Adversarial Training with Generated Data in High-Dimensional Regression: An Asymptotic Study

Yue Xing [1]

Abstract

In recent years, studies such as (Carmon et al., 2019; Gowal et al., 2021; Xing et al., 2022) have demonstrated that incorporating additional real or generated data with pseudo-labels can enhance adversarial training through a two-stage training approach. In this paper, we perform a theoretical analysis of the asymptotic behavior of this method in high-dimensional linear regression. While a double-descent phenomenon can be observed in ridgeless training, with an appropriate $\mathcal{L}_2$ regularization, the two-stage adversarial training achieves a better performance. Finally, we derive a shortcut cross-validation formula specifically tailored for the two-stage training method.

1. Introduction

The development of machine learning and deep learning methods has led to breakthrough performance in various applications. However, recent studies, e.g., (Goodfellow et al., 2014), observe that these models are vulnerable when the data are perturbed by adversaries. Attacked inputs can be imperceptibly different from clean inputs to humans but can cause the model to make incorrect predictions.

To defend against adversarial attacks, adversarial training is a popular and promising way to improve the adversarial robustness of modern machine learning models. Adversarial training first generates attacked samples, then calculates the gradient of the model based on these augmented data. Such a procedure can make the model less susceptible to adversarial attacks in real-world situations.

There are fruitful results in the theoretical justification and methodology development in adversarial training. Among various research directions, one interesting aspect is to improve adversarial training with extra unlabeled data. Recent works successfully demonstrate great improvements in the adversarial robustness with additional unlabeled data. For example, (Xing et al., 2021), show that additional external real data help improve adversarial robustness; (Gowal et al., 2021; Wang et al., 2023) use synthetic data to improve the adversarial robustness and achieve the highest 65% to 70% adversarial testing accuracy for CIFAR-10 dataset under AutoAttack (AA) in (Croce et al., 2020)[1].

A recent study (Xing et al., 2022) reveals that adversarial training gains greater benefits from unlabeled data than clean (natural) training. The key observation is that adversarially robust models rely on the conditional distribution of the response given the features $(Y|X)$ and the marginal distribution of the features (X). In contrast, clean training only depends on $Y|X$ in their study. As a result, adversarial training can benefit more than clean training from unlabeled data.

Besides adversarial training, high dimensional statistics is another important field of traditional machine learning to solve real-world problems from genomics, neuroscience to image processing. While many studies focus on obtaining a better performance via regularization, one surprising phenomenon in this field is the double descent phenomenon (Belkin et al., 2019; Hastie et al., 2019), which refers to a U-shaped curve in the test error as a function of the model complexity, together with a second descent phase occurring in the over-parameterized regime. This phenomenon challenges the conventional wisdom that increasing model complexity always leads to over-fitting. It provides significant implications for designing and analyzing machine learning algorithms in high-dimensional settings.

Given the substantial achievements in high-dimensional statistics, this paper aims to extend the analysis of (Xing et al., 2022) to a high-dimensional regression setup, in which both the data dimension d and the sample size of the labeled data n_1 increase and $d/n_1 \to \gamma$ asymptotically. Although (Xing et al., 2022) provides a theoretical explanation for the benefits of unlabeled data in the large sample regime ($n_1 \gg d$), the asymptotic behavior of the two-stage method

[1]Department of Statistics and Probability, Michigan State University, United States. Correspondence to: Yue Xing <xingyue1@msu.edu>.

2^{nd} AdvML Frontiers workshop at 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]https://robustbench.github.io

in other scenarios remains unclear.

Our contributions are summarized as follows:

- We derived the asymptotic convergence of the two-stage adversarial training when $d/n_1 \to \gamma$ for some constant $\gamma > 0$. (Section 3.1).
- It is observed that a proper ridge penalty in the clean training stage benefits the two-stage method. However, the optimal ridge penalty for the clean estimate in the first stage of (Xing et al., 2022) differs from the one yielding the best clean performance. We conjecture that this discrepancy arises from the change in the error decomposition from clean training to two-stage adversarial training. To facilitate more efficient hyper-parameter tuning, we propose adaptations to existing cross validation (CV) methods, improving the time-consuming vanilla CV approach (Sections 3.2 and 3.3).

1.1. Related Works

Below is a summary of related works in adversarial training, high-dimensional statistics, and cross validation.

Adversarial Training. There are many studies in the area of adversarial training. Some studies, e.g., (Goodfellow et al., 2014; Zhang et al., 2019; Wang et al., 2019b; Cai et al., 2018; Zhang et al., 2020a; Carmon et al., 2019; Gowal et al., 2021), work in methodology. Theoretical investigations have also been conducted from different perspectives. For instance, Chen et al. (2020); Javanmard et al. (2020); Taheri et al. (2021); Yin et al. (2018); Raghunathan et al. (2019); Najafi et al. (2019); Min et al. (2020); Hendrycks et al. (2019); Dan et al. (2020); Wu et al. (2020); Deng et al. (2021) study the statistical properties of adversarial training; Sinha et al. (2018); Wang et al. (2019a); Xiao et al. (2022) study the optimization perspective; Gao et al. (2019); Zhang et al. (2020b); Zhang and Li (2023); Mianjy and Arora (2022); Lv and Zhu (2021); Xiao et al. (2021) work on deep learning.

Double Descent and High-Dimensional Statistics. Double descent phenomenon is an observation in the learning curves of machine learning models. It describes the behavior of the generalization gap, i.e., the difference between the model performance on the training data and testing data. In a typical learning curve, the generalization error decreases and then increases with larger model complexity. However, in the double descent phenomenon, after the first decrease-increase pattern, the error decreases again when further enlarging the model complexity in the over-fitting regime. This non-monotonic behavior of the learning curve has been observed in various machine learning settings. Comprehensive investigations into the double descent phenomenon can be found in (Belkin et al., 2019; Hastie et al., 2019; Ba et al., 2020; d'Ascoli et al., 2020; Adlam and Pennington, 2020; Liu et al., 2021; Rocks and Mehta, 2022).

Cross Validation. Cross validation (CV) is a resampling procedure used to evaluate the performance of machine learning models. This paper mainly considers leave-one-out CV. For leave-one-out CV, it trains the model using all-but-one samples and repeats this process so that every sample is left in the estimation once. The final model performance is then averaged across all the models. The model can generalize better to new data by optimizing the hyperparameters in the model, e.g., regularization, through CV.

However, although a leave-one-out CV is an effective method for selecting hyperparameters, it is time-consuming by its design. Consequently, some studies propose shortcut formulas for the leave-one-out CV to reuse some terms when estimating the model using different data. Studies related to CV can be found in (Stone, 1978; Picard and Cook, 1984; Shao, 1993; Browne, 2000; Berrar, 2019).

2. Model Setup

In this section, we present the data generation model and the two-stage adversarial training framework.

Data generation model. We assume that the attributes $X \sim N(\mathbf{0}, \mathbf{\Sigma})$ with covariance matrix $\mathbf{\Sigma} = \boldsymbol{I}_d$, and the response Y satisfies $Y = X^\top \boldsymbol{\theta}_0 + \varepsilon$ for $\|\boldsymbol{\theta}_0\| = r = O(1)$ and a Gaussian noise ε with $Var(\varepsilon) = \sigma^2$.

Two-stage adversarial training. There are two stages in this training framework. In the first stage, we utilize n_1 i.i.d. labeled samples, i.e., $(\boldsymbol{x}_i, y_i)$ for $i = 1, \ldots, n_1$. We consider the scenario where $d \asymp n_1$. The first stage solves the following clean training problem

$$\frac{1}{n_1}\sum_{i=1}^{n_1}(\boldsymbol{x}_i^\top \boldsymbol{\theta} - y_i)^2 + \lambda\|\boldsymbol{\theta}\|^2 \tag{2.1}$$

and obtain the clean estimate $\widehat{\boldsymbol{\theta}}_0(\lambda)$.

In the second stage, we use the trained model $\widehat{\boldsymbol{\theta}}_0(\lambda)$ to generate a pseudo response for a set of unlabeled data, i.e.,

$$\widehat{y}_i = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\theta}}_0(\lambda) + \varepsilon_i$$

for $i = n_1 + 1, \ldots, n_1 + n_2$. In this paper, we consider the scenario where $n_2 = \infty$. We also assume σ^2 is known and ε_i are generated from $N(0, \sigma^2)$. Finally we use the extra data with pseudo response to do adversarial training and minimize the following loss w.r.t $\boldsymbol{\theta}$:

$$\frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}\sup_{\boldsymbol{z}\in\mathcal{B}_2(\boldsymbol{x}_i,\epsilon)}(\boldsymbol{z}^\top\boldsymbol{\theta} - \widehat{y}_i)^2. \tag{2.2}$$

Denote the final solution as $\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)$.

Remark 1. *The two-stage method in this paper is slightly different from the original one in (Gowal et al., 2021; Xing et al., 2022). We only utilize the generated data in the second stage. This simplifies the theoretical analysis. In addition, when $d/n_1 = \gamma$ is a large constant, we empirically observe that the two-stage method is better than an adversarial training with only labeled data, i.e., the right of Figure 1.*

Remark 2. *Our initial trial indicates that adding additional regularization in equation (2.2) does not help much. Thus, we only inject a penalty in the clean training stage.*

Expected Adversarial Risk Under the model assumption of (X, Y), the population adversarial risk for any given estimate $\boldsymbol{\theta}$ becomes

$$\begin{aligned} R_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}_0) &= \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}^2 \\ &\quad + 2c_0\epsilon\|\boldsymbol{\theta}\|\sqrt{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}^2 + \sigma^2} + \epsilon^2\|\boldsymbol{\theta}\|^2, \end{aligned}$$

where $\|\cdot\|$ is the $\mathcal{L}_2$ norm, and $c_0 = \sqrt{2/\pi}$ is derived from the exact distribution of (X, Y). We rewrite $R_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ as $R_\epsilon(\boldsymbol{\theta})$ for simplicity when no confusion arises.

Remark 3. *One can denote $\boldsymbol{\theta}_\epsilon = \arg\min_{\boldsymbol{\theta}} R_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ as the best robust model. However, from $R_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$, we are interested in $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}$ and $\|\boldsymbol{\theta}\|$ rather than $\|\boldsymbol{\theta} - \boldsymbol{\theta}_\epsilon\|$.*

Based on (Xing et al., 2021), when an estimate $\boldsymbol{\theta} \to \boldsymbol{\theta}_\epsilon$, the excess adversarial risk $R_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}_0) - R_\epsilon(\boldsymbol{\theta}_\epsilon, \boldsymbol{\theta}_0)$ can be approximated by a function of $\boldsymbol{\theta} - \boldsymbol{\theta}_\epsilon$. However, when $\boldsymbol{\theta} - \boldsymbol{\theta}_\epsilon$ diverges in the high-dimensional setup, such an approximation leads to a large error.

3. Analyzing the Two-Stage Adversarial Training Framework

This section presents the main theoretical results and simulation studies. We first demonstrate the main theory of the convergence of the two-stage method in Section 3.1, take different λ under different attack strength ϵ in Section 3.2, and finally introduce a CV method in Section 3.3.

3.1. Convergence Result

For the two-stage adversarial framework, to study $\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)$, we denote the following function

$$m_\gamma(-\lambda) = \frac{-(1-\gamma+\lambda) + \sqrt{(1-\gamma+\lambda)^2 + 4\lambda\gamma}}{2\gamma\lambda},$$

which is used to describe the asymptotic behavior of $tr\left((\sum_{i=1}^{n_1} \boldsymbol{x}_i\boldsymbol{x}_i^\top + \lambda \boldsymbol{I}_d)^{-1}\right)$ as in (Hastie et al., 2019).

After defining m_γ, one can obtain the convergence of $\widehat{\boldsymbol{\theta}}_0(\lambda)$, and further figure out the asymptotic behavior of $\widehat{\boldsymbol{\theta}}_\epsilon(\lambda)$. The convergence of the two-stage adversarial training framework is as follows:

Theorem 1 (Convergence of Two-Stage Adversarial Training)**.** *With probability tending to 1, $\widehat{\boldsymbol{\theta}}_0(\lambda)$ satisfies*

$$\begin{aligned} \|\widehat{\boldsymbol{\theta}}_0(\lambda) - \boldsymbol{\theta}_0\|^2 &\to \lambda^2 r^2 m_\gamma'(-\lambda) \\ &\quad + \sigma^2\gamma\left(m_\gamma(-\lambda) - \lambda m_\gamma'(-\lambda)\right), \\ \|\widehat{\boldsymbol{\theta}}_0(\lambda)\|^2 &\to r^2[1 - 2\lambda m_\gamma(-\lambda) + \lambda^2 m_\gamma'(-\lambda)] \\ &\quad + \sigma^2\gamma[m_\gamma(-\lambda) - \lambda m_\gamma'(-\lambda)]. \end{aligned}$$

For the two-stage adversarial estimate $\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)$, assuming $n_2 = \infty$, $\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)$ satisfies

$$\begin{aligned} \|\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda) - \boldsymbol{\theta}_0\|^2 &\to \frac{1}{(1+\alpha_\epsilon(\lambda))^2}\|\widehat{\boldsymbol{\theta}}_0(\lambda)\|^2 \\ &\quad + r^2 - \frac{2}{(1+\alpha_\epsilon(\lambda))}\widehat{\boldsymbol{\theta}}_0(\lambda)^\top\boldsymbol{\theta}_0, \\ \|\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)\|^2 &\to \frac{1}{(1+\alpha_\epsilon(\lambda))^2}\|\widehat{\boldsymbol{\theta}}_0(\lambda)\|^2, \end{aligned}$$

where $2\widehat{\boldsymbol{\theta}}_0(\lambda)^\top\boldsymbol{\theta}_0$ can be calculated via

$$2\widehat{\boldsymbol{\theta}}_0(\lambda)^\top\boldsymbol{\theta}_0 = \|\boldsymbol{\theta}_0\|^2 + \|\widehat{\boldsymbol{\theta}}_0(\lambda)\|^2 - \|\widehat{\boldsymbol{\theta}}_0(\lambda) - \boldsymbol{\theta}_0\|^2,$$

and $\alpha_\epsilon(\lambda)$ is the solution of α in

$$\begin{aligned} &\alpha + \epsilon c_0\frac{\alpha\|\widehat{\boldsymbol{\theta}}_0(\lambda)\|}{\sqrt{\|\widehat{\boldsymbol{\theta}}_0(\lambda)\|^2\alpha^2 + \sigma^2(1+\alpha)^2}} \\ = \; &\epsilon c_0\frac{\sqrt{\|\widehat{\boldsymbol{\theta}}_0(\lambda)\|^2\alpha^2 + \sigma^2(1+\alpha)^2}}{\|\widehat{\boldsymbol{\theta}}_0(\lambda)\|} + \epsilon^2. \end{aligned}$$

The proof of Theorem 1 is in the appendix. We first study the convergence of $\widehat{\boldsymbol{\theta}}_0(\lambda)$, and then evaluate $\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)$.

From Theorem 1, similar to $\widehat{\boldsymbol{\theta}}_0$, one can see that $\|\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda) - \boldsymbol{\theta}_0\|^2$ and $\|\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)\|^2$ converges to some value as a function of $(\gamma, \lambda, \epsilon, \sigma^2)$ asymptotically.

We conduct a simulation to verify Theorem 1 and study the risk of the two-stage adversarial training. In the experiment, we take $n_1 = 100$ and $n_2 = \infty$, i.e., we directly use the population adversarial risk in the second stage. We change the data dimension d to obtain different $\gamma = d/n_1$. The data follows $X \sim N(\boldsymbol{0}, \boldsymbol{I}_d)$, $Y = X^\top\boldsymbol{\theta}_0 + \varepsilon$ with $\boldsymbol{\theta}_0 \sim N(0, \boldsymbol{I}_d/d)$ and $\varepsilon \sim N(0,1)$. The adversarial attack is taken as $\epsilon = 0.3$. We repeat the experiment 100 times to obtain the average performance. We use the excess adversarial risk, i.e., $R_\epsilon(\boldsymbol{\theta}) - R_\epsilon(\boldsymbol{\theta}_\epsilon)$ for $\boldsymbol{\theta} \in \{\widehat{\boldsymbol{\theta}}_0(\lambda), \widehat{\boldsymbol{\theta}}_\epsilon(\lambda), \widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)\}$, to evaluate the performance of the three methods. The model $\widehat{\boldsymbol{\theta}}_\epsilon(\lambda)$ refers to the vanilla adversarial training as an additional benchmark, i.e., we conduct adversarial training using

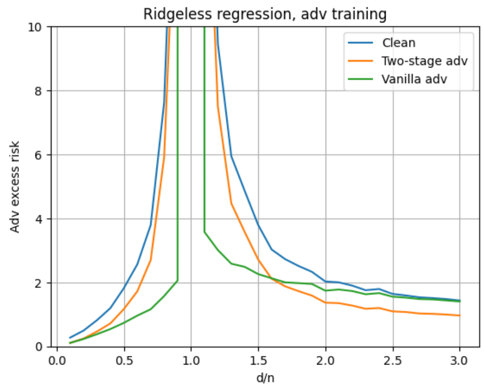

Figure 1. Simulation: Excess adversarial risk of clean training, vanilla adversarial training, and the two-stage adversarial training, without ridge penalty.

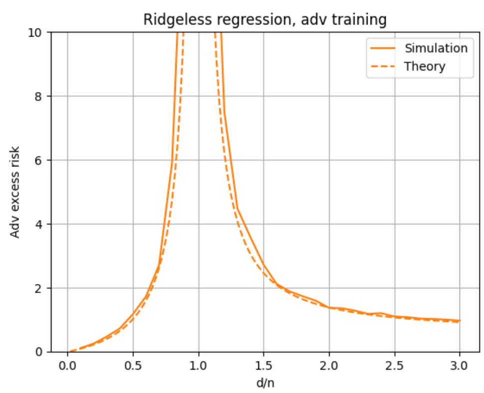

Figure 2. Theoretical value corresponding to Figure 1.

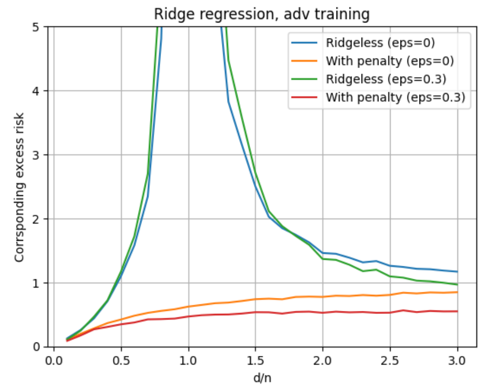

Figure 3. Simulation: Ridgeless regression and ridge regression with the best penalty in clean training and the two-stage adversarial training respectively. Adversarial training benefits more from a proper penalty.

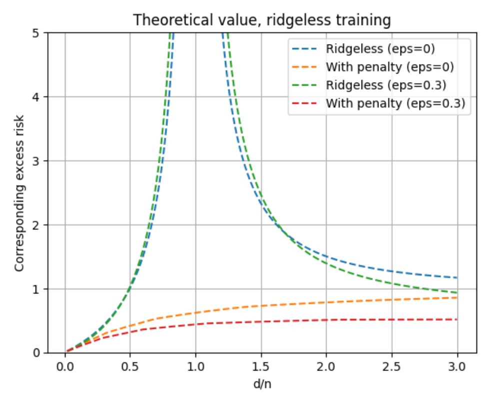

Figure 4. Theoretical value corresponding to Figure 3

the n_1 labeled samples. The simulation results are summarized in Figure 1, 2, 3, 4.

In Figure 1, we take $\lambda \to 0$ to align with the experiments in the double descent literature. There are several observations from Figure 1. First, if we compare the performance of the two-stage adversarial training and the clean training, the two-stage adversarial training is better than clean training. Second, when d/n_1 gets larger, the performance of the two-stage adversarial training is better than the vanilla adversarial training, indicating that the information of the additional extra data matters. Finally, for all the three training methods, they all observe a double-descent phenomenon.

In addition, we plot the theoretical curves for the excess adversarial risk associating with the two-stage adversarial training. From Figure 2, the theoretical curve and the simulation result match with each other.

Finally, we examine how the ridge penalty affects the performance. In the simulation in Figure 3, we take $\epsilon = 0, 0.3$ and compare the performance when $\lambda = 0$ and λ is taken to minimize the risk. In Figure 3, the y-axis is the corresponding excess adversarial risk, i.e., $\epsilon = 0, 0.3$ for the corresponding groups respectively. The corresponding theoretical curves can be found in Figure 4.

From Figure 3, one can see that the excess risk for the ridgeless regression is similar, while the two-stage adversarial training ($\epsilon = 0.3$) benefits more than clean training ($\epsilon = 0$) when taking a proper ridge penalty, which motivates us to further investigate in the penlaty term in the following sections. In addition, the theoretical curves in Figure 4 align with the simulation results in 3 as well.

3.2. A Better Clean Estimate May Not Be Preferred

Different from ridgeless regression in the large-sample regime, with high-dimensional data, it is essential to utilize ridge penalty or other regularization to improve the testing performance. While one can adjust the penalty to control the performance of the clean estimate, we would like to ask:

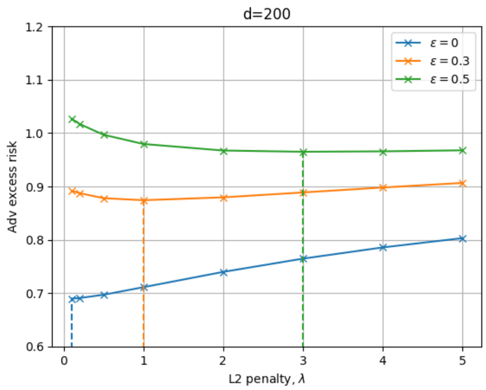

Figure 5. Simulation: How the tuning parameter λ in clean ridge regression affects the final adversarial robustness when using extra unlabeled data in training. While a small λ minimizes the population clean risk, this choice of λ is sub-optimal when using $\widehat{\boldsymbol{\theta}}_0(\lambda)$ to create pseudo response. Besides the cases of $\epsilon \in \{0, 0.3, 0.5\}$, when $\epsilon = 0.7$, the best penalty λ is extremely large and is not included in the figure.

Is a better clean estimate (measured by clean testing performance) always preferred in the two-stage method?

To answer the above question, it is essential to investigate the role of the clean estimate in the two-stage method. Recall that the population adversarial risk is written as

$$\begin{aligned} R_\epsilon(\boldsymbol{\theta}, \boldsymbol{\theta}_0) &= \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}^2 \\ &\quad + 2c_0\epsilon\|\boldsymbol{\theta}\|\sqrt{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}^2 + \sigma^2} + \epsilon^2\|\boldsymbol{\theta}\|^2, \end{aligned}$$

where taking expectation on training data we have

$$\mathbb{E}\|\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda) - \boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}^2 = \|\mathbb{E}\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda) - \boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}^2 + tr(Var(\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda))),$$

and

$$\mathbb{E}\|\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)\|^2 = \|\mathbb{E}\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)\|^2 + tr(Var(\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda))).$$

The above decompositions imply that the while ridge regression balances bias and variance of $\widehat{\boldsymbol{\theta}}_0(\lambda)$, the importance of bias and variance are changed in $\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)$. As a result, the optimal λ for the clean estimate may not be the best when applied in the two-stage adversarial training.

To investigate how the optimal λ changes in the two-stage method, a simulation study is conducted in Figure 5. We take $n_1 = 50$. The data $X \sim N(\mathbf{0}, \boldsymbol{I}_d)$ and $d = 200$. The response $Y = \boldsymbol{\theta}_0^\top X + \varepsilon$ with $\boldsymbol{\theta}_0 = \mathbf{1}/\sqrt{d}$ and $\varepsilon \sim N(0, 0.1^2)$. Besides the n_1 labeled data, we take $n_2 = \infty$. We repeat 30 times to get the average result and check the best λ under different attack strength ϵ.

From Figure 5, one can see that the optimal λ gets larger when the attack strength gets larger. When $\epsilon = 0$, the optimal λ is closed to zero. When $\epsilon = 0.3$, the best λ is around 1, and 3 when $\epsilon = 0.5$, both of which are much larger than the case for $\epsilon = 0$.

3.3. Cross Validation

Observing that the optimal λ for clean training is not the best for the two-stage adversarial training, we next investigate how to better select a proper λ.

While one can always use the leave-one-out procedure for any estimate, it is time-consuming. As a result, existing literature, e.g. (Hastie et al., 2019), utilize ways to approximate the leave-one-out CV procedure.

Recall that when $n_2 = \infty$, the second stage of the two-stage method minimizes

$$\begin{aligned} R_\epsilon(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_0(\lambda)) &= \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_0(\lambda)\|_{\boldsymbol{\Sigma}}^2 + \sigma^2 + \epsilon^2\|\boldsymbol{\theta}\|^2 \\ &\quad + 2c_0\epsilon\|\boldsymbol{\theta}\|\sqrt{\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_0(\lambda)\|_{\boldsymbol{\Sigma}}^2 + \sigma^2}, \end{aligned}$$

and the solution is

$$\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda) = (\boldsymbol{\Sigma} + \alpha_\epsilon(\lambda)\boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0(\lambda),$$

for some $\alpha_\epsilon(\lambda) \geq 0$. One needs to rerun the CV procedure for n_1 times and obtain different $\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)^{-j}$, the leave-one-out estimate of $\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)$ leaving the jth labeled sample.

Given that the above formula $\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)$ is a transformation of $\widehat{\boldsymbol{\theta}}_0(\lambda)$, one can borrow the idea of approximating CV in clean training to the two-stage adversarial training. To be specific, since both the $\alpha_\epsilon(\lambda)$ and $\widehat{\boldsymbol{\theta}}_0(\lambda)$ relate to each labeled sample, assuming the jth sample is discarded, the estimate of the two-stage method will be

$$(\boldsymbol{\Sigma} + \alpha^{-j}\boldsymbol{I}_d)\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0^{-j}(\lambda), \tag{3.1}$$

and we approximate both α^{-j} and $\widehat{\boldsymbol{\theta}}_0^{-j}(\lambda)$.

The following lemma shows how to approximate $\alpha_\epsilon(\lambda)$ in the leave-one-out CV:

Lemma 1. *Rewrite $\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)$ as $\widetilde{\boldsymbol{\theta}}$, $\widehat{\boldsymbol{\theta}}_0(\lambda)$ as $\widehat{\boldsymbol{\theta}}_0$, and $\alpha = \alpha_\epsilon(\lambda)$ for simplicity. Denote $\Delta_j = \widehat{\boldsymbol{\theta}}_0^{-j} - \widehat{\boldsymbol{\theta}}_0$, and*

$$\begin{aligned} A_1 &= \frac{1}{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}\|\widetilde{\boldsymbol{\theta}}\|}\widetilde{\boldsymbol{\theta}}^\top(\boldsymbol{\Sigma} + \alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0 \\ &\quad - \frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}^3}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0)^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0, \end{aligned}$$

$$\begin{aligned} A_2 &= \frac{1}{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}\|\widetilde{\boldsymbol{\theta}}\|}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0)^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0 \\ &\quad - \frac{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|^3}\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \alpha\boldsymbol{I}_d)^{-2}\widehat{\boldsymbol{\theta}}_0, \end{aligned}$$

ϵ	0.3	0.5	0.7
Cross validation (CV loss in training)	0.8750	0.9663	1.0300
Cross validation (corresponding population risk)	0.8871	0.9751	1.0270
Cross validation for clean regression (corresponding population risk)	0.8873	1.0076	1.1140
Best λ (corresponding population risk)	0.8741	0.9648	1.0185

Table 1. Adversarial risks using cross validation and the best λ.

$$\begin{aligned} A_3 &= \left(I_d + \epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}} - \theta_0\|_{\boldsymbol{\Sigma}}}\right) \Sigma(\boldsymbol{\Sigma} + \alpha \boldsymbol{I}_d)^{-1}\widetilde{\boldsymbol{\theta}} \\ &\quad + \left(\epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|} + \epsilon^2\right) (\boldsymbol{\Sigma} + \alpha \boldsymbol{I}_d)^{-1}\widetilde{\boldsymbol{\theta}}, \end{aligned}$$

$$\begin{aligned} A_4 &= \frac{1}{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}\|\widetilde{\boldsymbol{\theta}}\|}\widetilde{\boldsymbol{\theta}}^\top (\boldsymbol{\Sigma} + \alpha \boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma} \\ &\quad + \frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}^3} \alpha (\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0)^\top \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \alpha \boldsymbol{I}_d)^{-1}, \end{aligned}$$

$$\begin{aligned} A_5 &= \frac{1}{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}\|\widetilde{\boldsymbol{\theta}}\|} \alpha (\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0)^\top \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \alpha \boldsymbol{I}_d)^{-1} \\ &\quad - \frac{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|^3}\widetilde{\boldsymbol{\theta}}^\top (\boldsymbol{\Sigma} + \alpha \boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}, \end{aligned}$$

then when $\|\widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\theta}}_0^{-j}\| = o(1)$, the leave-one-out estimate of α satisfies

$$\begin{aligned} \alpha^{-j} - \alpha &= \frac{\left(\epsilon c_0 A_1 \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0) + \epsilon c_0 A_2 \widetilde{\boldsymbol{\theta}} + A_3\right)^\top}{\|\epsilon c_0 A_1 \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0) + \epsilon c_0 A_2 \widetilde{\boldsymbol{\theta}} + A_3\|^2} \\ &\quad \times \left(\epsilon c_0 A_4 \Delta_j \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0) + \epsilon c_0 A_5 \Delta_j \widetilde{\boldsymbol{\theta}}\right) + o, \end{aligned}$$

where o represents negligible terms.

The proof of Lemma 1 can be found in the appendix. Based on the result in Lemma 1, we can use

$$\begin{aligned} \widehat{\alpha}^{-j} - \alpha &= \frac{\left(\epsilon c_0 A_1 \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0) + \epsilon c_0 A_2 \widetilde{\boldsymbol{\theta}} + A_3\right)^\top}{\|\epsilon c_0 A_1 \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0) + \epsilon c_0 A_2 \widetilde{\boldsymbol{\theta}} + A_3\|^2} \\ &\quad \times \left(\epsilon c_0 A_4 \Delta_j \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0) + \epsilon c_0 A_5 \Delta_j \widetilde{\boldsymbol{\theta}}\right) \end{aligned}$$

to approximate α^{-j}.

In terms of the leave-on-out estimate of $\widehat{\boldsymbol{\theta}}_0(\lambda)$, i.e., $\widehat{\boldsymbol{\theta}}_0^{-j}(\lambda)$, one can use the Kailath Variant fomular (from 3.1.2 of Petersen and Pedersen, 2008) and obtain

$$\widehat{\boldsymbol{\theta}}_0(\lambda) - \widehat{\boldsymbol{\theta}}_0^{-j}(\lambda) = \frac{y_j - \widehat{y}_j(\lambda)}{1 - S_j(\lambda)} (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{I}_d)^{-1} \boldsymbol{x}_j,$$

where $\boldsymbol{X} \in \mathbb{R}^{n_1 \times d}$ denotes the labeled data matrix, and $\widehat{y}_j(\lambda) = \widehat{\boldsymbol{\theta}}_0(\lambda)^\top \boldsymbol{x}_j$ as the fitted value of the jth observation.

After obtaining the estimate $\widehat{\alpha}^{-j}$ and $\widehat{\boldsymbol{\theta}}_0^{-j}(\lambda)$, one can put them into (3.1) to obtain the leave-one-out estimate of $\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)$. The following theorem justifies the correctness of the above procedure:

Theorem 2. *Denote*

$$CV(\lambda, \epsilon) = \frac{1}{n_1} \sum \left(|\boldsymbol{x}_i^\top \widetilde{\boldsymbol{\theta}}_\epsilon^{-j}(\lambda) - y_i| + \epsilon \|\widetilde{\boldsymbol{\theta}}_\epsilon^{-j}(\lambda)\| \right)^2,$$

and $\breve{\boldsymbol{\theta}}_\epsilon^{-j}(\lambda) = (\boldsymbol{\Sigma} + \widehat{\alpha}^{-j} \boldsymbol{I}_d)\Sigma \widehat{\boldsymbol{\theta}}_0^{-j}(\lambda)$ as the approximation of the leave-one-out estimate using Lemma 1. Then under the Gaussian model assumption of (X, Y), the approximated CV converges to the actual CV result, i.e.,

$$\frac{1}{n_1} \sum \left(|\boldsymbol{x}_i^\top \breve{\boldsymbol{\theta}}_\epsilon^{-j}(\lambda) - y_i| + \epsilon \|\breve{\boldsymbol{\theta}}_\epsilon^{-j}(\lambda)\| \right)^2 \xrightarrow{P} CV(\lambda, \epsilon).$$

We use the simulation setting in Figure 5 to examine the performance of the above cross validation method. The results are summarized in Table 1.

From Table 1, there are two observations. First, one can see that using the cross validation, the CV loss in training is closed to the corresponding population risk.

In addition, the performance of the proposed algorithm is closed to the optimal λ, and using clean regression in cross validation leads to a worse performance.

4. Conclusion and Future Directions

This paper studies the asymptotics of the two-stage adversarial training in a high-dimensional linear regression setup. Double descent is observed for the ridge-less regression case, and a better performance can be achieved via $\mathcal{L}_2$ regularization. We also derive the shortcut cross validation formula for this two-stage method to simplify the computation for cross validation.

The results in this paper can be extended in some directions. First, in literature, e.g., (Ba et al., 2020), the double descent phenomenon is also related to two-layer neural networks. An interesting future direction is to extend the analysis in this paper to the neural network setup. Second, since the shortcut formula for cross validation is distribution specific and assumes $n_2 = \infty$, one may investigate in a more general cross validation procedure or relax to the scenario with a finite n_2.

References

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. *Advances in neural information processing systems*, 33:11022–11032, 2020.

Jimmy Ba, Murat Erdogdu, Taiji Suzuki, Denny Wu, and Tianzong Zhang. Generalization of two-layer neural networks: an asymptotic viewpoint. In *8th International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1gBsgBYwH.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019.

Daniel Berrar. Cross-validation., 2019.

Michael W Browne. Cross-validation methods. *Journal of mathematical psychology*, 44(1):108–132, 2000.

Qi-Zhi Cai, Min Du, Chang Liu, and Dawn Song. Curriculum adversarial training. *arXiv preprint arXiv:1805.04807*, 2018.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11192–11203, 2019.

Lin Chen, Yifei Min, Mingrui Zhang, and Amin Karbasi. More data can expand the generalization gap between adversarially robust and standard models. In *International Conference on Machine Learning*, pages 1670–1680. PMLR, 2020.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Chen Dan, Yuting Wei, and Pradeep Ravikumar. Sharp statistical guaratees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pages 2345–2355. PMLR, 2020.

Zhun Deng, Linjun Zhang, Amirata Ghorbani, and James Zou. Improving adversarial robustness via unlabeled out-of-domain data. In *International Conference on Artificial Intelligence and Statistics*, pages 2845–2853. PMLR, 2021.

Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290. PMLR, 2020.

Ruiqi Gao, Tianle Cai, Haochuan Li, Liwei Wang, Cho-Jui Hsieh, and Jason D Lee. Convergence of adversarial training in overparametrized networks. *arXiv preprint arXiv:1906.07916*, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34, 2021.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. *arXiv preprint arXiv:1901.09960*, 2019.

Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pages 2034–2078. PMLR, 2020.

Fanghui Liu, Zhenyu Liao, and Johan Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics*, pages 649–657. PMLR, 2021.

Bochen Lv and Zhanxing Zhu. Implicit bias of adversarial training for deep neural networks. In *International Conference on Learning Representations*, 2021.

Poorya Mianjy and Raman Arora. Robustness guarantees for adversarially trained neural networks. 2022.

Yifei Min, Lin Chen, and Amin Karbasi. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. *arXiv preprint arXiv:2002.11080*, 2020.

Amir Najafi, Shin-ichi Maeda, Masanori Koyama, and Takeru Miyato. Robustness to adversarial perturbations in learning from incomplete data. In *Advances in Neural Information Processing Systems*, pages 5542–5552, 2019.

KB Petersen and MS Pedersen. The matrix cookbook, vol. 7. 2008.

Richard R Picard and R Dennis Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583, 1984.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

Jason W Rocks and Pankaj Mehta. Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models. *Physical Review Research*, 4(1): 013201, 2022.

Jun Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494, 1993.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. 2018.

Mervyn Stone. Cross-validation: A review. *Statistics: A Journal of Theoretical and Applied Statistics*, 9(1):127–139, 1978.

Mahsa Taheri, Fang Xie, and Johannes Lederer. Statistical guarantees for regularized neural networks. *Neural Networks*, 142:148–161, 2021.

Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In *International Conference on Machine Learning*, pages 6586–6595, 2019a.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019b.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. *arXiv preprint arXiv:2302.04638*, 2023.

Dongxian Wu, Yisen Wang, and Shu-tao Xia. Adversarial weight perturbation helps robust generalization. *arXiv preprint arXiv:2004.05884*, 2020.

Jiancong Xiao, Yanbo Fan, Ruoyu Sun, and Zhi-Quan Luo. Adversarial rademacher complexity of deep neural networks. 2021.

Jiancong Xiao, Yanbo Fan, Ruoyu Sun, Jue Wang, and Zhi-Quan Luo. Stability analysis and generalization bounds of adversarial training. *arXiv preprint arXiv:2210.00960*, 2022.

Yue Xing, Ruizhi Zhang, and Guang Cheng. Adversarially robust estimate and risk analysis in linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 514–522. PMLR, 2021.

Yue Xing, Qifan Song, and Guang Cheng. Why do artificially generated data help adversarial robustness. *Neurips*, 2022.

Dong Yin, Kannan Ramchandran, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. *arXiv preprint arXiv:1810.11914*, 2018.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 2019.

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *International Conference on Machine Learning*, pages 11278–11287. PMLR, 2020a.

Teng Zhang and Kang Li. Understanding overfitting in adversarial training in kernel regression. *arXiv preprint arXiv:2304.06326*, 2023.

Yi Zhang, Orestis Plevrakis, Simon S Du, Xingguo Li, Zhao Song, and Sanjeev Arora. Over-parameterized adversarial training: An analysis overcoming the curse of dimensionality. *arXiv preprint arXiv:2002.06668*, 2020b.

A. Proofs

A.1. Theorem 1

Proof of Theorem 1. We first analyze $\|\widehat{\boldsymbol{\theta}}_0(\lambda)-\boldsymbol{\theta}_0\|^2$ and $\|\widehat{\boldsymbol{\theta}}_0(\lambda)\|^2$.

For $\|\widehat{\boldsymbol{\theta}}_0(\lambda)\|^2$, denoting $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ as the vector of response and noise, we have

$$\begin{aligned}\|\widehat{\boldsymbol{\theta}}_0(\lambda)\|^2 &= \boldsymbol{y}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda n_1\boldsymbol{I}_d)^{-2}\boldsymbol{X}^\top\boldsymbol{y}\\ &= \boldsymbol{\theta}_0^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda n_1\boldsymbol{I}_d)^{-2}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\theta}_0+\boldsymbol{\varepsilon}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda n_1\boldsymbol{I}_d)^{-2}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\\ &\quad +2\boldsymbol{\theta}_0^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda n_1\boldsymbol{I}_d)^{-2}\boldsymbol{X}^\top\boldsymbol{\varepsilon}.\end{aligned}$$

We look at the each term respectively. In probability, we have

$$\begin{aligned}&\boldsymbol{\theta}_0^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda n_1\boldsymbol{I}_d)^{-2}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\theta}_0\\ =\;& r^2-2\lambda n_1\boldsymbol{\theta}_0^\top(\boldsymbol{X}^\top\boldsymbol{X}+\lambda n_1\boldsymbol{I}_d)^{-1}\boldsymbol{\theta}_0+\lambda^2 n_1^2\boldsymbol{\theta}_0^\top(\boldsymbol{X}^\top\boldsymbol{X}+\lambda n_1\boldsymbol{I}_d)^{-2}\boldsymbol{\theta}_0\\ \to\;& r^2\left[1-2\lambda m_\gamma(-\lambda)+\lambda^2 m_\gamma'(-\lambda)\right],\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\varepsilon}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda n_1\boldsymbol{I}_d)^{-2}\boldsymbol{X}^\top\boldsymbol{\varepsilon} &\to \sigma^2\left[\frac{1}{n_1}tr((\widehat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I}_d)^{-1})-\frac{1}{n_1}\lambda tr((\widehat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I}_d)^{-2})\right]\\ &\to \sigma^2\left[\gamma m_\gamma(-\lambda)-\lambda\gamma m_\gamma'(-\lambda)\right],\end{aligned}$$

where the function m_γ is obtained from (Hastie et al., 2019). For the cross term, we also have

$$\begin{aligned}&\left[\boldsymbol{\theta}_0^\top\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda n_1\boldsymbol{I}_d)^{-2}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right]^2\\ \to\;& \sigma^2 tr\left[\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda n_1\boldsymbol{I}_d)^{-2}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda n_1\boldsymbol{I}_d)^{-2}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\right]\\ \xrightarrow{P}\;& 0.\end{aligned}$$

As a result,

$$\|\widehat{\boldsymbol{\theta}}_0(\lambda)\|^2\xrightarrow{P} r^2\left[1-2\lambda m_\gamma(-\lambda)+\lambda^2 m_\gamma'(-\lambda)\right]+\sigma^2\gamma\left[m_\gamma(-\lambda)-\lambda m_\gamma'(-\lambda)\right].$$

For $\|\widehat{\boldsymbol{\theta}}_0(\lambda)-\boldsymbol{\theta}_0\|^2$, we have

$$\|\widehat{\boldsymbol{\theta}}_0(\lambda)-\boldsymbol{\theta}_0\|^2=\|\widehat{\boldsymbol{\theta}}_0(\lambda)\|^2+\|\boldsymbol{\theta}_0\|^2-2\widehat{\boldsymbol{\theta}}_0(\lambda)^\top\boldsymbol{\theta}_0,$$

where in probability,

$$\left[\boldsymbol{\varepsilon}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda n_1\boldsymbol{I}_d)^{-1}\boldsymbol{\theta}_0\right]^2\to 0,$$

and

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_0(\lambda)^\top\boldsymbol{\theta}_0 &= \boldsymbol{\theta}_0\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda n_1\boldsymbol{I}_d)^{-1}\boldsymbol{\theta}_0+\boldsymbol{\varepsilon}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda n_1\boldsymbol{I}_d)^{-1}\boldsymbol{\theta}_0\\ &\to r^2-\lambda n_1\boldsymbol{\theta}_0(\boldsymbol{X}^\top\boldsymbol{X}+\lambda n_1\boldsymbol{I}_d)^{-1}\boldsymbol{\theta}_0\\ &\to r^2-\lambda r^2 m_\gamma(-\lambda).\end{aligned}$$

Consequently, in probability,

$$\|\widehat{\boldsymbol{\theta}}_0(\lambda)-\boldsymbol{\theta}_0\|^2\to r^2\lambda^2 m_\gamma'(-\lambda)+\sigma^2\gamma[m_\gamma(-\lambda)-\lambda m_\gamma'(-\lambda)].$$

For adversarial training, from (Javanmard et al., 2020; Xing et al., 2021) we know that the minimizer of $R_\epsilon(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_0(\lambda))$ is

$$\widetilde{\boldsymbol{\theta}}_\epsilon(\lambda)=(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0(\lambda),$$

where $c_0 = \sqrt{2/\pi}$ and α satisfies

$$\alpha\left(1+\epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}}\|}{\sqrt{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}^2+\sigma^2}}\right) = \left(\epsilon c_0 \frac{\sqrt{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}^2+\sigma^2}}{\|\widetilde{\boldsymbol{\theta}}\|}+\epsilon^2\right).$$

When $\boldsymbol{\Sigma} = \boldsymbol{I}_d$, the above is reduced to

$$\alpha + \epsilon c_0 \frac{\alpha\|\widehat{\boldsymbol{\theta}}_0\|}{\sqrt{\|\widehat{\boldsymbol{\theta}}_0\|^2\alpha^2+\sigma^2(1+\alpha)^2}} = \epsilon c_0 \frac{\sqrt{\|\widehat{\boldsymbol{\theta}}_0\|^2\alpha^2+\sigma^2(1+\alpha)^2}}{\|\widehat{\boldsymbol{\theta}}_0\|}+\epsilon^2.$$

Since $\|\widehat{\boldsymbol{\theta}}_0(\lambda)\|^2$ asymptotically converges to some fixed value, the solution of α also asymptotically converges. $\square$

A.2. Cross Validation

We present the proof of Lemma 1 and Theorem 2 in this section.

Proof of Lemma 1. To do cross validation, we know that α satisfies

$$\alpha\left(1+\epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}}\|}{\sqrt{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}^2+\sigma^2}}\right) = \left(\epsilon c_0 \frac{\sqrt{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}^2+\sigma^2}}{\|\widetilde{\boldsymbol{\theta}}\|}+\epsilon^2\right).$$

For the optimal solution in the adversarial training stage, we have

$$\begin{aligned}\mathbf{0} = \nabla R_\epsilon(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_0) &= 2\left[\Sigma(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_0)+\epsilon c_0\frac{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\boldsymbol{\theta}\|}\boldsymbol{\theta}+\epsilon c_0\frac{\|\boldsymbol{\theta}\|}{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}\boldsymbol{\Sigma}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_0)+(\epsilon^2)\boldsymbol{\theta}\right]\\ &= 2\left[\left(\boldsymbol{I}_d+\epsilon c_0\frac{\|\boldsymbol{\theta}\|}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}}\right)\Sigma(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_0)+\left(\epsilon c_0\frac{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\boldsymbol{\theta}\|}+\epsilon^2\right)\boldsymbol{\theta}\right].\end{aligned}$$

For leave-one-out CV, we have

$$\begin{aligned}\mathbf{0} = \nabla R_\epsilon(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_0^{-j}) &= 2\left[\Sigma(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_0^{-j})+\epsilon c_0\frac{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_0^{-j}\|_{\boldsymbol{\Sigma}}}{\|\boldsymbol{\theta}\|}\boldsymbol{\theta}+\epsilon c_0\frac{\|\boldsymbol{\theta}\|}{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_0^{-j}\|_{\boldsymbol{\Sigma}}}\boldsymbol{\Sigma}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_0^{-j})+(\epsilon^2)\boldsymbol{\theta}\right]\\ &= 2\left[\left(\boldsymbol{I}_d+\epsilon c_0\frac{\|\boldsymbol{\theta}\|}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^{-j}\|_{\boldsymbol{\Sigma}}}\right)\Sigma(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_0^{-j})+\left(\epsilon c_0\frac{\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_0^{-j}\|_{\boldsymbol{\Sigma}}}{\|\boldsymbol{\theta}\|}+\epsilon^2\right)\boldsymbol{\theta}\right].\end{aligned}$$

Consequently,

$$\begin{aligned}&\left(\boldsymbol{I}_d+\epsilon c_0\frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}}\right)\Sigma(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)+\left(\epsilon c_0\frac{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|}+\epsilon^2\right)\widetilde{\boldsymbol{\theta}}\\ =\ &\left(\boldsymbol{I}_d+\epsilon c_0\frac{\|\widetilde{\boldsymbol{\theta}}^{-j}\|}{\|\widetilde{\boldsymbol{\theta}}^{-j}-\boldsymbol{\theta}_0^{-j}\|_{\boldsymbol{\Sigma}}}\right)\Sigma(\widetilde{\boldsymbol{\theta}}^{-j}-\widehat{\boldsymbol{\theta}}_0^{-j})+\left(\epsilon c_0\frac{\|\widetilde{\boldsymbol{\theta}}^{-j}-\widehat{\boldsymbol{\theta}}_0^{-j}\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}^{-j}\|}+\epsilon^2\right)\widetilde{\boldsymbol{\theta}}^{-j}.\end{aligned}$$

Denote

$$\Delta_j = \widehat{\boldsymbol{\theta}}_0^{-j}-\widehat{\boldsymbol{\theta}}_0,$$

and denote α^{-j} as the best α without jth sample. Then

$$\begin{aligned}\widetilde{\boldsymbol{\theta}}^{-j}-\widetilde{\boldsymbol{\theta}} &= (\boldsymbol{\Sigma}+\alpha^{-j}\boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0^{-j}-(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0\\ &= (\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}\Delta_j-(\alpha^{-j}-\alpha)(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\widetilde{\boldsymbol{\theta}}+R_0.\end{aligned}$$

When $\|\widehat{\boldsymbol{\theta}}_0\|$ and $\|\widetilde{\boldsymbol{\theta}}\|$ are away from zero,

$$\|R_0\| = O(|\alpha^{-j} - \alpha|\|\Delta_j\|).$$

As a result,

$$\begin{aligned}
&\left(\boldsymbol{I}_d + \epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}}\right)\Sigma(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0) + \left(\epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|} + \epsilon^2\right)\widetilde{\boldsymbol{\theta}} \\
= \quad & \left(\boldsymbol{I}_d + \epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}}^{-j}\|}{\|\widetilde{\boldsymbol{\theta}}^{-j} - \boldsymbol{\theta}_0^{-j}\|_{\boldsymbol{\Sigma}}}\right)\Sigma(\widetilde{\boldsymbol{\theta}}^{-j} - \widehat{\boldsymbol{\theta}}_0^{-j}) + \left(\epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}}^{-j} - \widehat{\boldsymbol{\theta}}_0^{-j}\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}^{-j}\|} + \epsilon^2\right)\widetilde{\boldsymbol{\theta}}^{-j} \\
&+ \left(\boldsymbol{I}_d + \epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}}\right)\Sigma(\widetilde{\boldsymbol{\theta}}^{-j} - \widehat{\boldsymbol{\theta}}_0^{-j}) + \left(\epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|} + \epsilon^2\right)\widetilde{\boldsymbol{\theta}}^{-j} \\
&- \left(\boldsymbol{I}_d + \epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}}\right)\Sigma(\widetilde{\boldsymbol{\theta}}^{-j} - \widehat{\boldsymbol{\theta}}_0^{-j}) - \left(\epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|} + \epsilon^2\right)\widetilde{\boldsymbol{\theta}}^{-j} \\
= \quad & \epsilon c_0 \left(\frac{\|\widetilde{\boldsymbol{\theta}}^{-j}\|}{\|\widetilde{\boldsymbol{\theta}}^{-j} - \boldsymbol{\theta}_0^{-j}\|_{\boldsymbol{\Sigma}}} - \frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}}\right)\Sigma(\widetilde{\boldsymbol{\theta}}^{-j} - \widehat{\boldsymbol{\theta}}_0^{-j}) + \epsilon c_0 \left(\frac{\|\widetilde{\boldsymbol{\theta}}^{-j} - \widehat{\boldsymbol{\theta}}_0^{-j}\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}^{-j}\|} - \frac{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|}\right)\widetilde{\boldsymbol{\theta}}^{-j} \\
&+ \left(\boldsymbol{I}_d + \epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}}\right)\Sigma(\widetilde{\boldsymbol{\theta}}^{-j} - \widehat{\boldsymbol{\theta}}_0^{-j}) + \left(\epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|} + \epsilon^2\right)\widetilde{\boldsymbol{\theta}}^{-j},
\end{aligned}$$

and changing the order of the terms in the above, we have

$$\begin{aligned}
&\left(\boldsymbol{I}_d + \epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}}\right)\Sigma(\widetilde{\boldsymbol{\theta}}^{-j} - \widehat{\boldsymbol{\theta}}_0^{-j}) + \left(\epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|} + \epsilon^2\right)\widetilde{\boldsymbol{\theta}}^{-j} - \nabla R_\epsilon(\widetilde{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}_0) \\
= \quad & \left(\boldsymbol{I}_d + \epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}}\right)\Sigma(\widetilde{\boldsymbol{\theta}}^{-j} - \widetilde{\boldsymbol{\theta}} - \Delta_j) + \left(\epsilon c_0 \frac{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|} + \epsilon^2\right)(\widetilde{\boldsymbol{\theta}}^{-j} - \widetilde{\boldsymbol{\theta}}) \\
= \quad & -\epsilon c_0 \left(\frac{\|\widetilde{\boldsymbol{\theta}}^{-j}\|}{\|\widetilde{\boldsymbol{\theta}}^{-j} - \boldsymbol{\theta}_0^{-j}\|_{\boldsymbol{\Sigma}}} - \frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}\right)\Sigma(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0) - \epsilon c_0 \left(\frac{\|\widetilde{\boldsymbol{\theta}}^{-j} - \widehat{\boldsymbol{\theta}}_0^{-j}\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}^{-j}\|} - \frac{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|}\right)\widetilde{\boldsymbol{\theta}} + R_1,
\end{aligned}$$

for

$$\|R_1\| = O(\|\widetilde{\boldsymbol{\theta}}^{-j} - \widehat{\boldsymbol{\theta}}_0\|\|\Delta_j\|) = O(\|\Delta_j\|^2 + |\alpha^{-j} - \alpha|\|\Delta_j\|).$$

We know that

$$\begin{aligned}
&\frac{\|\widetilde{\boldsymbol{\theta}}^{-j}\|}{\|\widetilde{\boldsymbol{\theta}}^{-j} - \widehat{\boldsymbol{\theta}}_0^{-j}\|_{\boldsymbol{\Sigma}}} - \frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}} \\
= \quad & \frac{1}{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}} \frac{\widetilde{\boldsymbol{\theta}}^\top(\widetilde{\boldsymbol{\theta}}^{-j} - \widetilde{\boldsymbol{\theta}})}{\|\widetilde{\boldsymbol{\theta}}\|} - \frac{(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0)^\top \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}}^{-j} - \widetilde{\boldsymbol{\theta}} - \Delta_j)}{\|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}^3}\|\widetilde{\boldsymbol{\theta}}\| + O(\|R_0\|),
\end{aligned}$$

and rewriting $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}^{-j}$ as functions of α and α^{-j}, the first-order terms can be represented as

$$\begin{aligned}
& \frac{1}{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}} \frac{\widetilde{\boldsymbol{\theta}}^\top(\widetilde{\boldsymbol{\theta}}^{-j}-\widetilde{\boldsymbol{\theta}})}{\|\widetilde{\boldsymbol{\theta}}\|} - \frac{(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)^\top\boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}}^{-j}-\widetilde{\boldsymbol{\theta}}-\Delta_j)}{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|^3_{\boldsymbol{\Sigma}}}\|\widetilde{\boldsymbol{\theta}}\| \\
= \; & \frac{1}{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}} \frac{\widetilde{\boldsymbol{\theta}}^\top\left((\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}\Delta_j-(\alpha^{-j}-\alpha)(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0\right)}{\|\widetilde{\boldsymbol{\theta}}\|} \\
& -\frac{(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)^\top\boldsymbol{\Sigma}\left((\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}\Delta_j-(\alpha^{-j}-\alpha)(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0-\Delta_j\right)}{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|^3_{\boldsymbol{\Sigma}}}\|\widetilde{\boldsymbol{\theta}}\|+O(\|R_0\|) \\
= \; & \frac{1}{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}\|\widetilde{\boldsymbol{\theta}}\|}\left(\widetilde{\boldsymbol{\theta}}^\top(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}\Delta_j-(\alpha^{-j}-\alpha)\widetilde{\boldsymbol{\theta}}^\top(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0\right)+O(\|R_0\|) \\
& -\frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|^3_{\boldsymbol{\Sigma}}}\left((\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}\Delta_j-(\alpha^{-j}-\alpha)(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0-(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)^\top\boldsymbol{\Sigma}\Delta_j\right) \\
= \; & \underbrace{\frac{1}{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}\|\widetilde{\boldsymbol{\theta}}\|}\alpha\widetilde{\boldsymbol{\theta}}^\top(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0-\frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|^3_{\boldsymbol{\Sigma}}}\alpha(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0}_{:=A_1\alpha} \\
& +\underbrace{\left(\frac{1}{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}\|\widetilde{\boldsymbol{\theta}}\|}\widetilde{\boldsymbol{\theta}}^\top(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}+\frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|^3_{\boldsymbol{\Sigma}}}\alpha(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\right)}_{:=A_4}\Delta_j \\
& -\alpha^{-j}\underbrace{\left(\frac{1}{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}\|\widetilde{\boldsymbol{\theta}}\|}\widetilde{\boldsymbol{\theta}}^\top(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0-\frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|^3_{\boldsymbol{\Sigma}}}(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0\right)}_{=A_2}+O(\|R_0\|).
\end{aligned}$$

Similarly,

$$\begin{aligned}
& \frac{\|\widetilde{\boldsymbol{\theta}}^{-j}-\widehat{\boldsymbol{\theta}}_0^{-j}\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}^{-j}\|}-\frac{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|} \\
= \; & \frac{(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)^\top\boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}}^{-j}-\widetilde{\boldsymbol{\theta}}-\Delta_j)}{\|\widetilde{\boldsymbol{\theta}}\|\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}-\frac{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\theta}}^\top(\widetilde{\boldsymbol{\theta}}^{-j}-\widetilde{\boldsymbol{\theta}})}{\|\widetilde{\boldsymbol{\theta}}\|^3} \\
= \; & \underbrace{\frac{1}{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}\|\widetilde{\boldsymbol{\theta}}\|}\alpha(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0-\frac{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|^3}\alpha\widetilde{\boldsymbol{\theta}}^\top(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0}_{:=A_2\alpha} \\
& +\underbrace{\left(\frac{1}{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}\|\widetilde{\boldsymbol{\theta}}\|}\alpha(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}-\frac{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|^3}\widetilde{\boldsymbol{\theta}}^\top(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}\right)}_{:=A_5}\Delta_j \\
& -\alpha^{-j}\underbrace{\left(\frac{1}{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}\|\widetilde{\boldsymbol{\theta}}\|}(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0-\frac{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|^3}\widetilde{\boldsymbol{\theta}}^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-2}\widehat{\boldsymbol{\theta}}_0\right)}_{=A_2}+O(\|R_0\|).
\end{aligned}$$

As a result,

$$\begin{aligned}
&\left(\boldsymbol{I}_d+\epsilon c_0\frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}}\right)\Sigma(\widetilde{\boldsymbol{\theta}}^{-j}-\widetilde{\boldsymbol{\theta}}-\Delta_j)+\left(\epsilon c_0\frac{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|}+\epsilon^2\right)(\widetilde{\boldsymbol{\theta}}^{-j}-\widetilde{\boldsymbol{\theta}})\\
=\;&-\frac{1}{2}\left(\epsilon c_0\left(A_1+A_2\Delta_j-\alpha^{-j}A_3\right)\Sigma(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)+\epsilon c_0\left(A_4+A_5\Delta_j-\alpha^{-j}A_6\right)\widetilde{\boldsymbol{\theta}}\right)+O(\|R_0\|+\|R_0\|)\\
=\;&\left(\boldsymbol{I}_d+\epsilon c_0\frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}}\right)\Sigma\left((\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}\Delta_j-(\alpha^{-j}-\alpha)(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0-\Delta_j\right)\\
&+\left(\epsilon c_0\frac{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|}+\epsilon^2\right)\left((\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}\Delta_j-(\alpha^{-j}-\alpha)(\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-2}\boldsymbol{\Sigma}\widehat{\boldsymbol{\theta}}_0\right)\\
=\;&\underbrace{\left[-\alpha\left(\boldsymbol{I}_d+\epsilon c_0\frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}}\right)+\left(\epsilon c_0\frac{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|}+\epsilon^2\right)\right](\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\boldsymbol{\Sigma}\,\Delta_j}_{=\boldsymbol{0}}\\
&+\alpha\underbrace{\left[\left(\boldsymbol{I}_d+\epsilon c_0\frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}}\right)\Sigma+\left(\epsilon c_0\frac{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|}+\epsilon^2\right)\right](\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\widetilde{\boldsymbol{\theta}}}_{:=A_3}\\
&-\alpha^{-j}\left[\left(\boldsymbol{I}_d+\epsilon c_0\frac{\|\widetilde{\boldsymbol{\theta}}\|}{\|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_{\boldsymbol{\Sigma}}}\right)\Sigma+\left(\epsilon c_0\frac{\|\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0\|_{\boldsymbol{\Sigma}}}{\|\widetilde{\boldsymbol{\theta}}\|}+\epsilon^2\right)\right](\boldsymbol{\Sigma}+\alpha\boldsymbol{I}_d)^{-1}\widetilde{\boldsymbol{\theta}},
\end{aligned}$$

that is,

$$\begin{aligned}
&-\epsilon c_0A_1\Delta_j\boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)-\epsilon c_0\alpha A_3\Delta_j\widetilde{\boldsymbol{\theta}}\\
=\;&(\alpha^{-j}-\alpha)\left(\epsilon c_0A_2\Sigma(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)+\epsilon c_0A_4\widetilde{\boldsymbol{\theta}}-A_5\right)+O(\|R_0\|+\|R_0\|),
\end{aligned}$$

and

$$\alpha^{-j}-\alpha\quad\approx\quad\frac{\left(\epsilon c_0A_2\Sigma(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)+\epsilon c_0A_4\widetilde{\boldsymbol{\theta}}+A_5\right)^\top}{\|\epsilon c_0A_2\Sigma(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)+\epsilon c_0A_4\widetilde{\boldsymbol{\theta}}+A_5\|^2}\left(\epsilon c_0A_1\Delta_j\boldsymbol{\Sigma}(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_0)+\epsilon c_0A_3\Delta_j\widetilde{\boldsymbol{\theta}}\right).$$

□

Proof of Theorem 2. From Lemma 1, we know that when $\|\Delta_j\|=o(1)$, $\alpha^{-j}-\alpha=o(1)$. In this proof, we check whether $\|\Delta_j\|\to 0$ for all $j=1,\dots,n_1$.

One can use the Kailath Variant fomular (from 3.1.2 of (Petersen and Pedersen, 2008)) to obtain

$$\begin{aligned}
&\widehat{\boldsymbol{\theta}}_0(\lambda)-\widehat{\boldsymbol{\theta}}_0^{-j}(\lambda)\\
=\;&(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d)^{-1}\boldsymbol{X}^\top\boldsymbol{y}\\
&-\left[(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d)^{-1}+\frac{(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d)^{-1}\boldsymbol{x}_j\boldsymbol{x}_j^\top(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d)^{-1}}{1-\boldsymbol{x}_j^\top(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d)^{-1}\boldsymbol{x}_j}\right]\boldsymbol{X}_{-j}^\top\boldsymbol{y}_{-j}\\
=\;&y_j(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d)^{-1}\boldsymbol{x}_j-\frac{\widehat{y}_j(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d)^{-1}\boldsymbol{x}_j}{1-S_j(\lambda)}+\frac{y_jS_j(\lambda)(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d)^{-1}\boldsymbol{x}_j}{1-S_j(\lambda)}\\
=\;&\frac{y_j-\widehat{y}_j(\lambda)}{1-S_j(\lambda)}(\boldsymbol{X}^\top\boldsymbol{X}+n\lambda\boldsymbol{I}_d)^{-1}\boldsymbol{x}_j,
\end{aligned}$$

where $\widehat{y}_j(\lambda)=\widehat{\boldsymbol{\theta}}_0(\lambda)^\top\boldsymbol{x}_j$.

Based on (Hastie et al., 2019), almost surely, denote $\boldsymbol{A}_i = n_1(\boldsymbol{X}_{-i}^\top \boldsymbol{X}_{-i} + \lambda n_1 \boldsymbol{I}_d)^{-1}$, and $\delta_i = \frac{\boldsymbol{x}_i}{\sqrt{n_1}}$, then

$$
\begin{aligned}
\boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X} + \lambda n_1 \boldsymbol{I}_d)^{-2} \boldsymbol{x}_i &= \frac{1}{n_1} \delta_i \left(\boldsymbol{A}_i - \frac{\boldsymbol{A}_i \delta_i \delta_i^\top \boldsymbol{A}_i}{1 + \delta_i^\top \boldsymbol{A}_i \delta_i} \right)^2 \delta_i \\
&= \frac{1}{n_1} \delta_i \left(\boldsymbol{A}_i^2 - 2 \frac{\boldsymbol{A}_i^2 \delta_i \delta_i^\top \boldsymbol{A}_i}{1 + \delta_i^\top \boldsymbol{A}_i \delta_i} + \frac{\boldsymbol{A}_i \delta_i \delta_i^\top \boldsymbol{A}_i^2 \delta_i \delta_i^\top \boldsymbol{A}_i}{(1 + \delta_i^\top \boldsymbol{A}_i \delta_i)^2} \right) \delta_i \\
&= \frac{1}{n_1} \left(\delta_i \boldsymbol{A}_i^2 \delta_i - 2 \frac{\delta_i^\top \boldsymbol{A}_i^2 \delta_i \delta_i^\top \boldsymbol{A}_i \delta_i}{1 + \delta_i^\top \boldsymbol{A}_i \delta_i} + \frac{(\delta_i^\top \boldsymbol{A}_i \delta_i)^2 \delta_i^\top \boldsymbol{A}_i^2 \delta_i}{(1 + \delta_i^\top \boldsymbol{A}_i \delta_i)^2} \right) \\
&= \frac{1}{n_1} \frac{\delta_i^\top \boldsymbol{A}_i^2 \delta_i}{(1 + \delta_i^\top \boldsymbol{A}_i \delta_i)^2} \\
&\xrightarrow{a.s.} \frac{1}{n_1} \frac{\gamma m_\gamma'(-\lambda)}{(1 + \gamma m_\gamma(-\lambda))^2}.
\end{aligned}
$$

Finally, for $y_i - \widehat{y}_i$, we have

$$
\begin{aligned}
y_i - \widehat{y}_i &= y_i - \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X} + \lambda n_1 \boldsymbol{I}_d)^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}) \\
&= \varepsilon_i - \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X} + \lambda n_1 \boldsymbol{I}_d)^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon} + \lambda n_1 \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X} + \lambda n_1 \boldsymbol{I}_d)^{-1} \boldsymbol{\theta}_0.
\end{aligned}
$$

Using Sherman–Morrison formula, we have

$$
\boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X} + \lambda n_1 \boldsymbol{I}_d)^{-1} \boldsymbol{\theta}_0 = \boldsymbol{x}_i^\top \left[\frac{1}{n_1} \boldsymbol{A}_i - \frac{\boldsymbol{A}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{A}_i / n_1^2}{1 + \boldsymbol{x}_i^\top \boldsymbol{A}_i \boldsymbol{x}_i / n_1} \right] \boldsymbol{\theta}_0,
$$

thus

$$
\begin{aligned}
& \left(\boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X} + \lambda n_1 \boldsymbol{I}_d)^{-1} \boldsymbol{\theta}_0 \right)^2 \\
= & \boldsymbol{x}_i^\top \frac{1}{n_1^2} \boldsymbol{A}_i \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top \boldsymbol{A}_i \boldsymbol{x}_i - \frac{2}{n_1} \boldsymbol{x}_i^\top \boldsymbol{A}_i \boldsymbol{\theta}_0 \boldsymbol{\theta}_0 \frac{\boldsymbol{A}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{A}_i / n_1^2}{1 + \boldsymbol{x}_i^\top \boldsymbol{A}_i \boldsymbol{x}_i / n_1} \boldsymbol{x}_i + \boldsymbol{x}_i^\top \frac{\boldsymbol{A}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{A}_i / n_1^2}{1 + \boldsymbol{x}_i^\top \boldsymbol{A}_i \boldsymbol{x}_i / n_1} \boldsymbol{\theta}_0 \boldsymbol{\theta}_0 \frac{\boldsymbol{A}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{A}_i / n_1^2}{1 + \boldsymbol{x}_i^\top \boldsymbol{A}_i \boldsymbol{x}_i / n_1} \boldsymbol{x}_i \\
= & O_p(tr\left(\boldsymbol{A}_i^2 \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top \right) / n_1^2) \\
= & O_p\left(m_\gamma(-\lambda) / n_1^2 \right).
\end{aligned}
$$

In addition,

$$
\left(\boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X} + \lambda n_1 \boldsymbol{I}_d)^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon} \right)^2 \to \sigma^2 \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X} + \lambda n_1 \boldsymbol{I}_d)^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda n_1 \boldsymbol{I}_d)^{-1} \boldsymbol{x}_i,
$$

which converges to a constant.

Finally, given the distribution of ε, we have with probability tending to 1,

$$
\sup_j \|\widehat{\boldsymbol{\theta}}_0(\lambda) - \widehat{\boldsymbol{\theta}}_0^{-j}(\lambda)\|^2 = o((\log n_1)/n_1).
$$

$\square$