Unlabeled Data Help: Minimax Analysis and Adversarial Robustness

Yue Xing
Purdue University
xing49@purdue.edu

Qifan Song
Purdue University
qfsong@purdue.edu

Guang Cheng
University of California, Los Angeles
guangcheng@ucla.edu

Abstract

The recent proposed self-supervised learning (SSL) approaches successfully demonstrate the great potential of supplementing learning algorithms with additional unlabeled data. However, it is still unclear whether the existing SSL algorithms can fully utilize the information of both labelled and unlabeled data. This paper gives an affirmative answer for the reconstruction-based SSL algorithm (Lee et al., 2020) under several statistical models. While existing literature only focuses on establishing the upper bound of the convergence rate, we provide a rigorous minimax analysis, and successfully justify the rate-optimality of the reconstruction-based SSL algorithm under different data generation models. Furthermore, we incorporate the reconstruction-based SSL into the existing adversarial training algorithms and show that learning from unlabeled data helps improve the robustness.

1 Introduction

Modern learning algorithms (e.g., deep learning) have been the driving force of artificial intelligence. However, the success of these algorithms heavily relies on a huge volume of high-quality labeled training data, and these labeled data are expensive and not always available. To overcome this, the recently proposed idea of self-supervised learning (SSL) aims to supplement the training process with abundant unlabeled data, which are inexpensive and easily accessible (e.g., the images or footage captured by surveillance systems).

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

While SSL algorithms have gained increasing popularity, the corresponding theoretical investigations were not conducted until recent years. For instance, Lee et al. (2020); Teng and Huang (2021) reduce and reformulate the reconstruction-based SSL as a simple statistical model and show that it improves the estimation efficiency. Two other studies (Arora et al., 2019; Tosh et al., 2021) justify the effectiveness of contrastive learning in classification.

However, to our best knowledge, there is no existing study working on the fundamental information limits of SSL, i.e., the minimax lower bound of the estimation efficiency. Minimax lower bound does not directly inspire new methodologies, but it helps understand whether the existing methods achieve the best or not.

Minimax rate is the best possible convergence rate that can be achieved by **any** estimator in the worst case given finite samples, where "the worst-case" refers to the data distribution. Attaining minimax rate guarantees that the estimator achieves the best efficiency under the worst case. Failing to attain the minimax rate means that there must exist some scenarios where the estimator is not efficient. As a result, to understand the performance and optimality of SSL algorithms, it is essential to study the minimax bound.

Another motivation to study minimax lower bound is to understand the role of conditional independence (CI) for SSL. Lee et al. (2020) identifies that CI is a key factor yielding the estimation efficiency. A natural question would be how essential the CI condition is. Lee et al. (2020) suggests that the convergence rate of SSL estimate might be slower when CI does not hold. However, it is unclear whether it is caused by that the SSL is inefficient under conditional dependency, or that the fundamental information limit is worse under conditional dependency. To answer this question, minimax lower bound analysis is necessary.

Besides the minimax analysis of SSL, since the aforementioned works observe a great advantage of SSL over classical supervised learning methods, it is natural

to conjecture that the rationale behind SSL can potentially boost the performance of other techniques, e.g., adversarial training, differential privacy, and pruning. We consider adversarial training in this paper.

It is well known that deep learning models are vulnerable when they are fed with adversarial inputs (e.g., Zhang et al., 2017; Papernot et al., 2017). The adversarial inputs lead to serious concerns about AI safety. However, despite the existing literature in adversarial training with SSL that works on algorithm design and empirical study, e.g., Kim et al. (2020); Zeng et al. (2021); Ho and Vasconcelos (2020); Gowal et al. (2020); Chen et al. (2020b), there is little theoretical understanding towards this.

In summary, this paper aims to study two problems. First, we want to establish theoretical justifications on how and when the reconstruction-based algorithm in Lee et al. (2020) enhances the estimation efficiency in a training process. Second, we study an adversarially robust adaptation of this algorithm to show its effectiveness in the adversarial setting.

Model Setup To explain the details, we denote covariates $X_1 \in \mathbb{R}^{d_1}$, response $Y \in \mathbb{R}$, and some extra attributes $X_2 \in \mathbb{R}^{d_2}$. Following Lee et al. (2020), our target is to learn a model of Y from X_1. We use (X_1, X_2, Y) as the random variables, and (x_1, x_2, y) as observations. Assume we have the following datasets:

- Labeled Data: S_1 (n_1 samples of (x_1, x_2, y)) and S_2 (n_2 samples of (x_1, y)).
- Extra unlabeled data: S_3 (n_3 samples of (x_1, x_2)) and S_4 (n_4 samples of (x_1)).

In terms of the estimation procedure, in clean training, it can be summarized as follows:

- Pretext task: Learn some representation $\widehat{\phi} := \arg\min_{\phi \in \mathcal{F}} \frac{1}{n_1+n_3} \sum_{S_1,S_3} \|x_2 - \phi(x_1)\|_2^2$ for the mapping from x_1 to x_2, where $\mathcal{F}$ is some function space to be defined.
- Downstream task: estimate the coefficient $\widehat{W} = \arg\min_W \frac{1}{n_1+n_2} \sum_{S_1,S_2} l(W^\top \widehat{\phi}(x_1), y)$ for some loss function l depending on the specific task, e.g. regression, logistic regression. For classification task, the final classifier is $\mathrm{sgn}(\widehat{W}^\top \widehat{\phi}(x_1))$.

To set a concrete example, we want to predict the gender y of a person using the hairstyle x_1 and train from a set of front photos. One can first train a regression model from the hairstyle x_1 to predict the face x_2 (pretext task), then use this regression model to make a prediction for each photo, and finally, use the predicted face picture $\widehat{\phi}(x_1)$ to train on the gender label (downstream task). The gender label is not used in the pretext task, so we can use many unlabeled photos for the regression model in the pretext task.

In classical learning methods, the relationship between y and x_1 is learned solely based on the training data pairs of (x_1, y). The information from x_2 and the relationship between (x_1, x_2) are overlooked. The above reconstruction-based method utilizes these data so that it potentially improves the estimation efficiency. Note that we consider two conditions for possible extra data sets besides S_1: (1) whether the data are labeled or not, and (2) whether the data contain X_2. Thus, for the sake of completeness, it is natural to include all S_1 to S_4 in our framework of minimax analysis, despite that the usual clean SSL algorithm doesn't utilize data S_4. We will show that S_4 indeed does not affect the rate of the minimax lower bound.

Besides, as mentioned above, Lee et al. (2020) also discusses the importance of CI in SSL. The CI condition is formally defined as follows:

Definition 1. *The data generation model satisfies conditional independence if X_1 and X_2 are conditionally independent given $Y = y$.*

Contributions Our contributions are as follows:

First, we explain from a minimax perspective that the SSL method is generally well-behaved in clean training for classification models. We provide detailed characterizations in how the four datasets (S_1 to S_4) affect the estimation efficiency. From both lower bound and upper bound aspects, the size of S_1, S_2, and S_3 affects the efficiency, and whether SSL improves the efficiency depends on the comparison between $n_1 + n_2$ and $n_1 + n_3$. In addition, towards the CI condition, we reveal that when $n_1 + n_3$ is large enough, no matter whether CI holds or not, SSL is minimax optimal under a well-designed family of $\mathcal{F}$, though the rate is slower without CI. As for dataset S_4, its sample size n_4 does not affect the minimax convergence rate.

Secondly, we adapt the "pseudolabel" method (Carmon et al., 2019) so that adversarial training achieves the minimax lower bound in classification. We figure out the minimax lower bound of the convergence with the presence of unlabeled data (i.e., S_1 to S_4) and propose a way so that adversarial training achieves this lower bound with the help of SSL under the proper design of pseudolabel imputation. Again, when $n_1 + n_3$ is sufficiently large, SSL improves the adversarial robustness compared to a vanilla adversarial training.

Finally, as a by-product, we provide discussions about SSL for regression with ϕ being a linear function and SSL for classification with ϕ being a two-layer ReLU neural network (with lazy training). For the former one, we establish similar minimax results as the above.

For the latter one, we show that the neural network family serves as a good candidate for $\mathcal{F}$ in the pretext task when there is no parametric knowledge for $\mathcal{F}$. It can potentially accelerate the convergence.

2 Related Works

Below is a summary of other related articles in the areas of self-supervised learning, adversarial training, as well as statistical minimax lower bound analysis.

Self-Supervised Learning There are two popular types of self-supervised learning algorithms in the literature, i.e., reconstruction-based SSL and contrastive learning. Reconstruction-based SSL learns the reconstruction mapping from the large pool of unlabeled images and then employs it to reconstruct labeled images which are used in the downstream task (Noroozi and Favaro, 2016; Zhang et al., 2016; Pathak et al., 2016; Doersch et al., 2015; Gidaris et al., 2018). Contrastive learning uses the unlabeled images to train representations that distinguish different images invariant to non-semantic transformations (Mikolov et al., 2013; Oord et al., 2018; Arora et al., 2019; Dai and Lin, 2017; Chen et al., 2020a; Tian et al., 2020; Chen et al., 2020a; Khosla et al., 2020; HaoChen et al., 2021; Chuang et al., 2020; Xiao et al., 2020; Li et al., 2020).

Adversarial Training Many works consider the adversarial robustness of learning algorithms from different perspectives, e.g., the statistical properties or generalization performance of the global optimum of some well-designed adversarial loss function (Mehrabi et al., 2021; Javanmard et al., 2020; Javanmard and Soltanolkotabi, 2020; Dan et al., 2020; Taheri et al., 2020; Yin et al., 2018; Raghunathan et al., 2019; Schmidt et al., 2018; Najafi et al., 2019; Zhai et al., 2019; Hendrycks et al., 2019), or the algorithmic properties of optimizing the adversarial loss function (Sinha et al., 2018; Gao et al., 2019; Zhang et al., 2020; Allen-Zhu and Li, 2020; Xing et al., 2021a).

Related studies about semi-supervised learning with unlabeled data can be found in deep learning and other areas. For example, Carmon et al. (2019); Xing et al. (2021b) verify that unlabeled data helps in improving the estimation efficiency of adversarially robust models. Cannings et al. (2017) use unlabeled data to construct the local k-Nearest Neighbors algorithm.

Minimax Lower Bound Minimax lower bound is an important property in the area of statistics, and has been studied for different models, e.g. non-parametric model, linear regression, LASSO, as well as adversarially robust estimate (Audibert and Tsybakov, 2007; Raskutti et al., 2012; Yang and Tokdar, 2015; Sun et al., 2016; Dicker et al., 2016; Cai et al., 2010; Mourtada, 2019; Tony Cai and Zhang, 2019; Dan et al., 2020; Xu et al., 2020; Xing et al., 2021b).

3 Minimax Lower Bound

To reconcile the notation for both clean and adversarial training, for binary classification, we denote risk $R(f, \epsilon)$ as the population misclassification rate of the classifier $\text{sgn}[f(\widetilde{X}_1)]$, where $\widetilde{X}_1$ is the attacked input variable under strength ϵ. Specifically, given f and $X_1 = x_1$, $\widetilde{X}_1 = \arg\max_{z \in \mathcal{R}(x_1, \epsilon)} l(f(z), y)$ where l is the loss function for training[1]. The constraint $\mathcal{R}(x_1, \epsilon)$ is an $\mathcal{L}_2$ or $\mathcal{L}_\infty$ ball centering at x_1 with radius ϵ. Define $R^*(\epsilon) = \inf_f R(f, \epsilon)$ as the optimal misclassification rate under ϵ. To train a classifier, one minimizes an empirical loss function, where the loss can be different from R, e.g., square loss or cross-entropy.

To regulate the distribution of (X_1, X_2, Y), we impose the following assumption:

Assumption 1. *The distribution family $\mathcal{P}$ satisfies:*

(1) There is some known function $p(\cdot;\cdot)$ such that, any distribution in $\mathcal{P}$ satisfies $P(Y = 1|X_1 = x_1) = p(x_1; \beta)$ for some $\beta \in \mathbb{R}^{d_1}$;

(2) Assume (X_1, X_2, Y) satisfies (1) with $\beta = \beta^*$, then $R(2p(\cdot; \beta) - 1, \epsilon)$ is L-Lipschitz and is twice differentiable in β when $\beta \in B(\beta^*, r)$ for some small $r > 0$ for all $\epsilon = 0$ and the ϵ of interest.

The condition (1) in Assumption 1 is for the purpose of parametrization. Since we are doing parametric estimation, we need to consider the class of models whose parametric form exists. The generalized linear model is included by our assumption. The condition (2) in Assumption 1 describes how β is related to the misclassification rate. It should hold for $\epsilon = 0$ (clean training) and the ϵ of interest in adversarial training. We fix ϵ (which does not change with n) and do not consider it a changing parameter throughout training.

The following theorem presents the minimax lower bounds of the convergence of any estimator when CI holds/does not hold, for both clean ($\epsilon = 0$) and adversarial training ($\epsilon > 0$). Combining with upper bounds in the later section (i.e., Theorems 2, 3, 4), the presented rates indeed are optimal:

Theorem 1. *Assume Assumption 1 holds. Also assume $(d_1 + d_2)\log(n_i) = o(n_i^{1/3})$ for $i = 1, 2, 3$. The*

[1] In the models we consider in this paper, the attacks for l and $0 - 1$ loss are the same.

minimax lower bound is

$$\begin{aligned}&\inf_{\widehat{f}}\sup_{\mathcal{P}}\mathbb{E}R(\widehat{f},\epsilon)-R^*(\epsilon)\\ =\;&\Omega\left(\frac{d_1}{n_1+n_2}\wedge\left(\frac{d_2}{n_1+n_2}+\frac{d_1+d_2}{n_1+n_3}\right)\right).\end{aligned}$$

When CI holds, the lower bound becomes

$$\inf_{\widehat{f}}\sup_{\mathcal{P}}\mathbb{E}R(\widehat{f},\epsilon)-R^*(\epsilon)=\Omega\left(\frac{d_1}{n_1+n_2}\wedge\frac{d_1}{n_1+n_3}\right).$$

The proof of Theorem 1 is postponed to the appendix. To prove the minimax lower bound, one common way is to design a specific distribution so that the distribution parameters, e.g. mean and variance, always involve error given the finite training samples. The estimator $\widehat{f}$ will further inherit this error. Our examples used to prove the minimax lower bounds in Theorem 1 are more complicated compared to Dan et al. (2020); Xing et al. (2021b).

Although Lee et al. (2020) reveals that reconstruction-based SSL achieves a faster convergence rate under CI, the minimax lower bound gets much larger when CI does not hold based on our result. From this aspect, even if CI does not hold, the SSL algorithm is still good, and achieves the optimal convergence rate based on the results in later sections.

Besides, the rates are irrelevant to the sample size of S_4, indicating that the information of X_4 is not the bottleneck for this classification problem.

Remark 1. *Xing et al. (2021b) proves that introducing S_4 is helpful, and this does not contradict with our arguments in Theorem 1. Assume $Var(Y|X_1,X_2)$ and d_1 are all constants, and $n_1=n_3=0$, then the minimax lower bound in Xing et al. (2021b) is still $\Omega(1/n_2)$. The unlabeled data in S_4 improves the convergence in a multiplicative constant level, but not the rate of the convergence.*

Remark 2. *In this paper, we consider the upper bound and lower bound of $\mathbb{E}R(\widehat{f},\epsilon)-R^*(\epsilon)$. This is different from some literature in learning theory, e.g. Section 3 in Mohri et al. (2018), where they consider $R(f,\epsilon)-\sum l(f(x_1,x_2),y)/\sum n_i$. The former one focuses on the difference between the testing performance using the trained model and the true robust model, while the latter one considers the discrepancy between the training performance and the testing performance of the same model. Since we aim to study how the trained model performs compared to the true robust model, we use the former one in this paper. It is noteworthy that the latter one converges in a different rate from our results in this paper.*

4 Convergence Upper Bound

This section studies the convergence rate of SSL to see whether SSL achieves the optimal rate.

4.1 Convergence in Clean Training

We translate the results in Lee et al. (2020) into our format to match the minimax lower bounds above.

For the pretext task, under CI, we consider learning ϕ from the function space $\mathcal{F}:=\{\phi\mid\phi(x_1)=p(x_1;\beta)\mu_2+(1-p(x_1;\beta))\mu_2',\ \beta\in\mathbb{R}^{d_1},\ \mu_2,\mu_2'\in\mathbb{R}^{d_2}\}$, where p is the parametric form of $P_\beta(Y=1|X_1=x_1)$ as defined in Assumption 1. The rationale behind this choice of $\mathcal{F}$ is that, under CI condition, $\mathbb{E}[X_2|X_1=x_1]=P(Y=1|X_1=x_1)\mathbb{E}[X_2|Y=1]+P(Y=-1|X_1=x_1)\mathbb{E}[X_2|Y=-1]$, which matches the form of functions in $\mathcal{F}$. A concrete example will be provided in Example 1 later.

For the downstream task, we consider two estimators of W as follows. For both cases, the trained classifier is defined as $\mathrm{sgn}(\widehat{W}^\top\widehat{\phi}(x_1))$.

- Logistic regression on $(y,\widehat{\phi}(x_1))$.
- Plugin estimator in Dan et al. (2020), which is equivalent to square loss in clean training.

The following example analyzes the Gaussian mixture model when CI holds. It provides the basic analysis on how X_2 affects the convergence.

Example 1 (Classification under CI)**.** *Consider Gaussian mixture model defined as follows:*

$$\begin{aligned}&P(Y=1)=P(Y=-1)=\frac{1}{2},\\ &(X_1,X_2)|Y=y\sim N\left(y\begin{bmatrix}\mu_1^*\\ \mu_2^*\end{bmatrix},\begin{bmatrix}\Sigma_{1,1}^* & \Sigma_{1,2}^*\\ \Sigma_{2,1}^* & \Sigma_{2,2}^*\end{bmatrix}\right),\end{aligned}$$

where μ_i^*'s and $\Sigma_{i,j}^*$'s are unknown parameters. The conditional distribution of X_2 given $X_1=x_1,Y=1$ is

$$\begin{aligned}X_2|X_1=x_1,Y=1\quad\sim\quad&N(\mu_2^*+\Sigma_{2,1}^*(\Sigma_{1,1}^*)^{-1}(x_1-\mu_1^*),\\ &\Sigma_{2,2}^*-\Sigma_{2,1}^*(\Sigma_{1,1}^*)^{-1}\Sigma_{1,2}^*).\end{aligned}$$

Therefore, $\Sigma_{1,2}^*=\boldsymbol{0}$ is equivalent to CI condition in this model. Further, the probability $P(Y=1|X_1=x_1)$ is a function of x_1 and $(\Sigma_{1,1}^*)^{-1}\mu_1$ only, so the best ϕ to minimize $\mathbb{E}\|x_2-\phi(x_1)\|^2$ can be represented as $\phi^*(x_1)=(2p(x_1;(\Sigma_{1,1}^*)^{-1}\mu_1^*)-1)\mu_2^*$ under CI. Based on this, the family of ϕ, $\mathcal{F}=\{\phi\mid\phi(x)=(2p(x;\Sigma_{1,1}^{-1}\mu_1)-1)\mu_2,\ \forall\Sigma_{1,1}^{-1}\mu_1,\mu_2\}$ is a proper choice for the pretext task.

Solving the pretext task, we have

$$\begin{aligned}&\widehat{\phi}(x_1)=(2p(x_1;\widehat{\Sigma_{1,1}^{-1}\mu_1})-1)\widehat{\mu}_2,\ \textit{where}\\ &\mathbb{E}\|\widehat{\Sigma_{1,1}^{-1}\mu_1}-(\Sigma_{1,1}^*)^{-1}\mu_1^*\|^2=O\left(\frac{d_1}{n_1+n_3}\right).\end{aligned}$$

The detailed derivation is postponed to appendix.

In the downstream task, since the output of $\widehat{\phi}$ is always in the same direction (parallel to $\widehat{\mu}_2$), $\widehat{W}\widehat{\phi}(x_1)$ becomes $c(2p(x_1;\widehat{\Sigma_{1,1}^{-1}\mu_1})-1)$ for some constant c, and its sign only depends on $(2p(x_1;\widehat{\Sigma_{1,1}^{-1}\mu_1})-1)$. The estimation error in $\widehat{W}$ and $\widehat{\mu}_2$ therefore does not affect the final prediction, and the error in the prediction is only caused by the error in $\widehat{\Sigma_{1,1}^{-1}\mu_1}$. Consequently,

$$\mathbb{E}R(\widehat{W}^\top\widehat{\phi},0)-R^*(0)=O\left(\frac{d_1}{n_1+n_3}\right). \tag{1}$$

The proof for Example 1, and Theorem 2 and Theorem 3 below are postponed to the appendix. The basic idea is to use Taylor expansion on the estimation equation to obtain the Bahadur representation of the estimator. In general, via Bahadur representation, one can show that the estimator asymptotically converges to the true model in Gaussian.

The following theorem can be obtained via extending Example 1 to other models under CI:

Theorem 2. *Assume Assumption 1 together with some finite-variance condition (to be specified in the appendix) hold. If $(d_1+d_2)\log(n_i)=o(n_i^{1/3})$ for $i=1,2,3$, and $d_2=o(\sqrt{d_1(n_1+n_3)})$, then for both the two loss functions (logistic loss and square loss), when CI holds,*

$$\mathbb{E}R(\widehat{W}^\top\widehat{\phi},0)-R((W^*)^\top\phi^*,0)=O\left(\frac{d_1}{n_1+n_3}\right),$$

where ϕ^* is the population loss minimizer of the pretext task, and W^* is the population loss minimizer (logistic or square correspondingly) in the downstream task.

In contrast to Theorem 2, the following theorem studies the convergence of SSL when CI does not hold. For simplicity, we consider using linear ϕ in the pretext task, i.e., $\mathcal{F}$=$\{$linear mappings from $\mathbb{R}^{d_1}$ to $\mathbb{R}^{d_2}\}$

Theorem 3. *Assume Assumption 1 together with some finite-variance condition (to be specified in the appendix) hold. If $(d_1+d_2)\log(n_i)=o(n_i^{1/3})$ for $i=1,2,3$. For linear ϕ, if the singular values of $\mathbb{E}X_1X_2^\top$ are finite and bounded away from zero, then*

$$\begin{aligned}&\mathbb{E}R(\widehat{W}^\top\widehat{\phi},0)-R((W^*)^\top\phi^*,0)\\ =\;& O\left(\frac{d_2}{n_1+n_2}+\frac{d_1+d_2}{n_1+n_3}\right),\end{aligned}$$

where ϕ^* is the minimizer of the population loss of the pretext task, and W^* is the minimizer of the population loss in the downstream task.

Together with the lower bounds obtained in Theorem 1, the upper bounds in Theorem 2 and 3 indicate that SSL achieves minimax optimal for clean training, when $n_1+n_3\gg n_1+n_2$ under CI, or $n_1+n_3\gg(d_1+d_2)(n_1+n_2)/(d_1-d_2)$ without CI. This implies that SSL efficiently utilizes data information to achieve optimal convergence, while the deterioration of rate when CI fails is merely caused by information bottleneck of the data.

Furthermore, for commonly used model which only considers (X_1,Y), the upper bound is $O(d_1/(n_1+n_2))$, e.g. Xing et al. (2021a). Compared to this rate, the upper bounds in the Theorem 2 and 3 are faster when n_3 is large. These observations imply that the reconstruction-based SSL does perform better than only studying the relationship between X_1 and Y.

Simulation Study We use the model in Example 1 to numerically verify the effectiveness of SSL under CI. We take $d_1=5$, $d_2=2$, the mean vector $\mu=\mathbf{1}_{d_1+d_2}/\sqrt{d_1}$, and the covariance matrix $\Sigma=I_{d_1+d_2}$. We repeat 100 times to obtain the mean and variance of Regret. The sample size n_2 and n_4 are zero, and $n_1=100$. The results for plugin estimate are summarized in Table 1. From Table 1, SSL improves the performance when n_3 is large enough. The observations in logistic regression are similar (postponed to the appendix).

4.2 Adversarial Training

Intuitively, a straightforward way to adapt SSL in adversarially robust learning is to perform the downstream task with adversarial loss. However, a simple example below illustrates that such a procedure may lead to a bias:

Example 2. *Under the Gaussian mixture classification model in Example 1, following Dan et al. (2020), one can show that the population adversarial risk minimizer, for both the two loss functions, is a linear classifier whose coefficient vector is of the form $(A\Sigma_{1,1}^*+BI_{d_1})^{-1}\mu_1^*$. Using $\mathcal{F}$ considered in Example 1, the decision boundary implied from $W^\top\widehat{\phi}$ (for any W) is always parallel to $\widehat{(\Sigma_{1,1})^{-1}}\mu_1$ which is a biased estimation for $(A\Sigma_{1,1}^*+BI_{d_1})^{-1}\mu_1^*$ if $\Sigma_{1,1}^*$ is not proportional to I_{d_1} .*

To ensure the consistency of the adversarially robust estimator, one can borrow the idea of Carmon et al. (2019); Uesato et al. (2019): we first use SSL in clean training, and based on which, we create pseudolabel for data in S_3 and S_4, then we perform an adversarial training using S_1 to S_4 with the pseudolabels. Algorithm 1 summarizes this procedure.

With an abuse of notation, we denote $R(\theta,\epsilon)$ as the

Table 1: Regret in clean training under CI: SSL (plugin estimator/square loss) vs learning only labeled data.

n_3	SSL (mean)	labeled (mean)	SSL (var)	labeled (var)
500	0.01057	0.00959	7.07E-05	3.84E-05
1000	0.00529	0.01017	1.83E-05	7.53E-05
5000	0.00104	0.00970	2.01E-06	4.78E-05
10000	0.00042	0.00876	1.55E-06	4.88E-05
20000	0.00031	0.00974	6.94E-07	5.20E-05

Algorithm 1 Adversarial Training with SSL (adv+SSL)

Input: data S_i for $i = 1, ..., 4$. Adversarial training configuration $(\eta, T, \epsilon, ...)$.
Use S_1 and S_3 to obtain $\widehat{\phi}$.
Use S_1 and S_2 to obtain $\widehat{W}$.
Create pseudolabel for samples in S_3 and S_4 as $\widehat{y}$. Take $\widehat{y} = y$ for S_1, S_2.
Conduct adversarial training with configuration $(\eta, T, \epsilon, ...)$ to obtain $\widehat{\theta}$ where

$$\widehat{\theta} = \arg\min_\theta \frac{1}{\sum n_i} \sum_{S_1,...,S_4} \max_{\widetilde{x}_1 \in \mathcal{R}(x_1,\epsilon)} l(\theta^\top \widetilde{x}_1, \widehat{y}).$$

Output: the robust model $\widehat{\theta}$.

risk of the linear classifier $\text{sgn}(\theta^\top \widetilde{X}_1)$. Algorithm 1 focuses on the linear classifier $\theta^\top \widetilde{x}_1$, but in real practice, one may train a nonlinear model in the adversarial training stage (e.g., a neural network).

How to obtain reasonable pseudolabels? A key requirement of the pseudolabel is that the distribution of $(X_1, \widehat{Y})$ approximately matches (X_1, Y). Thus $\widehat{Y}$ is a simple plug-in estimator $p(x_1; \widehat{\beta})$ when we have the parametric form of p, i.e., $P(\widehat{Y} = 1|X_1 = x_1)$. The following Gaussian mixture model example illustrates how to construct pseudolables for unlabeled data:

Example 3 (Pseudolabel for Gaussian Mixture Model)**.** *When estimating ϕ, we are considering the class of function $\mathcal{F} = \{\phi \mid \phi(x_1) = (2p(x_1; \Sigma_{1,1}^{-1}\mu_1) - 1)\mu_2,\ \forall \Sigma_{1,1}^{-1}\mu_1, \mu_2\}$. An estimate of $P(Y = 1|X_1 = x_1)$, i.e. $p(x_1; \widehat{\Sigma_{1,1}^{-1}\mu_1})$, can be directly obtained from the pretext task. This construction method can also be applied to general models in $\mathcal{P}$.*

The following theorem evaluates the convergence rate of Algorithm 1 and shows its effectiveness:

Theorem 4. *Assume Assumption 1 and some finite-variance condition (in the appendix) hold.*

(I) Assume the conditions in Theorem 2 hold. Denote $\theta^* = \arg\min_\theta El(\theta^\top \widetilde{X}_1, Y)$ as the optimal linear classifier using square loss/logistic regression. Denote $\widehat{\theta}$ as the linear adversarially robust estimator obtained via Algorithm 1. If $\widehat{W}^\top \widehat{\phi}$ is unbiased, then for square loss/logistic regression,

$$\mathbb{E}R(\widehat{\theta}, \epsilon) - R(\theta^*, \epsilon) = O\left(\frac{d_1}{n_1 + n_3}\right).$$

(II) Assume the conditions in Theorem 3 hold, if $\widehat{W}^\top \widehat{\phi}$ is asymptotically unbiased,

$$\mathbb{E}R(\widehat{\theta}, \epsilon) - R(\theta^*, \epsilon) = O\left(\frac{d_2}{n_1 + n_2} + \frac{d_1 + d_2}{n_1 + n_3}\right).$$

The proof of Theorem 4 is postponed to the appendix. To build the connection between the clean training and the adversarial training, we borrow the idea from semi-parametric problems to expand $p(x_1; \widehat{\Sigma_{1,1}^{-1}\mu_1})$ in the Taylor expansion for the estimation equation, e.g., Wang and Wang (2009).

Theorem 4 shows the convergence rate of the estimator obtained in Algorithm 1. Again for adversarial training, SSL achieves minimax optimal when n_3 is large.

Effect of Accuracy of Imputed Labels Theorem 4 establishes the convergence rate of the whole procedure in Algorithm 1 where the SSL clean training stage helps estimate probability $p = P(Y = 1|X_1 = x_1)$. However, in real practice, when there is no parametric knowledge of the model (i.e., Assumption 1 fails), it is not easy to obtain an accurate $\widehat{p}$, and people may consider directly using the predicted label as the pseudolabel. The following result illustrates how the accuracy of $\widehat{p}$ affects the convergence in logistic regression. For square loss, the condition is slightly different, and we postpone the discussion to the appendix.

Proposition 1. *Under the conditions of Theorem 4(I), assume one obtains some consistent $\widehat{p}$ such that $\mathbb{E}\|X_1\|^2 \|\widehat{p}(X_1) - p(X_1)\|^2 \to 0$ in $n_1 + n_3$, then for logistic regression, (1) $\widehat{\theta}$ is consistent to θ^*; and (2) the convergence rate of $\widehat{\theta}$ is $O(d_1/(\sum n_i) + \mathbb{E}\|X_1\|^2 \|\widehat{p}(X_1) - p(X_1)\|^2)$.*

Simulation Study Our aim is to numerically verify: (1) Algorithm 1 improves the overall performance; and (2) the dataset S_4 is not the bottleneck of the convergence, which is an observation from the comparison among upper bounds and lower bounds as discussed Section 3. Similar to Table 1, we take $d_1 = 5$,

$d_2 = 2$. The mean and variance are $\mu = \mathbf{1}_{d_1+d_2}/\sqrt{d_1}$, $\Sigma = I_{d_1+d_2}$ respectively. We consider $\mathcal{L}_2$ attack with $\epsilon = 0.1$ in this experiment.

In Table 2, the benchmark algorithm is omnipotent and performs standard adversarial training on S_1 to S_3, with labels in S_3 known. For benchmark and the other methods except for adv+SSL(S_1, S_3, S_4), they do not use S_4, while for the method adv+SSL(S_1, S_3, S_4), $n_4 = n_3$. The method adv(S_1) means to use adversarial training on the dataset S_1 only. The method adv+pseudo label(S_1,S_3) is to use clean training in S_1 to impute the label for samples in S_3 and then conduct adversarial training.

From Table 2, we have some observations.

First, the quality of the imputed labels affects the adversarial robustness, and unlabeled data helps improve adversarial robustness. Comparing the benchmark and the other methods, the benchmark has a better clean training stage result, i.e., the true label, thus the final adversarial estimate is better. Comparing adv+SSL(S_1,S_3), adv(S_1), and adv+pseudo label(S_1,S_3), we claim that using unlabeled data helps improve the estimation efficiency.

In addition, S_4 only slightly contributes to the improvement of adversarial robustness. Comparing adv+SSL(S_1,S_3) and adv+SSL(S_1,S_3,S_4), we see that the additional data S_4 do not significantly improve the estimation efficiency. Similar observations can be found in Table 5 for logistic regression (in appendix).

5 Additional Discussions

We provide some additional discussions as by-products of the analysis above. In the main text, we provide theoretical results associated with two-layer neural networks in SSL. Due to the space limit, we postpone the discussion about the linear regression model with linear ϕ to the appendix.

5.1 Neural Networks

The design of ϕ in the previous sections is based on the parametric knowledge of the data generating model. We consider using a two-layer neural network as a "nonparametric" alternative to model ϕ while such knowledge is unavailable. In the literature, there are abundant results on the expressibility or fitting convergence of neural networks, e.g. Schmidt-Hieber et al. (2020); Bauer et al. (2019); Elbrächter et al. (2019); Hu et al. (2020, 2021); Farrell et al. (2021).

We follow Hu et al. (2021) to consider an easy-to-implement estimation procedure. To be specific, we use a two-layer neural network $\phi = (\phi_1, \ldots, \phi_{d_2})$ with

$$\phi_k(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} a_{j,k} \sigma(w_{j,k}^\top x), \quad w_{j,k} \in \mathbb{R}^{d_1},$$

where $a_{j,k}$ are generated from $\{\pm 1\}$ uniformly. The weights $w_{j,k}$'s are initialized from $N(0, \tau^2 I_{d_1})$ and trained with an $\mathcal{L}_2$ penalty with multiplier λ.

Proposition 2. *Let d_1 and d_2 be fixed and CI holds. Assume $\widehat{\phi}$ is trained under proper configurations and the distribution of (X_1, Y) satisfies some extra conditions. Assume n_3 is large enough so that $n_3 \gg \text{poly}(n_1+n_2)$ for some polynomial of n_1+n_2. Both $\widehat{W}$ and $\widehat{\phi}$ in the SSL procedure are regression estimator. Then with high probability,*

$$R(\widehat{W}\widehat{\phi}, 0) - R^*(0) = o\left(\frac{1}{n_1+n_2}\right).$$

The detailed conditions in Proposition 2 are postponed to Appendix E. From Proposition 2, when there are sufficient samples in S_3, the SSL procedure will improve the accuracy in clean training even if we do not have parametric knowledge for the model.

Real-Data Experiment We use the Yearbook dataset from Ginosar et al. (2015). We consider a two-layer ReLU network with lazy training (this network matches Proposition 2) for ϕ. Since all of our theoretical results are developed under large-sample asymptotics, we resize the images to 32x32, take the center 16x16 patch of an image as X_2, and take the rest as X_1. The goal is to classify the gender of each image. We minimize the square loss to obtain a classifier.

For the clean training task, in the pretext task, we randomly select 20,000 samples and regress X_2 on X_1 to learn the representation mapping $\widehat{\phi}$ using a two-layer ReLU network with lazy training and taking m, the number of hidden nodes, as 1000. Since the data dimension d_1 and d_2 are comparable to 20,000, we add an $\mathcal{L}_2$ penalty in the regression loss. The pretext task is trained by 100 epochs. For the downstream task, we take $10, 25, 50, 100$ samples with their labels to obtain $\widehat{W}$. Using square loss, there is an analytical solution of $\widehat{W}$. We add an $\mathcal{L}_2$ penalty to the square loss and tune it to achieve the best prediction accuracy. For adversarial training task, we use SSL in clean training and impute labels for unlabeled data, and use all data with labels/pseudolabels to train a two-layer ReLU network (with 1000 hidden node) as the adversarial classifier. We use $\mathcal{L}_\infty$ attack with $\epsilon = 4/255$.

To assess the performance of SSL for the clean training task, we compare it against a benchmark clean training algorithm: we directly use the two-layer ReLU network with 1000 hidden nodes (train all layers) on the

Table 2: Average Regret of adversarially robust estimate under CI condition. $n_1 = 100$. The variance information is in Table 3 in the appendix.

n_3	benchmark	adv+SSL(S_1,S_3)	adv(S_1)	adv+SSL(S_1,S_3,S_4)	adv+pseudo label(S_1,S_3)
500	0.00771	0.01264	0.01041	0.01211	0.00965
1000	0.00543	0.01195	0.01040	0.01046	0.00946
5000	0.00150	0.00543	0.00897	0.00526	0.00916
10000	0.00070	0.00332	0.01008	0.00330	0.00898
20000	0.00050	0.00213	0.00956	0.00185	0.00907

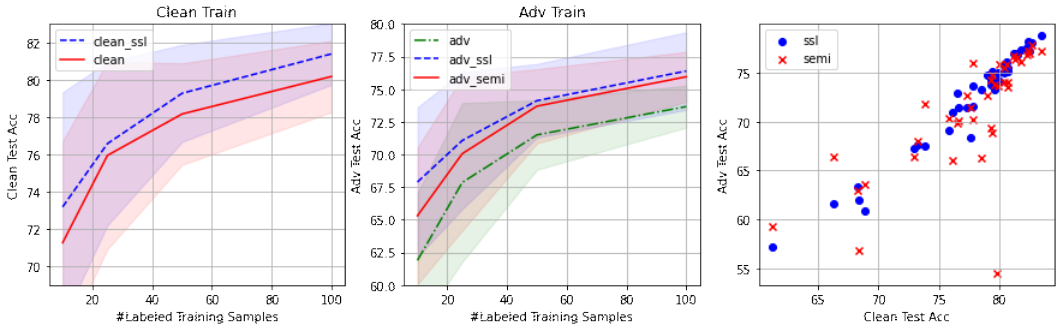

Figure 1: Clean and adversarial ($\mathcal{L}_\infty$ attack with $\epsilon = 4/255$) test accuracy (gender classification) using Yearbook data. Left: clean training accuracy. clean_ssl: SSL clean training. clean: benchmark clean training. Middle: adversarial training accuracy adv: standard adversarial training with only labeled data. adv_ssl: adv+SSL algorithm. adv_semi: Carmon et al. (2019). Right: clean test accuracy in clean training vs adversarial test accuracy in adversarial training.

labeled samples for 100 epochs with a learning rate of 0.1. For fairness, we also add and tune the $\mathcal{L}_2$ penalty to achieve the best performance. To assess the performance of adv+SSL algorithm for the adversarial training task, we compare it against two benchmark adversarial training algorithms: (1) We conduct the standard clean training using only the labeled data (x_1, y) and then impute labels for unlabeled data to do adversarial training, i.e., the exact algorithm in Carmon et al. (2019). (2) We only use the given labeled data to conduct adversarial training and do not use the unlabeled data. We use a two-layer ReLU network (with 1000 hidden nodes) for both benchmarks.

The experiment results are summarized in Figure 1, based on the average and variance of testing accuracies of over ten repeated runs. There are three figures in Figure 1. The left panel of Figure 1 compares the clean testing accuracy, and it is easy to see that SSL leads to a higher accuracy than the benchmark method. The middle panel of Figure 1 compares the adversarial testing accuracy, and one can see that utilizing unlabeled data helps improve the testing performance. If we compare the blue dashed curve and the red curve in both left and middle plots, it suggests that a better clean model (used to generate pseudolabels) leads to a better adversarial model. To confirm this, in the right panel of Figure 1, we plot the clean testing accuracy in clean training against the adversarial testing accuracy in adversarial training to study how the quality of the pseudolabels affects the final adversarial robustness. One can observe a positive correlation between these two accuracies, implying a positive correlation between the pseudolabel quality and final adversarial robustness. We conjecture that the improvements in Figure 1 is not as remarkable as Table 2 in simulation due to (1) from Proposition 2, the required n_3 is much larger than Theorem 4, and (2) the data dimension d_2 for real data is larger than simulated model, involving more error in estimating $\widehat{\phi}$.

6 Conclusion

In this paper, we investigate the statistical properties of reconstruction-based SSL. In particular, we study the minimax lower bound of estimation accuracy and the adversarial robustness. Through figuring out these properties, we argue that (1) in clean training, no matter CI holds or not, reconstruction-based SSL reaches the optimal rate of convergence in the models we consider; and (2) it is possible to design adversarially robust estimate such that it is also optimal. These advantages of the SSL method lead to a better performance of SSL compared with the vanilla training.

There are several potential directions for future de-

velopment. First, the procedure for adversarial training considered is tedious (i.e., pretext task and downstream task in clean training, and the adversarial training itself). It is interesting to simplify the procedure. Second, we only provide some light discussion using neural networks for the pretext task when the parametric form of $P(Y = 1|X_1 = x_1)$ is unknown. An in-depth investigation on this matter is definitely worthwhile, as deep neural networks have been viewed as a powerful nonparametric learning tool for modern data sciences. Third, this paper only discusses reconstruction-based SSL, and it is of great interest to generalize our analysis to other types of SSL-based methods.

7 Acknowledgements

This project is partially supported by NSF-SCALE MoDL (2134209).

References

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

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. (2019), "A theoretical analysis of contrastive unsupervised representation learning," *arXiv preprint arXiv:1902.09229*.

Audibert, J.-Y. and Tsybakov, A. B. (2007), "Fast learning rates for plug-in classifiers," *The Annals of statistics*, 35, 608–633.

Bauer, B., Kohler, M., et al. (2019), "On deep learning as a remedy for the curse of dimensionality in nonparametric regression," *Annals of Statistics*, 47, 2261–2285.

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., and Samworth, R. J. (2017), "Local nearest neighbour classification with applications to semi-supervised learning," *arXiv preprint arXiv:1704.00642*.

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. 11192–11203.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a), "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PMLR, pp. 1597–1607.

Chen, T., Liu, S., Chang, S., Cheng, Y., Amini, L., and Wang, Z. (2020b), "Adversarial robustness: From self-supervised pre-training to fine-tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 699–708.

Chuang, C.-Y., Robinson, J., Yen-Chen, L., Torralba, A., and Jegelka, S. (2020), "Debiased contrastive learning," *arXiv preprint arXiv:2007.00224*.

Dai, B. and Lin, D. (2017), "Contrastive learning for image captioning," *arXiv preprint arXiv:1710.02534*.

Dan, C., Wei, Y., and Ravikumar, P. (2020), "Sharp Statistical Guaratees for Adversarially Robust Gaussian Classification," in *International Conference on Machine Learning*, PMLR, pp. 2345–2355.

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

Doersch, C., Gupta, A., and Efros, A. A. (2015), "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430.

Elbrächter, D., Perekrestenko, D., Grohs, P., and Bölcskei, H. (2019), "Deep neural network approximation theory," *arXiv preprint arXiv:1901.02220*.

Farrell, M. H., Liang, T., and Misra, S. (2021), "Deep neural networks for estimation and inference," *Econometrica*, 89, 181–213.

Gao, R., Cai, T., Li, H., Wang, L., Hsieh, C.-J., and Lee, J. D. (2019), "Convergence of adversarial training in overparametrized networks," *arXiv preprint arXiv:1906.07916*.

Gidaris, S., Singh, P., and Komodakis, N. (2018), "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*.

Ginosar, S., Rakelly, K., Sachs, S., Yin, B., and Efros, A. A. (2015), "A century of portraits: A visual historical record of american high school yearbooks," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 1–7.

Gowal, S., Huang, P.-S., van den Oord, A., Mann, T., and Kohli, P. (2020), "Self-supervised Adversarial Robustness for the Low-label, High-data Regime," in *International Conference on Learning Representations*.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. (2021), "Provable Guarantees for Self-Supervised Deep Learning with Spectral Contrastive Loss," *arXiv preprint arXiv:2106.04156*.

Hendrycks, D., Lee, K., and Mazeika, M. (2019), "Using pre-training can improve model robustness and uncertainty," *arXiv preprint arXiv:1901.09960*.

Ho, C.-H. and Vasconcelos, N. (2020), "Contrastive learning with adversarial examples," *arXiv preprint arXiv:2010.12050*.

Hu, T., Shang, Z., and Cheng, G. (2020), "Optimal Rate of Convergence for Deep Neural Network Classifiers under the Teacher-Student Setting," *arXiv preprint arXiv:2001.06892*.

Hu, T., Wang, W., Lin, C., and Cheng, G. (2021), "Regularization Matters: A Nonparametric Perspective on Overparametrized Neural Network," in *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 829–837.

Javanmard, A. and Soltanolkotabi, M. (2020), "Precise statistical analysis of classification accuracies for adversarial training," *arXiv preprint arXiv:2010.11213*.

Javanmard, A., Soltanolkotabi, M., and Hassani, H. (2020), "Precise tradeoffs in adversarial training for linear regression," *arXiv preprint arXiv:2002.10477*.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020), "Supervised contrastive learning," *arXiv preprint arXiv:2004.11362*.

Kim, M., Tack, J., and Hwang, S. J. (2020), "Adversarial Self-Supervised Contrastive Learning," in *Advances in Neural Information Processing Systems*.

Lee, J. D., Lei, Q., Saunshi, N., and Zhuo, J. (2020), "Predicting what you already know helps: Provable self-supervised learning," *arXiv preprint arXiv:2008.01064*.

Li, J., Zhou, P., Xiong, C., and Hoi, S. C. (2020), "Prototypical contrastive learning of unsupervised representations," *arXiv preprint arXiv:2005.04966*.

Mehrabi, M., Javanmard, A., Rossi, R. A., Rao, A., and Mai, T. (2021), "Fundamental Tradeoffs in Distributionally Adversarial Training," *arXiv preprint arXiv:2101.06309*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013), "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018), *Foundations of machine learning*, MIT press.

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.

Noroozi, M. and Favaro, P. (2016), "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*, Springer, pp. 69–84.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018), "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*.

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.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. (2016), "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J. C., and Liang, P. (2019), "Adversarial training can hurt generalization," *arXiv preprint arXiv:1906.06032*.

Raskutti, G., Wainwright, M. J., and Yu, B. (2012), "Minimax-optimal rates for sparse additive models over kernel classes via convex programming," *The Journal of Machine Learning Research*, 13, 389–427.

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.

Schmidt-Hieber, J. et al. (2020), "Nonparametric regression using deep neural networks with ReLU activation function," *Annals of Statistics*, 48, 1875–1897.

Sinha, A., Namkoong, H., and Duchi, J. (2018), "Certifying some distributional robustness with principled adversarial training," .

Sun, W. W., Qiao, X., and Cheng, G. (2016), "Stabilized nearest neighbor classifier and its statistical properties," *Journal of the American Statistical Association*, 111, 1254–1265.

Taheri, H., Pedarsani, R., and Thrampoulidis, C. (2020), "Asymptotic Behavior of Adversarial Training in Binary Classification," *arXiv preprint arXiv:2010.13275*.

Teng, J. and Huang, W. (2021), "Can Pretext-Based Self-Supervised Learning Be Boosted by Downstream Data? A Theoretical Analysis," *arXiv preprint arXiv:2103.03568*.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. (2020), "What makes for good views for contrastive learning?" *arXiv preprint arXiv:2005.10243*.

Tony Cai, T. and Zhang, L. (2019), "High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81, 675–705.

Tosh, C., Krishnamurthy, A., and Hsu, D. (2021), "Contrastive learning, multi-view redundancy, and linear models," in *Algorithmic Learning Theory*, PMLR, pp. 1179–1206.

Uesato, J., Alayrac, J.-B., Huang, P.-S., Stanforth, R., Fawzi, A., and Kohli, P. (2019), "Are labels required for improving adversarial robustness?" *arXiv preprint arXiv:1905.13725*.

Wang, H. J. and Wang, L. (2009), "Locally weighted censored quantile regression," *Journal of the American Statistical Association*, 104, 1117–1128.

Xiao, T., Wang, X., Efros, A. A., and Darrell, T. (2020), "What should not be contrastive in contrastive learning," *arXiv preprint arXiv:2008.05659*.

Xing, Y., Song, Q., and Cheng, G. (2021a), "On the generalization properties of adversarial training," in *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 505–513.

Xing, Y., Zhang, R., and Cheng, G. (2021b), "Adversarially Robust Estimate and Risk Analysis in Linear Regression," in *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 514–522.

Xu, Q., Bello, K., and Honorio, J. (2020), "A Le Cam Type Bound for Adversarial Learning and Applications," *arXiv preprint arXiv:2007.00289*.

Yang, Y. and Tokdar, S. T. (2015), "Minimax-optimal nonparametric regression in high dimensions," *The Annals of Statistics*, 43, 652–674.

Yin, D., Ramchandran, K., and Bartlett, P. (2018), "Rademacher complexity for adversarially robust generalization," *arXiv preprint arXiv:1810.11914*.

Zeng, Z., He, K., Yan, Y., Xu, H., and Xu, W. (2021), "Adversarial self-supervised learning for out-of-domain detection," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5631–5639.

Zhai, R., Cai, T., He, D., Dan, C., He, K., Hopcroft, J., and Wang, L. (2019), "Adversarially robust generalization just requires more unlabeled data," *arXiv preprint arXiv: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, R., Isola, P., and Efros, A. A. (2016), "Colorful image colorization," in *European conference on computer vision*, Springer, pp. 649–666.

Zhang, Y., Plevrakis, O., Du, S. S., Li, X., Song, Z., and Arora, S. (2020), "Over-parameterized Adversarial Training: An Analysis Overcoming the Curse of Dimensionality," *arXiv preprint arXiv:2002.06668*.

Below is the list of contents in the appendix:

- Section A: discussion about regression.
- Section B: additional experiments, extra tables, and results regarding to neural networks.
- Section C: additional assumptions on the finite variance, and lemmas which provide minimax lower bound for some particular distributions.
- Section D: proofs for results in Section 2 and 3.
- Section E: proofs for results in Section 4 and A.
- The derivations for different losses are similar, so during the proofs, we firstly present the proof details for one loss, then display the proof for the other losses in a separate proof block to mention the differences.

A Regression

For regression, we assume X_1 and X_2 jointly follow some multivariate Gaussian distribution with $Var(X_1) = I_{d_1}$, $Var(X_2) = I_{d_1}$. The singular values of $Cov(X_1, X_2)$ are finite and bounded away from zero. The response satisfies $y = \theta_0^\top x_1 + \varepsilon$ for noise ε. The following theorem presents the convergence rate of the SSL estimate and the minimax lower bound:

Theorem 5. *For linear regression model described in above, assume $\theta_0 = Cov(X_1, X_2)a_0$, then*

$$\mathbb{E}R(\widehat{W}^\top \widehat{\phi}, \epsilon) - R^*(\epsilon) = O\left(\frac{\sigma^2 d_2}{n_1 + n_2} + \frac{\|a_0\|^2(d_1 + d_2)}{n_1 + n_3}\right),$$

which is minimax optimal when $n_1 + n_3 \gg (n_1 + n_2)\|a_0\|^2(d_1 + d_2)/(\sigma^2(d_1 - d_2))$.

Note that the condition $\theta_0 = Cov(X_1, X_2)a_0$ implies that the final estimate $\widehat{W}^\top \widehat{\phi}$ is asymptotically unbiased to θ_0. In comparison, Theorem 5 delivers a similar conclusion to the Theorem 4(II). Although the convergence rate of SSL estimate is slower without CI, it still reaches the minimax lower bound. The proof of Theorem 5 is postponed to Section E.

We only present the case where CI fails in Theorem 5 as CI is not appropriate in the model we consider. The following example illustrates this issue:

Example 4. *Assume (X_1, X_2, ε) follows multivariate Gaussian, then CI implies $\sigma^2 = 0$.*

B Extra Tables and Additional Experiments

Table 3: Variance table corresponding to Table 2

n_3	benchmark	adv+SSL(S_1,S_3)	adv(S_1)	adv+SSL(S_1,S_3,S_4)	adv+pseudo label(S_1,S_3)
500	1.14E-05	3.55E-04	7.09E-05	3.18E-04	3.53E-05
1000	8.43E-06	8.35E-05	6.71E-05	9.69E-05	3.07E-05
5000	7.85E-07	8.68E-06	4.58E-05	8.62E-06	2.27E-05
10000	3.01E-07	4.61E-06	6.12E-05	5.65E-06	1.62E-05
20000	2.06E-07	2.48E-06	5.11E-05	1.42E-06	1.79E-05

Table 4 is the clean training result using logistic regression, and the observations are similar to Table 1. Table 5 is the adversarial training result using logistic regression, and the observations are similar to Table 2. Table 6 summarizes the variance information for simulation in adversarial training.

n3	SSL(mean)	labeled(mean)	SSL(var)	labeled(var)
500	0.00946	0.02307	4.17E-05	3.19E-04
1000	0.00458	0.02076	1.24E-05	3.14E-04
5000	0.00106	0.02304	1.84E-06	3.42E-04
10000	0.00049	0.02375	1.25E-06	2.72E-04
20000	0.00015	0.02394	6.79E-07	2.23E-04

Table 4: Clean estimate under CI condition using logistic regression. $n_1 = 100$. Left: mean regret, right: the variance.

n_3	benchmark	adv+SSL(S_1,S_3)	adv(S_1)	adv+SSL(S_1,S_3,S_4)	adv+pseudo label(S_1,S_3)
500	0.00218	0.00570	0.01229	0.00441	0.00643
1000	0.00104	0.00325	0.01209	0.00291	0.00572
5000	0.00021	0.00064	0.01254	0.00072	0.00477
10000	0.00025	0.00051	0.01248	0.00027	0.00485
20000	0.00002	0.00020	0.01245	0.00017	0.00553

Table 5: Mean: Difference between the adversarial testing accuracy of estimators and θ^* using logistic regression. $n_1 = 100$. The variance information is in Table 6.

n_3	benchmark	adv+SSL(S_1,S_3)	adv(S_1)	adv+SSL(S_1,S_3,S_4)	adv+pseudo label(S_1,S_3)
500	3.78E-06	1.63E-05	7.20E-05	1.29E-05	2.09E-05
1000	1.08E-06	7.06E-06	9.72E-05	5.30E-06	2.06E-05
5000	4.06E-07	7.65E-07	8.69E-05	5.96E-07	9.63E-06
10000	1.99E-07	4.28E-07	9.69E-05	2.86E-07	9.09E-06
20000	1.32E-07	2.72E-07	8.62E-05	2.20E-07	2.07E-05

Table 6: Variance table corresponding to Table 5

C Lemmas and Extra Conditions

We first introduce some extra conditions and lemmas which are related to the minimax lower bounds. The extra conditions are technical assumptions regulating the behavior of the loss (and its Taylor expansions) to ensure a finite variance. The lemmas are some particular examples used in the minimax lower bound.

Assumption 2. *We further assume that $\mathcal{P}$ satisfies:*

(3) $\mathbb{E}\|X_1\|^2 X_1 X_1^\top / \mathbb{E}\|X_1\|^2$ have bounded eigenvalues.

(4) When CI holds, all the eigenvalues of the covariance of $\partial\|X_2 - \mathbb{E}(X_2|X_1)\|^2/\partial\beta$ and the expectation of $\partial^2\|X_2 - \mathbb{E}(X_2|X_1)\|^2/\partial\beta^2$ are of $\Theta(\gamma_{d_2})$ for some $\gamma_{d_2} = \Omega(1)$.

(5) When CI holds, if (X_1, X_2, Y) follows distribution P_{β^*} for some β^*, then $\mathbb{E}\partial\|X_2 - \mathbb{E}(X_2|X_1)\|^2/\partial\beta$ is L-Lipschitz for some constant $L > 0$ when $\beta \in B(\beta^*, r)$ where B denotes an L_2 ball and r is some constant.

Assumption 3. *We assume that*

1. *when l is logistic regression: the covariance of $\partial l/\partial\theta$ has all bounded and greater-than-zero eigenvalues, the expectation of $\partial^2 l/\partial\theta^2$ has all bounded and greater-than-zero eigenvalues. The density is finite and away from zero when x_1 is near the decision boundary. In addition, all the eigenvalues of*

$$\mathbb{E}_{X_1}\|X_1\|^2 \left[\left(\frac{\partial p(X_1;\beta^*)}{\partial\beta}\right)^\top \mathbb{E}_{X_1}\left\{\left(\frac{\partial p(X_1;\beta^*)}{\partial\beta}\right)\left(\frac{\partial p(X_1;\beta^*)}{\partial\beta}\right)^\top\right\}^{-1}\left(\frac{\partial p(X_1;\beta^*)}{\partial\beta}\right)\right]^2$$

are in $\Theta(d_1^3)$.

2. *when l is square loss: the covariance of $\partial l/\partial\theta$ has all bounded and greater-than-zero eigenvalues, the expectation of $\partial^2 l/\partial\theta^2$ has all bounded and greater-than-zero eigenvalues. The density is finite and away from zero when x_1 is near the decision boundary. In addition, all the eigenvalues of*

$$\mathbb{E}_{X_1}\|X_1\|^2\left[\left(\frac{\partial p(X_1;\beta^*)}{\partial\beta}\right)^\top \mathbb{E}_{X_1}\left\{\left(\frac{\partial p(X_1;\beta^*)}{\partial\beta}\right)\left(\frac{\partial p(X_1;\beta^*)}{\partial\beta}\right)^\top\right\}^{-1}\left(\frac{\partial p(X_1;\beta^*)}{\partial\beta}\right)\right]^2 (X_1^\top\theta^*)^2$$

are in $\Theta(d_1^3)$.

Conditions (3), (4), (5) in Assumption 2 regulate the distributions in $\mathcal{P}$, ensuring that a second-order Taylor expansion is accurate for the likelihood/loss function.

Assumption 3 is an extra assumption for adversarial training. Since we use SSL in clean training and use the vanilla method in adversarial training, some extra conditions are needed to supplement adversarial training. Similar to the idea in Assumption 2, Assumption 3 ensures that the Taylor expansions w.r.t. p and l are accurate.

Lemma 1. *Assume $(x_1, x_2) \in \mathbb{R}^{d_1+d_2}$, and $(x_1, x_2) \sim N(\mu, \Sigma^*)$ for some known Σ^* and unknown $\mu = (\mu_1, \mu_2)$. There are $n_1 + n_3$ samples of (x_1, x_2), $n_2 + n_4$ samples of (x_1). Then for estimators of μ_1,*

$$\inf_{\widehat{\mu}_1}\sup_{\|\mu_1\|\le R,\Sigma^*}\mathbb{E}\|\widehat{\mu}_1-\mu_1\|^2=\Omega\left(\frac{R^2 d_1}{\sum n_i}\right),$$

and for estimators of μ_2,

$$\inf_{\widehat{\mu}_2}\sup_{\|\mu_2\|\le R,\Sigma^*}\mathbb{E}\|\widehat{\mu}_2-\mu_2\|^2=\Omega\left(\frac{R^2 d_2}{n_1+n_3}\right).$$

Proof of Lemma 1. The proof is similar to Lemma 2 below using Gaussian prior. □

Lemma 2. *Assume $(x_1, x_2) \in \mathbb{R}^{d_1+d_2}$, and $(x_1, x_2) \sim N(\mathbf{0}, \Sigma^*)$ for some known Σ^*, and the response $\mathbb{E}[y|x_1] = x_1^\top\Sigma^*_{1,2}a$ for some vector a and $Var(y|x_1) = \sigma^2$ for any x_1. There are n_1 samples of (x_1, x_2, y), n_2 samples of (x_1, y), n_3 samples of (x_1, x_2) and n_4 samples of (x_1). Then*

$$\inf_{\widehat{a}}\sup_{a}\mathbb{E}\|\widehat{a}-a\|^2=\Theta\left(\frac{\sigma^2 d_2}{n_1+n_2}\right).$$

Proof of Lemma 2. Assume $y|x_1$ follows Gaussian distribution. We take $a \sim N(0, \sigma^2 I_{d_2}/(\alpha n))$ for some $\alpha > 0$ as the prior distribution of a. Then it is easy to see that only S_1 and S_2 are related to a. Denote $\widehat{\Sigma}_{n_1,n_2}$ as the sample covariance matrix, and $\widehat{a}_\alpha = (\widehat{\Sigma}_{n_1,n_2}+\alpha I_{d_1})^{-1}\frac{1}{n_1+n_2}\sum_{S_1,S_2}x_1 y$. The conditional distribution of $a|S_1, S_2$ becomes a multivariate Gaussian

$$a|S_1,S_2\sim N(\widehat{a}_\alpha,(\sigma^2/(n_1+n_2))(\widehat{\Sigma}_{n_1,n_2}+\alpha I_{d_2})^{-1}),$$

and therefore

$$\inf_{\widehat{a}}\sup_{a}\mathbb{E}\|\widehat{a}-a\|^2\ge\inf_{\widehat{a}}\mathbb{E}\|\widehat{a}-\widehat{a}_\alpha\|^2+\|\widehat{a}_\alpha-a\|^2=\Theta\left(\frac{\sigma^2 d_2}{n_1+n_2}\right).$$

□

Lemma 3. *Assume $(x_1, x_2) \in \mathbb{R}^{d_1+d_2}$, and $(x_1, x_2) \sim N(\mu, \Sigma^*)$, where $\Sigma^* = \begin{bmatrix}\Sigma^*_{1,1} & \Sigma^*_{1,2}\\ \Sigma^*_{2,1} & \Sigma^*_{2,2}\end{bmatrix}$. Assume there are $n_1 + n_3$ samples of (x_1, x_2), and $n_2 + n_4$ samples of x_1, and no response y is provided, then when $(d_1 + d_2) = o((n_1 + n_3)\log(n_1 + n_3))$,*

$$\inf_{\widehat{\Sigma}_{1,2}}\sup_{\Sigma^*_{1,2}}\mathbb{E}\|\widehat{\Sigma}_{1,2}-\Sigma^*_{1,2}\|^2=\Omega\left(\frac{d_1+d_2}{n_1+n_3}\right).$$

Assume the response $\mathbb{E}[y|x_1] = x_1^\top \Sigma_{1,2}^* a$ *for some known vector* a *with* $\|a\| = 1$ *and* $\mu = 0$. *There are* n_1 *samples of* (x_1, x_2, y), n_2 *samples of* (x_1, y), n_3 *samples of* (x_1, x_2) *and* n_4 *samples of* (x_1), *then when* $(n_1+n_2) = o(n_1+n_3)$, *for any estimator* $\widehat{\theta}$ *which estimates* $\Sigma_{1,2}^* a$,

$$\inf_{\widehat{\theta}} \sup_{\Sigma_{1,2}^*} \mathbb{E}\|\widehat{\theta} - \Sigma_{1,2}^* a\|^2 = \Omega\left(\frac{d_1 + d_2}{n_1 + n_3}\right).$$

Proof of Lemma 3. We directly prove the second argument of Lemma 3. We use Bayes method to show the minimax lower bound. Assume $\Sigma_{1,2} = \Sigma_{1,2}^*$ follows some prior distribution, then

$$\begin{aligned}
\inf_{\widehat{\Sigma}_{1,2}} \sup_{\Sigma_{1,2}^*} \mathbb{E}\|\widehat{\Sigma}_{1,2} - \Sigma_{1,2}^*\|^2 &\geq \inf_{\widehat{\Sigma}_{1,2}} \mathbb{E}_{S_i}\mathbb{E}_{\Sigma_{1,2}|S_i}\|\widehat{\Sigma}_{1,2} - \Sigma_{1,2}\|^2 \\
&\geq \inf_{\widehat{\Sigma}_{1,2}} \mathbb{E}_{S_i}\mathbb{E}_{\Sigma_{1,2}|S_i}\|[\mathbb{E}_{\Sigma_{1,2}|S_i}\Sigma_{1,2}] - \Sigma_{1,2}\|^2.
\end{aligned}$$

Denote the density of $\Sigma_{1,2}$ as $g(\Sigma_{1,2})$. Assume $\mathbb{E}\Sigma_{1,2} = \Sigma_{1,2}^*$ and $\Sigma_{1,2} - \Sigma_{1,2}^* = \Delta_1\Delta_2^\top$. And we take $\Delta_2 = a/\|a\|$. Assume $\Sigma_{1,1} = I_{d_1}$, $\Sigma_{1,2} = I_{d_2}$, and $\theta_0 = \Sigma_{1,2}a$. Denote $\Sigma^* = \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2}^* \\ \Sigma_{2,1}^* & \Sigma_{2,2} \end{bmatrix}$. Then the likelihood of the four types of samples $S_1, ..., S_4$ is proportional to

$$\frac{g(\Sigma_{1,2})}{|\Sigma|^{(n_1+n_3)/2}} \exp\left\{-\frac{1}{2}\sum_{S_1,S_3}\left[x_1^\top, x_2^\top\right]\Sigma^{-1}\begin{bmatrix}x_1\\x_2\end{bmatrix} - \frac{1}{2\sigma^2}\sum_{S_1,S_2}(y - x_1^\top\Sigma_{1,2}a)^2\right\}.$$

Since $\Sigma_{1,1}$ and $\Sigma_{2,2}$ are both identity matrix,

$$\begin{aligned}
|\Sigma| &= \left|\begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2}^* \\ \Sigma_{2,1}^* + \Delta_2\Delta_1^\top & \Sigma_{2,2} \end{bmatrix}\right| \left(1 + \begin{bmatrix}\mathbf{0} & \Delta_2^\top\end{bmatrix}\begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2}^* \\ \Sigma_{2,1}^* + \Delta_2\Delta_1^\top & \Sigma_{2,2} \end{bmatrix}^{-1}\begin{bmatrix}\Delta_1\\\mathbf{0}\end{bmatrix}\right) \\
&= |\Sigma^*|\left(1 + \begin{bmatrix}\mathbf{0} & \Delta_2^\top\end{bmatrix}(\Sigma^*)^{-1}\begin{bmatrix}\Delta_1\\\mathbf{0}\end{bmatrix}\right)\left(1 + \begin{bmatrix}\mathbf{0} & \Delta_2^\top\end{bmatrix}\begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2}^* \\ \Sigma_{2,1}^* + \Delta_2\Delta_1^\top & \Sigma_{2,2} \end{bmatrix}^{-1}\begin{bmatrix}\Delta_1\\\mathbf{0}\end{bmatrix}\right),
\end{aligned}$$

where

$$\begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2}^* \\ \Sigma_{2,1}^* + \Delta_2\Delta_1^\top & \Sigma_{2,2} \end{bmatrix}^{-1} = (\Sigma^*)^{-1} - \frac{(\Sigma^*)^{-1}\begin{bmatrix}\Delta_1\\0\end{bmatrix}\begin{bmatrix}\mathbf{0} & \Delta_2^\top\end{bmatrix}(\Sigma^*)^{-1}}{1 + \begin{bmatrix}\mathbf{0} & \Delta_2^\top\end{bmatrix}(\Sigma^*)^{-1}\begin{bmatrix}\Delta_1\\0\end{bmatrix}},$$

thus denoting $\xi = \begin{bmatrix}\mathbf{0} & \Delta_2^\top\end{bmatrix}(\Sigma^*)^{-1}\begin{bmatrix}\Delta_1\\\mathbf{0}\end{bmatrix}$, we obtain

$$\begin{aligned}
|\Sigma| &= |\Sigma^*|(1+\xi)\left(1 + \begin{bmatrix}\mathbf{0} & \Delta_2^\top\end{bmatrix}\left[(\Sigma^*)^{-1} - \frac{(\Sigma^*)^{-1}\begin{bmatrix}\Delta_1\\0\end{bmatrix}\begin{bmatrix}\mathbf{0} & \Delta_2^\top\end{bmatrix}(\Sigma^*)^{-1}}{1+\xi}\right]\begin{bmatrix}\Delta_1\\\mathbf{0}\end{bmatrix}\right) \\
&= |\Sigma^*|(1+\xi)\left(1 + \xi - \frac{\xi^2}{1+\xi}\right) \\
&= |\Sigma^*|(1+2\xi).
\end{aligned}$$

In terms of Σ^{-1}, denoting $\xi_1 = \begin{bmatrix}\Delta_1^\top & \mathbf{0}\end{bmatrix}(\Sigma^*)^{-1}\begin{bmatrix}\Delta_1\\ \mathbf{0}\end{bmatrix}$, and $\xi_2 = \begin{bmatrix}\mathbf{0} & \Delta_2^\top\end{bmatrix}(\Sigma^*)^{-1}\begin{bmatrix}\mathbf{0}\\ \Delta_2\end{bmatrix}$

$$
\begin{aligned}
\Sigma^{-1} &= \begin{bmatrix}\Sigma_{1,1} & \Sigma_{1,2}^*\\ \Sigma_{2,1}^*+\Delta_2\Delta_1^\top & \Sigma_{2,2}\end{bmatrix}^{-1} - \frac{\begin{bmatrix}\Sigma_{1,1} & \Sigma_{1,2}^*\\ \Sigma_{2,1}^*+\Delta_2\Delta_1^\top & \Sigma_{2,2}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{0}\\ \Delta_2\end{bmatrix}\begin{bmatrix}\Delta_1^\top & \mathbf{0}\end{bmatrix}\begin{bmatrix}\Sigma_{1,1} & \Sigma_{1,2}^*\\ \Sigma_{2,1}^*+\Delta_2\Delta_1^\top & \Sigma_{2,2}\end{bmatrix}^{-1}}{1+\xi-\xi^2/(1+\xi)}\\
&= (\Sigma^*)^{-1} - \frac{(\Sigma^*)^{-1}\begin{bmatrix}\Delta_1\\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \Delta_2^\top\end{bmatrix}(\Sigma^*)^{-1}}{1+\xi}\\
&\quad -\frac{(\Sigma^*)^{-1}\begin{bmatrix}\mathbf{0}\\ \Delta_2\end{bmatrix}\begin{bmatrix}\Delta_1^\top & \mathbf{0}\end{bmatrix}(\Sigma^*)^{-1}}{1+\xi-\xi^2/(1+\xi)} - \frac{\xi_1\xi_2(\Sigma^*)^{-1}\begin{bmatrix}\Delta_1\\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \Delta_2^\top\end{bmatrix}(\Sigma^*)^{-1}}{(1+\xi)^2(1+\xi-\xi^2/(1+\xi))}\\
&\quad +\frac{\xi_2(\Sigma^*)^{-1}\begin{bmatrix}\Delta_1\\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\Delta_1 & \mathbf{0}\end{bmatrix}(\Sigma^*)^{-1}}{(1+\xi)(1+\xi-\xi^2/(1+\xi))} + \frac{\xi_1(\Sigma^*)^{-1}\begin{bmatrix}\mathbf{0}\\ \Delta_2\end{bmatrix}\begin{bmatrix}\mathbf{0} & \Delta_2\end{bmatrix}(\Sigma^*)^{-1}}{(1+\xi)(1+\xi-\xi^2/(1+\xi))}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\frac{g(\Sigma_{1,2})}{|\Sigma|^{(n_1+n_3)/2}}\exp\left\{-\frac{1}{2}\sum_{S_1,S_3}\begin{bmatrix}x_1^\top, x_2^\top\end{bmatrix}\Sigma^{-1}\begin{bmatrix}x_1\\ x_2\end{bmatrix} - \frac{1}{2\sigma^2}\sum_{S_1,S_2}(y-x_1^\top\Sigma_{1,2}a)^2\right\}\\
=\;& \frac{g(\Sigma_{1,2})}{|\Sigma^*|^{(n_1+n_3)/2}}\exp\left\{-\frac{1}{2}\sum_{S_1,S_3}\begin{bmatrix}x_1^\top, x_2^\top\end{bmatrix}(\Sigma^*)^{-1}\begin{bmatrix}x_1\\ x_2\end{bmatrix} - \frac{1}{2\sigma^2}\sum_{S_1,S_2}(y-x_1^\top\Sigma_{1,2}^*a)^2\right\}\\
&\times\frac{1}{(1+2\xi)^{(n_1+n_3)/2}}\exp\left\{-\frac{1}{2}\sum_{S_1,S_3}\begin{bmatrix}x_1^\top, x_2^\top\end{bmatrix}(\Sigma^{-1}-(\Sigma^*)^{-1})\begin{bmatrix}x_1\\ x_2\end{bmatrix}\right\}\\
&\times\exp\left\{-\frac{1}{2\sigma^2}\sum_{S_1,S_2}(x_1^\top\Delta_1\Delta_2^\top a)^2 - 2(y-x_1^\top\Sigma_{1,2}^*a)(x_1^\top\Delta_1\Delta_2^\top a)\right\}\\
:=\;& g(\Sigma_{1,2})g_0g_1(\Delta_1,\Delta_2)g_2(\Delta_1,\Delta_2)
\end{aligned}
$$

Denoting $\begin{bmatrix}z_1\\ z_2\end{bmatrix} = (\Sigma^*)^{-1}\begin{bmatrix}x_1\\ x_2\end{bmatrix}$, then

$$
\begin{aligned}
&-\frac{1}{2}\sum_{S_1,S_3}\begin{bmatrix}x_1^\top, x_2^\top\end{bmatrix}(\Sigma^{-1}-(\Sigma^*)^{-1})\begin{bmatrix}x_1\\ x_2\end{bmatrix}\\
=\;& -\frac{1}{2}\sum_{S_1,S_3} -\frac{[z_1^\top, z_2^\top]\begin{bmatrix}\Delta_1\\ \mathbf{0}\end{bmatrix}[z_1^\top, z_2^\top]\begin{bmatrix}\mathbf{0}\\ \Delta_2\end{bmatrix}}{1+\xi} - \frac{[z_1^\top, z_2^\top]\begin{bmatrix}\Delta_1\\ \mathbf{0}\end{bmatrix}[z_1^\top, z_2^\top]\begin{bmatrix}\mathbf{0}\\ \Delta_2\end{bmatrix}}{1+\xi-\xi^2/(1+\xi)} - \frac{\xi_1\xi_2(\Sigma^*)^{-1}\begin{bmatrix}\Delta_1\\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \Delta_2^\top\end{bmatrix}(\Sigma^*)^{-1}}{(1+\xi)^2(1+\xi-\xi^2/(1+\xi))}\\
&-\frac{1}{2}\sum_{S_1,S_3}\frac{\xi_2\left([z_1^\top, z_2^\top]\begin{bmatrix}\Delta_1\\ \mathbf{0}\end{bmatrix}\right)^2}{(1+\xi)(1+\xi-\xi^2/(1+\xi))} + \frac{\xi_1\left([z_1^\top, z_2^\top]\begin{bmatrix}\mathbf{0}\\ \Delta_2\end{bmatrix}\right)^2}{(1+\xi)(1+\xi-\xi^2/(1+\xi))}.
\end{aligned}
$$

Now we consider another likelihood

$$
g(\Sigma_{1,2})g_0\widetilde{g}_1(\Delta_1,\Delta_2),
$$

where $\widetilde{g}_1$ is an approximation of g_1 and equals to

$$\begin{aligned}
\widetilde{g}_1(\Delta_1,\Delta_2) &= \exp\left\{tr\left(\begin{bmatrix}\Delta_1\\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \Delta_2\end{bmatrix}(\Sigma^*)^{-1}\left[\sum_{S_1,S_3}(\Sigma^*)^{-1}\begin{bmatrix}x_1\\ x_2\end{bmatrix}\begin{bmatrix}x_1^\top, x_2^\top\end{bmatrix} - I_{d_1+d_2}\right]\right)\right\}\\
&\quad\times\exp\left\{-\frac{1}{2}\sum_{S_1,S_3}\xi_2\left([z_1^\top, z_2^\top]\begin{bmatrix}\Delta_1\\ \mathbf{0}\end{bmatrix}\right)^2+\xi_1\left([z_1^\top, z_2^\top]\begin{bmatrix}\mathbf{0}\\ \Delta_2\end{bmatrix}\right)^2\right\}\\
&\quad\times 1\{\Delta_1\in\mathcal{S}\}\\
&= \exp\left\{tr\left(\begin{bmatrix}\Delta_1\\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \Delta_2\end{bmatrix}(\Sigma^*)^{-1}\left[\sum_{S_1,S_3}(\Sigma^*)^{-1}\begin{bmatrix}x_1\\ x_2\end{bmatrix}\begin{bmatrix}x_1^\top, x_2^\top\end{bmatrix} - I_{d_1+d_2}\right]\right)\right\}\\
&\quad\times\exp\left\{-\frac{1}{2}\begin{bmatrix}\Delta_1^\top & \mathbf{0}\end{bmatrix}\left(\sum_{S_1,S_3}\xi_2\begin{bmatrix}x_1\\ x_2\end{bmatrix}(\Sigma^*)^{-2}[x_1^\top, x_2^\top]+(x_1(\Sigma^*)^{-1}\Delta_2)^2 I_{d_1}\right)\begin{bmatrix}\Delta_1\\ \mathbf{0}\end{bmatrix}\right\}\\
&\quad\times 1\{\Delta_1\in\mathcal{S}\}.
\end{aligned}$$

Intuitively, when $\|\Delta_1\|\to 0$, $g_1\to\widetilde{g}_1$; otherwise $g_1\to 0$. As a result, we take $\mathcal{S}=\{\|\Delta_1\|_\infty\le 1/\sqrt{n_1+n_3}\}$.

From the generation of (x_1,x_2), assume $g_0(\Sigma_{1,2})\propto 1\{\|\Delta_1\|_\infty\le 1/\sqrt{n_1+n_3}\}$. We know that with probability tending to 1,

$$\left\|a\left(\sum_{S_1,S_3}(\Sigma^*)^{-1}\begin{bmatrix}x_1\\ x_2\end{bmatrix}\begin{bmatrix}x_1^\top, x_2^\top\end{bmatrix}-I_{d_1+d_2}\right)\right\|_2^2 = O\left((d_1+d_2)(n_1+n_3)\right),$$

and

$$\left(\sum_{S_1,S_3}\xi_2\begin{bmatrix}x_1\\ x_2\end{bmatrix}(\Sigma^*)^{-2}[x_1^\top, x_2^\top]+(x_1(\Sigma^*)^{-1}\Delta_2)^2 I_{d_1}\right) = O\left((n_1+n_3)I_{d_1}\right).$$

Therefore, Δ_1 follows truncated normal distribution and

$$\mathbb{E}_{\Delta_1\sim\widetilde{g}_1}\|(\Delta_1\Delta_2^\top-\mathbb{E}\Delta_1\Delta_2^\top)a\|^2=\Theta\left(\frac{d_1+d_2}{n_1+n_3}\right).$$

In the above analysis of A we investigate in the distribution $gg_0\widetilde{g}_1$ instead of the true distribution $gg_0g_1g_2$. Now we quantify the difference between $\widetilde{g}_1$ and g_1g_2.

When $\Delta_1 \in \mathcal{S}$, we have

$$\begin{aligned}
&\frac{g_1 g_2}{\widetilde{g}_1} \\
= \quad & \exp\left\{\frac{1}{2} tr\left(\begin{bmatrix}\Delta_1 \\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \Delta_2\end{bmatrix}(\Sigma^*)^{-1}\left[\sum_{S_1,S_3}(\Sigma^*)^{-1}\begin{bmatrix}x_1 \\ x_2\end{bmatrix}[x_1^\top, x_2^\top] - I_{d_1+d_2}\right]\right)\right. \\
& \qquad \left. \times\left(\frac{1}{1+\xi} + \frac{1}{1+\xi-\xi^2/(1+\xi)} - 2\right)\right\} \\
& \times \exp\left\{-\frac{1}{2}\sum_{S_1,S_3}\frac{\xi_2\left([z_1^\top, z_2^\top]\begin{bmatrix}\Delta_1 \\ \mathbf{0}\end{bmatrix}\right)^2}{(1+\xi)(1+\xi-\xi^2/(1+\xi))}\right. \\
& \qquad \left. + \frac{\xi_1\left([z_1^\top, z_2^\top]\begin{bmatrix}\mathbf{0} \\ \Delta_2\end{bmatrix}\right)^2}{(1+\xi)(1+\xi-\xi^2/(1+\xi))}\left(\frac{2}{(1+\xi)(1+\xi-\xi^2/(1+\xi))} - 2\right)\right\} \\
& \times \exp\left\{-\frac{1}{2}\sum_{S_1,S_3}\frac{\xi_1\xi_2(\Sigma^*)^{-1}\begin{bmatrix}\Delta_1 \\ \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \Delta_2^\top\end{bmatrix}(\Sigma^*)^{-1}}{(1+\xi)^2(1+\xi-\xi^2/(1+\xi))}\right\} \\
& \times \exp\left\{-\frac{1}{2\sigma^2}\sum_{S_1,S_2}(x_1^\top\Delta_1\Delta_2^\top a)^2 - 2(y - x_1^\top\Sigma_{1,2}^* a)(x_1^\top\Delta_1\Delta_2^\top a)\right\} \\
& \times 1\{\Delta_1 \in \mathcal{S}\} \\
:= \quad & A \times B \times C \times D \times 1\{\Delta_1 \in \mathcal{S}\}.
\end{aligned}$$

Recall that $\xi = \begin{bmatrix}\mathbf{0} & \Delta_2^\top\end{bmatrix}(\Sigma^*)^{-1}\begin{bmatrix}\Delta_1 \\ \mathbf{0}\end{bmatrix}$ and $\Delta_2 = a/\|a\|$, from the support of Δ_1, we have $\xi = O(\sqrt{d_1/(n_1+n_3)})$. Therefore,

$$\begin{aligned}
& \frac{1}{1+\xi} + \frac{1}{1+\xi-\xi^2/(1+\xi)} - 2 \\
= \quad & \frac{1}{(1+\xi)[1+\xi-\xi^2/(1+\xi)]}\left[1+\xi-\xi^2/(1+\xi)+1+\xi-2(1+\xi)[1+\xi-\xi^2/(1+\xi)]\right] \\
= \quad & \frac{1}{(1+\xi)[1+\xi-\xi^2/(1+\xi)]}\left[1+\xi-\xi^2/(1+\xi)+1+\xi-2[1+\xi-\xi^2/(1+\xi)]\right] \\
& -\frac{1}{(1+\xi)[1+\xi-\xi^2/(1+\xi)]}2[\xi+\xi^2-\xi^3/(1+\xi)] \\
= \quad & \frac{1}{(1+\xi)[1+\xi-\xi^2/(1+\xi)]}\left[\xi^2/(1+\xi)-2[\xi+\xi^2-\xi^3/(1+\xi)]\right] \\
= \quad & O(\xi),
\end{aligned}$$

so $A = \exp\{o(1)\}$. Similarly, we have $B = \exp\{o(1)\}$. In terms of C and D, one can show that both of them are in $\exp\{o(1)\}$ as well. Therefore, $g_1 g_2 = \widetilde{g}_1(1+o(1))$, which implies that

$$\sup_{\Sigma}\mathbb{E}\|(\Delta_1\Delta_2^\top - \mathbb{E}\Delta_1\Delta_2^\top)a\|^2 = \Theta\left(\frac{d_1+d_2}{n_1+n_3}\right).$$

As a result, we can conclude that

$$\inf_{\widehat{\theta}}\sup_{\Sigma}\mathbb{E}\|\widehat{\theta} - \Sigma_{1,2}a\|^2 = \Theta\left(\frac{d_1+d_2}{n_1+n_3}\right).$$

□

D Proofs for Section 2 and 3

D.1 Proof for Theorem 1

Proof of Theorem 1, CI holds. The basic idea is similar to Theorem 3. We impose a prior distribution on the parameter μ_0 associated with $P(y|x_1)$, then argue that we cannot exactly estimate μ_0.

Denote f_1 and f_2 are the density of x_1 and x_2, and $x_2|y \sim N(y\mu_2, \Sigma_{2,2})$ and is conditionally independent to x_1 given y. Also denote $\psi(x, \mu, \Sigma)$ as normal density of x given mean μ and covariance Σ. Assume $P(y = 1|x_1) = \psi(x_1, \mu_0, \Sigma_{1,1})/(\psi(x_1, \mu_0, \Sigma_{1,1}) + \psi(x_1, -\mu_0, \Sigma_{1,1}))$, then the likelihood becomes

$$f_0(\mu_0)\left(\prod_{S_1} f_1(x_1)P(y|x_1, \mu_0)f_2(x_2|y)\right) \times \left(\prod_{S_2} f_1(x_1)P(y|x_1, \mu_0)\right)$$
$$\times \left(\prod_{S_3} f_1(x_1)f_2(x_2|x_1, \mu_0)\right) \times \left(\prod_{S_4} f_1(x_1)\right).$$

Thus the posterior distribution of μ_0 given S_1 to S_4 is proportional to

$$f_0(\mu_0)\left(\prod_{S_1,S_2} P(y|x_1, \mu_0)\right) \times \left(\prod_{S_3} f_2(x_2|x_1, \mu_0)\right).$$

Denote $f_2(x_2|x_1, \mu_0) = m(\mu_0, x_1, x_2)$, then we have

$$\begin{aligned}\frac{\partial}{\partial\mu_0}\log m(\mu_0^*, x_1, x_2) &= \frac{\psi\left([x_1, x_2], [\mu_0^*, \mu_2], \Sigma\right) - \psi\left([x_1, x_2], -[\mu_0^*, \mu_2], \Sigma\right)}{\psi\left([x_1, x_2], [\mu_0^*, \mu_2], \Sigma\right) + \psi\left([x_1, x_2], -[\mu_0^*, \mu_2], \Sigma\right)}(x_1 - \mu_0^*) \\ &\quad - \frac{\psi(x_1, \mu_0^*, \Sigma_{1,1}) - \psi(x_1, -\mu_0^*, \Sigma_{1,1})}{\psi(x_1, \mu_0^*, \Sigma_{1,1}) + \psi(x_1, -\mu_0^*, \Sigma_{1,1})}(x_1 - \mu_0^*),\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2}{\partial\mu_0^2}\log m(\mu_0^*, x_1, x_2) &= -\frac{\psi\left([x_1, x_2], [\mu_0^*, \mu_2], \Sigma\right) - \psi\left([x_1, x_2], -[\mu_0^*, \mu_2], \Sigma\right)}{\psi\left([x_1, x_2], [\mu_0^*, \mu_2], \Sigma\right) + \psi\left([x_1, x_2], -[\mu_0^*, \mu_2], \Sigma\right)} I_{d_1} \\ &\quad + \frac{\psi(x_1, \mu_0^*, \Sigma_{1,1}) - \psi(x_1, -\mu_0^*, \Sigma_{1,1})}{\psi(x_1, \mu_0^*, \Sigma_{1,1}) + \psi(x_1, -\mu_0^*, \Sigma_{1,1})} I_{d_1} \\ &\quad - \left(\frac{\psi\left([x_1, x_2], [\mu_0^*, \mu_2], \Sigma\right) - \psi\left([x_1, x_2], -[\mu_0^*, \mu_2], \Sigma\right)}{\psi\left([x_1, x_2], [\mu_0^*, \mu_2], \Sigma\right) + \psi\left([x_1, x_2], -[\mu_0^*, \mu_2], \Sigma\right)}\right)^2 (x_1 - \mu_0^*)(x_1 - \mu_0^*)^\top \\ &\quad + \left(\frac{\psi(x_1, \mu_0^*, \Sigma_{1,1}) - \psi(x_1, -\mu_0^*, \Sigma_{1,1})}{\psi(x_1, \mu_0^*, \Sigma_{1,1}) + \psi(x_1, -\mu_0^*, \Sigma_{1,1})}\right)^2 (x_1 - \mu_0^*)(x_1 - \mu_0^*)^\top.\end{aligned}$$

When $(d_1 + d_2)^2 \log(n_1 + n_3) = o(n_1 + n_3)$, one can figure out that $\frac{\partial}{\partial\mu_0}\log m(\mu_0^*, x_1, x_2)$ and $\frac{\partial^2}{\partial\mu_0^2}\log m(\mu_0^*, x_1, x_2)$ converges to their mean respectively. In addition, from the model construction, we know that $\mathbb{E}\frac{\partial^2}{\partial\mu_0^2}\log m(\mu_0^*, x_1, x_2)$ is negatively definite.

Using Taylor expansion, we have

$$\begin{aligned}\log m(\mu_0, x_1, x_2) &= \log m(\mu_0^*, x_1, x_2) + \left(\frac{\partial}{\partial\mu_0}\log m(\mu_0^*, x_1, x_2)\right)^\top (\mu_0 - \mu_0^*) \\ &\quad + \frac{1}{2}(\mu_0 - \mu_0^*)^\top \left(\frac{\partial^2}{\partial\mu_0^2}\log m(\mu_0^*, x_1, x_2)\right)(\mu_0 - \mu_0^*) + O(\|\mu_0 - \mu_0^*\|^3).\end{aligned}$$

Thus when $\|\mu_0 - \mu_0^*\|^3 = O(1/(n_1 + n_3))$, $\prod_{S_3} f_2$ only changes in $O(1)$ proportion.

On the other hand, one can figure out that when $\|\mu_0 - \mu_0^*\|^3 = O(1/(n_1 + n_2))$, the posterior distribution of μ_0 is approximately a multivariate Gaussian. The final lower bound takes from the smaller one in the above two bounds.

When $\epsilon > 0$, assume $x_1 \sim N(\mathbf{0}, I_{d_1})$, then the adversarial risk minimizer still the linear classifier $\text{sgn}(x_1^\top \mu_0)$. As a result, the minimax lower bound of the estimation error of μ_0 is inherited in adversarial case. □

Proof of Theorem 1, CI does not hold. We consider several cases:

- Case 1: $n_1 + n_2 = o(n_1 + n_3)$ and $d_2 \le d_1$.
- Case 2: $n_1 + n_2 = o(n_1 + n_3)$ and $d_2 > d_1$.
- Case 3: $n_1 + n_3 = O(n_1 + n_2)$.

Case 1: The proof is similar to Theorem 5 for regression. Assume the optimal classifier w.r.t x_1 is of the form $\text{sgn}(a^\top \Sigma_{2,1} x_1)$ for some $a \in \mathbb{R}^{d_1}$. Based on Lemma 3, when a is known and $\|a\| = 1$, we have

$$\inf_{\widehat{\theta}} \sup_{\Sigma_{1,2}} \mathbb{E}\|\widehat{\theta} - \Sigma_{1,2} a\|^2 = \Omega\left(\frac{d_1 + d_2}{n_1 + n_3}\right).$$

When $\Sigma_{1,2}$ is known, the proof follows similar arguments as Theorem 1 when CI holds. Assume $P(y = 1|x_1) = \psi(x_1, \Sigma_{1,2}a, \Sigma_{1,1})/(\psi(x_1, \Sigma_{1,2}a, \Sigma_{1,1}) + \psi(x_1, -\Sigma_{1,2}a, \Sigma_{1,1}))$. Since there is no CI condition, there is no connection between the distribution of x_2 and the label y. As a result, the part of likelihood related to a is

$$f_0(\Sigma_{1,2}a) \prod_{S_1, S_2} P(y|x_1, \Sigma_{1,2}a).$$

Denote $P(y = 1|x_1) = m(\Sigma_{1,2}a, x_1)$, then we have

$$\begin{aligned}
\frac{\partial}{\partial \Sigma_{1,2}a} \log m(\Sigma_{1,2}a^*, x_1) &= (x_1 - \Sigma_{1,2}a^*) \\
&\quad - \frac{\psi(x_1, \Sigma_{1,2}a, \Sigma_{1,1}) - \psi(x_1, -\Sigma_{1,2}a, \Sigma_{1,1})}{\psi(x_1, \Sigma_{1,2}a, \Sigma_{1,1}) + \psi(x_1, -\Sigma_{1,2}a, \Sigma_{1,1})}(x_1 - \Sigma_{1,2}a^*) \\
\frac{\partial}{\partial \Sigma_{1,2}a} \log[1 - m(\Sigma_{1,2}a^*, x_1)] &= -(x_1 - \Sigma_{1,2}a^*) \\
&\quad - \frac{\psi(x_1, \Sigma_{1,2}a, \Sigma_{1,1}) - \psi(x_1, -\Sigma_{1,2}a, \Sigma_{1,1})}{\psi(x_1, \Sigma_{1,2}a, \Sigma_{1,1}) + \psi(x_1, -\Sigma_{1,2}a, \Sigma_{1,1})}(x_1 - \Sigma_{1,2}a^*)
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial^2}{\partial \Sigma_{1,2}a^2} \log m(\Sigma_{1,2}a^*, x_1) \\
= \; & -I_{d_1} - (x_1 - \Sigma_{1,2}a^*)(x_1 - \Sigma_{1,2}a^*)^\top \\
& + \frac{\psi(x_1, \Sigma_{1,2}a^*, \Sigma_{1,1}) - \psi(x_1, -\Sigma_{1,2}a^*, \Sigma_{1,1})}{\psi(x_1, \Sigma_{1,2}a,^* \Sigma_{1,1}) + \psi(x_1, -\Sigma_{1,2}a^*, \Sigma_{1,1})} I_{d_1} \\
& + \left(\frac{\psi(x_1, \Sigma_{1,2}a^*, \Sigma_{1,1}) - \psi(x_1, -\Sigma_{1,2}a^*, \Sigma_{1,1})}{\psi(x_1, \Sigma_{1,2}a^*, \Sigma_{1,1}) + \psi(x_1, -\Sigma_{1,2}a^*, \Sigma_{1,1})}\right)^2 (x_1 - \Sigma_{1,2}a^*)(x_1 - \Sigma_{1,2}a^*)^\top \\
& \frac{\partial^2}{\partial \Sigma_{1,2}(a^*)^2} \log[1 - m(\Sigma_{1,2}a^*, x_1)] \\
= \; & I_{d_1} - (x_1 - \Sigma_{1,2}a^*)(x_1 - \Sigma_{1,2}a^*)^\top \\
& + \frac{\psi(x_1, \Sigma_{1,2}a^*, \Sigma_{1,1}) - \psi(x_1, -\Sigma_{1,2}a^*, \Sigma_{1,1})}{\psi(x_1, \Sigma_{1,2}a^*, \Sigma_{1,1}) + \psi(x_1, -\Sigma_{1,2}a^*, \Sigma_{1,1})} I_{d_1} \\
& + \left(\frac{\psi(x_1, \Sigma_{1,2}a^*, \Sigma_{1,1}) - \psi(x_1, -\Sigma_{1,2}a^*, \Sigma_{1,1})}{\psi(x_1, \Sigma_{1,2}a^*, \Sigma_{1,1}) + \psi(x_1, -\Sigma_{1,2}a^*, \Sigma_{1,1})}\right)^2 (x_1 - \Sigma_{1,2}a^*)(x_1 - \Sigma_{1,2}a^*)^\top.
\end{aligned}$$

As a result, when the singular values of $\Sigma_{1,2}$ are all finite positive constants, if $\|a - a^*\|^3 = O(1/(n_1 + n_2))$, the posterior distribution of a is approximately a multivariate Gaussian when $d_1 > d_2$, and we obtain $\mathbb{E}\|a - a^*\|^2 = \Theta(d_2/(n_1 + n_2))$. The overall rate becomes $\Theta(d_2/(n_1 + n_2) + (d_1 + d_2)/(n_1 + n_3))$.

Case 2: the arguments are similar to *Case 1*. However, in the last step, the covariance of the posterior of a if not of full rank, so again we obtain $\mathbb{E}\|a - a^*\|^2 = \Theta(d_1/(n_1+n_2))$. On the other hand, since $n_1 + n_2 = o(n_1 + n_3)$, the overall minimax rate becomes $\Theta(d_1/(n_1+n_2))$.

Case 3: we directly assume a prior distribution on μ_0 and do not consider the relationship between μ_0 and $\Sigma_{1,2}$. Following similar arguments as in the previous cases, we obtain $\mathbb{E}\|\mu_0 - \mu_0^*\|^2 = \Theta(d_1/(n_1+n_2))$.

□

D.2 Proof of Example 1

Proof of Bahadur Representation in Example 1. In pretext task, we have

$$\frac{\partial}{\partial \Sigma_{1,1}^{-1}\mu_1}\|x_2 - \phi(x_1)\|_2^2 = -4(x_2 - \phi(x_1))^\top \mu_2 \frac{\partial p(x_1; \Sigma_{1,1}^{-1}\mu_1)}{\partial \Sigma_{1,1}^{-1}\mu_1},$$

$$\frac{\partial}{\partial \mu_2}\|x_2 - \phi(x_1)\|_2^2 = -2(2p(x_1) - 1)(x_2 - \phi(x_1)),$$

and

$$\begin{aligned}\frac{\partial^2}{\partial (\Sigma_{1,1}^{-1}\mu_1)^2}\|x_2 - \phi(x_1)\|_2^2 &= -4(x_2 - \phi(x_1))^\top \mu_2 \frac{\partial^2 p(x_1; \Sigma_{1,1}^{-1}\mu_1)}{\partial (\Sigma_{1,1}^{-1}\mu_1)^2} \\ &\quad +8\|\mu_2\|^2 \left(\frac{\partial p(x_1; \Sigma_{1,1}^{-1}\mu_1)}{\partial (\Sigma_{1,1}^{-1}\mu_1)}\right)\left(\frac{\partial p(x_1; \Sigma_{1,1}^{-1}\mu_1)}{\partial (\Sigma_{1,1}^{-1}\mu_1)}\right)^\top,\end{aligned}$$

$$\frac{\partial^2}{\partial \mu_2^2}\|x_2 - \phi(x_1)\|_2^2 = 2(2p(x_1) - 1)^2 I_{d_2},$$

$$\frac{\partial^2}{\partial (\Sigma_{1,1}^{-1}\mu_1)\partial \mu_2}\|x_2 - \phi(x_1)\|_2^2 = -4\frac{\partial p(x_1; \Sigma_{1,1}^{-1}\mu_1)}{\partial (\Sigma_{1,1}^{-1}\mu_1)}(x_2 - \phi(x_1))^\top + 4(2p(x_1) - 1)\frac{\partial p(x_1; \Sigma_{1,1}^{-1}\mu_1)}{\partial (\Sigma_{1,1}^{-1}\mu_1)}\mu_2^\top.$$

In addition,

$$\begin{aligned}\frac{\partial p(x_1; \Sigma_{1,1}^{-1}\mu_1)}{\partial (\Sigma_{1,1}^{-1}\mu_1)} &= \frac{\partial}{\partial (\Sigma_{1,1}^{-1}\mu_1)}\frac{\psi(x_1, \mu_1, \Sigma_{1,1})}{\psi(x_1, \mu_1, \Sigma_{1,1}) + \psi(x_1, -\mu_1, \Sigma_{1,1})} \\ &= \frac{\partial \psi(x_1, \mu_1, \Sigma_{1,1})}{\partial (\Sigma_{1,1}^{-1}\mu_1)}\frac{1}{\psi(x_1, \mu_1, \Sigma_{1,1}) + \psi(x_1, -\mu_1, \Sigma_{1,1})} \\ &\quad -\frac{\psi(x_1, \mu_1, \Sigma_{1,1})}{[\psi(x_1, \mu_1, \Sigma_{1,1}) + \psi(x_1, -\mu_1, \Sigma_{1,1})]^2}\frac{\partial}{\partial (\Sigma_{1,1}^{-1}\mu_1)}(\psi(x_1, \mu_1, \Sigma_{1,1}) + \psi(x_1, -\mu_1, \Sigma_{1,1})) \\ &= p(x_1; \Sigma_{1,1}^{-1}\mu_1)x_1 - p(x_1; \Sigma_{1,1}^{-1}\mu_1)(2p(x_1; \Sigma_{1,1}^{-1}\mu_1) - 1)x_1 \\ &= 2p(x_1; \Sigma_{1,1}^{-1}\mu_1)(1 - p(x_1; \Sigma_{1,1}^{-1}\mu_1))x_1.\end{aligned}$$

We know that $\sqrt{n_1+n_3}(\widehat{[\Sigma_{1,1}^{-1}\mu_1}, \widehat{\mu}_2] - [(\Sigma_{1,1}^*)^{-1}\mu_1^*, \mu_2^*])$ asymptotically follows $N(\mathbf{0}, A^{-1}B(A^{-1})^\top)$ where

$$\begin{aligned}A &= \frac{\partial^2}{\partial [(\Sigma_{1,1}^{-1}\mu_1), \mu_2]^2}\mathbb{E}\|x_2 - \phi(x_1)\|_2^2, \\ B &= \mathbb{E}\left(\frac{\partial}{\partial [(\Sigma_{1,1}^{-1}\mu_1), \mu_2]}\|x_2 - \phi(x_1)\|_2^2\right)\left(\frac{\partial}{\partial [(\Sigma_{1,1}^{-1}\mu_1), \mu_2]}\|x_2 - \phi(x_1)\|_2^2\right)^\top.\end{aligned}$$

Thus we have

$$A := \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$

with

$$\begin{aligned}
A_{1,1} &= \frac{\partial^2}{\partial(\Sigma_{1,1}^{-1}\mu_1)^2}\mathbb{E}\|x_2-\phi(x_1)\|_2^2 = 8\|\mu_2\|^2\mathbb{E}\left(\frac{\partial p(x_1;\Sigma_{1,1}^{-1}\mu_1)}{\partial(\Sigma_{1,1}^{-1}\mu_1)}\right)\left(\frac{\partial p(x_1;\Sigma_{1,1}^{-1}\mu_1)}{\partial(\Sigma_{1,1}^{-1}\mu_1)}\right)^\top \\
A_{1,2} &= \frac{\partial^2}{\partial(\Sigma_{1,1}^{-1}\mu_1)\partial\mu_2}\mathbb{E}\|x_2-\phi(x_1)\|_2^2 = 4\mathbb{E}\left[(2p(x_1)-1)\frac{\partial p(x_1;\Sigma_{1,1}^{-1}\mu_1)}{\partial(\Sigma_{1,1}^{-1}\mu_1)}\right]\mu_2^\top \\
A_{2,2} &= \frac{\partial^2}{\partial\mu_2^2}\mathbb{E}\|x_2-\phi(x_1)\|_2^2 = 2\mathbb{E}(2p(x_1)-1)^2 I_{d_2}
\end{aligned}$$

Using block matrix inversion on A, we have

$$A^{-1} = \begin{bmatrix} (A_{1,1}-A_{1,2}A_{2,2}^{-1}A_{2,1})^{-1} & -(A_{1,1}-A_{1,2}A_{2,2}^{-1}A_{2,1})^{-1}A_{1,2}A_{2,2}^{-1} \\ -A_{2,2}^{-1}A_{2,1}(A_{1,1}-A_{1,2}A_{2,2}^{-1}A_{2,1})^{-1} & \dots \end{bmatrix}.$$

As a result, the Bahadur representation of $\widehat{\Sigma_{1,1}^{-1}\mu_1}$ is

$$\begin{aligned}
\widehat{\Sigma_{1,1}^{-1}\mu_1} - (\Sigma_{1,1}^*)^{-1}\mu_1^* &= \frac{-(A_{1,1}-A_{1,2}A_{2,2}^{-1}A_{2,1})^{-1}}{n_1+n_3}\sum_{S_1,S_3}\frac{\partial}{\partial\Sigma_{1,1}^{-1}\mu_1}\|x_2-\phi(x_1,\mu_1^*,\Sigma_{1,1}^*)\|_2^2 \qquad (2)\\
&\quad + \frac{(A_{1,1}-A_{1,2}A_{2,2}^{-1}A_{2,1})^{-1}A_{1,2}A_{2,2}^{-1}}{n_1+n_3}\sum_{S_1,S_3}\frac{\partial}{\partial\mu_2}\|x_2-\phi(x_1,\mu_1^*,\Sigma_{1,1}^*)\|_2^2 + o. \qquad (3)
\end{aligned}$$

Expanding $A_{1,2}$ in the Bahadur representation and then

$$\mathbb{E}\|\widehat{\Sigma_{1,1}^{-1}\mu_1} - (\Sigma_{1,1}^*)^{-1}\mu_1^*\|^2 = O\left(\frac{d_1}{n_1+n_3}\right).$$

Finally, in terms of the regret, following Lemma 6.3 of Dan et al. (2020), we have

$$\mathbb{E}R(2p(\cdot,\widehat{\mu}_1)-1) - R(2p(\cdot,\mu_1^*)-1) = O\left(\mathbb{E}\|\widehat{\Sigma_{1,1}^{-1}\mu_1} - (\Sigma_{1,1}^*)^{-1}\mu_1^*\|^2\right) = O\left(\frac{d_1}{n_1+n_3}\right).$$

□

D.3 Proof in Section 3.1

Proof of Theorem 2. Theorem 2 is established upon Assumption 1 and 2.

When CI condition holds, the covariance matrix $A_{1,2}$ in Example 1 is of rank 1. From the pretext task and the family of ϕ we choose, following the same arguments as Example 1[2], under Assumption 2,

$$\frac{1}{\|\beta^*\|^2}\mathbb{E}\|\widehat{\beta}-\beta^*\|^2 = O\left(\frac{d_1}{n_1+n_3}\right).$$

On the other hand, as we mentioned before, the output of $\widehat{\phi}$ is always in the same direction, thus there is no further error involved in the downstream task in terms of the misclassification rate for both plugin estimator and logistic regression.

Denote $\widehat{\mu}_1$ and $\widehat{\mu}_2$ are the parameter associated with $\widehat{\phi}$, then since

$$R^* = R(\phi^* W^*) = R(2p(\cdot,\beta^*)-1),$$

based on Assumption 1, we have

$$R(\widehat{\phi}\widehat{W}) - R^* = O\left(\frac{d_1}{n_1+n_3}\right).$$

□

[2]The way of doing Taylor expansion is the same for Example 1 and other models, and the Bahadur representation (2) is the same.

Proof of Theorem 3, Square Loss. Theorem 3 is established upon Assumption 1 and 2.

Denote (X_1, X_2) as the data matrix for S_1, S_3 (without response), and (X_1', X_2', Y') as the data matrix for S_1, S_2. Denote $\Sigma_{i,j} = \mathbb{E}x_i x_j^\top$ for $i, j \in \{1, 2\}$. Also denote $\Sigma_{i,y}$ as $\mathbb{E}x_i y$ for $i \in \{1, 2\}$.

We first look at the asymptotics of $\widehat{\phi}\widehat{W}$. From the problem setup, we can directly solve $\widehat{\phi}$:

$$\widehat{\phi} = (X_1^\top X_1)^{-1} X_1^\top X_2 \to \Sigma_{1,1}^{-1}\Sigma_{1,2},$$

and further write down $\widehat{W}$:

$$\widehat{W} = (\widehat{\phi}(X_1')^\top \widehat{\phi}(X_1'))^{-1}\widehat{\phi}(X_1')^\top Y' \to (\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2})^{-1}\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,y}.$$

Thus

$$\widehat{\theta} = \widehat{\phi}\widehat{W} \to \Sigma_{1,1}^{-1}\Sigma_{1,2}(\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2})^{-1}\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,y}.$$

We then study the convergence rate. Denote θ_0 as the $\widehat{\theta}$ obtained when $n_i \to \infty$ for all $i = 1, 2, 3$. For the pretext task, one can see that

$$\begin{aligned}
\widehat{\phi} - \phi^* &= (\widehat{\phi}(X_1')^\top \widehat{\phi}(X_1'))^{-1}\widehat{\phi}(X_1')^\top Y' - \phi^* \\
&= \Sigma_{1,1}^{-1}\left(\frac{X_1^\top X_2}{n_1+n_3} - \Sigma_{1,2}\right) + \Sigma_{1,1}^{-1}\left(\frac{X_1^\top X_1}{n_1+n_3} - \Sigma_{1,1}\right)\Sigma_{1,1}^{-1}\Sigma_{1,2} + o.
\end{aligned}$$

$$\begin{aligned}
\widehat{W} - W^* &= (\widehat{\phi}(X_1')^\top \widehat{\phi}(X_1'))^{-1}\widehat{\phi}(X_1')^\top Y' - W^* \\
&= \left(\frac{\phi^*(X_1')^\top \phi^*(X_1')}{n_1+n_2} + (\widehat{\phi}-\phi^*)^\top\Sigma_{1,2} + \Sigma_{2,1}(\widehat{\phi}-\phi^*) + o\right)^{-1} \\
&\quad \cdot \left(\frac{\phi^*(X_1')^\top Y'}{n_1+n_2} + (\widehat{\phi}-\phi^*)^\top\Sigma_{1,1}\theta_0 + o\right),
\end{aligned}$$

where

$$\begin{aligned}
&\left(\frac{\phi^*(X_1')^\top \phi^*(X_1')}{n_1+n_2} + (\widehat{\phi}-\phi^*)^\top\Sigma_{1,2} + \Sigma_{2,1}(\widehat{\phi}-\phi^*) + o\right)^{-1} - \left(\frac{\phi^*(X_1')^\top \phi^*(X_1')}{n_1+n_2}\right)^{-1} \\
= &\left(\frac{\phi^*(X_1')^\top \phi^*(X_1')}{n_1+n_2}\right)^{-1}\left((\widehat{\phi}-\phi^*)^\top\Sigma_{1,2} + \Sigma_{2,1}(\widehat{\phi}-\phi^*)\right)\left(\frac{\phi^*(X_1')^\top \phi^*(X_1')}{n_1+n_2}\right)^{-1} + o.
\end{aligned}$$

As a result, denoting $\widehat{\theta} = \widehat{\phi}^\top\widehat{W}$, we have

$$\widehat{\theta} - \theta_0 = (\widehat{\phi}-\phi^*)^\top W^* + (\phi^*)^\top(\widehat{W} - W^*) + o,$$

therefore,

$$\mathbb{E}\|\widehat{\theta} - \theta_0\|^2 = O\left(\frac{d_2}{n_1+n_2} + \frac{d_1+d_2}{n_1+n_3}\right).$$

□

Proof of Theorem 3, Logistic Regression. Theorem 3 is established upon Assumption 1 and 2.

The derivation for the pretext task is the same as the one in square loss. In the downstream task, denote

$\xi_1(x_1,\phi,W)=\phi(x_1)\frac{1}{1+e^{\phi(x_1)^\top W}}$ and $\xi_2(x_1,\phi,W)=\phi(x_1)\frac{1}{1+e^{-\phi(x_1)^\top W}}$, then

$$
\begin{aligned}
& \frac{1}{n_1+n_2}\sum_{S_1,S_2}[p(x_1)\xi_1(x_1,\phi^*,W^*)-(1-p(x_1))\xi_2(x_1,\phi^*,W^*)] \\
= \; & \frac{1}{n_1+n_2}\sum_{S_1,S_2}[p(x_1)\xi_1(x_1,\phi^*,W^*)-(1-p(x_1))\xi_2(x_1,\phi^*,W^*)] \\
& -\frac{1}{n_1+n_2}\sum_{S_1,S_2}\left[1\{y=1\}\xi_1(x_1,\widehat{\phi},\widehat{W})-1\{y=-1\}\xi_2(x_1,\widehat{\phi},\widehat{W})\right] \\
= \; & \frac{1}{n_1+n_2}\sum_{S_1,S_2}[p(x_1)\xi_1(x_1,\phi^*,W^*)-(1-p(x_1))\xi_2(x_1,\phi^*,W^*)] \\
& -\frac{1}{n_1+n_2}\sum_{S_1,S_2}\left[p(x_1)\xi_1(x_1,\phi^*,\widehat{W})-(1-p(x_1))\xi_2(x_1,\phi^*,\widehat{W})\right] \\
& +\frac{1}{n_1+n_2}\sum_{S_1,S_2}\left[p(x_1)\xi_1(x_1,\phi^*,\widehat{W})-(1-p(x_1))\xi_2(x_1,\phi^*,\widehat{W})\right] \\
& -\frac{1}{n_1+n_2}\sum_{S_1,S_2}\left[p(x_1)\xi_1(x_1,\widehat{\phi},\widehat{W})-(1-p(x_1))\xi_2(x_1,\widehat{\phi},\widehat{W})\right] \\
& +\frac{1}{n_1+n_2}\sum_{S_1,S_2}\left[p(x_1)\xi_1(x_1,\widehat{\phi},\widehat{W})-(1-p(x_1))\xi_2(x_1,\widehat{\phi},\widehat{W})\right] \\
& -\frac{1}{n_1+n_2}\sum_{S_1,S_2}\left[1\{y=1\}\xi_1(x_1,\widehat{\phi},\widehat{W})-1\{y=-1\}\xi_2(x_1,\widehat{\phi},\widehat{W})\right].
\end{aligned}
$$

Observe that with probability tending to 1,

$$
\begin{aligned}
& \frac{1}{n_1+n_2}\sum_{S_1,S_2}[p(x_1)\xi_1(x_1,\phi^*,W^*)-(1-p(x_1))\xi_2(x_1,\phi^*,W^*)] \\
& -\frac{1}{n_1+n_2}\sum_{S_1,S_2}\left[p(x_1)\xi_1(x_1,\phi^*,\widehat{W})-(1-p(x_1))\xi_2(x_1,\phi^*,\widehat{W})\right] \\
= \; & \underbrace{\mathbb{E}\left[p(x_1)\frac{\partial\xi_1}{\partial W}(x_1,\phi^*,W^*)-(1-p(x_1))\frac{\partial\xi_2}{\partial W}(x_1,\phi^*,W^*)\right]}_{:=A}(W^*-\widehat{W})+o,
\end{aligned}
$$

and

$$
\begin{aligned}
& \frac{1}{n_1+n_2}\sum_{S_1,S_2}\left[p(x_1)\xi_1(x_1,\widehat{\phi},\widehat{W})-(1-p(x_1))\xi_2(x_1,\widehat{\phi},\widehat{W})\right] \\
& -\frac{1}{n_1+n_2}\sum_{S_1,S_2}\left[1\{y=1\}\xi_1(x_1,\widehat{\phi},\widehat{W})-1\{y=-1\}\xi_2(x_1,\widehat{\phi},\widehat{W})\right]
\end{aligned}
$$

is a random noise with variance $O(d_2/(n_1+n_2))$. In addition,

$$\begin{aligned}
&\frac{1}{n_1+n_2}\sum_{S_1,S_2}\Big[p(x_1)\xi_1(x_1,\phi^*,\widehat{W})-(1-p(x_1))\xi_2(x_1,\phi^*,\widehat{W})\Big]\\
&-\frac{1}{n_1+n_2}\sum_{S_1,S_2}\Big[p(x_1)\xi_1(x_1,\widehat{\phi},\widehat{W})-(1-p(x_1))\xi_2(x_1,\widehat{\phi},\widehat{W})\Big]\\
=\;&\frac{1}{n_1+n_2}\sum_{S_1,S_2}[p(x_1)\xi_1(x_1,\phi^*,W^*)-(1-p(x_1))\xi_2(x_1,\phi^*,W^*)]\\
&-\frac{1}{n_1+n_2}\sum_{S_1,S_2}\Big[p(x_1)\xi_1(x_1,\widehat{\phi},W^*)-(1-p(x_1))\xi_2(x_1,\widehat{\phi},W^*)\Big]+o\\
=\;&\underbrace{\mathbb{E}\left[p(x_1)\frac{\partial\xi_1}{\partial\phi W}(x_1,\phi^*,W^*)+(1-p(x_1))\frac{\partial\xi_2}{\partial\phi W}(x_1,\phi^*,W^*)\right]}_{:=B}(\widehat{\phi}-\phi^*)W^*+o
\end{aligned}$$

Thus we have

$$\begin{aligned}
\widehat{W}-W^* \;=\;&-\frac{A^{-1}}{n_1+n_2}\sum_{S_1,S_2}[p(x_1)\xi_1(x_1,\phi^*,W^*)-(1-p(x_1))\xi_2(x_1,\phi^*,W^*)]\\
&+A^{-1}B(\widehat{\phi}-\phi^*)W^*\\
&+\frac{A^{-1}}{n_1+n_2}\sum_{S_1,S_2}\Big[p(x_1)\xi_1(x_1,\widehat{\phi},\widehat{W})-(1-p(x_1))\xi_2(x_1,\widehat{\phi},\widehat{W})\Big]\\
&-\frac{A^{-1}}{n_1+n_2}\sum_{S_1,S_2}\Big[1\{y=1\}\xi_1(x_1,\widehat{\phi},\widehat{W})-1\{y=-1\}\xi_2(x_1,\widehat{\phi},\widehat{W})\Big]+o.
\end{aligned}$$

As a result, taking $\widehat{\theta}=\widehat{\phi}\widehat{W}$ and $\theta_0=\phi^*W^*$,

$$\mathbb{E}\|\widehat{\theta}-\theta_0\|^2=O\left(\frac{d_2}{n_1+n_2}+\frac{d_1+d_2}{n_1+n_3}\right).$$

□

D.4 Proof in Section 3.2

Proof of Theorem 4, Logistic Regression, Upper bound. Theorem 4 is built upon Assumption 1, 2, and 3. Assumption 1, 2 ensures the performance of the SSL in clean training, and Assumption 3 regulates the adversarial training.

Below is a summary of the proof:

- Part 1: we show that $\widehat{\theta}$ is consistent.
- Part 2: given the consistency results from Part 1, we can present the Bahadur representation of $\widehat{\theta}$.
- Part 3: we figure out $\widehat{p}$ from clean training, and take it into the Bahadur representation to get the final convergence result.

We first use the data model in Example 1 to go through the proof, then discuss on how to generalize it. The extra moment conditions mentioned in the theorem statement are mentioned when we generalize the proof.

Part 1: Our first aim is to show that $\widehat{\theta}$ is consistent, i.e., $\widehat{\theta}\to\theta^*$. To achieve this, Denoting $\xi_1(x_1,\theta)=\left(x_1-\epsilon\frac{\theta}{\|\theta\|}\right)\frac{1}{1+e^{x_1^\top\theta-\epsilon\|\theta\|}}$ and $\xi_2(x_1,\theta)=\left(x_1+\epsilon\frac{\theta}{\|\theta\|}\right)\frac{1}{1+e^{-x_1^\top\theta-\epsilon\|\theta\|}}$, since the adversarial logistic loss is strongly

convex, there exists some constant $C > 0$ such that

$$\begin{aligned} & \inf_{\|\theta-\theta^*\|\geq\varepsilon} \left\| \mathbb{E}\left[p(x_1)\xi_1(x_1,\theta^*) - (1-p(x_1))\xi_2(x_1,\theta^*)\right] - \mathbb{E}\left[p(x_1)\xi_1(x_1,\theta) - (1-p(x_1))\xi_2(x_1,\theta)\right] \right\| \\ = & \inf_{\|\theta-\theta^*\|\geq\varepsilon} \left\| \mathbb{E}\left[p(x_1)\xi_1(x_1,\theta) - (1-p(x_1))\xi_2(x_1,\theta)\right] \right\| \\ \geq & C\varepsilon^2. \end{aligned}$$

Furthermore, with probability tending to 1, we have

$$\left\| \mathbb{E}\left[p(x_1)\xi_1(x_1,\widehat{\theta}) - (1-p(x_1))\xi_2(x_1,\widehat{\theta})\right] - \frac{1}{\sum n_i}\sum_{S_1,S_2,S_3,S_4}\left[p(x_1)\xi_1(x_1,\widehat{\theta}) - (1-p(x_1))\xi_2(x_1,\widehat{\theta})\right] \right\|$$

is smaller than some $\tau \to 0$ in $\sum n_i$. We further compare

$$\frac{1}{\sum n_i}\sum_{S_1,S_2,S_3,S_4}\left[p(x_1)\xi_1(x_1,\widehat{\theta}) - (1-p(x_1))\xi_2(x_1,\widehat{\theta})\right]$$

to the first-order optimality condition, i.e.

$$\frac{1}{\sum n_i}\sum_{S_1,S_2,S_3,S_4}\left[1_{\{y=1\}}\xi_1(x_1,\widehat{\theta}) - 1_{\{y=-1\}}\xi_2(x_1,\widehat{\theta})\right] = \mathbf{0}.$$

Since S_1 and S_2 contains labels, we have with probability tending to 1, for some constant $r > 0$,

$$\begin{aligned} \sup_{\|\theta-\theta^*\|\leq r} \Bigg\| & \frac{1}{\sum n_i}\sum_{S_1,S_2}\left[p(x_1)\xi_1(x_1,\theta) - (1-p(x_1))\xi_2(x_1,\theta)\right] \\ & -\frac{1}{\sum n_i}\sum_{S_1,S_2}\left[1_{\{y=1\}}\xi_1(x_1,\theta) - 1_{\{y=-1\}}\xi_2(x_1,\theta)\right] \Bigg\| \to 0. \end{aligned}$$

In terms of S_3 and S_4, the labels are imputed from $\widehat{p}$, thus

$$\begin{aligned} & \sup_{\|\theta-\theta^*\|\leq r} \Bigg\| \frac{1}{\sum n_i}\sum_{S_3,S_4}\left[p(x_1)\xi_1(x_1,\theta) - (1-p(x_1))\xi_2(x_1,\theta)\right] \\ & \qquad -\frac{1}{\sum n_i}\sum_{S_3,S_4}\left[\widehat{p}(x_1)\xi_1(x_1,\theta) - (1-\widehat{p}(x_1))\xi_2(x_1,\theta)\right] \Bigg\| \\ = & \sup_{\|\theta-\theta^*\|\leq r} \left\| \frac{1}{\sum n_i}\sum_{S_3,S_4}\left[\left(\widehat{p}(x_1) - p(x_1)\right)\left(\xi_1(x_1,\theta) + \xi_2(x_1,\theta)\right)\right] \right\|, \end{aligned}$$

which also converges to zero since $\widehat{p} \to p$ and each dimension of x_1 has finite fourth moment. Further following similar argument as for S_1 and S_2, we have

$$\begin{aligned} \sup_{\|\theta-\theta^*\|\leq R} \Bigg\| & \frac{1}{\sum n_i}\sum_{S_3,S_4}\left[p(x_1)\xi_1(x_1,\theta) - (1-p(x_1))\xi_2(x_1,\theta)\right] \\ & -\frac{1}{\sum n_i}\sum_{S_3,S_4}\left[1_{\{y=1\}}\xi_1(x_1,\theta) - 1_{\{y=-1\}}\xi_2(x_1,\theta)\right] \Bigg\| \to 0. \end{aligned}$$

Therefore, combining all the above results, we have

$$Pr(\|\widehat{\theta} - \theta^*\| \geq \varepsilon) \quad \leq \quad P\left(\left\| \mathbb{E}\left[p(x_1)\xi_1(x_1,\widehat{\theta}) - (1-p(x_1))\xi_2(x_1,\widehat{\theta})\right] \right\| \geq C\varepsilon^2\right),$$

and $\left\| \mathbb{E}\left[p(x_1)\xi_1(x_1,\widehat{\theta}) - (1-p(x_1))\xi_2(x_1,\widehat{\theta})\right] \right\| \to 0$ in probability, thus with probability tending to 1, $\|\widehat{\theta}-\theta^*\| \to 0$.

Part 2: Given the consistency result, we further consider the convergence rate as a function of $\widehat{p}-p$. We have

$$\begin{aligned}
&\sum_{S_1,S_2,S_3,S_4}[p(x_1)\xi_1(x_1,\theta^*)-(1-p(x_1))\xi_2(x_1,\theta^*)]\\
=&\sum_{S_1,S_2,S_3,S_4}[p(x_1)\xi_1(x_1,\theta^*)-(1-p(x_1))\xi_2(x_1,\theta^*)]-\underbrace{\sum_{S_1,S_2,S_3,S_4}\left[1_{\{y=1\}}\xi_1(x_1,\widehat{\theta})-1_{\{y=-1\}}\xi_2(x_1,\widehat{\theta})\right]}_{=\mathbf{0}\text{ by optimality condition}}\\
=&\sum_{S_1,S_2}\left[p(x_1)\xi_1(x_1,\widehat{\theta})-(1-p(x_1))\xi_2(x_1,\widehat{\theta})\right]-\sum_{S_1,S_2}\left[1_{\{y=1\}}\xi_1(x_1,\widehat{\theta})-1_{\{y=-1\}}\xi_2(x_1,\widehat{\theta})\right]\\
&+\sum_{S_1,S_2,S_3,S_4}[p(x_1)\xi_1(x_1,\theta^*)-(1-p(x_1))\xi_2(x_1,\theta^*)]-\sum_{S_1,S_2,S_3,S_4}\left[p(x_1)\xi_1(x_1,\widehat{\theta})-(1-p(x_1))\xi_2(x_1,\widehat{\theta})\right]\\
&+\sum_{S_3,S_4}\left[p(x_1)\xi_1(x_1,\widehat{\theta})-(1-p(x_1))\xi_2(x_1,\widehat{\theta})\right]-\sum_{S_3,S_4}\left[\widehat{p}(x_1)\xi_1(x_1,\widehat{\theta})-(1-\widehat{p}(x_1))\xi_2(x_1,\widehat{\theta})\right]\\
&+\sum_{S_3,S_4}\left[\widehat{p}(x_1)\xi_1(x_1,\widehat{\theta})-(1-\widehat{p}(x_1))\xi_2(x_1,\widehat{\theta})\right]-\sum_{S_3,S_4}\left[1_{\{y=1\}}\xi_1(x_1,\widehat{\theta})-1_{\{y=-1\}}\xi_2(x_1,\widehat{\theta})\right]\\
:=&\ A_1+A_2+A_3+A_4=A_0
\end{aligned}$$

A_1 is a random variable with noise variance $O(n_1+n_2)$. A_2 measures the difference between θ and θ^*. A_3 measures the difference between $\widehat{p}$ and p. A_4 is a random variable with noise variance $O(n_3+n_4)$. With probability tending to 1,

$$\begin{aligned}
A_2 &= -\sum_{S_1,S_2,S_3,S_4}[p(x_1)\xi_1'(x_1,\theta^*)(\theta-\theta^*)-(1-p(x_1))\xi_2'(x_1,\theta^*)(\theta-\theta^*)]+o\\
&= -\left(\sum n_i\right)\mathbb{E}\left[p(x_1)\xi_1'(x_1,\theta^*)-(1-p(x_1))\xi_2'(x_1,\theta^*)\right](\theta-\theta^*)+O(\sqrt{\textstyle\sum n_i}\|\theta-\theta^*\|)+o.
\end{aligned}$$

Therefore we have

$$\begin{aligned}
\widehat{\theta}-\theta^* &= \frac{1}{\sum n_i}\underbrace{\left(\mathbb{E}\left[p(x_1)\xi_1'(\theta^*)-(1-p(x_1))\xi_2'(\theta^*)\right]\right)^{-1}}_{:=\Gamma^{-1}}(A_1+A_3+A_4-A_0)\\
&= \underbrace{\frac{\Gamma^{-1}}{\sum n_i}\sum_{S_1,S_2}\left(1_{\{y=1\}}-p(x_1)\right)\xi_1(x_1,\theta^*)-(1_{\{y=-1\}}-1+p(x_1))\xi_2(x_1,\theta^*)}_{=O_p\left(\sqrt{d_1(n_1+n_2)}/\sum n_i\right)}\\
&\quad+\underbrace{\frac{\Gamma^{-1}}{\sum n_i}\sum_{S_3,S_4}(p(x_1)-\widehat{p}(x_1))\left(\xi_1(x_1,\theta^*)+\xi_2(x_1,\theta^*)\right)}_{:=\Delta}\\
&\quad+\underbrace{\frac{\Gamma^{-1}}{\sum n_i}\sum_{S_3,S_4}\left(1_{\{y=1\}}-\widehat{p}(x_1)\right)\xi_1(x_1,\theta^*)-(1_{\{y=-1\}}-1+\widehat{p}(x_1))\xi_2(x_1,\theta^*)}_{=O_p\left(\sqrt{d_1(n_3+n_4)}/\sum n_i\right)}\\
&\quad-\underbrace{\frac{\Gamma^{-1}}{\sum n_i}\sum_{S_1,S_2,S_3,S_4}[p(x_1)\xi_1(x_1,\theta^*)-(1-p(x_1))\xi_2(x_1,\theta^*)]}_{=O_p\left(\sqrt{d_1/\sum n_i}\right)}+o.
\end{aligned}$$

Part 3: We further using the construction of $\widehat{p}$ to bound Δ. As mentioned in Example 3, we use $\widehat{\phi}$ to obtain μ_1. We know that $\|\widehat{\mu}_1-\mu_1\|/\|\mu_1\|=O_p(d_1/(n_1+n_3))$ and $\widehat{\mu}_1-\mu_1$ can be represented using Bahadur representation

as well. For (x_1) in S_4, it is independent to $\widehat{\Sigma_{1,1}^{-1}\mu_1}$, thus

$$\mathbb{E}\left\|\Gamma^{-1}\left\langle\widehat{\Sigma_{1,1}^{-1}\mu_1}-(\Sigma_{1,1}^*)^{-1}\mu_1,\frac{\partial p(x_1)}{\partial\Sigma_{1,1}^{-1}\mu_1}\right\rangle(\xi_1(x_1,\theta^*)+\xi_2(x_1,\theta^*))\right\|^2=O\left(\frac{d_1^2}{n_1+n_3}\right).$$

For two samples (x_1) and (x_1') in S_4, we have

$$\begin{aligned}&\mathbb{E}\Bigg[\left\langle\widehat{\Sigma_{1,1}^{-1}\mu_1}-(\Sigma_{1,1}^*)^{-1}\mu_1,\frac{\partial p(x_1)}{\partial\Sigma_{1,1}^{-1}\mu_1}\right\rangle\left\langle\widehat{\Sigma_{1,1}^{-1}\mu_1}-(\Sigma_{1,1}^*)^{-1}\mu_1,\frac{\partial p(x_1')}{\partial\Sigma_{1,1}^{-1}\mu_1}\right\rangle\\&\qquad\times(\xi_1(x_1,\theta^*)+\xi_2(x_1,\theta^*))^\top\Gamma^{-2}(\xi_1(x_1',\theta^*)+\xi_2(x_1',\theta^*))\Bigg]\\=\;&O\left(\frac{d_1}{n_1+n_3}\right).\end{aligned}$$

Thus

$$\begin{aligned}&\mathbb{E}\left\|\frac{\Gamma^{-1}}{\sum n_i}\sum_{S_4}\left\langle\widehat{\Sigma_{1,1}^{-1}\mu_1}-(\Sigma_{1,1}^*)^{-1}\mu_1,\frac{\partial p(x_1)}{\partial\Sigma_{1,1}^{-1}\mu_1}\right\rangle(\xi_1(x_1,\theta^*)+\xi_2(x_1,\theta^*))\right\|^2\\=\;&O\left(\frac{d_1}{n_1+n_3}\left(\frac{n_4}{\sum n_i}\right)^2+\frac{d_1^2}{n_1+n_3}\frac{n_4}{(\sum n_i)^2}\right)\\=\;&O\left(\frac{d_1}{n_1+n_3}\right).\end{aligned}$$

For (x_1,x_2) in S_3, it is correlated to $\widehat{\Sigma_{1,1}^{-1}\mu_1}$, thus

$$\mathbb{E}\left\|\Gamma^{-1}\left\langle\widehat{\Sigma_{1,1}^{-1}\mu_1}-(\Sigma_{1,1}^*)^{-1}\mu_1,\frac{\partial p(x_1)}{\partial\Sigma_{1,1}^{-1}\mu_1}\right\rangle(\xi_1(x_1,\theta^*)+\xi_2(x_1,\theta^*))\right\|^2=O\left(\frac{d_1^2}{n_1+n_3}\right)+O\left(\frac{d_1^3}{(n_1+n_3)^2}\right),$$

and for two samples (x_1,x_2) and (x_1',x_2') in S_3, it becomes

$$\begin{aligned}&\mathbb{E}\Bigg[\left\langle\widehat{\Sigma_{1,1}^{-1}\mu_1}-(\Sigma_{1,1}^*)^{-1}\mu_1,\frac{\partial p(x_1)}{\partial\Sigma_{1,1}^{-1}\mu_1}\right\rangle\left\langle\widehat{\Sigma_{1,1}^{-1}\mu_1}-(\Sigma_{1,1}^*)^{-1}\mu_1,\frac{\partial p(x_1')}{\partial\Sigma_{1,1}^{-1}\mu_1}\right\rangle\\&\qquad\times(\xi_1(x_1,\theta^*)+\xi_2(x_1,\theta^*))^\top\Gamma^{-2}(\xi_1(x_1',\theta^*)+\xi_2(x_1',\theta^*))\Bigg]\\=\;&O\left(\frac{d_1}{n_1+n_3}+\frac{d_1^2}{(n_1+n_3)^2}\right).\end{aligned}$$

Thus although S_3 is related to $\widehat{\mu}_1$, we still have

$$\mathbb{E}\left\|\frac{\Gamma^{-1}}{\sum n_i}\sum_{S_3,S_4}\left\langle\widehat{\Sigma_{1,1}^{-1}\mu_1}-(\Sigma_{1,1}^*)^{-1}\mu_1,\frac{\partial p(x_1)}{\partial\Sigma_{1,1}^{-1}\mu_1}\right\rangle(\xi_1(x_1,\theta^*)+\xi_2(x_1,\theta^*))\right\|^2=O\left(\frac{d_1}{n_1+n_3}\right).$$

Combining Part 2 and Part 3 we have

$$\mathbb{E}\|\widehat{\theta}-\theta^*\|^2=O\left(\frac{d_1}{n_1+n_3}\right),$$

and further we obtain

$$\mathbb{E}R(\widehat{\theta},\epsilon)-R(\theta^*,\epsilon)=O\left(\frac{d_1}{n_1+n_3}\right).$$

Assumption 3 guarantees that the above analysis can be generalized to other distributions. Furthermore, although the Bahadur representation of $\widehat{\Sigma_{1,1}^{-1}\mu_1}$ involves X_2, since X_1 and X_2 are conditionally independent given $Y=y$, there is no extra requirement on X_2. □

Proof of Theorem 4, Square Loss, Upper bound. Theorem 4 is built upon Assumption 1, 2, and 3. Assumption 1, 2 ensures the performance of the SSL in clean training, and Assumption 3 regulates the adversarial training.

The proof if similar to the one for logistic regression below and replace ξ_1 to $\xi(x_1, y, \theta) = \left(x_1 + \epsilon \frac{\theta}{\|\theta\|} \operatorname{sgn}(y - x_1^\top \theta)\right)(y - x_1^\top \theta - \operatorname{sgn}(y - x_1^\top \theta)\epsilon\|\theta\|)$. The adversarial square loss is strongly convex.

Assumption 3 ensures that the above analysis can generalize to other distributions. □

Proof of Theorem 4, Linear without CI, Upper bound. Theorem 4 is built upon Assumption 1, 2, and 3. Assumption 1, 2 ensures the performance of the SSL in clean training, and Assumption 3 regulates the adversarial training.

In the proof when CI holds, the Bahadur representation of $\widehat{\theta} - \theta^*$ does not directly utilize the CI condition. Instead, we use the convergence result of $\widehat{p}$. Therefore, similarly, we use the convergence result of SSL from Theorem 3 to obtain the convergence results of $\widehat{p}$ to apply to Part 3. □

Proof of Proposition 1. **Logistic Regression** Since $\widehat{p}$ is consistent, one can follow Part 1 of the proof of Theorem 4 to obtain the consistency result. In terms of the convergence rate, following Part 2 of the proof of Theorem 4, we have

$$\theta - \theta^* = \frac{\Gamma^{-1}}{\sum n_i} \sum_{S_3, S_4} (p(x_1) - \widehat{p}(x_1)) \left(\xi_1(x_1, \theta^*) + \xi_2(x_1, \theta^*)\right) + \Delta$$

for some $\mathbb{E}\|\Delta\|^2 = O(d_1/(\sum n_i))$. Since $\mathbb{E}\|\xi_1\|^2 = O(d_1)$ and $\mathbb{E}\|\xi_2\|^2 = O(d_1)$, we have $\mathbb{E}\|\widehat{\theta} - \theta^*\|^2 = O(d_1/(\sum n_i) + \mathbb{E}\|X_1\|^2 \|\widehat{p}(X_1) - p(X_1)\|^2)$.

Square Loss When $\mathbb{E}\|X_1\|^2 \|\widehat{p}(X_1) - p(X_1)\|^2 (X_1\theta^*)^2 \to 0$, the convergence rate of $\widehat{\theta} - \theta^*$ is $\mathbb{E}\|\widehat{\theta} - \theta^*\|^2 = O(d_1/(\sum n_i) + \mathbb{E}\|X_1\|^2 \|\widehat{p}(X_1) - p(X_1)\|^2 (X_1\theta^*)^2)$. □

D.5 Discussion about Logistic Regression

Our first goal is to investigate in what is the θ^* in logistic regression. Assume there are infinite labeled data, the first-order optimality condition is

$$\mathbb{E}1_{\{Y=1\}}(x + \epsilon\frac{\theta}{\|\theta\|})\frac{1}{1 + e^{x^\top\theta - \epsilon\|\theta\|}} - \mathbb{E}1_{\{Y=-1\}}(x + \epsilon\frac{\theta}{\|\theta\|})\frac{1}{1 + e^{-x^\top\theta + \epsilon\|\theta\|}} + \lambda\theta = \mathbf{0}$$

From the distribution of (X, Y), we have

$$\mathbb{E}1_{\{Y=1\}}(x + \epsilon\frac{\theta}{\|\theta\|})\frac{1}{1 + e^{x^\top\theta - \epsilon\|\theta\|}} + \mathbb{E}1_{\{Y=1\}}(x - \epsilon\frac{\theta}{\|\theta\|})\frac{1}{1 + e^{x^\top\theta + \epsilon\|\theta\|}} + \lambda\theta = \mathbf{0}$$

$$\begin{aligned}
& \mathbb{E}1_{\{Y=1\}} x \left(\frac{1}{1 + e^{x^\top\theta - \epsilon\|\theta\|}} + \frac{1}{1 + e^{x^\top\theta + \epsilon\|\theta\|}}\right) \\
= \; & \mathbb{E}\left\{\mathbb{E}\left[1_{\{Y=1\}} x \left(\frac{1}{1 + e^{x^\top\theta - \epsilon\|\theta\|}} + \frac{1}{1 + e^{x^\top\theta + \epsilon\|\theta\|}}\right) \middle| \theta^\top x = u\right]\right\} \\
= \; & \mathbb{E}\left\{\mathbb{E}\left[1_{\{Y=1\}} x \left(\frac{1}{1 + e^{u - \epsilon R}} + \frac{1}{1 + e^{u + \epsilon R}}\right) \middle| \theta^\top x = u\right]\right\}.
\end{aligned}$$

Since $(X, \theta^\top X)$ follows Gaussian with mean $(\mu, \theta^\top\mu)$ and variance $\begin{bmatrix} \Sigma & \Sigma\theta \\ \theta^\top\Sigma & \theta^\top\Sigma\theta \end{bmatrix}$, we have

$$\mathbb{E}(x|\theta^\top x = u) = \mu + \frac{\Sigma\theta}{\theta^\top\Sigma\theta}(u - \theta^\top\mu) = \mu + \frac{\Sigma\theta}{\theta^\top\Sigma\theta}u - \frac{\Sigma\theta\theta^\top\mu}{\theta^\top\Sigma\theta},$$

thus denote $\zeta = \zeta(\theta, \mu, \Sigma, \epsilon) = \mathbb{E}\left[\left(\frac{1}{1+e^{u-\epsilon R}} - \frac{1}{1+e^{u+\epsilon R}}\right)\right]$, we have

$$\begin{aligned} & \mathbb{E}\left\{\mathbb{E}\left[1_{\{Y=1\}} x \left(\frac{1}{1+e^{u-\epsilon R}} + \frac{1}{1+e^{u+\epsilon R}}\right) \middle| \theta^\top x = u\right]\right\} \\ = \quad & \frac{\Sigma\theta}{\theta^\top \Sigma \theta} \mathbb{E}\left[u\left(\frac{1}{1+e^{u-\epsilon R}} + \frac{1}{1+e^{u+\epsilon R}}\right)\right] + \mu\zeta - \frac{\Sigma\theta\theta^\top\mu}{\theta^\top\Sigma\theta}\zeta \end{aligned}$$

Return to the optimal condition, we have

$$\mu\zeta - \frac{\Sigma\theta\theta^\top\mu}{\theta^\top\Sigma\theta}\zeta + \frac{\Sigma\theta}{\theta^\top\Sigma\theta}\mathbb{E}\left[u\left(\frac{1}{1+e^{u-\epsilon R}} + \frac{1}{1+e^{u+\epsilon R}}\right)\right] + \epsilon\frac{\theta}{\|\theta\|}\zeta + \lambda\theta = \mathbf{0}$$

Dividing ζ,

$$\mu - \frac{\theta^\top\mu\Sigma\theta}{\theta^\top\Sigma\theta} + \frac{\Sigma\theta}{\theta^\top\Sigma\theta}\mathbb{E}\left[u\left(\frac{1}{1+e^{u-\epsilon R}} + \frac{1}{1+e^{u+\epsilon R}}\right)\right]\frac{1}{\zeta} + \epsilon\frac{\theta}{\|\theta\|} + \lambda\theta = \mathbf{0}$$

Thus for some constants A and B,

$$\theta = -\left(A\Sigma + BI\right)^{-1}\mu.$$

The above result reveals that, the convergence rate of logistic regression is the same as plugin estimator. However, the relationship between (A, B) in the above formula may be different from the one in plugin estimator, leading to potential bias in adversarial setup.

E Proof for Section 5 and A

E.1 Proof for Section 5.1

To provide detailed conditions on the neural network and configurations, we first define some quantities. For two unit vectors $s, t \in \mathbb{R}^{d_1}$, define a function h as

$$h(s,t) = \mathbb{E}_{w\sim N(0,I_{d_1})}(s^\top t 1\{w^\top s \geq 0, w^\top y \geq 0\}) = \frac{s^\top t(\pi - \arccos(s^\top t))}{2\pi}. \tag{4}$$

There are total $n_1 + n_3$ samples which have (x_1, x_2). We take $n = n_1 + n_2$ and index the samples as (x_1^i, x_2^i) for $i = 1, ..., n$. After indexing the samples, we then define $H^{(\infty)}$ as a $n \times n$ matrix such that $H_{i,j}^{(\infty)} = h(x_1^i, x_1^j)$.

Proof of Proposition 2. The detailed conditions for Proposition 2 are as follows:

- The learning rate $\eta = \Theta(n^{-\frac{3d_1-1}{2d_1-1}})$.
- The penalty $\lambda = \Theta(n^{\frac{d_1-1}{2d_1-1}})$.
- The number of hidden nodes $m \geq \tau^{-2}\text{poly}(n, 1/\lambda_0)$ for some initialization variance $\tau^2 = O(1)$ and $\lambda_0 = \lambda_{\min}^{-1}(^{(\infty)})$.
- The number of iterations T satisfies $\log(\text{poly}(n, \tau, 1/\lambda_0)) \ll \eta\mu T \ll \log(\text{poly}(\tau, 1/n, \sqrt{m}))$.
- The input x_1 is normalized such that $\|x_1\| = 1$, and this normalization does not change the minimal misclassification rate. Denoting μ_1 and μ_1' as the conditional expectation of X_1 (after normalization) under $y = \pm 1$, then both $\|\mu_1\|$ and $\|\mu_1'\|$ are nonzero.

Since d_2 is a constant, training a neural network with input dimension d_1 and output dimension d_2 is equivalent to training d_2 different neural networks. Therefore following Hu et al. (2021), one obtain that

$$\|\widehat{\phi} - \phi^*\|_2^2 = O_p\left(n^{-\frac{d}{2d-1}}\right). \tag{5}$$

Denote (X_1, X_2) as the data matrix for S_1, S_3 (without response), and (X_1', X_2', Y') as the data matrix for S_1, S_2. In terms of $\widehat{W}$, under CI,

$$\frac{\phi^*(X_1')^\top \phi^*(X_1')}{n_1+n_2} \tag{6}$$

is not a full rank matrix (at most rank two for binary classification). To avoid singular matrix problem, we take

$$\begin{aligned}
& \widehat{W} - W^* \\
= & \lim_{\lambda\to 0} (\widehat{\phi}(X_1')^\top \widehat{\phi}(X_1') + \lambda I_{d_1})^{-1} \widehat{\phi}(X_1')^\top Y' - W^* \\
= & \lim_{\lambda\to 0} \left(\frac{\phi^*(X_1')^\top \phi^*(X_1')}{n_1+n_2} + \frac{(\widehat{\phi}-\phi^*)^\top \phi^*}{n_1+n_2} + \frac{(\phi^*)^\top (\widehat{\phi}-\phi^*)}{n_1+n_2} + o + \lambda I_{d_1} \right)^{-1} \\
& \cdot \left(\frac{\phi^*(X_1')^\top Y'}{n_1+n_2} + \frac{(\widehat{\phi}-\phi^*)^\top Y'}{n_1+n_2} + o \right) - W^* \\
= & \lim_{\lambda\to 0} \left(\frac{\phi^*(X_1')^\top \phi^*(X_1')}{n_1+n_2} + \lambda I_{d_1} \right)^{-1} \left(\frac{(\widehat{\phi}-\phi^*)^\top \phi^*}{n_1+n_2} + \frac{(\phi^*)^\top (\widehat{\phi}-\phi^*)}{n_1+n_2} \right) \left(\frac{\phi^*(X_1')^\top \phi^*(X_1')}{n_1+n_2} + \lambda I_{d_1} \right)^{-1} \\
& \cdot \left(\frac{\phi^*(X_1')^\top Y'}{n_1+n_2} \right) \\
& + \lim_{\lambda\to 0} \left(\frac{\phi^*(X_1')^\top \phi^*(X_1')}{n_1+n_2} + \lambda I_{d_1} \right)^{-1} \left(\frac{(\widehat{\phi}-\phi^*)^\top Y'}{n_1+n_2} \right) + o.
\end{aligned}$$

As a result, $\widehat{W} \to W^*$.

Different from Lee et al. (2020), we are considering the regret (the difference on the misclassification rate between the estimated classifier and the Bayes classifier) as the final performance measure. Based on the definition of W^* and ϕ^*, if we use $\text{sgn}(\widetilde{W}^\top \phi^*(x_1))$ for some estimate $\widetilde{W}$ such that $\widetilde{W} \to W$, the classifier always makes the exact same decision as the Bayes classifier.

On the other hand, for the estimated output

$$\widehat{W}^\top \widehat{\phi}(x_1) = \widehat{W}^\top \phi^*(x_1) + (W^*)^\top (\widehat{\phi}(x_1) - \phi^*(x_1)) + o,$$

since we have argued that $\text{sgn}(\widehat{W}^\top \phi^*(x_1)) \equiv \text{sgn}((W^*)^\top \phi^*(x_1))$, we aims to study how $(W^*)^\top (\widehat{\phi}(x_1) - \phi^*(x_1))$ affects the regret.

The regret can be represented as

$$\begin{aligned}
& \int |1/2 - p(x_1)| 1\left\{ \text{sgn}(\widehat{W}^\top \widehat{\phi}(x_1)) \neq \text{sgn}((W^*)^\top \phi^*(x_1)) \right\} dP(x_1) \\
= & \int |1/2 - p(x_1)| 1\left\{ \text{sgn}\left(\widehat{W}^\top \phi^*(x_1) + (W^*)^\top (\widehat{\phi}(x_1) - \phi^*(x_1)) \right) \neq \text{sgn}\left((W^*)^\top \phi^*(x_1) \right) \right\} dP(x_1) + o \\
\leq & \int |1/2 - p(x_1)| 1\left\{ \|W^*\| \|\widehat{\phi}(x_1) - \phi^*(x_1))\| \geq \min\left(|\widehat{W}^\top \phi^*(x_1)|, |(W^*)^\top \phi^*(x_1)| \right) \right\} dP(x_1) + o \\
\leq & (1+o(1)) \int |1/2 - p(x_1)| 1\left\{ \|\widehat{\phi}(x_1) - \phi^*(x_1))\| \geq |(W^*)^\top \phi^*(x_1)| / \|W^*\| \right\} dP(x_1) + o.
\end{aligned}$$

Further, since $(W^*)^\top \phi^*(x_1) = p(x_1)(W^*)^\top \mu_1 + (1-p(x_1))(W^*)^\top \mu_1'$ and $\|\mu_1\|, \|\mu_1'\| = \Theta(1)$, there exists some

$c > 0$ such that

$$\begin{aligned}
& \int |1/2 - p(x_1)| 1\left\{\|\widehat{\phi}(x_1) - \phi^*(x_1))\| \geq |(W^*)^\top \phi^*(x_1)|/\|W^*\|\right\} dP(x_1) \\
\leq & \int |1/2 - p(x_1)| 1\left\{\|\widehat{\phi}(x_1) - \phi^*(x_1))\| \geq c|1/2 - p(x_1)|\right\} dP(x_1) \\
\leq & \frac{1}{c}\int \|\widehat{\phi}(x_1) - \phi^*(x_1))\| 1\left\{\|\widehat{\phi}(x_1) - \phi^*(x_1))\| \geq c|1/2 - p(x_1)|\right\} dP(x_1) \\
\leq & \frac{1}{c}\sqrt{\int \|\widehat{\phi}(x_1) - \phi^*(x_1))\|^2 dP(x_1)},
\end{aligned}$$

which becomes $O_p((n_1+n_3)^{-d/2(2d-1)})$ based on (5). □

E.2 Proof for Section A

Proof of Theorem 5, Linear ϕ, Regression, Lower Bound. Assume $\Sigma_{1,1} = Var(x_1) = I_{d_1}$ and $\Sigma_{2,2} = Var(x_2) = I_{d_2}$. When using linear ϕ, it is easy to see that $\phi^*(x_1) = \Sigma_{2,1}x_1$ for $\Sigma_{2,1} = Cov(x_2, x_1)$. Thus when SSL is unbiased, we have $\theta_0 x_1 \equiv a^\top \Sigma_{2,1} x_1$, i.e. $\theta_0 = a^\top \Sigma_{2,1}$ for some vector a. From Lemma 3, we have known fixed a,

$$\inf_{\widehat{\theta}} \sup_{\Sigma_{1,2}} \mathbb{E}\|\widehat{\theta} - \Sigma_{1,2}a\|^2 = \Omega\left(\frac{d_1+d_2}{n_1+n_3}\|a\|^2\right),$$

and for known $\Sigma_{1,2}$, from Lemma 2 we have

$$\inf_{\widehat{\theta}} \sup_{a} \mathbb{E}\|\widehat{a} - a\|^2 = \Omega\left(\frac{\sigma^2 d_2}{n_1+n_2}\right).$$

Using the strong convexity property of the risk, we then obtain

$$\inf_{\widehat{\theta}} \sup_{\Sigma_{1,2}} \mathbb{E}R(\widehat{\theta}) - R^* = \Omega\left(\frac{d_1+d_2}{n_1+n_3}\|a\|^2 + \frac{\sigma^2 d_2}{n_1+n_2}\right),$$

where $\|a\| = \Theta(\|\theta_0\|)$ based on assumption. □

Proof of Theorem 5, Linear ϕ, Regression, Upper Bound. The proof is similar to the square loss case in Theorem 3. Denote (X_1, X_2) as the data matrix for S_1, S_3 (without response), and (X_1', X_2', Y') as the data matrix for S_1, S_2. Denote $\Sigma_{i,j} = \mathbb{E}x_i x_j^\top$ for $i, j \in \{1, 2\}$. Also denote $\Sigma_{i,y}$ as $\mathbb{E}x_i y$ for $i \in \{1, 2\}$.

We first look at the asymptotics of $\widehat{\phi}\widehat{W}$. From the problem setup, we can directly solve $\widehat{\phi}$:

$$\widehat{\phi} = (X_1^\top X_1)^{-1} X_1^\top X_2 \to \Sigma_{1,1}^{-1}\Sigma_{1,2},$$

and further write down $\widehat{W}$:

$$\widehat{W} = (\widehat{\phi}(X_1')^\top \widehat{\phi}(X_1'))^{-1}\widehat{\phi}(X_1')^\top Y' \to (\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2})^{-1}\Sigma_{2,1}\theta_0.$$

Thus

$$\widehat{\theta} = \widehat{\phi}\widehat{W} \to \Sigma_{1,1}^{-1}\Sigma_{1,2}(\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2})^{-1}\Sigma_{2,1}\theta_0.$$

When $\theta_0 = \Sigma_{1,1}^{-1}\Sigma_{2,1}a_0$ for some a_0, we have

$$\Sigma_{1,1}^{-1}\Sigma_{1,2}(\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2})^{-1}\Sigma_{2,1}\theta_0 = \theta_0,$$

i.e., SSL is unbiased.

We next study the convergence rate. Denote θ_0 as the $\widehat{\theta}$ obtained when $n_i \to \infty$ for all $i = 1, 2, 3$. For classification task, there is no preference on the magnitude of θ_0 as it works as a linear classifier, so we take $\|\theta_0\| = 1$. For the pretext task, one can see that

$$\begin{aligned}\widehat{\phi} - \phi^* &= (\widehat{\phi}(X_1')^\top \widehat{\phi}(X_1'))^{-1} \widehat{\phi}(X_1')^\top Y' - \phi^* \\ &= \Sigma_{1,1}^{-1} \left(\frac{X_1^\top X_2}{n_1 + n_3} - \Sigma_{1,2} \right) + \Sigma_{1,1}^{-1} \left(\frac{X_1^\top X_1}{n_1 + n_3} - \Sigma_{1,1} \right) \Sigma_{1,1}^{-1} \Sigma_{1,2} + o.\end{aligned}$$

$$\begin{aligned}\widehat{W} - W^* &= (\widehat{\phi}(X_1')^\top \widehat{\phi}(X_1'))^{-1} \widehat{\phi}(X_1')^\top Y' - W^* \\ &= \left(\frac{\phi^*(X_1')^\top \phi^*(X_1')}{n_1 + n_2} + (\widehat{\phi} - \phi^*)^\top \Sigma_{1,2} + \Sigma_{2,1}(\widehat{\phi} - \phi^*) + o \right)^{-1} \\ &\quad \cdot \left(\frac{\phi^*(X_1')^\top Y'}{n_1 + n_2} + (\widehat{\phi} - \phi^*)^\top \Sigma_{1,1} \theta_0 + o \right) - W^*,\end{aligned}$$

where

$$\begin{aligned}&\left(\frac{\phi^*(X_1')^\top \phi^*(X_1')}{n_1 + n_2} + (\widehat{\phi} - \phi^*)^\top \Sigma_{1,2} + \Sigma_{2,1}(\widehat{\phi} - \phi^*) + o \right)^{-1} - \left(\frac{\phi^*(X_1')^\top \phi^*(X_1')}{n_1 + n_2} \right)^{-1} \\ = &\left(\frac{\phi^*(X_1')^\top \phi^*(X_1')}{n_1 + n_2} \right)^{-1} \left((\widehat{\phi} - \phi^*)^\top \Sigma_{1,2} + \Sigma_{2,1}(\widehat{\phi} - \phi^*) \right) \left(\frac{\phi^*(X_1')^\top \phi^*(X_1')}{n_1 + n_2} \right)^{-1} + o.\end{aligned}$$

As a result, denoting $\widehat{\theta} = \widehat{\phi}^\top \widehat{W}$, we have

$$\widehat{\theta} - \theta_0 = (\widehat{\phi} - \phi^*)^\top W^* + (\phi^*)^\top (\widehat{W} - W^*) + o,$$

therefore,

$$\mathbb{E}\|\widehat{\theta} - \theta_0\|^2 = O\left(\frac{d_2 \sigma_2^2}{n_1 + n_2} + \frac{d_1 + d_2}{n_1 + n_3} \|\theta_0\|^2 \right).$$

□

Proof of Theorem 5, Linear ϕ, Regression, Adversarial. For the convergence upper bound, following the decomposition of estimation error in Xing et al. (2021b), beside the part from $\|\widehat{\theta} - \theta_0\|^2$ in clean training, there is one extra part due to the information limit on $\Sigma_{1,1}$. However, since there are $\sum n_i$ samples to provide information of x_1, the new term can be ignored, so the minimax lower bound in adversarial training is the same to clean training.

In terms of the lower bound, we know that

$$\theta(\epsilon) = (\Sigma_{1,1} + \lambda I_{d_1})^{-1} \Sigma_{1,1} \theta(0),$$

thus following Xing et al. (2021b), we consider two scenarios: (3) $\Sigma_{1,1}$ is known and we impose prior distribution on $\theta(0)$; (4) $\theta(0)$ is known and we impose prior distribution on $\Sigma_{1,1}$. Following the arguments in clean training, we have scenario (3) reduces to clean training setup. For scenario (4), following Xing et al. (2021b) we obtain

$$\inf_{\widehat{\theta}} \sup_{\Sigma, a, \epsilon} \mathbb{E}\|\widehat{\theta} - \theta(\epsilon)\|^2 = \Theta\left(\frac{d_1}{n_1 + n_2 + n_3 + n_4} \right).$$

To conclude,

$$\inf_{\widehat{\theta}} \sup_{\Sigma, a, \epsilon} \mathbb{E}\|\widehat{\theta} - \theta(\epsilon)\|^2 = \Theta\left(\frac{\sigma^2 d_2}{n_1 + n_2} + \frac{d_1}{n_1 + n_3} + \frac{d_1}{n_1 + n_2 + n_3 + n_4} \right) = \Theta\left(\frac{\sigma^2 d_2}{n_1 + n_2} + \frac{d_1}{n_1 + n_3} \right).$$

□