace{\|\bar{\beta}^{t+1} - \beta^*\|_2}_{B}.
\end{aligned} \tag{29}$$

For the term A, we have

$$\begin{aligned}
\|(\beta^{t+0.5} - \bar{\beta}^{t+0.5})_{\hat{\mathcal{S}}^{t+0.5}}\|_2 &\leq \sqrt{s}\|\beta^{t+0.5} - \bar{\beta}^{t+0.5}\|_\infty \\
&= \eta\sqrt{s}\|\nabla\tilde{Q}_n(\beta^t;\beta^t) - \nabla Q(\beta^t;\beta^t)\|_\infty.
\end{aligned} \tag{30}$$

Thus, if $\beta^t \in \mathcal{B}$, *i.e.,* $\|\beta^t - \beta^*\| \leq k\|\beta^*\|_2$) and $\|\beta^t\|_0 = s$, then by the assumption and (26), we have

$$A \leq \eta\sqrt{s}\alpha. \tag{31}$$

Next, we will bound the term B. To do this, we need the following lemma, which follows (Wang et al., 2015).

Lemma 15 *If*

$$\|\bar{\beta}^{t+0.5} - \beta^*\|_2 \leq k\|\beta^*\|_2 \tag{32}$$

for some $k \in (0,1)$ *and*

$$s \geq \frac{4(1+k)^2}{(1-k)^2}s^* \text{ and } \sqrt{s}\|\beta^{t+0.5} - \bar{\beta}^{t+0.5}\|_\infty \leq \frac{(1-k)^2}{2(1+k)}\|\beta^*\|_2, \tag{33}$$

then, the following holds

$$\|\bar{\beta}^{t+1} - \beta^*\|_2 \leq \frac{C\sqrt{s^*}}{\sqrt{1-k}}\|\beta^{t+0.5} - \bar{\beta}^{t+0.5}\|_\infty + (1 + 4\sqrt{\frac{s^*}{s}})^{1/2}\|\bar{\beta}^{t+0.5} - \beta^*\|_2. \tag{34}$$

Proof (Proof of Lemma 15) By assumption (32), we have

$$(1-k)\|\beta^*\|_2 \leq \|\bar{\beta}^{t+0.5}\|_2 \leq (1+k)\|\beta^*\|_2. \tag{35}$$

We then denote

$$\bar{\theta} = \frac{\bar{\beta}^{t+0.5}}{\|\bar{\beta}^{t+0.5}\|_2}, \theta = \frac{\beta^{t+0.5}}{\|\bar{\beta}^{t+0.5}\|_2}, \text{ and } \theta^* = \frac{\beta^*}{\|\beta^*\|_2} \tag{36}$$

and the sets $\mathcal{I}_1, \mathcal{I}_2$ and $\mathcal{I}_3$ as the follows

$$\mathcal{I}_1 = \mathcal{S}^* \backslash \hat{\mathcal{S}}^{t+0.5}, \mathcal{I}_2 = \mathcal{S}^* \bigcap \hat{\mathcal{S}}^{t+0.5}, \text{ and } \mathcal{I}_3 = \hat{\mathcal{S}}^{t+0.5} \backslash \mathcal{S}^*, \tag{37}$$

where $S^* = \text{supp}(\beta^*)$. Let $s_i = |\mathcal{I}_i|$ for $i = 1,2,3$, respectively. Also, we define $\Delta = \langle\bar{\theta}, \theta^*\rangle$. Note that

$$\Delta = \langle\bar{\theta}, \theta^*\rangle = \sum_{j\in\mathcal{S}^*}\bar{\theta}_j\theta_j^* = \sum_{j\in\mathcal{I}_1}\bar{\theta}_j\theta_j^* + \sum_{j\in\mathcal{I}_2}\bar{\theta}_j\theta_j^* \leq \|\bar{\theta}_{\mathcal{I}_1}\|_2\|\theta^*_{\mathcal{I}_1}\|_2 + \|\bar{\theta}_{\mathcal{I}_2}\|_2\|\theta^*_{\mathcal{I}_2}\|_2. \tag{38}$$

By Cauchy-Schwartz inequality, we have

$$\begin{aligned}\Delta^2 &\leq (\|\bar{\theta}_{\mathcal{I}_1}\|_2\|\theta^*_{\mathcal{I}_1}\|_2 + \|\bar{\theta}_{\mathcal{I}_2}\|_2\|\theta^*_{\mathcal{I}_2}\|_2)^2 \\ &\leq (\|\bar{\theta}_{\mathcal{I}_1}\|_2^2 + \|\bar{\theta}_{\mathcal{I}_2}\|_2^2)(\|\theta^*_{\mathcal{I}_1}\|_2^2 + \|\theta^*_{\mathcal{I}_2}\|_2^2) \\ &= (1 - \|\bar{\theta}_{\mathcal{I}_3}\|_2^2)(1 - \|\theta^*_{\mathcal{I}_3}\|_2^2) \leq 1 - \|\bar{\theta}_{\mathcal{I}_3}\|_2^2. \end{aligned} \tag{39}$$

Since $\mathcal{I}_3 \subseteq \hat{\mathcal{S}}^{t+0.5}$ and $\mathcal{I}_1 \bigcap \hat{\mathcal{S}}^{t+0.5} = \emptyset$, we have

$$\frac{\|\beta^{t+0.5}_{\mathcal{I}_3}\|_2^2}{\|\beta^{t+0.5}_{\mathcal{I}_1}\|_2^2} \geq \frac{s_3}{s_1}, \text{ i.e., } \frac{\|\theta_{\mathcal{I}_3}\|_2}{\sqrt{s_3}} \geq \frac{\|\theta_{\mathcal{I}_1}\|_2}{\sqrt{s_1}}. \tag{40}$$

We let $\tilde{\epsilon} = 2\|\bar{\theta} - \theta\|_\infty = 2\frac{\|\bar{\beta}^{t+0.5} - \beta^{t+0.5}\|_\infty}{\|\bar{\beta}^{t+0.5}\|_2}$. Note that we have

$$\max\{\frac{\|\theta_{\mathcal{I}_3} - \bar{\theta}_{\mathcal{I}_3}\|_2}{\sqrt{s_3}}, \frac{\|\theta_{\mathcal{I}_1} - \bar{\theta}_{\mathcal{I}_1}\|_2}{\sqrt{s_1}}\} \leq \max\{\|\theta_{\mathcal{I}_3} - \bar{\theta}_{\mathcal{I}_3}\|_\infty, \|\theta_{\mathcal{I}_1} - \bar{\theta}_{\mathcal{I}_1}\|_\infty\} \leq \|\bar{\theta} - \theta\|_\infty = \frac{\tilde{\epsilon}}{2}, \tag{41}$$

which implies that

$$\begin{aligned}\frac{\|\bar{\theta}_{\mathcal{I}_3}\|_2}{\sqrt{s_3}} \geq \frac{\|\theta_{\mathcal{I}_3}\|_2}{\sqrt{s_3}} - \frac{\|\theta_{\mathcal{I}_3} - \bar{\theta}_{\mathcal{I}_3}\|_2}{\sqrt{s_3}} &\underset{(a)}{\geq} \frac{\|\theta_{\mathcal{I}_1}\|_2}{\sqrt{s_1}} - \frac{\|\theta_{\mathcal{I}_3} - \bar{\theta}_{\mathcal{I}_3}\|_2}{\sqrt{s_3}} \\ &\geq \frac{\|\bar{\theta}_{\mathcal{I}_1}\|_2}{\sqrt{s_1}} - \frac{\|\theta_{\mathcal{I}_3} - \bar{\theta}_{\mathcal{I}_3}\|_2}{\sqrt{s_3}} - \frac{\|\theta_{\mathcal{I}_1} - \bar{\theta}_{\mathcal{I}_1}\|_2}{\sqrt{s_1}} \geq \frac{\|\bar{\theta}_{\mathcal{I}_1}\|_2}{\sqrt{s_1}} - \tilde{\epsilon}, \end{aligned} \tag{42}$$

where inequality (a) is due to (40). Plugging (42) into (39), we have

$$\Delta^2 \leq 1 - \|\bar{\theta}_{\mathcal{I}_3}\|_2^2 \leq 1 - (\sqrt{\frac{s_3}{s_1}}\|\bar{\theta}_{\mathcal{I}_1}\|_2 - \sqrt{s_3}\tilde{\epsilon})^2. \tag{43}$$

Solving $\|\bar{\theta}_{\mathcal{I}_1}\|_2$ in (43), we get

$$\|\bar{\theta}_{\mathcal{I}_1}\|_2 \leq \sqrt{\frac{s_1}{s_3}}\sqrt{1 - \Delta^2} + \sqrt{s_1}\tilde{\epsilon} \leq \sqrt{\frac{s^*}{s}}\sqrt{1 - \Delta^2} + \sqrt{s^*}\tilde{\epsilon}. \tag{44}$$

The final inequality is due to the inequality $\frac{s_1}{s_3} \leq \frac{s_1+s_2}{s_3+s_2} = \frac{s^*}{s}$, which follows from $\frac{s^*}{s} \leq \frac{(1-k)^2}{4(1+k)^2} \leq 1$ and $s_3 \geq s - s^* \geq s^* \geq s_1$.

In the following, we will prove that the right hand side of (44) is upper bounded by Δ. To achieve this, it is sufficient to show that

$$\Delta \geq \frac{\sqrt{s^*}\tilde{\epsilon} + [s^*\tilde{\epsilon}^2 - (s^*/s+1)(s^*\epsilon^2 - \tilde{s^*}/s)]^{\frac{1}{2}}}{s^*/s+1} = \frac{\sqrt{s^*}\tilde{\epsilon} + [-(s^*\tilde{\epsilon})^2/s + (s^*/s+1)s^*/s]^{\frac{1}{2}}}{s^*/s+1}. \tag{45}$$

To prove (45), we first note that $\sqrt{s^*}\tilde{\epsilon} \leq \Delta$, which is due to

$$\sqrt{s^*}\tilde{\epsilon} \leq \sqrt{s}\tilde{\epsilon} = \frac{2\sqrt{s}\|\bar{\beta}^{t+0.5} - \beta^{t+0.5}\|_\infty}{\|\beta^*\|_2} \leq \frac{1-k}{1+k} \leq \Delta, \tag{46}$$

where the second inequality is due to assumption (33) and the final inequality is due to

$$\Delta = \langle\bar{\theta}, \theta^*\rangle = \frac{\langle\bar{\beta}^{t+0.5}, \beta^*\rangle}{\|\bar{\beta}^{t+0.5}\|_2\|\beta^*\|_2} \overset{(a)}{\geq} \frac{\|\bar{\beta}^{t+0.5}\|_2^2 + \|\beta^*\|_2^2 - k^2\|\beta^*\|_2^2}{2\|\bar{\beta}^{t+0.5}\|_2\|\beta^*\|_2} \geq \frac{(1-k)^2 + 1 - k^2}{2(1+k)} = \frac{1-k}{1+k},$$

where inequality (a) is due to assumption (32).

Now, we show that (45) holds. By (46), we have

$$\sqrt{s}\tilde{\epsilon} \leq \frac{1-k}{1+k} < 1 < \sqrt{\frac{s^*+s}{s}}, \tag{47}$$

which implies that $\tilde{\epsilon} \leq \frac{\sqrt{s^*+s}}{s}$.

For the right hand side of (45), we have

$$\begin{aligned}
\frac{\sqrt{s^*}\tilde{\epsilon} + [-(s^*\tilde{\epsilon})^2/s + (s^*/s+1)s^*/s]^{\frac{1}{2}}}{s^*/s+1} &\leq \frac{\sqrt{s^*}\tilde{\epsilon} + [(s^*/s+1)s^*/s]^{\frac{1}{2}}}{s^*/s+1} && (48)\\
&\leq 2\sqrt{\frac{s^*}{s^*+s}} \leq 2\sqrt{\frac{1}{1+4(1+k)^2/(1-k)^2}} && (49)\\
&\leq \frac{1-k}{1+k} \leq \Delta. && (50)
\end{aligned}$$

Thus, in total, by (44) we can get

$$\|\bar{\theta}_{\mathcal{I}_1}\|_2 \leq \Delta. \tag{51}$$

From (39), we can see that

$$\Delta \leq \|\bar{\theta}_{\mathcal{I}_1}\|_2\|\theta^*_{\mathcal{I}_1}\|_2 + \|\bar{\theta}_{\mathcal{I}_2}\|_2\|\theta^*_{\mathcal{I}_2}\|_2 \leq \|\bar{\theta}_{\mathcal{I}_1}\|_2\|\theta^*_{\mathcal{I}_1}\|_2 + \sqrt{(1-\|\bar{\theta}_{\mathcal{I}_1}\|_2^2)}\sqrt{(1-\theta^*_{\mathcal{I}_1}\|_2^2)},$$

that is,

$$(\Delta - \|\bar{\theta}_{\mathcal{I}_1}\|_2\|\theta^*_{\mathcal{I}_1}\|_2)^2 \leq (1-\|\bar{\theta}_{\mathcal{I}_1}\|_2^2)(1-\theta^*_{\mathcal{I}_1}\|_2^2).$$

Solving the above inequality, we get

$$\begin{aligned}
\|\theta^*_{\mathcal{I}_1}\|_2 \leq \|\bar{\theta}_{\mathcal{I}_1}\|_2\Delta + \sqrt{1-\|\bar{\theta}_{\mathcal{I}_1}\|_2^2}\sqrt{1-\Delta^2} &\leq \|\bar{\theta}_{\mathcal{I}_1}\|_2 + \sqrt{1-\Delta^2}\\
&\leq \sqrt{\frac{s^*}{s}}\sqrt{1-\Delta^2} + \sqrt{s^*}\tilde{\epsilon} + \sqrt{1-\Delta^2}, \quad (52)
\end{aligned}$$

where the final inequality is due to (44). Combining this with (44) and (52), we have

$$\|\theta^*_{\mathcal{I}_1}\|_2\|\bar{\theta}_{\mathcal{I}_1}\|_2 \leq [\sqrt{\frac{s^*}{s}}\sqrt{1-\Delta^2} + \sqrt{s^*}\tilde{\epsilon} + \sqrt{1-\Delta^2}] \cdot [\sqrt{\frac{s^*}{s}}\sqrt{1-\Delta^2} + \sqrt{s^*}\tilde{\epsilon}]. \tag{53}$$

Now, by the definition of $\bar{\theta}$, we have

$$\bar{\beta}^{t+1} = \mathrm{trunc}(\bar{\beta}^{t+0.5}, \hat{\mathcal{S}}^{t+0.5}) = \mathrm{trunc}(\bar{\theta}, \hat{\mathcal{S}}^{t+0.5})\|\hat{\beta}^{t+0.5}\|_2. \tag{54}$$

Therefore, we get

$$\langle \frac{\bar{\beta}^{t+1}}{\|\bar{\beta}^{t+0.5}\|_2}, \frac{\beta^*}{\|\beta^*\|_2} \rangle = \langle \mathrm{trunc}(\bar{\theta}, \hat{\mathcal{S}}^{t+0.5}), \theta^* \rangle = \langle \bar{\theta}_{\mathcal{I}_2}, \theta^*_{\mathcal{I}_2} \rangle \geq \langle \bar{\theta}, \theta^* \rangle - \|\bar{\theta}_{\mathcal{I}_1}\|_2\|\theta^*_{\mathcal{I}_1}\|_2. \tag{55}$$

Let $\chi = \|\bar{\beta}^{t+0.5}\|_2\|\beta^*\|_2$. Then, by (55) and (53) we have

$$\begin{aligned}
&\langle \bar{\beta}^{t+1}, \beta^* \rangle\\
&\geq \langle \bar{\beta}^{t+0.5}, \beta^* \rangle - [(\sqrt{\frac{s^*}{s}}+1)\sqrt{\chi(1-\Delta^2)} + \sqrt{s^*}\sqrt{\chi}\tilde{\epsilon}] \cdot [\sqrt{\frac{s^*}{s}}\sqrt{\chi(1-\Delta^2)} + \sqrt{s^*}\sqrt{\chi}\tilde{\epsilon}]\\
&= \langle \bar{\beta}^{t+0.5}, \beta^* \rangle - (\sqrt{\frac{s^*}{s}} + \frac{s^*}{s})\chi(1-\Delta^2) - (1+2\sqrt{\frac{s^*}{s}})\sqrt{\chi(1-\Delta^2)}\sqrt{s^*}\sqrt{\chi}\tilde{\epsilon} - (\sqrt{s^*}\sqrt{\chi}\tilde{\epsilon})^2.
\end{aligned} \tag{56}$$

For the term $\sqrt{\chi(1-\Delta^2)}$, we have

$$\sqrt{\chi(1-\Delta^2)} \leq \sqrt{2\chi(1-\Delta)} \leq \sqrt{2\|\bar{\beta}^{t+0.5}\|_2\|\beta^*\|_2 - 2\langle\bar{\beta}^{t+0.5},\beta^*\rangle} \leq \|\bar{\beta}^{t+0.5}-\beta^*\|_2. \tag{57}$$

For the term $\sqrt{\chi}\tilde{\epsilon}$, we have

$$\sqrt{\chi}\tilde{\epsilon} = 2\sqrt{\|\bar{\beta}^{t+0.5}\|_2\|\beta^*\|_2}\frac{\|\bar{\beta}^{t+0.5}-\beta^{t+0.5}\|_\infty}{\|\bar{\beta}^{t+0.5}\|_2} \leq \frac{2}{\sqrt{1-k}}\|\bar{\beta}^{t+0.5}-\beta^{t+0.5}\|_\infty. \tag{58}$$

Plugging (57) and (58) into (56), we get

$$\langle\bar{\beta}^{t+1},\beta^*\rangle \geq \langle\bar{\beta}^{t+0.5},\beta^*\rangle - (\sqrt{\frac{s^*}{s}}+\frac{s^*}{s})\|\bar{\beta}^{t+0.5}-\beta^*\|_2^2 -$$
$$(1+2\sqrt{\frac{s^*}{s}})\|\bar{\beta}^{t+0.5}-\beta^*\|_2\frac{2\sqrt{s^*}}{\sqrt{1-k}}\|\bar{\beta}^{t+0.5}-\beta^{t+0.5}\|_\infty - \frac{4s^*}{1-k}\|\bar{\beta}^{t+0.5}-\beta^{t+0.5}\|_\infty^2. \tag{59}$$

Also, since $\|\bar{\beta}^{t+1}\|_2^2 + \|\beta^*\|_2^2 \leq \|\bar{\beta}^{t+0.5}\|_2^2 + \|\beta^*\|_2^2$, subtracting (59), we obtain

$$\begin{aligned}\|\bar{\beta}^{t+1}-\beta^*\|_2^2 &\leq (1+\sqrt{\frac{s^*}{s}}+\frac{s^*}{s})\|\bar{\beta}^{t+0.5}-\beta^*\|_2^2 + \frac{8s^*}{1-k}\|\bar{\beta}^{t+0.5}-\beta^{t+0.5}\|_\infty^2 \\ &+ (1+2\sqrt{\frac{s^*}{s}})\|\bar{\beta}^{t+0.5}-\beta^*\|_2\frac{4\sqrt{s^*}}{\sqrt{1-k}}\|\bar{\beta}^{t+0.5}-\beta^{t+0.5}\|_\infty \\ &\leq (1+2\sqrt{\frac{s^*}{s}}+2\frac{s^*}{s})[\|\bar{\beta}^{t+0.5}-\beta^*\|_2 + \frac{2\sqrt{s^*}}{\sqrt{1-k}}\|\bar{\beta}^{t+0.5}-\beta^{t+0.5}\|_\infty]^2 \\ &+ \frac{8s^*}{1-k}\|\bar{\beta}^{t+0.5}-\beta^{t+0.5}\|_\infty^2. \end{aligned}\tag{60}$$

Thus, we have

$$\|\bar{\beta}^{t+1}-\beta^*\|_2 \leq (1+4\sqrt{\frac{s^*}{s}})^{\frac{1}{2}}\|\bar{\beta}^{t+0.5}-\beta^*\|_2 + \frac{2\sqrt{2}\sqrt{s^*}}{\sqrt{1-k}}\|\bar{\beta}^{t+0.5}-\beta^{t+0.5}\|_\infty. \tag{61}$$

This completes the proof of Lemma 15.

Next, we bound the term $\|\bar{\beta}^{t+0.5}-\beta^*\|_2$ in (34).

Lemma 16 *Under the assumptions in Theorem 1, the following inequality holds*

$$\|\bar{\beta}^{t+0.5}-\beta^*\|_2 \leq (1-2\frac{\upsilon-\gamma}{\upsilon+\mu})\|\beta^t-\beta^*\|_2. \tag{62}$$

Proof (Proof of Lemma 16) We first note that the self-consistent property in (McLachlan and Krishnan, 2007) implies that

$$\beta^* = \arg\max_\beta Q(\beta;\beta^*), \tag{63}$$

which means that β^* is a maximizer of $Q(\beta;\beta^*)$. Thus, the proof follows from the convergence rate of the strongly convex and smooth functions $Q(\beta;\beta^*)$ in Nesterov (2013). For the step size $\eta = \frac{2}{\mu+\upsilon}$, we have

$$\|\beta^t + \eta\nabla Q(\beta^t;\beta^*) - \beta^*\|_2 \leq (\frac{\mu-\upsilon}{\mu+\upsilon})\|\beta^T-\beta^*\|_2. \tag{64}$$

Thus, we get

$$\|\bar{\beta}^{t+0.5}-\beta^*\|_2 = \|\beta^t+\eta\nabla Q(\beta^t;\beta^t)-\beta^*\|_2 \tag{65}$$
$$= \|\beta^t+\eta\nabla Q(\beta^t;\beta^*)-\beta^*\|_2 + \eta\|\nabla Q(\beta^t;\beta^*)-\nabla Q(\beta^t;\beta^t)\|_2 \tag{66}$$
$$\leq (\frac{\mu-\upsilon}{\mu+\upsilon})\|\beta^T-\beta^*\|_2 + \eta\gamma\|\beta^t-\beta^*\|_2. \tag{67}$$

Taking $\eta = \frac{2}{\mu+\upsilon}$, we complete the proof.

Combining Lemmas 15, 16, and equation (31), we have the following lemma.

Lemma 17 *If*

$$\|\bar{\beta}^{t+0.5} - \beta^*\|_2 \leq k\|\beta^*\|_2 \tag{68}$$

for some $k \in (0,1)$ *and further assuming that*

$$s \geq \frac{4(1+k)^2}{(1-k)^2}s^* \text{ and } \sqrt{s}\alpha \leq \frac{(1-k)^2}{2(1+k)}\|\beta^*\|_2, \tag{69}$$

then it holds with probability at least $1 - d^{-3}$ *that*

$$\|\beta^{t+1} - \beta^*\|_2 \leq \frac{2}{\upsilon+\mu}\sqrt{s}\alpha + \frac{1}{\upsilon+\mu}\frac{4\sqrt{2}\sqrt{s^*}}{\sqrt{1-k}}\alpha + (1+4\sqrt{\frac{s^*}{s}})^{\frac{1}{2}}(1-2\frac{\upsilon-\gamma}{\upsilon+\mu})\|\beta^t - \beta^*\|_2, \tag{70}$$

where $\alpha = C_2\xi(\epsilon\log(nd) + \sqrt{\frac{\log d}{n}})$.

We now prove Theorem 1.

Proof (Proof of Theorem 1) By Lemma 17, we know that it is sufficient to prove (68), which can be shown by mathematical induction.

We first prove $\beta^0 \in \mathcal{B}$. By assumption, we have $\|\beta^{\text{init}} - \beta^*\|_2 \leq \frac{R}{2}$. By the same proof of Lemma 15, we can get $\|\beta^0 - \beta^*\|_2 \leq (1+4\sqrt{\frac{s^*}{s}})^{\frac{1}{2}}\|\beta^{\text{init}} - \beta^*\|_2 \leq (1+4\sqrt{\frac{1}{4}})^{\frac{1}{2}}\frac{R}{2} \leq R = k\|\beta^*\|_2$. Thus, by Lemma 16, we can see that (68) holds for $t = 0$.

Now suppose that (68) holds for all $t \leq k$. Then, we have

$$\|\beta^{k+1} - \beta^*\|_2 \leq \frac{2}{\upsilon+\mu}\sqrt{s}\alpha + \frac{1}{\upsilon+\mu}\frac{4\sqrt{2}\sqrt{s^*}}{\sqrt{1-k}}\alpha + (1+4\sqrt{\frac{s^*}{s}})^{\frac{1}{2}}(1-2\frac{\upsilon-\gamma}{\upsilon+\mu})\|\beta^k - \beta^*\|_2, \tag{71}$$

by assumption we can see that $(1+4\sqrt{\frac{s^*}{s}})^{\frac{1}{2}}(1-2\frac{\upsilon-\gamma}{\upsilon+\mu}) \leq \sqrt{1-2\frac{\upsilon-\gamma}{\upsilon+\mu}}$. Thus, we have

$$\|\beta^{k+1} - \beta^*\|_2 \leq \frac{1}{\upsilon+\mu}\frac{(2\sqrt{s}+4\sqrt{2}\sqrt{s^*}/\sqrt{1-k})\alpha}{1-\sqrt{1-2\frac{\upsilon-\gamma}{\upsilon+\mu}}} + (\sqrt{1-2\frac{\upsilon-\gamma}{\upsilon+\mu}})^k R. \tag{72}$$

By the assumption of $\frac{1}{\upsilon+\mu}\frac{(2\sqrt{s}+4\sqrt{2}\sqrt{s^*}/\sqrt{1-k})\alpha}{1-\sqrt{1-2\frac{\upsilon-\gamma}{\upsilon+\mu}}} \leq (1-\sqrt{1-2\frac{\upsilon-\gamma}{\upsilon+\mu}})R$, we have

$$\|\beta^{k+1} - \beta^*\|_2 \leq (1-\sqrt{1-2\frac{\upsilon-\gamma}{\upsilon+\mu}})R + \sqrt{1-2\frac{\upsilon-\gamma}{\upsilon+\mu}}R = R. \tag{73}$$

Hence, by Lemma 16, we obtain (68) for the case of $t = k+1$. This completes the proof.

B.2 Proof of Lemma 2

From (5) it is oblivious that $[\nabla q_i(\beta,\beta))]_j$ is independent of other $i \in [n]$ for fixed $j \in [d]$. Next, we prove the property of sub-exponential for each coordinate.

Note that

$$[\nabla q_i(\beta,\beta))]_j = [2w_\beta(y_i) - 1]y_{i,j} - \beta_j,$$

and

$$\mathbb{E}[\nabla q_i(\beta,\beta))]_j = \mathbb{E}(2w_\beta(Y)Y_j - Y_j) - \beta_j.$$

For convenience, we let $\nabla q_{i,j}$ denote $[\nabla q_i(\beta,\beta))]_j$ and ∇q_j denote $\mathbb{E}[\nabla q_i(\beta,\beta))]_j$.

By the symmetrization lemma in Lemma 8, we have the following for any $t > 0$

$$\mathbb{E}\{\exp(t|[\nabla q_{i,j} - \nabla q_j]|)\} \leq \mathbb{E}\{\exp(t|\epsilon[2w_\beta(y_i) - 1]y_{i,j}|)\}, \tag{74}$$

where ϵ is a Rademacher random variable.

Next, we use Lemma 9 with $f(y_{i,j}) = y_{i,j}$, $\mathcal{F} = \{f\}$, $\phi_i(v) = [2w_\beta(y_i) - 1]v$ and $\phi(v) = \exp(u \cdot v)$. It is easy to see that ϕ_i is 1-Lipschitz. Thus, by Lemma 9 we have

$$\mathbb{E}\{\exp(t|\epsilon[2w_\beta(y_i) - 1]y_{i,j}|)\} \leq \mathbb{E}\{\exp[2t|\epsilon y_{i,j}|]\}. \tag{75}$$

By the formulation of the model, we have $y_{i,j} = z_i\beta_j^* + v_{i,j}$, where z_i is a Rademacher random variable and $v_{i,j} \sim \mathcal{N}(0, \sigma^2)$. It is easy to see that $y_{i,j}$ is sub-Gaussian and

$$\|y_{i,j}\|_{\psi_2} = \|z_i \cdot \beta_j^* + v_{i,j}\|_{\psi_2} \leq C \cdot \sqrt{\|z_i \cdot \beta_j\|_{\psi_2}^2 + \|v_{i,j}\|_{\psi_2}^2} \leq C'\sqrt{|\beta_j^*|^2 + \sigma^2}, \tag{76}$$

for some absolute constants C, C', where the last inequality is due to the facts that $\|z_j\beta_j^*\|_{\psi_2} \leq |\beta_j^*|$ and $\|v_{i,j}\|_{\psi_2} \leq C''\sigma^2$ for some $C'' > 0$.

Since $|\epsilon y_{i,j}| = |y_{i,j}|$, $\|\epsilon y_{i,j}\|_{\psi_2} = \|y_{i,j}\|_{\psi_2}$ and $\mathbb{E}(\epsilon y_{i,j}) = 0$, by Lemma 5.5 in Vershynin (2010) we have that for any u' there exists a constant $C^{(4)} > 0$ such that

$$\mathbb{E}\{\exp(u' \cdot \epsilon \cdot y_{i,j})\} \leq \exp(u'^2 \cdot C^{(4)} \cdot (|\beta|_j^2 + \sigma^2)). \tag{77}$$

Thus, for any $t > 0$ we get

$$\mathbb{E}\{\exp(2t \cdot |\epsilon \cdot y_{i,j}|)\} \leq 2\exp(t^2 \cdot C^{(5)} \cdot (|\beta|_j^2 + \sigma^2)) \tag{78}$$

for some constant $C^{(5)}$. Therefore, in total we have the following for some constant $C^{(6)} > 0$

$$\mathbb{E}\{\exp(t|[\nabla q_{i,j} - \nabla q_j]|)\} \leq \exp(t^2 \cdot C^{(6)} \cdot (|\beta|_j^2 + \sigma^2)) \leq \exp(t^2 \cdot C^{(6)} \cdot (|\beta^*|_\infty^2 + \sigma^2)). \tag{79}$$

Combining this with Lemma 10 and the definition, we know that $\nabla q_{i,j}$ is $O(\sqrt{\|\beta^*\|_\infty^2 + \sigma^2})$-sub-exponential.

B.3 Proof of Lemma 4

From (7) it is oblivious that $[\nabla q_i(\beta, \beta))]_j$ is independent of other $i \in [n]$ for any fixed $j \in [d]$. Next, we prove the property of sub-exponential.

Note that $\mathbb{E}\nabla q_{i,j} = \mathbb{E}2w_\beta(x, y)y \cdot x_j - \beta_j$. Thus, we have

$$\nabla q_{i,j} - \nabla q_j = \underbrace{2w_\beta(x_i, y_i)y_i x_{i,j} - \mathbb{E}[]2w_\beta(x, y)yx_j]}_{A} + \underbrace{[x_i x_i^T\beta - \beta]_j}_{B} - \underbrace{y_i x_{i,j}}_{C}. \tag{80}$$

For term A and any $t > 0$, we have

$$\mathbb{E}\{\exp(t|A|)\} \leq \mathbb{E}\{\exp[t|2\epsilon w_\beta(x_i, y_i)y_i x_{i,j}|]\}. \tag{81}$$

Using Lemma 9 on $f(y_i x_{i,j}) = y_i x_{i,j}$, $\mathcal{F} = f$, $\phi_i(v) = 2w_\beta(x, y)v$ and $\phi(v) = \exp(uv)$, we have

$$\mathbb{E}\{\exp[t|2\epsilon w_\beta(x_i, y_i)y_i x_{i,j}|] \leq \mathbb{E}\{\exp[4t|\epsilon y_i x_{i,j}|]\}. \tag{82}$$

Note that since $y_i = z_i\langle\beta^*, x_i\rangle + v_i$ and $\|z_i\langle\beta^*, x_i\rangle\|_{\psi_2} = \|\langle\beta^*, x_i\rangle\|_{\psi_2} \leq C\|\beta^*\|_2$ and $\|v_i\|_{\psi_2} \leq C'\sigma$ for some constants $C, C' > 0$, by Lemma 14 we know that there exists a constant $C'' > 0$ such that

$$\|y_i\|_{\psi_2} \leq C''\sqrt{\|\beta^*\|_2^2 + \sigma^2}. \tag{83}$$

Thus, by Lemma 13 we have

$$\|y_i x_{i,j}\|_{\psi_1} \leq \max\{C''^2(\|\beta^*\|_2^2 + \sigma^2), C'''\} \leq C_4 \max\{\|\beta^*\|_2^2 + \sigma^2, 1\}. \tag{84}$$

For term B, we have

$$\mathbb{E}\{\exp[t|B|]\} = \mathbb{E}\{\exp[t|\sum_{k=1}^{d} x_j x_k \beta_k - \beta_j|]\}, \tag{85}$$

where $x_j, x_k \sim \mathcal{N}(0,1)$. Now, by Lemma 13 we have $\|x_j x_k \beta_k\|_{\psi_1} \leq |\beta_k| C^{(5)}$ for some constant $C^{(5)} > 0$. Thus, we get $\|\sum_{k=1}^{d} x_j x_k \beta_k\|_{\psi_1} \leq C^{(5)} \|\beta\|_1$.

Also, we know that $\|\beta\|_1 \leq \sqrt{s}\|\beta\|_2$, since by assumption $\|\beta\|_0 = s$. Furthermore, we have $\|\beta\|_2 \leq \|\beta^*\|_2 + \|\beta^* - \beta\|_2 \leq (1 + \frac{1}{32})\|\beta^*\|_2$, since $\beta \in \mathcal{B}$ (by assumption). From Lemma 12, we get $\|B\|_{\psi_1} \leq C^{(6)}\sqrt{s}\|\beta^*\|_2$ with some constant $C^{(6)} > 0$.

Thus, we know that there exist some constants $C^{(7)} > 0$ and $C^{(8)} > 0$ such that

$$\begin{aligned}\|\nabla q_{i,j} - \nabla q_j\|_{\psi_1} &\leq C^{(7)} \max\{\|\beta^*\|_2^2 + \sigma^2, 1\} + C^{(8)}\sqrt{s}\|\beta^*\|_2 \\ &\leq C^{(9)} \max\{\|\beta^*\|_2^2 + \sigma^2, 1, \sqrt{s}\|\beta^*\|_2\}.\end{aligned}$$

This means that $\nabla q_{i,j}$ is $O(\max\{\|\beta^*\|_2^2 + \sigma^2, 1, \sqrt{s}\|\beta^*\|_2\})$ sub-exponential.

B.4 Proof of Lemma 6

For simplicity, we use notations $\bar{m}^i = m_\beta(x_i^{\text{obs}}, y_i)$, $\bar{m} = \beta(x^{\text{obs}}, y)$, $\bar{K}^i = K_\beta(x_i^{\text{obs}}, y_i)$, and $\bar{K} = K_\beta(x^{\text{obs}}, y)$. Then, we have

$$\nabla q_i - \nabla q = \underbrace{m_\beta(x_i^{\text{obs}}, y_i) y_i - \mathbb{E}[m_\beta(x_i^{\text{obs}}, y_i) y_i]}_{A} + \overbrace{\big(K_\beta(x_i^{\text{obs}}, y_i) - \mathbb{E}K_\beta(x_i^{\text{obs}}, y_i)\big)\beta}^{B}. \tag{86}$$

For the j-th coordinate of A, we have

$$A_j = \bar{m}_j^i y_i - \mathbb{E}[\bar{m}_j y]. \tag{87}$$

We note that $\bar{m}_j$ is a zero-mean sub-Gaussian random variable with $\|\bar{m}_j\|_{\psi_2} \leq C(1 + kr)$ (see Lemma B.3 in Wang et al. (2015))

Lemma 18 *Under the assumption of Lemma 6, for each $j \in [d]$, $\bar{m}_j$ is sub-Gaussian with mean zero and $\|\bar{m}_j\|_{\psi_2} \leq C(1 + kr)$.*

Thus, by Lemma 13 we have

$$\|\bar{m}_j y_i\|_{\psi_1} \leq C \max\{\|\bar{m}_j\|_{\psi_2}^2, \|y\|_{\psi_2}^2\} \leq C' \max\{(1 + kr)^2, \sigma^2 + \|\beta^*\|_2^2\}, \tag{88}$$

where the last inequality is due to the fact that $y = \langle \beta^*, x \rangle + v$. Thus, $\|y\|_{\psi_2}^2 \leq C_3(\|\langle \beta^*, x \rangle\|_{\psi_2}^2 + \|v\|_{\psi_2}^2)$ for some C_3.

For term B, we have

$$\bar{K}_j^i = \underbrace{(1 - z_{i,j})\beta_j}_{C} + \underbrace{\sum_{k=1}^{d} \bar{m}_j^i \bar{m}_k^i \beta_k}_{D} - \underbrace{\sum_{k=1}^{d} [(1 - z_{i,j})\bar{m}_j^i][(1 - z_{i,k})\bar{m}_k^i]\beta_k}_{E}. \tag{89}$$

For term C, we have the following (by Example 5.8 in Vershynin (2010))

$$\|(1 - z_{i,j})\beta_j\|_{\psi_2} \leq |\beta_j| \leq \|\beta\|_\infty \leq (1 + k)\sqrt{s}\|\beta^*\|_2. \tag{90}$$

For term D, by Lemma 18 and 13 we have

$$\|\sum_{k=1}^{d}\bar{m}_j^i\bar{m}_k^i\beta_k\|_{\psi_1}\leq\sum_{k=1}^{d}|\beta_k|\|\bar{m}_j^i\bar{m}_k^i\|_{\psi_1}\leq\sum_{k=1}^{d}|\beta_k|C^2(1+kr)^2\leq C_4(1+kr)^2\|\beta\|_1. \quad (91)$$

Since $\beta\in\mathcal{B}$, we get $\|\beta\|_1\leq\sqrt{s}\|\beta\|_2\leq(1+k)\sqrt{s}\|\beta^*\|_2$. Thus, we have

$$\|\sum_{k=1}^{d}\bar{m}_j^i\bar{m}_k^i\beta_k\|_{\psi_1}\leq C_4\sqrt{s}(1+kr)^2\|\beta^*\|_2. \quad (92)$$

For term E, since $1-z_i\in[0,1]$, we have $\|(1-z_{i,j})\bar{m}_j^i\|_{\psi_2}\leq\|\bar{m}_j^i\|_{\psi_2}\leq C(1+kr)$. Hence, by Lemma 13 we get

$$\begin{aligned}\|\sum_{k=1}^{d}[(1-z_{i,j})\bar{m}_j^i][(1-z_{i,k})\bar{m}_k^i]\beta_k\|_{\psi_1}&\leq\sum_{k=1}^{d}|\beta_k|\|[(1-z_{i,j})\bar{m}_j^i][(1-z_{i,k})\bar{m}_k^i]\|_{\psi_1}\\&\leq\sum_{k=1}^{d}|\beta_k|C(1+kr)^2\leq C_6(1+kr)^2\sqrt{s}\|\beta^*\|_2.\end{aligned} \quad (93)$$

This gives us

$$\|\bar{K}_j^i\|_{\psi_1}\leq C_7\sqrt{s}(1+k)(1+kr)^2\|\beta^*\|_2. \quad (94)$$

By Lemma 12, we get

$$\|[\nabla q_i-\nabla q]_j\|_{\psi_1}\leq 2\|[\nabla q_i]_j\|_{\psi_1}\leq C_8[(1+k)(1+kr)^2\sqrt{s}\|\beta^*\|_2+\max\{(1+kr)^2,\sigma^2+\|\beta^*\|_2^2\}]. \quad (95)$$