# Sharp Statistical Guarantees for Adversarially Robust Gaussian Classification

**Chen Dan** [1] **Yuting Wei** [1] **Pradeep Ravikumar** [1]

## Abstract

Adversarial robustness has become a fundamental requirement in modern machine learning applications. Yet, there has been surprisingly little statistical understanding so far. In this paper, we provide the first result of the *optimal* minimax guarantees for the excess risk for adversarially robust classification, under Gaussian mixture model proposed by (Schmidt et al., 2018). The results are stated in terms of the *Adversarial Signal-to-Noise Ratio (AdvSNR)*, which generalizes a similar notion for standard linear classification to the adversarial setting. For the Gaussian mixtures with AdvSNR value of $r$, we establish an excess risk lower bound of order $\Theta(e^{-(\frac{1}{8}+o(1))r^2}\frac{d}{n})$ and design a computationally efficient estimator that achieves this optimal rate. Our results built upon minimal set of assumptions while cover a wide spectrum of adversarial perturbations including $\ell_p$ balls for any $p \geq 1$.

## 1. Introduction

Recent years, machine learning algorithms have revolutionized our life due to their tremendous success in a variety of different domains such as image classification, natural language processing and strategy games (e.g. Krizhevsky et al. (2012); Bahdanau et al. (2014); Silver et al. (2016)). These algorithms often achieve extremely accurate performances yet are susceptible to small perturbations of the inputs. In particular, Szegedy et al. (2013) (among others e.g. Goodfellow et al. (2014); Papernot et al. (2016)) noticed that small perturbations (nearly imperceptible) to images could cause neural network classifiers to make wrong predictions with high confidence. While a growing amount of effort has been made in order to empirically improve the robustness of these learning algorithms against adversarial attacks, the problems of assessing statistical optimality, understanding generalization and statistical significance are important but far less understood. In this paper, we take a step towards this end.

In this work, we consider the adversarially robust classification problem under the Gaussian mixture model proposed by Schmidt et al. (2018). While the classification for mixture of Gaussian distributions — which is also referred to as discriminant analysis — has now been standard in statistics and computer science literature (see, e.g. McLachlan & Peel (2004)), it is only until recently that researchers start to consider what can go wrong in the adversarial scenarios for this simple problem. It turns out (and as is shown in the sequel) that this simple yet instructive model demonstrates clear tradeoffs between adversarially robustness and the statistical complexities, and at the same time, capturing some of the features one would encounter in real applications.

Under minimal assumptions of the adversarial perturbations, we provide optimal minimax lower bounds, and show that a natural computationally efficient estimator achieves these minimax lower bounds in terms of the adversarial signal to noise ratio. Putting these together gives a sharp characterization of the intrinsic hardness of this problem in terms of how far one can push towards a robust estimator without any essential loss of statistical accuracy. These optimal lower and upper bounds are useful since that they provide a comprehensive view of the adversarially robust sample complexity of the conditional Gaussian model, which could then be contrasted with that of the rates of the classical conditional Gaussian model.

Despite of an extensive line of work considering this problem, Schmidt et al. (2018) and Bhagoji et al. (2019) lie most closely to this paper. In order to obtain tight statistical characterizations of the risk, they made a number of simplifications, which thus do not directly provide answers to the minimax sample complexity of the original problem. As one main contrast, they consider the Bayesian setting where the means of the conditional Gaussians have as prior an independent standard Gaussian distribution. For other simplifications, Schmidt et al. (2018) considered the spherical models so that the covariance is identity and also made additional simplifications such as large separation between two Gaussians and an upper bound on the noise level. These additional assumptions made it hard to compare with that of the adversary-free scenario. More detailed comparisons

[1]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. Correspondence to: Chen Dan <cdan@cs.cmu.edu>.

*Proceedings of the $37^{th}$ International Conference on Machine Learning*, Online, PMLR 119, 2020. 

and discussions are provided after our main results.

### 1.1. Our contributions

The main contributions of this paper are summarized below, all of which are built upon a careful analysis of the classification error for linear classifiers.

- We develop the first minimax lower bounds for the classification excess risk in the conditional Gaussian model, stated in Theorem 4.1. In terms of the Adversarial Signal-to-Noise Ratio (AdvSNR), this excess risk scales as $\Omega_P(\exp(-(\frac{1}{8}+o(1))r^2)\frac{d}{n})$ for $\mathrm{AdvSNR}=r$, dimension $d$ and sample size $n$.
- We construct a computationally efficient estimator based on the solution of a constrained quadratic optimization problem that has excess risk of order $O_P(\exp(-(\frac{1}{8}+o(1))r^2)\frac{d}{n})$. This result is given in Theorem 3.1. Hence, the upper bound is nearly tight (up to lower order terms in $r$) with the minimax lower bound in our regime of interest in terms of AdvSNR $r$, dimension $d$ and sample size $n$ .
- The recipe provided herein, works for a wide range of adversarial perturbations, generalizing the result by Schmidt et al. (2018) who focus only on the $\ell_\infty$-type perturbations.
- Finally, our results are built upon minimum set of assumptions, without assuming strong separations between two classes, allowing for unknown and arbitrary covariance structure and the rates are naturally adaptive to the true signal.

Our findings unveil new insights into the adversarially robust sample complexity of the conditional Gaussian model which goes beyond of what the current theory has to offer.

### 1.2. Other related works

The conditional Gaussian models or mixture of Gaussians has been studied a lot in statistics and computer science literature. An incomplete and more recent list includes Kim et al. (2006); Azizyan et al. (2013); Li et al. (2015; 2017); Cai & Zhang (2019). In the context of adversarial robustness, since the seminal work of (Schmidt et al., 2018), there are several other papers that studied the sample complexity issue in conditional Gaussian models. Bhagoji et al. (2019) also provided a slightly improved bound in the same setting. Carmon et al. (2019), Stanforth et al. (2019), Zhai et al. (2019) showed that with the help of unlabeled data, it is possible to achieve high robust accuracy with the same number of labeled data required for standard learning.

Another line of research study the sample complexity of adversarially robust learning under the PAC framework, using extensions of Rademacher complexity or VC dimension, including Attias et al. (2018), Khim & Loh (2018), Yin et al. (2018), Cullina et al. (2018), Montasser et al. (2019), Awasthi et al. (2020). The tradeoff in standard and robust accuracy has been theoretically and empirically studied in Zhang et al. (2019), Suggala et al. (2018), Tsipras et al. (2018), Raghunathan et al. (2020) and Javanmard et al. (2020).

Several previous works analyzed the robustness of specific family of classifiers. The early work of Xu et al. (2009a;b) estabilished the connections between robust optimization for linear models and certain types of regularization in classification and regression settings. Subsequently, Xu & Mannor (2012) also showed that under certain notion of robustness, robust algorithms can generalize well. Wang et al. (2017) studied the robustness of nearest neighbor classifiers.

From the aspect of computational complexity, some recent works showed that learning a robust model or even verifying robustness of a given model can be computationally hard, including (Bubeck et al., 2018a;b) and (Awasthi et al., 2019; Weng et al., 2018).

### 1.3. Notations

For the reader's convenience, we list here our notational conventions.

For positive semi-definite matrix $A$, we use $\|x\|_A := \sqrt{x^TAx}$. Let $\Phi(\cdot)$ the CDF of standard Gaussian distribution $\mathcal{N}(0,1)$ and $\bar{\Phi}(x) := 1-\Phi(x)$. The notation $f(n,d) = O\big(g(n,d)\big)$ means that there exits a universal constant $c>0$ that does not depend on the problem parameters such as $n, d$ etc, such that $|f(n,d)| \le c|g(n,d)|$. Similarly, we define $f(n,d) = \Omega\big(g(n,d)\big)$ when there exist constants $c_1, c_2 > 0$ such that $c_1|g(n,d)| \le |f(n,d)| \le c_2|g(n,d)|$. Notation $O_P, \Omega_P$ are used if the corresponding relations happen with probability converges to 1 as $n\to\infty$ (see e.g. Chapter 2 of (Van der Vaart, 2000)). We define the $\ell_p$ norm $\|x\|_p = (\sum_{i=1}^d x_i^p)^{1/p}$ and the corresponding $\ell_p$-ball as $\{x\in\mathbb{R}^d | \|x\|_p \le 1\}$.

## 2. Preliminaries

This section is devoted to setting up the adversarial robust classification problem that is considered in this paper. Along the way, we introduce necessary background and state several preliminary results for future comparisons.

**Conditional Gaussian Model** We consider the binary classification problem with data pair $(x,y)$ generated from the mixture of two Gaussian distributions $P_{\mu,\Sigma}$,

$$p(y=1)=\frac{1}{2},\quad p(y=-1)=\frac{1}{2},$$
$$p(x|y)=\mathcal{N}(x;y\mu,\Sigma).$$

Here $\mu\in\mathbb{R}^d$, $\Sigma\in\mathbb{R}^{d\times d}, \Sigma\succeq 0$ denote the mean and covariance of the Gaussian distribution. Given $n$ training

samples $(x_i, y_i) \sim_{i.i.d.} P_{\mu,\Sigma}$ for $1 \le i \le n$, the goal is to learn a classifier $\hat{f}(x)$ for predicting the class of a future data point that is drawn from the same distribution $P_{\mu,\Sigma}$.

**Adversarially Robust Classification** In the standard setting of classification, the optimal classifier is defined as the one that which minimizes the population classification error

$$R^{\text{std}}_{\mu,\Sigma}(f) := \mathbb{E}_{(x,y)\sim P_{\mu,\Sigma}}\left[\mathbb{I}(f(x) \neq y)\right].$$

which we refer to the standard error throughout. In this paper, we consider the classification problem under conditional Gaussian generative model in presence of an adversary — which is to say — at the testing stage, an adversary is allowed to add any perturbation $\delta$ to the input $x$, that has bounded magnitude $\|\delta\|_B \le \varepsilon$. The norm defined here is the standard Minkowski functional that associated with a convex set (Thompson & Thompson, 1996). Formally, given a closed and origin-symmetric convex set $B$, the Minkowski functional is defined as

$$\|x\|_B := \inf\{\lambda \in \mathbb{R}_{>0} : x \in \lambda B\}.$$

For instance, when $B$ is the $\ell_p$ unit ball, then $\|x\|_B$ boils down to the classical $\ell_p$ norm of $x$. In practice, the most widely considered norm for the adversary are $\ell_\infty$ and $\ell_2$ norms.

In the adversarially robust setting, a mapping $f : \mathbb{R}^d \to \{-1, +1\}$ classifies a sample $(x, y)$ correctly, if and only if the prediction agrees with the true label for *all* possible perturbations of the adversary. To put it in mathematical form,

$$\ell_{B,\varepsilon}(f; x, y) := \mathbb{I}\left(\exists \delta : \|\delta\|_B \le \varepsilon,\ f(x+\delta) \neq y\right).$$

Our goal is to obtain a classifier with minimal expected robust classification error, i.e. finding mapping $f$ that minimizes

$$\begin{aligned} R^{B,\varepsilon}_{\mu,\Sigma}(f) &= \mathbb{E}_{(x,y)\sim P_{\mu,\Sigma}}[\ell_{B,\varepsilon}(f; x, y)] \\ &= \mathbb{E}_{(x,y)\sim P_{\mu,\Sigma}}[\mathbb{I}\left(\exists \|\delta\|_B \le \varepsilon, f(x+\delta) \neq y\right)]. \end{aligned} \tag{1}$$

The optimal risk is then defined as the classification error regarding the optimal classifier, namely

$$R^{B,\varepsilon}_{\mu,\Sigma}* := R^{B,\varepsilon}_{\mu,\Sigma}(f_*), \tag{2}$$

and accordingly, we define the excess risk of any classifier $f$ as

$$R^{B,\varepsilon}_{\mu,\Sigma}(f) - R^{B,\varepsilon}_{\mu,\Sigma}*, \tag{3}$$

which by definition is always non-negative.

**Robust Bayes Optimal Classifier** To motivate the robust optimal classifiers, we start our discussion with the optimal risk and optimal classifier in the conditional Gaussian Model. We note that when $\varepsilon = 0$, i.e. there is no adversary, the classification problem reduces to the well-known *Fisher's Linear Discriminant Analysis* problem, where the Bayes optimal classifier is a simple linear classifier

$$f_{\text{Bayes}}(x) = \text{sign}(\mu^T x),$$

known as Fisher's linear discriminant rule (see, e.g. Johnson et al. (2002)). The Bayes optimal classifier minimizes the misclassification rate. However, the classifier that minimizes the *robust* classification error is not known until recently, where (Bhagoji et al., 2019) provided a tight lower bound on the minimal robust classification error via optimal transport techniques. It is also proved that the optimal risk can be written as the optimal value of a convex program, and the *oracle* optimal classifier is a linear classifier that has a closed form given the solution of the convex program.

We find it is useful to first simplify and restate this result in order to set the stage for our main result.

**Theorem 2.1** (Restated and simplified from Bhagoji et al. (2019))**.** *Let $z_\Sigma(\mu)$ be the solution of the following convex program:*

$$z_\Sigma(\mu) = \underset{\|z\|_B \le \varepsilon}{\text{argmin}} \|\mu - z\|^2_{\Sigma^{-1}}, \tag{4}$$

*where $\|x\|_A = \sqrt{x^T A x}$.* [1]*Then, the optimal robust classifier for $P_{\mu,\Sigma}$ is a linear classifier $f_*(x) = \text{sign}(w_0^T x)$, where*

$$w_0 := \Sigma^{-1}(\mu - z_\Sigma(\mu)), \tag{5}$$

*and the optimal robust classification error is*

$$R^{B,\varepsilon}_{\mu,\Sigma}* := \bar{\Phi}(\|w_0\|_\Sigma) = \bar{\Phi}(\|\mu - z_\Sigma(\mu)\|_{\Sigma^{-1}}).$$

We remark that the above mentioned classifier is indeed an oracle classifier since it is constructed using the unknown parameters $\mu$ and $\Sigma$.

**Adversarial Signal-To-Noise Ratio (AdvSNR).** In the context of standard classification in the conditional Gaussian model, the notion of Signal-To-Noise Ratio was introduced to measure the effective separation which is defined as the Mahalanobis distance between the means of two conditional distributions.

**Definition 2.1** (Standard Signal-To-Noise Ratio)**.** *The Standard Signal-To-Noise Ratio (StdSNR) of conditional Gaussian model $P_{\mu,\Sigma}$ is defined as*

$$\text{StdSNR}(\mu, \Sigma) := 2\|\mu\|_{\Sigma^{-1}}.$$

[1] Note that this notation is different with (Bhagoji et al., 2019), where in their notation $\|x\|_A = \sqrt{x^T A^{-1} x}$.

Here, the constant 2 is introduced to be consistent with the literature in Fisher's LDA, e.g. (Cai & Zhang, 2019), where SNR is defined as the Mahalanobis distance between means of two mixture components. We make the note that the StdSNR measures the difficulty of standard classification in the conditional Gaussian model, since the minimal misclassification error equals to $\bar{\Phi}(\frac{1}{2}\mathrm{StdSNR}(\mu,\Sigma))$ (Cai & Zhang, 2019). In fact, the misclassification error decreases exponentially as the StdSNR increases.

When it comes to the adversarial setting, StdSNR, however, is no longer a proper metric for the classification difficulty. Specifically, conditional Gaussian models with the same StdSNR can have very different levels of hardness in the adversarially robust classification problem. In order to illustrate this, we demonstrate a simple example.

**Example 2.1.** *Consider an adversary which is allowed to perturb the input with budget $\varepsilon = \frac{6}{\sqrt{d}}$ in terms the $\ell_\infty$ norm. Set the covariance $\Sigma$ to be the identity matrix $I_d$. We examine two conditional Gaussian models, $P_{\mu_1,\Sigma}$ and $P_{\mu_2,\Sigma}$ with different means $\mu_1$ and $\mu_2$, where*

$$\mu_1 = \frac{6}{\sqrt{d}} \cdot (1,1,1,\cdots,1)^T, \quad \mu_2 = (6,0,0,\cdots,0)^T.$$

It is easily seen that $\|\mu_1\|_{\Sigma^{-1}} = \|\mu_2\|_{\Sigma^{-1}} = 6$, therefore $P_{\mu_1,\Sigma}$ and $P_{\mu_2,\Sigma}$ have the same StdSNR. However, by Theorem 2.1, these two distributions actually exhibit completely different minimal robust classification error, indeed,

$$R^{B,\varepsilon}_{\mu_1,\Sigma} = \bar{\Phi}(0) = \frac{1}{2}, \quad R^{B,\varepsilon}_{\mu_2,\Sigma} = \bar{\Phi}(6 - \frac{6}{\sqrt{d}}).$$

When the dimension $d$ is sufficiently large, the optimal risk $R^{B,\varepsilon}_{\mu_2,\Sigma}$ approaches $\bar{\Phi}(6) \approx 10^{-8}$, which means there exists a very good robust classifier for $P_{\mu_2,\Sigma}$. In contrast, the optimal risk $R^{B,\varepsilon}_{\mu_1,\Sigma} = \frac{1}{2}$, i.e. no classifier can achieve a robust accuracy better than a uninformative predictor that classifies everything as the same class. From this simple example, it is safe to conclude that StdSNR is not an ideal measurement for the difficulty in the adversarially robust classification problem.

To address the above issue, one need a proper definition of the signal-to-noise-ratio that is suitable for the adversarial robust setting. Therefore we introduce the Adversarial Signal-To-Noise Ratio (AdvSNR) for any $(B,\varepsilon)$ adversary.

**Definition 2.2** (Adversarial Signal-To-Noise Ratio)**.** *Define the $(B,\varepsilon)$ Adversarial Signal-To-Noise Ratio (AdvSNR) of conditional Gaussian model $P_{\mu,\Sigma}$ as*

$$\mathrm{AdvSNR}_{B,\varepsilon}(\mu,\Sigma) := 2\|\mu - z_\Sigma(\mu)\|_{\Sigma^{-1}} = 2\|w_0\|_\Sigma,$$

*where $w_0$ is defined in (5).*

As a consequence of Theorem 2.1, the minimal robust classification error satisfies

$$R^{B,\varepsilon}_{\mu,\Sigma}* = \bar{\Phi}\left(\frac{1}{2}\mathrm{AdvSNR}(\mu,\Sigma)\right). \tag{6}$$

Consequently, the AdvSNR fully characterizes the difficulty for the adversarially robust setting as the StdSNR in the standard setting. We also note that when $\varepsilon = 0$, i.e. there is no adversary, the AdvSNR reduces to the traditional definition of the StdSNR. Thus, AdvSNR is a reasonable generalization for StdSNR.

Naturally, for every $r > 0$, one can consider a class of distributions where each of them has the same $(B,\varepsilon)$-AdvSNR equal to $r$. Within each class, they should enjoy the same hardness of the classification problem. Formally, let us define the class $D_{B,\varepsilon}(r)$.

**Definition 2.3.** *The family of conditional Gaussian models with $(B,\varepsilon)$-AdvSNR value of $r$, is defined as:*

$$D_{B,\varepsilon}(r) := \{(\mu,\Sigma) | \mathrm{AdvSNR}_{B,\varepsilon}(\mu,\Sigma) = r\}.$$

In the sequel, we develop our minimax lower bounds over these classes of distributions. To assist our analysis, we also define the family of conditional Gaussian models with a standard SNR value of $r$ similarly.

**Definition 2.4.** *The family of conditional Gaussian models with a standard SNR value of $r$, is defined as:*

$$D_{\mathrm{std}}(r) := \{(\mu,\Sigma) | \mathrm{StdSNR}(\mu,\Sigma) = r\}.$$

In the derivations of our upper bounds and minimax lower bounds, we make the assumption that the AdvSNR $r$ is strictly bounded away from zero by a universal constant [2], otherwise as a result of Theorem 2.1, no classifier can achieve accuracy much better than $\frac{1}{2}$, the robust risk of a constant classifier $f(x) \equiv 1$.

## 3. A Coputationally Efficient Estimator and Risk Upper Bound

Thus far, we introduce the notion of $\mathrm{AdvSNR}$ which is known to characterize the minimal robust classification error as in expression (6). However, whether there exists a computation-efficient classifier that behaves similarly to the oracle best classifier is still unclear.

This section, we aim to answer this question in the affirmative by constructing such a classifier. For the classifier that we shall define in the sequel, we give an exact characterization of its excess robust classification error compared with the oracle best classifier. Motivated by the fact that the

[2] for instance, $r \geq 10^{-9}$

optimal robust classifier has the form of (5), we design a "plug-in" estimator for $w_0$. The estimator is described in the following algorithm.

**Algorithm 1** A plug-in estimator of $w_0$

**Input:** Data pairs $\{(x_i, y_i)\}_{i=1}^n$.
**Output:** $\widehat{w}$.
**Step 1:** Define $\widehat{\mu}$ and $\widehat{\Sigma}$ as

$$\widehat{\mu} := \frac{1}{n}\sum_{i=1}^n y_i x_i, \qquad \widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^n x_i x_i^T - \widehat{\mu}\widehat{\mu}^T.$$

**Step 2:** Solve for $\widehat{z}$ in the following

$$\widehat{z} := z_{\widehat{\Sigma}}(\widehat{\mu}) = \operatorname*{argmin}_{\|z\|_B \le \varepsilon} \|\widehat{\mu} - z\|_{\widehat{\Sigma}^{-1}}^2.$$

**Step 3:** Define $\widehat{w} := \widehat{\Sigma}^{-1}(\widehat{\mu} - \widehat{z})$.

The main theorem of this section is to characterize the excess risk bound of the classifier induced by $\widehat{w}$.

**Theorem 3.1.** *For the $(\|\cdot\|_B, \varepsilon)$ adversary, suppose the adversarial signal-to-noise ratio $\mathrm{AdvSNR}_{B,\varepsilon}(\mu, \Sigma) = r$, then the excess risk of $f_{\widehat{w}}$ is upper bounded by*

$$R^{B,\varepsilon}_{\mu,\Sigma}(f_{\widehat{w}}) - R^{B,\varepsilon}_{\mu,\Sigma}* \le O_P\Big(e^{-\frac{1}{8}r^2} \cdot r \cdot \frac{d}{n}\Big).$$

We take a moment to make several remarks. First recall that the $\mathrm{AdvSNR}$ is defined as a measurement for the hardness of the classification problem. Indeed, as the above result shows, the excess risk vanishes exponentially with the $\mathrm{AdvSNR}$. Moreover, our estimator is *adaptive* in the sense that it does not require knowing any information about the value of $r$, but the theoretical guarantee improves automatically with larger AdvSNRs. We also note that the dependency with sample size $n$ is $O\left(\frac{1}{n}\right)$, which is the same as the rate of Fisher's LDA, but faster than the typical $O\left(\frac{1}{\sqrt{n}}\right)$ rate.

**Comparisons to (Schmidt et al., 2018)** We note that our result generalizes the one showed in (Schmidt et al., 2018) in many different aspects:

1. In terms of the perturbations, Schmidt et al. (2018) considered perturbations in $\ell_\infty$ balls, while ours allow for any convex, closed and origin-symmentric perturbaion set $B$, including all $\ell_p$ balls for $p \ge 1$.

2. Our upper and lower bounds hold for both spherical and non-spherical Gaussians, without the knowledge of the population covariance structure.

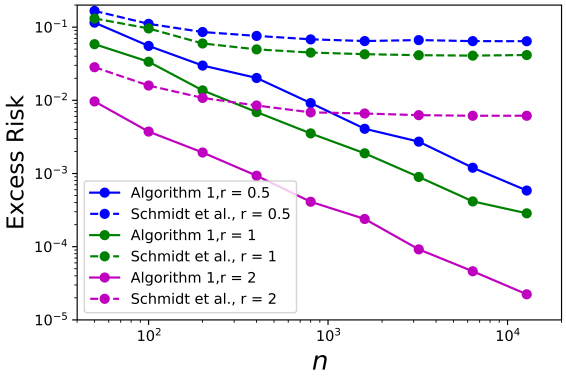


*Figure 1.* A simple simulation on the performance of Algorithm 1 and the algorithm proposed in (Schmidt et al., 2018) is shown here with different values of AdvSNR $r$. Here we consider a 50-dimensional example under $\ell_\infty$ adversary with $\varepsilon = 0.1$. The covariance matrix is fixed to be $\Sigma = I$, and the mean parameter $\mu$ is set as $\mu = (r + \varepsilon, \varepsilon, \varepsilon, \cdots, \varepsilon)$ for $r \in \{0.5, 1.0, 2.0\}$. We evaluate the excess risk $R^{B,\varepsilon}_{\mu,\Sigma}(f_{\widehat{w}}) - R^{B,\varepsilon}_{\mu,\Sigma}*$ returned by the two algorithms using $n$ i.i.d. training data pairs, where $n \in \{100, 200, 400, 800, 1600, 3200, 6400, 12800\}$. For each combination of $(n, r)$, the averaged excess risk over 10 random repetitions is reported respectively.

3. We impose no restrictions on the separation between Gaussian distributions. Schmidt et al. (2018) studied a very specific regime, where the budget of $\ell_\infty$ adversary is bounded by $\frac{1}{4}$ , the separation between the means of two Gaussians is $\sqrt{d}$, and the spherical covariance matrix $\Sigma = \sigma^2 I$ satisfies $\sigma \le \frac{1}{32} d^{1/4}$. This regime is low-noise by design, while our analysis applies to any regime whenever there exists a classifier with robust accuracy slightly better than $\frac{1}{2}$ .

4. Our estimator is consistent, i.e. the excess risk converges to zero as sample size $n \to \infty$. The classifier used in Schmidt et al. (2018) is actually $\mathrm{sign}(\hat{\mu}^T x)$. While this classifier achieve near-optimal classification error in the regime of their interst (the low noise regime mentioned above with Gaussian prior on $\mu$), the excess risk does not converge to zero in general. This is due to the fact that the large-sample limit of their classifier is actually $\mathrm{sign}(\mu^T x)$, i.e. the Bayes optimal classifier for the standard setting. As we can see from Theorem 2.1 and a simple simulation in Figure 1, the excess risk of their algorithm saturates at a level above zero, which is very different from the behavior of Algorithm 1.

**Proof Sketch:** Here we provide a brief sketch of the proof. More details can be found in the Section 6.

**Step 1: First order approximation of the risk.** Since both the learned $f_{\widehat{w}}$ and the optimal robust classifier $f_*$ are linear classifiers, we can calculate the robust excess risk in closed form using Lemma 6.2 (also shown in (Bhagoji et al., 2019)):

$$R^{B,\varepsilon}_{\mu,\Sigma}(f_{\widehat{w}}) - R^{B,\varepsilon}_{\mu,\Sigma}* = \bar{\Phi}\left(\frac{\widehat{w}^T\mu - \varepsilon\|\widehat{w}\|_{B*}}{\|\widehat{w}\|_\Sigma}\right) - \bar{\Phi}(\frac{1}{2}r).$$

By the Taylor expansion of $\bar{\Phi}(\cdot)$, we have

$$\bar{\Phi}\left(\frac{\widehat{w}^T\mu - \varepsilon\|\widehat{w}\|_{B*}}{\|\widehat{w}\|_\Sigma}\right) - \bar{\Phi}(\frac{1}{2}r) \approx \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{8}r^2}\delta_n,$$

where

$$\delta_n = \frac{1}{2}r - \frac{\widehat{w}^T\mu - \varepsilon\|\widehat{w}\|_{B*}}{\|\widehat{w}\|_\Sigma}.$$

Therefore, it is sufficient to show that $\delta_n = O_P(r \cdot \frac{d}{n})$.

**Step 2: Controlling $\delta_n$.** To give an upper bound of $\delta_n$, we will use the fact that sample mean $\widehat{\mu}$ and sample covariace $\widehat{\Sigma}$ converge to $\mu$ and $\Sigma$ respectively. Furthermore, the convergence rate is well known as $O_P(\sqrt{\frac{d}{n}})$.

From a high level, the upper bound of $\delta_n$ is estabished (see Lemma 6.3) by carefully decomposing $\delta_n$ into four terms and each term is in the form of the differences between population and sample quantities like $\Sigma$ vs $\widehat{\Sigma}$, $\widehat{\mu}$ vs $\mu$. Invoking the convergence rates of $\widehat{\mu}$ and $\widehat{\Sigma}$, we are able to bound each of these terms and complete the proof.

## 4. Minimax Lower Bounds

This section is dedicated to developing minimax excess risk lower bounds for the adversarially robust classification with conditional Gaussian models.

As is mentioned above, we consider a class of distributions $D_{B,\varepsilon}(r)$ which have the same $\mathrm{AdvSNR}_{B,\varepsilon} = r$, as in Definition 2.3. As quantity AdvSNR characterizes the minimal robust classification error, this class of distributions $D_{B,\varepsilon}(r)$ all share the same adversarially robust classification error. Therefore, our lower bounds here measure the fundamental information-theoretic limit of this problem, namely, no estimator can achieve an essential improvement in terms of the adversarial classification error.

**Theorem 4.1.** *Let $\widehat{f}$ be any estimator based on $n$ samples $(x_1, y_1), \cdots, (x_n, y_n) \sim_{i.i.d.} P_{\mu,\Sigma}$. We have the following lower bound on the minimax excess risk:*

$$\min_{\widehat{f}} \max_{(\mu,\Sigma)\in D_{B,\varepsilon}(r)} [R^{B,\varepsilon}_{\mu,\Sigma}(\widehat{f}) - R^{B,\varepsilon}_{\mu,\Sigma}*] \geq \Omega_P\left(e^{-(\frac{1}{8}+o(1))r^2}\frac{d}{n}\right).$$

Putting together with the upper bound in Theorem 3.1, this lower bound matches almost exactly with the upper bound in the regime of interest, therefore they are both optimal up to lower order terms.

The main technique we used for this lower bound is with a flavor of black-box reduction. In particular, we show that the minimax *robust* excess risk in $D_{B,\varepsilon}(r)$ cannot be smaller than the minimax *standard* excess risk in $D_{\mathrm{std}}(r)$. In other words,

**Lemma 4.1.** *The minimax excess error satisfies*

$$\begin{aligned}&\min_{\widehat{f}} \max_{(\mu,\Sigma)\in D_{B,\varepsilon}(r)} [R^{B,\varepsilon}_{\mu,\Sigma}(\widehat{f}) - R^{B,\varepsilon}_{\mu,\Sigma}*] \\ \geq\ &\min_{\widehat{f}} \max_{(\mu',\Sigma)\in D_{\mathrm{std}}(r)} [R^{\mathrm{std}}_{\mu',\Sigma}(\widehat{f}) - R^{\mathrm{std}}_{\mu',\Sigma}*]. \qquad (7)\end{aligned}$$

The right hand side of (7), i.e. the minimax rate for standard classification, is well-studied in the existing literature of Fisher's LDA. For example, (Li et al., 2017) proved the following lower bound:

**Theorem 4.2** (Theorem 1 of (Li et al., 2017))**.** *Suppose the covariance matrix satisfies $\Sigma = I$ and is known to the learner, then we have the minimax lower bound*

$$\begin{aligned}&\min_{\widehat{f}} \max_{(\mu',I)\in D_{\mathrm{std}}(r)} [R^{\mathrm{std}}_{\mu',\Sigma}(\widehat{f}) - R^{\mathrm{std}}_{\mu',\Sigma}*] \\ \geq\ &\Omega_P\left(e^{-\frac{1}{8}r^2}\cdot\frac{1}{r}\cdot\frac{d}{n}\right).\end{aligned}$$

Since the parameter space considered in (Li et al., 2017) is a subset of $D_{\mathrm{std}}(r)$, we have (7) is also lower bounded by $\Omega_P\left(e^{-\frac{1}{8}r^2}\cdot\frac{1}{r}\cdot\frac{d}{n}\right)$, therefore proves Theorem 4.1.

**Comparisons to (Schmidt et al., 2018) and (Bhagoji et al., 2019)** To the best of our knowlege, Theorem 4.1 is the first minimax-type lower bound in adversarially robust classification. Existing works (Schmidt et al., 2018) and (Bhagoji et al., 2019) also studied the sample complexity of robust learning in conditional Gaussian model. However, both of them simplified the problem and considered the case when $\mu$ follows from a prior distribution $\mathcal{N}(0, I)$. This assumption is crucial to their analysis, otherwise the posterior distribution of $\mu$ given training data is intractable. Hence, the technical tool used in prior works is not sufficient for developing such a minimax lower bound of our interest.

**Proof Sketch:** Here we also provide a proof sketch to Lemma 4.1. More details can be found in the Section 6.

**Step 1: Connecting standard and robust risks** In Lemma 6.4, we prove that for any classifier $f$ and a perturbed distribution $P_{\mu',\Sigma}$, where $\|\mu' - \mu\|_B \leq \varepsilon$, the robust risk of $f$ on $P_{\mu,\Sigma}$ is always lower bounded by the standard risk on $P_{\mu',\Sigma}$.

As a consequence, in Corollary 6.1 we show that if we choose $\mu' = \mu - z_\Sigma(\mu)$, then the robust excess risk of $f$ on

$P_{\mu,\Sigma}$ is always lower bounded by the standard excess risk on $P_{\mu',\Sigma}$.

**Step 2: A mapping fron $D_{\mathrm{std}}(r)$ to $D_{B,\varepsilon}(r)$** To prove Lemma 4.1, we only need to answer the following question: for any $(\mu',\Sigma) \in D_{\mathrm{std}}(r)$, can we find a $(\mu,\Sigma) \in D_{B,\varepsilon}(r)$, so that the robust excess risk on $P_{\mu,\Sigma}$ is always lower bounded by the standard excess risk on $P_{\mu',\Sigma}$? We give an affirmative answer to this question. The proof in a combination of Corollary 6.1 showed in Step 1 and an examination of optimality condition in the optimization problem 4.

## 5. Comparing Adversarial and Standard Rates

Putting the upper and lower bounds together provides a comprehensive view of the statistical aspect of the adversarially robust classification. A key question to ask is that: How much does the classification error blows up as the price of being adversarially robust?

To answer this question, it is sufficient to compare the optimal risks in both cases. Informally, one can write the logarithm ratio between two rates as

$$\log\left(\frac{\mathrm{AdvRate}}{\mathrm{StdRate}}\right) \approx \frac{1}{2}\left(\|\mu - z_\Sigma(\mu)\|_{\Sigma^{-1}}^2 - \|\mu\|_{\Sigma^{-1}}^2\right). \tag{8}$$

From the definition of $z_\Sigma(\mu)$ in (4), we can see that $\|\mu - z_\Sigma(\mu)\|_{\Sigma^{-1}}^2 \le \|\mu\|_{\Sigma^{-1}}^2$, hence adversarial rate is always slower.

To analyze this difference quantitively and interpretably, we consider the special case where $\Sigma = I$ and the adversary is $\ell_2$ bounded. Similar results hold for other adversaries as well. The key observation is that depending on the different scale of $\|\mu\|_2$ and the budget of perturbation $\varepsilon$, this difference can be as small as $O(1)$, or as large as $\Omega(\exp(d))$.

**Proposition 5.1.** *When $\Sigma = I$ and the adversarial perturbation satisfies $\|\delta\|_2 \le \varepsilon$, then*

- *When $\varepsilon \le O(\frac{1}{\|\mu\|_2})$, the adversarial rate is at most $O(1)$ times slower than the standard rate.*
- *When $\|\mu\|_2 \ge \Omega(\log d)$ and $\varepsilon \ge \Omega(\frac{\log d}{\|\mu\|_2})$, the adversarial rate can be slower than the standard rate by a $\mathrm{poly}(d)$ factor.*
- *When $\|\mu\|_2 \ge \Omega(\sqrt{d})$ and $\varepsilon \ge \Omega(\frac{d}{\|\mu\|_2})$, the adversarial rate can be slower than the standard rate by an $\exp(d)$ factor.*

In general, the difference is more significant when $\varepsilon$ or $\|\mu\|_2$ is larger. This example demonstrates a clear tradeoff between being adversarial robust and obtaining the optimal accuracy, in particular in the case of large perturbations.

## 6. Proofs and further details

In this section, we provide detailed proofs for our main results. The proof details of some lemmas are deferred to our supplementary file.

### 6.1. Proof of Theorem 3.1

Before presenting our analysis, we first state a standard lemma about the convergence of empirical mean and covariance.

**Lemma 6.1** (Convergence of the empirical mean and covariance (see, e.g. Wainwright (2019)))**.** *The convergence rates of the empirical mean $\widehat{\mu}$ and $\widehat{\Sigma}$ to the corresponding ground truth satisfy*

$$\|\widehat{\mu} - \mu\|_{\Sigma^{-1}} = O_P\left(\sqrt{\frac{d}{n}}\right),$$

*and*

$$\|\Sigma^{-\frac{1}{2}}\widehat{\Sigma}\Sigma^{-\frac{1}{2}} - I\|_{op} = O_P\left(\sqrt{\frac{d}{n}}\right).$$

The following lemma about the classification error of linear classifiers will also be useful for us.

**Lemma 6.2** (Robust classification error of linear classifier, (see e.g. in (Bhagoji et al., 2019), Appendix B.3))**.** *For a linear classifier $f_w(x) = \mathrm{sign}(w^T x)$, the robust classification error with a $B,\varepsilon$ adversary is*

$$R_{\mu,\Sigma}^{B,\varepsilon}(f_w) = \bar{\Phi}\left(\frac{w^T\mu - \varepsilon\|w\|_{B*}}{\|w\|_\Sigma}\right).$$

*Here, $\|\cdot\|_{B*}$ is the dual norm of $\|\cdot\|_B$. We use $R_{\mu,\Sigma}^{B,\varepsilon}(w)$ as a shorthand for $R_{\mu,\Sigma}^{B,\varepsilon}(f_w)$ when the meaning is clear from context.*

*Proof of Theorem 3.1.* By Lemma 6.2 and Taylor expansion of $\bar{\Phi}(t)$ around $t = \frac{1}{2}r = \|w_0\|_\Sigma$, the excess risk can be written as:

$$\begin{aligned} R_{\mu,\Sigma}^{B,\varepsilon}(\widehat{w}) - R_{\mu,\Sigma}^{B,\varepsilon}* &= \bar{\Phi}\left(\frac{\widehat{w}^T\mu - \varepsilon\|\widehat{w}\|_{B*}}{\|\widehat{w}\|_\Sigma}\right) - \bar{\Phi}(\|w_0\|_\Sigma) \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{8}r^2}\delta_n + O(\delta_n^2), \end{aligned}$$

where

$$\delta_n = \|w_0\|_\Sigma - \frac{\widehat{w}^T\mu - \varepsilon\|\widehat{w}\|_{B*}}{\|\widehat{w}\|_\Sigma}.$$

Therefore, to analyze the convergence rate of the excess risk, we only need to analyze the convergence rate of $\delta_n$. We would like to prove that

$$\delta_n = O_P\left(r \cdot \frac{d}{n}\right).$$

The following lemma is the key of our analysis: it decomposes $\delta_n$ into four terms, each in the form of the difference between population and sample quantities like $\Sigma$ vs $\widehat{\Sigma}$, $\widehat{\mu}$ vs $\mu$.

**Lemma 6.3.** *We have the following decomposition for $\delta_n$:*

$$\|\widehat{w}\|_\Sigma \delta_n = \underbrace{-\frac{1}{2}\left(\|w_0\|_\Sigma - \|\widehat{w}\|_\Sigma\right)^2}_{T_1} \underbrace{+w_0^T(\widehat{z} - z_\Sigma(\mu))}_{T_2}$$
$$\underbrace{-\frac{1}{2}\|\widehat{z} - z_\Sigma(\mu)\|_{\Sigma^{-1}}^2}_{T_3} \underbrace{+\frac{1}{2}\|(\Sigma - \widehat{\Sigma})\widehat{w} + (\widehat{\mu} - \mu)\|_{\Sigma^{-1}}^2}_{T_4}.$$

*where $\hat{z}$ is the shorthand for $\hat{z} = z_{\widehat{\Sigma}}(\widehat{\mu})$.*

The proof of Lemma 6.3 is provided in Appendix D.1. Based on this decomposition, our goal is to establish the following relations.

$$T_1 \le 0,\ T_2 \le 0,\ T_3 \le 0,\ T_4 \le O_P\left(r^2 \frac{d}{n}\right).$$

It is obvious that $T_1 \le 0, T_3 \le 0$. For the second term $T_2$, consider $\phi(z) = \|\mu - z\|_{\Sigma^{-1}}^2$. Since $z_\Sigma(\mu) = \operatorname{argmin}_{\|z\|_B \le \varepsilon} \|\mu - z\|_{\Sigma^{-1}}^2 = \operatorname{argmin}_{\|z\|_B \le \varepsilon} \phi(z)$ , by the first order optimality condition, we have $(z' - z_\Sigma(\mu))^T \nabla\phi(z_\Sigma(\mu)) \le 0$ holds for any $\|z'\|_B \le \varepsilon$. Choosing $z' = \hat{z}$ gives:

$$(\mu - z_\Sigma(\mu))^T \Sigma^{-1}(\widehat{z} - z_\Sigma(\mu)) \le 0 \Leftrightarrow w_0^T(\widehat{z} - z_\Sigma(\mu)) \le 0.$$

Therefore, $T_2 \le 0$ as we desired.

The remaining work is to prove that $T_4 \le O_P\left((1+r)^2 \frac{d}{n}\right)$. By triangle's inequality,

$$\|(\Sigma - \widehat{\Sigma})\widehat{w} + (\widehat{\mu} - \mu)\|_{\Sigma^{-1}} \le \|(\Sigma - \widehat{\Sigma})\widehat{w}\|_{\Sigma^{-1}} + \|\widehat{\mu} - \mu\|_{\Sigma^{-1}}.$$

Both terms can be controled using covergence of sample mean and covariance. By Lemma 6.1, one has

$$\|\widehat{\mu} - \mu\|_{\Sigma^{-1}} \le O_P\left(\sqrt{\frac{d}{n}}\right),$$

and direct calculations give

$$\begin{aligned}\|(\Sigma - \widehat{\Sigma})\widehat{w}\|_{\Sigma^{-1}} &= \|(I - \Sigma^{-\frac{1}{2}}\widehat{\Sigma}\Sigma^{-\frac{1}{2}})(\Sigma^{\frac{1}{2}}\widehat{w})\|_2 \\ &\le \|I - \Sigma^{-\frac{1}{2}}\widehat{\Sigma}\Sigma^{-\frac{1}{2}}\|_{op}\|\Sigma^{\frac{1}{2}}\widehat{w}\|_2 \\ &= O_P\left(\sqrt{\frac{d}{n}}\right)\|\widehat{w}\|_\Sigma.\end{aligned}$$

Combined pieces together, triangle's inequality further guarantees that

$$\|(\Sigma - \widehat{\Sigma})\widehat{w} + (\widehat{\mu} - \mu)\|_{\Sigma^{-1}} \le O_P\left(\sqrt{\frac{d}{n}}(\|\widehat{w}\|_\Sigma + 1)\right).$$

Since $\widehat{\mu} \to \mu, \widehat{\Sigma} \to \Sigma$, we have $\widehat{w} \to w_0$, therefore $\|\widehat{w}\|_\Sigma = (1 + o(1))\|w_0\|_\Sigma = (\frac{1}{2} + o(1))r$, hence,

$$\begin{aligned}T_4 &= \frac{1}{2}\|(\Sigma - \widehat{\Sigma})\widehat{w} + (\widehat{\mu} - \mu)\|_{\Sigma^{-1}}^2 \\ &\le \frac{1}{2}\left(O_P(\sqrt{\frac{d}{n}})(\|\widehat{w}\|_\Sigma + 1)\right)^2 \\ &= O_P\left(r^2 \cdot \frac{d}{n}\right).\end{aligned}$$

Putting things together and recall that $r = \Omega(1)$, we have

$$\delta_n = O_P\left(r \cdot \frac{d}{n}\right).$$

Therefore we have completed the proof. □

### 6.2. Proof of Lemma 4.1

To prove Lemma 4.1, we start with a simple observation: for any classifier $f$, its standard error on any perturbed distribution $P_{\mu',\Sigma}$ is always a lower bound on robust error of the original distribution $P_{\mu,\Sigma}$, as long as the perturbation has bounded $B$-norm $\|\mu' - \mu\|_B \le \varepsilon$:

**Lemma 6.4.** *For any classifier $f : \mathbb{R}^d \to \{-1, +1\}$ and any $\mu' \in \mathbb{R}^d, \|\mu' - \mu\|_B \le \varepsilon$*

$$R_{\mu,\Sigma}^{B,\varepsilon}(f) \ge R_{\mu',\Sigma}^{\text{std}}(f).$$

*Proof.* By the definition of robust classification error (1), we can decompose the error into two parts: the error on positive class ($y = 1$) and negative class ($y = -1$), namely,

$$\begin{aligned}R_{\mu,\Sigma}^{B,\varepsilon}(f) =& \mathbb{E}_{(x,y)\sim P_{\mu,\Sigma}}[\mathbb{I}\left(\exists\|\delta\|_B \le \varepsilon, f(x+\delta) \ne y\right)] \\ =& \frac{1}{2}\mathbb{E}_{x\sim N(\mu,\Sigma)}[\mathbb{I}\left(\exists\|\delta\|_B \le \varepsilon, f(x+\delta) \ne 1\right)] + \\ & \frac{1}{2}\mathbb{E}_{x\sim N(-\mu,\Sigma)}[\mathbb{I}\left(\exists\|\delta\|_B \le \varepsilon, f(x+\delta) \ne -1\right)].\end{aligned} \tag{9}$$

By choosing the adversarial perurbation as $\delta = \mu' - \mu$, we have the error on positive class is lower bounded by:

$$\begin{aligned}&\mathbb{E}_{x\sim N(\mu,\Sigma)}[\mathbb{I}\left(\exists\|\delta\|_B \le \varepsilon, f(x+\delta) \ne 1\right)] \\ \ge& \mathbb{E}_{x\sim N(\mu,\Sigma)}[\mathbb{I}\left(f(x - \mu + \mu') \ne 1\right)] \qquad (10) \\ =& \mathbb{E}_{x'\sim N(\mu',\Sigma)}[\mathbb{I}\left(f(x') \ne 1\right)]. \qquad (11)\end{aligned}$$

Similarly, by choosing $\delta = \mu - \mu'$, we have the error on negative class is lower bounded by:

$$\begin{aligned}&\mathbb{E}_{x\sim N(-\mu,\Sigma)}[\mathbb{I}\left(\exists\|\delta\|_B \le \varepsilon, f(x+\delta) \ne 1\right)] \\ \ge& \mathbb{E}_{x'\sim N(-\mu',\Sigma)}[\mathbb{I}\left(f(x') \ne -1\right)]. \qquad (12)\end{aligned}$$

Hence, combining (9) , (10) and (12), we get

$$\begin{aligned} R^{B,\varepsilon}_{\mu,\Sigma}(f) \geq & \frac{1}{2}\mathbb{E}_{x'\sim N(\mu',\Sigma)}[\mathbb{I}\left(f(x')\neq 1\right)]+ \\ & \frac{1}{2}\mathbb{E}_{x'\sim N(-\mu',\Sigma)}[\mathbb{I}\left(f(x')\neq -1\right)] \\ = & R^{\text{std}}_{\mu',\Sigma}(f), \end{aligned}$$

where the last step is by the definition of standard error (2). Therefore we have completed the proof. □

Next, we show more connections between robust and standard classification. Namely, the robust Bayes classifier of $P_{\mu,\Sigma}$ coincides with the standard Bayes classifier of $P_{\mu-z_\Sigma(\mu),\Sigma}$, as stated in the following Lemma:

**Lemma 6.5.** *Let $z_\Sigma(\mu)$ be the solution of (4), then the robust Bayes classifier of $P_{\mu,\Sigma}$, $f_*(x)=\text{sign}(w_0^T x)$, satisfies the following conditions:*

1. $R^{B,\varepsilon}_{\mu,\Sigma}(f_*) = R^{\text{std}}_{\mu-z_\Sigma(\mu),\Sigma}(f_*)$.
2. *$f_*$ is the standard Bayes Optimal Classifier of $P_{\mu-z_\Sigma(\mu),\Sigma}$.*

*Proof.* Note that by setting $\varepsilon = 0$ in Theorem 2.1, we get the characterization of the standard Bayes error and Bayes optimal classifier for conditional Gaussian models. Applying this result for the distribution $P_{\mu-z_\Sigma(\mu),\Sigma}$, we have

1. The standard Bayes Optimal Classifier of $P_{\mu-z_\Sigma(\mu),\Sigma}$ is $\text{sign}((\mu-z_\Sigma(\mu))^T\Sigma^{-1}x)$, which is exactly $f_*(x)$.
2. The standard Bayes error of of $P_{\mu-z_\Sigma(\mu),\Sigma}$ is $\bar{\Phi}(\sqrt{(\mu-z_\Sigma(\mu))^T\Sigma^{-1}(\mu-z_\Sigma(\mu))})$, which is exactly $R^{B,\varepsilon}_{\mu,\Sigma}*$.

Hence we have completed the proof. □

As a direct consequence of Lemma 6.4 and Lemma 6.5, we have the robust excess risk under $P_{\mu,\Sigma}$ is lower bounded by the standard excess risk under $P_{\mu-z_\Sigma(\mu),\Sigma}$:

**Corollary 6.1.** *For any classifier $f:\mathbb{R}^d\to\{-1,+1\}$,*

$$\begin{aligned} R^{B,\varepsilon}_{\mu,\Sigma}(f) - R^{B,\varepsilon}_{\mu,\Sigma}* &\geq R^{\text{std}}_{\mu-z_\Sigma(\mu),\Sigma}(f) - R^{\text{std}}_{\mu-z_\Sigma(\mu),\Sigma}(f_*) \\ &= R^{\text{std}}_{\mu-z_\Sigma(\mu),\Sigma}(f) - R^{\text{std}}_{\mu-z_\Sigma(\mu),\Sigma}*, \end{aligned}$$

*where*

$$R^{\text{std}}_{\mu',\Sigma}* = \inf_g R^{\text{std}}_{\mu',\Sigma}(g)$$

*is the optimal standard risk.*

The last piece of tool needed for proving Lemma 4.1 is a mapping from $D_{\text{std}}(r)$ to $D_{B,\varepsilon}(r)$ that keeps the excess risk non-decreasing. This is established via the following lemma:

**Lemma 6.6.** *For any $(\mu',\Sigma)\in D_{\text{std}}(r)$, there exists $(\mu,\Sigma)\in D_{B,\varepsilon}(r)$, such that $\mu - z_\Sigma(\mu)=\mu'$, here $z_\Sigma(\mu)$ is the optimal solution of (4).*

*Proof.* The proof is constructive: we choose $\mu = \mu' + \widetilde{z}_\Sigma(\mu')$, where $\widetilde{z}_\Sigma(\mu')$ is the maximizer of the following convex program (which is maximizing a linear function over a convex set):

$$\widetilde{z}_\Sigma(\mu') = \underset{\|z\|_B\leq\varepsilon}{\text{argmax}}\, \mu'^T\Sigma^{-1}z. \tag{13}$$

We want to prove that $\mu - z_\Sigma(\mu)=\mu'$. By our choice of $\mu$, we also have $\mu=\mu'+\widetilde{z}_\Sigma(\mu')$. Hence, we only need to prove that

$$\widetilde{z}_\Sigma(\mu') = z_\Sigma(\mu).$$

In other words, we only need to show that $\widetilde{z}_\Sigma(\mu')$ is the minimizer of (4).

Since (4) is a convex program with a strongly convex objective, it suffices to prove the following first order optimality condition holds for any $\forall\|z'\|_B\leq\varepsilon$:

$$(\mu-\widetilde{z}_\Sigma(\mu'))^T\Sigma^{-1}(z'-\widetilde{z}_\Sigma(\mu'))\leq 0.$$

Since $\mu-\widetilde{z}_\Sigma(\mu')=\mu'$, the inequality is equivalent to:

$$\mu'^T\Sigma^{-1}z' \leq \mu'^T\Sigma^{-1}\widetilde{z}_\Sigma(\mu'),$$

which is correct by the definition of $\widetilde{z}_\Sigma(\mu')$. Hence we have completed the proof. □

Equipped with Lemma 6.6, now we can prove the important lemma:

*Proof of Lemma 4.1.* By Lemma 6.6, for any $(\mu',\Sigma)\in D_{\text{std}}(r)$, there exists $(\mu,\Sigma)\in D_{B,\varepsilon}(r)$, such that $\mu - z_\Sigma(\mu)=\mu'$, where $z_\Sigma(\mu)$ is the optimal solution of (4). By Corollary 6.1, we have the following inequality holds for any fixed $\widehat{f}$:

$$R^{\text{std}}_{\mu',\Sigma}(\widehat{f}) - R^{\text{std}}_{\mu',\Sigma}* \leq R_{\mu,\Sigma}(\widehat{f}) - R^{B,\varepsilon}_{\mu,\Sigma}*\,.$$

Therefore,

$$R^{\text{std}}_{\mu',\Sigma}(\widehat{f}) - R^{\text{std}}_{\mu',\Sigma}* \leq \max_{(\mu,\Sigma)\in D_{B,\varepsilon}(r)}[R_{\mu,\Sigma}(\widehat{f}) - R^{B,\varepsilon}_{\mu,\Sigma}*].$$

holds for all $(\mu,\Sigma)\in D_{B,\varepsilon}(r)$, which means

$$\begin{aligned} &\max_{(\mu,\Sigma)\in D_{B,\varepsilon}(r)}[R^{B,\varepsilon}_{\mu,\Sigma}(\widehat{f}) - R^{B,\varepsilon}_{\mu,\Sigma}*] \\ \geq &\max_{(\mu',\Sigma)\in D_{\text{std}}(r)}[R^{\text{std}}_{\mu',\Sigma}(\widehat{f}) - R^{\text{std}}_{\mu',\Sigma}*]. \end{aligned}$$

Then, taking minimum over $\widehat{f}$ on both sides proves the theorem.

□

## Acknowledgements

Y.W. is supported in part by the NSF grant DMS-2015447 and CCF-2007911. C.D. and P.R. are supported by DARPA via HR00112020006, and NSF via IIS1909816.

The authors would also like to thank Kaizheng Wang for many helpful discussions, Tianle Cai and Justin Khim for pointing us toward the work of (Bhagoji et al., 2019; Cai & Zhang, 2019), and annonymous reviewer for many suggestions about improving the presentation of the paper.

## References

Attias, I., Kontorovich, A., and Mansour, Y. Improved generalization bounds for robust learning. *arXiv preprint arXiv:1810.02180*, 2018.

Awasthi, P., Dutta, A., and Vijayaraghavan, A. On robustness to adversarial examples and polynomial optimization. In *Advances in Neural Information Processing Systems*, pp. 13760–13770, 2019.

Awasthi, P., Frank, N., and Mohri, M. Adversarial learning guarantees for linear hypotheses and neural networks. *arXiv preprint arXiv:2004.13617*, 2020.

Azizyan, M., Singh, A., and Wasserman, L. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems*, pp. 2139–2147, 2013.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Bhagoji, A. N., Cullina, D., and Mittal, P. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems*, pp. 7496–7508. 2019.

Bubeck, S., Lee, Y. T., Price, E., and Razenshteyn, I. Adversarial examples from cryptographic pseudo-random generators. *arXiv preprint arXiv:1811.06418*, 2018a.

Bubeck, S., Price, E., and Razenshteyn, I. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018b.

Cai, T. and Zhang, L. High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):675–705, 2019.

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

Cullina, D., Bhagoji, A. N., and Mittal, P. Pac-learning in the presence of adversaries. In *Advances in Neural Information Processing Systems*, pp. 230–241, 2018.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

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

Johnson, R. A., Wichern, D. W., et al. *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ, 2002.

Khim, J. and Loh, P.-L. Adversarial risk bounds for binary classification via function transformation. *arXiv preprint arXiv:1810.09519*, 2, 2018.

Kim, S.-J., Magnani, A., and Boyd, S. Robust fisher discriminant analysis. In *Advances in neural information processing systems*, pp. 659–666, 2006.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Li, T., Prasad, A., and Ravikumar, P. K. Fast classification rates for high-dimensional gaussian generative models. In *Advances in Neural Information Processing Systems*, pp. 1054–1062, 2015.

Li, T., Yi, X., Carmanis, C., and Ravikumar, P. Minimax gaussian classification & clustering. In *Artificial Intelligence and Statistics*, pp. 1–9, 2017.

McLachlan, G. J. and Peel, D. *Finite mixture models*. John Wiley & Sons, 2004.

Montasser, O., Hanneke, S., and Srebro, N. Vc classes are adversarially robustly learnable, but only improperly. *arXiv preprint arXiv:1902.04217*, 2019.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pp. 372–387. IEEE, 2016.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J., and Liang, P. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*, 2020.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pp. 5019–5031, 2018.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

Stanforth, R., Fawzi, A., Kohli, P., et al. Are labels required for improving adversarial robustness? *arXiv preprint arXiv:1905.13725*, 2019.

Suggala, A. S., Prasad, A., Nagarajan, V., and Ravikumar, P. Revisiting adversarial risk. *arXiv preprint arXiv:1806.02924*, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Thompson, A. C. and Thompson, A. C. *Minkowski geometry*. Cambridge University Press, 1996.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, Y., Jha, S., and Chaudhuri, K. Analyzing the robustness of nearest neighbors to adversarial examples. *arXiv preprint arXiv:1706.03922*, 2017.

Weng, T.-W., Zhang, H., Chen, H., Song, Z., Hsieh, C.-J., Boning, D., Dhillon, I. S., and Daniel, L. Towards fast computation of certified robustness for relu networks. *arXiv preprint arXiv:1804.09699*, 2018.

Xu, H. and Mannor, S. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

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

Xu, H., Caramanis, C., and Mannor, S. Robustness and regularization of support vector machines. *Journal of machine learning research*, 10(7), 2009b.

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

Zhai, R., Cai, T., He, D., Dan, C., He, K., Hopcroft, J., and Wang, L. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.