Model Repair: Robust Recovery of Over-Parameterized Statistical Models

Chao Gao
Department of Statistics
University of Chicago

John Lafferty
Department of Statistics and Data Science
Yale University

May 16, 2020

1. Introduction

In this paper we introduce a new type of robust estimation problem—how to recover a statistical model that has been corrupted after estimation. Traditional robust estimation assumes that the data are corrupted, and studies methods of estimation that are immune to these corruptions or outliers in the data. In contrast, we explore the setting where the data are "clean" but a statistical model is corrupted after it has been estimated using the data. We study methods for recovering the model that do not require re-estimation from scratch, using only the design and not the original response values.

The problem of model repair is motivated from several different perspectives. First, it can be formulated as a well-defined statistical problem that is closely related to, but different from, traditional robust estimation, and that deserves study in its own right. From a more practical perspective, modern machine learning practice is increasingly working with very large statistical models. For example, artificial neural networks having several million parameters are now routinely estimated. It is anticipated that neural networks having trillions of parameters will be built in the coming years, and that large models will be increasingly embedded in systems, where they may be subject to errors and corruption of the parameter values. In this setting, the maintenance of models in a fault tolerant manner becomes a concern. A different perspective takes inspiration from plasticity in brain function, with the human brain in particular having a remarkable ability to repair itself after trauma. The framework for model repair that we introduce in this paper can be viewed as a simple but mathematically rigorous formulation of this ability in neural networks.

At a high level, our findings reveal that two important ingredients are necessary for model repair. First, the statistical model must be over-parameterized, meaning that there should be many more parameters than observations. While over-parameterization leads to issues of identifiability from traditional perspectives, here it is seen as a necessary property of the model. Second, the estimator must incorporate redundancy in some form; for instance, sparse estimators of over-parameterized models will not in general be repairable. Notably, we show that estimators based on gradient descent and stochastic gradient descent are well suited to model repair.

At its core, our formulation and analysis of model repair rests upon representing an estimator in terms of the row space of functions of the data design matrix. This leads to a view of model repair as a form of robust estimation. The recovery algorithms that we propose are based on solving a linear program that is equivalent to median regression. Our key technical lemma, which may be of independent interest, gives sharp bounds on the probability that this linear program successfully recovers the model, which in turn determines the level of over-parameterization that is required. An interesting facet of this formulation is that the response vector is not required by the repair process. Because the model is over-parameterized, the estimator effectively encodes the response. This phenomenon can be viewed from the perspective of communication theory, where the corruption process is seen as a noisy channel, and the design matrix is seen as a linear error-correcting code for communication over this channel.

After formulating the problem and establishing the key technical lemma, we present a series of results for repair of over-parameterized linear models, random feature models, and artificial neural networks. These form the main technical contributions of this paper. A series of simulation experiments are presented that corroborate and illustrate our theoretical results. In the following section we give a more detailed overview of our results, including the precise formulation of the model repair problem, its connection to robust estimation and error correcting codes, and an example of the repair algorithm in simulation. We then present the key lemma, followed by detailed analysis of model repair for specific model classes. We present the proof of the key lemma in Section 3, and the proofs of the neural network repair results with hyperbolic tangent activation are given in Section 6. Proofs of technical lemmas and the results for neural networks with ReLU activation are presented in the appendix, to make the presentation more readable. Section 8 gives a discussion of directions for further research and potential implications of our findings for applications.

2. Problem formulation and overview of results

In this section we formulate the problem of model repair, and give an overview of our results. Suppose that $\widehat{\theta} \in \mathbb{R}^p$ is a model with p parameters estimated on n data points $\{(x_i, y_i)\}_{i=1}^n$ as a classification or regression model. The model $\widehat{\theta}$ is then corrupted by noise. The primary noise model we study in this paper is

$$\eta = \widehat{\theta} + z \tag{2.1}$$

where $z_j \sim (1-\varepsilon)\delta_0 + \varepsilon Q_j$ and Q_j is an arbitrary distribution. In other words, each component $\widehat{\theta}_j$ of $\widehat{\theta}$ is corrupted by additive noise from an arbitrary distribution Q_j with probability ε, where $0 \leq \varepsilon \leq 1$, and is uncorrupted with probability $1-\varepsilon$. The noise vector z is assumed to be independent of the design $\{x_i\}_{i=1}^n$. We discuss alternative error models later in the paper. The goal is to recover $\widehat{\theta}$ from η, without reestimating the model from scratch; in particular, without using the response values $\{y_i\}$ when the model is estimated in a supervised learning setting.

Overparameterized linear models. To explain the main ideas, let us first consider the setting of under-determined linear regression. Let $X \in \mathbb{R}^{n\times p}$ be the design matrix and $y \in \mathbb{R}^n$ a vector of response values, and suppose that we wish to minimize the squared error $\|y - X\theta\|_2^2$. If $n > p$

then this is an under-determined optimization problem. Among all solutions to the linear system $y = X\theta$, the solution of minimal norm $\|\theta\|_2$ is given by

$$\widehat{\theta} = X^T(XX^T)^{-1}y \tag{2.2}$$

assuming that X has full rank n (Boyd and Vandenberghe, 2004). Thus, $\widehat{\theta}$ lies in the row space of the $n \times p$ design matrix X. The risk behavior of (2.2) has been well studied in the recent literature (Bartlett et al., 2020; Belkin et al., 2019; Hastie et al., 2019).

Now suppose that $\eta = \widehat{\theta} + z$ where $z_j \sim (1-\varepsilon)\delta_0 + \varepsilon Q_j$. The method we propose to recover $\widehat{\theta}$ from η is to let $\widetilde{u} \in \mathbb{R}^n$ be the solution to the optimization

$$\widetilde{u} = \underset{u}{\operatorname{argmin}} \|\eta - X^T u\|_1 \tag{2.3}$$

and define the repaired model as $\widetilde{\theta} = X^T\widetilde{u}$. The linear program defined in (2.3) can be thought of as performing median regression of η onto the rows of X. Our analysis shows that, under appropriate assumptions, the estimated model is exactly recovered with high probability, so that $\widetilde{\theta} = \widehat{\theta}$, as long as $n/p \leq c(1-\varepsilon)^2$ for some sufficiently small constant c.

Figure 1 shows the performance of the repair algorithm in simulation. The design is sampled as $x_{ij} \sim N(0,1)$ and the corruption distribution is $Q_j = N(1,1)$ for each j. With the sample size fixed at $n = 50$, the dimension p is varied according to $p_k/n = 200/k^2$ with k ranging from 1 to 6. The plots show the empirical probability of exact repair $\widetilde{\theta} = \widehat{\theta}$ as a function of ε. The roughly equal spacing of the curves agrees with our theory, which indicates that $\sqrt{n/p}/(1-\varepsilon)$ should be sufficiently small for successful repair. The theory indicates that the repair probability for dimension p_k as a function of the adjusted value $\varepsilon_k = \varepsilon + c' \cdot k - \frac{1}{2}$ should exhibit a threshold at $\varepsilon_k = 1/2$ for the constant $c' = \frac{\sqrt{2}}{20c}$; this is seen in the right plot of Figure 1.

Robust regression. This procedure can be viewed in terms of robust regression. Specifically, η can be viewed as a corrupted response vector, and $A = X^T \in \mathbb{R}^{p\times n}$ can be viewed as a design matrix that is *not corrupted*. Our result makes precise conditions under which this robust regression problem can be successfully carried out. In particular, we show that model repair is possible even if $\varepsilon \to 1$, so that the proportion of corrupted model components approaches one. This is in contrast to the traditional Huber model where the design is also corrupted (Huber, 1964), under which consistent estimation is only possible if ε is below some small constant (Chen et al., 2016; Gao, 2020). The problem of robust regression with uncorrupted design has a rich literature; we review some of the relevant work in Section 3.

Error-correcting codes. Model repair can also be viewed in terms of error-correcting codes. Specifically, viewing the response vector $y \in \mathbb{R}^n$ as a message to be communicated over a noisy channel, the minimum norm model $\widehat{\theta} = X^T u = X^T(XX^T)^{-1}y$ redundantly encodes y since $p > n$ (see Figure 2). The decoding algorithm $\widetilde{u} = \operatorname{argmin}_u \|\eta - X^T u\|$ then recovers the data y according to $y = (XX^T)\widetilde{u}$. The inequality $n/p < c(1-\varepsilon)^2$ gives a condition on the rate of the code, that is, the level of redundancy that is sufficient for this decoding procedure to recover the message with high probability.

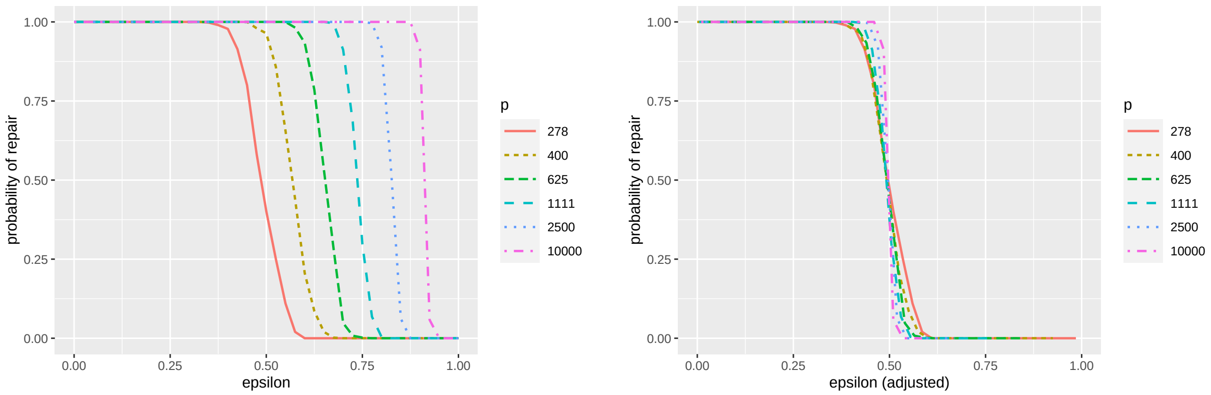

FIG 1. *Left: Empirical probability of exact repair as a function of* ε. *The sample size is* $n = 50$ *and the model dimension* p *varies as* $p_k/n = 200/k^2$, *for* $k = 1, \ldots, 6$; *each point is an average over 500 trials. The plot on the right shows the repair probability as a function of the adjusted value* $\varepsilon_k = \varepsilon + c' \cdot k - \frac{1}{2}$ *for dimension* p_k, *where the constant is* $c' = \frac{\sqrt{2}}{20c} = 0.085$.

When X is a random Gaussian matrix, the mapping $u \rightarrow X^T u = \sum_{i=1}^n u_i X_i^T$ can be viewed as a superposition of random codewords in $\mathbb{R}^p$ (Joseph and Barron, 2012; Rush et al., 2017). The fundamental difference with channel coding is that in our regression setting the design matrix X is fixed, and is not chosen for optimal channel coding. Indeed, the noise model $w \rightarrow w + z$ that we consider, with $z_j \sim (1 - \varepsilon)\delta_0 + \varepsilon Q_j$ corresponds to a channel having infinite capacity, and a simple repetition code would suffice for identifying components that are uncorrupted (Cover and Thomas, 2006).

Estimators based on gradient descent. The comments made above carry over to estimators of linear models based on gradient descent and stochastic gradient descent for arbitrary loss functions. Consider objective functions of the form

$$\ell(\theta) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(y_i, x_i^T \theta) \tag{2.4}$$

where $\mathcal{L}(y, f)$ is a general loss function; this includes a broad range of estimators for problems such as linear least squares and logistic regression, robust regression, support vector machines, and others. The gradient descent update rule is

$$\begin{aligned} \theta^{(t+1)} &= \theta^{(t)} - \gamma_t \frac{1}{n} \sum_{i=1}^n \nabla_\theta \mathcal{L}(y_i, x_i^T \theta^{(t-1)}) & (2.5) \\ &= \theta^{(t)} - \gamma_t \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial f} \mathcal{L}(y_i, x_i^T \theta^{(t-1)}) x_i & (2.6) \\ &= \theta^{(t)} - \sum_{i=1}^n w_i^{(t)} x_i, & (2.7) \end{aligned}$$

where γ_t is a step size parameter. If the model is initialized at $\theta^{(0)} = 0 \in \mathbb{R}^p$ then the estimate at

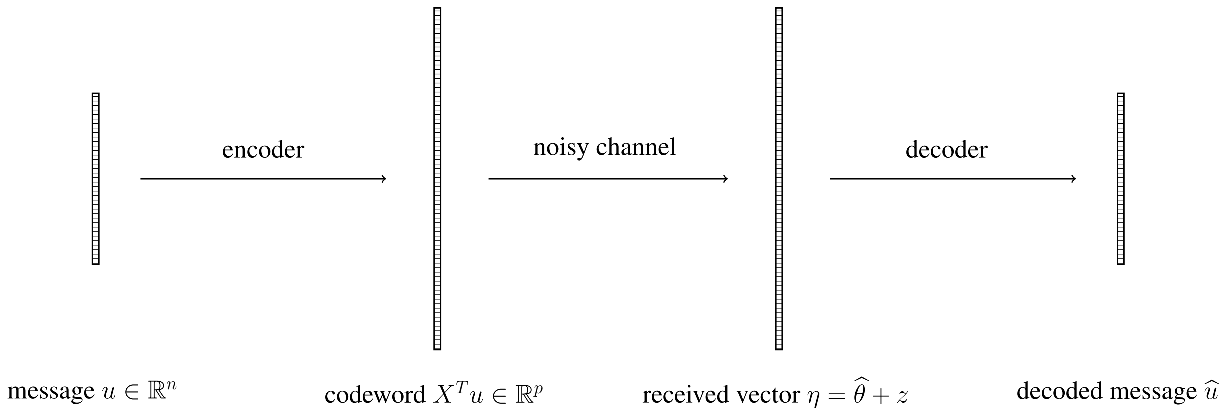
encoder
noisy channel
decoder
message $u \in \mathbb{R}^n$
codeword $X^T u \in \mathbb{R}^p$
received vector $\eta = \widehat{\theta} + z$
decoded message $\widehat{u}$

FIG 2. *Model repair viewed in terms of error-correcting codes. The model $\widehat{\theta} = X^T u \in \mathbb{R}^p$ is in the row-space of the design matrix, which gives a redundant representation of the "message" $u \in \mathbb{R}^n$ for $n < p$. The model is received as a noisy version η with each entry corrupted with probability ε. The received vector is decoded by solving a linear program.*

time t can thus be written as

$$\theta^{(t)} = X^T u^{(t)} \tag{2.8}$$

for some $u^{(t)} \in \mathbb{R}^n$. After contamination, we have $\eta = X^T u^{(t)} + z$. Therefore, we can recover the model by computing $\widetilde{\theta} = X^T \widetilde{u}$ where $\widetilde{u}$ is the solution to (2.3).

The same conclusion holds for estimators based on stochastic gradient descent. With B_t denoting the set of samples used in the mini-batch of the tth iteration, we can write

$$\theta^{(t)} = \sum_{i \in B_1 \cup \cdots \cup B_t} w_i^{(t)} x_i. \tag{2.9}$$

We repair the corrupted model η by computing $\widetilde{\theta} = X^T_{B_1 \cup \cdots B_t} \widetilde{u}$ with

$$\widetilde{u} = \operatorname*{argmin}_u \|\eta - X^T_{B_1 \cup \cdots B_t} u\|_1 \tag{2.10}$$

where the submatrix $X_{B_1 \cup \cdots B_t}$ only includes rows of X for indices that were visited during some stochastic gradient descent step. Our theory then establishes that the model is recovered with high probability in case

$$\frac{\sqrt{|B_1 \cup \cdots B_t|/p}}{1 - \varepsilon} < c. \tag{2.11}$$

Typically the training takes place in "epochs" where all n data points are visited in each epoch.

Random features and neural networks. Our theory extends to random features models (Rahimi and Recht, 2008), where the covariates are $\widetilde{X} = \psi(XW) \in \mathbb{R}^{n \times p}$ where $X \in \mathbb{R}^{n \times d}$, the matrix $W \in \mathbb{R}^{d \times p}$ is a random Gaussian matrix that is not trained, and ψ is a threshold function such as the hyperbolic tangent function or rectified linear unit. In particular, when the model is trained using gradient descent, the parameters $\widehat{\theta}$ lie in the row space of the matrix $\widetilde{X}$. We also show how the ideas can be extended to neural networks, where the weights W are trained. This requires modifications to the training and recovery algorithms that we detail below.

In the following section we present the key technical lemma that explains how these results are possible; this is a result in robust regression that may be of independent interest. In Sections 4 and 5 we state the theoretical results for over-complete linear models and neural networks trained with gradient descent. The proofs of the neural network results with hyperbolic tangent activation are given in Section 6.

3. Key lemma: Robust regression with uncorrupted design

Consider a regression model $\eta = Au^* + z \in \mathbb{R}^m$, where $A^T = (a_1, a_2, ..., a_m)^T \in \mathbb{R}^{m\times k}$ is a design matrix and $u^* \in \mathbb{R}^k$ is a vector of regression coefficients to be recovered. We consider a random design setting, and the distribution of A will be specified later. For the noise vector $z \in \mathbb{R}^m$, we assume it is independent of the design matrix A, and

$$z_i \sim (1-\varepsilon)\delta_0 + \varepsilon Q_i, \tag{3.1}$$

independently for all $i \in [m]$. In other words, a fraction ε of the components η_i are contaminated by additive noise z_i that is drawn from an arbitrary and unknown distribution. To robustly recover u^*, we propose the estimator

$$\widehat{u} = \underset{u\in\mathbb{R}^k}{\operatorname{argmin}} \|\eta - Au\|_1.$$

The estimator $\widehat{u}$ can be computed using a standard linear programming. In order that $\widehat{u}$ successfully recovers the true regression coefficients u^*, we need to impose the following conditions on the design matrix A.

Condition A. There exists some σ^2, such that for any fixed (not random) $c_1, ..., c_m$ satisfying $\max_i |c_i| \leq 1$,

$$\left\|\frac{1}{m}\sum_{i=1}^m c_i a_i\right\|^2 \leq \frac{\sigma^2 k}{m}, \tag{3.2}$$

with high probability.

Condition B. There exist $\underline{\lambda}$ and $\overline{\lambda}$, such that

$$\inf_{\|\Delta\|=1} \frac{1}{m}\sum_{i=1}^m |a_i^T\Delta| \quad \geq \quad \underline{\lambda}, \tag{3.3}$$

$$\sup_{\|\Delta\|=1} \frac{1}{m}\sum_{i=1}^m |a_i^T\Delta|^2 \quad \leq \quad \overline{\lambda}^2, \tag{3.4}$$

with high probability.

Theorem 3.1. *Assume the design matrix A satisfies Condition A and Condition B. Then if*

$$\frac{\overline{\lambda}\sqrt{\frac{k}{m}\log\left(\frac{em}{k}\right)} + \varepsilon\sigma\sqrt{\frac{k}{m}}}{\underline{\lambda}(1-\varepsilon)} \tag{3.5}$$

is sufficiently small, we have $\widehat{u} = u^*$ with high probability.

Proof. Define $L_m(u) = \frac{1}{m}\sum_{i=1}^m (|a_i^T(u^* - u) + z_i| - |z_i|)$, and $L(u) = \mathbb{E}(L_m(u) \,|\, A)$. Letting $t \geq 0$ be arbitrary, we must have $\inf_{\|u-u^*\|\geq t} L_m(u) \leq L_m(u^*) = 0$. By the convexity of $L_m(u)$, this leads to $\inf_{\|u-u^*\|=t} L_m(u) \leq 0$, and thus

$$\begin{aligned}\inf_{\|u-u^*\|=t} L(u) &\leq \inf_{\|u-u^*\|=t} L_m(u) + \sup_{\|u-u^*\|=t} (L(u) - L_m(u)) \\ &\leq \sup_{\|u-u^*\|=t} |L_m(u) - L(u)|.\end{aligned}$$

Introducing the notation $f_i(x) = \mathbb{E}_{z_i \sim Q_i}(|x + z_i| - |z_i|)$ and $Q_i(x) = Q_i(z_i \leq x)$, it is easy to see that $f_i(0) = 0$ and $f_i'(x) = 1 - 2Q_i(-x)$. Observe that we can write

$$L(u) = (1-\varepsilon)\frac{1}{m}\sum_{i=1}^m |a_i^T(u - u^*)| + \varepsilon\frac{1}{m}\sum_{i=1}^m f_i(a_i^T(u^* - u)). \tag{3.6}$$

For any u such that $\|u - u^*\| = t$, the first term of (3.6) can be lower bounded by

$$(1-\varepsilon)\frac{1}{m}\sum_{i=1}^m |a_i^T(u - u^*)| \geq \underline{\lambda}(1-\varepsilon)t,$$

by Condition B. To analyze the second term of (3.6), we note that f_i is a convex function, and therefore for any u such that $\|u - u^*\| = t$,

$$\begin{aligned}\varepsilon\frac{1}{m}\sum_{i=1}^m f_i(a_i^T(u^* - u)) &\geq \varepsilon\frac{1}{m}\sum_{i=1}^m f_i(0) + \varepsilon\frac{1}{m}\sum_{i=1}^m f_i'(0)a_i^T(u^* - u) \\ &= \varepsilon\frac{1}{m}\sum_{i=1}^m (1 - 2Q_i(0))\, a_i^T(u^* - u) \\ &\geq -\varepsilon t\left\|\frac{1}{m}\sum_{i=1}^m (1 - 2Q_i(0))\, a_i\right\| \\ &\geq -\varepsilon t\sigma\sqrt{\frac{k}{m}},\end{aligned}$$

where the first inequality uses Cauchy-Schwarz, and the second inequality uses Condition A. By Condition B and an empirical process result proved as Lemma A.6 in Appendix A.2, we have

$$\sup_{\|u-u^*\|=t} |L_m(u) - L(u)| \lesssim t\overline{\lambda}\sqrt{\frac{k}{m}\log\left(\frac{em}{k}\right)}, \tag{3.7}$$

with high probability. Therefore, we have shown that $\|\widehat{u} - u^*\| \geq t$ implies

$$\underline{\lambda}(1-\varepsilon)t - \varepsilon t\sigma\sqrt{\frac{k}{m}} \lesssim t\overline{\lambda}\sqrt{\frac{k}{m}\log\left(\frac{em}{k}\right)},$$

which is impossible when $\frac{\overline{\lambda}\sqrt{\frac{k}{m}\log\left(\frac{em}{k}\right)}+\varepsilon\sigma\sqrt{\frac{k}{m}}}{\underline{\lambda}(1-\varepsilon)}$ is sufficiently small, and thus $\|\widehat{u} - u^*\| < t$ with high probability. Since t is arbitrary, we must have $\widehat{u} = u^*$. □

The theorem gives a sufficient condition for the exact recovery of the regression coefficients. When both $(\sigma + \overline{\lambda})/\underline{\lambda}$ and $1-\varepsilon$ are constants, the condition becomes that k/m is sufficiently small.

A notable feature of this theorem is that it allows for $\varepsilon \to 1$; that is, an arbitrarily large fraction of the components of the response Au^* can be corrupted. This is in contrast to robust regression where both the response and design are contaminated. To be specific, consider independent observations $(a_i, \eta_i) \sim (1-\varepsilon)P_{u^*} + \varepsilon Q_i$, where the probability distribution P_{u^*} encodes the linear model $\eta_i = a_i^T u_i$, and for each $i \in [m]$, there is probability ε that the pair (a_i, η_i) is drawn from some arbitrary distribution Q_i. In this setting, consistent or exact recovery of the regression coefficient is only possible when $\varepsilon < c$ for some small constant $c > 0$ (Gao, 2020). The reason Theorem 3.1 allows $\varepsilon \to 1$ is that there is no contamination for the design matrix A.

Another distinguishing feature of Theorem 3.1 is that there is no assumption imposed on the contamination distribution Q_i, even though the median regression procedure apparently requires the noise to be symmetric around zero. To understand this phenomenon, consider a population objective function

$$\ell(u) = \mathbb{E}|\eta - a^T u|,$$

where $\eta = a^T u^* + z$, and the expectation is over both a and z. In order for the minimizer of $\ell(u)$ to recover u^* in the population, a criterion usually called Fisher consistency, it is required that $\nabla\ell(u^*) = 0$. Under the assumption that a and z are independent, this gives

$$\nabla\ell(u^*) = \mathbb{E}[\text{sign}(z)a] = \mathbb{E}\,\text{sign}(z)\mathbb{E}a = 0. \tag{3.8}$$

This means we should be able to achieve consistency without any assumption on the noise variable z as long as we assume $\mathbb{E}a_i = 0$.

But Condition A can be viewed as a general assumption that covers $\mathbb{E}a_i = 0$ as a special case. As a concrete example, let us suppose the design matrix A has m uncorrelated rows and its entries all have mean zero and variance at most one. Then,

$$\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^m c_i a_i\right\|^2 = \sum_{j=1}^k \mathbb{E}\left(\frac{1}{m}\sum_{i=1}^m c_i a_{ij}\right)^2 = \sum_{j=1}^k \frac{1}{m^2}\sum_{i=1}^m c_i^2 \mathbb{E}a_{ij}^2 \leq \frac{k}{m},$$

and thus Condition A holds with some constant σ^2, by an additional argument using Markov's inequality.

More generally, Condition A also allows a design matrix with entries whose means are not necessarily zero. This will in general lead to a term σ^2 that may not be of constant order. However, since the condition of Theorem 3.1 involves an additional ε factor in front of σ, the robust estimator $\widehat{u}$ can still recover u^* as long as the contamination proportion is vanishing at an appropriate rate. As an important application, the result of Theorem 3.1 also applies to design matrices with an intercept.

We also introduce an alternative of Condition A. By (3.8), we observe that Fisher consistency also follows if $\mathbb{E}\,\text{sign}(z_i) = 0$. However, this does not mean that we have to assume the distribution of z_i is symmetric. It turns out we only need the distribution of a_i to be symmetric by applying a symmetrization argument. Note that with the help of independent Rademacher random variables $\delta_i \sim \text{Uniform}\{\pm 1\}$, we can write the data generating process as $\delta_i \eta_i = \delta_i a_i^T u^* + \delta_i z_i$. With this new representation, we can also view $\delta_i \eta_i$, $\delta_i a_i$ and $\delta_i z_i$ as the response, covariate, and noise. Now the noise $\delta_i z_i$ is symmetric around zero, and it can be shown that $\delta_i a_i$ and $\delta_i z_i$ are still independent as long as the distribution of a_i is symmetric. Since for any $u \in \mathbb{R}^k$,

$$\sum_{i=1}^{m} |\delta_i \eta_i - \delta_i a_i^T u| = \sum_{i=1}^{m} |\eta_i - a_i^T u|,$$

we obtain an equivalent median regression after symmetrization. This alternative condition is stated as follows.

Condition $\widetilde{A}$. Given i.i.d. Rademacher random variables $\delta_1, ..., \delta_m$, the distribution of

$$\widetilde{A}^T = (\delta_1 a_1, \delta_2 a_2, ..., \delta_m a_m)^T$$

is identical to that of A^T.

Theorem 3.2. *Assume the design matrix A satisfies Condition $\widetilde{A}$ and Condition B. Then if*

$$\frac{\overline{\lambda}\sqrt{\frac{k}{m}\log\left(\frac{em}{k}\right)}}{\underline{\lambda}(1-\varepsilon)}$$

is sufficiently small, we have $\widehat{u} = u^*$ with high probability.

To close this section, we note that the problem of robust regression with uncorrupted design is also recognized as outlier-robust regression in the literature. This problem has been studied previously by Karmalkar and Price (2018); Nguyen and Tran (2013a,b); Tsakonas et al. (2014); Wright and Ma (2010). In particular, Bhatia et al. (2017) proposed a hard-thresholding algorithm that consistently recovers the regression coefficients as long as ε is below some small constant. The recent work Suggala et al. (2019) has established consistent recovery while allowing $\varepsilon \to 1$. Compared with their algorithm, our method based on ℓ_1 minimization is much simpler. Moreover, we allow $\varepsilon = 1 - \Theta\left(\sqrt{\frac{k}{m}\log\left(\frac{em}{k}\right)}\right)$, compared with the requirement $\varepsilon \leq 1 - \Theta\left(\frac{1}{\log\log m}\right)$ in Suggala et al. (2019).

4. Repair of linear and random feature models

Consider a linear model with $X \in \mathbb{R}^{n\times p}$ being the design matrix and $y \in \mathbb{R}^n$ being a vector of response values. We assume that each entry of the design matrix is i.i.d. $N(0,1)$ and do not impose any assumption on the response y. A machine learning algorithm learns a linear model $X\widehat{\theta}$ with some $\widehat{\theta} \in \mathbb{R}^p$. The vector $\widehat{\theta}$ is either computed via the formula (2.2) or through a gradient-based algorithm with the objective (2.4) initialized from 0. Either case implies $\widehat{\theta}$ belongs to the row space of X. Suppose we observe a contaminated version of $\widehat{\theta}$ through $\eta = \widehat{\theta} + z$, where z is independent of $\widehat{\theta}$ and $z_j \sim (1-\varepsilon)\delta_0 + \varepsilon Q_j$ independently for all $j \in [p]$. We then propose to recover $\widehat{\theta}$ via

$$\widetilde{u} = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \|\eta - X^T u\|_1,$$

and define the repaired model as $\widetilde{\theta} = X^T \widetilde{u}$. This turns out to be the same robust regression problem studied in Section 3, and thus we only need to check the design matrix $A = X^T$ satisfies Condition A and Condition B.

Lemma 4.1. *Assume n/p is sufficiently small. Then, Condition A and Condition B hold for $A = X^T$, $m = p$ and $k = n$ with some constants σ^2, $\underline{\lambda}$ and $\overline{\lambda}$.*

Combine Lemma 4.1 and Theorem 3.1, and we obtain the following guarantee for model repair.

Corollary 4.1. *Assume $\frac{\sqrt{\frac{n}{p}}\log\left(\frac{ep}{n}\right)}{1-\varepsilon}$ is sufficiently small. We then have $\widetilde{\theta} = \widehat{\theta}$ with high probability.*

We note that compared with the robust regression setting, the roles of the sample size and dimension are switched in model repair. Corollary 4.1 requires that the linear model to be overparametrized in the sense of $p \gg n(1-\varepsilon)^2$ (with logarithmic factors ignored) in order that repair is successful.

Besides an overparametrized model, we also require that the estimator $\widehat{\theta}$ lies in the row space of the design matrix X, so that the redundancy of a overparametrized model is preserved in the estimator.

Remark 4.1. To understand the requirement on the estimator $\widehat{\theta}$, let us consider a simple toy example. We assume that X has p identical columns, which is clearly an overparametrized model. Consider two estimators:

$$\begin{aligned} \widehat{\theta}_{\mathsf{min-norm}} &\in \operatorname{argmin}\left\{\|\theta\| : y = X\theta\right\}, \\ \widehat{\theta}_{\mathsf{sparse}} &\in \operatorname{argmin}\left\{\|\theta\|_0 : y = X\theta\right\}. \end{aligned}$$

It is clear that $\widehat{\theta}_{\mathsf{min-norm}}$ has identical entries and $\widehat{\theta}_{\mathsf{sparse}}$ has one nonzero entry. Since the contamination will change an ε-proportion of the entries, $\widehat{\theta}_{\mathsf{sparse}}$ cannot be repaired if its only nonzero entry is changed. On the other hand, $\widehat{\theta}_{\mathsf{min-norm}}$ is resilient to the contamination, and its redundant structure leads to consistent model repair. It is known that gradient based algorithms lead to implicit ℓ_2 norm regularizations (Neyshabur et al., 2014), which then explains the result of Corollary 4.1.

We also study a random feature model with design $\{\psi(W_j^T x_i)\}_{i\in[n],j\in[p]}$, where $x_i \sim N(0, I_d)$ and $W_j \sim N(0, d^{-1}I_d)$ independently for all $i \in [n]$ and $j \in [p]$. We choose the nonlinear activation function to be $\psi(t) = \tanh(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}}$, the hyperbolic tangent unit. The design matrix can thus be written as $\widetilde{X} = \psi(XW) \in \mathbb{R}^{n\times p}$ with $X \in \mathbb{R}^{n\times d}$ and $W \in \mathbb{R}^{d\times p}$. This is an important model, and its asymptotic risk behavior under overparametrization has recently been studied by Mei and Montanari (2019). We show that the design matrix $\widetilde{X}^T = \psi(W^T X^T)$ satisfies Condition A and Condition B so that model repair is possible.

Lemma 4.2. *Assume n/p^2 and n/d are sufficiently small. Then, Condition A and Condition B hold for $A = \widetilde{X}^T$, $m = p$ and $k = n$ with some constants σ^2, $\underline{\lambda}$ and $\overline{\lambda}$.*

Now consider a model $\widehat{\theta}$ that lies in the row space of $\widetilde{X}$. We observe a contaminated version $\eta = \widehat{\theta} + z$. We can then compute the procedure $\widetilde{u} = \operatorname{argmin}_{u \in \mathbb{R}^n} \|\eta - \widetilde{X}^T u\|_1$ and use $\widetilde{\theta} = \widetilde{X}^T \widetilde{u}$ for model repair.

Corollary 4.2. *Assume* $\frac{\sqrt{\frac{n}{p}}\log\left(\frac{ep}{n}\right)}{1-\varepsilon}$, n/p^2 *and* n/d *are sufficiently small. We then have* $\widetilde{\theta} = \widehat{\theta}$ *with high probability.*

The results in this section are stated for the hyperbolic tangent nonlinear activation. They can be extended to other activation functions. In practice, the most popular choice is the rectified linear unit (ReLU) $\psi(t) = \max(0, t)$. The results for ReLU will be given in the appendix.

5. Repair of neural networks

In this section we show how to use robust regression to repair neural networks. We consider a neural network with one hidden layer,

$$f(x) = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \beta_j \psi(W_j^T x),$$

where ψ is either the rectified linear unit (ReLU) function $\psi(t) = \max(t, 0)$, or the hyperbolic tangent $\psi(t) = \tanh(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}}$. The factor $p^{-1/2}$ in the definition above is convenient for our theoretical analysis. In this section we present the analysis for the hyperbolic tangent activation function, with the ReLU deferred to the appendix.

With the squared error loss function

$$\mathcal{L}(\beta, W) = \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \beta_j \psi(W_j^T x_i) \right)^2,$$

we consider training the model using a standard gradient descent algorithm (Algorithm 1).

Algorithm 1: Gradient descent for neural nets

```
Input: Data $(y, X)$ and the number of iterations $t_{\max}$.
Initialization: $W_j(0) \sim N(0, d^{-1} I_d)$ and $\beta_j(0) \sim N(0, 1)$ independently for all $j \in [p]$.
Iterate: For $t$ in $1 : t_{\max}$, compute
    $\beta_j(t) = \beta_j(t-1) - \gamma \frac{\partial \mathcal{L}(\beta, W)}{\partial \beta_j}\Big|_{(\beta, W) = (\beta(t-1), W(t-1))} \quad j \in [p],$
    $W_j(t) = W_j(t-1) - \frac{\gamma}{d} \frac{\partial \mathcal{L}(\beta, W)}{\partial W_j}\Big|_{(\beta, W) = (\beta(t), W(t-1))} \quad j \in [p].$
Output: Trained parameters $\beta(t_{\max})$ and $W(t_{\max})$.
```

Based on this standard gradient descent algorithm, we consider two estimators of the parameters $(\widehat{\beta}, \widehat{W})$. The first is simply to set $\widehat{\beta} = \beta(t_{\max})$ and $\widehat{W} = W(t_{\max})$; this is the usual estimator.

In the second approach, one fixes $\widehat{W} = W(t_{\max})$ and then retrains β using gradient descent for the objective $\|y - \psi(X\widehat{W})\beta\|^2$ after initializing at zero. In this way, $\widehat{\beta}$ is an approximation to the minimal ℓ_2 norm solution of $\|y - \psi(X\widehat{W})\beta\|^2$. This estimator can be viewed as a linear model that uses features extracted from the data by the neural network.

Now consider the contaminated model $\eta = \widehat{\beta} + z$ and $\Theta_j = \widehat{W}_j + Z_j$, where each entry of z and Z_j is zero with probability $1 - \varepsilon$ and follows an arbitrary distribution with the complementary probability ε. We analyze the following repair algorithm.

Algorithm 2: Model repair for neural networks

```
Input: Contaminated model $(\eta, \Theta)$, design matrix $X$, and initializations $\beta(0)$, $W(0)$.
Repair of the hidden layer: For $j \in [p]$, compute
    $$\widetilde{v}_j = \operatorname*{argmin}_v \|\Theta_j - W_j(0) - X^T v_j\|_1,$$
    and set $\widetilde{W}_j = W_j(0) + X^T \widetilde{v}_j$.
Repair of the output layer: Compute
    $$\widetilde{u} = \operatorname*{argmin}_u \|\eta - \beta(0) - \psi(\widetilde{W}^T X^T)u\|_1,$$
    and set $\widetilde{\beta} = \beta(0) + \psi(\widetilde{W}^T X^T)\widetilde{u}$.
Output: The repaired parameters $\widetilde{\beta}$ and $\widetilde{W}$.
```

Remark 5.1. Algorithm 2 adopts a layerwise repair strategy. This algorithm extends naturally to multilayer networks, repairing the parameters in stages with a forward pass through the layers. We leave the multilayer extension of our analysis to future work.

Remark 5.2. It is important to note that the repair of neural networks not only requires X, but it also requires the initializations $\beta(0)$ and $W(0)$. From a practical perspective, this can be easily achieved by setting a seed using a pseudorandom number generator to initialize the parameters, and making the seed available to the repair algorithm. We also note that when $\widehat{\beta}$ is trained after fixing $\widehat{W}$, one can replace $\beta(0)$ by 0 in Algorithm 2.

Since the gradient $\frac{\partial \mathcal{L}(\beta, W)}{\partial W_j}$ lies in the row space of X, the vector $\widehat{W}_j - W_j(0)$ also lies in the row space of X. Thus, the theoretical guarantee of the hidden layer repair directly follows Corollary 4.1. The repair of the output layer is more complicated, because the gradient $\frac{\partial \mathcal{L}(\beta, W)}{\partial \beta_j}|_{W = W(t-1)}$ lies in the row space of $\psi(XW(t-1))$, which changes over time. Thus, we cannot directly apply the result of Corollary 4.2 for the random feature model. However, when the neural network is overparametrized, it can be shown that the gradient descent algorithm (Algorithm 1) leads to $W(t)$

that is close to the initialization $W(0)$ for all $t \geq 0$. We establish this result in the following theorem by assuming that x_i is i.i.d. $N(0, I_d)$ and $|y_i| \leq 1$ for all $i \in [n]$. Define $u(t) \in \mathbb{R}^n$ with its ith entry given by $u_i(t) = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \beta_j(t)\psi(W_j(t)^T x_i)$, the function value of x_i at time t.

Theorem 5.1. *Assume $\frac{n}{d}$, $\frac{n^3(\log p)^2}{p}$, and $\gamma\left(1 + \frac{n^4(\log p)^2}{p}\right)$ are all sufficiently small. Then, we have*

$$\|y - u(t)\|^2 \leq \left(1 - \frac{\gamma}{8}\right)^t \|y - u(0)\|^2, \tag{5.1}$$

and

$$\max_{1\leq j\leq p} \|W_j(t) - W_j(0)\| \leq R_1, \tag{5.2}$$

$$\max_{1\leq j\leq p} |\beta_j(t) - \beta_j(0)| \leq R_2, \tag{5.3}$$

for all $t \geq 1$ with high probability, where $R_1 = \frac{100 n \log p}{\sqrt{pd}}$ and $R_2 = 32\sqrt{\frac{n^2 \log p}{p}}$.

Theorem 5.1 assumes that the width of the neural network is large compared with the sample size in the sense that $\frac{p}{(\log p)^4} \gg n^3$. For fixed n, the limit of the neural network as $p \to \infty$ is known as the neural tangent kernel (NTK) regime, and the behavior of gradient descent under this limit has been studied by Jacot et al. (2018). The result of Theorem 5.1 follows the explicit calculation in Du et al. (2018b), and we are able to sharpen some of the asymptotic conditions in Du et al. (2018b).

The theorem has two conclusions. The first conclusion shows the gradient descent algorithm has global convergence in the sense of (5.1) even though the loss $\mathcal{L}(\beta, W)$ is nonconvex. The second conclusion shows that the trajectory of the algorithm $(W(t), \beta(t))$ is bounded within some radius of the initialization. This allows us to characterize the repaired model $\widetilde{\beta}$ for the output layer.

Let us first consider the case $\widehat{\beta} = \beta(t_{\max})$ and $\widehat{W} = W(t_{\max})$. Since the vector $\beta(t) - \beta(t-1)$ lies in the row space of $\psi(XW(t-1))$ for every t, one can show that $\widehat{\beta} - \beta(0)$ approximately lies in the row space of $\psi(XW(0))$ by Theorem 5.1. Therefore, by extending the result of Corollary 4.2 that includes the bias induced by the row space approximation, we are able to obtain the following guarantee for the model repair.

Theorem 5.2. *Under the conditions of Theorem 5.1, additionally assume that $\frac{\log p}{d}$, $\frac{\sqrt{\frac{n}{d}\log\left(\frac{ed}{n}\right)}}{1-\varepsilon}$ and $\frac{n^2 \log p}{p(1-\varepsilon)}$ are sufficiently small. We then have $\widetilde{W} = \widehat{W}$ and $\frac{1}{p}\|\widetilde{\beta} - \widehat{\beta}\|^2 \lesssim \frac{n^2 \log p}{p(1-\varepsilon)}$ with high probability.*

We also consider the case where $\widehat{W} = W(t_{\max})$ and $\widehat{\beta}$ is obtained by retraining β using the features $\psi(X\widehat{W})$. Recall that in this case we shall replace $\beta(0)$ by 0 in Algorithm 2. Note that the vector $\widehat{\beta}$ exactly lies in the row space of $\psi(X\widehat{W})$. This allows us to extend the result of Lemma 4.2 to the matrix $\psi(\widehat{W}^T X^T)$ with the help of Theorem 5.1. Then, we can directly apply Theorem 3.2. We are able to obtain exact recovery of both $\widehat{\beta}$ and $\widehat{W}$ in this case.

Theorem 5.3. *Under the conditions of Theorem 5.1, additionally assume that $\frac{\log p}{d}$, $\frac{\sqrt{\frac{n}{d}\log\left(\frac{ed}{n}\right)}}{1-\varepsilon}$,*

$\frac{n\log p}{p(1-\varepsilon)}$ *and* $\frac{n}{p}\left(\frac{\log p}{1-\varepsilon}\right)^{4/3}$ *are sufficiently small. We then have* $\widetilde{W}=\widehat{W}$ *and* $\widetilde{\beta}=\widehat{\beta}$ *with high probability.*

Remark 5.3. As long as the rate that ε tends to 1 is not so fast, the conditions of Theorem 5.2 and Theorem 5.3 can be simplified to $p \gg n^3$ and $d \gg n$ by ignoring the logarithmic factors. The condition $p \gg n^3$ ensures the good property of gradient descent in the NTK regime, but our experimental results show that it can potentially be weakened by an improved analysis.

6. Proofs of Theorem 5.2 and Theorem 5.3

We give proofs of Theorem 5.2 and Theorem 5.3 in this section. To prove Theorem 5.2, we need to extend Theorem 3.1. Consider $\eta = b + Au^* + z \in \mathbb{R}^m$, where the noise vector z satisfies (3.1), and $b \in \mathbb{R}^m$ is an arbitrary bias vector. Then, the estimator $\widehat{u} = \operatorname{argmin}_{u\in\mathbb{R}^k} \|\eta - Au\|_1$ satisfies the following theoretical guarantee.

Theorem 6.1. *Assume the design matrix A satisfies Condition A and Condition B. Then, as long as* $\frac{\overline{\lambda}\sqrt{\frac{k}{m}\log\left(\frac{em}{k}\right)}+\varepsilon\sigma\sqrt{\frac{k}{m}}}{\underline{\lambda}(1-\varepsilon)}$ *is sufficiently small and* $\frac{8\frac{1}{m}\sum_{i=1}^m |b_i|}{\underline{\lambda}(1-\varepsilon)} < 1$*, we have*

$$\|\widehat{u} - u^*\| \leq \frac{4\frac{1}{m}\sum_{i=1}^m |b_i|}{\underline{\lambda}(1-\varepsilon)},$$

with high probability.

It is easy to see that Theorem 3.1 is a special case when $b = 0$. Now we are ready to prove Theorem 5.2.

Proof of Theorem 5.2. We first analyze $\widehat{v}_1, ..., \widehat{v}_p$. The idea is to apply the result of Theorem 3.1 to each of the p robust regression problems. Thus, it suffices to check if the conditions of Theorem 3.1 hold for the p regression problems simultaneously. Since the p regression problems share the same Gaussian design matrix, Lemma 4.1 implies that Conditions A and B hold for all the p regression problems. Next, by scrutinizing the proof of Theorem 3.1, the randomness of the conclusion is from the noise vector Z_j through the empirical process bound given by Lemma A.6. With an additional union bound argument applied to (A.2) in its proof, Lemma A.6 can be extended to Z_j simultaneously for all $j \in [p]$ with an additional assumption that $\frac{\log p}{d}$ is sufficiently small. Then, by the same argument that leads to Corollary 4.1, we have $\widetilde{W}_j = \widehat{W}_j$ for all $j \in [p]$ with high probability.

To analyze $\widehat{u}$, we apply Theorem 6.1. Note that

$$\begin{aligned} \eta_j - \beta_j(0) &= \beta_j(t_{\max}) - \beta_j(0) + z_j \\ &= \sum_{t=0}^{t_{\max}-1} (\beta_j(t+1) - \beta_j(t)) + z_j \\ &= \frac{\gamma}{\sqrt{p}} \sum_{t=0}^{t_{\max}-1} \sum_{i=1}^{n} (y_i - u_i(t))\psi(W_j(t)^T x_i) + z_j \\ &= \frac{\gamma}{\sqrt{p}} \sum_{t=0}^{t_{\max}-1} \sum_{i=1}^{n} (y_i - u_i(t))(\psi(W_j(t)^T x_i) - \psi(W_j(0)^T x_i)) \\ &\quad + \frac{\gamma}{\sqrt{p}} \sum_{t=0}^{t_{\max}-1} \sum_{i=1}^{n} (y_i - u_i(t))\psi(W_j(0)^T x_i) + z_j. \end{aligned}$$

Thus, in the framework of Theorem 6.1, we can view $\eta - \beta(0)$ as the response, $\psi(X^T W(0)^T)$ as the design, z as the noise, and $b_j = \frac{\gamma}{\sqrt{p}} \sum_{t=0}^{t_{\max}-1} \sum_{i=1}^{n} (y_i - u_i(t))(\psi(W_j(t)^T x_i) - \psi(W_j(0)^T x_i))$ as the jth entry of the bias vector. By Lemma 4.2, we know that the design matrix $\psi(X^T W(0)^T)$ satisfies Condition A and Condition B. So it suffices to bound $\frac{1}{p}\sum_{j=1}^{p} |b_j|$. With the help of Theorem 5.1, we have

$$\begin{aligned} \frac{1}{p}\sum_{j=1}^{p} |b_j| &\leq \frac{\gamma}{p^{3/2}} \sum_{j=1}^{p} \sum_{t=0}^{t_{\max}-1} \sum_{i=1}^{n} |y_i - u_i(t)| |(W_j(t) - W_j(0))^T x_i| \\ &\leq \frac{R_1 \gamma}{p^{1/2}} \sum_{t=0}^{t_{\max}-1} \sum_{i=1}^{n} |y_i - u_i(t)| \|x_i\| \\ &\leq \frac{R_1 \gamma}{p^{1/2}} \sum_{t=0}^{t_{\max}-1} \|y - u(t)\| \sqrt{\sum_{i=1}^{n} \|x_i\|^2} \\ &\lesssim \frac{R_1}{p^{1/2}} \|y - u(0)\| \sqrt{\sum_{i=1}^{n} \|x_i\|^2} \\ &\lesssim \frac{n^2 \log p}{p}, \end{aligned}$$

where the last inequality is by $\sum_{i=1}^{n} \|x_i\|^2 \lesssim nd$ due to a standard chi-squared bound (Lemma A.4), and $\|u(0)\|^2 \lesssim n$ is due to Markov's inequality and $\mathbb{E}|u_i(0)|^2 = \mathbb{E}\mathsf{Var}(u_i(0)|X) \leq 1$. By Theorem 6.1 and Lemma A.7, we have $\frac{1}{p}\|\widetilde{\beta} - \widehat{\beta}\|^2 \lesssim \frac{n^3 \log p}{p}$, which is the desired conclusion. □

Proof of Theorem 5.3. The analysis of $\widehat{v}_1, ..., \widehat{v}_p$ is the same as that in the proof of Theorem 5.2, and we have $\widetilde{W}_j = \widehat{W}_j$ for all $j \in [p]$ with high probability.

To analyze $\widehat{u}$, we apply Theorem 3.2. It suffices to check Condition $\widetilde{A}$ and Condition B for the design matrix $\psi(X^T \widetilde{W}^T) = \psi(X^T \widehat{W}^T)$. To check Condition $\widetilde{A}$, we consider i.i.d. Rademacher random variables $\delta_1, ..., \delta_m$. Then, we define a different gradient update with initialization $\check{W}_j(0) = \delta_j W_j(0)$ and $\check{\beta}_j(0) = \delta_j \beta_j(0)$, and

$$\begin{aligned}
\check{\beta}_j(t) &= \check{\beta}_j(t-1) - \gamma \frac{\partial L(\beta, W)}{\partial \beta_j}|_{(\beta,W)=(\check{\beta}(t-1),\check{W}(t-1))}, \\
\check{W}_j(t) &= \check{W}_j(t-1) - \frac{\gamma}{d} \frac{\partial L(\beta, W)}{\partial W_j}|_{(\beta,W)=(\check{\beta}(t),\check{W}(t-1))}.
\end{aligned}$$

In other words, $(W(t), \beta(t))$ and $(\check{W}(t), \check{\beta}(t))$ only differ in terms of the initialization. Recall that $u_i(t) = \frac{1}{\sqrt{p}} \sum_{j=1}^p \beta_j(t) \psi(W_j(t)^T x_i)$. We also define

$$\begin{aligned}
\check{u}_i(t) &= \frac{1}{\sqrt{p}} \sum_{j=1}^p \check{\beta}_j(t) \psi(\check{W}_j(t)^T x_i), \\
v_i(t) &= \frac{1}{\sqrt{p}} \sum_{j=1}^p \beta_j(t) \psi(W_j(t-1)^T x_i), \\
\check{v}_i(t) &= \frac{1}{\sqrt{p}} \sum_{j=1}^p \check{\beta}_j(t) \psi(\check{W}_j(t-1)^T x_i).
\end{aligned}$$

It is easy to see that

$$\check{u}_i(t) = \frac{1}{\sqrt{p}} \sum_{j=1}^p \delta_j \beta_j(t) \psi(\delta_j W_j(t)^T x_i) = \frac{1}{\sqrt{p}} \sum_{j=1}^p \beta_j(t) \psi(W_j(t)^T x_i) = u_i(t).$$

Similarly, we also have

$$\check{v}_i(t) = \frac{1}{\sqrt{p}} \sum_{j=1}^p \delta_j \beta_j(t) \psi(\delta_j W_j(t-1)^T x_i) = \frac{1}{\sqrt{p}} \sum_{j=1}^p \beta_j(t) \psi(W_j(t-1)^T x_i) = v_i(t).$$

Suppose $\check{W}_j(k) = \delta_j W_j(k)$ and $\check{\beta}_j(k) = \delta_j \beta_j(k)$ are true. Since

$$\begin{aligned}
\frac{\partial L(\beta, W)}{\partial \beta_j}|_{(\beta,W)=(\check{\beta}(k),\check{W}(k))} &= \frac{1}{\sqrt{p}} \sum_{i=1}^n (\check{u}_i(k) - y_i) \psi(\check{W}_j(k)^T x_i) \\
&= \frac{1}{\sqrt{p}} \sum_{i=1}^n (u_i(k) - y_i) \psi(\delta_j W_j(k)^T x_i) \\
&= \delta_j \frac{1}{\sqrt{p}} \sum_{i=1}^n (u_i(k) - y_i) \psi(W_j(k)^T x_i) \\
&= \delta_j \frac{\partial L(\beta, W)}{\partial \beta_j}|_{(\beta,W)=(\beta(k),W(k))},
\end{aligned}$$

we have $\breve{\beta}_j(k+1) = \delta_j \beta_j(k+1)$. Then,

$$\begin{aligned}
\frac{\partial L(\beta, W)}{\partial W_j}|_{(\beta,W)=(\breve{\beta}(k+1),\breve{W}(k))} &= \frac{1}{\sqrt{p}} \breve{\beta}_j(k+1) \sum_{i=1}^{n} (\breve{v}_i(k+1) - y_i) \psi'(\breve{W}_j(k)^T x_i) x_i \\
&= \frac{1}{\sqrt{p}} \delta_j \beta_j(k+1) \sum_{i=1}^{n} (v_i(k+1) - y_i) \psi'(\delta_j W_j(k)^T x_i) x_i \\
&= \frac{1}{\sqrt{p}} \delta_j \beta_j(k+1) \sum_{i=1}^{n} (v_i(k+1) - y_i) \psi'(W_j(k)^T x_i) x_i \\
&= \delta_j \frac{\partial L(\beta, W)}{\partial W_j}|_{(\beta,W)=(\beta(k+1),W(k))},
\end{aligned}$$

and thus we also have $\breve{W}_j(k+1) = \delta_j W_j(k+1)$. A mathematical induction argument leads to $\breve{W}_j(t) = \delta_j W_j(t)$ and $\breve{\beta}_j(t) = \delta_j \beta_j(t)$ for all $t \geq 1$. Since $(\breve{W}(0), \breve{\beta}(0))$ and $(W(0), \beta(0))$ have the same distribution, we can conclude that $(\breve{W}(t), \breve{\beta}(t))$ and $(W(t), \beta(t))$ also have the same distribution. Therefore, Condition $\widetilde{A}$ holds for the design matrix $\psi(X^T \widehat{W}^T) = \psi(X^T W(t_{\max})^T)$.

We also need to check Condition B. By Theorem 5.1, we have

$$\begin{aligned}
&\left| \frac{1}{p} \sum_{j=1}^{p} \left| \sum_{i=1}^{n} \psi(\widehat{W}_j^T x_i) \Delta_i \right| - \frac{1}{p} \sum_{j=1}^{p} \left| \sum_{i=1}^{n} \psi(W_j(0)^T x_i) \Delta_i \right| \right| \\
\leq\ & \frac{1}{p} \sum_{j=1}^{p} \sum_{i=1}^{n} |\widehat{W}_j^T x_i - W_j(0)^T x_i| |\Delta_i| \\
\leq\ & R_1 \sum_{i=1}^{n} \|x_i\| |\Delta_i| \leq R_1 \sqrt{\sum_{i=1}^{n} \|x_i\|^2} \lesssim \frac{n^{3/2} \log p}{\sqrt{p}},
\end{aligned}$$

where $\sum_{i=1}^{n} \|x_i\|^2 \lesssim nd$ is by Lemma A.4. By Lemma 4.2, we can deduce that

$$\inf_{\|\Delta\|=1} \frac{1}{p} \sum_{j=1}^{p} \left| \sum_{i=1}^{n} \psi(\widehat{W}_j^T x_i) \Delta_i \right| \gtrsim 1,$$

as long as $\frac{n^{3/2} \log p}{\sqrt{p}}$ is sufficiently small. According to Lemma A.7, we also have

$$\sup_{\|\Delta\|=1} \frac{1}{p} \sum_{j=1}^{p} \left| \sum_{i=1}^{n} \psi(\widehat{W}_j^T x_i) \Delta_i \right|^2 \lesssim 1 + \frac{n^2 \log p}{\sqrt{p}}.$$

See (A.46) in the appendix for details of derivation. Therefore, Condition B holds with $\overline{\lambda}^2 \asymp 1 + \frac{n^2 \log p}{\sqrt{p}}$ and $\underline{\lambda} \asymp 1$. Apply Theorem 3.2, we have $\widetilde{\beta} = \widehat{\beta}$ with high probability as desired. □

7. Simulation studies

In this section we discuss experimental results for over-paramaterized linear models, random feature models, and neural networks, illustrating and confirming the theoretical results presented in the previous sections. In all of our experiments, the `quantreg` package in R is used to carry out the ℓ_1 optimization of (2.3) as quantile regression for quantile level $\tau = \frac{1}{2}$ using the Frisch-Newton interior point algorithm to solve the linear program (method `fn` in this package).

7.1. Over-parameterized linear models

We begin by giving further details of the simulation briefly discussed in Section 2. In this experiment we simulate underdetermined linear models where $p > n$. We generate n data points (x_i, y_i) where $y_i = x_i^T \theta^* + w_i$ with w_i an additive noise term. We then compute the minimum norm estimator

$$\widehat{\theta} = X^T (XX^T)^{-1} y.$$

The estimated model is corrupted to

$$\eta = \widehat{\theta} + z$$

where $z_j \sim (1 - \varepsilon)\delta_0 + \varepsilon Q$. The corrupted estimator is then repaired by performing median regression:

$$\begin{aligned} \widetilde{u} &= \operatorname{argmin} \|\eta - X^T u\|_1, \\ \widetilde{\theta} &= X^T \widetilde{u}. \end{aligned}$$

The design is sampled as $X_{ij} \sim N(0, 1)$ and we take $\theta_j^* \sim N(0, 1)$ and $Q = N(1, 1)$. In the plots shown in Figure 3 the sample size is fixed at $n = 100$ and the dimension p is varied according to $p/n = 200/j^2$ for a range of values of j. The plots show the empirical probability of exact repair $\widetilde{\theta} = \widehat{\theta}$ as a function of ε. Each point on the curves is the average repair success over 500 random trials. The roughly equal spacing of the curves agrees with the theory, which indicates that $\sqrt{n/p}/(1 - \varepsilon)$ should be sufficiently small for successful repair. The right plot in Figure 4 shows the per-coefficient repair probability, and the left plot shows the probability that the entire model is repaired; in this plot the sample size is $n = 50$. The per-coefficient repair probability is the empirical probability that $\widetilde{\theta}_j = \widehat{\theta}_j$, averaged over $j = 1, \dots, p$.

7.2. Varying the mean

In this experiment we simulate over-parameterized linear models with nonzero mean. The data are generated as $X_{ij} \sim N(\mu, 1)$ independently, where we vary the mean μ and fix the dimension $p = 500$ and sample size $n = 50$. The probability of successful repair is shown in Figure 5.

As expected, the fraction ε that allows successful repair decreases; it appears to saturate at some fixed value $\varepsilon_{\min}$. The mean μ cannot be made too large because it causes the design X to become ill-conditioned, and the median regression fails.

The inequality in Condition A in this case takes the form

$$\mathbb{E} \left\| \frac{1}{m} \sum_{i=1}^{m} c_i a_i \right\|^2 \leq \frac{k}{m} + \mu^2 k \equiv \frac{n}{p} + \mu^2 n.$$

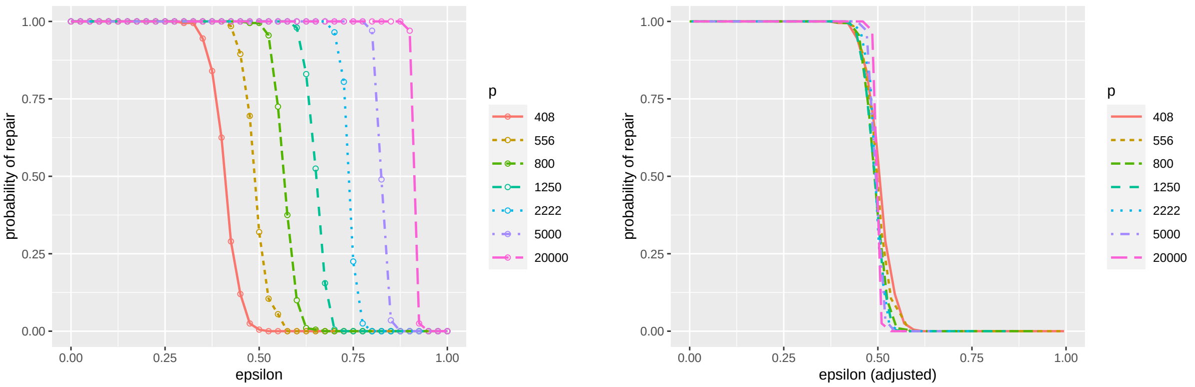

FIG 3. *Model repair for underdetermined linear models $y = X^T\theta + w$ with $p > n$. The left plot shows the empirical probability of successful model repair for $n = 100$ with the model dimension p varying as $p/n = 200/j^2$, for $j = 1, \ldots, 7$. Each point is an average over 500 random trials. The covariates are sampled as $N(0,1)$ and the corruption distribution is $Q = N(1,1)$. The right plot shows the repair probablity as a function of the adjusted corruption probability $\widetilde{\varepsilon}_j = \varepsilon + c' \cdot j - \frac{1}{2}$ for $c' = 0.085$.*

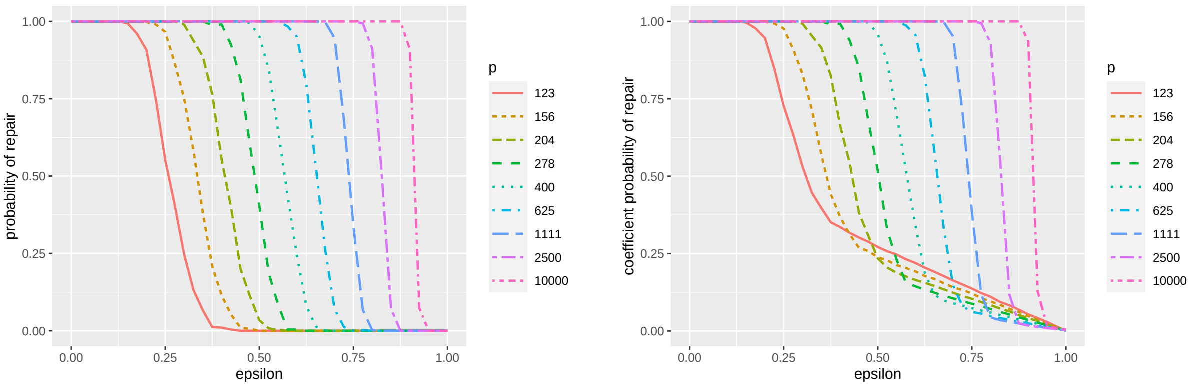

FIG 4. *Left: Empirical probability of successful model repair for $n = 50$ with the model dimension p varying as $p/n = 200/j^2$. Right: Per-coefficient probability of successful repair.*

The means μ in Figure 5 are taken to be $\mu_j = c_j/\sqrt{n}$ as c_j varies between zero and two.

7.3. *Random features models trained with gradient descent*

In this experiment we simulate over-parameterized random features models. We generate n data points $(\widetilde{x}_i, y_i)$ where $y_i = \widetilde{x}_i^T \theta^* + w_i$ with w_i an additive noise term. The covariates are generated as a layer of a random neural network, with $\widetilde{x}_i = \tanh(W^T x_i)$ where $x_i \in \mathbb{R}^d$ with $x_{ij} \sim N(0,1)$ and $W \in \mathbb{R}^{d \times p}$ with $W_{ij} \sim N(0, 1/d)$. We then approximate the least squares solution using gradient descent intialized at zero, with updates

$$\widehat{\theta}^{(t)} = \widehat{\theta}^{(t-1)} + \frac{\eta}{n} \widetilde{X}^T R^{(t-1)}$$

where the residual vector $R^{(t-1)} \in \mathbb{R}^n$ is given by $R_i^{(t-1)} = (y_i - \widetilde{x}_i^T \widehat{\theta}^{(t-1)})$. The step size η is selected empirically to insure convergence in under $1{,}000$ iterations. Figure 6 shows two sets of

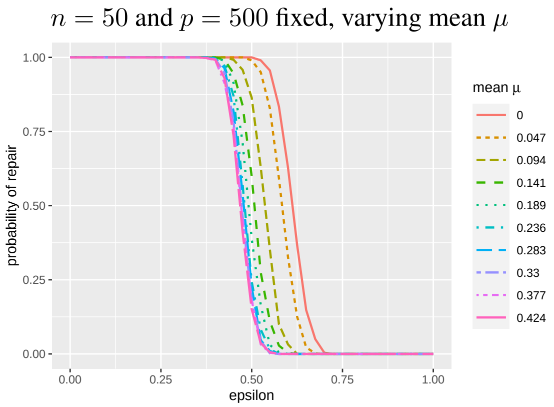

FIG 5. *Model repair for linear models with design entries $X_{i,j} \sim N(\mu, 1)$, where the mean μ is varied and the sample size and dimension are fixed at $n = 50$ and $p = 500$. Consistent with Theorem 3.1, a smaller corruption fraction ε is tolerated as the mean μ increases. In the plot above, the means are chosen as $\mu_j = c_j/\sqrt{n}$ for c_j varying between zero and two.*

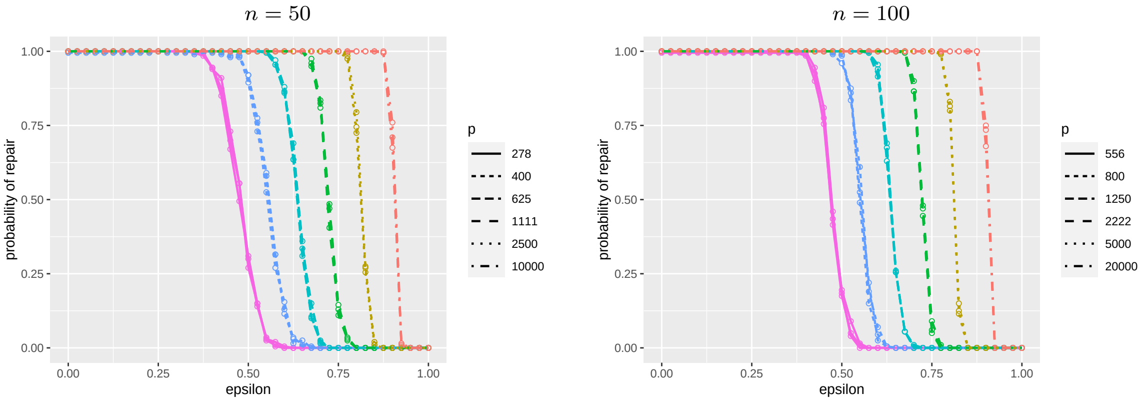

FIG 6. *Model repair for random feature models $y = \psi(XW)\theta + w$ with $p > n$, where $\psi(\cdot) = \tanh(\cdot)$ for $n = 50$ (left) and $n = 100$ (right). For each value of p, three values of d are evaluated, $d = p$, $d = \lceil 2p/3 \rceil$, and $d = \lceil p/2 \rceil$; the results are effectively the same for each d. The curves are very similar when $\tanh$ is replaced by ReLU, as long as the population mean is subtracted from the features.*

a

results, for $n = 50$ and $n = 100$. For each value of the final dimension p, three values of the original data dimension d are selected: $d = p$, $d = \lceil 2p/3 \rceil$, and $d = \lceil p/2 \rceil$. The recovery success curves for gradient descent are similar to those obtained for the minimal norm solution. The results are also similar if the ReLU activation function is used, as long as the population mean of the features is subtracted.

7.4. Neural networks

In this final set of simulations we investigate repair algorithms for neural networks. We report results using a single hidden layer and the use of the hyperbolic tangent activation function. Results

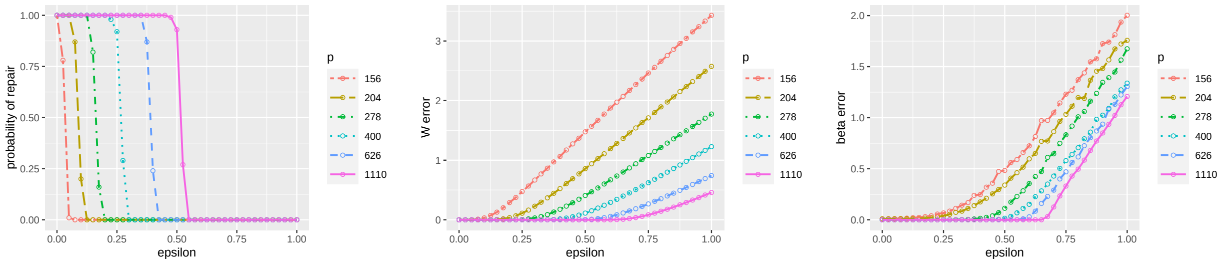

FIG 7. *Model repair for neural networks with a single hidden layer. The sample size is fixed at $n = 50$, the number of hidden units is p, and the input dimension is $d = p/2$; the dimenson of the design is $X \in \mathbb{R}^{n\times d}$, and the hidden layer is generated as $\tanh(XW)$ where $W \in \mathbb{R}^{d\times p}$. The predicted values are $\widehat{y} = \tanh(XW)\beta$. Left: The neural network is trained with gradient descent after initializing W_{ij} as $N(0, 1/d)$ and $\beta \sim N(0, I_p)$. After training, a forward pass is made where $\widehat{W}$ is fixed and the parameters β are retrained using gradient descent initialized at zero. This allows exact repair by running the linear program in stages, first repairing $\widehat{W}$, and then repairing $\widehat{\beta}$. Center and right: The neural network is trained with gradient descent using random initialization of W and β; no forward pass is made after training. The weight matrix W and weight vector β are then not repaired exactly. The center and right plots show the average L_2 error in the estimated coefficients W_{ij} and β_j as a function of the corruption fraction ε. In all plots, each point is the average over 100 random trials.*

using the ReLU activation are similar as long as the features are centered.

As described in Section 5, we consider two ways of training the models—with or without a forward pass to train the network parameters in stages. In the first approach, the network is trained using gradient descent, and the weights $\widehat{W} \in \mathbb{R}^{d\times p}$ are then fixed. Next, the weights $\beta \in \mathbb{R}^p$ are initialized at zero, and gradient descent over β is carried out using features $\widetilde{X} = \psi(X\widehat{W})$. This two-pass approach allows for exact repair, and only the initial weights $W(0)$ need to be accessed by the repair algorithm, using the seed value used in the pseudorandom number generator.

The left plot in Figure 7 shows the behavior of the linear program for exact repair when using this two-stage training algorithm. The sample size is fixed at $n = 50$, the number of "neurons" W_j varies as p, and we take $d = p/2$. It can been seen that similar repair curves are obtained as for linear models, but the curves are shifted toward the left, indicating an overall smaller probability of successful repair. This is because successful repair requires that $O(p^2)$ parameters are recovered, each of the p columns $W_j \in \mathbb{R}^d$ in addition to the vector $\beta \in \mathbb{R}^p$.

In the second approach, the neural network is trained using standard gradient descent, without a final forward pass. As described in Section 5, when trained in this manner the column space of $\widetilde{X} = \psi(XW)$ is continually changing. However, the "neural tangent kernel" analysis ensures that the linear program will approximately recover the trained parameters after they are corrupted by additive noise. This is seen in the center and right plots of Figure 7, which show the squared errors $(\widetilde{W}_{ij} - \widehat{W_{ij}})^2$ and $(\widetilde{\beta}_j - \widehat{\beta}_j)^2$, averaged over i and j. These results are consistent with our analysis, and suggest that the growth conditions on d and p in the results of our theorems are conservative.

8. Discussion

In this paper we introduced the problem of model repair, related it to robust estimation, and established a series of results showing the theoretical performance of a repair algorithm that is based on median regression. The specific models treated include linear models and families of neural networks trained using gradient descent. The experimental results largely validate the theory, quantifying how model repair requires over-parameterization in the model and redundancy in the estimator.

This work suggests several directions to explore in future research. A natural problem is to establish lower bounds for model repair. In particular, our results show the level of over-parameterization sufficient for repair algorithms based on ℓ_1 optimization. What level is required if the algorithm is not specified? Answering this question might exploit the rich literature on depth functions and multivariate generalizations of the median, together with minimax analysis for estimation and testing under the classical Huber model (Chen et al., 2018; Diakonikolas et al., 2017; Diakonikolas and Kane, 2019). In a different direction, Gao et al. (2019) introduces a connection between these optimizations and certain learning algorithms for adversarial neural networks called f-GANs, giving a variational characterization of robust estimation that could lead to new algorithmic procedures for model repair.

The repair problem also could be formulated in other ways. For example, the corruption model could be modified, allowing a dependence between z and X; a simple form of this dependence would be $\eta_j \,|\, X \sim (1-\varepsilon)\delta_{\widehat{\theta}_j} + \varepsilon Q_j$. What if the repair algorithm does not have access to the original training inputs $x_1, \ldots, x_n$? If a new unlabeled dataset $x'_1, \ldots, x'_m$ is available for which $\text{span}(x_1, \ldots, x_n) \subset \text{span}(x'_1, \ldots, x'_m)$, the results proven here will carry over. One could consider other formulations that make different assumptions on the information that is available.

It can be expected that the results for neural networks with a single layer established in the current paper can be extended to multiple layers, based on results for multilayer networks that extend the analysis of gradient descent of Du et al. (2018b), including Allen-Zhu et al. (2018) and Du et al. (2018a). It would be interesting to consider model repair for other architectures and estimation algorithms, including convolutional networks and deep generative networks (Dinh et al., 2017; Goodfellow et al., 2014; Kingma and Dhariwal, 2018).

Another natural direction to explore is repair for other families of statistical models, where over-parameterization and redundancy may take different forms. For instance, in classical Gaussian sequence models for orthonormal bases, additional coefficients could provide insurance against corrupted estimates. It would also be interesting to explore sequence models such as isotonic and shape-constrained regression. For example, consider the piecewise constant signals with k pieces,

$$\Theta_k = \{\theta : \theta_i = \mu_j \text{ for } i \in (a_{j-1}, a_j] \text{ for some } 0 = a_0 \leq a_1 \leq \cdots \leq a_k = n\}.$$

Adaptivity of the least–squares estimator to k has been well-established (Bellec and Tsybakov, 2015; Chatterjee, 2014; Chatterjee et al., 2015); but the redundancy in the sequence could also be exploited in model repair. If an initial estimator $\widehat{\theta}$ is corrupted to $\eta = \widehat{\theta} + z$, a natural repair procedure is

$$\widetilde{\theta}_i = \text{mode}(\{\eta_{i-h}, \ldots, \eta_{i+h}\})$$

with h acting as a bandwidth parameter. We conjecture that $\inf_h \mathbb{E}\|\widetilde{\eta} - \theta\|^2 = O(k \log(n/k))$. In this setting, the piecewise constant signal acts as a simple repetition code, with majority vote serving as a natural decoding procedure.

Returning to some of the motivation mentioned in the introduction, when training increasingly large neural networks it becomes necessary to estimate the models in a distributed manner, and erasures and errors may occur when communicating parameters across nodes, or after the trained model has been embedded in an application. Instead of running the repair program on a central hub, which would require sharing data and potentially compromising privacy, the linear program might also be distributed (Hong et al., 2012). Finally, drawing an analogy to brain plasticity and repair after trauma, if a spatially localized part of a multilayer network is permanently corrupted, the repair problem needs to be reformulated to allow "rewiring" the parameters to obtain a model whose predictions are close to those of the original model, possibly through specialized training. With appropriate formalization, these and other extensions might permit statistical analysis.

Acknowledgments

Research of CG is supported in part by NSF grant DMS-1712957 and NSF CAREER award DMS-1847590. Research of JL is supported in part by NSF grant CCF-1839308.

References

Allen-Zhu, Z., Li, Y., and Song, Z. (2018). A convergence theory for deep learning via over-parameterization. *arXiv:1811.03962*.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*.

Belkin, M., Hsu, D., and Xu, J. (2019). Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*.

Bellec, P. C. and Tsybakov, A. B. (2015). Sharp oracle bounds for monotone and convex regression through aggregation. *Journal of Machine Learning Research*, 16(56):1879–1892.

Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P. (2017). Consistent robust regression. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 2110–2119.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA.

Chatterjee, S. (2014). A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42(6):2340–2381.

Chatterjee, S., Guntuboyina, A., and Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics*, 43(4):1774–1800.

Chen, M., Gao, C., and Ren, Z. (2016). A general decision theory for Huber's ε-contamination model. *Electron. J. Statist.*, 10(2):3752–3774.

Chen, M., Gao, C., and Ren, Z. (2018). Robust covariance and scatter matrix estimation under huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960.

Cirel'son, B. S., Ibragimov, I. A., and Sudakov, V. (1976). Norms of Gaussian sample functions. In *Proceedings of the Third Japan—USSR Symposium on Probability Theory*, pages 20–41. Springer.

Cover, T. and Thomas, J. (2006). *Elements of Information Theory*. Wiley, 2nd edition.

Davidson, K. R. and Szarek, S. J. (2001). Local operator theory, random matrices and Banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2017). Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 999–1008. JMLR.org.

Diakonikolas, I. and Kane, D. M. (2019). Recent advances in algorithmic high-dimensional robust statistics. arXiv:1911.05911.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017). Density estimation using real NVP. arXiv:1605.08803.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. (2018a). Gradient descent finds global minima of deep neural networks. *arXiv:1811.03804*.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. (2018b). Gradient descent provably optimizes overparameterized neural networks. *arXiv:1810.02054*.

Gao, C. (2020). Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139–1170.

Gao, C., Liu, J., Yao, Y., and Zhu, W. (2019). Robust estimation via generative adversarial networks. In *International Conference on Learning Representations*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Hong, Y., Vaidya, J., and Lu, H. (2012). Secure and efficient distributed linear programming. *Journal of Computer Security*, 20:583–63.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580.

Joseph, A. and Barron, A. R. (2012). Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity. *IEEE Transactions on Information Theory*, 58(5):2541–2557.

Karmalkar, S. and Price, E. (2018). Compressed sensing with adversarial sparse noise via l1 regression. *arXiv preprint arXiv:1809.08055*.

Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 10215–10224. Curran Associates, Inc.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338.

Mei, S. and Montanari, A. (2019). The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*.

Neyshabur, B., Tomioka, R., and Srebro, N. (2014). In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*.

Nguyen, N. H. and Tran, T. D. (2013a). Exact recoverability from dense corrupted observations via ℓ_1 minimization. *IEEE Trans. Info. Theory*, 59(4):2017–2035.

Nguyen, N. H. and Tran, T. D. (2013b). Robust lasso with missing and grossly corrupted observations. *IEEE Trans. Info. Theory*, 59(4):2036–2056.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc.

Ross, S. M. and Peköz, E. A. (2007). *A second course in probability*. www. ProbabilityBookstore. com.

Rush, C., Greig, A., and Venkataramanan, R. (2017). Capacity-achieving sparse superposition codes via approximate message passing decoding. *IEEE Trans. Info. Theory*, 63(3):1476–1500.

Suggala, A. S., Bhatia, K., Ravikumar, P., and Jain, P. (2019). Adaptive hard thresholding for near-optimal consistent robust regression. *arXiv preprint arXiv:1903.08192*.

Tsakonas, E., Jaldén, J., Sidiropoulos, N. D., and Ottersten, B. (2014). Convergence of the huber regression m-estimate in the presence of dense outliers. *IEEE Signal Processing Letters*, 21(10):1211–1214.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Wright, J. A. and Ma, Y. (2010). Dense error correction via ℓ_1-minimization. *IEEE Trans. Info. Theory*, 56(7).

Appendix A: Proofs

A.1. Technical Lemmas

We present a few technical lemmas that will be used in the proofs. The first lemma is Hoeffding's inequality.

Lemma A.1 (Hoeffding (1963))**.** *Consider independent random variables $X_1, ..., X_n$ that satisfy $X_i \in [a_i, b_i]$ for all $i \in [n]$. Then, for any $t > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| > t\right) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Next, we need a central limit theorem with an explicit third moment bound. The following lemma is Theorem 2.20 of Ross and Peköz (2007).

Lemma A.2. *If $Z \sim N(0,1)$ and $W = \sum_{i=1}^{n} X_i$ where X_i are independent mean 0 and $\mathsf{Var}(W) = 1$, then*

$$\sup_z |\mathbb{P}(W \le z) - \mathbb{P}(Z \le z)| \le 2\sqrt{3\sum_{i=1}^{n}\mathbb{E}|X_i|^3}.$$

We also need a Talagrand Gaussian concentration inequality. The following version has explicit constants.

Lemma A.3 (Cirel'son et al. (1976))**.** *Let $f : \mathbb{R}^k \to \mathbb{R}$ be a Lipschitz function with constant $L > 0$. That is, $|f(x) - f(y)| \le L\|x - y\|$ for all $x, y \in \mathbb{R}^k$. Then, for any $t > 0$,*

$$\mathbb{P}\left(|f(Z) - \mathbb{E}f(Z)| > t\right) \le 2\exp\left(-\frac{t^2}{2L^2}\right),$$

where $Z \sim N(0, I_k)$.

Finally, we present two lemmas on the concentration of norms and inner products of multivariate Gaussians.

Lemma A.4 (Laurent and Massart (2000))**.** *For any $t > 0$, we have*

$$\begin{aligned} \mathbb{P}\left(\chi_k^2 \ge k + 2\sqrt{tk} + 2t\right) &\le e^{-t}, \\ \mathbb{P}\left(\chi_k^2 \le k - 2\sqrt{tk}\right) &\le e^{-t}. \end{aligned}$$

Lemma A.5. *Consider independent $Y_1, Y_2 \sim N(0, I_k)$. For any $t > 0$, we have*

$$\begin{aligned} \mathbb{P}\left(|\|Y_1\|\|Y_2\| - k| \ge 2\sqrt{tk} + 2t\right) &\le 4e^{-t}, \\ \mathbb{P}\left(|Y_1^T Y_2| \ge \sqrt{2kt} + 2t\right) &\le 2e^{-t}. \end{aligned}$$

Proof. By Lemma A.4, we have

$$
\begin{aligned}
\mathbb{P}\big(\|Y_1\|\|Y_2\|-k \geq & 2\sqrt{tk}+2t\big)\\
&\leq \mathbb{P}\left(\|Y_1\|^2 \geq k+2\sqrt{tk}+2t\right)+\mathbb{P}\left(\|Y_2\|^2 \geq k+2\sqrt{tk}+2t\right)\\
&\leq 2e^{-t},
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{P}\Big(\|Y_1\|\|Y_2\|-k \leq & -2\sqrt{tk}-2t\Big)\\
&\leq \mathbb{P}\left(\|Y_1\|^2 \leq k-2\sqrt{tk}\right)+\mathbb{P}\left(\|Y_2\|^2 \leq k-2\sqrt{tk}\right)\\
&\leq 2e^{-t}.
\end{aligned}
$$

Summing up the two bounds above, we obtain the first conclusion. For the second conclusion, note that

$$
\mathbb{P}\left(Y_1^TY_2 \geq x\right) \leq e^{-\lambda x}\mathbb{E}e^{\lambda Y_1^TY_2} = \exp\left(-\lambda x - \frac{k}{2}\log(1-\lambda^2)\right) \leq \exp\left(-\lambda x + \frac{k}{2}\lambda^2\right),
$$

for any $x>0$ and $\lambda \in (0,1)$. Optimize over $\lambda \in (0,1)$, and we obtain $\mathbb{P}\left(Y_1^TY_2 > x\right) \leq e^{-\frac{1}{2}\left(\frac{x^2}{k}\wedge x\right)}$. Take $x=\sqrt{2kt}+2t$, and then we obtain the bound

$$
\mathbb{P}\left(Y_1^TY_2 \geq \sqrt{2kt}+2t\right) \leq e^{-t},
$$

which immediately implies the second conclusion. □

A.2. Proofs of Theorem 3.2 and Theorem 6.1

We first establish an empirical process result.

Lemma A.6. *Consider independent random variables $z_1,...,z_m$. Assume $k/m \leq 1$. Then, for any $t \in (0,1/2)$ and any fixed $A^T=(a_1,...,a_m)^T \in \mathbb{R}^{m\times k}$ such that (3.4) holds, we have*

$$
\sup_{\|\Delta\|\leq t}\left|\frac{1}{m}\sum_{i=1}^m[(|a_i^T\Delta - z_i|-|z_i|)-\mathbb{E}(|a_i^T\Delta - z_i|-|z_i|)]\right| | \lesssim t\bar{\lambda}\sqrt{\frac{k}{m}\log\left(\frac{em}{k}\right)},
$$

with high probability.

Proof. We use the notation $G_m(\Delta)=\frac{1}{m}\sum_{i=1}^m[(|a_i^T\Delta - z_i|-|z_i|)-\mathbb{E}(|a_i^T\Delta - z_i|-|z_i|)]$, and we apply a discretization argument. For the Euclidian ball $B_k(t)=\{\Delta \in \mathbb{R}^k : \|\Delta\| \leq t\}$, there exists a subset $\mathcal{N}_{t,\zeta} \subset B_k(t)$, such that for any $\Delta \in B_k(t)$, there exists a $\Delta' \in \mathcal{N}_{t,\zeta}$ that satisfies $\|\Delta - \Delta'\| \leq \zeta$, and we also have the bound $\log|\mathcal{N}_{t,\zeta}| \leq k\log(1+2t/\zeta)$ according to Lemma 5.2 of

Vershynin (2010). For any $\Delta \in B_k(t)$ and the corresponding $\Delta' \in \mathcal{N}_{t,\zeta}$ that satisfies $\|\Delta - \Delta'\| \leq \zeta$, we have

$$\begin{aligned} |G_m(\Delta) - G_m(\Delta')| &\leq 2\frac{1}{m}\sum_{i=1}^{m} |a_i^T(\Delta - \Delta')| \\ &\leq 2\sqrt{\frac{1}{m}\sum_{i=1}^{m} |a_i^T(\Delta - \Delta')|^2} \leq 2\overline{\lambda}\zeta, \end{aligned}$$

where the last line is due to the condition (3.4). Thus,

$$|G_m(\Delta)| \leq |G_m(\Delta')| + 2\overline{\lambda}\zeta.$$

Taking the supremum over both sides of the inequality, we obtain

$$\sup_{\|\Delta\| \leq t} |G_m(\Delta)| \leq \max_{\Delta \in \mathcal{N}_{t,\zeta}} |G_m(\Delta)| + 2\overline{\lambda}\zeta. \tag{A.1}$$

For any $\Delta \in B_k(t)$, we have

$$\frac{1}{m}\sum_{i=1}^{m} \left(|a_i^T\Delta - z_i| - |z_i|\right)^2 \leq \frac{1}{m}\sum_{i=1}^{m} |a_i^T\Delta|^2 \leq \overline{\lambda}^2 t^2.$$

By Lemma A.1, we have

$$\mathbb{P}\left(|G_m(\Delta)| > x\right) \leq 2\exp\left(-\frac{2mx^2}{\overline{\lambda}^2 t^2}\right).$$

A union bound argument leads to

$$\mathbb{P}\left(\max_{\Delta \in \mathcal{N}_{t,\zeta}} |G_m(\Delta)| > x\right) \leq 2\exp\left(-\frac{2mx^2}{\overline{\lambda}^2 t^2} + k\log\left(1 + \frac{2t}{\zeta}\right)\right). \tag{A.2}$$

By choosing $x^2 \asymp \frac{t^2\bar{\lambda}^2 k\log(1+2t/\zeta)}{m}$, we have

$$\max_{\Delta \in \mathcal{N}_{t,\zeta}} |G_m(\Delta)| \lesssim t\bar{\lambda}\sqrt{\frac{k\log(1+2t/\zeta)}{m}},$$

with high probability. Together with the bound (A.1), we have

$$\sup_{\|\Delta\| \leq t} |G_m(\Delta)| \lesssim t\bar{\lambda}\sqrt{\frac{k\log(1+2t/\zeta)}{m}} + \bar{\lambda}\zeta,$$

with high probability. The choice $\zeta = t\sqrt{k/m}$ leads to the desired result. □

Proof of Theorem 3.2. Recall the definition of $L_m(u)$ in the proof of Theorem 3.1. We introduce i.i.d. Rademacher random variables $\delta_1, ..., \delta_m$. With the notation $\widetilde{a}_i = \delta_i a_i$, $\widetilde{b}_i = \delta_i b_i$ and $\widetilde{z}_i = \delta_i z_i$, we can write

$$L_m(u) = \frac{1}{m}\sum_{i=1}^{m}\left(|\widetilde{a}_i^T(u^* - u) + \widetilde{z}_i| - |\widetilde{z}_i|\right).$$

Let $\widetilde{A} \in \mathbb{R}^{m\times k}$ be the matrix whose ith row is $\widetilde{a}_i^T$. By the symmetry of A, we have $\mathbb{P}(\widetilde{A} \in U|\delta) = \mathbb{P}(\widetilde{A} \in U) = \mathbb{P}(A \in U)$ for any measurable set U. Therefore, for any measurable sets U and V, we have

$$\begin{aligned}\mathbb{P}(\widetilde{A} \in U, \widetilde{z} \in V) &= \mathbb{E}\mathbb{P}(\widetilde{A} \in U, \widetilde{z} \in V|\delta) \\ &= \mathbb{E}\mathbb{P}(\widetilde{A} \in U|\delta)\mathbb{P}(\widetilde{z} \in V|\delta) \\ &= \mathbb{E}\mathbb{P}(\widetilde{A} \in U)\mathbb{P}(\widetilde{z} \in V|\delta) \\ &= \mathbb{P}(\widetilde{A} \in U)\mathbb{P}(\widetilde{z} \in V),\end{aligned}$$

and thus $\widetilde{A}$ ad $\widetilde{z}$ are independent. Define $L(u) = \mathbb{E}(L_m(u)|\widetilde{A})$. Suppose $\|\widehat{u} - u^*\| \geq t$, we must have

$$\inf_{\|u-u^*\|\geq t} L_m(u) \leq L_m(u^*).$$

By the convexity of $L_m(u)$, we can replace $\|u - u^*\| \geq t$ by $\|u - u^*\| = t$ and the above inequality still holds, and therefore $\inf_{\|u-u^*\|=t} L_m(u) \leq 0$. This implies

$$\inf_{\|u-u^*\|=t} L(u) \leq \sup_{\|u-u^*\|=t} |L_m(u) - L(u)|. \tag{A.3}$$

Now we study $L(u)$. Introduce the function $f_i(x) = \mathbb{E}(|x + \widetilde{z}_i| - |\widetilde{z}_i|)$ so that we can write $L(u) = \frac{1}{m}\sum_{i=1}^m f_i(\widetilde{a}_i^T(u^* - u))$. For any $x \geq 0$,

$$\begin{aligned}f_i(x) &= \mathbb{E}(|x + \widetilde{z}_i| - |\widetilde{z}_i|)\mathbb{I}\{\widetilde{z}_i < -x\} + \mathbb{E}(|x + \widetilde{z}_i| - |\widetilde{z}_i|)\mathbb{I}\{\widetilde{z}_i > 0\} \\ &\quad + \mathbb{E}(|x + \widetilde{z}_i| - |\widetilde{z}_i|)\mathbb{I}\{-x \leq \widetilde{z}_i < 0\} + x\mathbb{P}(\widetilde{z}_i = 0) \\ &= -x\mathbb{P}(\widetilde{z}_i < -x) + x\mathbb{P}(\widetilde{z}_i > 0) + \mathbb{E}(x + 2\widetilde{z}_i)\mathbb{I}\{-x \leq \widetilde{z}_i < 0\} + x\mathbb{P}(\widetilde{z}_i = 0) \\ &\geq -x\mathbb{P}(\widetilde{z}_i < -x) + x\mathbb{P}(\widetilde{z}_i > 0) - x\mathbb{P}(-x \leq \widetilde{z}_i < 0) + x\mathbb{P}(\widetilde{z}_i = 0) \\ &= -x\mathbb{P}(\widetilde{z}_i < -x) + x\mathbb{P}(\widetilde{z}_i < 0) - x\mathbb{P}(-x \leq \widetilde{z}_i < 0) + x\mathbb{P}(\widetilde{z}_i = 0) \\ &\geq x\mathbb{P}(\widetilde{z}_i = 0) \\ &\geq (1 - \varepsilon)x.\end{aligned}$$

By the symmetry of $\widetilde{z}_i$, we also have

$$f_i(-x) = \mathbb{E}(|-x + \widetilde{z}_i| - |\widetilde{z}_i|) = \mathbb{E}(|x - \widetilde{z}_i| - |\widetilde{z}_i|) = \mathbb{E}(|x + \widetilde{z}_i| - |\widetilde{z}_i|) = f_i(x),$$

which implies $f_i(x) \geq (1-\varepsilon)|x|$. Therefore, for any u such that $\|u - u^*\| = t$, we have

$$\begin{aligned} L(u) &= \frac{1}{m}\sum_{i=1}^m f_i(\widetilde{a}_i^T(u^* - u)) \\ &\geq (1-\varepsilon)\frac{1}{m}\sum_{i=1}^m |\widetilde{a}_i^T(u^* - u)| \\ &= (1-\varepsilon)\frac{1}{m}\sum_{i=1}^m |a_i^T(u^* - u)| \\ &\geq \underline{\lambda}(1-\varepsilon)t, \end{aligned}$$

where the last inequality is by (3.3). Together with (A.3), we have

$$\mathbb{P}\left(\|\widehat{u} - u\| \geq t\right) \leq \mathbb{P}\left(\sup_{\|u-u^*\|=t} |L_m(u) - L(u)| \geq \underline{\lambda}(1-\varepsilon)t/2\right). \tag{A.4}$$

Since the condition (3.4) continues to hold with A replaced by $\widetilde{A}$, we can apply Lemma A.6 and obtain that

$$\sup_{\|u-u^*\|=t} |L_m(u) - L(u)| \lesssim t\overline{\lambda}\sqrt{\frac{k}{m}\log\left(\frac{em}{k}\right)},$$

with high probability. Under the conditions of the theorem, we know that $\frac{t\overline{\lambda}\sqrt{\frac{k}{m}\log\left(\frac{em}{k}\right)}}{\underline{\lambda}(1-\varepsilon)t}$ is sufficiently small, and thus by (A.4), $\|\widehat{u} - u^*\| < t$ with high probability. Since t is arbitrary, we must have $\widehat{u} = u^*$. □

Proof of Theorem 6.1. Recall the definitions of $L_m(u)$ and $L(u)$ in the proof of Theorem 3.1. Define

$$K_m(u) = \frac{1}{m}\sum_{i=1}^m \left(|b_i + a_i^T(u^* - u) + z_i| - |z_i|\right).$$

It is easy to see that

$$\sup_u |L_m(u) - K_m(u)| \leq \frac{1}{m}\sum_{i=1}^m |b_i|. \tag{A.5}$$

Suppose $\|\widehat{u} - u^*\| \geq t$, we must have $\inf_{\|u-u^*\|\geq t} K_m(u) \leq K_m(u^*)$. By the convexity of $K_m(u)$, we can replace $\|u - u^*\| \geq t$ by $\|u - u^*\| = t$ and the inequality still holds. By (A.5), we have $K_m(u^*) \leq \frac{1}{m}\sum_{i=1}^m |b_i|$, and therefore $\inf_{\|u-u^*\|=t} K_m(u) \leq \frac{1}{m}\sum_{i=1}^m |b_i|$. Since

$$\begin{aligned} \inf_{\|u-u^*\|=t} K_m(u) &\geq \inf_{\|u-u^*\|=t} L_m(u) - \frac{1}{m}\sum_{i=1}^m |b_i| \\ &\geq \inf_{\|u-u^*\|=t} L(u) + \inf_{\|u-u^*\|=t} (L_m(u) - L(u)) - \frac{1}{m}\sum_{i=1}^m |b_i|, \end{aligned}$$

we then have

$$\inf_{\|u-u^*\|=t} L(u) \le \sup_{\|u-u^*\|=t} |L_m(u) - L(u)| + 2\frac{1}{m}\sum_{i=1}^m |b_i|. \tag{A.6}$$

With the lower bound for (3.6) and the upper bound (3.7), we obtain

$$\underline{\lambda}(1-\varepsilon)t - \varepsilon t\sigma\sqrt{\frac{k}{m}} - 2\frac{1}{m}\sum_{i=1}^m |b_i| \lesssim t\overline{\lambda}\sqrt{\frac{k}{m}\log\left(\frac{em}{k}\right)},$$

which is impossible with the choice $t = \frac{4\frac{1}{m}\sum_{i=1}^m |b_i|}{\underline{\lambda}(1-\varepsilon)}$ when $\frac{\overline{\lambda}\sqrt{\frac{k}{m}\log\left(\frac{em}{k}\right)}+\varepsilon\sigma\sqrt{\frac{k}{m}}}{\underline{\lambda}(1-\varepsilon)}$ is sufficiently small. Thus, we obtain the desired conclusion. □

A.3. Proofs of Lemma 4.1, Corollary 4.1, Lemma 4.2 and Corollary 4.2

Proof of Lemma 4.1. Condition A is obvious. For Condition B, we have

$$\inf_{\|\Delta\|=1} \frac{1}{p}\sum_{j=1}^p |a_j^T\Delta| \ge \sqrt{\frac{2}{\pi}} - \sup_{\|\Delta\|=1}\left|\frac{1}{p}\sum_{j=1}^p |a_j^T\Delta| - \sqrt{\frac{2}{\pi}}\right|,$$

and we will analyze the second term on the right hand side of the inequality above via a discretization argument for the Euclidean sphere $S^{n-1} = \{\Delta \in \mathbb{R}^n : \|\Delta\| = 1\}$. There exists a subset $\mathcal{N}_\zeta \subset S^{n-1}$, such that for any $\Delta \in S^{n-1}$, there exists a $\Delta' \in \mathcal{N}_\zeta$ that satisfies $\|\Delta - \Delta'\| \le \zeta$, and we also have the bound $\log|\mathcal{N}_\zeta| \le n\log(1 + 2/\zeta)$ according to Lemma 5.2 of Vershynin (2010). For any $\Delta \in S^{n-1}$ and the corresponding $\Delta' \in \mathcal{N}_\zeta$ that satisfies $\|\Delta - \Delta'\| \le \zeta$, we have

$$\begin{aligned}\left|\frac{1}{p}\sum_{j=1}^p |a_j^T\Delta| - \sqrt{\frac{2}{\pi}}\right| &\le \left|\frac{1}{p}\sum_{j=1}^p |a_j^T\Delta'| - \sqrt{\frac{2}{\pi}}\right| + \zeta\sup_{\|\Delta\|=1}\frac{1}{p}\sum_{j=1}^p |a_j^T\Delta| \\ &\le \left|\frac{1}{p}\sum_{j=1}^p |a_j^T\Delta'| - \sqrt{\frac{2}{\pi}}\right| + \zeta\sup_{\|\Delta\|=1}\left|\frac{1}{p}\sum_{j=1}^p |a_j^T\Delta| - \sqrt{\frac{2}{\pi}}\right| + \zeta\sqrt{\frac{2}{\pi}}.\end{aligned}$$

Taking the supremum on both sides of the inequality, with some rearrangement, we obtain

$$\sup_{\|\Delta\|=1}\left|\frac{1}{p}\sum_{j=1}^p |a_j^T\Delta| - \sqrt{\frac{2}{\pi}}\right| \le (1-\zeta)^{-1}\max_{\Delta\in\mathcal{N}_\zeta}\left|\frac{1}{p}\sum_{j=1}^p |a_j^T\Delta| - \sqrt{\frac{2}{\pi}}\right| + \frac{\zeta}{1-\zeta}\sqrt{\frac{2}{\pi}}.$$

Setting $\zeta = 1/3$, we then have

$$\inf_{\|\Delta\|=1}\frac{1}{p}\sum_{j=1}^p |a_j^T\Delta| \ge (2\pi)^{-1} - \frac{3}{2}\max_{\Delta\in\mathcal{N}_{1/3}}\left|\frac{1}{p}\sum_{j=1}^p |a_j^T\Delta| - \sqrt{\frac{2}{\pi}}\right|.$$

Lemma A.3 together with a union bound argument leads to

$$\mathbb{P}\left(\max_{\Delta\in\mathcal{N}_{1/3}}\left|\frac{1}{p}\sum_{j=1}^p |a_j^T\Delta| - \sqrt{\frac{2}{\pi}}\right| > t\right) \le 2\exp\left(n\log(7) - \frac{pt^2}{2}\right),$$

which implies $\max_{\Delta\in\mathcal{N}_{1/3}}\left|\frac{1}{p}\sum_{j=1}^p|a_j^T\Delta|-\sqrt{\frac{2}{\pi}}\right|\lesssim\sqrt{\frac{n}{p}}$ with high probability. Since n/p is sufficiently small, we have $\inf_{\|\Delta\|=1}\frac{1}{p}\sum_{j=1}^p|a_j^T\Delta|\gtrsim 1$ with high probability as desired. The high probability bound $\sup_{\|\Delta\|=1}\frac{1}{p}\sum_{j=1}^p|a_j^T\Delta|^2=\|A\|_{\text{op}}^2/p\lesssim 1+n/p$ is by Davidson and Szarek (2001), and the proof is complete. □

Proof of Corollary 4.1. Since $\widehat{\theta}$ belongs to the row space of X, there exists some $u^*\in\mathbb{R}^n$ such that $\widehat{\theta}=X^Tu^*$. By Theorem 3.1 and Lemma 4.1, we know that $\widetilde{u}=u^*$ with high probability, and therefore $\widetilde{\theta}=X^T\widetilde{u}=X^Tu^*=\widehat{\theta}$. □

Now we state the proof of Lemma 4.2. Note that Condition A holds because

$$\sum_{i=1}^n\mathbb{E}\left(\frac{1}{p}\sum_{j=1}^p c_j\psi(W_j^Tx_i)\right)^2\leq\sum_{i=1}^n\frac{1}{p^2}\sum_{j=1}^p\mathbb{E}|\psi(W_j^Tx_i)|^2\leq\frac{n}{p},$$

and we only need to prove Condition B. We present the proofs of (3.3) and (3.4) separately.

Proof of (3.3) of Lemma 4.2. Let us adopt the notation that

$$f(W,X,\Delta)=\frac{1}{p}\sum_{j=1}^p\left|\sum_{i=1}^n\psi(W_j^Tx_i)\Delta_i\right|.$$

Define $g(X,\Delta)=\mathbb{E}(f(W,X,\Delta)|X)$. We then have

$$\begin{aligned}\inf_{\|\Delta\|=1}f(W,X,\Delta) &\geq \inf_{\|\Delta\|=1}\mathbb{E}f(W,X,\Delta)-\sup_{\|\Delta\|=1}|f(W,X,\Delta)-\mathbb{E}f(W,X,\Delta)|\\ &\geq \inf_{\|\Delta\|=1}\mathbb{E}f(W,X,\Delta) &\text{(A.7)}\\ &\quad-\sup_{\|\Delta\|=1}|f(W,X,\Delta)-\mathbb{E}(f(W,X,\Delta)|X)| &\text{(A.8)}\\ &\quad-\sup_{\|\Delta\|=1}|g(X,\Delta)-\mathbb{E}g(X,\Delta)|. &\text{(A.9)}\end{aligned}$$

We will analyze the three terms above separately.

Analysis of (A.7). For any Δ such that $\|\Delta\|=1$, we have

$$\begin{aligned}
\mathbb{E}f(W,X,\Delta) &= \mathbb{E}\left|\sum_{i=1}^n \psi(W^Tx_i)\Delta_i\right| \\
&\geq \mathbb{E}\left(\left|\sum_{i=1}^n \psi(W^Tx_i)\Delta_i\right| \mathbb{I}\left\{\left|\sum_{i=1}^n \psi(W^Tx_i)\Delta_i\right| \geq 1, 1/2\leq \|W\|^2\leq 2\right\}\right) \\
&\geq \mathbb{P}\left(\left|\sum_{i=1}^n \psi(W^Tx_i)\Delta_i\right| \geq 1, 1/2\leq \|W\|^2\leq 2\right) \\
&= \mathbb{P}\left(\left|\sum_{i=1}^n \psi(W^Tx_i)\Delta_i\right| \geq 1 \middle| 1/2\leq \|W\|^2\leq 2\right)\mathbb{P}\left(1/2\leq \|W\|^2\leq 2\right) \\
&\geq \mathbb{P}\left(\left|\sum_{i=1}^n \psi(W^Tx_i)\Delta_i\right| \geq 1 \middle| 1/2\leq \|W\|^2\leq 2\right)\left(1-2\exp(-d/16)\right),
\end{aligned}$$

where the last inequality is by Lemma A.4. It is easy to see that

$$\mathsf{Var}\left(\psi(W^Tx)|W\right) \leq \mathbb{E}(|\psi(W^Tx)|^2|W) \leq 1.$$

Moreover, for any W such that $1/2\leq \|W\|^2\leq 2$,

$$\mathsf{Var}\left(\psi(W^Tx)|W\right) = \mathbb{E}(|\psi(W^Tx)|^2|W) \geq \frac{1}{5}\mathbb{P}\left(|W^Tx|>1/2|W\right) \geq \frac{1}{5}\mathbb{P}(|N(0,1)|\geq 1/\sqrt{2}),$$

which is at least $1/20$. In summary, we have

$$1/20 \leq \mathsf{Var}\left(\psi(W^Tx)|W\right) \leq 1,$$

for any W such that $1/2\leq \|W\|^2\leq 2$. By Lemma A.2, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{i=1}^n \psi(W^Tx_i)\Delta_i\right| \geq 1 \middle| 1/2\leq \|W\|^2\leq 2\right) \\
\geq\ & \mathbb{P}\left(\frac{\left|\sum_{i=1}^n \psi(W^Tx_i)\Delta_i\right|}{\sqrt{\mathsf{Var}\left(\psi(W^Tx)|W\right)}} \geq \sqrt{20} \middle| 1/2\leq \|W\|^2\leq 2\right) \\
\geq\ & \mathbb{P}\left(N(0,1)>\sqrt{20}\right) - \sup_{1/2\leq\|W\|^2\leq 2} 2\sqrt{3\sum_{i=1}^n |\Delta_i|^3 \frac{\mathbb{E}\left(|\psi(W^Tx_i)|^3|W\right)}{\left(\mathsf{Var}\left(\psi(W^Tx)|W\right)\right)^{3/2}}} \\
\geq\ & \mathbb{P}\left(N(0,1)>\sqrt{20}\right) - 35\sqrt{\sum_{i=1}^n |\Delta_i|^3} \\
\geq\ & \mathbb{P}\left(N(0,1)>\sqrt{20}\right) - 35\max_{1\leq i\leq n}|\Delta_i|^{3/2}.
\end{aligned}$$

Hence, when $\max_{1\leq i\leq n}|\Delta_i|^{3/2}\leq \delta_0^{3/2} := \mathbb{P}\left(N(0,1)>\sqrt{20}\right)/70$, we can lower bound the expectation $\mathbb{E}f(W,X,\Delta)$ by an absolute constant, and we conclude that

$$\inf_{\|\Delta\|=1, \max_{1\leq i\leq n}|\Delta_i|\leq\delta_0} \mathbb{E}f(W,X,\Delta) \gtrsim 1. \tag{A.10}$$

We also need to consider the case when $\max_{1\le i\le n}|\Delta_i| > \delta_0$. Without loss of generality, we can assume $\Delta_1 > \delta_0$. We then lower bound $\mathbb{E}f(W,X,\Delta)$ by

$$\begin{aligned}
& \mathbb{E}\left(\left|\sum_{i=1}^n \psi(W^T x_i)\Delta_i\right| \mathbb{I}\left\{\sum_{i=1}^n \psi(W^T x_i)\Delta_i \ge \delta_0/2, 1/2 \le \|W\|^2 \le 2\right\}\right) \\
\ge\ & \frac{\delta_0}{2}\mathbb{P}\left(\sum_{i=1}^n \psi(W^T x_i)\Delta_i \ge \delta_0/2 \middle| 1/2 \le \|W\|^2 \le 2\right)\mathbb{P}\left(1/2 \le \|W\|^2 \le 2\right) \\
\ge\ & \frac{\delta_0}{2}\mathbb{P}\left(\psi(W^T x_1)\Delta_1 \ge \delta_0/2 \middle| 1/2 \le \|W\|^2 \le 2\right) \\
& \times\mathbb{P}\left(\sum_{i=2}^n \psi(W^T x_i)\Delta_i \ge 0 \middle| 1/2 \le \|W\|^2 \le 2\right)(1-2\exp(-d/16)) \\
=\ & \frac{\delta_0}{4}\mathbb{P}\left(\psi(W^T x_1)\Delta_1 \ge \delta_0/2 \middle| 1/2 \le \|W\|^2 \le 2\right)(1-2\exp(-d/16)).
\end{aligned}$$

For any W that satisfies $1/2 \le \|W\|^2 \le 2$, we have

$$\begin{aligned}
\mathbb{P}\left(\psi(W^T x_1)\Delta_1 \ge \delta_0/2 \middle| W\right) &\ge \mathbb{P}\left(\psi(W^T x_1) \ge 1/2 \middle| W\right) \\
&\ge \mathbb{P}\left(W^T x_1 \ge 1 \middle| W\right) \\
&\ge \mathbb{P}\left(N(0,1) \ge \sqrt{2}\right),
\end{aligned}$$

which is a constant. Therefore, we have

$$\mathbb{E}f(W,X,\Delta) \ge \frac{\delta_0}{4}(1-2\exp(-d/16))\,\mathbb{P}\left(N(0,1)\ge\sqrt{2}\right) \gtrsim 1,$$

and we can conclude that

$$\inf_{\|\Delta\|=1,\max_{1\le i\le n}|\Delta_i|\ge\delta_0} \mathbb{E}f(W,X,\Delta) \gtrsim 1. \tag{A.11}$$

Combining the two cases (A.10) and (A.11), we obtain the conclusion that

$$\inf_{\|\Delta\|=1} \mathbb{E}f(W,X,\Delta) \gtrsim 1.$$

Analysis of (A.8). We now denote the conditional expectation operator $\mathbb{E}(\cdot|X)$ by $\mathbb{E}^X$. Letting $\widetilde{W}$ be an independent copy of W, we first bound the moment generating function via a standard symmetrization argument. For any $\lambda > 0$,

$$\begin{aligned}
& \mathbb{E}^X \exp\left(\lambda \sup_{\|\Delta\|=1} \left|f(W,X,\Delta) - \mathbb{E}^X f(W,X,\Delta)\right|\right) \\
\leq\ & \mathbb{E}^X \exp\left(\lambda \mathbb{E}^{X,W} \sup_{\|\Delta\|=1} \left|f(W,X,\Delta) - f(\widetilde{W},X,\Delta)\right|\right) \\
\leq\ & \mathbb{E}^X \exp\left(\lambda \sup_{\|\Delta\|=1} \left|f(W,X,\Delta) - f(\widetilde{W},X,\Delta)\right|\right) \\
=\ & \mathbb{E}^X \exp\left(\lambda \sup_{\|\Delta\|=1} \left|\frac{1}{p}\sum_{j=1}^p \varepsilon_j \left(\left|\sum_{i=1}^n \psi(W_j^T x_i)\Delta_i\right| - \left|\sum_{i=1}^n \psi(\widetilde{W}_j^T x_i)\Delta_i\right|\right)\right|\right) \\
\leq\ & \mathbb{E}^X \exp\left(2\lambda \sup_{\|\Delta\|=1} \left|\frac{1}{p}\sum_{j=1}^p \varepsilon_j \left|\sum_{i=1}^n \psi(W_j^T x_i)\Delta_i\right|\right|\right), \qquad \text{(A.12)}
\end{aligned}$$

where $\varepsilon_1, ..., \varepsilon_p$ are independent Rademacher random variables. Let us adopt the notation

$$F(\varepsilon, W, X, \Delta) = \frac{1}{p}\sum_{j=1}^p \varepsilon_j \left|\sum_{i=1}^n \psi(W_j^T x_i)\Delta_i\right|.$$

We use a discretization argument. For the Euclidean sphere $S^{n-1} = \{\Delta \in \mathbb{R}^n : \|\Delta\| = 1\}$, there exists a subset $\mathcal{N} \subset S^{n-1}$, such that for any $\Delta \in S^{n-1}$, there exists a $\Delta' \in \mathcal{N}$ that satisfies $\|\Delta - \Delta'\| \leq 1/2$, and we also have the bound $\log|\mathcal{N}| \leq 2n$. See, for example, Lemma 5.2 of Vershynin (2010). For any $\Delta \in S^{n-1}$ and the corresponding $\Delta' \in \mathcal{N}$ that satisfies $\|\Delta - \Delta'\| \leq 1/2$, we have

$$\begin{aligned}
|F(\varepsilon, W, X, \Delta)| &\leq |F(\varepsilon, W, X, \Delta')| + |F(\varepsilon, W, X, \Delta - \Delta')| \\
&\leq |F(\varepsilon, W, X, \Delta')| + \frac{1}{2}\sup_{\|\Delta\|=1} |F(\varepsilon, W, X, \Delta)|,
\end{aligned}$$

which, by taking supremum over both sides, implies

$$\sup_{\|\Delta\|=1} |F(\varepsilon, W, X, \Delta)| \leq 2\max_{\Delta\in\mathcal{N}} |F(\varepsilon, W, X, \Delta)|.$$

Define $\bar{F}(\varepsilon, X, \Delta) = \mathbb{E}^{\varepsilon,X} F(\varepsilon, W, X, \Delta)$, and then

$$\max_{\Delta\in\mathcal{N}} |F(\varepsilon, W, X, \Delta)| \leq \max_{\Delta\in\mathcal{N}} |F(\varepsilon, W, X, \Delta) - \bar{F}(\varepsilon, X, \Delta)| + \max_{\Delta\in\mathcal{N}} |\bar{F}(\varepsilon, X, \Delta)|.$$

In view of (A.12), we obtain the bound

$$\begin{aligned}
&\mathbb{E}^X \exp\Big(\lambda \sup_{\|\Delta\|=1}\Big|f(W,X,\Delta) - \mathbb{E}^X f(W,X,\Delta)\Big|\Big) \\
&\qquad \le \mathbb{E}^X \exp\left(4\lambda \max_{\Delta\in\mathcal{N}} |F(\varepsilon,W,X,\Delta) - \bar{F}(\varepsilon,X,\Delta)| + 4\lambda \max_{\Delta\in\mathcal{N}} |\bar{F}(\varepsilon,X,\Delta)|\right) \\
&\qquad \le \frac{1}{2}\sum_{\Delta\in\mathcal{N}} \mathbb{E}^X \exp\big(4\lambda|F(\varepsilon,W,X,\Delta) - \bar{F}(\varepsilon,X,\Delta)|\big) \qquad \text{(A.13)} \\
&\qquad\quad + \frac{1}{2}\sum_{\Delta\in\mathcal{N}} \mathbb{E}^X \exp\big(4\lambda|\bar{F}(\varepsilon,X,\Delta)|\big)\,. \qquad \text{(A.14)}
\end{aligned}$$

We will bound the two terms above on the event $E = \{\sum_{i=1}^n \|x_i\|^2 \le 3nd\}$. For any $W, \widetilde{W}$, we have

$$\begin{aligned}
\Big|F(\varepsilon,W,X,\Delta) - F(\varepsilon,\widetilde{W},X,\Delta)\Big| &\le \frac{1}{p}\sum_{j=1}^p\sum_{i=1}^n \Big|(\psi(W_j^T x_i) - \psi(\widetilde{W}_j^T x_i))\Delta_i\Big| \\
&\le \frac{1}{p}\sum_{j=1}^p\sum_{i=1}^n |(W_j - \widetilde{W}_j)^T x_i||\Delta_i| \\
&\le \frac{1}{p}\sum_{j=1}^p\sum_{i=1}^n \|W_j - \widetilde{W}_j\|\|x_i\||\Delta_i| \\
&\le \frac{1}{\sqrt{p}}\sqrt{\sum_{j=1}^p \|W_j - \widetilde{W}_j\|^2}\sqrt{\sum_{i=1}^n \|x_i\|^2} \\
&\le \sqrt{\frac{3n}{p}}\sqrt{\sum_{j=1}^p \|\sqrt{d}W_j - \sqrt{d}\widetilde{W}_j\|^2},
\end{aligned}$$

where the last inequality holds under the event E. By Lemma A.3, we have for any X such that E holds,

$$\mathbb{P}\left(|F(\varepsilon,W,X,\Delta) - \bar{F}(\varepsilon,X,\Delta)| > t\big|X\right) \le 2\exp\left(-\frac{pt^2}{6n}\right),$$

for any $t > 0$. The sub-Gaussian tail implies a bound for the moment generating function. By Lemma 5.5 of Vershynin (2010), we have

$$\mathbb{E}^X \exp\big(4\lambda|F(\varepsilon,W,X,\Delta) - \bar{F}(\varepsilon,X,\Delta)|\big) \le \exp\left(C_1\frac{n}{p}\lambda^2\right),$$

for some constant $C_1 > 0$. To bound the moment generating function of $\bar{F}(\varepsilon, X, \Delta)$, we note that

$$\begin{aligned}
|\bar{F}(\varepsilon, X, \Delta)| &\leq \left|\frac{1}{p}\sum_{j=1}^{p}\varepsilon_j\right| \mathbb{E}^X \left|\sum_{i=1}^{n}\psi(W^T x_i)\Delta_i\right| \\
&\leq \left|\frac{1}{p}\sum_{j=1}^{p}\varepsilon_j\right| \sqrt{\sum_{i=1}^{n}\mathbb{E}^X|\psi(W^T x_i)|^2} \\
&\leq \left|\frac{1}{p}\sum_{j=1}^{p}\varepsilon_j\right| \sqrt{\sum_{i=1}^{n}\|x_i\|^2/d} \leq \sqrt{3n}\left|\frac{1}{p}\sum_{j=1}^{p}\varepsilon_j\right|,
\end{aligned}$$

where the last inequality holds under the event E. With an application of Hoeffding-type inequality (Lemma 5.9 of Vershynin (2010)), we have

$$\mathbb{E}^X \exp\left(4\lambda|\bar{F}(\varepsilon, X, \Delta)|\right) \leq \mathbb{E}\exp\left(4\lambda\sqrt{3n}\left|\frac{1}{p}\sum_{j=1}^{p}\varepsilon_j\right|\right) \leq \exp\left(C_1\frac{n}{p}\lambda^2\right).$$

Note that we can use the same constant C_1 by making its value sufficiently large. Plug the two moment generating function bounds into (A.13) and (A.14), and we obtain the bound

$$\mathbb{E}^X \exp\left(\lambda \sup_{\|\Delta\|=1}\left|f(W, X, \Delta) - \mathbb{E}^X f(W, X, \Delta)\right|\right) \leq \exp\left(C_1\frac{n}{p}\lambda^2 + 2n\right),$$

for any X such that E holds. To bound (B.5), we apply Chernoff bound, and then

$$\mathbb{P}\left(\sup_{\|\Delta\|=1}|f(W, X, \Delta) - \mathbb{E}(f(W, X, \Delta)|X)| > t\right) \leq \exp\left(-\lambda t + C_1\frac{n}{p}\lambda^2 + 2n\right).$$

Optimize over λ, set $t \asymp \sqrt{\frac{n^2}{p}}$, and we have

$$\sup_{\|\Delta\|=1}|f(W, X, \Delta) - \mathbb{E}(f(W, X, \Delta)|X)| \lesssim \sqrt{\frac{n^2}{p}},$$

with high probability.

Analysis of (A.9). We use a discretization argument. There exists a subset $\mathcal{N}_\zeta \subset S^{n-1}$, such that for any $\Delta \in S^{n-1}$, there exists a $\Delta' \in \mathcal{N}_\zeta$ that satisfies $\|\Delta - \Delta'\| \leq \zeta$, and we also have the bound $\log|\mathcal{N}| \leq n\log(1 + 2/\zeta)$ according to Lemma 5.2 of Vershynin (2010). For any $\Delta \in S^{n-1}$ and the corresponding $\Delta' \in \mathcal{N}_\zeta$ that satisfies $\|\Delta - \Delta'\| \leq \zeta$, we have

$$\begin{aligned}
|g(X, \Delta) - \mathbb{E}g(X, \Delta)| &\leq |g(X, \Delta') - \mathbb{E}g(X, \Delta')| \\
&\quad + |g(X, \Delta - \Delta') - \mathbb{E}g(X, \Delta - \Delta')| \\
&\quad + 2\mathbb{E}g(X, \Delta - \Delta') \\
&\leq |g(X, \Delta') - \mathbb{E}g(X, \Delta')| \\
&\quad + \zeta \sup_{\|\Delta\|=1}|g(X, \Delta) - \mathbb{E}g(X, \Delta)| \\
&\quad + 2\zeta \sup_{\|\Delta\|=1}\mathbb{E}g(X, \Delta).
\end{aligned}$$

Take supremum over both sides, arrange the inequality, and we obtain the bound

$$\begin{aligned}\sup_{\|\Delta\|=1}|g(X,\Delta)-\mathbb{E}g(X,\Delta)| &\leq (1-\zeta)^{-1}\max_{\Delta\in\mathcal{N}_\zeta}|g(X,\Delta)-\mathbb{E}g(X,\Delta)| &\text{(A.15)}\\ &\quad 2\zeta(1-\zeta)^{-1}\mathbb{E}g(X,\Delta). &\text{(A.16)}\end{aligned}$$

To bound (A.15), we will use Lemma A.3 together with a union bound argument. For any $X, \widetilde{X}$, we have

$$\begin{aligned}|g(X,\Delta)-g(\widetilde{X},\Delta)| &\leq \mathbb{E}^X\left|\sum_{i=1}^n(\psi(W_j^Tx_i)-\psi(W_j^T\widetilde{x}_j))\Delta_i\right|\\ &\leq \mathbb{E}^X\sqrt{\sum_{i=1}^n\left(\psi(W_j^Tx_i)-\psi(W_j^T\widetilde{x}_j)\right)^2}\\ &\leq \sqrt{\sum_{i=1}^n\mathbb{E}^X\left(W_j^T(x_i-\widetilde{x}_i)\right)^2}\\ &= \frac{1}{\sqrt{d}}\sqrt{\sum_{i=1}^n\|x_i-\widetilde{x}_i\|^2}.\end{aligned}$$

Therefore, by Lemma A.3,

$$\mathbb{P}\left(|g(X,\Delta)-g(\widetilde{X},\Delta)|>t\right)\leq 2\exp\left(-\frac{dt^2}{2}\right),$$

for any $t>0$. A union bound argument leads to

$$\mathbb{P}\left(\max_{\Delta\in\mathcal{N}_\zeta}|g(X,\Delta)-\mathbb{E}g(X,\Delta)|>t\right)\leq 2\exp\left(-\frac{dt^2}{2}+n\log\left(1+\frac{2}{\zeta}\right)\right),$$

which implies that

$$\max_{\Delta\in\mathcal{N}_\zeta}|g(X,\Delta)-\mathbb{E}g(X,\Delta)|\lesssim\sqrt{\frac{n\log(1+2/\zeta)}{d}},$$

with high probability. For (A.16), we have

$$\mathbb{E}g(X,\Delta)\leq\sqrt{\mathbb{E}\mathsf{Var}\left(\sum_{i=1}^n\psi(W^Tx_i)\Delta_i\Big|W\right)}\leq\sqrt{\mathbb{E}|\psi(W^Tx)|^2}\leq 1.$$

Combining the bounds for (A.15) and (A.16), we have

$$\sup_{\|\Delta\|=1}|g(X,\Delta)-\mathbb{E}g(X,\Delta)|\lesssim\sqrt{\frac{n\log(1+2/\zeta)}{d}}+\zeta,$$

with high probability as long as $\zeta\leq 1/2$. We choose $\zeta=\sqrt{n/d}$, and thus the bound is sufficiently small as long as n/d is sufficiently small.

Finally, combine results for (A.7), (A.8) and (A.9), and we obtain the desired conclusion as long as n^2/p and n/d are sufficiently small. □

To prove (3.4) of Lemma 4.2, we establish the following stronger result.

Lemma A.7. *Consider independent $W_1, \dots, W_p \sim N(0, d^{-1}I_d)$ and $x_1, \dots, x_n \sim N(0, I_d)$. We define the matrices $G, \bar{G} \in \mathbb{R}^{n\times n}$ by*

$$\begin{aligned} G_{il} &= \frac{1}{p}\sum_{j=1}^{p}\psi(W_j^T x_i)\psi(W_j^T x_l), \\ \bar{G}_{il} &= |\mathbb{E}\psi'(Z)|^2 \frac{x_i^T x_l}{\|x_i\|\|x_l\|} + \left(\mathbb{E}|\psi(Z)|^2 - |\mathbb{E}\psi'(Z)|^2\right)\mathbb{I}\{i = l\}, \end{aligned}$$

where $Z \sim N(0,1)$. Assume $d/\log n$ is sufficiently large, and then

$$\|G - \bar{G}\|_{\text{op}}^2 \lesssim \frac{n^2}{p} + \frac{\log n}{d} + \frac{n^2}{d^2},$$

with high probability. Therefore, if we further assume n^2/p and n/d are sufficiently small, we also have

$$1 \lesssim \lambda_{\min}(G) \leq \lambda_{\max}(G) \lesssim 1, \tag{A.17}$$

with high probability.

Proof. Define $\widetilde{G} \in \mathbb{R}^{n\times n}$ with entries $\widetilde{G}_{il} = \mathbb{E}\left(\psi(W^T x_i)\psi(W^T x_l)|X\right)$, and we first bound the difference between G and $\widetilde{G}$. Note that

$$\mathbb{E}(G_{il} - \widetilde{G}_{il})^2 = \mathbb{E}\mathsf{Var}(G_{il}|X) \leq \frac{1}{p}\mathbb{E}|\psi(W^T x_i)\psi(W^T x_l)|^2 \leq p^{-1}.$$

We then have

$$\mathbb{E}\|G - \widetilde{G}\|_{\text{op}}^2 \leq \mathbb{E}\|G - \widetilde{G}\|_{\text{F}}^2 \leq \frac{n^2}{p}.$$

By Markov's inequality,

$$\|G - \widetilde{G}\|_{\text{op}}^2 \lesssim \frac{n^2}{p}, \tag{A.18}$$

with high probability.

Next, we study the diagonal entries of $\widetilde{G}$. For any $i \in [n]$,

$$\widetilde{G}_{ii} = \mathbb{E}(|\psi(W^T x_i)|^2|X) = \mathbb{E}_{U\sim N(0,\|x_i\|^2/d)}|\psi(U)|^2.$$

Therefore,

$$\max_{1\leq i\leq n}|\widetilde{G}_{ii} - \bar{G}_{ii}| \leq \max_{1\leq i\leq n}\mathsf{TV}\left(N(0,\|x_i\|^2/d), N(0,1)\right) \leq \frac{3}{2}\max_{1\leq i\leq n}\left|\frac{\|x_i\|^2}{d} - 1\right|.$$

By Lemma A.4 and a union bound argument, we have

$$\max_{1\leq i\leq n}|\widetilde{G}_{ii} - \bar{G}_{ii}| \lesssim \sqrt{\frac{\log n}{d}}, \tag{A.19}$$

with high probability.

Now we analyze the off-diagonal entries. We use the notation $\bar{x}_i = \frac{\sqrt{d}}{\|x_i\|}x_i$. For any $i \neq l$, we have

$$\begin{aligned}\widetilde{G}_{il} &= \mathbb{E}\left(\psi(W^T\bar{x}_i)\psi(W^T\bar{x}_l)|X\right) &\text{(A.20)}\\ &+\mathbb{E}\left((\psi(W^Tx_i)-\psi(W^T\bar{x}_i))\psi(W^T\bar{x}_l)|X\right) &\text{(A.21)}\\ &+\mathbb{E}\left(\psi(W^T\bar{x}_i)(\psi(W^Tx_l)-\psi(W^T\bar{x}_l))|X\right) &\text{(A.22)}\\ &+\mathbb{E}\left((\psi(W^Tx_i)-\psi(W^T\bar{x}_i))(\psi(W^Tx_l)-\psi(W^T\bar{x}_l))|X\right). &\text{(A.23)}\end{aligned}$$

For first term on the right hand side of (A.20), we observe that $\mathbb{E}\left(\psi(W^T\bar{x}_i)\psi(W^T\bar{x}_l)|X\right)$ is a function of $\frac{\bar{x}_i^T\bar{x}_l}{d}$, and thus we can write

$$\mathbb{E}\left(\psi(W^T\bar{x}_i)\psi(W^T\bar{x}_l)|X\right) = f\left(\frac{\bar{x}_i^T\bar{x}_l}{d}\right),$$

where

$$f(\rho) = \begin{cases}\mathbb{E}\psi(\sqrt{1-\rho}U+\sqrt{\rho}Z)\psi(\sqrt{1-\rho}V+\sqrt{\rho}Z), & \rho \geq 0,\\ \mathbb{E}\psi(\sqrt{1+\rho}U-\sqrt{-\rho}Z)\psi(\sqrt{1+\rho}V+\sqrt{-\rho}Z), & \rho < 0,\end{cases}$$

with $U, V, Z \overset{iid}{\sim} N(0,1)$. By some direct calculations, we have $f(0) = 0$, $f'(0) = (\mathbb{E}\psi'(Z))^2$, and $\sup_{|\rho|\leq 0.2}|f''(\rho)| \lesssim 1$. Therefore, as long as $|\bar{x}_i^T\bar{x}_l|/d \leq 1/5$,

$$\left|f\left(\frac{\bar{x}_i^T\bar{x}_l}{d}\right) - (\mathbb{E}\psi'(Z))^2\frac{\bar{x}_i^T\bar{x}_l}{d}\right| \leq C_1\left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|^2,$$

for some constant $C_1 > 0$. By Lemma A.5, we know that $\max_{i\neq l}|\bar{x}_i^T\bar{x}_l|/d \lesssim \sqrt{\frac{\log n}{d}} \leq 1/5$ with high probability, which then implies

$$\sum_{i\neq l}\left(\mathbb{E}\left(\psi(W^T\bar{x}_i)\psi(W^T\bar{x}_l)|X\right) - \bar{G}_{il}\right)^2 \leq C_1\sum_{i\neq l}\left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|^4. \tag{A.24}$$

The term on the right hand side can be bounded by

$$\sum_{i\neq l}\left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|^4 \leq \frac{d}{\min_{1\leq i\leq n}\|x_i\|^2}\sum_{i\neq l}\left|\frac{x_i^Tx_l}{d}\right|^4.$$

By Lemma A.4, $\frac{d}{\min_{1\leq i\leq n}\|x_i\|^2} \lesssim 1$ with high probability. By integrating out the probability tail bound of $|x_i^Tx_l|$ given in Lemma A.5, we have $\sum_{i\neq l}\mathbb{E}\left|\frac{x_i^Tx_l}{d}\right|^4 \lesssim \frac{n^2}{d^2}$, and by Markov's inequality, we have $\sum_{i\neq l}\left|\frac{x_i^Tx_l}{d}\right|^4 \lesssim \frac{n^2}{d^2}$ with high probability.

We also need to analyze the contributions of (A.21) and (A.22). We can write (A.21) as

$$\begin{aligned}&\mathbb{E}\left[\psi(W^T\bar{x}_l)\psi'(W^T\bar{x}_i)W^T(x_i-\bar{x}_i)|X\right] &\text{(A.25)}\\ &+\frac{1}{2}\mathbb{E}\left[\psi(W^T\bar{x}_i)\psi''(t_i)|W^T(x_i-\bar{x}_i)|^2|X\right], &\text{(A.26)}\end{aligned}$$

where t_i is some random variable between $W^T x_i$ and $W^T \bar{x}_i$. The first term (A.25) can be expressed as

$$\left(\frac{\|x_i\|}{\sqrt{d}} - 1\right) \mathbb{E}\left[\psi(W^T\bar{x}_l)\psi'(W^T\bar{x}_i)W^T\bar{x}_i|X\right] = \left(\frac{\|x_i\|}{\sqrt{d}} - 1\right) g\left(\frac{\bar{x}_i^T\bar{x}_l}{d}\right),$$

where the function g satisfies $g(0) = 0$ and $\sup_{|\rho|\leq 0.2} |g'(\rho)| \lesssim 1$, and thus

$$\left|g\left(\frac{\bar{x}_i^T\bar{x}_l}{d}\right)\right| \lesssim \left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right| \lesssim \left|\frac{x_i^T x_l}{d}\right|,$$

because of the high probability bound $\max_{i\neq l} |\bar{x}_i^T\bar{x}_l|/d \lesssim \sqrt{\frac{\log n}{d}} \leq 1/5$. Therefore,

$$\begin{aligned}
& \sum_{i\neq l} \left(\mathbb{E}\left[\psi(W^T\bar{x}_l)\psi'(W^T\bar{x}_i)W^T(x_i - \bar{x}_i)|X\right]\right)^2 \\
\lesssim & \sum_{i\neq l} \left|\frac{\|x_i\|}{\sqrt{d}} - 1\right|^2 \left|\frac{x_i^T x_l}{d}\right|^2 \\
\lesssim & \; n\sum_{i=1}^{n} \left|\frac{\|x_i\|}{\sqrt{d}} - 1\right|^4 + \sum_{i\neq l} \left|\frac{x_i^T x_l}{d}\right|^4. \qquad \text{(A.27)}
\end{aligned}$$

By integrating out the probability tail bound of Lemma A.4, we have $\mathbb{E}\left|\frac{\|x_i\|}{\sqrt{d}} - 1\right|^4 \lesssim d^{-2}$. We also have $\mathbb{E}\left|\frac{x_i^T x_l}{d}\right|^4 \lesssim d^{-2}$. Hence, $\sum_{i\neq l} \left(\mathbb{E}\left[\psi(W^T\bar{x}_l)\psi'(W^T\bar{x}_i)W^T(x_i - \bar{x}_i)|X\right]\right)^2 \lesssim \frac{n^2}{d^2}$ with high probability. To bound (A.26), we observe that

$$\frac{1}{2}\mathbb{E}\left[\psi(W^T\bar{x}_i)\psi''(t_i)|W^T(x_i - \bar{x}_i)|^2|X\right] \leq \mathbb{E}(|W^T(x_i - \bar{x}_i)|^2|X) = \left|\frac{\|x_i\|}{\sqrt{d}} - 1\right|^2,$$

where the inequality above is by $\sup_x |\psi(x)| \leq 1$ and $\sup_x |\psi''(x)| \leq 2$. Since $\mathbb{E}\left|\frac{\|x_i\|}{\sqrt{d}} - 1\right|^4 \lesssim d^{-2}$, we then have

$$\sum_{i\neq l} \left(\frac{1}{2}\mathbb{E}\left[\psi(W^T\bar{x}_i)\psi''(t_i)|W^T(x_i - \bar{x}_i)|^2|X\right]\right)^2 \lesssim \frac{n^2}{d^2},$$

with high probability. With a similar analysis of (A.22), we conclude that the contributions of (A.21) and (A.22) is at most at the order of $\frac{n^2}{d^2}$ with respect to the squared Frobenius norm.

Finally, we show that the contribution of (A.23) is negligible. Note that

$$\begin{aligned}
& \left|\mathbb{E}\left((\psi(W^T x_i) - \psi(W^T\bar{x}_i))(\psi(W^T x_l) - \psi(W^T\bar{x}_l))|X\right)\right| \\
\leq & \left|\frac{\|x_i\|}{\sqrt{d}} - 1\right|\left|\frac{\|x_l\|}{\sqrt{d}} - 1\right| \mathbb{E}\left(|W^T\bar{x}_i||W^T\bar{x}_l||X\right) \\
\leq & \left|\frac{\|x_i\|}{\sqrt{d}} - 1\right|\left|\frac{\|x_l\|}{\sqrt{d}} - 1\right|,
\end{aligned}$$

where the last inequality is by $\mathbb{E}\left(|W^T\bar{x}_i||W^T\bar{x}_l||X\right) \leq \frac{1}{2}\mathbb{E}(|W^T\bar{x}_i|^2 + |W^T\bar{x}_l|^2|X) = 1$. Since

$$\sum_{i\neq l} \mathbb{E}\left(\frac{\|x_i\|}{\sqrt{d}} - 1\right)^2 \mathbb{E}\left(\frac{\|x_l\|}{\sqrt{d}} - 1\right)^2 \lesssim \frac{n^2}{d^2},$$

we can conclude that (A.23) is bounded by $O\left(\frac{n^2}{d^2}\right)$ with high probability by Markov's inequality.

Combining the analyses of (A.20), (A.21), (A.22) and (A.23), we conclude that $\sum_{i\neq l}(\widetilde{G}_{il} - \bar{G}_{il})^2 \lesssim \frac{n^2}{d^2}$ with high probability. Together with (A.18) and (A.19), we obtain the desired bound for $\|G - \bar{G}\|_{\text{op}}$. For the last conclusion, since $\|G-\bar{G}\|_{\text{op}}$ is sufficiently small, it is sufficient to show $1 \lesssim \lambda_{\min}(\bar{G}) \leq \lambda_{\max}(\bar{G}) \lesssim 1$. The bound $1 \lesssim \lambda_{\min}(\bar{G})$ is a direct consequence of the definition of $\bar{G}$. To upper bound $\lambda_{\max}(\bar{G})$, we have

$$\begin{aligned} \lambda_{\max}(\bar{G}) &\lesssim 1 + \max_{\|v\|=1} \sum_{i=1}^n \sum_{l=1}^n v_i v_l \frac{x_i^T x_l}{\|x_i\|\|x_l\|} \\ &\lesssim 1 + \max_{\|v\|=1} \sum_{i=1}^n \sum_{l=1}^n v_i v_l \frac{x_i^T x_l}{d} \\ &= 1 + \|X\|_{\text{op}}^2/d \\ &\lesssim 1 + \frac{n}{d}, \end{aligned}$$

with high probability, where the last inequality is by Davidson and Szarek (2001). The proof is complete. □

Proof of Corollary 4.2. Since $\widehat{\theta}$ belongs to the row space of $\widetilde{X}$, there exists some $u^* \in \mathbb{R}^n$ such that $\widehat{\theta} = \widetilde{X}^T u^*$. By Theorem 3.1 and Lemma 4.2, we know that $\widetilde{u} = u^*$ with high probability, and therefore $\widetilde{\theta} = \widetilde{X}^T \widetilde{u} = \widetilde{X}^T u^* = \widehat{\theta}$. □

A.4. *Proof of Theorem 5.1*

To prove Theorem 5.1, we need the following kernel random matrix result.

Lemma A.8. *Consider independent $W_1, \ldots, W_p \sim N(0, d^{-1}I_d)$, $x_1, \ldots, x_n \sim N(0, I_d)$, and parameters $\beta_1, \ldots, \beta_p \sim N(0,1)$. We define the matrices $H, \bar{H} \in \mathbb{R}^{n\times n}$ by*

$$\begin{aligned} H_{il} &= \frac{x_i^T x_l}{d} \frac{1}{p} \sum_{j=1}^p \beta_j^2 \psi'(W_j^T x_i)\psi'(W_j^T x_l), \\ \bar{H}_{il} &= |\mathbb{E}\psi'(Z)|^2 \frac{x_i^T x_l}{\|x_i\|\|x_l\|} + \left(\mathbb{E}|\psi'(Z)|^2 - |\mathbb{E}\psi'(Z)|^2\right) \mathbb{I}\{i = l\}, \end{aligned}$$

where $Z \sim N(0,1)$. Assume $d/\log n$ is sufficiently large, and then

$$\|H - \bar{H}\|_{\text{op}}^2 \lesssim \frac{n^2}{pd} + \frac{n}{p} + \frac{\log n}{d} + \frac{n^2}{d^2},$$

with high probability. If we further assume that d/n and p/n are sufficiently large, we will also have

$$0.09 \leq \lambda_{\min}(H) \leq \lambda_{\max}(H) \lesssim 1, \tag{A.28}$$

with high probability.

Proof. Define $\widetilde{H} \in \mathbb{R}^{n\times n}$ with entries $\widetilde{H}_{il} = \frac{x_i^T x_l}{d}\mathbb{E}\left(\psi'(W^T x_i)\psi'(W^T x_l)\middle|X\right)$, and we first bound the difference between H and $\widetilde{H}$. Note that

$$\mathbb{E}(H_{il} - \widetilde{H}_{il})^2 = \mathbb{E}\mathsf{Var}(H_{il}|X) \leq \frac{1}{p}\mathbb{E}\left(\frac{|x_i^T x_l|^2}{d^2}\beta^4\right) \leq \begin{cases} \frac{3}{pd}, & i \neq l, \\ 9p^{-1}, & i = l. \end{cases}$$

We then have

$$\mathbb{E}\|H - \widetilde{H}\|_{\mathrm{op}}^2 \leq \mathbb{E}\|H - \widetilde{H}\|_{\mathrm{F}}^2 \leq \frac{3n^2}{pd} + \frac{9n}{p}.$$

By Markov's inequality,

$$\|H - \widetilde{H}\|_{\mathrm{op}}^2 \lesssim \frac{n^2}{pd} + \frac{n}{p}, \tag{A.29}$$

with high probability.

Next, we study the diagonal entries of $\widetilde{H}$. For any $i \in [n]$,

$$\widetilde{H}_{ii} = \frac{\|x_i\|^2}{d}\mathbb{E}(|\psi'(W^T x_i)|^2|X) = \frac{\|x_i\|^2}{d}\mathbb{E}_{U\sim N(0,\|x_i\|^2/d)}|\psi'(U)|^2.$$

Since $\sup_x |\psi'(x)| \leq 1$ and $\sup_x |\psi''(x)| \leq 2$, we have

$$\begin{aligned} |\widetilde{H}_{ii} - \bar{H}_{ii}| &\leq \left|\frac{\|x_i\|^2}{d} - 1\right| + \left|\mathbb{E}_{U\sim N(0,\|x_i\|^2/d)}|\psi'(U)|^2 - \mathbb{E}_{U\sim N(0,1)}|\psi'(U)|^2\right| \\ &\leq \left|\frac{\|x_i\|^2}{d} - 1\right| + 2\mathsf{TV}\left(N(0,\|x_i\|^2/d), N(0,1)\right) \\ &\leq 4\left|\frac{\|x_i\|^2}{d} - 1\right| \end{aligned}$$

Similar to (A.19), Lemma A.4 and a union bound argument imply

$$\max_{1\leq i\leq n} |\widetilde{H}_{ii} - \bar{H}_{ii}| \lesssim \sqrt{\frac{\log n}{d}}, \tag{A.30}$$

with high probability.

Now we analyze the off-diagonal entries. Recall the notation $\bar{x}_i = \frac{\sqrt{d}}{\|x_i\|}x_i$. For any $i \neq l$, we have

$$\begin{aligned} \widetilde{H}_{il} &= \frac{\bar{x}_i^T \bar{x}_l}{d}\mathbb{E}\left(\psi'(W^T\bar{x}_i)\psi'(W^T\bar{x}_l)\middle|X\right) && \text{(A.31)} \\ &\quad + \frac{x_i^T x_l}{d}\mathbb{E}\left(\psi'(W^T x_i)\psi'(W^T x_l) - \psi'(W^T\bar{x}_i)\psi'(W^T\bar{x}_l)\middle|X\right) && \text{(A.32)} \\ &\quad + \left(\frac{\|x_i\|\|x_l\|}{d} - 1\right)\frac{\bar{x}_i^T\bar{x}_l}{d}\mathbb{E}\left(\psi'(W^T\bar{x}_i)\psi'(W^T\bar{x}_l)\middle|X\right). && \text{(A.33)} \end{aligned}$$

For the first term on the right hand side of (A.31), we observe that $\frac{\bar{x}_i^T\bar{x}_l}{d}\mathbb{E}\left(\psi'(W^T\bar{x}_i)\psi'(W^T\bar{x}_l)\middle|X\right)$ is a function of $\frac{\bar{x}_i^T\bar{x}_l}{d}$, and thus we can write

$$\frac{\bar{x}_i^T\bar{x}_l}{d}\mathbb{E}\left(\psi'(W^T\bar{x}_i)\psi'(W^T\bar{x}_l)\middle|X\right) = f\left(\frac{\bar{x}_i^T\bar{x}_l}{d}\right),$$

where

$$f(\rho) = \begin{cases} \rho \mathbb{E}\psi'(\sqrt{1-\rho}U + \sqrt{\rho}Z)\psi'(\sqrt{1-\rho}V + \sqrt{\rho}Z), & \rho \geq 0, \\ \rho \mathbb{E}\psi'(\sqrt{1+\rho}U - \sqrt{-\rho}Z)\psi'(\sqrt{1+\rho}V + \sqrt{-\rho}Z), & \rho < 0, \end{cases}$$

with $U, V, Z \overset{iid}{\sim} N(0,1)$. By some direct calculations, we have $f(0) = 0$, $f'(0) = (\mathbb{E}\psi'(Z))^2$, and $\sup_{|\rho|\leq 0.2} |f''(\rho)| \lesssim 1$. Therefore, using the same analysis that leads to the bound for (A.24), we have

$$\sum_{i\neq l} \left(\frac{\bar{x}_i^T \bar{x}_l}{d} \mathbb{E}\left(\psi'(W^T\bar{x}_i)\psi'(W^T\bar{x}_l)\big|X\right) - \bar{H}_{il} \right)^2 \lesssim \sum_{i\neq l} \left| \frac{\bar{x}_i^T \bar{x}_l}{d} \right|^4 \lesssim \frac{n^2}{d^2},$$

with high probability.

For (A.32), we note that

$$\begin{aligned} \mathbb{E}\Big(\psi'(W^T x_i)\psi'(W^T x_l) - \psi'(W^T\bar{x}_i)\psi'(W^T\bar{x}_l)\big|X\Big) \\ &\leq \mathbb{E}\left(|\psi'(W^T x_i) - \psi'(W^T\bar{x}_i)|\big|X\right) + \mathbb{E}\left(|\psi'(W^T x_l) - \psi'(W^T\bar{x}_l)|\big|X\right) \\ &\leq 2\mathbb{E}\left(|W^T(x_i - \bar{x}_i)|\big|X\right) + 2\mathbb{E}\left(|W^T(x_l - \bar{x}_l)|\big|X\right) \\ &= 2\left|\frac{\|x_i\|}{\sqrt{d}} - 1\right| + 2\left|\frac{\|x_l\|}{\sqrt{d}} - 1\right|, \end{aligned}$$

where we have used $\sup_x |\psi'(x)| \leq 1$ and $\sup_x |\psi''(x)| \leq 2$ in the above inequalities. Therefore, the contribution of (A.32) in terms of squared Frobenius norm is bounded by

$$\begin{aligned} & \sum_{i\neq l} \left|\frac{x_i^T x_l}{d}\right|^2 \left(2\left|\frac{\|x_i\|}{\sqrt{d}} - 1\right| + 2\left|\frac{\|x_l\|}{\sqrt{d}} - 1\right|\right)^2 \\ \lesssim\ & \sum_{i\neq l} \left|\frac{x_i^T x_l}{d}\right|^4 + n\sum_{i=1}^{n} \left|\frac{\|x_l\|}{\sqrt{d}} - 1\right|^4 \\ \lesssim\ & \frac{n^2}{d^2}, \end{aligned}$$

with high probability, and the last inequality above uses the same analysis that bounds (A.27).

Finally, since (A.33) can be bounded by $\left|\frac{\|x_i\|\|x_l\|}{d} - 1\right| \left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|$, its contribution in terms of squared Frobenius norm is bounded by

$$\begin{aligned} & \sum_{i\neq l} \left|\frac{\|x_i\|\|x_l\|}{d} - 1\right|^2 \left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|^2 \\ \lesssim\ & \sum_{i\neq l} \left|\frac{\|x_i\|\|x_l\|}{d} - 1\right|^4 + \sum_{i\neq l} \left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|^4. \end{aligned}$$

We have already shown that $\sum_{i\neq l} \left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|^4 \lesssim \frac{n^2}{d^2}$ in the analysis of (A.24). For the first term on the right hand side of the above inequality, we use Lemma A.5 and obtain a probability tail bound for

$|\|x_i\|\|x_l\| - d|$. By integrating out this tail bound, we have

$$\sum_{i\neq l} \mathbb{E}\left(\frac{\|x_i\|\|x_l\|}{d} - 1\right)^4 \lesssim \frac{n^2}{d^2},$$

which, by Markov's inequality, implies $\sum_{i\neq l}\left(\frac{\|x_i\|\|x_l\|}{d} - 1\right)^4 \lesssim \frac{n^2}{d^2}$ with high probability.

Combining the analyses of (A.31), (A.32), and (A.33), we conclude that $\sum_{i\neq l}(\widetilde{H}_{il} - \bar{H}_{il})^2 \lesssim \frac{n^2}{d^2}$ with high probability. Together with (A.29) and (A.30), we obtain the desired bound for $\|H - \bar{H}\|_{\text{op}}$. The last conclusion (A.28) follows a similar argument used in the proof of Lemma A.7. □

Now we are ready to prove Theorem 5.1.

Proof of Theorem 5.1. We first establish some high probability events:

$$\begin{aligned}
\max_{1\leq j\leq p} |\beta_j(0)| &\leq 2\sqrt{\log p}, && \text{(A.34)}\\
\max_{k\in\{1,2,3\}} \frac{1}{p}\sum_{j=1}^{p} |\beta_j(0)|^k &\lesssim 1, && \text{(A.35)}\\
\sum_{i=1}^{n} \|x_i\|^2 &\leq 2nd, && \text{(A.36)}\\
\max_{1\leq i\leq n} \|x_i\| &\lesssim \sqrt{d}, && \text{(A.37)}\\
\max_{1\leq i\neq l\leq n} \left|\frac{x_i^T x_l}{d}\right| &\lesssim d^{-1/2}, && \text{(A.38)}\\
\max_{1\leq l\leq n} \sum_{i=1}^{n} \left|\frac{x_i^T x_l}{d}\right| &\lesssim 1 + \frac{n}{\sqrt{d}}, && \text{(A.39)}\\
\|u(0)\| &\leq \sqrt{n}(\log p)^{1/4}, && \text{(A.40)}\\
\max_{1\leq j\leq p} \sum_{i=1}^{n} |W_j(0)^T x_i|^2 &\leq 6n + 18\log p, && \text{(A.41)}\\
\max_{1\leq i\leq n} \frac{1}{p}\sum_{j=1}^{p} \mathbb{I}\{|W_j(0)^T x_i| \leq R_1\|x_i\|\} &\lesssim \sqrt{d}R_1 + \sqrt{\frac{\log n}{p}}. && \text{(A.42)}
\end{aligned}$$

The bound (A.34) is a consequence of a standard Gaussian tail inequality and a union bound argument. The second bound (A.35) is by Markov's inequality and the fact that $\frac{1}{p}\sum_{j=1}^{p} \mathbb{E}|\beta_j(0)|^k \lesssim 1$. Then, we have (A.36), (A.37) and (A.41) derived from Lemma A.4 and a union bound. Similarly, (A.38) is by Lemma A.5 and a union bound. The bound (A.39) is a direct consequence of (A.37) and (A.38). To obtain (A.40), we note that $\mathbb{E}|u_i(0)|^2 = \mathbb{E}\mathsf{Var}(u_i(0)|X) \lesssim 1$, which then implies

(A.40) by Markov's inequality. Finally, for (A.42), we have

$$\max_{1\le i\le n}\frac{1}{p}\sum_{j=1}^{p}\mathbb{I}\{|W_j(0)^Tx_i|\le R_1\|x_i\|\}\le\mathbb{P}\left(|N(0,1)|\le\sqrt{d}R_1\right)$$
$$+\max_{1\le i\le n}\frac{1}{p}\sum_{j=1}^{p}\left(\mathbb{I}\{|W_j(0)^Tx_i|\le R_1\|x_i\|\}-\mathbb{P}\left(|N(0,1)|\le\sqrt{d}R_1\right)\right),$$

where the first term $\mathbb{P}\left(|N(0,1)|\le\sqrt{d}R_1\right)$ can be bounded by $O(\sqrt{d}R_1)$, and the second term can be bounded by $\sqrt{\frac{\log n}{p}}$ according to Lemma A.1 and a union bound.

Now we are ready to prove the main result. We introduce the function

$$v_i(t)=\frac{1}{\sqrt{p}}\sum_{j=1}^{p}\beta_j(t)\psi(W_j(t-1)^Tx_i).$$

Besides (5.2), (5.3) and (5.1), we will also establish

$$\|y-v(t)\|^2\le\left(1-\frac{\gamma}{8}\right)^t\|y-v(0)\|^2. \tag{A.43}$$

It suffices to show the following to claims are true.

Claim A. With high probability, for any integer $k\ge 1$, as long as (A.43), (5.1), (5.2) and (5.3) hold for all $t\le k$, then (5.3) holds for $t=k+1$.

Claim B. With high probability, for any integer $k\ge 1$, as long as (A.43), (5.1) and (5.2) hold for all $t\le k$, and (5.3) holds for all $t\le k+1$, then (A.43) holds for $t=k+1$.

Claim C. With high probability, for any integer $k\ge 1$, as long as (5.1) and (5.2) hold for all $t\le k$, and (5.3) and (A.43) hold for all $t\le k+1$, then (5.2) holds for $t=k+1$.

Claim D. With high probability, for any integer $k\ge 1$, as long as (5.1) holds for all $t\le k$, and (A.43), (5.2) and (5.3) hold for all $t\le k+1$, then (5.1) holds for $t=k+1$.

With all the claims above being true, we can then deduce (5.2), (5.3), (5.1) and (A.43) for all $t\ge 1$ by mathematical induction.

Proof of Claim A. By triangle inequality and the gradient formula,

$$\begin{aligned}
|\beta_j(k+1)-\beta_j(0)| &\leq \sum_{t=0}^{k}|\beta_j(t+1)-\beta_j(t)| \\
&\leq \frac{\gamma}{\sqrt{p}}\sum_{t=0}^{k}\left|\sum_{i=1}^{n}(u_i(t)-y_i)\psi(W_j(t)^Tx_i)\right| \\
&\leq \frac{\gamma}{\sqrt{p}}\sum_{t=0}^{k}\sum_{i=1}^{n}|y_i-u_i(t)||W_j(t)^Tx_i| \\
&\leq \frac{\gamma}{\sqrt{p}}\sum_{t=0}^{k}\|y-u(t)\|\sqrt{\sum_{i=1}^{n}|W_j(t)^Tx_i|^2} \\
&\leq \frac{\gamma}{\sqrt{p}}\sum_{t=0}^{k}\|y-u(t)\|\left(R_1\sqrt{\sum_{i=1}^{n}\|x_i\|^2}+\sqrt{\sum_{i=1}^{n}|W_j(0)^Tx_i|^2}\right) \\
&\leq \gamma\sqrt{\frac{7n+18\log p}{p}}\sum_{t=0}^{k}\|y-u(t)\| \\
&\leq 16\sqrt{\frac{7n+18\log p}{p}}\|y-u(0)\| \\
&\leq 32\sqrt{\frac{n^2\log p}{p}}=R_2,
\end{aligned}$$

where we have used (A.37), (A.40) and (A.41). Hence, (5.3) holds for $t=k+1$, and Claim A is true.

Proof of Claim B. We omit this step, because the analysis uses the same argument as that of the proof of Claim D.

Proof of Claim C. We bound $\|W_j(k+1) - W_j(0)\|$ by $\sum_{t=0}^{k} \|W_j(t+1) - W_j(t)\|$. Then by the gradient descent formula, we have

$$\begin{aligned}
\|W_j(k+1) - W_j(0)\| &\leq \frac{\gamma}{d\sqrt{p}} \sum_{t=0}^{k} \left\| \beta_j(t+1) \sum_{i=1}^{n} (v_i(t+1) - y_i)\psi'(W_j(t)^T x_i) x_i \right\| \\
&\leq \frac{\gamma}{d\sqrt{p}} \sum_{t=0}^{k} |\beta_j(t+1)| \sum_{i=1}^{n} |y_i - v_i(t+1)| \|x_i\| \\
&\leq \frac{\gamma}{d\sqrt{p}} (|\beta_j(0)| + R_2) \sqrt{\sum_{i=1}^{n} \|x_i\|^2} \sum_{t=0}^{k} \|y - v(t+1)\| \\
&\leq \frac{16}{d\sqrt{p}} (|\beta_j(0)| + R_2) \sqrt{\sum_{i=1}^{n} \|x_i\|^2} \|y - v(0)\| \\
&\leq \frac{100 n \log p}{\sqrt{pd}} = R_1,
\end{aligned}$$

where we have used (A.34), (A.36) and (A.40) in the above inequalities. Thus, Claim C is true.

Proof of Claim D. We first analyze $u(k+1) - u(k)$. For each $i \in [n]$, we have

$$\begin{aligned}
& u_i(k+1) - u_i(k) \\
= \ & \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \beta_j(k+1) \left(\psi(W_j(k+1)^T x_i) - \psi(W_j(k)^T x_i) \right) \\
& + \frac{1}{\sqrt{p}} \sum_{j=1}^{p} (\beta_j(k+1) - \beta_j(k)) \psi(W_j(k)^T x_i) \\
= \ & \frac{1}{\sqrt{p}} \sum_{j=1}^{p} \beta_j(k+1) (W_j(k+1) - W_j(k))^T x_i \psi'(W_j(k)^T x_i) \\
& + \frac{1}{\sqrt{p}} \sum_{j=1}^{p} (\beta_j(k+1) - \beta_j(k)) \psi(W_j(k)^T x_i) + r_i(k) \\
= \ & \gamma \sum_{l=1}^{n} (H_{il}(k) + G_{il}(k))(y_l - u_l(k)) + r_i(k),
\end{aligned}$$

where

$$\begin{aligned}
G_{il}(k) &= \frac{1}{p} \sum_{j=1}^{p} \psi(W_j(k)^T x_l) \psi(W_j(k)^T x_i), \\
H_{il}(k) &= \frac{x_i^T x_l}{d} \frac{1}{p} \sum_{j=1}^{p} \beta_j(k+1)^2 \psi'(W_j(k)^T x_i) \psi'(W_j(k)^T x_l),
\end{aligned}$$

and

$$r_i(k) = \frac{1}{2\sqrt{p}} \sum_{j=1}^{p} \beta_j(k+1)|(W_j(k+1) - W_j(k))^T x_i|^2 \psi''(\xi_{ijk}).$$

Note that ξ_{ijk} is some random variable whose value is between $W_j(k)^T x_i$ and $W_j(k+1)^T x_i$. The above iteration formula can be summarized in a vector form as

$$u(k+1) - u(k) = \gamma(H(k) + G(k))(y - u(k)) + r(k). \tag{A.44}$$

We need to understand the eigenvalues of $G(k)$ and $H(k)$, and bound the absolute value of $r_i(k)$.

To analyze $G(k)$, we first control the difference between $G(k)$ and $G(0)$. Since

$$\begin{aligned} |G_{il}(k) - G_{il}(0)| &\leq \frac{1}{p} \sum_{j=1}^{p} |\psi(W_j(k)^T x_l) - \psi(W_j(0)^T x_l)| \\ &\quad + \frac{1}{p} \sum_{j=1}^{p} |\psi(W_j(k)^T x_i) - \psi(W_j(0)^T x_i)| \\ &\leq \frac{1}{p} \sum_{j=1}^{p} |(W_j(k) - W_j(0))^T x_l| + \frac{1}{p} \sum_{j=1}^{p} |(W_j(k) - W_j(0))^T x_i| \\ &\leq R_1 \left(\|x_l\| + \|x_i\|\right), \end{aligned}$$

then, by (A.37),

$$\|G(k) - G(0)\|_{\text{op}} \leq \max_{1\leq l\leq n} \sum_{i=1}^{n} |G_{il}(k) - G_{il}(0)| \leq 2R_1 n \max_{1\leq i\leq n} \|x_i\| \lesssim \frac{n^2 \log p}{\sqrt{p}}. \tag{A.45}$$

By Lemma A.7, we have

$$0 \leq \lambda_{\min}(G(k)) \leq \lambda_{\max}(G(k)) \lesssim 1 + \frac{n^2 \log p}{\sqrt{p}}. \tag{A.46}$$

For the matrix $H(k)$, we show its eigenvalues can be controlled by those of $H(0)$. We have

$$\begin{aligned} |H_{il}(k) - H_{il}(0)| &\leq \left|\frac{x_i^T x_l}{d}\right| \frac{1}{p} \sum_{j=1}^{p} |\beta_j(k+1)^2 - \beta_j^2(0)| \\ &\quad + \left|\frac{x_i^T x_l}{d}\right| \frac{1}{p} \sum_{j=1}^{p} \beta_j^2(0) |\psi'(W_j(k)^T x_i) - \psi'(W_j(0)^T x_i)| \\ &\quad + \left|\frac{x_i^T x_l}{d}\right| \frac{1}{p} \sum_{j=1}^{p} \beta_j^2(0) |\psi'(W_j(k)^T x_l) - \psi'(W_j(0)^T x_l)| \\ &\leq \left|\frac{x_i^T x_l}{d}\right| \frac{1}{p} \sum_{j=1}^{p} R_2(R_2 + 2|\beta_j(0)|) \\ &\quad + 2R_1 \left(\|x_l\| + \|x_i\|\right) \left|\frac{x_i^T x_l}{d}\right| \frac{1}{p} \sum_{j=1}^{p} \beta_j^2(0). \end{aligned}$$

Thus, by (A.35) and (A.39),

$$\max_{1\le l\le n}\sum_{i=1}^{n}|H_{il}(k)-H_{il}(0)|\lesssim \max_{1\le l\le n}\sum_{i=1}^{n}(R_2+R_1\sqrt{d})\left|\frac{x_i^Tx_l}{d}\right|\lesssim \frac{n\log p}{\sqrt{p}}\left(1+\frac{n}{\sqrt{d}}\right).$$

Then, we have

$$\|H(k)-H(0)\|_{\mathrm{op}}\le \max_{1\le l\le n}\sum_{i=1}^{n}|H_{il}(k)-H_{il}(0)|\lesssim \frac{n\log p}{\sqrt{p}}\left(1+\frac{n}{\sqrt{d}}\right).$$

Together with Lemma A.8, we obtain

$$0.089\le \lambda_{\min}(H(k))\le \lambda_{\max}(H(k))\lesssim 1. \tag{A.47}$$

Next, we give a bound for $r_i(k)$. By $\sup_x|\psi''(x)|\le 2$ and $\sup_x|\psi'(x)|\le 1$, we have

$$\begin{aligned}
|r_i(k)| &\le \frac{1}{\sqrt{p}}\sum_{j=1}^{p}|\beta_j(k+1)||(W_j(k+1)-W_j(k))^Tx_i|^2\\
&\le \frac{\|x_i\|^2}{\sqrt{p}}\sum_{j=1}^{p}|\beta_j(k+1)|\|W_j(k+1)-W_j(k)\|^2\\
&\le \frac{\gamma^2}{pd^2}\frac{\|x_i\|^2}{\sqrt{p}}\sum_{j=1}^{p}|\beta_j(k+1)||\beta_j(k)|^2\left(\sum_{l=1}^{n}|y_l-u_l(k)|\|x_l\|\right)^2\\
&\le \frac{\gamma^2}{pd^2}\frac{\|x_i\|^2\sum_{l=1}^{n}\|x_l\|^2}{\sqrt{p}}\|y-u(k)\|^2\sum_{j=1}^{p}|\beta_j(k+1)||\beta_j(k)|^2\\
&\lesssim \frac{\gamma^2 n}{\sqrt{p}}\|y-u(k)\|^2\\
&\lesssim \frac{\gamma^2 n\sqrt{n\log p}}{\sqrt{p}}\|y-u(k)\|,
\end{aligned}$$

where we have used (A.35), (A.36), (A.37) and (A.40) in the above inequalities. This leads to the bound

$$\|r(k)\|=\sqrt{\sum_{i=1}^{n}|r_i(k)|^2}\lesssim \frac{\gamma^2n^2\sqrt{\log p}}{\sqrt{p}}\|y-u(k)\|. \tag{A.48}$$

By (A.44), we have

$$\begin{aligned}
\|y-u(k+1)\|^2 &= \|y-u(k)\|^2-2\gamma(y-u(k))^T(H(k)+G(k))(y-u(k))\\
&\quad -2\langle y-u(k),r(k)\rangle+\|u(k)-u(k+1)\|^2.
\end{aligned}$$

The bounds (A.46) and (A.47) imply

$$-2\gamma(y-u(k))^T(H(k)+G(k))(y-u(k))\le -\frac{\gamma}{6}\|y-u(k)\|^2. \tag{A.49}$$

The bound (A.48) implies

$$-2\left\langle y-u(k),r(k)\right\rangle \leq 2\|y-u(k)\|\|r(k)\| \lesssim \frac{\gamma^2 n^2\sqrt{\log p}}{\sqrt{p}}\|y-u(k)\|^2.$$

Using (A.46), (A.47) and (A.48), we have

$$\begin{aligned}\|u(k)-u(k+1)\|^2 &\leq 2\gamma^2\|(H(k)+G(k))(y-u(k))\|^2+2\|r(k)\|^2 \\ &\lesssim \gamma^2\left(1+\frac{n^4(\log p)^2}{p}\right)\|y-u(k)\|^2+\frac{\gamma^4 n^4\log p}{p}\|y-u(k)\|^2.\end{aligned}$$

Therefore, as long as $\gamma\frac{n^4(\log p)^2}{p}$ is sufficiently small, we have

$$-2\left\langle y-u(k),r(k)\right\rangle+\|u(k)-u(k+1)\|^2 \leq \frac{\gamma}{24}\|y-u(k)\|^2.$$

Together with the bound (A.49), we have

$$\|y-u(k+1)\|^2 \leq \left(1-\frac{\gamma}{8}\right)\|y-u(k)\|^2 \leq \left(1-\frac{\gamma}{8}\right)^{k+1}\|y-u(0)\|^2,$$

and thus Claim D is true. The proof is complete. □

Appendix B: Results with ReLU activation

B.1. Repair of random feature model and neural nets

In this section, we present analogous results of Sections 4 and 5 with ReLU activation. First, consider the random feature model with design $\widetilde{X}=\psi(XW)=\{\psi(W_j^Tx_i)\}_{i\in[n],j\in[p]}$, where $\psi(t)=\max(0,t)$. Recall that $x_i\sim N(0,I_d)$ and $W_j\sim N(0,d^{-1}I_d)$ independently for all $i\in[n]$ and $j\in[p]$. The random matrix $\widetilde{X}$ has good properties, which is given by the following lemma.

Lemma B.1. *Assume n/p^2 and $n\log n/d$ are sufficiently small. Then, Condition A and Condition B hold for $A=\widetilde{X}^T$, $m=p$ and $k=n$ with some $\sigma^2\asymp p$, $\overline{\lambda}^2\asymp n$ and $\underline{\lambda}\asymp 1$.*

Now consider a model $\widehat{\theta}$ that lies in the row space of $\widetilde{X}$. For example, $\widehat{\theta}$ can be computed from a gradient-based algorithm initialized at 0. We observe a contaminated version $\eta=\widehat{\theta}+z$. We can then compute the procedure $\widetilde{u}=\operatorname{argmin}_{u\in\mathbb{R}^n}\|\eta-\widetilde{X}^Tu\|_1$ and use $\widetilde{\theta}=\widetilde{X}^T\widetilde{u}$ for model repair.

Corollary B.1. *Assume $\varepsilon\sqrt{n}$, n/p^2 and $n\log n/d$ are sufficiently small. We then have $\widetilde{\theta}=\widehat{\theta}$ with high probability.*

Next, we study the repair of neural network $f(x)=\frac{1}{\sqrt{p}}\sum_{j=1}^p\beta_j\psi(W_j^Tx)$ with ReLU activation $\psi(t)=\max(0,t)$. The gradient descent algorithm (Algorithm 1) enjoys the following property. Recall the notation that $u_i(t)=\frac{1}{\sqrt{p}}\sum_{j=1}^p\beta_j(t)\psi(W_j(t)^Tx_i)$. We assume x_i is i.i.d. $N(0,I_d)$ and $|y_i|\leq 1$ for all $i\in[n]$.

Theorem B.1. *Assume $\frac{n\log n}{d}$, $\frac{n^3(\log p)^4}{p}$ and γn are all sufficiently small. Then we have*

$$\max_{1\leq j\leq p}\|W_j(t)-W_j(0)\| \;\leq\; R_1, \tag{B.1}$$

$$\max_{1\leq j\leq p}|\beta_j(t)-\beta_j(0)| \;\leq\; R_2, \tag{B.2}$$

and

$$\|y-u(t)\|^2 \leq \left(1-\frac{\gamma}{8}\right)^t\|y-u(0)\|^2, \tag{B.3}$$

for all $t\geq 1$ with high probability, where $R_1=\frac{100n\log p}{\sqrt{pd}}$ and $R_2=32\sqrt{\frac{n^2\log p}{p}}$.

Consider the contaminated model $\eta=\widehat{\beta}+z$ and $\Theta_j=\widehat{W}_j+Z_j$, where each entry of z and Z_j is zero with probability $1-\varepsilon$ and follows an arbitrary distribution with the complementary probability ε. We apply Algorithm 2 to repair the neural net model. We study two situations. In the first situation, $\widehat{\beta}=\beta(t_{\max})$ and $\widehat{W}=W(t_{\max})$ are the direct output of Algorithm 1.

Theorem B.2. *Under the conditions of Theorem B.1, additionally assume that $\frac{\log p}{d}$ and $\varepsilon\sqrt{n}$ are sufficiently small. We then have $\widetilde{W}=\widehat{W}$ and $\frac{1}{p}\|\widetilde{\beta}-\widehat{\beta}\|^2\lesssim\frac{n^3\log p}{p}$ with high probability.*

In the second situation, we have $\widehat{W}=W(t_{\max})$ and then $\widehat{\beta}$ is obtained by carrying out gradient descent over β using features $\widetilde{X}=\psi(X\widehat{W})$. Since the gradient descent over β is initialized at 0, we shall replace the $\beta(0)$ by 0 in Algorithm 2 as well.

Theorem B.3. *Under the conditions of Theorem B.1, additionally assume that $\frac{\log p}{d}$ and $\varepsilon\sqrt{n}$ are sufficiently small. We then have $\widetilde{W}=\widehat{W}$ and $\widetilde{\beta}=\widehat{\beta}$ with high probability.*

Remark B.1. When $\varepsilon\sqrt{n}$ is sufficiently small, the conditions of Theorem B.2 and Theorem B.3 can be simplified as $p\gg n^3$ and $d\gg n$ by ignoring the logarithmic factors. The more stringent requirement on ε is due to the fact that the design matrix $\psi(XW)$ does not have approximate zero mean with the ReLU activation. This results in a large σ^2 in Condition A. In contrast, the hyperbolic tangent activation is an odd function, a property of symmetry that leads to Condition A with a constant σ^2.

B.2. Proofs of Lemma B.1 and Corollary B.1

We first state the proof of Lemma B.1. The conclusion of Condition A is obvious by

$$\sum_{i=1}^n\mathbb{E}\left(\frac{1}{p}\sum_{j=1}^p c_j\psi(W_j^Tx_i)\right)^2\leq\sum_{i=1}^n\frac{1}{p}\sum_{j=1}^p\mathbb{E}|W_j^Tx_i|^2=n,$$

and Markov's inequality. To check Condition B, we prove (3.3) and (3.4) separately.

Proof of (3.3) of Lemma B.1. We adopt a similar strategy to the proof of Lemma 4.2. Define

$$f(W,X,\Delta)=\frac{1}{p}\sum_{j=1}^p\left|\sum_{i=1}^n\psi(W_j^Tx_i)\Delta_i\right|,$$

and $g(X,\Delta)=\mathbb{E}(f(W,X,\Delta)|X)$. We then have

$$\begin{aligned}\inf_{\|\Delta\|=1} f(W,X,\Delta) &\geq \inf_{\|\Delta\|=1} \mathbb{E}f(W,X,\Delta) - \sup_{\|\Delta\|=1} |f(W,X,\Delta)-\mathbb{E}f(W,X,\Delta)| \\ &\geq \inf_{\|\Delta\|=1} \mathbb{E}f(W,X,\Delta) \qquad \text{(B.4)} \\ &\quad - \sup_{\|\Delta\|=1} |f(W,X,\Delta)-\mathbb{E}(f(W,X,\Delta)|X)| \qquad \text{(B.5)} \\ &\quad - \sup_{\|\Delta\|=1} |g(X,\Delta)-\mathbb{E}g(X,\Delta)|. \qquad \text{(B.6)}\end{aligned}$$

We will analyze the three terms above separately.

Analysis of (B.4). Define $h(W_j)=\mathbb{E}(\psi(W_j^Tx_i)|W_j)$ and $\bar{\psi}(W_j^Tx_i)=\psi(W_j^Tx_i)-h(W_j)$. We then have

$$\mathbb{E}f(W,X,\Delta)=\mathbb{E}\left|\sum_{i=1}^n \bar{\psi}(W^Tx_i)\Delta_i + h(W)\sum_{i=1}^n \Delta_i\right|. \tag{B.7}$$

A lower bound of (B.7) is

$$\mathbb{E}f(W,X,\Delta)\geq \left|\sum_{i=1}^n \Delta_i\right| |\mathbb{E}h(W)| - \mathbb{E}\left|\sum_{i=1}^n \bar{\psi}(W^Tx_i)\Delta_i\right|,$$

where the second term can be bounded by

$$\begin{aligned}\mathbb{E}\left|\sum_{i=1}^n \bar{\psi}(W^Tx_i)\Delta_i\right| &\leq \sqrt{\mathbb{E}\left|\sum_{i=1}^n \bar{\psi}(W^Tx_i)\Delta_i\right|^2} \\ &= \sqrt{\mathbb{E}\mathsf{Var}\left(\left|\sum_{i=1}^n \psi(W^Tx_i)\Delta_i\right|\Bigg|W\right)} \\ &= \sqrt{\mathbb{E}\sum_{i=1}^n \Delta_i^2\mathsf{Var}(\psi(W^Tx_i)|W)} \\ &= \sqrt{\mathbb{E}\sum_{i=1}^n \Delta_i^2\mathbb{E}(|\psi(W^Tx_i)|^2|W)} \\ &= \sqrt{\mathbb{E}|\psi(W^Tx)|^2}\leq\sqrt{\mathbb{E}|W^Tx|^2}=1.\end{aligned}$$

Since

$$\mathbb{E}h(W)=\frac{1}{\sqrt{2\pi}}\mathbb{E}\|W\|=\frac{1}{\sqrt{\pi}}\frac{\Gamma((d+1)/2)}{\sqrt{d}\Gamma(d/2)}\geq\frac{1}{\sqrt{2\pi}}\sqrt{\frac{d-1}{d}}.$$

Therefore, as long as $d\geq 3$ and $|\sum_{i=1}^n \Delta_i|\geq 7$, we have $\mathbb{E}f(W,X,\Delta)\geq 1$, and we thus can conclude that

$$\inf_{\|\Delta\|=1,|\sum_{i=1}^n \Delta_i|\geq 7} \mathbb{E}f(W,X,\Delta)\gtrsim 1. \tag{B.8}$$

Now we consider the case $|\sum_{i=1}^n \Delta_i| < 7$. A lower bound for $\left|\sum_{i=1}^n \psi(W^T x_i)\Delta_i\right|$ is

$$\left|\sum_{i=1}^n \psi(W^T x_i)\Delta_i\right| \geq \left|\sum_{i=1}^n \bar{\psi}(W^T x_i)\Delta_i\right| - 7h(W) = \left|\sum_{i=1}^n \bar{\psi}(W^T x_i)\Delta_i\right| - \frac{7}{\sqrt{2\pi}}\|W\|. \tag{B.9}$$

Thus,

$$\begin{aligned}
\mathbb{E}f(W,X,\Delta) &\geq \mathbb{E}\left(\left|\sum_{i=1}^n \psi(W^T x_i)\Delta_i\right| \mathbb{I}\left\{\left|\sum_{i=1}^n \bar{\psi}(W^T x_i)\Delta_i\right| \geq 6, 1/2 \leq \|W\|^2 \leq 2\right\}\right) \\
&\geq \mathbb{P}\left(\left|\sum_{i=1}^n \bar{\psi}(W^T x_i)\Delta_i\right| \geq 6, 1/2 \leq \|W\|^2 \leq 2\right) \\
&= \mathbb{P}\left(\left|\sum_{i=1}^n \bar{\psi}(W^T x_i)\Delta_i\right| \geq 6 \middle| 1/2 \leq \|W\|^2 \leq 2\right) \mathbb{P}\left(1/2 \leq \|W\|^2 \leq 2\right) \\
&\geq \mathbb{P}\left(\left|\sum_{i=1}^n \bar{\psi}(W^T x_i)\Delta_i\right| \geq 6 \middle| 1/2 \leq \|W\|^2 \leq 2\right) \left(1 - 2\exp(-d/16)\right),
\end{aligned}$$

where the last inequality is by Lemma A.4. By direct calculation, we have

$$\mathsf{Var}\left(\bar{\psi}(W^T x)|W\right) = \|W\|^2 \mathsf{Var}(\max(0, W^T x/\|W\|)|W) = \|W\|^2 \frac{1-\pi^{-1}}{2}, \tag{B.10}$$

and

$$\mathbb{E}\left(|\bar{\psi}(W^T x)|^3|W\right) \leq 3\mathbb{E}\left(|\psi(W^T x)|^3|W\right) + 3|h(W)|^3 \leq \frac{3}{2}\|W\|^3.$$

Therefore, by Lemma A.2, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{i=1}^n \bar{\psi}(W^T x_i)\Delta_i\right| \geq 6 \middle| 1/2 \leq \|W\|^2 \leq 2\right) \\
\geq\ & \mathbb{P}\left(\frac{\left|\sum_{i=1}^n \bar{\psi}(W^T x_i)\Delta_i\right|}{\|W\|\sqrt{\frac{1-\pi^{-1}}{2}}} \geq 21 \middle| 1/2 \leq \|W\|^2 \leq 2\right) \\
\geq\ & \mathbb{P}\left(N(0,1) > 21\right) - \sup_{1/2 \leq \|W\|^2 \leq 2} 2\sqrt{3\sum_{i=1}^n |\Delta_i|^3 \frac{\mathbb{E}\left(|\bar{\psi}(W^T x_i)|^3|W\right)}{\|W\|^3\left(\frac{1-\pi^{-1}}{2}\right)^{3/2}}} \\
\geq\ & \mathbb{P}\left(N(0,1) > 21\right) - 10\sqrt{\sum_{i=1}^n |\Delta_i|^3} \\
\geq\ & \mathbb{P}\left(N(0,1) > 21\right) - 10\max_{1 \leq i \leq n} |\Delta_i|^{3/2}.
\end{aligned}$$

Hence, when $\max_{1\leq i\leq n} |\Delta_i|^{3/2} \leq \delta_0^{3/2} := \mathbb{P}\left(N(0,1) > 21\right)/20$ and $|\sum_{i=1}^n \Delta_i| < 7$, we can lower bound $\mathbb{E}f(W,X,\Delta)$ by an absolute constant, and we conclude that

$$\inf_{\|\Delta\|=1, |\sum_{i=1}^n \Delta_i| \leq 7, \max_{1\leq i\leq n} |\Delta_i| \leq \delta_0} \mathbb{E}f(W,X,\Delta) \gtrsim 1. \tag{B.11}$$

Finally, we consider the case when $\max_{1\leq i\leq n}|\Delta_i| > \delta_0$ and $|\sum_{i=1}^n \Delta_i| < 7$. Without loss of generality, we can assume $\Delta_1 > \delta_0$. Note that the lower bound (B.9) still holds, and thus we have

$$\left|\sum_{i=1}^n \psi(W^Tx_i)\Delta_i\right| \geq \bar{\psi}(W^Tx_1)\Delta_1 - \left|\sum_{i=2}^n \bar{\psi}(W^Tx_i)\Delta_i\right| - \frac{7}{\sqrt{2\pi}}\|W\|.$$

We then lower bound $\mathbb{E}f(W,X,\Delta)$ by

$$\begin{aligned}
& \mathbb{E}\left(\left|\sum_{i=1}^n \psi(W^Tx_i)\Delta_i\right| \mathbb{I}\left\{\bar{\psi}(W^Tx_1)\Delta_1 \geq 8, \left|\sum_{i=2}^n \bar{\psi}(W^Tx_i)\Delta_i\right| \leq 2, 1/2 \leq \|W\|^2 \leq 2\right\}\right) \\
\geq\ & \mathbb{P}\left(\bar{\psi}(W^Tx_1)\Delta_1 \geq 8, \left|\sum_{i=2}^n \bar{\psi}(W^Tx_i)\Delta_i\right| \leq 2 \middle| 1/2 \leq \|W\|^2 \leq 2\right) \mathbb{P}\left(1/2 \leq \|W\|^2 \leq 2\right) \\
\geq\ & \mathbb{P}\left(\bar{\psi}(W^Tx_1)\Delta_1 \geq 8 \middle| 1/2 \leq \|W\|^2 \leq 2\right) \\
& \times\mathbb{P}\left(\left|\sum_{i=2}^n \bar{\psi}(W^Tx_i)\Delta_i\right| \leq 2 \middle| 1/2 \leq \|W\|^2 \leq 2\right)\left(1 - 2\exp(-d/16)\right).
\end{aligned}$$

For any W that satisfies $1/2 \leq \|W\|^2 \leq 2$, we have

$$\begin{aligned}
\mathbb{P}\left(\bar{\psi}(W^Tx_1)\Delta_1 \geq 8 \middle| W\right) &\geq \mathbb{P}\left(\bar{\psi}(W^Tx_1) \geq 8/\delta_0 \middle| W\right) \\
&\geq \mathbb{P}\left(\psi(W^Tx_1) \geq 8/\delta_0 + 1/\sqrt{\pi} \middle| W\right) \\
&\geq \mathbb{P}\left(W^Tx_1 \geq 8/\delta_0 + 1/\sqrt{\pi} \middle| W\right) \\
&\geq \mathbb{P}\left(N(0,1) \geq \sqrt{2}8/\delta_0 + \sqrt{2/\pi}\right),
\end{aligned}$$

which is a constant. We also have

$$\begin{aligned}
& \mathbb{P}\left(\left|\sum_{i=2}^n \bar{\psi}(W^Tx_i)\Delta_i\right| \leq 2 \middle| 1/2 \leq \|W\|^2 \leq 2\right) \\
\geq\ & 1 - \frac{1}{4}\mathsf{Var}\left(\sum_{i=2}^n \bar{\psi}(W^Tx_i)\Delta_i \middle| W\right) \\
\geq\ & \frac{1}{2},
\end{aligned}$$

where the last inequality is by (B.10). Therefore, we have

$$\mathbb{E}f(W,X,\Delta) \geq \frac{1}{2}\left(1 - 2\exp(-d/16)\right)\mathbb{P}\left(N(0,1) \geq \sqrt{2}8/\delta_0 + \sqrt{2/\pi}\right) \gtrsim 1,$$

and we can conclude that

$$\inf_{\|\Delta\|=1, |\sum_{i=1}^n \Delta_i| \leq 7, \max_{1\leq i\leq n}|\Delta_i| \geq \delta_0} \mathbb{E}f(W,X,\Delta) \gtrsim 1. \tag{B.12}$$

In the end, we combine the three cases (B.8), (B.11), and (B.12), and we obtain the conclusion that $\inf_{\|\Delta\|=1}\mathbb{E}f(W,X,\Delta) \gtrsim 1$.

Analysis of (B.5). This step follows the same analysis of (A.8) in the proof of Lemma 4.2, and we have

$$\sup_{\|\Delta\|=1} |f(W,X,\Delta) - \mathbb{E}(f(W,X,\Delta)|X)| \lesssim \sqrt{\frac{n^2}{p}},$$

with high probability.

Analysis of (B.6). This step follows a similar analysis of (A.9) in the proof of Lemma 4.2. The only difference is that the bound $\mathbb{E}g(X,\Delta) \leq 1$ there can be replaced by $\mathbb{E}g(X,\Delta) \leq \sqrt{n}$, because

$$\mathbb{E}g(X,\Delta) \leq \mathbb{E}\sqrt{\sum_{i=1}^n |\psi(W^Tx_i)|^2} \leq \sqrt{\sum_{i=1}^n \mathbb{E}|\psi(W^Tx_i)|^2} \leq \sqrt{n}.$$

Therefore,

$$\sup_{\|\Delta\|=1} |g(X,\Delta) - \mathbb{E}g(X,\Delta)| \lesssim \sqrt{\frac{n\log(1+2/\zeta)}{d}} + \sqrt{n}\zeta,$$

with high probability as long as $\zeta \leq 1/2$. We choose $\zeta = \frac{c}{\sqrt{n}}$ with a sufficiently small constant $c > 0$, and thus the bound is sufficiently small as long as $\frac{n\log n}{d}$ is sufficiently small.

Finally, combine results for (B.4), (B.5) and (B.6), and we obtain the desired conclusion as long as n^2/p and $n\log n/d$ are sufficiently small. □

To prove (3.4) of Lemma B.1, we establish the following stronger result.

Lemma B.2. *Consider independent* $W_1, ..., W_p \sim N(0, d^{-1}I_d)$ *and* $x_1, ..., x_n \sim N(0, I_d)$*. We define the matrices* $G, \bar{G} \in \mathbb{R}^{n\times n}$ *by*

$$G_{il} = \frac{1}{p}\sum_{j=1}^p \psi(W_j^Tx_i)\psi(W_j^Tx_l),$$

and

$$\bar{G}_{il} = \begin{cases} \frac{1}{2}, & i = l, \\ \frac{1}{2\pi} + \frac{1}{4}\frac{\bar{x}_i^T\bar{x}_l}{d} + \frac{1}{2\pi}\left(\frac{\|x_i\|}{\sqrt{d}} - 1 + \frac{\|x_l\|}{\sqrt{d}} - 1\right), & i \neq l. \end{cases}$$

Assume $d/\log n$ *is sufficiently large, and then*

$$\|G - \bar{G}\|_{\text{op}}^2 \lesssim \frac{n^2}{p} + \frac{\log n}{d} + \frac{n^2}{d^2},$$

with high probability. Moreover, we also have $\|G\|_{\text{op}} \lesssim n$ *with high probability.*

Proof. Define $\widetilde{G} \in \mathbb{R}^{n\times n}$ with entries $\widetilde{G}_{il} = \mathbb{E}\left(\psi(W^Tx_i)\psi(W^Tx_l)|X\right)$, and we first bound the difference between G and $\widetilde{G}$. Note that

$$\mathbb{E}(G_{il} - \widetilde{G}_{il})^2 = \mathbb{E}\mathsf{Var}(G_{il}|X) \leq \frac{1}{p}\mathbb{E}|\psi(W^Tx_i)\psi(W^Tx_l)|^2 = \frac{3}{2p}\mathbb{E}\|W\|^4 \leq 5p^{-1}.$$

We then have

$$\mathbb{E}\|G-\widetilde{G}\|_{\mathrm{op}}^2 \leq \mathbb{E}\|G-\widetilde{G}\|_{\mathrm{F}}^2 \leq \frac{5n^2}{p}.$$

By Markov's inequality,

$$\|G-\widetilde{G}\|_{\mathrm{op}}^2 \lesssim \frac{n^2}{p}, \tag{B.13}$$

with high probability.

Next, we study the diagonal entries of $\widetilde{G}$. For any $i \in [n]$, $\widetilde{G}_{ii} = \mathbb{E}(|\psi(W^T x_i)|^2|X) = \frac{\|x_i\|^2}{2d}$. By Lemma A.4 and a union bound argument, we have

$$\max_{1\leq i\leq n} |\widetilde{G}_{ii} - \bar{G}_{ii}| \lesssim \sqrt{\frac{\log n}{d}}, \tag{B.14}$$

with high probability.

Now we analyze the off-diagonal entries. We use the notation $\bar{x}_i = \frac{\sqrt{d}}{\|x_i\|}x_i$. For any $i \neq l$, we have

$$\begin{aligned}
\widetilde{G}_{il} \;\; = \;\; & \mathbb{E}\left(\psi(W^T\bar{x}_i)\psi(W^T\bar{x}_l)|X\right) && \text{(B.15)}\\
& +\mathbb{E}\left((\psi(W^T x_i)-\psi(W^T\bar{x}_i))\psi(W^T\bar{x}_l)|X\right) && \text{(B.16)}\\
& +\mathbb{E}\left(\psi(W^T\bar{x}_i)(\psi(W^T x_l)-\psi(W^T\bar{x}_l))|X\right) && \text{(B.17)}\\
& +\mathbb{E}\left((\psi(W^T x_i)-\psi(W^T\bar{x}_i))(\psi(W^T x_l)-\psi(W^T\bar{x}_l))|X\right). && \text{(B.18)}
\end{aligned}$$

For the first term on the right hand side of (B.15), we observe that $\mathbb{E}\left(\psi(W^T\bar{x}_i)\psi(W^T\bar{x}_l)|X\right)$ is a function of $\frac{\bar{x}_i^T\bar{x}_l}{d}$, and thus we can write

$$\mathbb{E}\left(\psi(W^T\bar{x}_i)\psi(W^T\bar{x}_l)|X\right) = f\left(\frac{\bar{x}_i^T\bar{x}_l}{d}\right),$$

where

$$f(\rho) = \begin{cases} \mathbb{E}\psi(\sqrt{1-\rho}U+\sqrt{\rho}Z)\psi(\sqrt{1-\rho}V+\sqrt{\rho}Z), & \rho \geq 0, \\ \mathbb{E}\psi(\sqrt{1+\rho}U-\sqrt{-\rho}Z)\psi(\sqrt{1+\rho}V+\sqrt{-\rho}Z), & \rho < 0, \end{cases}$$

with $U, V, Z \overset{iid}{\sim} N(0,1)$. By some direct calculations, we have $f(0) = \frac{1}{2\pi}$, $f'(0) = \frac{1}{4}$, and $\sup_{|\rho|\leq 0.2}\frac{|f'(\rho)-f'(0)|}{|\rho|} \lesssim 1$. Therefore, as long as $|\bar{x}_i^T\bar{x}_l|/d \leq 1/5$,

$$\left| f\left(\frac{\bar{x}_i^T\bar{x}_l}{d}\right) - \frac{1}{2\pi} - \frac{1}{4}\frac{\bar{x}_i^T\bar{x}_l}{d}\right| \leq C_1\left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|^2,$$

for some constant $C_1 > 0$. By Lemma A.5, we know that $\max_{i\neq l}|\bar{x}_i^T\bar{x}_l|/d \lesssim \sqrt{\frac{\log n}{d}} \leq 1/5$ with high probability, which then implies

$$\sum_{i\neq l}\left(\mathbb{E}\left(\psi(W^T\bar{x}_i)\psi(W^T\bar{x}_l)|X\right) - \bar{G}_{il}\right)^2 \leq C_1\sum_{i\neq l}\left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|^4. \tag{B.19}$$

The term on the right hand side has been analyzed in (A.24), and we have $\sum_{i\neq l}\left|\frac{x_i^Tx_l}{d}\right|^4 \lesssim \frac{n^2}{d^2}$ with high probability.

We also need to analyze the contributions of (B.16) and (B.17). Observe the fact that $\mathbb{I}\{W^Tx_i \geq 0\} = \mathbb{I}\{W^T\bar{x}_i \geq 0\}$, which implies

$$\begin{aligned}\psi(W^Tx_i) - \psi(W^T\bar{x}_i) &= W^T(x_i - \bar{x}_i)\mathbb{I}\{W^T\bar{x}_i \geq 0\}\psi(W^T\bar{x}_l) \\ &= \left(\frac{\|x_i\|}{\sqrt{d}} - 1\right)\psi(W^T\bar{x}_i)\psi(W^T\bar{x}_l). \qquad \text{(B.20)}\end{aligned}$$

Then, the sum of (B.16) and (B.17) can be written as

$$\left(\frac{\|x_i\|}{\sqrt{d}} - 1 + \frac{\|x_l\|}{\sqrt{d}} - 1\right) f\left(\frac{\bar{x}_i^T\bar{x}_l}{d}\right).$$

Note that

$$\begin{aligned}&\sum_{i\neq l}\left(\frac{\|x_i\|}{\sqrt{d}} - 1 + \frac{\|x_l\|}{\sqrt{d}} - 1\right)^2\left[f\left(\frac{\bar{x}_i^T\bar{x}_l}{d}\right) - \frac{1}{2\pi}\right]^2 \\ \lesssim\ &\sum_{i\neq l}\left(\frac{\|x_i\|}{\sqrt{d}} - 1 + \frac{\|x_l\|}{\sqrt{d}} - 1\right)^4 + \sum_{i\neq l}\left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|^4.\end{aligned}$$

We have already shown that $\sum_{i\neq l}\left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|^4 \lesssim \frac{n^2}{d^2}$ with high probability. By integrating out the probability tail bound of Lemma A.4, we have $\mathbb{E}\left(\frac{\|x_i\|}{\sqrt{d}} - 1\right)^4 \lesssim d^{-2}$, which then implies

$$\sum_{i\neq l}\mathbb{E}\left(\frac{\|x_i\|}{\sqrt{d}} - 1 + \frac{\|x_l\|}{\sqrt{d}} - 1\right)^4 \lesssim \frac{n^2}{d^2}$$

and the corresponding high-probability bound by Markov's inequality.

Finally, we show that the contribution of (B.18) is negligible. By (B.20), we can write (B.18) as

$$\left(\frac{\|x_i\|}{\sqrt{d}} - 1\right)\left(\frac{\|x_l\|}{\sqrt{d}} - 1\right)\mathbb{E}\left(\psi(W^T\bar{x}_i)^2\psi(W^T\bar{x}_l)^2\middle|X\right),$$

whose absolute value can be bounded by $\frac{3}{2}\left|\frac{\|x_i\|}{\sqrt{d}} - 1\right|\left|\frac{\|x_l\|}{\sqrt{d}} - 1\right|$. Since

$$\sum_{i\neq l}\mathbb{E}\left(\frac{\|x_i\|}{\sqrt{d}} - 1\right)^2\mathbb{E}\left(\frac{\|x_l\|}{\sqrt{d}} - 1\right)^2 \lesssim \frac{n^2}{d^2},$$

we can conclude that (B.18) is bounded by $O\left(\frac{n^2}{d^2}\right)$ with high probability by Markov's inequality.

Combining the analyses of (B.15), (B.16), (B.17) and (B.18), we conclude that $\sum_{i\neq l}(\widetilde{G}_{il} - \bar{G}_{il})^2 \lesssim \frac{n^2}{d^2}$ with high probability. Together with (B.13) and (B.14), we obtain the desired bound for $\|G - \bar{G}\|_{\text{op}}$.

To prove the last conclusion $\|\bar{G}\|_{\text{op}} \lesssim n$, it suffices to analyze $\lambda_{\max}(\bar{G})$. We bound this quantity by $\mathbb{E}\lambda_{\max}(\bar{G})^2 \leq \mathbb{E}\|\bar{G}\|_{\text{F}}^2 \lesssim n^2$, which leads to the desired conclusion. □

Proof of Corollary B.1. Since $\widehat{\theta}$ belongs to the row space of $\widetilde{X}$, there exists some $u^* \in \mathbb{R}^n$ such that $\widehat{\theta} = \widetilde{X}^T u^*$. By Theorem 3.1 and Lemma B.1, we know that $\widetilde{u} = u^*$ with high probability, and therefore $\widetilde{\theta} = \widetilde{X}^T \widetilde{u} = \widetilde{X}^T u^* = \widehat{\theta}$. $\square$

B.3. Proof of Theorem B.1

To prove Theorem B.1, we need the following kernel random matrix result.

Lemma B.3. *Consider independent $W_1, \ldots, W_p \sim N(0, d^{-1} I_d)$, $x_1, \ldots, x_n \sim N(0, I_d)$, and parameters $\beta_1, \ldots, \beta_p \sim N(0,1)$. We define the matrices $H, \bar{H} \in \mathbb{R}^{n \times n}$ by*

$$\begin{aligned} H_{il} &= \frac{x_i^T x_l}{d} \frac{1}{p} \sum_{j=1}^{p} \beta_j^2 \mathbb{I}\{W_j^T x_i \geq 0, W_j^T x_l \geq 0\}, \\ \bar{H}_{il} &= \frac{1}{4} \frac{x_i^T x_l}{\|x_i\| \|x_l\|} + \frac{1}{4} \mathbb{I}\{i = l\}. \end{aligned}$$

Assume $d / \log n$ is sufficiently large, and then

$$\|H - \bar{H}\|_{\mathrm{op}}^2 \lesssim \frac{n^2}{pd} + \frac{n}{p} + \frac{\log n}{d} + \frac{n^2}{d^2},$$

with high probability. If we additionally assume that d/n and p/n are sufficiently large, we will also have

$$\frac{1}{5} \leq \lambda_{\min}(H) \leq \lambda_{\max}(H) \lesssim 1, \tag{B.21}$$

with high probability.

Proof. Define $\widetilde{H} \in \mathbb{R}^{n \times n}$ with entries $\widetilde{H}_{il} = \frac{x_i^T x_l}{d} \mathbb{E}\left(\beta^2 \mathbb{I}\{W^T x_i \geq 0, W^T x_l \geq 0\} \middle| X\right)$, and we first bound the difference between H and $\widetilde{H}$. Note that

$$\mathbb{E}(H_{il} - \widetilde{H}_{il})^2 = \mathbb{E}\mathsf{Var}(H_{il}|X) \leq \frac{1}{p} \mathbb{E}\left(\frac{|x_i^T x_l|^2}{d^2} \beta^4\right) \leq \begin{cases} \frac{3}{pd}, & i \neq l, \\ 9p^{-1}, & i = l. \end{cases}$$

We then have

$$\mathbb{E}\|H - \widetilde{H}\|_{\mathrm{op}}^2 \leq \mathbb{E}\|H - \widetilde{H}\|_{\mathrm{F}}^2 \leq \frac{3n^2}{pd} + \frac{9n}{p}.$$

By Markov's inequality,

$$\|H - \widetilde{H}\|_{\mathrm{op}}^2 \lesssim \frac{n^2}{pd} + \frac{n}{p}, \tag{B.22}$$

with high probability.

Next, we study the diagonal entries of $\widetilde{H}$. For any $i \in [n]$, $\widetilde{H}_{ii} = \frac{\|x_i\|^2}{d} \mathbb{E}(\beta^2 \mathbb{I}\{W^T x_i \geq 0\}|X) = \frac{\|x_i\|^2}{2d}$. The same analysis that leads to the bound (B.14) also implies that

$$\max_{1 \leq i \leq n} |\widetilde{H}_{ii} - \bar{H}_{ii}| \lesssim \sqrt{\frac{\log n}{d}}, \tag{B.23}$$

with high probability.

Now we analyze the off-diagonal entries. Recall the notation $\bar{x}_i = \frac{\sqrt{d}}{\|x_i\|}x_i$. For any $i \neq l$, we have

$$\begin{aligned}
\widetilde{H}_{il} &= \frac{\|x_i\|\|x_l\|}{d}\frac{\bar{x}_i^T\bar{x}_l}{d}\mathbb{P}\left(W^T\bar{x}_i \geq 0, W^T\bar{x}_l \geq 0|X\right) \\
&= \frac{\bar{x}_i^T\bar{x}_l}{d}\mathbb{P}\left(W^T\bar{x}_i \geq 0, W^T\bar{x}_l \geq 0|X\right) \\
&\quad + \left(\frac{\|x_i\|\|x_l\|}{d} - 1\right)\frac{\bar{x}_i^T\bar{x}_l}{d}\mathbb{P}\left(W^T\bar{x}_i \geq 0, W^T\bar{x}_l \geq 0|X\right).
\end{aligned} \tag{B.24}$$

Since $\mathbb{P}\left(W^T\bar{x}_i \geq 0, W^T\bar{x}_l \geq 0|X\right)$ is a function of $\frac{\bar{x}_i^T\bar{x}_l}{d}$, we can write

$$\frac{\bar{x}_i^T\bar{x}_l}{d}\mathbb{P}\left(W^T\bar{x}_i \geq 0, W^T\bar{x}_l \geq 0|X\right) = f\left(\frac{\bar{x}_i^T\bar{x}_l}{d}\right), \tag{B.25}$$

where for $\rho > 0$,

$$\begin{aligned}
f(\rho) &= \rho\mathbb{P}\left(\sqrt{1-\rho}U + \sqrt{\rho}Z \geq 0, \sqrt{1-\rho}V + \sqrt{\rho}Z \geq 0\right) \\
&= \rho\mathbb{E}\mathbb{P}\left(\sqrt{1-\rho}U + \sqrt{\rho}Z \geq 0, \sqrt{1-\rho}V + \sqrt{\rho}Z \geq 0|Z\right) \\
&= \rho\mathbb{E}\Phi\left(\sqrt{\frac{\rho}{1-\rho}}Z\right)^2,
\end{aligned}$$

with $U, V, Z \overset{iid}{\sim} N(0,1)$ and $\Phi(\cdot)$ being the cumulative distribution function of $N(0,1)$. Similarly, for $\rho < 0$,

$$f(\rho) = \rho\mathbb{E}\left[\Phi\left(\sqrt{\frac{-\rho}{1+\rho}}Z\right)\left(1 - \Phi\left(\sqrt{\frac{-\rho}{1+\rho}}Z\right)\right)\right].$$

By some direct calculations, we have $f(0) = 0$, $f'(0) = \frac{1}{4}$, and

$$\sup_{|\rho|\leq 1/5}|f''(\rho)| \lesssim \sup_{|t|\leq 1/2}|\mathbb{E}\phi(tZ)\Phi(tZ)Z/t| + \sup_{|t|\leq 1/2}|\mathbb{E}\phi(tZ)Z/t|,$$

where $\phi(x) = (2\pi)^{-1/2}e^{-x^2/2}$. For any $|t| \leq 1/2$,

$$|\mathbb{E}\phi(tZ)Z/t| = \left|\mathbb{E}\frac{\phi(tZ)-\phi(0)}{tZ}Z^2\right| = \left|\mathbb{E}\xi\phi(\xi)Z^2\right| \leq \frac{|t|}{\sqrt{2\pi}}\mathbb{E}|Z|^3 \lesssim 1,$$

where ξ is a scalar between 0 and tZ so that $|\xi| \leq |tZ|$. By a similar argument, we also have $\sup_{|t|\leq 1/2}|\mathbb{E}\phi(tZ)\Phi(tZ)Z/t| \lesssim 1$ so that $\sup_{|\rho|\leq 1/5}|f''(\rho)| \lesssim 1$. Therefore, as long as $|\bar{x}_i^T\bar{x}_l|/d \leq 1/5$,

$$\left|f\left(\frac{\bar{x}_i^T\bar{x}_l}{d}\right) - \frac{1}{4}\frac{\bar{x}_i^T\bar{x}_l}{d}\right| \leq C_1\left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|^2,$$

for some constant $C_1 > 0$. By Lemma A.5, we know that $\max_{i\neq l}|\bar{x}_i^T\bar{x}_l|/d \lesssim \sqrt{\frac{\log n}{d}} \leq 1/5$ with high probability. In view of the identities (B.24) and (B.25), we then have the high probability bound,

$$\begin{aligned}\sum_{i\neq l}\left(\widetilde{H}_{il}-\frac{1}{4}\frac{\bar{x}_i^T\bar{x}_l}{d}\right)^2 &\leq 2\sum_{i\neq l}\left(\frac{\|x_i\|\|x_l\|}{d}-1\right)^2\left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|^2+2C_1\sum_{i\neq l}\left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|^4 \\ &\leq \sum_{i\neq l}\left(\frac{\|x_i\|\|x_l\|}{d}-1\right)^4+(2C_1+1)\sum_{i\neq l}\left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|^4. \qquad \text{(B.26)}\end{aligned}$$

For the first term on the right hand side of (B.26), we use Lemma A.5 and obtain a probability tail bound for $|\|x_i\|\|x_l\|-d|$. By integrating out this tail bound, we have

$$\sum_{i\neq l}\mathbb{E}\left(\frac{\|x_i\|\|x_l\|}{d}-1\right)^4\lesssim\frac{n^2}{d^2},$$

which, by Markov's inequality, implies $\sum_{i\neq l}\left(\frac{\|x_i\|\|x_l\|}{d}-1\right)^4\lesssim\frac{n^2}{d^2}$ with high probability. Using the same argument in the proof of Lemma B.2, we have $\sum_{i\neq l}\left|\frac{x_i^Tx_l}{d}\right|^4\lesssim\frac{n^2}{d^2}$ with high probability. Finally, combining (B.22), (B.23), and the bound for (B.26), we obtain the desired bound for $\|H-\bar{H}\|_{\text{op}}$. The last conclusion (B.21) follows a similar argument in the proof of Lemma A.7. The proof is complete. □

Now we are ready to prove Theorem B.1.

Proof of Theorem B.1. The proof is similar to that of Theorem 5.1, and we will omit repeated arguments. We will use the high-probability inequalities (A.34)-(A.42). Then, it suffices to establish Claims A, B, C and D in the proof of Theorem 5.1. Since Claims A and C follow the same argument, we only need to check Claims B and D. Given the similarity of Claims B and D, we only present the proof of Claim D. We have

$$u(k+1)-u(k)=\gamma(H(k)+G(k))(y-u(k))+r(k), \qquad \text{(B.27)}$$

where

$$\begin{aligned}G_{il}(k) &= \frac{1}{p}\sum_{j=1}^p\psi(W_j(k)^Tx_l)\psi(W_j(k)^Tx_i), \\ H_{il}(k) &= \frac{x_i^Tx_l}{d}\frac{1}{p}\sum_{j=1}^p\beta_j(k+1)^2\psi'(W_j(k)^Tx_i)\psi'(W_j(k)^Tx_l),\end{aligned}$$

and

$$\begin{aligned}r_i(k) &= \frac{1}{\sqrt{p}}\sum_{j=1}^p\beta_j(k+1)\left(\psi(W_j(k+1)^Tx_i)-\psi(W_j(k)^Tx_i)\right) \\ &\quad -\frac{1}{\sqrt{p}}\sum_{j=1}^p\beta_j(k+1)(W_j(k+1)-W_j(k))^Tx_i\psi'(W_j(k)^Tx_i).\end{aligned}$$

With the same argument, the bound (A.45) still holds. By Lemma B.2 and the fact that $G(k)$ is positive semi-definite, we have

$$0 \leq \lambda_{\min}(G(k)) \leq \lambda_{\max}(G(k)) \lesssim n. \tag{B.28}$$

We also need to control the difference between $H(k)$ and $H(0)$. By the definition, we have

$$\begin{aligned}
|H_{il}(k) - H_{il}(0)| \;\; &\leq \;\; \left|\frac{x_i^T x_l}{d}\right| \frac{1}{p}\sum_{j=1}^{p} |\beta_j(k+1)^2 - \beta_j^2(0)| && \text{(B.29)}\\
&\quad + \left|\frac{x_i^T x_l}{d}\right| \frac{1}{p}\sum_{j=1}^{p} \beta_j^2(0)|\psi'(W_j(k)^T x_i) - \psi'(W_j(0)^T x_i)| && \text{(B.30)}\\
&\quad + \left|\frac{x_i^T x_l}{d}\right| \frac{1}{p}\sum_{j=1}^{p} \beta_j^2(0)|\psi'(W_j(k)^T x_l) - \psi'(W_j(0)^T x_l)|. && \text{(B.31)}
\end{aligned}$$

We can bound (B.29) by $\left|\frac{x_i^T x_l}{d}\right| \frac{1}{p}\sum_{j=1}^{p} R_2(R_2 + 2|\beta_j(0)|)$. To bound (B.30), we note that

$$\begin{aligned}
|\psi'(W_j(k)^T x_i) - \psi'(W_j(0)^T x_i)| \;\; &\leq \;\; \mathbb{I}\{|W_j(0)^T x_i| \leq |(W_j(k) - W_j(0))^T x_i|\} \\
&\leq \;\; \mathbb{I}\{|W_j(0)^T x_i| \leq R_1\|x_i\|\}, && \text{(B.32)}
\end{aligned}$$

which implies

$$\begin{aligned}
&\left|\frac{x_i^T x_l}{d}\right| \frac{1}{p}\sum_{j=1}^{p} \beta_j^2(0)|\psi'(W_j(k)^T x_i) - \psi'(W_j(0)^T x_i)| \\
\leq \;\; &\left|\frac{x_i^T x_l}{d}\right| \frac{1}{p}\sum_{j=1}^{p} \beta_j^2(0)\mathbb{I}\{|W_j(0)^T x_i| \leq R_1\|x_i\|\},
\end{aligned}$$

and a similar bound holds for (B.31). Then,

$$\begin{aligned}
\|H(k) - H(0)\|_{\mathrm{op}} \;\; &\leq \;\; \max_{1\leq i\leq n} |H_{ii}(k) - H_{ii}(0)| + \max_{1\leq l\leq n} \sum_{i\in[n]\backslash\{l\}} |H_{il}(k) - H_{il}(0)| \\
&\lesssim \;\; \max_{1\leq i\leq n} \frac{1}{p}\sum_{j=1}^{p} \beta_j^2(0)\mathbb{I}\{|W_j^T(0)^T x_i| \leq R_1\|x_i\|\} \\
&\quad\;\; + d^{-1/2} n \max_{1\leq i\leq n} \frac{1}{p}\sum_{j=1}^{p} \beta_j^2(0)\mathbb{I}\{|W_j^T(0)^T x_i| \leq R_1\|x_i\|\} \\
&\quad\;\; + \max_{1\leq l\leq n} \sum_{i=1}^{n} \left|\frac{x_i^T x_l}{d}\right| R_2 \frac{1}{p}\sum_{j=1}^{p} (R_2 + 2|\beta_j(0)|) \\
&\lesssim \;\; \left(1 + \frac{n}{\sqrt{d}}\right)\left(\sqrt{d}R_1 \log p + \frac{\sqrt{\log n}\log p}{\sqrt{p}} + R_2^2 + R_2\sqrt{\log p}\right) \\
&\lesssim \;\; \left(1 + \frac{n}{\sqrt{d}}\right)\frac{n(\log p)^2}{\sqrt{p}},
\end{aligned}$$

where we have used (A.34), (A.37), (A.38), (A.39) and (A.42). In view of Lemma B.3, we then have

$$\frac{1}{6} \leq \lambda_{\min}(H(k)) \leq \lambda_{\max}(H(k)) \lesssim 1, \tag{B.33}$$

under the conditions of d, p and n.

Next, we give a bound for $r_i(k)$. Observe that

$$\psi(W_j(k+1)^T x_i) - \psi(W_j(k)^T x_i) = (W_j(k+1) - W_j(k))^T x_i \psi'(W_j(k)^T x_i),$$

when $\mathbb{I}\{W_j(k+1)^T x_i > 0\} = \mathbb{I}\{W_j(k)^T x_i > 0\}$. Thus, we only need to sum over those $j \in [p]$ that $\mathbb{I}\{W_j(k+1)^T x_i > 0\} \neq \mathbb{I}\{W_j(k)^T x_i > 0\}$. By (B.32), we have

$$\begin{aligned}
& \left|\mathbb{I}\{W_j(k+1)^T x_i > 0\} - \mathbb{I}\{W_j(k)^T x_i > 0\}\right| \\
\leq\ & \left|\mathbb{I}\{W_j(k+1)^T x_i > 0\} - \mathbb{I}\{W_j(0)^T x_i > 0\}\right| + \left|\mathbb{I}\{W_j(k)^T x_i > 0\} - \mathbb{I}\{W_j(0)^T x_i > 0\}\right| \\
\leq\ & 2\mathbb{I}\{|W_j^T(0)^T x_i| \leq R_1\|x_i\|\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \left|\psi(W_j(k+1)^T x_i) - \psi(W_j(k)^T x_i) - (W_j(k+1) - W_j(k))^T x_i \psi'(W_j(k)^T x_i)\right| \\
\leq\ & 4|(W_j(k+1) - W_j(k))^T x_i|\mathbb{I}\{|W_j^T(0)^T x_i| \leq R_1\|x_i\|\} \\
\leq\ & \frac{4\gamma}{d\sqrt{p}}|\beta_j(k+1)|\|y - u(k)\|\|x_i\|\sqrt{\sum_{l=1}^n \|x_l\|^2 \mathbb{I}\{|W_j^T(0)^T x_i| \leq R_1\|x_i\|\}},
\end{aligned}$$

which implies

$$\begin{aligned}
|r_i(k)| \ &\leq\ \frac{4\gamma}{dp}\sum_{j=1}^p |\beta_j(k+1)|^2 \|y - u(k)\|\|x_i\|\sqrt{\sum_{l=1}^n \|x_l\|^2 \mathbb{I}\{|W_j^T(0)^T x_i| \leq R_1\|x_i\|\}} \\
&\lesssim\ \sqrt{n}\|y - u(k)\|\gamma\frac{1}{p}\sum_{j=1}^p (\beta_j(0)^2 + R_2^2)\mathbb{I}\{|W_j^T(0)^T x_i| \leq R_1\|x_i\|\} \\
&\lesssim\ \gamma\sqrt{n}\log p\left(R_1 + \sqrt{\frac{\log n}{p}}\right)\|y - u(k)\|.
\end{aligned}$$

This leads to the bound

$$\|r(k)\| = \sqrt{\sum_{i=1}^n |r_i(k)|^2} \lesssim \gamma n \log p\left(R_1 + \sqrt{\frac{\log n}{p}}\right)\|y - u(k)\|. \tag{B.34}$$

Now we are ready to analyze $\|y - u(k+1)\|^2$. Given the relation (B.27), we have

$$\begin{aligned}
\|y - u(k+1)\|^2 \ &=\ \|y - u(k)\|^2 - 2\langle y - u(k), u(k+1) - u(k)\rangle + \|u(k) - u(k+1)\|^2 \\
&=\ \|y - u(k)\|^2 - 2\gamma(y - u(k))^T(H(k) + G(k))(y - u(k)) \\
&\quad -2\langle y - u(k), r(k)\rangle + \|u(k) - u(k+1)\|^2.
\end{aligned}$$

By (B.28) and (B.33), we have

$$-2\gamma(y-u(k))^T(H(k)+G(k))(y-u(k)) \leq -\frac{\gamma}{6}\|y-u(k)\|^2. \tag{B.35}$$

The bound (B.34) implies

$$-2\left\langle y-u(k), r(k)\right\rangle \leq 2\|y-u(k)\|\|r(k)\| \lesssim \gamma n \log p\left(R_1+\sqrt{\frac{\log n}{p}}\right)\|y-u(k)\|^2.$$

By (B.28), (B.33) and (B.34), we also have

$$\begin{aligned}
\|u(k)-u(k+1)\|^2 &\leq 2\gamma^2\|(H(k)+G(k))(y-u(k))\|^2+2\|r(k)\|^2 \\
&\lesssim \gamma^2 n\|y-u(k)\|^2+(\gamma n\log p)^2\left(R_1+\sqrt{\frac{\log n}{p}}\right)^2\|y-u(k)\|^2.
\end{aligned}$$

Therefore, as long as $\frac{n\log n}{d}$, $\frac{n^3(\log p)^4}{p}$ and γn are all sufficiently small, we have

$$-2\left\langle y-u(k), r(k)\right\rangle+\|u(k)-u(k+1)\|^2 \leq \frac{\gamma}{24}\|y-u(k)\|^2.$$

Together with the bound (B.35), we have

$$\|y-u(k+1)\|^2 \leq \left(1-\frac{\gamma}{8}\right)\|y-u(k)\|^2 \leq \left(1-\frac{\gamma}{8}\right)^{k+1}\|y-u(0)\|^2,$$

and thus Claim D is true. The proof is complete. □

B.4. Proofs of Theorem B.2 and Theorem B.3

Proof of Theorem B.2. The proof is the same as that of Theorem 5.2. The only exception here is that we apply Lemma B.1 and Lemma B.2 instead of Lemma 4.2 and Lemma A.7. □

Proof of Theorem B.3. The analysis of $\widehat{v}_1, ..., \widehat{v}_p$ is the same as that in the proof of Theorem 5.2, and we have $\widetilde{W}_j = \widehat{W}_j$ for all $j \in [p]$ with high probability.

To analyze $\widehat{u}$, we apply Theorem 3.1. It suffices to check Condition A and Condition B for the design matrix $\psi(X^T\widetilde{W}^T) = \psi(X^T\widehat{W}^T)$. Since

$$\sum_{i=1}^n \mathbb{E}\left(\frac{1}{p}\sum_{j=1}^p c_j\psi(\widehat{W}_j^T x_i)\right)^2 \leq \sum_{i=1}^n \frac{1}{p}\sum_{j=1}^p \mathbb{E}\psi(\widehat{W}^T x_i)^2,$$

and $\mathbb{E}\psi(\widehat{W}^T x_i)^2 \leq \mathbb{E}|\widehat{W}_j^T x_i|^2 \lesssim 1+R_1 d \lesssim 1$, Condition A holds with $\sigma^2 \asymp p$. We also need to check Condition B. By Theorem B.1, we have

$$\begin{aligned}
& \left| \frac{1}{p}\sum_{j=1}^{p}\left|\sum_{i=1}^{n}\psi(\widehat{W}_j^T x_i)\Delta_i\right| - \frac{1}{p}\sum_{j=1}^{p}\left|\sum_{i=1}^{n}\psi(W_j(0)^T x_i)\Delta_i\right|\right| \\
\leq\ & \frac{1}{p}\sum_{j=1}^{p}\sum_{i=1}^{n}|\widehat{W}_j^T x_i - W_j(0)^T x_i||\Delta_i| \\
\leq\ & R_1\sum_{i=1}^{n}\|x_i\||\Delta_i| \\
\leq\ & R_1\sqrt{\sum_{i=1}^{n}\|x_i\|^2} \\
\lesssim\ & \frac{n^{3/2}\log p}{\sqrt{p}},
\end{aligned}$$

where $\sum_{i=1}^{n}\|x_i\|^2 \lesssim nd$ is by Lemma A.4. By Lemma B.1, we can deduce that

$$\inf_{\|\Delta\|=1}\frac{1}{p}\sum_{j=1}^{p}\left|\sum_{i=1}^{n}\psi(\widehat{W}_j^T x_i)\Delta_i\right| \gtrsim 1,$$

as long as $\frac{n^{3/2}\log p}{\sqrt{p}}$ is sufficiently small. By (B.28), we also have

$$\sup_{\|\Delta\|=1}\frac{1}{p}\sum_{j=1}^{p}\left|\sum_{i=1}^{n}\psi(\widehat{W}_j^T x_i)\Delta_i\right|^2 \lesssim n.$$

Therefore, Condition B holds with $\overline{\lambda}^2 \asymp n$ and $\underline{\lambda} \asymp 1$. Applying Theorem 3.1, we have $\widetilde{\beta} = \widehat{\beta}$ with high probability, as desired. □