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

April 6, 2020

Abstract: A new type of robust estimation problem is introduced where the goal is to recover a statistical model that has been corrupted after it has been estimated from data. Methods are proposed for "repairing" the model using only the design and not the response values used to fit the model in a supervised learning setting. Theory is developed which reveals that two important ingredients are necessary for model repair—the statistical model must be over-parameterized, and the estimator must incorporate redundancy. In particular, estimators based on stochastic gradient descent are seen to be well suited to model repair, but sparse estimators are not in general repairable. After formulating the problem and establishing a key technical lemma related to robust estimation, a series of results are presented for repair of over-parameterized linear models, random feature models, and artificial neural networks. Simulation studies are presented that corroborate and illustrate the theoretical findings.

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 crude 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. We also explain how the concepts of over-paramaterization and redundancy apply to repair of nonparametric models, including Gaussian sequence models for Sobolev spaces and isotonic regression. 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. Section 6 presents further simulation results that confirm the theory. In Section 7 we discuss 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 $z_j \sim (1-\varepsilon)\delta_0 + \varepsilon Q$ and Q 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 with probability ε, where $0 \leq \varepsilon \leq 1$, and is uncorrupted with probability $1-\varepsilon$. We discuss alternative error models later in the paper. The goal is to recover $\widehat{\theta}$ from η, without reestimating the model using the response

values $\{y_i\}$.

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.

Now suppose that $\eta = \widehat{\theta} + z$ where $z_j \sim (1-\varepsilon)\delta_0 + \varepsilon Q$. 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

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

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 model is repaired with high probability, so that $\widetilde{\theta} = \widehat{\theta}$, as long as $n/p \le c(1-\varepsilon)^2$ 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 $Q = N(1,1)$. With the sample size fixed at $n = 50$, the dimension p is varied according to $p_j/n = 200/j^2$ with j 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_j as a function of the adjusted value $\varepsilon_j = \varepsilon + c' \cdot j - \frac{1}{2}$ should exhibit a threshold at $\varepsilon_j = 1/2$ 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 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 stark contrast to the traditional Huber model where the design is corrupted (Huber, 1964), under which consistent estimation is only possible if $\varepsilon \to 0$ (Chen et al., 2016; Gao, 2020).

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

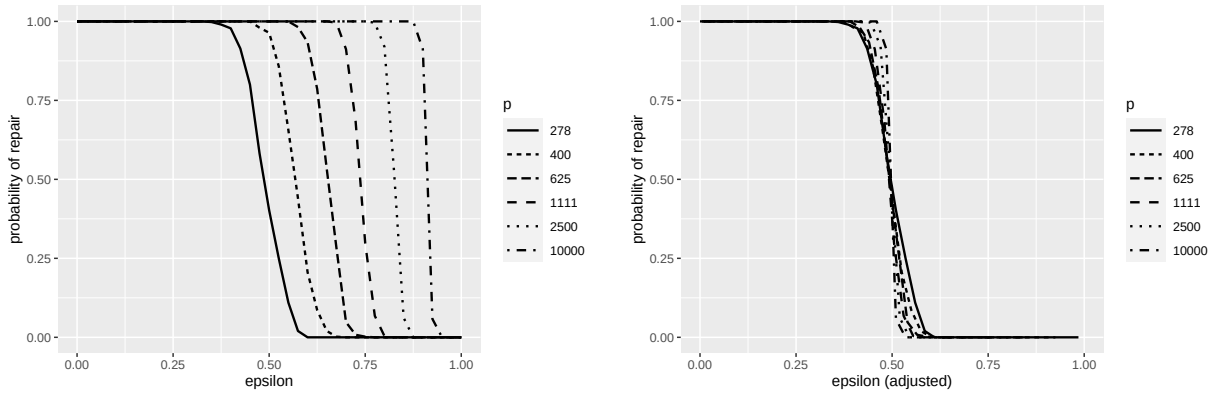

FIG 1. *Left: Empirical probability of exact repair as a function of ε. The sample size is $n = 50$ and the model dimension p varies as $p_j/n = 200/j^2$, for $j = 1, \ldots, 6$; each point is an average over 500 trials. The plot on the right shows the repair probability as a function of the adjusted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

high probability. When X is a random Gaussian matrix, the mapping $u \to X^T u = \sum_{i=1}^n u_i X_i^T$ 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 $w \to w + z$ that we consider corresponds to a channel having infinite capacity, and a simple repetition code would suffice for identifying components that are uncorrupted.

Estimators based on gradient descent. The observations made above carry over to estimators of linear models based on gradient descent. 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

$$\theta^{(t+1)} = \theta^{(t)} - \gamma_t \frac{1}{n} \sum_{i=1}^n \nabla_\theta \mathcal{L}(y_i, x_i^T \theta^{(t-1)}) \tag{2.5}$$

$$= \theta^{(t)} - \gamma_t \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial f} \mathcal{L}(y_i, x_i^T \theta^{(t-1)}) x_i \tag{2.6}$$

$$= \theta^{(t)} - \sum_{i=1}^n w_i^{(t)} x_i \tag{2.7}$$

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 time t can thus be written as

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

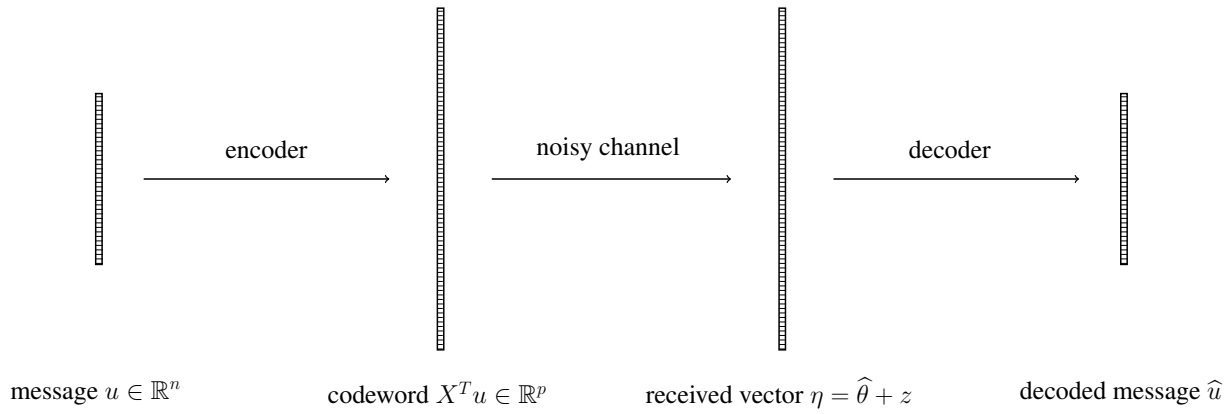

FIG 2. *Model repair viewed in terms of error-correcting codes. The model $\widehat{\theta} = X^T u \in \mathbb{R}^p$ is in the row-space of the design matrix, which gives a redundant representation of the "message" $u \in \mathbb{R}^n$ for $n < p$. The model is received as a noisy version η with each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

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. Let B_t be the set of samples used in the mini-batch of the tth iteration. Then, we can write

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

We can recover $\theta^{(t)}$ from a 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 takes 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.

3. Background on robust regression

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, there is an ε-proportion of η_i's that are contaminated by z_i's that are drawn from some arbitrary unknown distributions. To robustly recover u^*, we propose the estimator

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

It 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. 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.

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

$$\inf_{\|\Delta\|=1} \frac{1}{m}\sum_{i=1}^m |a_i^T\Delta| \;\geq\; \underline{\lambda}, \tag{3.2}$$

$$\sup_{\|\Delta\|=1} \frac{1}{m}\sum_{i=1}^m |a_i^T\Delta|^2 \;\leq\; \overline{\lambda}^2, \tag{3.3}$$

with high probability.

Theorem 3.1. *Assume the design matrix A satisfies Condition A and Condition B. Then, as long as $\frac{\overline{\lambda}\sqrt{\frac{k}{m}\log\left(\frac{em}{k}\right)}}{\underline{\lambda}(1-\varepsilon)}$ is sufficiently small, we have $\widehat{u} = u^*$ with high probability.*

The theorem gives a sufficient condition on the exact recovery of the regression coefficient. When both $\overline{\lambda}/\underline{\lambda}$ and $1-\varepsilon$ are constants, the condition becomes k/m sufficiently small. One remarkable feature of this theorem is that it even allows the situation $\varepsilon \to 1$. This is in contrast to robust regression with both response and design 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 an ε-probability 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 why Theorem 3.1 allows $\varepsilon \to 1$ is because there is no contamination for the design matrix A.

Another distinguished feature of Theorem 3.1 is that there is no assumption imposed on the contamination distribution Q_i, even though the median regression procedure naturally requires

the noise to be symmetric around zero. To understand this phenomenon, note that with the help of the independent Rademacher random variables, 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 because of Condition A. 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 equivalent median regression after symmetrization, which explains why Theorem 3.1 does not require any assumption on Q_i.

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

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

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 $\underline{\lambda}$ 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 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:

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

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 $\psi(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}}$. 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 $\underline{\lambda}$ 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.*

5. Repair of neural networks

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

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

The factor $p^{-1/2}$ in the definition above is convenient for our theoretical analysis. Consider the loss function

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

and we train the neural network model via a standard gradient descent algorithm (Algorithm 1).

Algorithm 1: Gradient descent for neural nets

Input : The data (y, X) and the number of iterations $t_{\max}$.

Output: The trained parameters $\beta(t_{\max})$ and $W(t_{\max})$.

1 Initialization: $W_j(0) \sim N(0, d^{-1}I_d)$ and $\beta_j(0) \sim N(0,1)$ independently for all $j \in [p]$.

2 For t in $1 : t_{\max}$, compute

$$\begin{aligned} W_j(t) &= W_j(t-1) - \frac{\gamma}{d}\frac{\partial L(\beta, W)}{\partial W_j}|_{(\beta,W)=(\beta(t-1),W(t-1))}, \quad j \in [p], \\ \beta_j(t) &= \beta_j(t-1) - \gamma\frac{\partial L(\beta, W)}{\partial \beta_j}|_{(\beta,W)=(\beta(t-1),W(t-1))}, \quad j \in [p]. \end{aligned}$$

Based on Algorithm 1, we consider two variations of $(\widehat{\beta}, \widehat{W})$:

1. Implement Algorithm 1, and set $\widehat{\beta} = \beta(t_{\max})$ and $\widehat{W} = W(t_{\max})$;
2. Implement Algorithm 1, and set $\widehat{W} = W(t_{\max})$. Retrain β using the learned feature $\psi(X\widehat{W})$. That is, take $\widehat{\beta}$ to be minimal ℓ_2 norm solution of $\|y - \psi(X\widehat{W})\beta\|^2$, compute $\widehat{\beta}$ using a gradient based algorithm initialized at 0 for the objective $\|y - \psi(X\widehat{W})\beta\|^2$.

Both variations of $(\widehat{\beta}, \widehat{W})$ are widely used in practice. The second option 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 0 with probability $1 - \varepsilon$ and follows an arbitrary distribution with the other probability $1 - \varepsilon$. We present the following algorithm that repairs the model.

Algorithm 2: Model repair for neural nets

Input : Contaminated model (η, Θ), design matrix X, and initialization $(\beta(0), W(0))$.

Output: The rapaired parameters $\widetilde{\beta}$ and $\widetilde{W}$.

1 Repair of the hidden layer: for $j \in [p]$, compute

$$\widehat{v}_j = \underset{v}{\operatorname{argmin}} \|\Theta_j - W_j(0) - X^T v_j\|_1,$$

and set $\widetilde{W}_j = W_j(0) + X^T\widehat{v}_j$.

2 Repair of the output layer: compute

$$\widehat{u} = \underset{u}{\operatorname{argmin}} \|\eta - \beta(0) - \psi(\widetilde{W}^T X^T)u\|_1,$$

and set $\widetilde{\beta} = \beta(0) + \psi(\widetilde{W}^T X^T)\widehat{u}$.

Algorithm 2 adopts a layerwise repair strategy. It is important to note that the repair of neural networks not only require information of X, but also that of the initialization $(\beta(0), W(0))$. It is

thus crucial for practitioners to always store $(\beta(0), W(0))$ after training in case model repair is needed later.

Since the gradient $\frac{\partial L(\beta,W)}{\partial W_j}$ 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. On the other hand, the repair of the output layer is more complicated, because the gradient $\frac{\partial L(\beta,W)}{\partial \beta_j}|_{(\beta,W)=(\beta(t-1),W(t-1))}$ 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 assume that x_i is i.i.d. $N(0, I_d)$ and $|y_i| \leq 1$ for all $i \in [n]$.

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

$$\max_{1\leq j\leq p}\left(\frac{\|W_j(t) - W_j(0)\|}{R_1} \vee \frac{|\beta_j(t) - \beta_j(0)|}{R_2}\right) \leq 1, \tag{5.2}$$

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

Theorem 5.1 assumes that the width of the neural network to be wide compared with the sample size in the sense that $\frac{p}{(\log p)^2} \gg n^3$. For a fixed n, the limit of the neural network when $p \to \infty$ is known as the neural tangent kernel (NTK), and the behavior of the 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. (2018), and we are able to sharpen some of the asymptotic conditions in Du et al. (2018).

The theorem has two conclusions. The first conclusion shows the gradient descent algorithm has global convergence in the sense of (5.2) even though the loss $L(\beta, W)$ 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 feature $\psi(X\widehat{W})$. In this case, the vector $\widehat{\beta} - \beta(0)$ 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 Corollary 4.2. Compared with Theorem 5.2, we are able to obtain exact recover 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 5.1. When $1-\varepsilon$ is a constant, 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. Simulation studies

6.1. 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

$$\widehat{\theta} = X^T(XX^T)^{-1}y \tag{6.1}$$

The estimated model is corrupted to

$$\eta = \widehat{\theta} + z \tag{6.2}$$

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

$$\widetilde{u} = \operatorname{argmin} \|\eta - X^T u\|_1 \tag{6.3}$$

$$\widetilde{\theta} = X^T\widetilde{u}. \tag{6.4}$$

The `quantreg` package in R is used to carry out the median regression (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).

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.

6.2. Random features models trained with gradient descent

In this experiment we simulate over-parameterized random features models. 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. The covariates are generated as a layer of a random neural network, with $X_i = \tanh(WZ_i)$ where $Z_i \in \mathbb{R}^d$ with $Z_{ij} \sim N(0,1)$

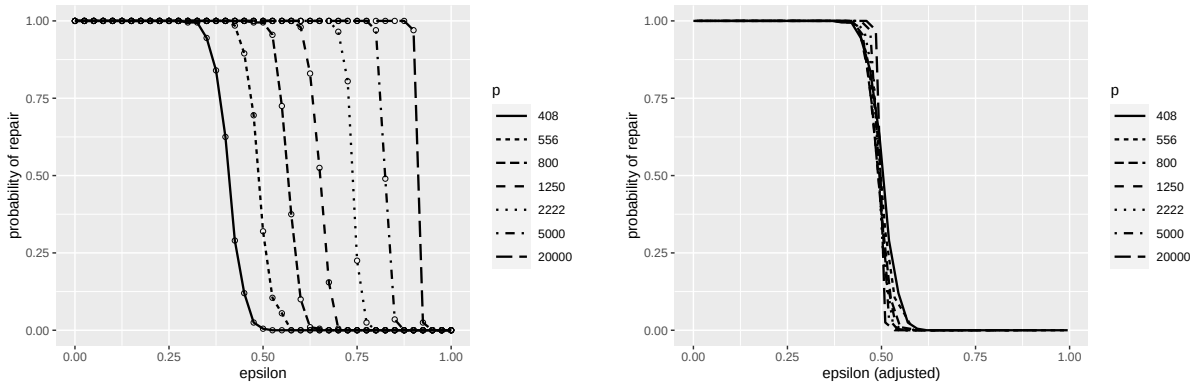

FIG 3. *Model repair for underdetermined linear models $y = X^T\theta + w$ with $p > n$. The left plot shows the empirical probability of successful model repair for $n = 100$ with the model dimension p varying as $p/n = 200/j^2$, for $j = 1, \ldots, 7$. Each point is an average over 500 random trials. 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 $\widetilde{\varepsilon}_j = \varepsilon + c' \cdot j - \frac{1}{2}$.*

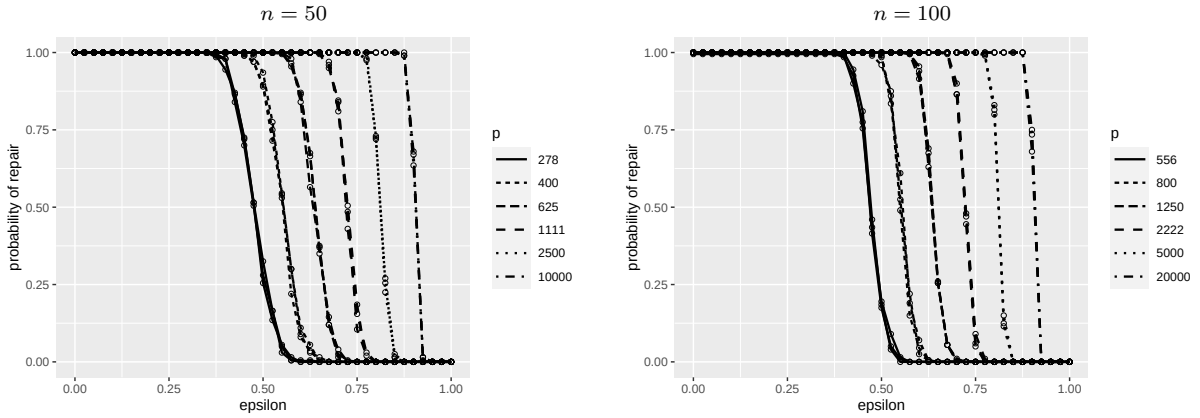

FIG 4. *Model repair for random feature models $y = \psi(XW)\theta + w$ with $p > n$, where $\psi(\cdot) = \tanh(\cdot)$ for $n = 50$ (left) and $n = 100$ (right). For each value of p, three values of d are evaluated, $d = p$, $d = \lceil 2p/3 \rceil$, and $d = \lceil p/2 \rceil$; the results are effectively the same for each d.*

and $W \in \mathbb{R}^{p \times d}$ with $W_{ij} \sim N(0, 1/d)$. We then approximate the least squares solution using gradient descent intialized at zero, with updates

$$\widehat{\theta}^{(t)} = \widehat{\theta}^{(t-1)} + \frac{\eta}{n} X^T R^{(t-1)} \tag{6.5}$$

where the residual vector $R^{(t-1)} \in \mathbb{R}^n$ is given by $R_i = (y_i - X_i^T \widehat{\theta}^{(t-1)})$. The step size η is selected empirically to insure convergence in under $1{,}000$ iterations. Figure 4 shows two sets of 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.

7. Discussion

8. Proofs

8.1. Technical Lemmas

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

Lemma 8.1 (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 $t > 0$,*

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

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 8.2. *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*

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

Lemma 8.3 (Cirel'son et al. (1976))**.** *Let $f : \mathbb{R}^k \to \mathbb{R}$ be a Lipschitz function with constant $L > 0$. That is, $|f(x) - f(y)| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^k$. Then, for any $t > 0$,*

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

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

Lemma 8.4 (Laurent and Massart (2000))**.** *For any $t > 0$, we have*

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

Lemma 8.5. *Consider independent $Y_1, Y_2 \sim N(0, I_k)$. For any $t > 0$, we have*

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

Proof. By Lemma 8.4, we have

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

and

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

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

$$\mathbb{P}\left(Y_1^T Y_2 \geq x\right) \leq e^{-\lambda x}\mathbb{E}e^{\lambda Y_1^T Y_2} = \exp\left(-\lambda x - \frac{k}{2}\log(1-\lambda^2)\right) \leq \exp\left(-\lambda x + \frac{k}{2}\lambda^2\right),$$

for any $x > 0$ and $\lambda \in (0,1)$. Optimize over $\lambda \in (0,1)$, and we obtain $\mathbb{P}\left(Y_1^T Y_2 > x\right) \leq e^{-\frac{1}{2}\left(\frac{x^2}{k} \wedge x\right)}$. Take $x = \sqrt{2kt} + 2t$, and then we obtain the bound

$$\mathbb{P}\left(Y_1^T Y_2 \geq \sqrt{2kt} + 2t\right) \leq e^{-t},$$

which immediately implies the second conclusion. □

8.2. Proof of Theorem 3.1

In order to prove Theorem 3.1, we establish a more general result. Consider $\eta = b + Au^* + z \in \mathbb{R}^m$, where the noise vector z satisfies (3.1), and $b \in \mathbb{R}^m$ is any bias vector. Then, the estimator $\widehat{u} = \operatorname{argmin}_{u\in\mathbb{R}^k} \|\eta - Au\|_1$ satisfies the following theoretical guarantee.

Theorem 8.1. *Assume the design matrix A satisfies Condition A and Condition B. Then, as long as $\frac{\overline{\lambda}\sqrt{\frac{k}{m}\log\left(\frac{em}{k}\right)}}{\underline{\lambda}(1-\varepsilon)}$ 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 $b = 0$. To prove Theorem 8.1, we need the following empirical process result.

Lemma 8.6. *Consider independent random variables $z_1, ..., z_m$. Assume $k/m \leq 1$. Then, for any $t \in (0, 1/2)$ and any fixed $A^T = (a_1, ..., a_m)^T$ such that (3.3) holds, we have*

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

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 $B_k(t) = \{\Delta \in \mathbb{R}^k : \|\Delta\| \le t\}$, 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 $\|\Delta - \Delta'\| \le \zeta$, and we also have the bound $\log|\mathcal{N}_{t,\zeta}| \le k\log(1+2t/\zeta)$ 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 $\|\Delta - \Delta'\| \le \zeta$, we have

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

where the last line is due to the condition (3.3). Thus,

$$|G_m(\Delta)| \le |G_m(\Delta')| + 2\bar{\lambda}\zeta.$$

Taking supremum over both sides of the inequality, we obtain

$$\sup_{\|\Delta\|\le t} |G_m(\Delta)| \le \max_{\Delta\in\mathcal{N}_{t,\zeta}} |G_m(\Delta)| + 2\bar{\lambda}\zeta. \tag{8.1}$$

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

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

By Lemma 8.1, we have

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

A union bound argument leads to

$$\mathbb{P}\left(\max_{\Delta\in\mathcal{N}_{t,\zeta}} |G_m(\Delta)| > x\right) \le 2\exp\left(-\frac{2mx^2}{\bar{\lambda}^2 t^2} + k\log\left(1 + \frac{2t}{\zeta}\right)\right). \tag{8.2}$$

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 (8.1), we have

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

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

Proof of Theorem 8.1. Define

$$\begin{aligned} L_m(u) &= \frac{1}{m}\sum_{i=1}^{m}\left(|\gamma_i - a_i^T u| - |z_i|\right) \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(|a_i^T(u^* - u) + z_i| - |z_i|\right), \end{aligned}$$

and $K_m(u) = \frac{1}{m}\sum_{i=1}^{m}\left(|b_i + a_i^T(u^* - u) + z_i| - |z_i|\right)$. It is easy to see that

$$\sup_u |L_m(u) - K_m(u)| \le \frac{1}{m}\sum_{i=1}^{m}|b_i|. \tag{8.3}$$

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

$$\begin{aligned} K_m(u) &= \frac{1}{m}\sum_{i=1}^{m}\left(|\widetilde{b}_i + \widetilde{a}_i^T(u^* - u) + \widetilde{z}_i| - |\widetilde{z}_i|\right), \\ L_m(u) &= \frac{1}{m}\sum_{i=1}^{m}\left(|\widetilde{a}_i^T(u^* - u) + \widetilde{z}_i| - |\widetilde{z}_i|\right). \end{aligned}$$

Let $\widetilde{A} \in \mathbb{R}^{m\times k}$ be the matrix whose ith row is $\widetilde{a}_i$. 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 $\|\widehat{u} - u^*\| \ge t$, we must have

$$\inf_{\|u-u^*\|\ge t} K_m(u) \le K_m(u^*).$$

By the convexity of $K_m(u)$, we can replace $\|u - u^*\| \ge t$ by $\|u - u^*\| = t$ and the above inequality still holds. By (8.3), we have $K_m(u^*) \le \frac{1}{m}\sum_{i=1}^{m}|b_i|$, and therefore $\inf_{\|u-u^*\|=t} K_m(u) \le \frac{1}{m}\sum_{i=1}^{m}|b_i|$. Since

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

we then have

$$\inf_{\|u-u^*\|=t} L(u) \le \sup_{\|u-u^*\|=t} |L_m(u) - L(u)| + 2\frac{1}{m}\sum_{i=1}^m |b_i|. \tag{8.4}$$

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 $x \ge 0$,

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

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

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

where the last inequality is by (3.2). Together with (8.4), we have

$$\underline{\lambda}(1-\varepsilon)t \le \sup_{\|u-u^*\|=t} |L_m(u) - L(u)| + 2\frac{1}{m}\sum_{i=1}^m |b_i|.$$

Set $t = \frac{4\frac{1}{m}\sum_{i=1}^m |b_i|}{\underline{\lambda}(1-\varepsilon)}$, and we then have

$$\mathbb{P}\left(\|\widehat{u}-u\| \ge t\right) \le \mathbb{P}\left(\sup_{\|u-u^*\|=t} |L_m(u) - L(u)| \ge \underline{\lambda}(1-\varepsilon)t/2\right). \tag{8.5}$$

Since the condition (3.3) continues to hold with A replaced by $\widetilde{A}$, we can apply Lemma 8.6 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 (8.5), $\|\widehat{u}-u^*\| < t$ with high probability. □

8.3. *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

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

and we will analyze the second term on the right hand side of the inequality above via a discretization argument. 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 $\log|\mathcal{N}_\zeta| \leq n\log(1+2/\zeta)$ 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

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

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

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

Set $\zeta = 1/3$, and we then have

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

Lemma 8.3 together with a union bound argument leads to

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

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 $\sup_{\|\Delta\|=1}\frac{1}{p}\sum_{j=1}^{p}|a_j^T\Delta|^2 = \|A\|_{\mathrm{op}}^2/p \lesssim 1 + n/p$ 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 $\widehat{\theta} = X^T u^*$. By Theorem 3.1 and Lemma 4.1, we know that $\widetilde{u} = u^*$ with high probability, and therefore $\widetilde{\theta} = X^T\widetilde{u} = X^T u^* = \widehat{\theta}$. □

Now we state the proof of Lemma 4.2. Note that Condition A is obvious, and we only need to prove Condition B. We present the proofs of (3.2) and (3.3) separately.

Proof of (3.2) 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

$$\begin{aligned}\inf_{\|\Delta\|=1}f(W,X,\Delta) &\geq \inf_{\|\Delta\|=1}\mathbb{E}f(W,X,\Delta)-\sup_{\|\Delta\|=1}|f(W,X,\Delta)-\mathbb{E}f(W,X,\Delta)| \\ &\geq \inf_{\|\Delta\|=1}\mathbb{E}f(W,X,\Delta) && (8.6)\\ &\quad -\sup_{\|\Delta\|=1}|f(W,X,\Delta)-\mathbb{E}(f(W,X,\Delta)|X)| && (8.7)\\ &\quad -\sup_{\|\Delta\|=1}|g(X,\Delta)-\mathbb{E}g(X,\Delta)|\,. && (8.8)\end{aligned}$$

We will analyze the three terms above separately.

Analysis of (8.6). For any Δ such that $\|\Delta\|=1$, we have

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

where the last inequality is by Lemma 8.4. 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 $1/2\leq\|W\|^2\leq 2$,

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

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 $1/2 \le \|W\|^2 \le 2$. By Lemma 8.2, we have

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

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

$$\inf_{\|\Delta\|=1, \max_{1\le i\le n}|\Delta_i|\le\delta_0} \mathbb{E}f(W,X,\Delta) \gtrsim 1. \tag{8.9}$$

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

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

and we can conclude that

$$\inf_{\|\Delta\|=1,\max_{1\leq i\leq n}|\Delta_i|\geq\delta_0} \mathbb{E}f(W,X,\Delta) \gtrsim 1. \tag{8.10}$$

In the end, we combine the two cases (8.9) and (8.10), and we obtain the conclusion that $\inf_{\|\Delta\|=1}\mathbb{E}f(W,X,\Delta)\gtrsim 1$.

Analysis of (8.7). We shorthand the conditional expectation operator $\mathbb{E}(\cdot|X)$ by $\mathbb{E}^X$. Let $\widetilde{W}$ be an independent copy of W, and we first bound the moment generating function via a standard symmetrization argument. For any $\lambda>0$,

$$\begin{aligned}
& \mathbb{E}^X \exp\left(\lambda \sup_{\|\Delta\|=1}\left|f(W,X,\Delta)-\mathbb{E}^X f(W,X,\Delta)\right|\right) \\
\leq\ & \mathbb{E}^X \exp\left(\lambda\mathbb{E}^{X,W}\sup_{\|\Delta\|=1}\left|f(W,X,\Delta)-f(\widetilde{W},X,\Delta)\right|\right) \\
\leq\ & \mathbb{E}^X \exp\left(\lambda \sup_{\|\Delta\|=1}\left|f(W,X,\Delta)-f(\widetilde{W},X,\Delta)\right|\right) \\
=\ & \mathbb{E}^X \exp\left(\lambda \sup_{\|\Delta\|=1}\left|\frac{1}{p}\sum_{j=1}^p \varepsilon_j\left(\left|\sum_{i=1}^n \psi(W_j^Tx_i)\Delta_i\right|-\left|\sum_{i=1}^n\psi(\widetilde{W}_j^Tx_i)\Delta_i\right|\right)\right|\right) \\
\leq\ & \mathbb{E}^X \exp\left(2\lambda \sup_{\|\Delta\|=1}\left|\frac{1}{p}\sum_{j=1}^p \varepsilon_j\left|\sum_{i=1}^n \psi(W_j^Tx_i)\Delta_i\right|\right|\right),
\end{aligned} \tag{8.11}$$

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

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

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 (8.11), we obtain the bound

$$\begin{aligned}
& \mathbb{E}^X \exp\left(\lambda \sup_{\|\Delta\|=1} \left| f(W, X, \Delta) - \mathbb{E}^X f(W, X, \Delta) \right| \right) \\
\leq \; & \mathbb{E}^X \exp\left(4\lambda \max_{\Delta \in \mathcal{N}} |F(\varepsilon, W, X, \Delta) - \bar{F}(\varepsilon, X, \Delta)| + 4\lambda \max_{\Delta \in \mathcal{N}} |\bar{F}(\varepsilon, X, \Delta)| \right) \\
\leq \; & \frac{1}{2} \sum_{\Delta \in \mathcal{N}} \mathbb{E}^X \exp\left(4\lambda |F(\varepsilon, W, X, \Delta) - \bar{F}(\varepsilon, X, \Delta)|\right) \qquad (8.12) \\
& + \frac{1}{2} \sum_{\Delta \in \mathcal{N}} \mathbb{E}^X \exp\left(4\lambda |\bar{F}(\varepsilon, X, \Delta)|\right). \qquad (8.13)
\end{aligned}$$

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

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

where the last inequality holds under the event E. By Lemma 8.3, we have for any X such that E holds,

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

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

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

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

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

With an application of Hoeffding-type inequality (Lemma 5.9 of Vershynin (2010)), we have

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

Note that we can use the same constant C_1 by making its value sufficiently large. Plug the two moment generating function bounds into (8.12) and (8.13), 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 (8.7), 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 (8.8). 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

$$\sup_{\|\Delta\|=1} |g(X,\Delta) - \mathbb{E}g(X,\Delta)| \leq (1-\zeta)^{-1} \max_{\Delta\in\mathcal{N}_\zeta} |g(X,\Delta) - \mathbb{E}g(X,\Delta)| \tag{8.14}$$

$$2\zeta(1-\zeta)^{-1}\mathbb{E}g(X,\Delta). \tag{8.15}$$

To bound (8.14), we will use Lemma 8.3 together with a union bound argument. For any $X, \widetilde{X}$, we have

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

Therefore, by Lemma 8.3,

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

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 (8.15), we have

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

Combining the bounds for (8.14) and (8.15), 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 (8.6), (8.7) and (8.8), and we obtain the desired conclusion as long as n^2/p and n/d are sufficiently small. □

To prove (3.3) of Lemma 4.2, we establish the following stronger result.

Lemma 8.7. *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*

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

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

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

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

$$1 \lesssim \lambda_{\min}(G) \le \lambda_{\max}(G) \lesssim 1, \tag{8.16}$$

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

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

We then have

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

By Markov's inequality,

$$\|G - \widetilde{G}\|_{\mathrm{op}}^2 \lesssim \frac{n^2}{p}, \tag{8.17}$$

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,

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

By Lemma 8.4 and a union bound argument, we have

$$\mathbb{P}\left(\max_{1\le i\le n}|\widetilde{G}_{ii} - \bar{G}_{ii}| > 3\sqrt{\frac{t}{d}} + 3\frac{t}{d}\right) \le 2ne^{-t},$$

for any $t > 0$. Choosing $t \asymp \log n$, we obtain the bound

$$\max_{1\le i\le n} |\widetilde{G}_{ii} - \bar{G}_{ii}| \lesssim \sqrt{\frac{\log n}{d}}, \tag{8.18}$$

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

$$\begin{aligned}
\widetilde{G}_{il} &= \mathbb{E}\left(\psi(W^T\bar{x}_i)\psi(W^T\bar{x}_l)|X\right) && (8.19)\\
&\quad +\mathbb{E}\left((\psi(W^Tx_i)-\psi(W^T\bar{x}_i))\psi(W^T\bar{x}_l)|X\right) && (8.20)\\
&\quad +\mathbb{E}\left(\psi(W^T\bar{x}_i)(\psi(W^Tx_l)-\psi(W^T\bar{x}_l))|X\right) && (8.21)\\
&\quad +\mathbb{E}\left((\psi(W^Tx_i)-\psi(W^T\bar{x}_i))(\psi(W^Tx_l)-\psi(W^T\bar{x}_l))|X\right). && (8.22)
\end{aligned}$$

For first term on the right hand side of (8.19), 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

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

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 $\sup_{|\rho|\le 0.2} |f''(\rho)| \lesssim 1$. Therefore, as long as $|\bar{x}_i^T\bar{x}_l|/d \le 1/5$,

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

for some constant $C_1 > 0$. By Lemma 8.5, we know that $\max_{i\neq l}|\bar{x}_i^T\bar{x}_l|/d \lesssim \sqrt{\frac{\log n}{d}} \le 1/5$ with high probability, which then implies

$$\sum_{i\neq l}\left(\mathbb{E}\left(\psi(W^T\bar{x}_i)\psi(W^T\bar{x}_l)|X\right) - \bar{G}_{il}\right)^2 \le C_1\sum_{i\neq l}\left|\frac{\bar{x}_i^T\bar{x}_l}{d}\right|^4. \tag{8.23}$$

The term on the right hand side can be bounded by

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

By Lemma 8.4, $\frac{d}{\min_{1\le i\le n}\|x_i\|^2} \lesssim 1$ with high probability. By integrating out the probability tail bound of $|x_i^Tx_l|$ given in Lemma 8.5, 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 (8.20) and (8.21). We can write (8.20) as

$$\mathbb{E}\left[\psi(W^T\bar{x}_l)\psi'(W^T\bar{x}_i)W^T(x_i-\bar{x}_i)|X\right] \tag{8.24}$$

$$+\frac{1}{2}\mathbb{E}\left[\psi(W^T\bar{x}_i)\psi''(t_i)|W^T(x_i-\bar{x}_i)|^2|X\right], \tag{8.25}$$

where t_i is some random variable between W^Tx_i and $W^T\bar{x}_i$. The first term (8.24) 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 $\sup_{|\rho|\le 0.2}|g'(\rho)|\lesssim 1$, and thus

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

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

$$\begin{aligned}
&\sum_{i\neq l}\left(\mathbb{E}\left[\psi(W^T\bar{x}_l)\psi'(W^T\bar{x}_i)W^T(x_i-\bar{x}_i)|X\right]\right)^2\\
\lesssim\ &\sum_{i\neq l}\left|\frac{\|x_i\|}{\sqrt{d}}-1\right|^2\left|\frac{x_i^Tx_l}{d}\right|^2\\
\lesssim\ &n\sum_{i=1}^n\left|\frac{\|x_i\|}{\sqrt{d}}-1\right|^4+\sum_{i\neq l}\left|\frac{x_i^Tx_l}{d}\right|^4.
\end{aligned}\tag{8.26}$$

By integrating out the probability tail bound of Lemma 8.4, we have $\mathbb{E}\left|\frac{\|x_i\|}{\sqrt{d}}-1\right|^4\lesssim d^{-2}$. We also have $\mathbb{E}\left|\frac{x_i^Tx_l}{d}\right|^4\lesssim d^{-2}$. 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 (8.25), we observe that

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

where the inequality above is by $\sup_x|\psi(x)|\le 1$ and $\sup_x|\psi''(x)|\le 2$. 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 (8.21), we conclude that the contributions of (8.20) and (8.21) 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 (8.22) is negligible. Note that

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

where the last inequality is by $\mathbb{E}\left(|W^T\bar{x}_i||W^T\bar{x}_l||X\right) \le \frac{1}{2}\mathbb{E}(|W^T\bar{x}_i|^2 + |W^T\bar{x}_l|^2|X) = 1$. 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 (8.22) is bounded by $O\left(\frac{n^2}{d^2}\right)$ with high probability by Markov's inequality.

Combining the analyses of (8.19), (8.20), (8.21) and (8.22), 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 (8.17) and (8.18), we obtain the desired bound for $\|G-\bar{G}\|_{\text{op}}$.

To prove the last conclusion (8.16), it suffices to show $1 \lesssim \lambda_{\min}(\bar{G}) \le \lambda_{\max}(\bar{G}) \lesssim 1$. The lower bound of the smallest eigenvalue is because $\lambda_{\min}(\bar{G}) \ge \mathbb{E}|\psi(Z)|^2 - |\mathbb{E}\psi'(Z)|^2 \gtrsim 1$. The largest eigenvalue can be bounded by

$$\begin{aligned}
\lambda_{\max}(\bar{G}) &\le \left(\mathbb{E}|\psi(Z)|^2 - |\mathbb{E}\psi'(Z)|^2\right) + |\mathbb{E}\psi'(Z)|^2 \max_{\|v\|=1}\sum_{i=1}^n\sum_{l=1}^n v_i v_l \frac{x_i^T x_l}{\|x_i\|\|x_l\|} \\
&\lesssim 1 + \max_{\|v\|=1}\sum_{i=1}^n\sum_{l=1}^n v_i v_l \frac{x_i^T x_l}{d} \\
&= 1 + \|X\|_{\text{op}}^2/d \\
&\lesssim 1 + \frac{n}{d},
\end{aligned}$$

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 $\widetilde{\theta} = \widetilde{X}^T\widetilde{u} = \widetilde{X}^T u^* = \widehat{\theta}$. □

8.4. Proof of Theorem 5.1

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

Lemma 8.8. *Consider independent $W_1, ..., W_p \sim N(0, d^{-1}I_d)$, $x_1, ..., x_n \sim N(0, I_d)$, and $\beta_1, ..., \beta_p \sim N(0,1)$. 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 assume that d/n and p/n are sufficiently large, we will also have

$$0.09 \le \lambda_{\min}(H) \le \lambda_{\max}(H) \lesssim 1, \tag{8.27}$$

with high probability.

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

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

We then have

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

By Markov's inequality,

$$\|H - \widetilde{H}\|_{\rm op}^2 \lesssim \frac{n^2}{pd} + \frac{n}{p}, \tag{8.28}$$

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

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

Similar to (8.18), Lemma 8.4 and a union bound argument imply

$$\max_{1\le i\le n} |\widetilde{H}_{ii} - \bar{H}_{ii}| \lesssim \sqrt{\frac{\log n}{d}} + \frac{\log n}{d}, \tag{8.29}$$

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 $i \ne l$, we have

$$\widetilde{H}_{il} = \frac{\bar{x}_i^T \bar{x}_l}{d}\mathbb{E}\left(\psi'(W^T\bar{x}_i)\psi'(W^T\bar{x}_l)\big|X\right) \tag{8.30}$$

$$+\frac{x_i^T x_l}{d}\mathbb{E}\left(\psi'(W^T x_i)\psi'(W^T x_l) - \psi'(W^T\bar{x}_i)\psi'(W^T\bar{x}_l)\big|X\right) \tag{8.31}$$

$$+\left(\frac{\|x_i\|\|x_l\|}{d} - 1\right)\frac{\bar{x}_i^T\bar{x}_l}{d}\mathbb{E}\left(\psi'(W^T\bar{x}_i)\psi'(W^T\bar{x}_l)\big|X\right). \tag{8.32}$$

For the first term on the right hand side of (8.30), we observe that $\frac{\bar{x}_i^T \bar{x}_l}{d} \mathbb{E}\left(\psi'(W^T \bar{x}_i)\psi'(W^T \bar{x}_l)\big|X\right)$ is a function of $\frac{\bar{x}_i^T \bar{x}_l}{d}$, and thus we can write

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

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 (8.23), 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 (8.31), we note that

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

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 (8.31) 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 (8.26).

Finally, since (8.32) 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 (8.23). For the first term on the right hand side of the above inequality, we use Lemma 8.5 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 (8.30), (8.31), and (8.32), 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 (8.28) and (8.29), we obtain the desired bound for $\|H-\bar{H}\|_{\mathrm{op}}$.

The last conclusion (8.27) follows a similar argument used in the proof of Lemma 8.7. □

Now we are ready to prove Theorem 5.1.

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

$$\max_{1\leq j\leq p}|\beta_j(0)| \leq 2\sqrt{\log p}, \tag{8.33}$$

$$\max_{k\in\{1,2,3\}}\frac{1}{p}\sum_{j=1}^{p}|\beta_j(0)|^k \lesssim 1, \tag{8.34}$$

$$\sum_{i=1}^{n}\|x_i\|^2 \leq 2nd, \tag{8.35}$$

$$\max_{1\leq i\leq n}\|x_i\| \lesssim \sqrt{d}, \tag{8.36}$$

$$\max_{1\leq l\leq n}\sum_{i=1}^{n}\left|\frac{x_i^Tx_l}{d}\right| \lesssim 1+\frac{n}{\sqrt{d}}, \tag{8.37}$$

$$\|u(0)\| \leq \sqrt{n\log p}. \tag{8.38}$$

The bound (8.33) is a consequence of a standard Gaussian tail inequality and a union bound argument. The second bound (8.34) 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 (8.35) and (8.36) derived from Lemma 8.4 and a union bound. For (8.37), we first have the bound

$$\begin{aligned}\max_{1\leq l\leq n}\sum_{i=1}^{n}\left|\frac{x_i^Tx_l}{d}\right| &\leq \max_{1\leq l\leq n}\frac{\|x_l\|^2}{d}+\max_{1\leq l\leq n}\frac{\|x_l\|}{d}\sum_{i\in[n]\setminus\{l\}}\left|\frac{x_i^Tx_l}{\|x_l\|}\right| \\ &\lesssim 1+\frac{1}{\sqrt{d}}\max_{1\leq l\leq n}\sum_{i\in[n]\setminus\{l\}}\left|\frac{x_i^Tx_l}{\|x_l\|}\right|,\end{aligned}$$

where the last inequality is by (8.36). Since

$$\sum_{i\in[n]\setminus\{l\}}\left|\frac{x_i^Tx_l}{\|x_l\|}\right| \leq \sqrt{n}\sqrt{\sum_{i\in[n]\setminus\{l\}}\left|\frac{x_i^Tx_l}{\|x_l\|}\right|^2},$$

and $\sum_{i\in[n]\setminus\{l\}}\left|\frac{x_i^Tx_l}{\|x_l\|}\right|^2\sim\chi^2_{n-1}$, we can then use Lemma 8.4 and a union bound and obtain

$$\max_{1\le l\le n}\sum_{i\in[n]\setminus\{l\}}\left|\frac{x_i^Tx_l}{\|x_l\|}\right|\lesssim n,$$

with high probability, which then leads to (8.37). To obtain the last bound (8.38), we note that $\mathbb{E}|u_i(0)|^2=\mathbb{E}\mathsf{Var}(u_i(0)|X)\le 1$, which then implies (8.38) by Markov's inequality.

Now we are ready to prove the main result. It suffices to show the following to claims are true.

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

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

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

Proof of Claim A. 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

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

where we have used (8.33), (8.35) and (8.38) in the above inequalities. Similarly, we also have

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

where we have used (8.38). Hence, (5.1) holds for $t = k+1$, and Claim A is true.

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

$$\begin{aligned}
& u_i(k+1)-u_i(k) \\
= \; & \frac{1}{\sqrt{p}}\sum_{j=1}^{p}\beta_j(k+1)\left(\psi(W_j(k+1)^Tx_i)-\psi(W_j(k)^Tx_i)\right) \\
& +\frac{1}{\sqrt{p}}\sum_{j=1}^{p}(\beta_j(k+1)-\beta_j(k))\psi(W_j(k)^Tx_i) \\
= \; & \frac{1}{\sqrt{p}}\sum_{j=1}^{p}\beta_j(k+1)(W_j(k+1)-W_j(k))^Tx_i\psi'(W_j(k)^Tx_i) \\
& +\frac{1}{\sqrt{p}}\sum_{j=1}^{p}(\beta_j(k+1)-\beta_j(k))\psi(W_j(k)^Tx_i) \\
& +\frac{1}{2\sqrt{p}}\sum_{j=1}^{p}\beta_j(k+1)|(W_j(k+1)-W_j(k))^Tx_i|^2\psi''(\xi_{ijk}) \\
= \; & \gamma\sum_{l=1}^{n}(H_{il}(k)+G_{il}(k))(y_l-u_l(k))+r_i(k),
\end{aligned}$$

where

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

and

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

We need to understand the eigenvalues of $G(k)$ and $H(k)$, and bound the absolute value of $r_i(k)$. By its definition, it is easy to see that $\lambda_{\min}(G(k)) \geq 0$. To analyze $\lambda_{\max}(G(k))$, we note that

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

Thus, by (8.36),

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

Then, we have

$$\|G(k) - G(0)\|_{\text{op}} \leq \max_{1\leq l\leq n}\sum_{i=1}^{n}|G_{il}(k) - G_{il}(0)| \lesssim \frac{n^2\log p}{\sqrt{p}}, \tag{8.39}$$

which leads to the bound $\lambda_{\max}(G(k)) \lesssim 1 + \frac{n^2 \log p}{\sqrt{p}}$ for all k by Lemma 8.7. For the matrix $H(k)$, we show its eigenvalues can be controlled by those of $H(0)$ We have

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

Thus, by (8.36) and (8.37),

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

Then, we have

$$\|H(k) - H(0)\|_{\text{op}} \leq \max_{1\leq l\leq n}\sum_{i=1}^{n} |H_{il}(k) - H_{il}(0)| \lesssim \frac{n\log p}{\sqrt{p}}\left(1 + \frac{n}{\sqrt{d}}\right). \tag{8.40}$$

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

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

where we have used (8.34), (8.35), (8.36) and (8.38) in the above inequalities. This leads to the bound

$$\|r(k)\| = \sqrt{\sum_{i=1}^{n} |r_i(k)|^2} \lesssim \frac{\gamma^2 n^2 \sqrt{\log p}}{\sqrt{p}} \|y - u(k)\|. \tag{8.41}$$

Given the relation that

$$u(k+1) - u(k) = \gamma(H(k) + G(k))(y - u(k)) + r(k),$$

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 $\lambda_{\min}(G(k)) \geq 0$, the operator norm bound (8.40), and Lemma 8.8, we have

$$-2\gamma(y - u(k))^T(H(k) + G(k))(y - u(k)) \leq -\frac{\gamma}{6}\|y - u(k)\|^2. \tag{8.42}$$

The bound (8.41) implies

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

and together with $\lambda_{\max}(H(0)) \lesssim 1$, (8.40), and the bound $\lambda_{\max}(G(k)) \lesssim 1 + \frac{n^2 \log p}{\sqrt{p}}$, 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

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

Together with the bound (8.42), 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 B is true. The proof is complete. □

8.5. Proofs of Theorem 5.2 and Theorem 5.3

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 8.6. With an additional union bound argument applied to (8.2), Lemma 8.6 can be extended to Z_j for all $j \in [p]$ with an additional assumption that $\frac{\log p}{d}$ is sufficiently small. Then, by the same argument in the proof of 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 8.1. 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 8.1, 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

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

where the last inequality is by $\sum_{i=1}^{n} \|x_i\|^2 \lesssim nd$ due to Lemma 8.4, and $\|u(0)\|^2 \lesssim n$ due to Markov's inequality and $\mathbb{E}|u_i(0)|^2 = \mathbb{E}\mathsf{Var}(u_i(0)|X) \le 1$. By Theorem 8.1 and Lemma 8.7, we have $\frac{1}{p}\|\widetilde{\beta} - \widehat{\beta}\|^2 \lesssim \frac{n^2 \log p}{p(1-\varepsilon)}$, 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 3.1. It suffices to check Condition A and Condition B for the design matrix $\psi(X^T\widetilde{W}^T) = \psi(X^T\widehat{W}^T)$. To check Condition 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

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

In other words, $(W(t), \beta(t))$ and $(\check{W}(t), \check{\beta}(t))$ only differ in terms of the initialization. We also define $\check{u}_i(t) = \frac{1}{\sqrt{p}}\sum_{j=1}^p \check{\beta}_j(t)\psi(\check{W}_j(t)^T x_i)$. It is easy to see that

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

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

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

and

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

we then have $\check{W}_j(k+1) = \delta_j W_j(k+1)$ and $\check{\beta}_j(k+1) = \delta_j\beta_j(k+1)$. A mathematical induction argument leads to $\check{W}_j(t) = \delta_j W_j(t)$ and $\check{\beta}_j(t) = \delta_j\beta_j(t)$ for all $t \geq 1$. Since $(\check{W}(0), \check{\beta}(0))$ and

$(W(0), \beta(0))$ have the same distribution, we can conclude that $(\check{W}(t), \check{\beta}(t))$ and $(W(t), \beta(t))$ also have the same distribution. Therefore, Condition 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 8.4. 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. By (8.39) and the result of (8.7), we also have

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

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

References

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

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.

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.

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.

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

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

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

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.

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*.

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.

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