High-dimensional (Group) Adversarial Training in Linear Regression

Yiling Xie
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia, USA
yxie350@gatech.edu

Xiaoming Huo
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia, USA
huo@gatech.edu

Abstract

Adversarial training can achieve robustness against adversarial perturbations and has been widely used in machine-learning models. This paper delivers a non-asymptotic consistency analysis of the adversarial training procedure under ℓ_∞-perturbation in high-dimensional linear regression. It will be shown that, under the restricted eigenvalue condition, the associated convergence rate of prediction error can achieve the *minimax* rate up to a logarithmic factor in the high-dimensional linear regression on the class of sparse parameters. Additionally, the group adversarial training procedure is analyzed. Compared with classic adversarial training, it will be proved that the group adversarial training procedure enjoys a better prediction error upper bound under certain group-sparsity patterns.

1 Introduction

Adversarial training is proposed to hedge against adversarial perturbations and has attracted much research interest in recent years. Adversarial training has been widely used in Large Language Models [14, 24], computer vision [13], cybersecurity [33], etc. While the empirical risk minimization procedure optimizes the empirical loss, the adversarial training procedure seeks conservative solutions that optimize the worst-case loss under a given magnitude of perturbation. People have actively investigated model modifications and algorithmic frameworks to improve performance and training efficiency for adversarial training under different problem settings [1, 10, 12, 22, 23, 27, 29].

We are interested in understanding the fundamental properties of adversarial training from a statistical viewpoint. A standard approach for statisticians to evaluate statistical or machine-learning models is to investigate whether the estimator obtained from the model can achieve the minimax rate [25]. In this paper, we will give the non-asymptotic convergence rate of the prediction error in high-dimensional adversarial training. The associated convergence rate achieves the minimax rate under certain settings, which will be clarified in Section 2.2.

In machine-learning models, adversarial training has the following mathematical formulation:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \sup_{\|\Delta\| \leq \delta} L\left(\boldsymbol{X}_i + \Delta, Y_i, \beta\right),$$

where $(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_n, Y_n)$ are given samples, Δ is the perturbation, $\|\cdot\|$ is the perturbation norm, δ is the perturbation magnitude, β is the model parameter, and $L(\boldsymbol{x}, y, \beta)$ is the loss function with $\boldsymbol{x}$ being the input variable and y being the response variable.

Regarding the choice of the perturbation norm, we focus on ℓ_∞-perturbation, i.e., $\|\Delta\| = \|\Delta\|_\infty$. Some literature has pointed out that ℓ_∞-perturbation could help recover the model sparsity [21, 28].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

For example, [28] has proved that the asymptotic distribution of adversarial training estimator under ℓ_∞-perturbation has a positive mass at 0 when the underlying parameter is 0. Since the sparsity assumption could improve the model interpretation and reduce problem complexity [11], especially in high-dimensional regimes, ℓ_∞-perturbation will be studied, and certain sparsity patterns will be assumed in this paper. In terms of the loss function, we focus on the loss in the linear regression, i.e., $L(\boldsymbol{x}, y, \beta) = (\boldsymbol{x}^\top \beta - y)^2$. In particular, many of the existing theoretical explorations on adversarial training are based on linear models [15, 16, 21, 28, 30, 31], which admits advanced analytical analysis and sheds light on the characteristics of adversarial training in more general settings and applications. In this regard, the linear regression is considered in this paper. In short, we will focus on the adversarial-trained linear regression as shown in the following:

$$\widehat{\beta} \in \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \max_{\|\Delta\|_\infty \leq \delta} \left((\boldsymbol{X}_i + \Delta)^\top \beta - Y_i \right)^2, \tag{1}$$

where $\boldsymbol{X}_i, \beta \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$. For the convenience of analysis, we write the given n samples $(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_n, Y_n)$ in the matrix form: $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_n)^\top \in \mathbb{R}^{n \times p}$ and $\boldsymbol{Y} = (Y_1, ..., Y_n)^\top \in \mathbb{R}^{n \times 1}$, where we call $\boldsymbol{X}$ the design matrix.

This paper delivers the convergence analysis under the high-dimension setting, where we suppose the dimension of the model parameter β is larger than the sample size, i.e., $p > n$. Further, the parameter sparsity is assumed. Specifically, we suppose that only a subset of the elements of the p-dimensional ground-truth parameter β_* is nonzero. If the size of the nonzero subset is s, it will be shown that the resulting prediction error of problem (1), i.e., $\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2^2/n$, is of the order $s \log p/n$ under the restricted eigenvalue condition. The restricted eigenvalue condition is a standard assumption in the literature of sparse high-dimensional linear regression [2, 3, 4, 17]. Notably, the rate $s \log p/n$ is optimal in the minimax sense, up to a logarithmic factor, for all estimators over a class of s-sparse p-dimensional vectors if there are n training samples [20].

Our aforementioned results have the following implications. Firstly, in addition to robustness towards perturbation, our results show that adversarial training is beneficial regarding statistical optimality. This means the resulting estimator can achieve the minimax convergence rate for prediction error. To the best of our knowledge, we are the first to prove the minimax optimality of the adversarial training estimator. Secondly, our analysis illustrates that the ℓ_∞-perturbation is recommended if the sparsity condition, i.e., the ground truth β_* is supported on a subset of $\{1, ..., p\}$, is known and a fast theoretical error convergence rate is required.

The convergence rate of the group adversarial training is also investigated. In the literature, the group effect has been studied in (finite-type) Wasserstein distributionally robust optimization problems [5, 7]. Since adversarial training is equivalent to ∞-type Wasserstein distributionally robust optimization problem [9], the formulation of group Wasserstein distributionally robust optimization problem discussed in [5, 7] can be generalized to the adversarial training problem. We give a formal formulation of the group adversarial training problem based on frameworks in [5, 7]. Further, we derive the non-asymptotic convergence rate of the prediction error for the group adversarial training problem. It will be shown that group adversarial training can achieve a faster convergence upper bound if certain group sparsity structures are satisfied. The details can be found in Section 3.2.

1.1 Related Work

We review and compare some related work in this subsection.

The asymptotic behavior of ℓ_∞-adversarial training estimator in the generalized linear model has been discussed in [28]. Notably, the paper [28] studies the behavior of adversarial training estimator from an asymptotic point of view, while our paper delivers a non-asymptotic analysis. More specifically, analysis in [28] is based on the asymptotic distribution of the adversarial training estimator, while our work is to give a non-asymptotic upper bound of the prediction error of the adversarial training estimator. More discussions can be found in Remark 2.7.

The prediction error of ℓ_∞-adversarial training estimator has been briefly analyzed in [21], where the proven convergence is of the order $1/\sqrt{n}$ in terms of n. The results in our paper are different in the following two perspectives. Firstly, a faster convergence rate of the order $1/n$ in terms of n is given, and the associated rate is minimax optimal up to a logarithmic factor. Secondly, we have incorporated the sparsity setting in the model analysis, while no sparsity pattern is considered in

theoretical analysis for ℓ_∞-adversarial training in [21]. More discussions can be found in Remark 2.8.

The paper [30] also investigates the convergence of adversarial training estimator in linear regression. The derivations in [30] are based on the assumption that the input variable $\boldsymbol{X}$ follows p-dimensional Gaussian distribution while our analysis imposes the restricted eigenvalue condition. In addition, notice that [30] argues the superiority of incorporating the sparsity information by deriving lower bounds for the estimator error while we directly prove the rate optimality of the adversarial training estimator under the sparsity assumption. Also, [30] applies ℓ_2-perturbation while our work focuses on ℓ_∞-perturbation.

In the literature, it has been proven that multiple estimators, including LASSO, Dantzig selector, and square-root LASSO, can achieve the minimax rate (up to a logarithmic factor) in high-dimensional sparse linear regression [3, 4, 11, 20]. However, to the best of our knowledge, no literature has investigated this property for the widely used adversarial training model. We are the first to study whether the adversarial training estimator can be minimax optimal, and our theoretical analysis implies that the answer is yes, i.e., the adversarial training estimator under ℓ_∞-perturbation enjoys rate optimality. In addition, the group lasso has been intensely studied to explore the parameter group structure [17, 32] while group adversarial training imposes group structure on the perturbation. It will be shown that the group adversarial training estimator shares a similar convergence rate with the group LASSO estimator. Our proof technique is developed upon and extends the technical methods in the aforementioned papers [3, 4, 11, 17].

1.2 Notations and Preliminaries

We introduce some notations, which will be used in the rest of the paper. For vector $\boldsymbol{z} \in \mathbb{R}^p$, we use $\|\boldsymbol{z}\|_q$ to denote the ℓ_q norm of the vector $\boldsymbol{z}$, i.e., $\|\boldsymbol{z}\|_q^q = \sum_{j=1}^p |\boldsymbol{z}_j|^q, 1 \leq q < \infty, \|\boldsymbol{z}\|_\infty = \max_{1\leq j\leq p} |\boldsymbol{z}_j|$. We use $e_j \in \mathbb{R}^p, 1 \leq j \leq p$, to denote the basis vectors where the jth component is 1 and 0 otherwise. $I_n \in \mathbb{R}^{n\times n}$ denotes the identity matrix. For some set S, we use S^c to denote the complement set of S and $|S|$ to denote the cardinality of S. If the set S is the subset of $\{1, ..., p\}$, we use $\boldsymbol{z}_S \in \mathbb{R}^{|S|}$ to denote the subvector indexed by elements of S.

We clarify some preliminary settings that will be used in this paper. Throughout this paper, we suppose the high-dimension setting holds, the samples are generated from the Gaussian linear model, and the design matrix is normalized. We conclude these conditions with the following assumptions:

Assumption 1.1 (High-dimension)**.** *The parameter dimension p is larger than the sample size n, i.e., we have $n < p$.*

Assumption 1.2 (Gaussian linear model)**.** *The design matrix $\boldsymbol{X}$ is fixed and the response vector $\boldsymbol{Y}$ is generated by the following: $\boldsymbol{Y} = \boldsymbol{X}\beta_* + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ has i.i.d. entries $\mathcal{N}(0, \sigma^2)$.*

Assumption 1.3 (Normalization)**.** *The design matrix $\boldsymbol{X}$ is normalized such that $\|\boldsymbol{X}e_j\|_2 \leq \sqrt{n}$, for $1 \leq j \leq p$.*

1.3 Organization of this Paper

The remainder of this paper is organized as follows. In Section 2, we derive the convergence rate of the adversarial training estimator in high-dimensional linear regression. In Section 3, we derive the convergence rate of group adversarial training and compare it with the existing adversarial training. Numerical experiments are conducted in Section 4. Possible future work is discussed in Section 5. The proofs are relegated to the Appendix whenever possible.

2 ℓ_∞-Adversarial-Trained Linear Regression

In this section, we will first introduce the problem formulation of the adversarial training in linear regression under ℓ_∞-perturbation and then deliver the convergence analysis of the prediction error under ℓ_∞-perturbation in the high-dimensional setting.

2.1 Problem Formulation

In this subsection, we give the problem formulation of ℓ_∞-adversarial-trained linear regression and discuss its dual.

Recall that the ℓ_∞-adversarial training problem in linear regression has the formulation shown in (1). The solution $\widehat{\beta}$ to the optimization problem (1) is used to estimate the ground-truth data generating parameter β_*, seeing Assumption 1.2. In the inner optimization problem, we compute the worst-case square loss between the response variable and the linear prediction among the perturbations. The perturbations are added to the input variable and with the largest ℓ_∞-norm δ. In the outer optimization problem, we optimize the empirical expectation of the worst-case loss of given samples.

The optimization problem (1) can be further simplified by considering its dual formulation, which is shown as follows [21, Proposition 1].

Proposition 2.1 (Dual Formulation of problem (1))**.** *If we denote the optimal value of problem* (1) *by* $R(\delta)$*, then we have that*

$$R(\delta) = \min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(|\boldsymbol{X}_i^\top \beta - Y_i| + \delta \|\beta\|_1 \right)^2. \tag{2}$$

We discuss the advantages and theoretical insights we could get by considering the dual problem (2). Note that the dual formulation (2) removes the inner maximization of problem (1), and the associated objective function is a convex function of β. Thus, it will be more convenient to solve the dual problem (2) than the primal problem (1). Also, the expansion of the objective function in (2) yields the following:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{X}_i^\top \beta - Y_i)^2 + \delta \|\beta\|_1 \left(\frac{2}{n} \sum_{i=1}^{n} |\boldsymbol{X}_i^\top \beta - Y_i| \right) + \delta^2 \|\beta\|_1^2, \tag{3}$$

where the residual term $\delta^2 \|\beta\|_1^2$ will be of high order if we let δ, for example, be proportional to the inverse of a positive power of n. Regardless of the high order residual term, the objective function in problem (2) can be viewed as the sum of the loss function in linear regression and a regularization term depending on $\|\beta\|_1$. This implies that ℓ_∞-adversarial-trained linear regression has a regularization effect. We refer to [21, 28] and references therein for more discussions about the regularization effect of adversarial training. Since the well-known LASSO is formulated by imposing the ℓ_1-norm regularization term and enjoys the minimax convergence rate of the prediction error [20], the dual formulation (2) of ℓ_∞-adversarial training in linear regression and its expansion (3) may indicate a fast convergence of its prediction error for the adversarial training estimator.

2.2 Convergence Analysis

In this subsection, we will first introduce the restricted eigenvalue condition and then derive the convergence rate of the prediction error for the adversarial training estimator in high-dimensional linear regression under the restricted eigenvalue condition and ℓ_∞-perturbation. We will also discuss the high-probability arguments upon which we prove the optimality of the associated adversarial training estimator.

Before we deliver the convergence analysis, we make the following assumption.

Assumption 2.2 (Restricted Eigenvalue Condition)**.** *The matrix* $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ *satisfies the restricted eigenvalue condition if there exists a positive number* $\gamma = \gamma(s, c_1) > 0$ *such that*

$$\min \left\{ \frac{\|\boldsymbol{X} v\|_2}{\sqrt{n} \|v\|_2} : |S| \leq s, v \in \mathbb{R}^p \backslash \{\mathbf{0}\}, \|v_{S^c}\|_1 \leq c_1 \|v_S\|_1 \right\} \geq \gamma,$$

where S *is some subset of* $\{1, ..., p\}$.

In the sequel, we use the notation $\mathrm{RE}(s, c_1)$ to denote the restricted eigenvalue condition w.r.t. the cardinality s of the index set S and the constant c_1 in the constrained cone, i.e., $\|v_{S^c}\|_1 \leq c_1 \|v_S\|_1$. The restricted eigenvalue condition can be considered as a relaxation of the positive semidefiniteness of the gram matrix $\boldsymbol{X}^\top \boldsymbol{X}$ and is a useful technique in theoretical analysis in the sparse high-dimensional analysis [11].

Equipped with Assumption 2.2, we have the following convergence result of prediction error.

Theorem 2.3 (Prediction Error Analysis for Adversarial Training)**.** *Suppose the adversarial training problem* (1) *satisfies*

$$\frac{2\|\boldsymbol{X}^\top \boldsymbol{\epsilon}\|_\infty}{\|\boldsymbol{\epsilon}\|_1} \leq \delta. \tag{4}$$

If β_* is supported on a subset S of $\{1, ..., p\}$ where $|S| \leq s$, and the design matrix $\boldsymbol{X}$ satisfies $\mathrm{RE}(s, 3)$ *with parameter $\gamma(s, 3)$, then we have that*

$$\frac{1}{2n}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2^2 \leq 3\delta^2 s \max\left\{\frac{9}{\gamma^2(s,3)}\left(\frac{\|\boldsymbol{\epsilon}\|_1}{n}\right)^2, 164\|\beta_*\|_2^2\right\}$$

Theorem 2.3 shows that the upper bound of the prediction error mainly depends on the sparsity cardinality s of the ground-truth parameter β_* and perturbation magnitude δ. The perturbation magnitude δ is assumed to be equal to or larger than $2\|\boldsymbol{X}^\top \boldsymbol{\epsilon}\|_\infty / \|\boldsymbol{\epsilon}\|_1$. We could apply the concentration inequalities to give a closed-form expression of the perturbation magnitude, based on which the convergence rate of the prediction error is derived. The convergence rate holds with a high probability and can be found in the following corollary.

Corollary 2.4. *Consider the adversarial training problem* (1) *with perturbation magnitude*

$$\delta = \frac{4}{\sqrt{\frac{2}{\pi}} - \frac{1}{10}}\sqrt{\frac{\log p}{n}}. \tag{5}$$

If β_* is supported on a subset S of $\{1, ..., p\}$ where $|S| \leq s$, and the design matrix $\boldsymbol{X}$ satisfies $\mathrm{RE}(s, 3)$ *with parameter $\gamma(s, 3)$, then we have that*

$$\frac{1}{2n}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2^2 \leq 192\frac{s\log p}{n}\max\left\{\frac{9}{\gamma^2(s,3)}\left(\frac{\|\boldsymbol{\epsilon}\|_1}{n}\right)^2, 164\|\beta_*\|_2^2\right\} \tag{6}$$

holds with a probability greater than $1 - 2\exp(-C_1 n) - 2/p$, where C_1 is a positive constant.

Remark 2.5. *We discuss the choice of δ. Corollary 2.4 implies that the perturbation magnitude δ should be of the order $1/\sqrt{n}$ in order to derive the non-asymptotic convergence rate* (6)*. The associated order is consistent with the asymptotic analysis in [28], where the sparsity-recovery ability could be proven in the asymptotic sense if the sample size is of the order $1/\sqrt{n}$.*

In Corollary 2.4, we choose δ as is shown in (5). Under this setting, it can be proven that the inequality (4) holds with a high probability. Then, we adopt Theorem 2.3 and could have the expression of the prediction error in terms of p and n as shown in (6).

The convergence rate could be further simplified in the following corollary if both the error variance σ^2 and the ℓ_2-norm of the ground-truth parameter β_* are bounded.

Corollary 2.6. *Under the assumptions stated in Corollary 2.4, suppose there exists a finite positive constant R such that*

$$2\sqrt{41}\|\beta_*\|_2 \leq R, \quad \sigma < \frac{1}{6}\gamma(s,3)R,$$

then we have that

$$\frac{1}{2n}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2^2 \leq 192\frac{s\log p}{n}R^2$$

holds with a probability greater than $1 - 2/p - 2\exp(-C_1 n) - \exp(-n)$, where C_1 is a positive constant.

Remark 2.7. *Corollary 2.6 investigates the behavior of the adversarial training estimator under ℓ_∞-perturbation by computing the resulting prediction error while [28] studies the behavior of the adversarial training estimator under ℓ_∞-perturbation by deriving the associated limiting distribution. Both [28] and our results consider the sparsity condition. [28] proves that ℓ_∞-adversarial training can help recover sparsity asymptotically if the parameter sparsity is known while our paper, i.e., Corollary 2.6, provides a fast non-asymptotic convergence rate for prediction error under the sparsity condition.*

Remark 2.8. *Corollary 2.6 illustrates that the convergence rate of prediction error for ℓ_∞-adversarial training in linear regression is of the order $s \log p/n$ while the prediction error shown in [21] has a lower order $1/\sqrt{n}$ in terms of n. Our paper achieves a faster rate by incorporating the sparsity information and applying the restricted eigenvalue condition.*

Corollary 2.6 implies that the prediction error of high-dimensional ℓ_∞-adversarial-trained estimator is of the order $s \log p/n$. This order is optimal up to a logarithmic factor in the minimax sense for any estimators over a class of s-sparse vectors in $\mathbb{R}^p$ when n samples are given [2, 20].

3 Group Adversarial Training

This section will elaborate on the formulation of group adversarial training and the associated convergence rate. Also, we compare group adversarial training under $(2, \infty)$-perturbation with classic adversarial training under ℓ_∞-perturbation.

Since the adversarial training forces the perturbation with uniform magnitude to each component of the input variable, it may not perform well if the input variable has a group effect. The group structure exists in many real-world problems. For example, groups of genes act together in pathways in gene-expression arrays [18], and financial data can be grouped by different sectors and industries to help market prediction [6]. Also, if an input variable is a multilevel factor and dummy variables are introduced, then these dummy variables act in a group [32]. Group adversarial training can tackle the group effect by adding group-structured perturbation. The detailed formulation can be seen in Section 3.1.

3.1 Problem Formulation

In this subsection, we describe the formulation of the group adversarial training.

Suppose the input variable $\boldsymbol{x}$ can be divided into L non-overlapped groups. Then, we have the definition of the group-structured weighted norm accordingly in the following proposition, where the associated dual norm is also stated.

Proposition 3.1 (Proposition 5 in [5], Theorem 2.2 in [7])**.** *Consider a vector $\boldsymbol{x} = (\boldsymbol{x}^1, ..., \boldsymbol{x}^L)$, where each $\boldsymbol{x}^l \in \mathbb{R}^{p_l}$, and $\sum_{l=1}^L p_l = p$. Define the weighted (r, s)-norm of $\boldsymbol{x}$ with the L-dimensional weight vector $\boldsymbol{\omega} = (\omega_1, ..., \omega_L)$ to be:*

$$\|\boldsymbol{x}_{\boldsymbol{\omega}}\|_{r,s} = \left(\sum_{l=1}^{L} \|\omega_l \boldsymbol{x}^l\|_r^s \right)^{1/s}, 1 \le s < \infty,$$

$$\|\boldsymbol{x}_{\boldsymbol{\omega}}\|_{r,\infty} = \max_{1 \le l \le L} \|\omega_l \boldsymbol{x}^l\|_r, \quad s = \infty,$$

where $\omega_l > 0, \forall l$ and $r \ge 1$. Then, the dual norm of (r, s)-norm with weight $\boldsymbol{\omega}$ is the (q, t)-norm with weight $\boldsymbol{\omega}^{-1} = (1/\omega_1, ..., 1/\omega_L)$, i.e. $\|\boldsymbol{x}_{\boldsymbol{\omega}^{-1}}\|_{q,t}$, where $1/r + 1/q = 1$ and $1/s + 1/t = 1$.

To handle the group structure in the input variable, the weighted (r, s)-norm is applied to add group structure in the perturbation accordingly, and the group adversarial training is formulated as follows,

$$\min_\beta \frac{1}{n} \sum_{i=1}^{n} \sup_{\|\Delta_{\boldsymbol{\omega}}\|_{r,s} \le \delta} \left((\boldsymbol{X}_i + \Delta)^\top \beta - Y_i \right)^2,$$

where $\boldsymbol{\omega} = (\omega_1, ..., \omega_L)$.

Recall we focus on adversarial training problems under ℓ_∞-perturbation, high-dimension setting, and sparsity condition. Under this consideration, we let $s = \infty$ and $r = 2$, and then the associated group adversarial training problem under $(2, \infty)$-perturbation has the following expression:

$$\min_\beta \frac{1}{n} \sum_{i=1}^{n} \sup_{\|\Delta_{\boldsymbol{\omega}}\|_{2,\infty} \le \delta} \left((\boldsymbol{X}_i + \Delta)^\top \beta - Y_i \right)^2. \tag{7}$$

To facilitate convenience for the computation and analysis, similar to our study towards classic adversarial training in Section 2.1, we derive the dual formulation of problem (7) in the following proposition. One can check that the corresponding objective in the dual formulation (8) is also convex.

Proposition 3.2 (Dual Formulation of problem (7))**.** *If we denote the optimal value of problem* (7) *by* $\widetilde{R}(\delta)$*, then we have that*

$$\widetilde{R}(\delta) = \min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(|(\boldsymbol{X}_i + \Delta)^\top \beta - Y_i| + \delta \|\beta_{\boldsymbol{\omega}^{-1}}\|_{2,1} \right)^2, \tag{8}$$

where

$$\|\beta_{\boldsymbol{\omega}^{-1}}\|_{2,1} = \sum_{l=1}^{L} \frac{1}{\omega_l} \|\beta^l\|_2.$$

3.2 Convergence Analysis

In this subsection, we deliver the convergence analysis of the prediction error of the estimator obtained from group adversarial training under $(2, \infty)$-perturbation, i.e., problem (7).

First, we clarify some notations for subsequent analysis. In terms of the group structure of the input variable and the perturbation, we focus on non-overlapped cases. Assume that the index set $\{1, ..., p\}$ has the prescribed (disjoint) partition $\{1, ..., p\} = \bigcup_{l=1}^{L} G_l$. We use p_l to denote the cardinality of each group, i.e., $|G_l| = p_l$.

Consider the group sparsity structure in the ground-truth parameter $\beta_* \in \mathbb{R}^p$, where sparsity is imposed at the group level instead of on individual components. Specially, the set $J \subset \{1, ..., L\}$ denotes a set of groups and β_* is supported at these J groups, i.e., β_* is supported on the $G_J = \bigcup_{l \in J} G_l$.

We make the following assumption before we proceed to derive the convergence analysis.

Assumption 3.3 (Group Restricted Eigenvalue Condition)**.** *The matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ satisfies the group restricted eigenvalue condition if there exists a positive number $\kappa = \kappa(g, c_2) > 0$ such that*

$$\min \left\{ \frac{\|\boldsymbol{X} v\|_2}{\sqrt{n} \|v_{G_J}\|_2} : |J| \leq g, v \in \mathbb{R}^p \backslash \{\mathbf{0}\}, \sum_{l \in J^c} \frac{1}{\omega_l} \|v^l\|_2 \leq c_2 \sum_{l \in J} \frac{1}{\omega_l} \|v^l\|_2 \right\} \geq \kappa,$$

where J is some subset of $\{1, ..., L\}$.

In the sequel, we use the notation $\mathrm{GRE}(g, c_2)$ to denote the restricted eigenvalue condition w.r.t. the cardinality g of the index set J and the constant c_2 in the constrained cone, i.e., $\sum_{l \in J^c} \frac{1}{\omega_l} \|v^l\|_2 \leq c_2 \sum_{l \in J} \frac{1}{\omega_l} \|v^l\|_2$. Group restricted eigenvalue condition is an extension of the restricted eigenvalue condition and can be used in the theoretical analysis for the group LASSO, seeing [17].

Theorem 3.4 (Prediction Error Analysis for Group Adversarial Training)**.** *Consider the group adversarial training problem* (7) *satisfying*

$$\frac{2\| \left(\boldsymbol{X}^\top \boldsymbol{\epsilon} \right)^l \|_2}{\|\boldsymbol{\epsilon}\|_1} \leq \frac{\delta}{\omega_l}, \quad \forall l.$$

If β_* is supported on a subset G_J of $\{1, ..., p\}$ where $|J| \leq g$, and the design matrix $\boldsymbol{X}$ satisfies $\mathrm{GRE}(g, 3)$ *with parameter $\kappa(g, 3)$, then we have that*

$$\frac{1}{2n} \|\boldsymbol{X}(\widetilde{\beta} - \beta_*)\|_2^2 \leq 3\delta^2 \sum_{l \in J} \frac{1}{\omega_l^2} \max \left\{ \frac{9}{\kappa^2(g, 3)} \left(\frac{\|\boldsymbol{\epsilon}\|_1}{n} \right)^2, 164 \|\beta_*\|_2^2 \right\},$$

where $\widetilde{\beta}$ is the estimator obtained from solving problem (7).

Theorem 3.4 shows that the upper bound of the prediction error mainly depends on the weight $\boldsymbol{\omega}$ and perturbation magnitude δ. We apply the arguments in concentration inequalities and obtain the convergence rate in the following corollary.

Corollary 3.5. *Consider the group adversarial training problem* (7) *satisfying*

$$\frac{\delta}{\omega_l} = \frac{2}{\sqrt{\frac{2}{\pi}} - \frac{1}{10}} \sqrt{\frac{3p_l + 9 \log L}{n}}, \forall l,$$

and $\Psi_l = \boldsymbol{X}_{G_l}^\top \boldsymbol{X}_{G_l}/n = I_{p_l \times p_l}$, where $\boldsymbol{X}_{G_l}$ denotes the $n \times p_l$ sub-matrix of $\boldsymbol{X}$ formed by the columns indexed by G_l. If β_* is supported on a subset G_J of $\{1, ..., p\}$ where $|J| \leq g$, and the design matrix $\boldsymbol{X}$ satisfies $\mathrm{GRE}(g, 3)$ with parameter $\kappa(g, 3)$, then we have that

$$\frac{1}{2n}\|\boldsymbol{X}(\widetilde{\beta} - \beta_*)\|_2^2 \leq 432\frac{|G_J| + g\log L}{n}\max\left\{\frac{9}{\kappa^2(s,3)}\left(\frac{\|\boldsymbol{\epsilon}\|_1}{n}\right)^2, 164\|\beta_*\|_2^2\right\}$$

holds with a probability greater than $1 - 2\exp(-C_1 n) - 2/L$, where C_1 is a positive constant.

Remark 3.6. *Note that we assume the gram matrix satisfies that $\boldsymbol{X}_{G_l}^\top \boldsymbol{X}_{G_l}/n = I_{p_l \times p_l}$. This is a standard assumption in the theoretical analysis in sparse high-dimensional linear regression, seeing [17, 19].*

Similar to the analytic investigations in Section 2, the convergence rate of the prediction error could be further simplified in the following corollary if the ℓ_2-norm of the ground-truth parameter β_* and error variance σ^2 are bounded.

Corollary 3.7. *Under the assumptions stated in Corollary 3.5, suppose there exists a finite positive constant R such that*

$$2\sqrt{41}\|\beta_*\|_2 \leq R, \quad \sigma < \frac{1}{6}\kappa(g, 3)R,$$

then we have that

$$\frac{1}{2n}\|\boldsymbol{X}(\widetilde{\beta} - \beta_*)\|_2^2 \leq 432\frac{|G_J| + g\log L}{n}R^2$$

holds with a probability greater than $1 - 2/L - 2\exp(-C_1 n) - \exp(-n)$, where C_1 is a positive constant.

Remark 3.8. *If we make the number of groups equal to p, i.e., each group only has one component, then we will have that $L = p, p_l = 1, |G_J| = g$. The resulting error bound is $g\log p/n$, where g denotes the number of nonzero components of β_*. This order matches what is derived in Corollary 2.6.*

Corollary 3.7 indicates that the upper bound of the associated prediction error in group adversarial training under $(2, \infty)$-perturbation is of the order $(|G_J| + g\log L)/n$. Recall that L is the number of prescribed groups for the p-dimensional variable, the ground-truth parameter $\beta_* \in \mathbb{R}^p$ is supported by a subset of the L groups, and the subset is denoted by $J \subset \{1, ..., L\}$. The cardinality of the support subset J is g. We also use $G_J \subset \{1, ..., p\}$ to denote all the indexes included in J.

It follows from Corollary 2.6 that the convergence rate of the prediction error for the classic adversarial training under ℓ_∞-perturbation is of the order $s\log p/n$, where s denotes the cardinality of the support set of β_*. Then, we can conclude that if $|G_J|/s \ll \log p$ and $g \ll s$, the group adversarial training is superior to the classic adversarial training. In essence, if the sparsity pattern of β_* has a good group structure, i.e., most of the nonzero components can be captured in J, then the group adversarial training procedure can provide an improved upper bound for the prediction error.

4 Numerical Experiments

In this section, we will run numerical experiments to observe the empirical performances of (group) adversarial training in high-dimensional linear regression.

We consider the following models to generate synthetic data: The response variable Y is generated by the Gaussian linear model, as stated in Assumption 1.2. The standard deviation of the error $\boldsymbol{\epsilon}$ is chosen as 0.1. In Model 1, the ground truth parameter β_* is a 500-dimensional vector. The first four components are $[0.1, 0.2, 0.15, 0.25]$, the last four components are $[0.9, 0.95, 1, 1.05]$, and the other components are zero. In Model 2, the ground truth parameter β_* is a 600-dimensional vector. The first three components of β_* are $[0.4, 0.5, 0.6]$. The last three components of β_* correspond to dummy variables generated from a four-level categorical factor. These dummy variable components are assigned values $[0.2, 0.3, 0.7]$. The other components are zero.

We run adversarial training under ℓ_∞-perturbation and group adversarial training under $(2, \infty)$-perturbation to give the estimations for the ground-truth parameter β_*, respectively. As suggested in Corollary 2.4 and Corollary 3.5, the perturbation magnitude is chosen in the order of $1/\sqrt{n}$ in the

adversarial training; the ratio of the perturbation magnitude and the perturbation weight is chosen in the order of $1/\sqrt{n}$ in the group adversarial training. For the constant, we selected 1 for simplicity and experimental convenience. For the group adversarial training, we divide the parameter equally into 125 groups of size 4 for Model 1 and 200 groups of size 3 for Model 2. The sample sizes are chosen $\{50, 100, 150, 200, 250, 300, 350, 400\}$ for Model 1 and $\{50, 100, 150, 200, 250, 350, 450, 550\}$ for Model 2. In terms of computation, we apply the dual formulations, i.e., problem (2) and problem (8), and solve these convex optimization problems using the CVXPY toolbox [8]. Five random samples are generated at each sample size, and we run (group) adversarial training for each sample. The mean and standard error of the coefficient estimations and prediction errors are computed and recorded.

We first plot the coefficient estimation paths of adversarial training with error bars in Figure 1 and Figure 2. Both adversarial training and group adversarial training can shrink the parameter estimation, while group adversarial training performs a better shrinkage effect, In addition, the final values that the coefficients converge to are annotated in the figures. Given the ground-truth non-zero values $[0.1, 0.15, 0.2, 0.25, 0.9, 0.95, 1, 1.05]$ and $[0.4, 0.5, 0.6, 0.2, 0.3, 0.7]$, the final values of group adversarial training are closer to the ground-truth, indicating that the group adversarial training output more accurate estimations.

We also plot the curve of $\log_{10}(\text{prediction error})$ versus $\log_{10}(\text{sample size})$ with error bars in Figure 3 and Figure 4. We can observe that the slopes of two curves are approximately equal to -1, which is consistent with our theoretical analysis, where we have proved that the prediction error for high-dimensional (group) adversarial training is of the order $1/n$ in terms of the sample size n. Further, the curves and error bars of group adversarial training are below those of adversarial training, indicating the superiority of group adversarial training. This conclusion is also consistent with our theoretical analysis that if the model has a good group structure, group adversarial training has a lower order of prediction error.

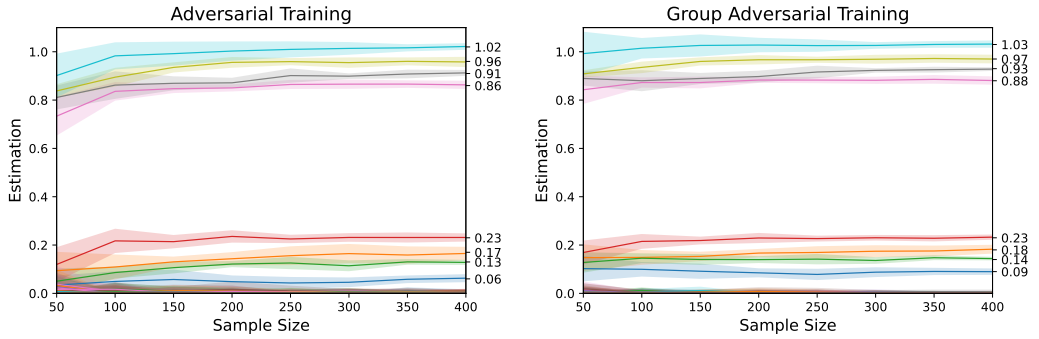

Figure 1: Coefficient Estimation Path in Model 1

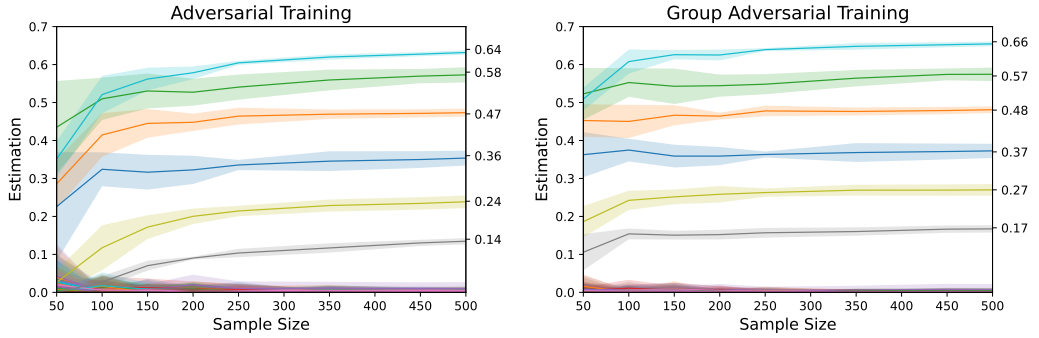

Figure 2: Coefficient Estimation Path in Model 2

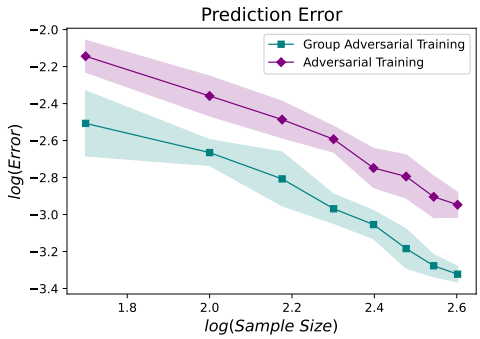

Figure 3: Prediction Error in Model 1

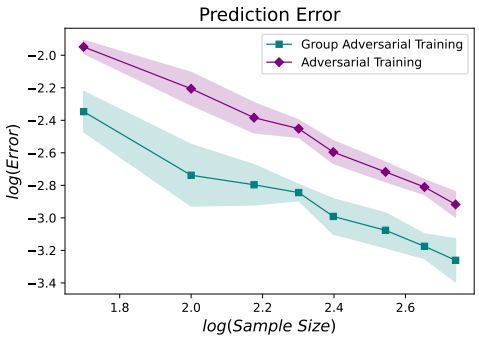

Figure 4: Prediction Error in Model 2

5 Discussions

This paper reveals the statistical optimality of adversarial training under ℓ_∞-perturbation in high dimensional linear regression and discusses potential improvements that can be achieved by group adversarial training. In the future, we may generalize the analysis in linear regression to broader statistical models, e.g., the generalized linear model and other parametric models. Also, since the prediction errors are investigated in this paper, we will consider analyzing estimation errors as our next step. More advanced analytical future work may use the primal-dual witness technique for the non-asymptotic variable-selection analysis.

Acknowledgments and Disclosure of Funding

The authors would like to thank the chairs and anonymous reviewers for their careful comments, which helped enhance the presentation of the manuscript. The authors are partially sponsored by a subcontract of NSF grant 2229876, the A. Russell Chandler III Professorship at Georgia Institute of Technology, and NIH-sponsored Georgia Clinical & Translational Science Alliance.

References

[1] M. Andriushchenko and N. Flammarion. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33:16048–16059, 2020.

[2] P. C. Bellec, G. Lecué, and A. B. Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018.

[3] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[4] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[5] J. Blanchet and Y. Kang. Distributionally robust groupwise regularization estimator. In *Asian Conference on Machine Learning*, pages 97–112. PMLR, 2017.

[6] M. J. A. Chan-Lau. *Lasso regressions and forecasting models in applied stress testing*. International Monetary Fund, 2017.

[7] R. Chen and I. C. Paschalidis. Robust grouped variable selection using distributionally robust optimization. *Journal of Optimization Theory and Applications*, 194(3):1042–1071, 2022.

[8] S. Diamond and S. Boyd. CVXPY: A python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[9] R. Gao and A. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

[10] Y. Gao, L. Qin, Z. Song, and Y. Wang. A sublinear adversarial training algorithm. In *The Twelfth International Conference on Learning Representations*, 2024.

[11] T. Hastie, R. Tibshirani, and M. Wainwright. Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143(143):8, 2015.

[12] D. Hendrycks, K. Lee, and M. Mazeika. Using pre-training can improve model robustness and uncertainty. In *International conference on machine learning*, pages 2712–2721. PMLR, 2019.

[13] C. Herrmann, K. Sargent, L. Jiang, R. Zabih, H. Chang, C. Liu, D. Krishnan, and D. Sun. Pyramid adversarial training improves vit performance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13419–13429, 2022.

[14] X. Hu, P.-Y. Chen, and T.-Y. Ho. Radar: Robust AI-text detection via adversarial learning. *Advances in Neural Information Processing Systems*, 36:15077–15095, 2023.

[15] A. Javanmard and M. Soltanolkotabi. Precise statistical analysis of classification accuracies for adversarial training. *The Annals of Statistics*, 50(4):2127–2156, 2022.

[16] A. Javanmard, M. Soltanolkotabi, and H. Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pages 2034–2078. PMLR, 2020.

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

[18] G. Malenová, D. Rowson, and V. Boeva. Exploring pathway-based group lasso for cancer survival analysis: a special case of multi-task learning. *Frontiers in Genetics*, 12:771301, 2021.

[19] Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.

[20] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on information theory*, 57(10):6976–6994, 2011.

[21] A. Ribeiro, D. Zachariah, F. Bach, and T. Schön. Regularization properties of adversarially-trained linear regression. *Advances in Neural Information Processing Systems*, 36, 2023.

[22] A. Robey, F. Latorre, G. J. Pappas, H. Hassani, and V. Cevher. Adversarial training should be cast as a non-zero-sum game. In *The Twelfth International Conference on Learning Representations*, 2024.

[23] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein. Adversarial training for free! *Advances in neural information processing systems*, 32, 2019.

[24] E. Shayegani, M. A. A. Mamun, Y. Fu, P. Zaree, Y. Dong, and N. Abu-Ghazaleh. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*, 2023.

[25] V. N. Vapnik, V. Vapnik, et al. Statistical learning theory. 1998.

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

[27] Z. Wang, T. Pang, C. Du, M. Lin, W. Liu, and S. Yan. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, pages 36246–36263. PMLR, 2023.

[28] Y. Xie and X. Huo. Asymptotic behavior of adversarial training estimator under ℓ_∞-perturbation. *arXiv preprint arXiv:2401.15262*, 2024.

[29] Y. Xing, Q. Song, and G. Cheng. On the algorithmic stability of adversarial training. *Advances in neural information processing systems*, 34:26523–26535, 2021.

[30] Y. Xing, R. Zhang, and G. Cheng. Adversarially robust estimate and risk analysis in linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 514–522. PMLR, 2021.

[31] D. Yin, R. Kannan, and P. Bartlett. Rademacher complexity for adversarially robust generalization. In *International conference on machine learning*, pages 7085–7094. PMLR, 2019.

[32] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

[33] S. Zhou, C. Liu, D. Ye, T. Zhu, W. Zhou, and P. S. Yu. Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. *ACM Computing Surveys*, 55(8):1–39, 2022.

Appendix

A Proof of Theorem 2.3

We first give the upper bound of the $\|\widehat{\beta}\|_1$ in terms of $\|\beta_*\|_1$.

Lemma A.1 (Upper Bound of $\|\widehat{\beta}\|_1$)**.** *Under conditions stated in Theorem 2.3, we have that*

$$\|\widehat{\beta}\|_1 \leq 9\|\beta_*\|_1.$$

Proof. The proof of this lemma follows a similar approach to the proof of Theorem 2 in [21]. It follows from the first-order condition of dual formulation (2) of the adversarial training problem that

$$\mathbf{0} = \frac{1}{n}\boldsymbol{X}^\top(\boldsymbol{X}\widehat{\beta} - \boldsymbol{Y}) + \delta^2\|\widehat{\beta}\|_1 w + \frac{\delta}{n}\|\boldsymbol{X}\widehat{\beta} - \boldsymbol{Y}\|_1 w + \frac{\delta}{n}\|\widehat{\beta}\|_1 \boldsymbol{X}^\top z, \tag{9}$$

where

$$z_i = \partial|\boldsymbol{X}_i^\top\widehat{\beta} - Y|, w_i = \partial|\widehat{\beta}|_i.$$

Then, we take the dot product of both sides of equation (9) with $\widehat{\beta} - \beta_*$ and could have the following:

$$\begin{aligned}
&\frac{1}{n}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2^2 \\
=&\frac{1}{n}(\boldsymbol{X}(\widehat{\beta} - \beta_*))^\top\boldsymbol{\epsilon} - \delta^2\|\widehat{\beta}\|_1 w^\top(\widehat{\beta} - \beta_*) - \frac{\delta}{n}\|\boldsymbol{X}\widehat{\beta} - \boldsymbol{Y}\|_1 w^\top(\widehat{\beta} - \beta_*) - \frac{\delta}{n}\|\widehat{\beta}\|_1(\boldsymbol{X}(\widehat{\beta} - \beta_*))^\top z, \\
=&\frac{1}{n}(\boldsymbol{X}(\widehat{\beta} - \beta_*))^\top\boldsymbol{\epsilon} - \delta^2\|\widehat{\beta}\|_1^2 + \delta^2\|\widehat{\beta}\|_1 w^\top\beta_* - \frac{\delta}{n}\|\boldsymbol{X}\widehat{\beta} - \boldsymbol{Y}\|_1\|\widehat{\beta}\|_1 + \frac{\delta}{n}\|\boldsymbol{X}\widehat{\beta} - \boldsymbol{Y}\|_1 w^\top\beta_* \\
&- \frac{\delta}{n}\|\widehat{\beta}\|_1\|\boldsymbol{X}\widehat{\beta} - \boldsymbol{Y}\|_1 - \frac{\delta}{n}\|\widehat{\beta}\|_1\boldsymbol{\epsilon}^\top z \\
\overset{(a)}{\leq}&\frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1\|\widehat{\beta} - \beta_*\|_1 - \delta^2\|\widehat{\beta}\|_1^2 + \delta^2\|\widehat{\beta}\|_1\|\beta_*\|_1 - \frac{\delta}{n}\|\boldsymbol{X}\widehat{\beta} - \boldsymbol{Y}\|_1\|\widehat{\beta}\|_1 + \frac{\delta}{n}\|\boldsymbol{X}\widehat{\beta} - \boldsymbol{Y}\|_1\|\beta_*\|_1 \\
&- \frac{\delta}{n}\|\widehat{\beta}\|_1\|\boldsymbol{X}\widehat{\beta} - \boldsymbol{Y}\|_1 + \frac{\delta}{n}\|\widehat{\beta}\|_1\|\boldsymbol{\epsilon}\|_1 \\
\overset{(b)}{\leq}&\frac{\delta}{n}\|\widehat{\beta}\|_1(\|\boldsymbol{\epsilon}\|_1 - 2\|\boldsymbol{X}\widehat{\beta} - \boldsymbol{Y}\|_1) + \frac{\delta}{n}\|\boldsymbol{X}\widehat{\beta} - \boldsymbol{Y}\|_1\|\beta_*\|_1 + \frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1(\|\widehat{\beta}\|_1 + \|\beta_*\|_1) \\
&- \delta^2\|\widehat{\beta}\|_1^2 + \delta^2\|\widehat{\beta}\|_1\|\beta_*\|_1 \\
\overset{(c)}{\leq}&\frac{\delta}{n}\|\widehat{\beta}\|_1(\frac{5}{3}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_1 - \frac{2}{3}\|\boldsymbol{\epsilon}\|_1) + \frac{\delta}{n}\|\boldsymbol{X}\widehat{\beta} - \boldsymbol{Y}\|_1\|\beta_*\|_1 + \frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1(\|\widehat{\beta}\|_1 + \|\beta_*\|_1) \\
&- \delta^2\|\widehat{\beta}\|_1^2 + \delta^2\|\widehat{\beta}\|_1\|\beta_*\|_1 \\
\overset{(d)}{\leq}&\frac{\delta}{\sqrt{n}}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2(\frac{5}{3}\|\widehat{\beta}\|_1 + \|\beta_*\|_1) + \frac{\delta}{n}\|\boldsymbol{\epsilon}\|_1(-\frac{1}{6}\|\widehat{\beta}\|_1 + \frac{3}{2}\|\beta_*\|_1) - \delta^2\|\widehat{\beta}\|_1^2 + \delta^2\|\widehat{\beta}\|_1\|\beta_*\|_1,
\end{aligned} \tag{10}$$

where (a) comes from the Hölder's inequality, i.e.,

$$(\boldsymbol{X}(\widehat{\beta} - \beta_*))^\top\boldsymbol{\epsilon} \leq \|\widehat{\beta} - \beta_*\|_1\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\|_\infty,\ w^\top\beta_* \leq \|w\|_\infty\|\beta_*\|_1 = \|\beta_*\|_1,\ -\boldsymbol{\epsilon}^\top z \leq \|\boldsymbol{\epsilon}\|_1\|z\|_\infty = \|\boldsymbol{\epsilon}\|_1,$$

and the condition $2\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\|_\infty/\|\boldsymbol{\epsilon}\|_1 \leq \delta$, (b) comes from $\|\widehat{\beta} - \beta_*\|_1 \leq \|\widehat{\beta}\|_1 + \|\beta_*\|_1$, (c) comes from the relationship that

$$2\|\boldsymbol{X}\widehat{\beta} - \boldsymbol{Y}\|_1 \geq \frac{5}{3}\|\boldsymbol{X}\widehat{\beta} - \boldsymbol{Y}\|_1 \geq -\frac{5}{3}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_1 + \frac{5}{3}\|\boldsymbol{\epsilon}\|_1,$$

and (d) comes from the inequality that $\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_1 \leq \sqrt{n}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2$ and $\|\boldsymbol{X}\widehat{\beta} - \boldsymbol{Y}\|_1 \leq \|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_1 + \|\boldsymbol{\epsilon}\|_1$.

One may observe that (10) is a second-order inequality of variable $\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2/\sqrt{n}$, then the associate discriminant should be equal to or larger than 0, resulting in

$$\frac{1}{9}\delta^2(-11\|\widehat{\beta}\|_1^2 + 66\|\beta_*\|_1\|\widehat{\beta}\|_1 + 9\|\beta_*\|_1^2) + \frac{\delta}{3n}\|\boldsymbol{\epsilon}\|_1(-2\|\widehat{\beta}\|_1 + 18\|\beta_*\|_1) \geq 0,$$

from which we could conclude that at least one of two terms should be equal or larger than 0, implying

$$\|\widehat{\beta}\|_1 \leq 9\|\beta_*\|_1.$$

□

Then, we proceed to prove Theorem 2.3.

Proof. We write the objective function in (2) in the matrix norm and then have that

$$\frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\beta}\|_2^2+\frac{1}{2}\delta^2\|\widehat{\beta}\|_1^2+\frac{\delta}{n}\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\beta}\|_1\|\widehat{\beta}\|_1 \leq \frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{X}\beta_*\|_2^2+\frac{1}{2}\delta^2\|\beta_*\|_1^2+\frac{\delta}{n}\|\boldsymbol{Y}-\boldsymbol{X}\beta_*\|_1\|\beta_*\|_1. \tag{11}$$

It follows from $\boldsymbol{Y} = \boldsymbol{X}\beta_* + \boldsymbol{\epsilon}$, i.e., Assumption 1.2, that

$$\|\boldsymbol{Y} - \boldsymbol{X}\beta_*\|_2^2 = \|\boldsymbol{\epsilon}\|_2^2,$$

$$\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}\|_2^2 = \|\boldsymbol{X}(\widehat{\beta} - \beta_*) - \boldsymbol{\epsilon}\|_2^2 = \|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2^2 + \|\boldsymbol{\epsilon}\|_2^2 - 2\boldsymbol{\epsilon}^\top \boldsymbol{X}(\widehat{\beta} - \beta_*),$$

$$\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}\|_1 = \|\boldsymbol{X}(\widehat{\beta} - \beta_*) - \boldsymbol{\epsilon}\|_1 \geq \|\boldsymbol{\epsilon}\|_1 - \|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_1.$$

In this way, the inequality (11) could be reformulated as

$$\begin{aligned}
&\frac{1}{2n}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2^2 \\
&\leq \frac{1}{n}\boldsymbol{\epsilon}^\top \boldsymbol{X}(\widehat{\beta} - \beta_*) + \frac{\delta}{n}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_1\|\widehat{\beta}\|_1 + \frac{\delta}{n}\|\boldsymbol{\epsilon}\|_1\left(\|\beta_*\|_1 - \|\widehat{\beta}\|_1\right) + \frac{1}{2}\delta^2\|\beta_*\|_1^2 - \frac{1}{2}\delta^2\|\widehat{\beta}\|_1^2 \\
&\overset{(a)}{\leq} \frac{1}{n}\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\|_\infty\|\widehat{\beta} - \beta_*\|_1 + \frac{\delta}{n}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_1\|\widehat{\beta}\|_1 + \frac{\delta}{n}\|\boldsymbol{\epsilon}\|_1\left(\|\beta_*\|_1 - \|\widehat{\beta}\|_1\right) + \frac{1}{2}\delta^2\|\beta_*\|_1^2 - \frac{1}{2}\delta^2\|\widehat{\beta}\|_1^2 \\
&\overset{(b)}{\leq} \frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1\|\widehat{\beta} - \beta_*\|_1 + \frac{\delta}{n}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_1\|\widehat{\beta}\|_1 + \frac{\delta}{n}\|\boldsymbol{\epsilon}\|_1\left(\|\beta_*\|_1 - \|\widehat{\beta}\|_1\right) + \frac{1}{2}\delta^2\|\beta_*\|_1^2 - \frac{1}{2}\delta^2\|\widehat{\beta}\|_1^2 \\
&\overset{(c)}{\leq} \frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1\|\widehat{\beta} - \beta_*\|_1 + \frac{\delta}{\sqrt{n}}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2\|\widehat{\beta}\|_1 + \frac{\delta}{n}\|\boldsymbol{\epsilon}\|_1\left(\|\beta_*\|_1 - \|\widehat{\beta}\|_1\right) + \frac{1}{2}\delta^2\|\beta_*\|_1^2 - \frac{1}{2}\delta^2\|\widehat{\beta}\|_1^2,
\end{aligned} \tag{12}$$

where (a) comes from the Hölder's inequality, (b) comes from the condition $2\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\|_\infty/\|\boldsymbol{\epsilon}\|_1 \leq \delta$, (c) comes from $\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_1 \leq \sqrt{n}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2$.

Then, we begin to give the upper bound of the prediction error. Two cases should be discussed.

First Case:

$$\|\widehat{\beta} - \beta_*\|_1 + 2\|\beta_*\|_1 - 2\|\widehat{\beta}\|_1 \leq 0. \tag{13}$$

In this case, we have that

$$\|\beta_*\|_1 - \|\widehat{\beta}\|_1 = \|\widehat{\beta}\|_1 - \|\beta_*\|_1 + 2\|\beta_*\|_1 - 2\|\widehat{\beta}\|_1 \leq \|\widehat{\beta} - \beta_*\|_1 + 2\|\beta_*\|_1 - 2\|\widehat{\beta}\|_1 \leq 0. \tag{14}$$

Then, it follows from (12) that

$$\begin{aligned}
&\frac{1}{2n}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2^2 \\
&\leq \frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1\|\widehat{\beta} - \beta_*\|_1 + \frac{\delta}{\sqrt{n}}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2\|\widehat{\beta}\|_1 + \frac{\delta}{n}\|\boldsymbol{\epsilon}\|_1\left(\|\beta_*\|_1 - \|\widehat{\beta}\|_1\right) + \frac{1}{2}\delta^2\|\beta_*\|_1^2 - \frac{1}{2}\delta^2\|\widehat{\beta}\|_1^2 \\
&= \frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1\left(\|\widehat{\beta} - \beta_*\|_1 + 2\|\beta_*\|_1 - 2\|\widehat{\beta}\|_1\right) + \frac{\delta}{\sqrt{n}}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2\|\widehat{\beta}\|_1 + \frac{1}{2}\delta^2\|\beta_*\|_1^2 - \frac{1}{2}\delta^2\|\widehat{\beta}\|_1^2 \\
&\leq \frac{\delta}{\sqrt{n}}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2\|\widehat{\beta}\|_1,
\end{aligned}$$

where the last inequality comes from (13) and (14).

Lemma A.1 indicates that

$$\frac{1}{\sqrt{2n}}\|\boldsymbol{X}(\widehat{\beta} - \beta_*)\|_2 \leq \sqrt{2}\delta\|\widehat{\beta}\|_1 \leq 9\sqrt{2}\delta\|\beta_*\|_1,$$

which is equivalent to

$$\frac{1}{2n}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2^2 \leq 162\delta^2\|\beta_*\|_1^2 \leq 162\delta^2|S|\|\beta_*\|_2^2 \leq 162\delta^2 s\|\beta_*\|_2^2. \tag{15}$$

Second Case:

$$\|\widehat{\beta}-\beta_*\|_1 + 2\|\beta_*\|_1 - 2\|\widehat{\beta}\|_1 \geq 0. \tag{16}$$

If we let $\widehat{v} = \widehat{\beta} - \beta_*$, then we have that

$$\|\beta_* - \widehat{\beta}\|_1 = \|\widehat{v}\|_1 = \|\widehat{v}_S\|_1 + \|\widehat{v}_{S^c}\|_1,$$

$$\|\beta_*\|_1 - \|\widehat{\beta}\|_1 = \|\beta_{*S}\|_1 - (\|\beta_{*S} + \widehat{v}_S\|_1 + \|\widehat{v}_{S^c}\|_1) \leq \|\widehat{v}_S\|_1 - \|\widehat{v}_{S^c}\|_1.$$

The inequality (16) indicates that

$$3\|\widehat{v}_S\|_1 - \|\widehat{v}_{S^c}\|_1 \geq \|\widehat{\beta}-\beta_*\|_1 + 2\|\beta_*\|_1 - 2\|\widehat{\beta}\|_1 \geq 0, \tag{17}$$

implying

$$\|\widehat{v}_{S^c}\|_1 \leq 3\|\widehat{v}_S\|_1.$$

In this way, the $\mathrm{RE}(s,3)$ condition can be applied.

In addition, it follows the inequality (12) that

$$\begin{aligned}
&\frac{1}{2n}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2^2 \\
&\leq \frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1\|\widehat{\beta}-\beta_*\|_1 + \frac{\delta}{\sqrt{n}}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2\|\widehat{\beta}\|_1 + \frac{\delta}{n}\|\boldsymbol{\epsilon}\|_1\left(\|\beta_*\|_1 - \|\widehat{\beta}\|_1\right) + \frac{1}{2}\delta^2\|\beta_*\|_1^2 - \frac{1}{2}\delta^2\|\widehat{\beta}\|_1^2 \\
&\overset{(a)}{\leq} \frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1\|\widehat{\beta}-\beta_*\|_1 + \frac{1}{4n}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2^2 + \delta^2\|\widehat{\beta}\|_1^2 + \frac{\delta}{n}\|\boldsymbol{\epsilon}\|_1\left(\|\beta_*\|_1 - \|\widehat{\beta}\|_1\right) + \frac{1}{2}\delta^2\|\beta_*\|_1^2 - \frac{1}{2}\delta^2\|\widehat{\beta}\|_1^2 \\
&= \frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1\left(\|\widehat{\beta}-\beta_*\|_1 + 2\|\beta_*\|_1 - 2\|\widehat{\beta}\|_1\right) + \frac{1}{4n}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2^2 + \frac{1}{2}\delta^2\|\beta_*\|_1^2 + \frac{1}{2}\delta^2\|\widehat{\beta}\|_1^2 \\
&\overset{(b)}{\leq} \frac{3\delta}{2n}\|\boldsymbol{\epsilon}\|_1\|\widehat{v}_S\|_1 + \frac{1}{4n}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2^2 + 41\delta^2\|\beta_*\|_1^2,
\end{aligned}$$

where (a) comes from the inequality

$$\frac{1}{4n}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2^2 + \delta^2\|\widehat{\beta}\|_1^2 \geq \frac{\delta}{\sqrt{n}}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2\|\widehat{\beta}\|_1,$$

(b) comes from (17) and Lemma A.1.

Further, we have that

$$\begin{aligned}
\frac{1}{4n}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2^2 &\leq \frac{3\delta}{2n}\sqrt{|S|}\|\boldsymbol{\epsilon}\|_1\|\widehat{v}\|_2 + 41\delta^2|S|\|\beta_*\|_2^2 \\
&\leq \frac{3\delta}{2\sqrt{n}}\frac{\sqrt{s}}{\gamma(s,3)}\frac{\|\boldsymbol{\epsilon}\|_1}{n}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2 + 41\delta^2 s\|\beta_*\|_2^2,
\end{aligned} \tag{18}$$

where the first inequality comes from $\|\widehat{v}_S\|_1 \leq \sqrt{|S|}\|\widehat{v}\|_2$ and $\|\beta_*\|_1 \leq \sqrt{|S|}\|\beta_*\|_2$, and the last inequality comes from the $\mathrm{RE}(s,3)$ condition and $|S| \leq s$.

Then, if we solve the inequality (18), we could have that

$$\frac{1}{\sqrt{n}}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2 \leq \frac{\delta\sqrt{s}}{\gamma(s,3)}\left(3\frac{\|\boldsymbol{\epsilon}\|_1}{n} + \sqrt{9\left(\frac{\|\boldsymbol{\epsilon}\|_1}{n}\right)^2 + 164\gamma^2(s,3)\|\beta_*\|_2^2}\right),$$

which is equivalent to

$$\frac{1}{\sqrt{n}}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2 \leq (1+\sqrt{2})\frac{\delta\sqrt{s}}{\gamma(s,3)}\max\left\{3\frac{\|\boldsymbol{\epsilon}\|_1}{n}, \sqrt{164}\gamma(s,3)\|\beta_*\|_2\right\},$$

indicating that

$$\frac{1}{2n}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2^2 \leq 3\frac{\delta^2 s}{\gamma^2(s,3)}\max\left\{9\left(\frac{\|\boldsymbol{\epsilon}\|_1}{n}\right)^2, 164\gamma^2(s,3)\|\beta_*\|_2^2\right\}. \tag{19}$$

Combining (15) and (19), we have that

$$\frac{1}{2n}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2^2 \leq 3\delta^2 s\max\left\{\frac{9}{\gamma^2(s,3)}\left(\frac{\|\boldsymbol{\epsilon}\|_1}{n}\right)^2, 164\|\beta_*\|_2^2\right\}.$$

□

B Proof of Corollary 2.4

Proof. To give the high probability result, we should analyze the bound of the term $\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\|_\infty/n$ and $\|\boldsymbol{\epsilon}\|_1/n$. The associated arguments are discussed in the following two parts, respectively. (Note that we assume $\boldsymbol{\epsilon}$ has i.i.d. Gaussian entries. However, our high-probability result can be extended to sub-Gaussian cases, as sub-Gaussian variables share similar tail decay behavior.)

Part I: We focus on the tail bound of $\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\|_\infty/n$. Since the design matrix $\boldsymbol{X}$ is normalized, the random variable $\boldsymbol{x}_j^\top\boldsymbol{\epsilon}/n$ is stochastically dominated by $\mathcal{N}(0,\sigma^2/n)$. As shown in Theorem 11.1 in [11], it follows from the Gaussian tail bound and the union bound that

$$\mathbb{P}\left(\frac{\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\|_\infty}{n}\geq t\right)\leq 2\exp\left(-\frac{nt^2}{2\sigma^2}+\log p\right).$$

In this way, with a probability greater than $1-2/p$, the following holds

$$\frac{\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\|_\infty}{n}\leq 2\sigma\sqrt{\frac{\log p}{n}}.$$

Part II: We focus on the concentration inequality of $\|\boldsymbol{\epsilon}\|_1/n$. It follows from general Hoeffding's inequality [26] that

$$\mathbb{P}\left(\left|\|\boldsymbol{\epsilon}\|_1-n\sigma\sqrt{\frac{2}{\pi}}\right|\geq t\right)\leq 2\exp\left(-C_2\frac{t^2}{n\sigma^2}\right),$$

indicating

$$\mathbb{P}\left(\left|\frac{1}{n}\|\boldsymbol{\epsilon}\|_1-\sigma\sqrt{\frac{2}{\pi}}\right|\geq t\right)\leq 2\exp\left(-C_2\frac{nt^2}{\sigma^2}\right),$$

where C_2 is some positive constant. By choosing $t=\sigma/10$, we have that

$$\mathbb{P}\left(\sigma\left(\sqrt{\frac{2}{\pi}}-\frac{1}{10}\right)\leq\frac{1}{n}\|\boldsymbol{\epsilon}\|_1\right)\geq 1-2\exp\left(-C_1 n\right),$$

where C_1 is some positive constant. This is to say, with a probability greater than $1-2\exp(-C_1 n)$, we have that

$$\sigma\left(\sqrt{\frac{2}{\pi}}-\frac{1}{10}\right)\leq\frac{1}{n}\|\boldsymbol{\epsilon}\|_1.$$

Suppose we let

$$\delta=\frac{4}{\sqrt{\frac{2}{\pi}}-\frac{1}{10}}\sqrt{\frac{\log p}{n}},$$

with a probability greater than $1-2\exp(-C_1 n)-2/p$, we have that

$$\frac{2\|\boldsymbol{X}^\top\boldsymbol{\epsilon}\|_\infty}{\|\boldsymbol{\epsilon}\|_1}\leq\delta.$$

It follows from Theorem 2.3 that

$$\frac{1}{2n^2}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2^2\leq 192\frac{s\log p}{n}\max\left\{\frac{9}{\gamma^2(s,3)}\left(\frac{\|\boldsymbol{\epsilon}\|_1}{n}\right)^2, 164\|\beta_*\|_2^2\right\}$$

holds with a probability greater $1-2\exp(-C_1 n)-2/p$. □

C Proof of Corollary 2.6

Proof. We focus on the tail bound of $\|\boldsymbol{\epsilon}\|_1/n$. We apply the Chernoff bound to give the tail bound of $\|\boldsymbol{\epsilon}\|_1$, and we have that

$$\mathbb{P}\left(\|\boldsymbol{\epsilon}\|_1 \geq t\right) \leq \inf_{s>0} M(s)\exp(-ts),$$

where $M(s)$ is the moment-generating function of $\|\boldsymbol{\epsilon}\|_1$. We also could obtain that

$$M_i(s) = \mathbb{E}[\exp(s|\boldsymbol{\epsilon}_i|)] \leq 2\mathbb{E}[\exp(s\boldsymbol{\epsilon}_i)] = 2\exp\left(\frac{\sigma^2 s^2}{2}\right),$$

indicating

$$M(s) \leq 2^n \exp\left(\frac{n\sigma^2 s^2}{2}\right).$$

Then, we have that

$$\mathbb{P}\left(\|\boldsymbol{\epsilon}\|_1 \geq t\right) \leq \inf_{s>0} 2^n \exp\left(\frac{n\sigma^2 s^2}{2} - ts\right),$$

indicating

$$\mathbb{P}\left(\frac{1}{n}\|\boldsymbol{\epsilon}\|_1 \geq t\right) \leq \exp\left(-\frac{nt^2}{2\sigma^2} + n\log 2\right).$$

In this way, with a probability greater $1-\exp(-n)$, the following holds:

$$\frac{1}{n}\|\boldsymbol{\epsilon}\|_1 \leq \sqrt{2\log 2 + 2}\sigma \leq 2\sigma.$$

Since we have that $\frac{1}{6}\gamma(s,3)R \geq \sigma$,

$$\frac{9}{\gamma^2(s,3)}\left(\frac{\|\boldsymbol{\epsilon}\|_1}{n}\right)^2 \leq R^2,$$

holds with a probability greater $1-\exp(-n)$. Due to Corollary 2.4, we have the following

$$\frac{1}{2n^2}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2^2 \leq 192\frac{s\log p}{n}R^2$$

holds with a probability greater $1-2/p-2\exp(-C_1 n)-\exp(-n)$. □

D Proof of Proposition 3.2

It follows from Proposition 1 in [21] that

$$\widetilde{R}(\delta) = \min_{\beta}\frac{1}{n}\sum_{i=1}^{n}\left(|(\boldsymbol{X}_i+\Delta)^\top\beta - Y_i| + \delta\|\beta_{\boldsymbol{\omega}^{-1}}\|_*\right)^2,$$

where $\|\cdot\|_*$ denotes the dual norm of $(2,\infty)$-norm of $\beta_{\boldsymbol{\omega}}$. Due to Proposition 3.1, we conclude that (8) holds.

E Proof of Theorem 3.4

We first give the upper bound of the $\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}$ in terms of $\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}$.

Lemma E.1 (Upper Bound of $\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}$)**.** *Under conditions stated in Theorem 3.4, we have that*

$$\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} \leq 9\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}.$$

Proof. The proof of this lemma follows a similar approach to the proof of Lemma A.1. It follows from the first-order condition of the dual formulation (8) of the group adversarial training problem that

$$\mathbf{0} = \frac{1}{n}\boldsymbol{X}^\top(\boldsymbol{X}\widetilde{\beta}-\boldsymbol{Y}) + \delta^2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}t + \frac{\delta}{n}\|\boldsymbol{X}\widetilde{\beta}-\boldsymbol{Y}\|_1 t + \frac{\delta}{n}\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}\boldsymbol{X}^\top z, \tag{20}$$

where

$$z_i = \partial |X_i^\top \widetilde{\beta} - Y|,$$
$$t = \partial \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}, \quad t^l = \frac{1}{\omega_l} \frac{\widetilde{\beta}^l}{\|\widetilde{\beta}^l\|_2}.$$

Notice we have that

$$t^\top \widetilde{\beta} = \sum_{l=1}^{L} \frac{1}{\omega_l} \frac{(\widetilde{\beta}^l)^\top \widetilde{\beta}^l}{\|\widetilde{\beta}^l\|_2} = \sum_{l=1}^{L} \frac{1}{\omega_l} \|\widetilde{\beta}^l\|_2 = \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}, \tag{21}$$

and it follows from the Hölder's inequality that

$$t^\top \beta_* = \sum_{l=1}^{L} \frac{1}{\omega_l} \frac{(\widetilde{\beta}^l)^\top \beta_*^l}{\|\widetilde{\beta}^l\|_2} \leq \sum_{l=1}^{L} \frac{1}{\omega_l} \|\beta_*^l\|_2 = \|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}. \tag{22}$$

Also, we have the following from the Hölder's inequality:

$$(\boldsymbol{X}(\widetilde{\beta} - \beta_*))^\top \boldsymbol{\epsilon} \leq \sum_{l=1}^{L} \|(\boldsymbol{X}^\top \boldsymbol{\epsilon})^l\|_2 \|\widetilde{\beta}^l - \beta_*^l\|_2 \leq \frac{1}{2}\delta \|\boldsymbol{\epsilon}\|_1 \sum_{l=1}^{L} \frac{1}{\omega_l} \|\widetilde{\beta}^l - \beta_*^l\|_2 = \frac{1}{2}\delta \|\boldsymbol{\epsilon}\|_1 \|(\widetilde{\beta}^l - \beta_*^l)_{\boldsymbol{\omega}^{-1}}\|_{2,1}, \tag{23}$$

where the second inequality comes from the condition $2\| \left(\boldsymbol{X}^\top \boldsymbol{\epsilon}\right)^l \|_2 / \|\boldsymbol{\epsilon}\|_1 \leq \delta / \omega_l$.

Then, we take the dot product of both sides of equation (20) with $\widetilde{\beta} - \beta_*$ and could have the following:

$$\begin{aligned}
&\frac{1}{n}\|\boldsymbol{X}(\widetilde{\beta} - \beta_*)\|_2^2 \\
=&\frac{1}{n}(\boldsymbol{X}(\widetilde{\beta} - \beta_*))^\top \boldsymbol{\epsilon} - \delta^2 \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} t^\top (\widetilde{\beta} - \beta_*) - \frac{\delta}{n}\|\boldsymbol{X}\widetilde{\beta} - \boldsymbol{Y}\|_1 t^\top (\widetilde{\beta} - \beta_*) \\
&- \frac{\delta}{n}\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} (\boldsymbol{X}(\widetilde{\beta} - \beta_*))^\top z, \\
\overset{(a)}{=}&\frac{1}{n}(\boldsymbol{X}(\widetilde{\beta} - \beta_*))^\top \boldsymbol{\epsilon} - \delta^2 \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}^2 + \delta^2 \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} t^\top \beta_* - \frac{\delta}{n}\|\boldsymbol{X}\widetilde{\beta} - \boldsymbol{Y}\| \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} \\
&+ \frac{\delta}{n}\|\boldsymbol{X}\widetilde{\beta} - \boldsymbol{Y}\|_1 t^\top \beta_* - \frac{\delta}{n}\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} \|\boldsymbol{X}\widetilde{\beta} - \boldsymbol{Y}\|_1 + \frac{\delta}{n}\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} \boldsymbol{\epsilon}^\top z \\
\overset{(b)}{\leq}&\frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1 \|(\widetilde{\beta}^l - \beta_*^l)_{\boldsymbol{\omega}^{-1}}\|_{2,1} - \delta^2 \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}^2 + \delta^2 \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} \|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1} - \frac{\delta}{n}\|\boldsymbol{X}\widetilde{\beta} - \boldsymbol{Y}\|_1 \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} \\
&+ \frac{\delta}{n}\|\boldsymbol{X}\widetilde{\beta} - \boldsymbol{Y}\|_1 \|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1} - \frac{\delta}{n}\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} \|\boldsymbol{X}\widetilde{\beta} - \boldsymbol{Y}\|_1 + \frac{\delta}{n}\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} \boldsymbol{\epsilon}^\top z \\
\overset{(c)}{\leq}&\frac{\delta}{n}\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} (\|\boldsymbol{\epsilon}\|_1 - 2\|\boldsymbol{X}\widetilde{\beta} - \boldsymbol{Y}\|_1) + \frac{\delta}{n}\|\boldsymbol{X}\widetilde{\beta} - \boldsymbol{Y}\|_1 \|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1} \\
&+ \frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1 (\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} + \|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}) - \delta^2 \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}^2 + \delta^2 \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} \|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1} \\
\overset{(d)}{\leq}&\frac{\delta}{n}\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} (\frac{5}{3}\|\boldsymbol{X}(\widetilde{\beta} - \beta_*)\|_1 - \frac{2}{3}\|\boldsymbol{\epsilon}\|_1) + \frac{\delta}{n}\|\boldsymbol{X}\widetilde{\beta} - \boldsymbol{Y}\|_1 \|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1} \\
&+ \frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1 (\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} + \|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}) - \delta^2 \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}^2 + \delta^2 \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} \|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1} \\
\overset{(e)}{\leq}&\frac{\delta}{\sqrt{n}}\|\boldsymbol{X}(\widetilde{\beta} - \beta_*)\|_2 (\frac{5}{3}\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} + \|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}) + \frac{\delta}{n}\|\boldsymbol{\epsilon}\|_1 (-\frac{1}{6}\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} + \frac{3}{2}\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}) \\
&- \delta^2 \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}^2 + \delta^2 \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} \|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1},
\end{aligned} \tag{24}$$

where (a) comes from (21) and (22), (b) comes from (23), (c) comes from $\|(\widetilde{\beta}^l - \beta_*^l)_{\boldsymbol{\omega}^{-1}}\|_{2,1} \leq \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} + \|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}$, (d) comes from $2\|\boldsymbol{X}\widetilde{\beta} - \boldsymbol{Y}\|_1 \geq \frac{5}{3}\|\boldsymbol{X}\widetilde{\beta} - \boldsymbol{Y}\|_1 \geq -\frac{5}{3}\|\boldsymbol{X}(\widetilde{\beta} - \beta_*)\|_1 + \frac{5}{3}\|\boldsymbol{\epsilon}\|_1$, (e) comes from the inequality that $\|\boldsymbol{X}(\widetilde{\beta} - \beta_*)\|_1 \leq \sqrt{n}\|\boldsymbol{X}(\widetilde{\beta} - \beta_*)\|_2$ and $\|\boldsymbol{X}\widetilde{\beta} - \boldsymbol{Y}\|_1 \leq \|\boldsymbol{X}(\widetilde{\beta} - \beta_*)\|_1 + \|\boldsymbol{\epsilon}\|_1$.

Since the inequality (24) is a second-order inequality of variable $\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2/\sqrt{n}$, then the associate discriminant should be equal to or larger than 0, resulting in

$$\frac{1}{9}\delta^2(-11\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}^2+66\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}+9\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}^2) + \frac{\delta}{3n}\|\boldsymbol{\epsilon}\|_1(-2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}+18\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1})\geq 0,$$

from which we could conclude that

$$\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}\leq 9\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}.$$

□

Then, we proceed to prove Theorem 3.4.

Proof. We write the objective function in (8) in the matrix norm and then have that

$$\begin{aligned}&\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\widetilde{\beta}\|_1^2+\frac{1}{2}\delta^2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}^2+\frac{\delta}{n}\|\boldsymbol{y}-\boldsymbol{X}\widetilde{\beta}\|\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}\\ \leq&\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\beta_*\|_1^2+\frac{1}{2}\delta^2\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}^2+\frac{\delta}{n}\|\boldsymbol{y}-\boldsymbol{X}\beta_*\|_1\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}.\end{aligned} \tag{25}$$

In this way, we have the following reformulation:

$$\begin{aligned}&\frac{1}{2n}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2^2\\ &\leq\frac{1}{n}\boldsymbol{\epsilon}^\top\boldsymbol{X}(\widetilde{\beta}-\beta_*)+\frac{\delta}{n}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_1\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}+\frac{\delta}{n}\|\boldsymbol{\epsilon}\|_1\left(\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}-\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}\right)\\ &+\frac{1}{2}\delta^2\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}^2-\frac{1}{2}\delta^2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}^2\\ &\leq\frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1\|(\widetilde{\beta}-\beta_*)_{\boldsymbol{\omega}^{-1}}\|_{2,1}+\frac{\delta}{\sqrt{n}}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}+\frac{\delta}{n}\|\boldsymbol{\epsilon}\|_1\left(\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}-\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}\right)\\ &+\frac{1}{2}\delta^2\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}^2-\frac{1}{2}\delta^2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}^2,\end{aligned} \tag{26}$$

where the last inequality comes from (23) and $\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_1\leq\sqrt{n}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2$.

Similar to the proof of Theorem 2.3, we begin to give the upper bound of the prediction error. Two cases should be discussed.

First Case:

$$\|(\widetilde{\beta}-\beta_*)_{\boldsymbol{\omega}^{-1}}\|_{2,1}+2\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}-2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}\leq 0. \tag{27}$$

In this case, we have that

$$\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}-\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}\leq\|(\widetilde{\beta}-\beta_*)_{\boldsymbol{\omega}^{-1}}\|_{2,1}+2\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}-2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}\leq 0. \tag{28}$$

It follow from (26) that

$$\begin{aligned}&\frac{1}{2n}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2^2\\ &\leq\frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1\|(\widetilde{\beta}-\beta_*)_{\boldsymbol{\omega}^{-1}}\|_{2,1}+\frac{\delta}{\sqrt{n}}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}+\frac{\delta}{n}\|\boldsymbol{\epsilon}\|_1\left(\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}-\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}\right)\\ &+\frac{1}{2}\delta^2\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}^2-\frac{1}{2}\delta^2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}^2,\\ &=\frac{\delta}{\sqrt{n}}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}+\frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1\left(\|(\widetilde{\beta}-\beta_*)_{\boldsymbol{\omega}^{-1}}\|_{2,1}+2\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}-2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}\right)\\ &+\frac{1}{2}\delta^2\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}^2-\frac{1}{2}\delta^2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}^2,\\ &\leq\frac{\delta}{\sqrt{n}}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1},\end{aligned} \tag{29}$$

where the last inequality comes from (27) and (28).

Notice we have that

$$\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1} = \sum_{l\in J}\frac{1}{\omega_l}\|\beta_*^l\|_2 \le \sqrt{\sum_{l\in J}\frac{1}{\omega_l^2}}\|\beta_*\|_2. \tag{30}$$

Lemma E.1, (29) and (30) indicate that

$$\frac{1}{2n^2}\|\boldsymbol{X}(\widehat{\beta}-\beta_*)\|_2^2 \le 162\delta^2\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}^2 \le 162\delta^2\sum_{l\in J}\frac{1}{\omega_l^2}\|\beta_*\|_2^2. \tag{31}$$

Second Case:

$$\|(\widetilde{\beta}-\beta_*)_{\boldsymbol{\omega}^{-1}}\|_{2,1} + 2\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1} - 2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} \ge 0.$$

Notice we have that

$$\|(\widetilde{\beta}-\beta_*)_{\boldsymbol{\omega}^{-1}}\|_{2,1} = \sum_{l=1}^{L}\frac{1}{\omega_l}\|\widetilde{\beta}^l-\beta_*^l\|_2 = \sum_{l\in J^c}\frac{1}{\omega_l}\|(\widetilde{\beta}-\beta_*)^l\|_2 + \sum_{l\in J}\frac{1}{\omega_l}\|(\widetilde{\beta}-\beta_*)^l\|_2,$$

$$\begin{aligned}
\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1} - \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} =& \sum_{l\in J}\frac{1}{\omega_l}\|\beta_*^l\|_2 - \sum_{l=1}^{L}\frac{1}{\omega_l}\|\widetilde{\beta}^l\|_2 \\
=& \sum_{l\in J}\frac{1}{\omega_l}\|\beta_*^l\|_2 - \left(\sum_{l\in J}\frac{1}{\omega_l}\|\widetilde{\beta}^l\|_2 + \sum_{l\in J^c}\frac{1}{\omega_l}\|(\widetilde{\beta}-\beta_*)^l\|_2\right) \\
\le& \sum_{l\in J}\frac{1}{\omega_l}\|(\widetilde{\beta}-\beta_*)^l\|_2 - \sum_{l\in J^c}\frac{1}{\omega_l}\|(\widetilde{\beta}-\beta_*)^l\|_2.
\end{aligned}$$

If we let $\widetilde{v} = \widetilde{\beta}-\beta_*$,then we have that

$$3\sum_{l\in J}\frac{1}{\omega_l}\|\widetilde{v}^l\|_2 - \sum_{l\in J^c}\frac{1}{\omega_l}\|\widetilde{v}^l\|_2 \ge \|(\widetilde{\beta}-\beta_*)_{\boldsymbol{\omega}^{-1}}\|_{2,1} + 2\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1} - 2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} \ge 0, \tag{32}$$

indicating

$$\sum_{l\in J^c}\frac{1}{\omega_l}\|\widetilde{v}^l\|_2 \le 3\sum_{l\in J}\frac{1}{\omega_l}\|\widetilde{v}^l\|_2.$$

In this way, the $\mathrm{GRE}(g,3)$ condition can be applied.

We also have the following from (26)

$$\begin{aligned}
&\frac{1}{2n}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2^2 \\
&\le \frac{\delta}{n}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_1\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} + \frac{\delta}{2n}\|\boldsymbol{\epsilon}\|_1\left(\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1} - \|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1} + 2\|(\widetilde{\beta}-\beta_*)_{\boldsymbol{\omega}^{-1}}\|_{2,1}\right) \\
&+ \frac{1}{2}\delta^2\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}^2 - \frac{1}{2}\delta^2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}^2 \\
&\overset{(a)}{\le} \frac{1}{4n}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2^2 + \frac{3\delta}{2n}\frac{\|\boldsymbol{\epsilon}\|_1}{n}\sum_{l\in J}\frac{1}{\omega_l}\|\widetilde{v}^l\|_2 + \frac{1}{2}\delta^2\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}^2 + \frac{1}{2}\delta^2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}^2, \\
&\overset{(b)}{\le} \frac{1}{4n}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2^2 + \frac{3\delta}{2n}\frac{\|\boldsymbol{\epsilon}\|_1}{n}\sqrt{\sum_{l\in J}\frac{1}{\omega_l^2}}\|\widetilde{v}_{G_J}\|_2 + \frac{1}{2}\delta^2\|\beta_{*\boldsymbol{\omega}^{-1}}\|_{2,1}^2 + \frac{1}{2}\delta^2\|\widetilde{\beta}_{\boldsymbol{\omega}^{-1}}\|_{2,1}^2
\end{aligned}$$

where (a) comes from (32), and (b) comes from

$$\sum_{l\in J}\frac{1}{\omega_l}\|\widetilde{v}^l\|_2 \le \sqrt{\sum_{l\in J}\frac{1}{\omega_l^2}}\|\widetilde{v}_{G_J}\|_2.$$

It follows from $\mathrm{GRE}(g,3)$ and Lemma E.1 that

$$\frac{1}{4n}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2^2 \le \frac{3\delta}{2\sqrt{n}}\sqrt{\sum_{l\in J}\frac{1}{\omega_l^2}}\frac{1}{\kappa(g,3)}\frac{\|\boldsymbol{\epsilon}\|_1}{n}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2 + 41\delta^2\sum_{l\in J}\frac{1}{\omega_l^2}\|\beta_*\|_2^2.$$

Then, we could have that

$$\frac{1}{2n}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2^2 \le 3\frac{\delta^2}{\kappa^2(g,3)}\sum_{l\in J}\frac{1}{\omega_l^2}\max\left\{9\left(\frac{\|\boldsymbol{\epsilon}\|_1}{n}\right)^2, 164\kappa^2(g,3)\|\beta_*\|_2^2\right\} \tag{33}$$

Combining (31) and (33), we have that

$$\frac{1}{2n}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2^2 \le 3\delta^2\sum_{l\in J}\frac{1}{\omega_l^2}\max\left\{\frac{9}{\kappa^2(g,3)}\left(\frac{\|\boldsymbol{\epsilon}\|_1}{n}\right)^2, 164\|\beta_*\|_2^2\right\}.$$

□

F Proof of Corollary 3.5

It follows from Lemma 3.1 in [17] that

$$\frac{2}{n}\|(\boldsymbol{X}^\top\boldsymbol{\epsilon})^l\|_2 \le \frac{2\sigma}{\sqrt{n}}\sqrt{\mathrm{tr}(\Psi_l)+2|||\Psi_l|||(4\log L+\sqrt{2p_l\log L})}$$

holds with a probability greater than $1-2/L$, $\mathrm{tr}(\Psi_l)$ denotes the trace of Ψ_l, and $|||\Psi_l|||$ denotes the maximum eigenvalue of Ψ_l. Since $\Psi_l = I_{p_l\times p_l}$, we have that $|||\Psi_l||| = 1$ and $\mathrm{tr}(\Psi_l) = p_l$. Consequently, we have that

$$\frac{2}{n}\|(\boldsymbol{X}^\top\epsilon)^l\|_2 \le \frac{2\sigma}{\sqrt{n}}\sqrt{p_l+2(4\log L+\sqrt{2p_l\log L})} \le \frac{2\sigma}{\sqrt{n}}\sqrt{3p_l+9\log L},$$

holds with a probability greater than $1-2/L$.

Also, it follows from the proof of Corollary 2.4 that

$$\sigma\left(\sqrt{\frac{2}{\pi}}-\frac{1}{10}\right) \le \frac{1}{n}\|\boldsymbol{\epsilon}\|_1$$

holds with a probability greater $1-2\exp(-C_1 n)$. Suppose we let

$$\frac{\delta}{\omega_l} = \frac{2}{\sqrt{\frac{2}{\pi}}-\frac{1}{10}}\sqrt{\frac{3p_l+9\log L}{n}},$$

we have that

$$\frac{2\|(\boldsymbol{X}^\top\boldsymbol{\epsilon})^l\|_2}{\|\boldsymbol{\epsilon}\|_1} \le \frac{\delta}{\omega_l}.$$

holds with a probability greater than $1-2\exp(-C_1 n)-2/L$.

It follows from Theorem 3.4 that

$$\begin{aligned}\frac{1}{2n}\|\boldsymbol{X}(\widetilde{\beta}-\beta_*)\|_2^2 &\le 48\delta^2\sum_{l\in J}\frac{1}{\omega_l^2}\max\left\{\frac{9}{\kappa^2(g,3)}\left(\frac{\|\boldsymbol{\epsilon}\|_1}{n}\right)^2, 164\|\beta_*\|_2^2\right\}\\ &\le 48\sum_{l\in J}\frac{3p_l+9\log L}{n}\max\left\{\frac{9}{\kappa^2(s,3)}\left(\frac{\|\boldsymbol{\epsilon}\|_1}{n}\right)^2, 164\|\beta_*\|_2^2\right\}\\ &= 432\frac{|G_J|+g\log L}{n}\max\left\{\frac{9}{\kappa^2(s,3)}\left(\frac{\|\boldsymbol{\epsilon}\|_1}{n}\right)^2, 164\|\beta_*\|_2^2\right\}.\end{aligned}$$

G Proof of Corollary 3.7

Corollary 3.7 is straightforward due to Corollary 3.5 and the upper bound arguments in the proof in Corollary 2.6.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claim is that the convergence rate of the prediction error of adversarial training estimator under ℓ_∞-perturbation can achieve the minimax rate up to a logarithmic factor in the high-dimensional linear regression on the class of sparse parameters. We also discuss the group adversarial training. These contributions, including important assumptions, are clearly stated in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have mentioned the limitations and future work in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results in this paper are written mathematically rigorous with full set of assumptions, and all associated proofs can be seen in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have mainly made theoretical contributions. However, some numerical experiments are also provided. Full experimental settings have been stated in the paper. Our results can be reproduced from the submitted code in the Supplementary Materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our numerical experiments are based on synthetic data. The code has been submitted in the Supplementary Materials.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All settings are specified in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper has reported the error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our numerical experiments do not require any workers, and our numerical experiments compare prediction errors achieved by different approaches instead of computation speed or storage.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification:[NA]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.