Adversarially Robust Estimate and Risk Analysis in Linear Regression

Yue Xing
Purdue University

Ruizhi Zhang
University of Nebraska-Lincoln

Guang Cheng
Purdue University

Abstract

Adversarially robust learning aims to design algorithms that are robust to small adversarial perturbations on input variables. Beyond the existing studies on the predictive performance to adversarial samples, our goal is to understand the statistical properties of adversarially robust estimates and analyze adversarial risk in the setup of linear regression models. By discovering the statistical minimax rate of convergence of adversarially robust estimators, we emphasize incorporating model information, e.g., sparsity, in adversarially robust learning. Further, we reveal an explicit connection between adversarial and standard estimates and propose a straightforward two-stage adversarial learning framework that facilitates utilizing model structure information to improve adversarial robustness. In theory, the consistency of the adversarially robust estimator is proven and its Bahadur representation is also developed for the statistical inference purpose. The proposed estimator converges in a sharp rate under either a low-dimensional or a sparse scenario. Moreover, our theory confirms two phenomena in adversarially robust learning: adversarial robustness hurts generalization, and unlabeled data improves generalization. In the end, we conduct numerical simulations to verify our theory.

1 INTRODUCTION

The development of machine/deep learning methods has led to breakthrough performance in various areas of application. However, some recent research revealed that these powerful but delicate models are vulnerable to random perturbation and adversarial attacks. For example, well-designed malicious adversarial input may induce wrong decision making when filtering junk emails or detecting malicious binary programs Zhang et al. (2017); Papernot et al. (2017). On the other hand, by studying adversarial samples, one can improve the adversarial robustness of algorithms in practice. The existing literature focused on generating adversarial samples, e.g., Papernot et al. (2016, 2017), adversarial training, e.g., Goodfellow et al. (2015); Kurakin et al. (2017); Wang et al. (2019), invariance/interpretability to detect adversarial samples, e.g., Xu et al. (2018); Tao et al. (2018); Ma et al. (2019); Etmann et al. (2019); Carmon et al. (2019) and theoretical studies of adversarially robust learning, e.g., Xu et al. (2009a,b); Xu and Mannor (2012). In particular, some studies Yin et al. (2019); Raghunathan et al. (2019) showed that adversarial training leads to a worse generalization performance, while Schmidt et al. (2018); Zhai et al. (2019); Najafi et al. (2019) argued that the adversarial robustness requires more (labeled/unlabeled) data to enhance generalization performance. Besides, the trade-off between standard performance and adversarial performance is carefully characterized in Zhang et al. (2019); Javanmard et al. (2020).

Adversarially robust estimation in the literature is often formulated as an empirical "min-max" problem: minimizing the empirical risk under the worst-case attack (which maximizes the loss) on the training data. Unfortunately, this formulation does not directly consider the structural information of the model such as sparsity and grouping, e.g., Shaham et al. (2015); Sinha et al. (2018); Wang et al. (2019), which may be utilized to improve adversarial robustness. The structure information is particularly needed in the high-dimensional regime, i.e., data dimension p is much larger than sample size n, where the empirical (adversarial) risk may no longer converge to the population risk Mei et al. (2018).

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

The above concern raises two questions: (1) whether the statistical minimax[1] rate of the estimation error of *any* linear adversarial estimator will get changed given certain structure information for the standard model, and (2) whether we can utilize this information to get a better adversarially robust estimator.

Our contributions can be summarized as follows:

- In Section 3, by studying the form of adversarial risk, we figure out the minimax lower bound of estimation error, which reveals the potential to improve the estimation efficiency through utilizing model information.
- In Section 4, we design a two-stage adversarially robust learning framework that nicely connects adversarially robust estimation with standard estimation. The model structure information can be easily embedded into the standard estimator, and is further carried over to the adversarially robust estimate through this two-stage learning procedure. For statistical inference, we develop the Bahadur representation result (He and Shao, 1996) that implies the asymptotic normality of the proposed estimate under certain conditions. Besides, by analyzing the upper bound for the estimation error, we reveal the benefit of incorporating sparsity information into the adversarial estimation procedure, in which the estimator reaches the minimax optimal rate of convergence.
- Besides the above two main contributions, in Section 5, we utilize our theory to verify two arguments in adversarially robust learning: adversarially robust learning hurts generalization, and adversarial robustness can be improved using unlabeled data.

Two related works are appearing very recently. The first one Javanmard et al. (2020) mainly investigated the trade-off between adversarial risk and standard risk under an isotropic condition of the covariate. Rather, we focus on improving adversarial robustness by utilizing prior knowledge on the model and studying statistical properties of the adversarially robust estimate itself, in contrast with the generalization studies by Schmidt et al. (2018); Zhang et al. (2019); Zhai et al. (2019); Najafi et al. (2019). Another recent work Dan et al. (2020) studied the sharp statistical bound in adversarially robust *classification*. In the regression setup, our theorems reveal that an adversarially robust estimate is different from a standard estimate even in the rate of convergence: for noiseless case, standard model estimators can exactly recover the correct model, but the lower bound for adversarially robust model is always nonzero. Our lower bound for sparse model is also new. **Notation.** We use boldface font for vectors, e.g., $\mathbf{x}$, and capital letters for matrices, e.g., $\mathbf{A}$. The ℓ_2 norm of a vector $\mathbf{u}$ is denoted as $\|\mathbf{u}\|_2$ (or $\|\mathbf{u}\|$ for simplicity). The $p\times p$ identity matrix is denoted by $\mathbf{I}_p$.The induced spectral norm of a matrix $\mathbf{A}\in\mathbb{R}^{p\times p}$ is denoted by $\|\mathbf{A}\|$, i.e., $\|\mathbf{A}\| := \sup\{\|\mathbf{A}\mathbf{x}\| : \|\mathbf{x}\| = 1\}$. We denote by $\lambda_i(\mathbf{A}), i\in\{1,2,\cdots,p\}$, its eigenvalues in decreasing order. For a symmetric matrix $\mathbf{A}$, denote $\|\mathbf{x}\|_{\mathbf{A}}^2 = \mathbf{x}^\top\mathbf{A}\mathbf{x}$. For two matrices $\mathbf{A},\mathbf{B}$, we denote $\langle\mathbf{A},\mathbf{B}\rangle_F$ as the Frobenius inner product, which is the sum of component-wise inner product of two matrices. The Frobenius norm of $\mathbf{A}$ is denoted by $\|\mathbf{A}\|_F$.

2 PROPERTIES OF ADVERSARIAL RISK

Consider a linear regression model

$$y = \mathbf{x}^\top\theta_0 + \epsilon, \tag{1}$$

where $\mathbb{E}\mathbf{x} = \mathbf{0}$, $\mathrm{Var}(\mathbf{x}) = \Sigma$, and ϵ is a noise term (independent of $\mathbf{x}$) with $\mathbb{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$. Throughout this paper, we assume that $\mathbf{x}\in\mathbb{R}^p$ follows a p-dimensional Gaussian distribution and Σ has a bounded largest eigenvalue (away from ∞) and a bounded smallest eigenvalue (away from 0) as p increases. The noise variance σ^2 and $\|\theta_0\|$ are allowed to diverge in p, and the signal-to-noise ratio $\|\theta_0\|_\Sigma/\sigma$ needs to be large enough, say bounded away from 0.

The (population) adversarial risk is defined as follows

$$\begin{aligned} R_0(\theta,\delta) &:= \mathbb{E}_{\mathbf{x}}\max_{\|\mathbf{x}^*-\mathbf{x}\|_2\le\delta}\left[((\mathbf{x}^*)^\top\theta - \mathbf{x}^\top\theta_0)^2\right] \qquad (2)\\ &= \|\theta-\theta_0\|_\Sigma^2 + 2\delta c_0\|\theta-\theta_0\|_\Sigma\|\theta\| + \delta^2\|\theta\|^2, \end{aligned}$$

where $c_0 := \sqrt{2/\pi}$. The corresponding minimizer of (2) is denoted by $\theta^*(\delta)$, i.e.,

$$\theta^*(\delta) := \arg\min_\theta R_0(\theta,\delta).$$

We may just use θ^* when no confusion arises.

In the proposition below, we study the shape of R_0, and establish an analytical form of $\theta^*(\delta)$, which suggests the construction of adversarially robust estimator (to be specified later). Define

$$\theta(\lambda) := (\Sigma+\lambda\mathbf{I}_p)^{-1}\Sigma\theta_0,$$

and two thresholds of δ:

$$\delta_1 = \frac{c_0\|\theta_0\|}{\|\theta_0\|_{\Sigma^{-1}}} \text{ and } \delta_2 = \frac{\|\theta_0\|_{\Sigma^2}}{c_0\|\theta_0\|_\Sigma}.$$

[1] In this paper, "min-max" refers to the optimization problem considered in adversarially robust learning, while "minimax" refers to the statistical lower bound on the estimation error.

Proposition 1. *The risk $R_0(\theta, \delta)$ is a convex function w.r.t. θ, and has positive definite Hessian for any $\theta \neq \mathbf{0}, \theta \neq \theta_0$. In addition, the global minimizer of $R_0(\theta, \delta)$ can be written as*

$$\theta^*(\delta) := \theta(\lambda^*(\delta)) \tag{3}$$

where $\lambda^*(\delta)$ depends on $(\delta, \Sigma, \theta_0)$. (1) If $\delta \leq \delta_1$, then $\lambda^*(\delta) = 0$ such that $\theta^* = \theta_0$, and there is no stationary point for $R_0(\theta, \delta)$. (2) If $\delta \geq \delta_2$, then $\lambda^*(\delta) = \infty$ such that $\theta^* = \mathbf{0}$, and there is no stationary point for $R_0(\theta, \delta)$. (3) If $\delta_1 < \delta < \delta_2$, then there is a unique stationary point $\theta(\lambda^*(\delta))$ of $R_0(\theta, \delta)$, which is the global optimum. Here $\lambda^*(\delta)$ is the solution of the following equation w.r.t. λ:

$$\lambda\left(1 + \frac{\delta c_0 \|\theta(\lambda)\|}{\|\theta(\lambda) - \theta_0\|_\Sigma}\right) = \delta c_0 \frac{\|\theta(\lambda) - \theta_0\|_\Sigma}{\|\theta(\lambda)\|} + \delta^2. \tag{4}$$

The proof of Proposition 1 is in Appendix B.

For a general Σ, it is hard to obtain an explicit solution for θ^* by solving (4). However, when $\Sigma = \mathbf{I}_p$, one can write down the explicit formula of $\theta^*(\delta)$, which is actually a re-scaled version of θ_0. In this case, $\delta_1 = c_0$, $\delta_2 = 1/c_0$, and $\lambda^*(\delta) = (\delta^2 - \delta c_0)/(1 - \delta c_0)$ when $\delta \in (\delta_1, \delta_2)$. Moreover, the adversarial risk and standard risk of the adversarially robust model become

$$R_0(\theta^*(\delta), \delta) = \begin{cases} \delta^2 \|\theta_0\|^2 & \delta \leq c_0 \\ \frac{\delta^2(1-c_0^2)}{\delta^2 + 1 - 2\delta c_0} \|\theta_0\|^2 & c_0 \leq \delta \leq 1/c_0 \\ \|\theta_0\|^2 & \delta \geq 1/c_0 \end{cases}$$

$$R_0(\theta^*(\delta), 0) = \begin{cases} 0 & \delta \leq c_0 \\ \frac{\delta^2(\delta - c_0)^2}{(\delta^2 + 1 - 2\delta c_0)^2} \|\theta_0\|^2 & c_0 \leq \delta \leq 1/c_0 \\ \|\theta_0\|^2 & \delta \geq 1/c_0 \end{cases}.$$

Similar as $R_0(\theta^*(\delta), \delta)$, the standard risk of the adversarially robust model $R_0(\theta^*(\delta), 0)$ also increases as δ and reaches the same level as $R_0(\theta^*(\delta), \delta)$ when $\delta > 1/c_0$; see Figure 1 below. This result echoes with Javanmard et al. (2020); Raghunathan et al. (2019) that the adversarially robust model leads to a worse performance when testing data is un-corrupted.

Remark 1. *Besides adversarial risk, we define adversarial* **prediction** *risk as*

$$R(\theta, \delta) := \mathbb{E}_{\mathbf{x}, y} \max_{\|\mathbf{x}^* - \mathbf{x}\| \leq \delta} \left[\left((\mathbf{x}^*)^\top \theta - y\right)^2\right].$$

The properties of R are similar as R_0 when $\epsilon \sim N(0, \sigma^2)$, and we focus on R_0 in this paper.

3 MINIMAX LOWER BOUND

In this section, through figuring out the minimax lower bounds of the estimation error, we argue that it is essential to incorporate sparsity information of (θ_0, Σ) in $(\widehat{\theta}_0, \widehat{\Sigma})$ in sparse model. For minimax lower bound in standard learning problems, studies can be found in Dicker et al. (2016); Mourtada (2019) for dense case and Verzelen (2010); Ye and Zhang (2010); Raskutti et al. (2011) for sparse case.

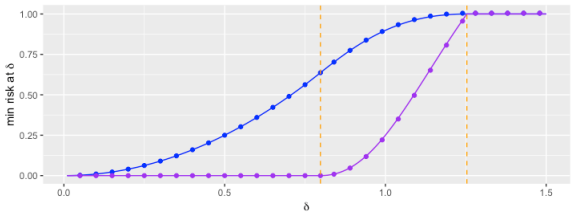

Figure 1: $R_0(\theta^*(\delta), \delta)$ and $R_0(\theta^*(\delta), 0)$ correspond to blue and purple curves, respectively. Here, $\Sigma = \mathbf{I}_p$ and $\|\theta_0\|^2 = 1$. Dashed lines represent the two thresholds $\delta_1 = c_0$ (left) and $\delta_2 = 1/c_0$ (right). Curve: theoretical values. Dots: simulations with $p = 10$ and $n = 10000$.

The following two theorems present the lower bounds of $\mathbb{E}\|\hat{\theta} - \theta^*\|^2$ for dense/sparse models respectively.

Theorem 1. *When $\sigma/\|\theta_0\| < \infty$, $\sigma^2 p/(\|\theta_0\|^2 n) \to 0$, and $(p \log^2 n)/n \to 0$, if $\|\theta_0\| \leq R$, $0 < c_1 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq c_2 < \infty$, $\delta > 0$, then there exists some constant $\delta > 0$ such that*

$$\inf_{\widehat{\theta}} \sup_{\Sigma, \theta_0, \delta} \mathbb{E}\|\widehat{\theta} - \theta^*\|^2 = \Omega\left(\frac{p\sigma^2}{n} \vee \frac{pR^2}{n}\right),$$

The estimator $\widehat{\theta}$ refers to any estimator $\widehat{\theta}(X, Y, \delta)$, and θ^* is a function of $(\theta_0, \Sigma, \delta)$.

For sparse model, the sparsity of θ_0 is directly controlled through the size of active set of θ_0. In terms of the sparsity of Σ, we follow Cai et al. (2010) to consider a family of sparse covariance matrix as follows:

$$\mathcal{F}_\alpha = \Big\{\Sigma : \max_j \sum_i \{|\sigma_{ij}| : |i - j| > k\} \leq M k^{-\alpha} \ \forall k, \\ \lambda_{\max}(\Sigma) \leq M_0, \ \lambda_{\min}(\Sigma) \geq m_0 > 0\Big\}.$$

Theorem 2. *When $\sigma/\|\theta_0\| < \infty$, if $\|\theta_0\| \leq R$ and $\|\theta_0\|_0 \leq s$, $0 < c_1 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq c_2 < \infty$, $\delta > 0$, then for any $0 < s < p$ and $\alpha > 0$, there exists some constant $\delta > 0$ such that*

$$\inf_{\widehat{\theta}} \sup_{\Sigma \in \mathcal{F}_\alpha, \theta_0, \delta} \mathbb{E}\|\widehat{\theta} - \theta^*\|^2 \\ = \Omega\left(s\sigma^2 \frac{1 + \log(p/s)}{n} \vee R^2 n^{-\frac{2\alpha}{2\alpha+1}}\right).$$

The proof of the above two theorems utilizes some tools in Mourtada (2019); Verzelen (2010); Cai et al.

(2010). A difficulty compared with existing literature in standard learning is that the relationship between θ_0 and θ^* is nonlinear, and θ^* further depends on Σ. The details are in Appendix C.

To compare Theorem 1 and 2, the lower bound for sparse model is much smaller than the one for dense model. This indicates a potential improvement for adversarially robust estimators if the algorithm can utilize the sparsity information (if there is). As discussed in Belkin et al. (2019); Xing et al. (2020), for the high-dimensional model, if we do not consider the sparsity information, the resulting model is not consistent in both standard and adversarially robust learning problems.

To compare with standard learning problem, the results in Theorem 1 and 2 are different from those in standard learning. Such a difference implies it is hard to train adversarially robust models. In standard learning, when $\sigma^2 = 0$, the lower bound is exactly zero since some estimators of θ_0 can achieve zero estimation error. However, when $\delta > 0$, even if $\sigma^2 = 0$, the lower bound is not zero.

Remark 2. *Similar to our results, Dan et al. (2020) provided a minimax lower bound of generalization error under the adversarially robust classification setup. However, they only considered the dense case corresponding to our Theorem 1, but not for the sparse case.*

4 TWO-STAGE ADVERSARIAL ROBUST ESTIMATOR

In this section, we demonstrate a two-stage procedure for constructing adversarially robust estimators based on the explicit relation pointed out in the previous section. This relation allows us to incorporate specific model information, such as sparsity, into adversarially robust estimates through standard estimates. The idea of the proposed method is similar to the estimators in Dan et al. (2020); Carmon et al. (2019) and the method is straightforward. We emphasize that such a simple two-stage method is powerful enough to achieve minimax optimal.

4.1 Estimator description

There are two stages in the proposed method. In the first stage, consistent estimators of the true parameter θ_0, denoted as $\widehat{\theta}_0$, and matrix Σ, denoted as $\widehat{\Sigma}$, are obtained from standard statistical procedures. In the second stage, the robust estimator of θ^*, which minimizes the adversarial risk, is constructed as follows:

$$\widehat{\theta}(\delta) := \widehat{\theta}(\widehat{\lambda}^*(\delta)) := (\widehat{\Sigma} + \widehat{\lambda}^*(\delta)\mathbf{I}_p)^{-1}\widehat{\Sigma}\widehat{\theta}_0, \tag{5}$$

where $\widehat{\lambda}^*(\delta)$ is a plug-in estimate of $\lambda^*(\delta)$ depending on $\widehat{\theta}_0$ and $\widehat{\Sigma}$. Alternatively speaking, $\widehat{\theta}(\delta)$ may be obtained by minimizing an empirical version of (2):

$$\begin{aligned}\widehat{R}_0(\theta,\delta) &:= \widehat{R}_0(\theta,\widehat{\theta}_0,\widehat{\Sigma},\delta) \\ &= \|\theta-\widehat{\theta}_0\|_{\widehat{\Sigma}}^2 + 2\delta c_0\|\theta-\widehat{\theta}_0\|_{\widehat{\Sigma}}\|\theta\| + \|\theta\|^2.\end{aligned} \tag{6}$$

According to the proof of Proposition 1, the empirical risk $\widehat{R}_0(\theta,\delta)$ shares similar properties as adversarial risk $R_0(\theta,\delta)$ in Proposition 1. We may simply use $\widehat{\theta}$ instead of $\widehat{\theta}(\delta)$ when no confusion arises.

4.2 Consistency

We first show that for any level of attack δ, the adversarial excess risk converges to zero, i.e., (7), as long as the standard estimates of θ_0 and Σ are consistent with proper rates and p does not grow too fast. Next, combining with the convex properties of R_0, the upper bound in (7) implies the consistency of $\widehat{\theta}$ in estimating θ^*; see Theorem 4. This consistency result will be used in deriving the generalization error in Theorem 5 later.

Theorem 3. *For any consistent estimators $\widehat{\theta}_0$ and $\widehat{\Sigma}$, with probability tending to 1,*

$$\begin{aligned}&\sup_{\delta\geq 0}\left|R_0(\theta^*(\delta),\delta) - R_0(\widehat{\theta}(\delta),\delta)\right| \\ =\ & O\left(\|\widehat{\theta}_0-\theta_0\|\|\theta_0\|\right) + O\left(\|\theta_0\|^2\sqrt{\|\widehat{\Sigma}-\Sigma\|}\right).\end{aligned} \tag{7}$$

To illustrate Theorem 3 in details, we use $\widehat{\theta}_0 = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ and $\widehat{\Sigma} = \mathbf{X}^\top\mathbf{X}/n$ to construct $\widehat{\theta}$. Based on Theorem 2 in Hsu et al. (2012) (taking ridge penalty as zero) and Theorem 3, with probability tending to 1, we have

$$\frac{R_0(\widehat{\theta},\delta) - R_0(\theta^*,\delta)}{\|\theta_0\|_\Sigma^2 + \sigma^2} = o(1), \tag{8}$$

which implies the adversarial excess risk of $\widehat{\theta}$ converges to zero as long as $(p\log n)/n \to 0$.

The proof of Theorem 3 is postponed to Appendix C. We also postpone an analog of Theorem 3 for the adversarial prediction risk R to Appendix A (for the statement) and C (for the proof). Note that the upper bound in (7) is not tight, but enough to justify the adversarial risk consistency of $\widehat{\theta}(\delta)$.

We next use an example to illustrate how sparsity information can be utilized in the proposed framework.

Example 1 (Sparse Standard Estimates)**.** *Assume matrix belongs to the family $\mathcal{F}_\alpha$, then using the sparse estimator $\widehat{\Sigma}$ in Cai et al. (2010), we have*

$$\mathbb{E}\|\widehat{\Sigma}-\Sigma\|^2 = O\left(n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}\right).$$

Assume $\widehat{\theta}_0$ is the LASSO estimate obtained under proper penalization. Denote $s < n$ as the number of nonzero coefficients in θ_0. When $\mathbf{x}$ follows Gaussian and the noise ϵ satisfies $\mathbb{E}\exp\{t\epsilon^2\} < \infty$ for some $t > 0$, based on Bickel et al. (2009); Jeng et al. (2018), we have with probability tending to 1,

$$\|\widehat{\theta}_0 - \theta_0\| = O\left(\sigma\sqrt{\frac{s\log p}{n}}\right).$$

Therefore, (8) holds under weaker conditions, say $(\sigma s\log p)/n \to 0$ and $(\log p)/n \to 0$. On the other hand, we point out that $\widehat{\theta}$ (θ^*) does not inherit the sparsity of $\widehat{\theta}_0$ (θ_0) according to (5) and (3).

4.3 Bahadur representation and convergence rate

We next study statistical properties of the adversarially robust estimator $\widehat{\theta}$ by establishing its Bahadur representation He and Shao (1996) that implies asymptotic normality in some cases.

Theorem 4. *Assume both $\|\widehat{\theta}_0-\theta_0\|/\|\theta_0\|$ and $\|\widehat{\Sigma}-\Sigma\|$ converge to zero in probability.*

(1) If $\delta \in (\delta_1, \delta_2)$, then $\widehat{\theta} - \theta^*$ is a linear combination of $\widehat{\theta}_0 - \theta_0$ and $\widehat{\Sigma} - \Sigma$ in the main term:

$$\begin{aligned}&\widehat{\theta} - \theta^*\\ =\ &\mathbf{M}_1(\theta^*, \theta_0, \Sigma)(\widehat{\theta}_0 - \theta_0)\\ &+(\theta^* - \theta_0)^\top(\widehat{\Sigma} - \Sigma)(\theta^* - \theta_0)\mathbf{M}_2(\theta^*, \theta_0, \Sigma)\\ &+\mathbf{M}_3(\theta^*, \theta_0, \Sigma)(\widehat{\Sigma} - \Sigma)(\theta^* - \theta_0) + o_p(\|\widehat{\theta} - \theta^*\|),\end{aligned}$$

where $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$ are functions of $(\delta, \theta_0, \Sigma, \theta^*)$, and detailed formulas are postponed to Appendix A.

(2) If $\delta < \delta_1$, then $\widehat{\theta} - \theta^* = \widehat{\theta}_0 - \theta_0 + o_p(\|\widehat{\Sigma} - \Sigma\|) + o_p(\|\widehat{\theta}_0 - \theta_0\|)$.

(3) If $\delta > \delta_2$, we have $\widehat{\theta} - \theta^* = o_p(\|\widehat{\Sigma} - \Sigma\|) + o_p(\|\widehat{\theta}_0 - \theta_0\|)$.

The proof for Theorem 4 is postponed to Appendix C. We next illustrate how the Bahadur representation can be used to infer the asymptotic normality of $\widehat{\theta}$.

Example 2 (Least Square Estimate)**.** *Consider the least square estimate (OLS)*

$$\widehat{\theta}_0 = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}, \qquad \widehat{\Sigma} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}.$$

It is trivial to see that $\widehat{\theta} = \mathbf{0}$ in probability when $\delta > \delta_2$ based on Theorems 1 and 4. When $\delta \in [0, \delta_1)$, the asymptotic normality of $\sqrt{n/p}(\widehat{\theta} - \theta^*)$ trivially follows the fact that $\widehat{\theta} = \widehat{\theta}_0$ in probability and $\theta^* = \theta_0$. When $\delta \in (\delta_1, \delta_2)$,

$$\widehat{\theta} - \theta^* = \mathbf{m}_1 + \mathbf{m}_2 + \mathbf{m}_3 + o_p(\|\widehat{\theta} - \theta^*\|),$$

where

$$\begin{aligned}\mathbf{m}_1 &= \mathbf{M}_1\left[\frac{\Sigma^{-1}}{n}\sum_{i=1}^n x_i\epsilon_i\right],\\ \mathbf{m}_2 &= \mathbf{M}_2\left[\frac{1}{n}\sum_{i=1}^n(\theta^* - \theta_0)^\top(x_ix_i^\top - \Sigma)(\theta^* - \theta_0)\right],\\ \mathbf{m}_3 &= \mathbf{M}_3\left[\frac{1}{n}\sum_{i=1}^n(x_ix_i^\top - \Sigma)(\theta^* - \theta_0)\right].\end{aligned}$$

If p is fixed and $\delta \in (\delta_1, \delta_2)$, then $\sqrt{n}(\widehat{\theta} - \theta^*)$ asymptotically converges to a zero-mean Gaussian. For inference purpose, we need to estimate $Var(\widehat{\theta})$. Since $x_i\epsilon_i$ and $(x_ix_i^\top - \Sigma)$ in $\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3$ are both i.i.d. random variables, and x_i follows Gaussian distribution, one can figure out the variance of $\widehat{\theta}$. As a result, replacing $(\theta^*, \theta_0, \Sigma, \delta)$ with $(\widehat{\theta}, \widehat{\theta}_0, \widehat{\Sigma}, \delta)$, one can obtain an estimate of $Var(\widehat{\theta})$. As a side remark, if p diverges in n, we have $\|\widehat{\theta} - \theta^*\|/\sqrt{\|\theta_0\|_\Sigma^2 + \sigma^2} = O_p(\sqrt{p/n})$.

Furthermore, when using dense/sparse estimators of (θ_0, Σ), our proposed two-stage estimator achieves minimax rate optimal in dense/sparse models respectively. The upper bound of $\mathbb{E}\|\widehat{\theta} - \theta^*\|^2$ can be developed from Theorem 4:

Corollary 1. *Denote $v^2 = \|\theta_0\|_\Sigma^2 + \sigma^2$. When $(p\log n)/n \to 0$, $\widehat{\theta}_0$ is the OLS estimate, and $\widehat{\Sigma}$ is the sample matrix, we have*

$$\mathbb{E}\|\widehat{\theta} - \theta^*\|^2 = \Theta\left(\frac{v^2p}{n}\right).$$

Combining upper bound result in the above corollary and lower bound in Theorem 1 together, one can see that using OLS estimate as $\widehat{\theta}_0$ and sample covariance matrix as $\widehat{\Sigma}$ in the two-stage method reaches minimax optimal in dense models. Besides, as stated in the following result, using the sparse estimators in Example 1, our proposed two-stage estimator reaches the minimax rate as in Theorem 2:

Corollary 2. *For sparse models, when $(\log p)/n \to 0$, $\sigma^2(s\log p)/(n\|\theta_0\|^2) \to 0$, $\widehat{\theta}_0$ is the LASSO estimate and $\widehat{\Sigma}$ is the sparse covariance estimator in Cai et al. (2010), it satisfies that*

$$\mathbb{E}\|\widehat{\theta} - \theta^*\|^2 = O\left(\frac{s\sigma^2\log p}{n} + v^2n^{-\frac{2\alpha}{2\alpha+1}}\right).$$

If $\log_s(p) > 1 + c_s$ for some constant $c_s > 0$, the above results are minimax-optimal.

5 PROPERTIES OF THE METHOD

This section provides additional properties of the proposed method beyond the consistency and convergence rate. In particular, we use theorems associated

with our method to verify two arguments in the existing literature: (1) generalization of adversarially robust learning is worse than standard learning; (2) one can improve the generalization of adversarially robust learning through utilizing extra unlabeled data.

5.1 Adversarial learning hurts generalization

We study the generalization of our proposed estimator. From the minimax lower bound theorems in Section 3, it is easy to see that the excess risk when $\delta > 0$ may converge in a slower rate than the one when $\epsilon = 0$. Besides this, we work on the multiplicative constants of excess risk and generalization error and reveal that those constants are larger when $\epsilon > 0$ as well.

Based on Theorem 4, the generalization error (9) and the estimation error of minimal adversarial risk (10) can be decomposed as follows:

$$\begin{aligned} & R_0(\widehat{\theta},\delta) - \widehat{R}_0(\widehat{\theta},\delta) \qquad (9)\\ = \;& e_{1,\Sigma}(\widehat{\Sigma},\delta) + e_{1,\theta_0}(\widehat{\theta}_0,\delta) + o_p(R_0(\widehat{\theta},\delta) - \widehat{R}_0(\widehat{\theta},\delta)),\\ & R_0(\theta^*,\delta) - \widehat{R}_0(\widehat{\theta},\delta) \qquad (10)\\ = \;& e_{2,\Sigma}(\widehat{\Sigma},\delta) + e_{2,\theta_0}(\widehat{\theta}_0,\delta) + o_p(R_0(\theta^*,\delta) - \widehat{R}_0(\widehat{\theta},\delta)). \end{aligned}$$

The term e_{j,θ_0} ($e_{j,\Sigma}$) represents the error component that is *only* caused by the estimation error of $\widehat{\theta}_0$ ($\widehat{\Sigma}$). We next characterizes the forms of $e_{j,\Sigma}$ and e_{j,θ_0} with precise multiplicative constants.

Theorem 5. *Under the same conditions as in Proposition 1, if $\|\widehat{\Sigma}-\Sigma\| \to 0$ and $\|\widehat{\theta}_0-\theta_0\|/\|\theta_0\| \to 0$, then when $\delta < \delta_1$,*

$$\begin{aligned} e_{1,\Sigma}(\widehat{\Sigma},\delta) &= o_p(\|\widehat{\Sigma}-\Sigma\|\|\theta_0\|^2),\\ e_{1,\theta_0}(\widehat{\theta}_0,\delta) &= \|\widehat{\theta}_0-\theta_0\|_\Sigma^2 + 2c_0\delta\|\theta_0\|\|\widehat{\theta}_0-\theta_0\|_\Sigma\\ &\quad + o_p(\|\widehat{\theta}_0-\theta_0\|\|\theta_0\|),\\ e_{2,\Sigma}(\widehat{\Sigma},\delta) &= o_p(\|\widehat{\Sigma}-\Sigma\|\|\theta_0\|^2),\\ e_{2,\theta_0}(\widehat{\theta}_0,\delta) &= -2\delta^2\theta_0^\top(\widehat{\theta}_0-\theta_0) + o_p(\|\widehat{\theta}_0-\theta_0\|\|\theta_0\|). \end{aligned}$$

If $\delta > \delta_1$, we have

$$\begin{aligned} e_{1,\Sigma}(\widehat{\Sigma},\delta) &= -c_\Sigma(\delta)\frac{(\theta^*-\theta_0)^\top(\widehat{\Sigma}-\Sigma)(\theta^*-\theta_0)}{\|\theta^*-\theta_0\|_\Sigma^2}\\ &\quad + o_p(\|\widehat{\Sigma}-\Sigma\|\|\theta_0\|^2),\\ e_{1,\theta_0}(\widehat{\theta}_0,\delta) &= 2c_{\theta_0}(\delta)\frac{(\widehat{\theta}_0-\theta_0)^\top\Sigma(\theta^*-\theta_0)}{\|\theta^*-\theta_0\|_\Sigma}\\ &\quad + o_p(\|\widehat{\theta}_0-\theta_0\|\|\theta_0\|),\\ e_{2,\Sigma}(\widehat{\Sigma},\delta) &= e_{1,\Sigma}(\widehat{\Sigma},\delta) + o_p(\|\widehat{\Sigma}-\Sigma\|\|\theta_0\|^2),\\ e_{2,\theta_0}(\widehat{\theta}_0,\delta) &= e_{1,\theta_0}(\widehat{\theta}_0,\delta) + + o_p(\|\widehat{\theta}_0-\theta_0\|\|\theta_0\|). \end{aligned}$$

where the multiplicative constants $c_\Sigma(\delta) := \|\theta^*-\theta_0\|_\Sigma^2 + \delta c_0\|\theta^*\|\|\theta^*-\theta_0\|_\Sigma$ and $c_{\theta_0}(\delta) := \|\theta^*-\theta_0\|_\Sigma + \delta c_0\|\theta^*\|$ are monotone increasing functions in δ. Recall that θ^* is a function of δ.

The proof of Theorem 5 is postponed to Appendix C.

To better understand Theorem 5, we plot the changes of $|e_{1,\theta_0}|$, $|e_{1,\Sigma}|$, and $|e_{2,\theta_0}|$ w.r.t. δ by assuming $\Sigma = \mathbf{I}_p$ in Figure 2. In the left plot, $|e_{1,\theta_0}|$ firstly increases in δ linearly until $\delta = \delta_1$, then jumps to the second regime and grows until it converges to $2|(\widehat{\theta}_0-\theta_0)^\top\Sigma\theta_0|$ after $\delta > \delta_2$. In the middle plot, $|e_{1,\Sigma}|$ is almost zero when $\delta < \delta_1$, then increases when $\delta \in (\delta_1,\delta_2)$ and finally converges when $\delta > \delta_2$. And, $|e_{2,\Sigma}|$ shares a similar pattern. The pattern of $|e_{2,\theta_0}|$ is similar as $|e_{1,\theta_0}|$ except that it smoothly transits into the second regime, as shown in the right plot. The empirical and theoretical curves match very well in Figure 2.

5.2 Reducing estimation error through additional unlabeled data

Unlabeled data is commonly used in semi-supervised learning, e.g. locally-weighted nearest neighbors algorithm (Cannings et al., 2020). Besides, in the context of adversarially robust learning, some studies also observed the benefits of using extra unlabeled data (Raghunathan et al., 2019).

We study the effect of extra unlabeled data on the minimax lower bounds and the upper bounds of our proposed method under different scenarios. With the existence of extra unlabeled data, the minimax lower bounds become smaller. Besides, these data also help reduce the upper bounds by improving the accuracy of $\widehat{\Sigma}$:

Theorem 6. *Under the conditions in Theorem 1, if there are extra n_1 samples of unlabeled data, the lower bound becomes $\Omega((p\sigma^2/n) \vee (pR^2/(n+n_1)))$.*

Under the conditions in Theorem 2, if there are extra n_1 samples of unlabeled data, the lower bound becomes

$$\Omega\left(s\sigma^2\frac{\log(p/s)}{n} \vee R^2(n+n_1)^{-\frac{2\alpha}{2\alpha+1}}\right).$$

In terms of the upper bounds, since the estimation of $\widehat{\Sigma}$ is only related to $\mathbf{x}$, one can directly utilize these extra unlabeled data into the two-stage framework. The following result is extended from Theorem 4:

Corollary 3. *Under the conditions in Corollary 1, if there are extra n_1 samples of unlabeled data, the upper bound becomes $O((p\sigma^2/n) \vee (pR^2/(n+n_1)))$.*

Under the conditions in Corollary 2, if there are extra n_1 samples of unlabeled data, the bound becomes

$$O\left(s\sigma^2\frac{\log(p/s)}{n} \vee R^2(n+n_1)^{-\frac{2\alpha}{2\alpha+1}}\right).$$

To summarize, as both lower bounds and upper bounds are reduced, it is essential to utilize extra un-

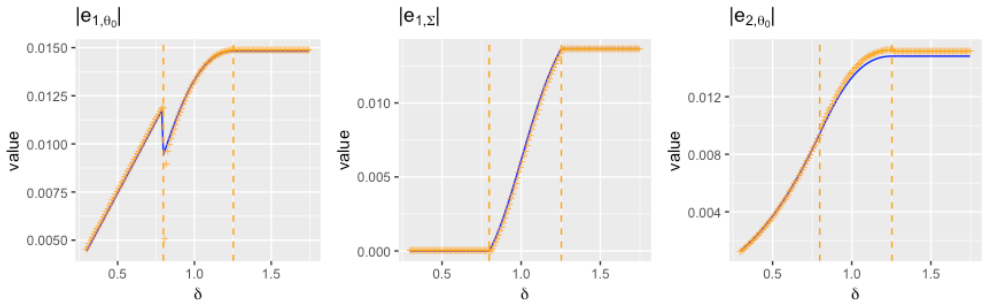

Figure 2: The value of $|e_{1,\theta_0}|$, $|e_{1,\Sigma}|$, and $|e_{2,\theta_0}|$ as functions of δ. Assume $\|\theta_0\| = 1$, $\Sigma = \mathbf{I}_p$. Blue curve is obtained from Theorem 5 given $(\widehat{\theta}_0, \widehat{\Sigma})$. Orange points are obtained from simulation. $n = 1000$, $\sigma^2 = 1$. The two vertical dashed lines in each figure represent δ_1 and δ_2.

labeled data for adversarially robust learning. A numerical illustration is also given in the next section of the experiments.

6 NUMERICAL EXPERIMENTS

In numerical experiments, we consider Example 2, and adopt LASSO/sparse estimators in the first stage to improve adversarial robustness.

We consider the following specifications of (θ_0, Σ): θ_0 is randomly generated from $\partial B(0,1)$, the sphere of a $\mathcal{L}_2$ ball; the diagonal elements in Σ are $\Sigma_{ii} = 2r + |\tau_i|$, where τ_i's follow i.i.d. standard Gaussian, and the other elements in Σ are r. Under this design of Σ, coordinates of $\mathbf{x}$ are correlated with each other, and the smallest and largest eigenvalues are within a reasonable range as p increases. Each experiment was repeated 500 times with $\sigma^2 = 1$. Define $\widehat{\Sigma} = \mathbf{X}^\top \mathbf{X}/n$ for non-sparse Σ.

Empirical coverage when p is fixed. As mentioned in Example 2, $\sqrt{n}(\widehat{\theta} - \theta^*)$ asymptotically converges to a zero-mean Gaussian when $\delta < \delta_2$. We use empirical coverage to verify this statement. In this experiment, $\theta_0 = (1,2)^\top$ and $\Sigma_{ii} = i$ for $i = 1, 2$ with $\Sigma_{12} = 0.5$. For each δ, we repeat the experiment of estimating θ^* for 1000 times using 1000 samples, and calculate the 95% empirical coverage for θ_1^* and θ_2^*. In Figure 3, when $\delta < 1.9$, the magnitude of θ_i^*'s are away from zero, and the empirical coverage for both θ_i^*'s are close to 0.95. When $\delta > 1.9$, θ_i^*'s are almost zero, and the corresponding empirical coverages are a little bit away from 0.95.

Sparse coefficients. In this experiment, we verify that LASSO helps to obtain a better adversarially robust estimate. We take $p = 50$, $n = 300$, and assume Σ is known. Cross-validation is applied to choose the penalty that minimizes the (standard) prediction risk. This is implemented by library `glmnet` in R.

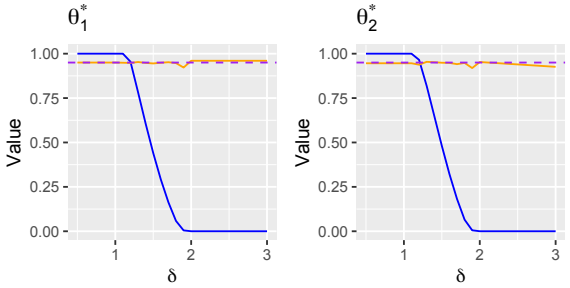

Figure 3: Value of θ_i^* and the 95% Empirical Coverage. Blue line: θ_i^*/θ_{0i}. Orange line: 95% Empirical Coverage. Purple dased line: 0.95. The 95% coverage for both θ_1^* and θ_2^* are close to 0.95 when $\delta < 1.9$.

We consider both lower-dimensional dense (Table 1) case with $(p, n) = (50, 300)$ and high-dimensional sparse scenario with $(p, n) = (300, 200)$. For high-dimensional sparse model, to make it clear on the difference between $\widehat{\theta}_{OLS}$ and $\widehat{\theta}_{LASSO}$, we present the results given Σ is known/unknown. In the dense coefficient model, although we can select a λ such that the LASSO estimator leads to a smaller standard risk than the OLS estimator, its corresponding adversarial risk gets worse with an increasing δ. For the sparse model, for all choices of δ, LASSO has a smaller adversarial risk than OLS. The results for unknown Σ are similar to the case when Σ is known, in the sense that LASSO is also better than OLS.

In addition, $R_0(\widehat{\theta}_{LASSO}, \delta)$ is always smaller when Σ is known than when Σ is unknown. This also verifies that unlabeled data helps improve the adversarial robustness (the comparison is not applicable to $R_0(\widehat{\theta}_{OLS}, \delta)$ since $\widehat{\theta}_{OLS}$ is not consistent).

Sparse matrix. We use sparse matrix estimator to verify that it helps enhancing adversarial robustness. To generate sparse matrix, we consider Σ such that $\Sigma_{ii} = 1$, and $\Sigma_{ij} = r|i-j|^{-\alpha-1}$ when $j \neq i$, where $r = 0.6$ and $\alpha = 0.2$. This choice of (r, α) ensures that all eigenvalues of Σ are positive. We take $p = 300$, $n =$

Table 1: Comparison between OLS and LASSO for dense θ_0 with known Σ. $p = 50$, $n = 300$, $r = 0.1$, $\sigma^2 = 1$. Σ is known. Standard deviation is provided for $R_0(\theta^*, \delta) - R_0(\widehat{\theta}_{OLS}, \delta)$ and $R_0(\theta^*, \delta) - R_0(\widehat{\theta}_{LASSO}, \delta)$.

δ	$R_0(\theta^*, \delta)$	$R_0(\widehat{\theta}_{OLS}, \delta)$	$R_0(\widehat{\theta}_{LASSO}, \delta)$	$R_0(\theta_0, \delta)$	$R_0(0, \delta)$
0.5	0.2489	0.8545(0.1413)	**0.633**(0.0795)	0.25	0.9997
0.8	0.5847	**0.8436**(0.0867)	0.8516(0.0858)	0.64	0.9997
0.9	0.6862	**0.8715**(0.65)	0.8888(0.0762)	0.81	0.9997

Table 2: Comparison between OLS and LASSO for sparse θ_0 . The first 10 elements of θ_0 are $1/\sqrt{10}$. $p = 300$, $n = 200$, $r = 0.1$, $\sigma^2 = 1$.

Σ	δ	$R_0(\theta^*, \delta)$	$R_0(\widehat{\theta}_{OLS}, \delta)$	$R_0(\widehat{\theta}_{LASSO}, \delta)$	$R_0(\theta_0, \delta)$	$R_0(0, \delta)$
known	0.5	0.25	6.1134(1.0171)	**0.7486**(0.1200)	0.25	1.8943
	1	0.7847	2.7114(0.4124)	**0.9941**(0.0752)	1	1.8943
	2	1.3088	1.4912(0.0431)	**1.3684**(0.0453)	4	1.8943
	3	1.6088	1.7522(0.0641)	**1.6435**(0.1033)	9	1.8943
unknown	0.5	0.25	2.3533(0.2551)	**0.8212**(0.0984)	0.25	1.8943
	1	0.7847	1.5830(0.1368)	**1.1414**(0.0732)	1	1.8943
	2	1.3088	1.5023(0.0341)	**1.4716**(0.0358)	4	1.8943
	3	1.6088	1.7040(0.0250)	**1.6930**(0.0494)	9	1.8943

Table 3: Comparison between $\widehat{\Sigma}$ and $\widehat{\Sigma}_{sparse}$. $p = 300$, $n = 200$, $\sigma^2 = 1$. θ_0 is known. $\widehat{\Sigma}_{sparse}$ performs slightly better when Σ is sparse.

$\delta = 2$	$R_0(\theta^*, \delta)$	$R_0(\theta^*_{\widehat{\Sigma}}, \delta)$	$R_0(\theta^*_{\widehat{\Sigma}_{sparse}}, \delta)$	$R_0(\theta_0, \delta)$
Dense	1.8865	**2.0576**(0.1841)	4.8769(0.1044)	4.0000
Sparse	2.9807	3.0652(0.0279)	**3.0293**(0.0279)	4.0000

200 so that the difference between $\widehat{\Sigma}$ and $\widehat{\Sigma}_{sparse}$ is obvious. The attack level δ is set to be 2 in this comparison. For simplicity, we assume θ_0 is known in the comparison of matrix estimators. The sparse covariance estimator $\widehat{\Sigma}_{sparse}$ was obtained based on the method in Cai et al. (2010). In Table 3, the adversarial excess risk is reduced from 0.0845 ($R_0(\theta^*_{\widehat{\Sigma}}, \delta) - R_0(\theta^*, \delta)$) to 0.0486 ($R_0(\theta^*_{\widehat{\Sigma}_{sparse}}, \delta) - R_0(\theta^*, \delta)$), which shows the effectiveness of $\widehat{\Sigma}_{sparse}$. In addition to the sparse matrix, we also consider dense covariance matrix generated in the same way as previous experiments by taking $r = 0.6$. When the true matrix is dense, using a sparse estimate is not appropriate; thus, the corresponding adversarial risk is much higher.

7 CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we figure out the minimax lower bound of estimation error of adversarially robust model in linear regression setup, which indicates the importance of incorporating model information in adversarially robust learning. In addition, we propose a two-stage adversarially robust learning method based on an explicit relation between adversarially robust estimator and standard estimator. The proposed two-stage estimator can encode model information (e.g., sparsity) into standard estimators, through which the robustness of adversarially robust estimator could be improved and reach minimax optimal convergence rate. Our investigation in the generalization error also verifies that adversarial robustness hurts generalization.

One future direction is to relax the distributional assumption on $(\mathbf{x}, y)$, say $\mathbf{x}$ follows non-Gaussian distribution. Although there is a wide range of data that may follow Gaussian assumption, e.g., abalone data and other biological data, many other data may not follow Gaussian, e.g., image data. The constant c_0 in our framework currently depends on the Gaussian assumption, and there is potential to relax it. Another direction is concerned with sparse adversarially robust learning, say sparse $\widehat{\theta}$, which could be useful in both compressing and robustifying deep neural networks Guo et al. (2018). The first step is to understand how the sparsity of θ_0 (together with other model assumptions) implies the sparsity of θ^*, which in turn determines the sparsity of $\widehat{\theta}$. An example can be found in Allen-Zhu and Li (2020) for linear sparse coding model. However, more careful studies would be needed in the future.

References

Allen-Zhu, Z. and Li, Y. (2020), "Feature Purification: How Adversarial Training Performs Robust Deep Learning," *arXiv preprint arXiv:2005.10190*.

Belkin, M., Hsu, D., and Xu, J. (2019), "Two models of double descent for weak features," *CoRR*, abs/1903.07571.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous analysis of Lasso and Dantzig selector," *The Annals of Statistics*, 37, 1705–1732.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010), "Optimal rates of convergence for covariance matrix estimation," *The Annals of Statistics*, 38, 2118–2144.

Cannings, T. I., Berrett, T. B., Samworth, R. J., et al. (2020), "Local nearest neighbour classification with applications to semi-supervised learning," *Annals of Statistics*, 48, 1789–1814.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. (2019), "Unlabeled data improves adversarial robustness," in *Advances in Neural Information Processing Systems*, pp. 11190–11201.

Dan, C., Wei, Y., and Ravikumar, P. (2020), "Sharp Statistical Guarantees for Adversarially Robust Gaussian Classification," vol. abs/2006.16384.

Dicker, L. H. et al. (2016), "Ridge regression and asymptotic minimax estimation over spheres of growing dimension," *Bernoulli*, 22, 1–37.

Etmann, C., Lunz, S., Maass, P., and Schönlieb, C. (2019), "On the Connection Between Adversarial Robustness and Saliency Map Interpretability," in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, vol. 97 of *Proceedings of Machine Learning Research*, pp. 1823–1832.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015), "Explaining and Harnessing Adversarial Examples," in *3rd International Conference on Learning Representations*.

Guo, Y., Zhang, C., Zhang, C., and Chen, Y. (2018), "Sparse dnns with improved adversarial robustness," in *Advances in neural information processing systems*, pp. 242–251.

He, X. and Shao, Q.-M. (1996), "A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs," *The Annals of Statistics*, 24, 2608–2630.

Hsu, D., Kakade, S. M., and Zhang, T. (2012), "Random design analysis of ridge regression," in *Conference on Learning Theory*, pp. 9–1.

Javanmard, A., Soltanolkotabi, M., and Hassani, H. (2020), "Precise Tradeoffs in Adversarial Training for Linear Regression," *CoRR*, abs/2002.10477.

Jeng, X. J., Peng, H., and Lu, W. (2018), "Post-Lasso Inference for High-Dimensional Regression," *CoRR, abs/1806.06304*.

Kurakin, A., Goodfellow, I. J., and Bengio, S. (2017), "Adversarial Machine Learning at Scale," in *5th International Conference on Learning Representations*, OpenReview.net.

Ma, S., Liu, Y., Tao, G., Lee, W.-C., and Zhang, X. (2019), "NIC: Detecting Adversarial Samples with Neural Network Invariant Checking." in *Network and Distributed System Security Symposium*.

Mei, S., Bai, Y., and Montanari, A. (2018), "The landscape of empirical risk for nonconvex losses," *The Annals of Statistics*, 46, 2747–2774.

Mourtada, J. (2019), "Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices," *arXiv preprint arXiv:1912.10754*.

Najafi, A., Maeda, S.-i., Koyama, M., and Miyato, T. (2019), "Robustness to adversarial perturbations in learning from incomplete data," in *Advances in Neural Information Processing Systems*, pp. 5542–5552.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. (2017), "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ACM, pp. 506–519.

Papernot, N., McDaniel, P., Swami, A., and Harang, R. (2016), "Crafting adversarial input sequences for recurrent neural networks," in *Military Communications Conference, MILCOM 2016-2016 IEEE*, IEEE, pp. 49–54.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J. C., and Liang, P. (2019), "Adversarial Training Can Hurt Generalization," *CoRR*, abs/1906.06032.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011), "Minimax rates of estimation for high-dimensional linear regression over Lq-balls," *IEEE transactions on information theory*, 57, 6976–6994.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. (2018), "Adversarially robust generalization requires more data," in *Advances in Neural Information Processing Systems*, pp. 5014–5026.

Shaham, U., Yamada, Y., and Negahban, S. (2015), "Understanding Adversarial Training: Increasing Local Stability of Neural Nets through Robust Optimization," *CoRR*, abs/1511.05432.

Sinha, A., Namkoong, H., and Duchi, J. (2018), "Certifiable Distributional Robustness with Principled Adversarial Training," in *International Conference on Learning Representations*.

Tao, G., Ma, S., Liu, Y., and Zhang, X. (2018), "Attacks meet interpretability: Attribute-steered detection of adversarial samples," in *Advances in Neural Information Processing Systems*, pp. 7717–7728.

Verzelen, N. (2010), "High-dimensional Gaussian model selection on a Gaussian design," in *Annales de l'IHP Probabilités et statistiques*, vol. 46, pp. 480–524.

Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., and Gu, Q. (2019), "On the convergence and robustness of adversarial training," in *Proceedings of the 35th International Conference on Machine Learning*, PMLR, vol. 97 of *Proceedings of Machine Learning Research*, pp. 6586–6595.

Xing, Y., Song, Q., and Cheng, G. (2020), "On the Generalization Properties of Adversarial Training," *arXiv preprint arXiv:2008.06631*.

Xu, H., Caramanis, C., and Mannor, S. (2009a), "Robust regression and lasso," in *Advances in neural information processing systems*, pp. 1801–1808.

— (2009b), "Robustness and Regularization of Support Vector Machines." *Journal of machine learning research*, 10.

Xu, H. and Mannor, S. (2012), "Robustness and generalization," *Machine learning*, 86, 391–423.

Xu, W., Evans, D., and Qi, Y. (2018), "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks," in *25th Annual Network and Distributed System Security Symposium*, The Internet Society.

Ye, F. and Zhang, C.-H. (2010), "Rate minimaxity of the Lasso and Dantzig selector for the lq loss in lr balls," *The Journal of Machine Learning Research*, 11, 3519–3540.

Yin, D., Ramchandran, K., and Bartlett, P. L. (2019), "Rademacher Complexity for Adversarially Robust Generalization," in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, vol. 97 of *Proceedings of Machine Learning Research*, pp. 7085–7094.

Zhai, R., Cai, T., He, D., Dan, C., He, K., Hopcroft, J. E., and Wang, L. (2019), "Adversarially Robust Generalization Just Requires More Unlabeled Data," *CoRR*, abs/1906.00555.

Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., and Xu, W. (2017), "Dolphinattack: Inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ACM, pp. 103–117.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. (2019), "Theoretically Principled Trade-off between Robustness and Accuracy," in *Proceedings of the 36th International Conference on Machine Learning*, PMLR, vol. 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482.