2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834778

Binary Classification Under ℓ_0 Attacks for General Noise Distribution

Payam Delgosha[1], Hamed Hassani[2], and Ramtin Pedarsani[3]

***Abstract*—Adversarial examples have recently drawn considerable attention in the field of machine learning due to the fact that small perturbations in the data can result in major performance degradation. This phenomenon is usually modeled by a malicious adversary that can apply perturbations to the data in a constrained fashion, such as being bounded in a certain norm. In this paper, we study this problem when the adversary is constrained by the ℓ_0 norm; i.e., it can perturb a certain number of coordinates in the input, but has no limit on how much it can perturb those coordinates. Due to the combinatorial nature of this setting, we need to go beyond the standard techniques in robust machine learning to address this problem. We consider a binary classification scenario where d noisy data samples of the true label are provided to us after adversarial perturbations. We introduce a classification method which employs a nonlinear component called truncation, and show in an asymptotic scenario, as long as the adversary is restricted to perturb no more than $\sqrt{d}$ data samples, we can almost achieve the optimal classification error in the absence of the adversary, i.e. we can completely neutralize adversary's effect. Surprisingly, we observe a phase transition in the sense that using a converse argument, we show that if the adversary can perturb more than $\sqrt{d}$ coordinates, no classifier can do better than a random guess.**

I. Introduction

It is well-known that machine learning models are susceptible to adversarial attacks that can cause classification error. These attacks are typically in the form of a small norm-bounded perturbation to the input data that are carefully designed to incur misclassification – e.g. they can be form of an additive ℓ_p-bounded perturbation for some $p \geq 0$ [1], [2], [3], [4], [5].

There is an extensive body of prior work studying adversarial machine learning, most of which have focused on ℓ_2 and ℓ_∞ attacks [6], [7], [8], [9]. To train models that are more robust against such attacks, adversarial training is the state-of-the-art defense method. However, the success of the current adversarial training methods is mainly based on empirical evaluations [5]. It is therefore imperative to study the fundamental limits of robust machine learning under different classification settings and attack models.

In this paper, we focus on the important case of ℓ_0-bounded attacks that has been less investigated so far. In such attacks, given an ℓ_0 budget k, an adversary can change k entries of the input vector in an arbitrary fashion – i.e. the adversarial perturbations belong to the so-called ℓ_0 ball of radius k. In contrast with ℓ_p-balls ($p \geq 1$), the ℓ_0-ball is non-convex and non-smooth. Moreover, the ℓ_0 ball contains inherent discrete (combinatorial) structures that can be exploited by both the learner and the adversary. As a result, the ℓ_0-adversarial setting bears various challenges that are absent in common ℓ_p-adversarial settings. In thus regard, it has recently been shown that any piece-wise linear classifier, e.g. a feed-forward deep neural network with ReLu activations, completely fails in the ℓ_0 setting [10].

Perturbing only a few components of the data or signal has many real-world applications including natural language processing [11], malware detection [12], and physical attacks in object detection [13]. There have been several prior works on ℓ_0-adversarial attacks including white-box attacks that are gradient-based, e.g. [4], [14], [15], and black-box attacks based on zeroth-order optimization, e.g. [16], [17]. Defense strategies against ℓ_0-bounded attacks have also been proposed, e.g. defenses based on randomized ablation [18] and defensive distillation [19]. None of the above works have studied the fundamental limits of the ℓ_0-adversarial setting theoretically. In our prior work, we have studied the ℓ_0-adversarial setting for the case of Gaussian mixture model [20]. In this paper, we generalize our results to the case of binary classification with general noise distribution. We note that a line of work in distributed hypothesis testing has considered Byzantine attacks where a fraction of compromised nodes may cooperatively transmit fictitious observations according to different arbitrary distributions. This is different from the ℓ_0 attack setting, where k of the observations can be arbitrarily and adversarially changed (as opposed to their distribution getting adversarially changed) [21], [22], [23].

The goal of this paper is to characterize the optimal classifier and the corresponding robust classification error as a function of the adversary's budget k. More precisely, we focus on the binary classification setting with general but i.i.d. noise distributions, where the input is generated according to the following model: $x_i = y\mu + z_i$, where $y \in \{-1, 1\}$ is the true label, z_i is a zero-mean i.i.d. random noise process, and μ is its mean vector. We seek to find the robust classification error of the optimal classifier in this setting. In other words, we would like to study "how robust" we can design a classifier given a certain budget for an ℓ_0 adversary. Specifically, we consider the asymptotic regime that the dimension of the input gets large, and ask the following fundamental question: What is the maximum adversary's budget for which the optimal error in the absence of an adversary (standard error) can still be achieved and how does this limit scale with the input's dimension?

The main contributions of the paper to answer the above

We thank the NSF grants CNS-2003035 and CIF-1909320.

[1]Uni. of Illinois at Urbana-Champaign, `delgosha@illinois.edu`
[2]Uni. of Pennsylvania, `hassani@seas.upenn.edu`
[3]Uni. of California Santa Barbara, `ramtin@ucsb.edu`

questions are as follows.

- We prove an achievability result by introducing a classifier and characterizing its performance. Our proposed classification method finds the likelihood of each data sample, and applies *truncation* by removing a few of the largest and a few of the smallest values. This truncation phase effectively removes the "outliers" present in the input due to adversarial modification. We have shown in a previous work [20] that truncation is effective to robustify against ℓ_0 attacks in a Gaussian mixture setting. The present work shows the effectiveness of this method in a much broader setting for general noise distributions.
- We prove a converse result by finding a lower bound on the optimal robust error, and show that the two bounds asymptotically match as the dimension $d \to \infty$, hence our proposed classification method is optimally robust against such adversarial attacks. The key idea behind the converse proof is to use techniques from the optimal transport theory and studying the asymptotic behavior of the maximal coupling between the data distribution under the two labels $+1$ and -1. We use such a coupling to design a strategy for the adversary by making the distribution "look almost the same" under the two labels, hence removing the information about the true label.
- Surprisingly, we observe a phase transition for the optimal robust error in terms of the adversary's budget. Roughly speaking, we observe that if the adversary's budget is below $\sqrt{d}$, we can asymptotically achieve the optimal standard error which corresponds to the case where there is no adversary, while if the adversary's budget is above $\sqrt{d}$, no classifier can do better than a random guess. In other words, we can totally compensate for the presence of the adversary as long as its budget is below $\sqrt{d}$ and achieve a performance *as if there were no adversary*. On the other hand, above this threshold $\sqrt{d}$, the adversary can perturb the data in such a way that the information about the true label is lost and hence no classifier can do better than a random guess. Consequently, *there is no trade-off between robustness and accuracy in this setting*.

Truncation has been proved to be useful in robustifying learning algorithms against sparse attacks in other scenarios as well, such as sparse recovery [24] and learning of graphical models [25].

In Section II, we give the problem formulation, in Section III we discuss the main results, and in Section IV we conclude the paper. Proof ideas are discussed in the appendices, and the full proofs are given in [26].

We close this section by introducing some notation. We denote the set of integers $\{1, \dots, n\}$ by $[n]$. $\bar{\Phi}(x) := \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-t^2/2) dt$ denotes the complementary CDF of a standard normal distribution. $\mathcal{N}(\mu, \sigma^2)$ denotes a real-valued normal distribution with mean μ and variance σ^2. $\xrightarrow{\text{dist}}$ and $\xrightarrow{\text{prob}}$ denote convergence in distribution and convergence in probability, respectively. $X \sim p(.)$ means that the random variable X has distribution $p(.)$. We use the boldface notation for vectors in the Euclidean space, e.g. $\boldsymbol{x} \in \mathbb{R}^d$.

II. Problem Formulation

We consider the binary classification setting where the true label is $Y \sim \text{Unif}\{\pm 1\}$ and conditioned on a realization y, d independent real-valued data samples $x_1^{(d)}, \dots, x_d^{(d)}$ are generated such that $x_i^{(d)} = y\mu_d + z_i$. Here, $\mu_d \in \mathbb{R}$ is the conditional expectation of $x_i^{(d)}$ given $y = 1$ and $z_1, \dots, z_d$ are i.i.d. samples of a zero-mean real-valued noise distribution which has a density $q(.)$. We consider a high-dimensional setting where the dimension $d \to \infty$, and μ_d can depend on the data dimension d. However, we assume that the noise density $q(.)$ is fixed and known. Note that since the ℓ_0 norm is invariant under scalar multiplication, we can arbitrarily normalize the quantities, and this assumption is made without loss of generality. We denote the vector of the input data samples by $\boldsymbol{x}^{(d)} = (x_i^{(d)} : i \in [d])$. Throughout this paper, the superscript (d) emphasizes the dependence on the dimension d. However, we may drop it from the notations whenever the dimension is clear from the context. A classifier is a measurable function $\mathcal{C} : \boldsymbol{x} \mapsto \{\pm 1\}$ which predicts the true label from the input $\boldsymbol{x}$. We consider the 0-1 loss $\ell(\mathcal{C}; \boldsymbol{x}, y) := \mathbb{1}[\mathcal{C}(\boldsymbol{x}) \neq y]$ as a metric for discrepancy between the prediction of the classifier on the input $\boldsymbol{x}$ and the true label y.

We assume that an adversary is allowed to perturb the input $\boldsymbol{x}$ within the ℓ_0 ball of radius k:

$$\mathcal{B}_0(\boldsymbol{x}^{(d)}, k) := \{\boldsymbol{x'}^{(d)} \in \mathbb{R}^d : \|\boldsymbol{x}^{(d)} - \boldsymbol{x'}^{(d)}\|_0 \leq k\},$$

where $\|\boldsymbol{x}^{(d)}\|_0 := \sum_{i=1}^d \mathbb{1}\left[x_i^{(d)} \neq 0\right]$. Effectively, the adversary can change at most k data samples. The parameter k is called the *adversary's budget*. Similar to the above, whenever the dimension d is clear from the context, we may denote the adversary's perturbed data samples as $\boldsymbol{x}' = (x'_i : i \in [d])$. In this setting, the *robust classification error* (or *robust error* for short) associated to a classifier $\mathcal{C}$ is defined to be

$$\mathcal{L}^{(d)}_{\mu_d,q}(\mathcal{C}, k) := \mathbb{E}\left[\max_{\boldsymbol{x}' \in \mathcal{B}_0(\boldsymbol{x},k)} \ell(\mathcal{C}; \boldsymbol{x}', y)\right], \tag{1}$$

where the expectation is taken with respect to the above mentioned distribution parametrized by d, μ_d, and q. The *optimal robust classification error* (or *optimal robust error* for short) is defined by optimizing the robust error over all possible (measurable) classifiers:

$$\mathcal{L}^{*(d)}_{\mu_d,q}(k) := \inf_{\mathcal{C}} \mathcal{L}^{(d)}_{\mu_d,q}(\mathcal{C}, k). \tag{2}$$

In words, $\mathcal{L}^{*(d)}_{\mu_d,q}(k)$ is the minimum error that any classifier can achieve in the presence of an adversary with an ℓ_0 budget k. In other words, no classifier can obtain a robust error smaller than $\mathcal{L}^{*(d)}_{\mu_d,q}(k)$ in this setting. Whenever the problem parameters are clear from the context, we may drop them from the notation and write $\mathcal{L}^{(d)}(\mathcal{C}, k)$ or $\mathcal{L}(\mathcal{C}, k)$, and $\mathcal{L}^{*(d)}(k)$ or $\mathcal{L}^*(k)$.

In the absence of the adversary, or equivalently when $k = 0$, $\mathcal{L}^*(0)$ reduces to the *optimal standard error*, which is optimal Bayes error of estimating Y upon observing the noisy samples

$x_1, \ldots, x_d$. In order to fix the baseline, specifically to have a meaningful asymptotic discussion as $d \to \infty$, we assume that μ_d is such that the optimal standard error $\mathcal{L}^{*(d)}_{\mu_d,q}(0)$ remains constant as $d \to \infty$. As we will see later (see Theorem 2 in Section III-A), this is achieved when $\mu_d = c/\sqrt{d}$ for some $c > 0$. Motivated by this, we study the setting where $\mu_d = c/\sqrt{d}$ for some constant $c > 0$ throughout this paper.

III. Main Results

In order to prove our main results, we need the following assumptions on the noise distribution $q(.)$. We will show later (see Section III-D) that all of these assumptions are satisfied for a large class of distributions, including the exponential family of distributions with polynomial exponents, e.g. the normal distribution.

Assumption 1. *We have $q(z) > 0$ for all $z \in \mathbb{R}$, $q(.)$ is three times continuously differentiable, and*

$$\int_{-\infty}^{\infty} q'(z)dz = \int_{-\infty}^{\infty} q''(z)dz = 0,$$

where $q'(.)$ and $q''(.)$ denote the first and second derivatives of $q(.)$. Furthermore, the location family of distributions

$$q(z;\theta) := q(z-\theta), \tag{3}$$

parameterized by $\theta \in \mathbb{R}$ has well-defined and finite Fisher information $\{\mathcal{I}_q(\theta)\}_{\theta\in\mathbb{R}}$.

The Fisher information of the parametric family of distributions $q(z;\theta)$ where $z, \theta \in \mathbb{R}$ is defined to be

$$\mathcal{I}_q(\theta) := \int \left(\frac{\partial}{\partial\theta}\log q(z;\theta)\right)^2 q(z;\theta)dz.$$

See, for instance, [27] for more details. Since $q(z;\theta) = q(z-\theta)$ is a location family, it turns out that $\mathcal{I}_q(\theta)$ is independent of θ. The common value, which we denote by $\mathcal{I}_q$ by an abuse of notation, is given by

$$\mathcal{I}_q := \int_{-\infty}^{\infty} \frac{(q'(z))^2}{q(z)}dz. \tag{4}$$

Assumption 2. *There exists $\zeta > 0$ such that*

$$\mathbb{E}_{Z\sim q(.)}\left[\sup_{t\in[Z,Z+\zeta]}\left|\frac{d^3}{dt^3}\log q(t)\right|\right] < \infty. \tag{5}$$

Assumption 3. *There exist $\zeta > 0$ such that*

$$\mathbb{E}_{Z\sim q(.)}\left[\sup_{t\in[Z,Z+\zeta]}\left|\frac{d^2}{dt^2}\log q(t)\right|^2\right] < \infty. \tag{6}$$

Assumption 4. *There exist constants $\gamma > 0$ and $C_4 > 0$ such that*

$$\lim_{d\to\infty}\mathbb{P}\left(\max_{1\le i\le d}\left|\frac{d}{dz}\log q(Z_i)\right| > C_4(\log d)^\gamma\right) = 0,$$

where Z_i are i.i.d. with distribution $q(.)$.

The following theorem formalizes the phase transition we discussed previously, i.e. if adversary's budget is orderwise below $\sqrt{d}$, we can totally compensate for its presence, while if adversary's budget is orderwise above $\sqrt{d}$, no classifier can do better than a random guess. As we discusses previously, we assume that $\mu_d = c/\sqrt{d}$ for a constant $c > 0$ to ensure that the standard error is asymptotically constant (see Theorem 2 in Section III-A).

Theorem 1. *Assume that $\mu_d = c/\sqrt{d}$ for some constant $c > 0$, and the assumptions 1-4 are satisfied for the noise density $q(.)$. Then, if k_d is a sequence of adversary's ℓ_0 budget, then we have*

1) *If $\limsup_{d\to\infty}\log_d k_d < 1/2$, there exists a sequence of classifiers $\mathcal{C}^{(d)}_{k_d}$ such that*

$$\limsup_{d\to\infty}\mathcal{L}^{(d)}_{\mu_d,q}(\mathcal{C}^{(d)}_{k_d}, k_d) - \mathcal{L}^{*(d)}_{\mu_d,q}(0) = 0.$$

In other words, the excess risk of this sequence of classifiers as compared to the optimal standard error (when there is no adversary) converges to zero.

2) *If $\liminf_{d\to\infty}\log_d k_d > 1/2$, we have*

$$\liminf_{d\to\infty}\mathcal{L}^{*(d)}_{\mu_d,q}(k_d) \ge 1/2.$$

In other words, no classifier can asymptotically do better than a random guess.

The proof of this result essentially follows from Theorems 3 and 4 below. More precisely, in Section III-B, we prove an achievability result by introducing a sequence of robust classifiers in the sub-$\sqrt{d}$ regime (first part of the theorem), while in Section III-C, we prove a converse result by introducing a strategy for the adversary in the super-$\sqrt{d}$ regime which perturbs the data in such a way that the information about the true label is asymptotically removed (second part of the theorem). See [26] for a complete proof of Theorem 1.

A. Asymptotic Standadrd Error

Recall that in the absence of the adversary, or equivalently when adversary's budget k is zero, the optimal robust error $\mathcal{L}^{*(d)}_{\mu_d,q}(0)$ reduces to the optimal Bayes error of estimating Y upon observing the noisy samples $x_1, \ldots, x_d$. With an abuse of notation, we write $\mathcal{L}^{*(d)}_{\mu_d,q}$ (or $\mathcal{L}^*$ for short) for this optimal Bayes error. Our goal in this section is to find the appropriate scaling of μ_d with d such that $\mathcal{L}^{*(d)}_{\mu_d,q}$ converges to a constant as $d \to \infty$.

In order to characterize $\mathcal{L}^*$, note that since there is no adversary, and the prior on Y is uniform, the optimal Bayes classifier is the maximum likelihood estimator that computes the likelihood

$$\sum_{i=1}^{d}\widetilde{x}_i^{(d)} \qquad \text{where} \qquad \widetilde{x}_i^{(d)} := \log\frac{q(x_i^{(d)}-\mu_d)}{q(x_i^{(d)}+\mu_d)}, \tag{7}$$

and returns the estimate $\hat{y}$ of y as

$$\hat{y} = \begin{cases} 1 & \sum_{i=1}^{d}\widetilde{x}_i^{(d)} > 0 \\ -1 & \text{otherwise.}\end{cases} \tag{8}$$

The following Theorem 2 shows that if $\mu_d = c/\sqrt{d}$, then the optimal Bayes error converges to a constant. See Appendix A for proof ideas and [26] for a full proof.

Theorem 2. *Assume that assumptions 1 and 2 are satisfied for the noise density $q(.)$. Then, if $\mu_d = \frac{c}{\sqrt{d}}$ for some constant $c > 0$, we have*

$$\lim_{d\to\infty} \mathcal{L}^{*(d)}_{\mu_d,q} = \bar{\Phi}(c\sqrt{\mathcal{I}_q}).$$

Furthermore, in this case, as d $\to \infty$, conditioned on $Y = +1$, the log likelihood $\sum_{i=1}^{d} \widetilde{x}^{(d)}_i$ converges in distribution to a normal $\mathcal{N}(2c^2\mathcal{I}_q, 4c^2\mathcal{I}_q)$ where $\mathcal{I}_q$ was defined in (4) and is the Fisher information associated to the location family defined in (3). Moreover, conditioned on $Y = -1$, $\sum_{i=1}^{d} \widetilde{x}^{(d)}_i$ converges in distribution to a normal $\mathcal{N}(-2c^2\mathcal{I}_q, 4c^2\mathcal{I}_q)$.

B. Achievability: Upper Bound on the Optimal Robust Error

In this section, we introduce a classifier and study its robustness against ℓ_0 adversarial perturbations. Recall that if k is the adversary's budget, the input to the classifier is $\boldsymbol{x}' = (x'_1, \ldots, x'_d)$ which is different from the original sequence $x_1, \ldots, x_d$ in at most k coordinates. Recall from Section III-A that in the absence of the adversary, the optimal Bayes classifier is the maximum likelihood estimator based on $\sum_{i=1}^{d} \widetilde{x}_i$, as was defined in (7). Motivated by this, we define

$$\widetilde{x}'^{(d)}_i := \log \frac{q(x'^{(d)}_i - \mu_d)}{q(x'^{(d)}_i + \mu_d)}. \tag{9}$$

Note that if $\widetilde{\boldsymbol{x}}'^{(d)}$ denotes the vector $(\widetilde{x}'^{(d)}_i : i \in [d])$, since $\|\boldsymbol{x}'^{(d)} - \boldsymbol{x}^{(d)}\|_0 \le k$, we have

$$\|\widetilde{\boldsymbol{x}}'^{(d)} - \widetilde{\boldsymbol{x}}^{(d)}\|_0 \le k. \tag{10}$$

We define the truncated classifier $\mathcal{C}^{(d)}_k$ as follows. Given a vector $\boldsymbol{u} = (u_i : i \in [d]) \in \mathbb{R}^d$ and an integer $k \ge 0$, we define the truncated summation $\mathsf{TSum}_k(\boldsymbol{u})$ to be the summation of coordinates in $\boldsymbol{u}$ except for the top and bottom k coordinates. More precisely, let $\boldsymbol{s} = (s_i : i \in [d]) = \mathrm{sort}(\boldsymbol{u})$ be obtained by sorting the coordinates of $\boldsymbol{u}$ in descending order. We then define

$$\mathsf{TSum}_k(\boldsymbol{u}) := \sum_{i=k+1}^{d-k} s_i. \tag{11}$$

When $k = 0$, this indeed reduces to the normal summation. Motivated by (10), we replace $\sum_{i=1}^{d} \widetilde{x}^{(d)}_i$ with its *robustified* version $\mathsf{TSum}_k(\sum_{i=1}^{d} \widetilde{x}'^{(d)}_i)$ and define

$$\mathcal{C}^{(d)}_k(\boldsymbol{x}'^{(d)}) := \begin{cases} +1 & \mathsf{TSum}_k(\widetilde{\boldsymbol{x}}'^{(d)}) > 0 \\ -1 & \text{otherwise.} \end{cases} \tag{12}$$

This method essentially removes the "outliers" introduced by the adversary into the data.

The following theorem shows that this classifier is asymptotically robust against adversarial attacks with ℓ_0 budget of at most $\sqrt{d}$. A matching lower bound is provided in Section III-C. The proof outline of Theorem 3 below is given in Appendix B. See [26] for a full proof.

Theorem 3. *Assume that Assumptions 1-4 are satisfied for the noise density $q(.)$, and $\mu_d = c/\sqrt{d}$ for some $c > 0$. Then if k_d is a sequence of adversary's budgets so that $k_d < d^{\frac{1}{2}-\epsilon}$ for some $\epsilon > 0$, then we have*

$$\limsup_{d\to\infty} \mathcal{L}^{(d)}_{\mu_d,q}(\mathcal{C}^{(d)}_{k_d}, k_d) \le \bar{\Phi}(c\sqrt{\mathcal{I}_q}). \tag{13}$$

In particular, we have

$$\limsup_{d\to\infty} \mathcal{L}^{(d)}_{\mu_d,q}(\mathcal{C}^{(d)}_{k_d}, k_d) - \mathcal{L}^{*(d)}_{\mu_d,q} = 0. \tag{14}$$

Note that $\mathcal{L}^{*(d)}_{\mu_d,q}$, as was defined in Section III-A above, is the optimal Bayes error in an ideal scenario when there is no adversary, and $\mathcal{L}^{(d)}_{\mu_d,q}(\mathcal{C}^{(d)}_{k_d}, k_d) - \mathcal{L}^{*(d)}_{\mu_d,q}$ is the excess error of our truncated classifier with respect to this ideal scenario. In fact, (14) implies that our truncated classifier is asymptotically optimal in the specified regime of adversary's budget. The truncated classifier manages to compensate for the presence of the adversary, and performs as if there is no adversary.

C. Converse: Lower Bound on the Optimal Robust Error

In this section, we provide a lower bound on the optimal robust error. We do this by introducing an attack strategy for the adversary. In this strategy, the adversary with a sufficiently large budget, perturbs the input data in such a way that all the information about the true label Y is lost, resulting in a perturbed data which has a vanishing correlation with the true label. The proof outline of Theorem 4 is given in Appendix C. See [26] for a complete proof.

Theorem 4. *Assume that Assumptions 1-4 are satisfied for the noise density $q(.)$, and $\mu_d = c/\sqrt{d}$ for some $c > 0$. Then, if k_d is a sequence of adversary's budgets so that $k_d > d^{1/2+\epsilon}$ for some $\epsilon > 0$, then we have* $\liminf_{d\to\infty} \mathcal{L}^{*(d)}_{\mu_d,q}(k_d) \ge 1/2$.

D. Exponential Family of Distributions

In this section, we show that assumptions 1-4 are all satisfied for a large class of distributions, namely the exponential family of noise distributions of the form $q(z) = \frac{\exp(\psi(z))}{A}$ where $\psi(z) = -a_{2n}z^{2n} + a_{2n-1}z^{2n-1} + \ldots a_1 z + a_0$ is a polynomial in z with even degree $2n > 0$ such that $a_{2n} > 0$. Here, $A := \int_{-\infty}^{\infty} \psi(z)dz$ is the normalizing constant. Note that since $\psi(.)$ has an even degree with a negative leading coefficient, we have $A < \infty$. Proof ideas of Theorem 5 below are given in Appendix D. See [26] for a complete proof.

Theorem 5. *Assumptions 1- 4 are all satisfied for the density $q(.)$ of the form discussed above.*

IV. Conclusion

We studied the binary classification problem in the presence of an adversary constrained by the ℓ_0 norm. We introduced a robust classification method which employs truncation on the log likelihood. We showed that this classification method can asymptotically compensate for the presence of the adversary as long as adversary's budget is orderwise below $\sqrt{d}$. Moreover, we showed a phase transition through a converse argument in the sense that no classifier can asymptotically do better than a random guess if adversary's budget is orderwise above $\sqrt{d}$.

APPENDIX A
THEOREM 2: PROOF IDEAS

We have ${\mathcal{L}^*}^{(d)}_{\mu_d,q} = \frac{1}{2}\mathbb{P}\left(\sum_{i=1}^d \widetilde{x}_i^{(d)} \leq 0|Y=+1\right) + \frac{1}{2}\mathbb{P}\left(\sum_{i=1}^d \widetilde{x}_i^{(d)} > 0|Y=-1\right)$. From now on, we focus on the term conditioned on $Y=+1$, since the second term can be analyzed similarly. In this case, with $\mu_d = c/\sqrt{d}$, we may write

$$\sum_{i=1}^d \widetilde{x}_i^{(d)} = \sum_{i=1}^d \log\frac{q(z_i)}{q(z_i+2\mu_d)} = \frac{c}{\sqrt{d}}\sum_{i=1}^d \frac{1}{\mu_d}\log\frac{q(z_i)}{q(z_i+2\mu_d)}. \tag{15}$$

It can be seen that writing the Taylor expansion of $\log q(z_i + 2\mu_d)$ around z_i and simplifying, we get

$$\sum_{i=1}^d \log\frac{q(z_i)}{q(z_i+2\mu_d)} = \underbrace{\frac{-2c}{\sqrt{d}}\sum_{i=1}^d \frac{d}{dz}\log q(z_i)}_{=:T_1} + \underbrace{\frac{-2c\mu_d}{\sqrt{d}}\sum_{i=1}^d \frac{d^2}{dz^2}\log q(z_i)}_{=:T_2} + \underbrace{\frac{-4c\mu_d^2}{3\sqrt{d}}\sum_{i=1}^d \frac{d^3}{dz^3}\log q(z_i+\epsilon_i)}_{=:T_3}, \tag{16}$$

with $\epsilon_i \in (0, 2\mu_d)$. The rest of the proof follows by asymptotically studying the above three terms. More precisely, it can be shown that using the central limit theorem, we have $T_1 \xrightarrow[d\to\infty]{\text{dist}} \mathcal{N}(0, 4c^2\mathcal{I}_q)$. Moreover, law of large numbers implies that T_2 converges to $2c^2\mathcal{I}_q$ almost surely. Additionally, it can be seen that assumption 2 together with the law of large numbers ensure that T_3 converges to zero almost surely. Using these in (15) and (16), we realize that conditioned on $Y=+1$, $\sum_{i=1}^d \widetilde{x}_i^{(d)}$ converges in distribution to $\mathcal{N}(2c^2\mathcal{I}_q, 4c^2\mathcal{I}_q)$, and therefore $\mathbb{P}\left(\sum_{i=1}^d \widetilde{x}_i^{(d)} \leq 0|Y=+1\right)$ converges to $\bar{\Phi}(c\sqrt{\mathcal{I}_q})$.

APPENDIX B
THEOREM 3: PROOF IDEAS

We have

$$\begin{aligned}&\mathcal{L}^{(d)}_{\mu_d,q}(\mathcal{C}_k^{(d)}, k_d)\\ &= \frac{1}{2}\mathbb{P}\left(\exists \boldsymbol{x}'^{(d)} \in \mathcal{B}_0(\boldsymbol{x}^{(d)}, k_d) : \mathsf{TSum}_k(\widetilde{\boldsymbol{x}}'^{(d)}) \leq 0|Y=+1\right)\\ &\quad + \frac{1}{2}\mathbb{P}\left(\exists \boldsymbol{x}'^{(d)} \in \mathcal{B}_0(\boldsymbol{x}^{(d)}, k_d) : \mathsf{TSum}_k(\widetilde{\boldsymbol{x}}'^{(d)}) \geq 0|Y=-1\right).\end{aligned} \tag{17}$$

From this point forward, we focus on the term conditioned on $Y=+1$, and the other term can be analyzed similarly. Note that for $\boldsymbol{x}'^{(d)} \in \mathcal{B}_0(\boldsymbol{x}^{(d)}, k_d)$, we have $\|\widetilde{\boldsymbol{x}}'^{(d)} - \widetilde{\boldsymbol{x}}^{(d)}\|_0 \leq 0$. Therefore, using [20, Lemma 1], for all $\boldsymbol{x}'^{(d)} \in \mathcal{B}_0(\boldsymbol{x}^{(d)}, k_0)$, we have $\mathsf{TSum}_k(\widetilde{\boldsymbol{x}}'^{(d)}) \geq \left(\sum_{i=1}^d \widetilde{x}_i^{(d)}\right) - 8k_d\|\widetilde{\boldsymbol{x}}^{(d)}\|_\infty$. This implies that the first term on the right hand side of (17) is bounded from above by $\frac{1}{2}\mathbb{P}\left(\sum_{i=1}^d \widetilde{x}_i^{(d)} \leq 8k_d\|\widetilde{\boldsymbol{x}}^{(d)}\|_\infty|Y=+1\right)$. Note that from Theorem 2, conditioned on $Y=+1$, we have $\sum_{i=1}^d \widetilde{x}_i^{(d)} \xrightarrow[d\to\infty]{\text{dist}} \mathcal{N}(2c^2\mathcal{I}_q, 4c^2\mathcal{I}_q)$. Therefore, it suffices to show that $k_d\|\widetilde{\boldsymbol{x}}^{(d)}\|_\infty \xrightarrow[d\to\infty]{\text{prob}} 0$ provided that $k_d < d^{1/2-\epsilon}$. This can be done by writing the Taylor expansion similar to the proof of Theorem 2 up to the second order term. More precisely, it can be shown that

$$\begin{aligned}\|\widetilde{\boldsymbol{x}}^{(d)}\|_\infty \leq\ & \underbrace{\frac{2c}{\sqrt{d}}\max_{1\leq i\leq d}\left|\frac{d}{dz}\log q(z_i)\right|}_{=:T_1(d)}\\ &+ \underbrace{\frac{2c^2}{d}\max_{1\leq i\leq d}\sup_{t\in[z_i,z_i+2\mu_d]}\left|\frac{d^2}{dt^2}\log q(t)\right|}_{=:T_2(d)},\end{aligned} \tag{18}$$

where $\epsilon_i \in (0, 2\mu_d)$. It can be shown that from assumption 4, $k_d T_1(d) \xrightarrow[d\to\infty]{\text{prob}} 0$. Also, it can be shown that assumption 3 together with the law of large numbers, $k_d T_2(d) \xrightarrow[d\to\infty]{\text{prob}} 0$. Hence, $k_d\|\widetilde{\boldsymbol{x}}^{(d)}\|_\infty \xrightarrow[d\to\infty]{\text{prob}} 0$ and we obtain the desired bound.

APPENDIX C
THEOREM 4: PROOF IDEAS

Let $q_+^{(d)}$ and $q_-^{(d)}$ be the densities of $Z+\mu_d$ and $Z-\mu_d$, when $Z \sim q(.)$. Using ideas from the optimal transport theory, we can couple these two distributions with a mismatch probability bounded by $d_{\mathrm{TV}}(q_+^{(d)}, q_-^{(d)})$, where d_{TV} denotes the total variation distance. Now we can introduce a strategy for the adversary using this optimal coupling. Roughly speaking, in case of a mismatch, the adversary changes the data value to zero. This ensures that the information about the true label is completely removed. Since we have d data samples, the average required ℓ_0 budget for this strategy is $d \times d_{\mathrm{TV}}(q_+^{(d)}, q_-^{(d)})$, which using the Pinsker's inequality is bounded by $d\sqrt{\frac{1}{2}D(q_+^{(d)}\|q_-^{(d)})}$, where $D(.\|.)$ denotes the KL divergence. It can be shown that $D(q_+^{(d)}\|q_-^{(d)})$ asymptotically behaves like $2\mu_d^2\mathcal{I}_q + o(\mu_d^2)$, which scales as $1/d$ since $\mu_d = c/\sqrt{d}$. Hence, with an average adversary's budget of order $d/\sqrt{d} = \sqrt{d}$, we can effectively remove the information about the true label.

APPENDIX D
THEOREM 5: PROOF IDEAS

Assumption 1 is easy to verify, since $q'(z) = \psi'(z)q(z)$, $\psi'(z)$ is a polynomial in z, and $\int \text{poly}(z)\exp(\psi(z)) < \infty$ for any polynomial $\text{poly}(z)$. To verify assumptions 2 and 3, note that $d^r/dt^r \log q(z)$ is a polynomial in z for any integer r. Furthermore, it can be show that for any polynomial $p(z)$ and $\epsilon > 0$, the function $z \mapsto \sup_{t\in[z,z+\epsilon]} p(t)$ can be bounded above by another polynomial in $|z|$ with the same degree as $p(.)$. Additionally, the expectation of any polynomial with respect to the density $q(.)$ is finite. These together are sufficient to verify assumptions 2 and 3. To verify assumption 4, we first study the tail behavior of $q(.)$ and show that with high probability, $\max_{1\leq i\leq d}|Z_i|$ is bounded by $O(\log d)^{1/2n}$. On the other hand, since $\frac{d}{dz}\log q(z)$ is a polynomial in z, $\max_{1\leq i\leq d}|Z_i|$ can be used to obtain the desired upper bound for $\max_{1\leq i\leq d}|\frac{d}{dz}\log q(Z_i)|$ which holds with high probability.

References

[1] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations, 2014, Banff, AB, Canada, April 14-16*, 2014. [Online]. Available: http://arxiv.org/abs/1312.6199

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

[4] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy, San Jose, CA, USA, May 22-26,*, 2017, pp. 39–57.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=rJzIBfZAb

[6] A. Athalye, N. Carlini, and D. A. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning, ICML, Stockholm, Sweden, July 10-15*, 2018, pp. 274–283. [Online]. Available: http://proceedings.mlr.press/v80/athalye18a.html

[7] Z. Marzi, S. Gopalakrishnan, U. Madhow, and R. Pedarsani, "Sparsity-based defense against adversarial attacks on linear classifiers," in *2018 IEEE International Symposium on Information Theory, ISIT, Vail, CO, USA, June 17-22, 2018*, pp. 31–35.

[8] R. Bhattacharjee and K. Chaudhuri, "Consistent non-parametric methods for adaptive robustness," *arXiv preprint arXiv:2102.09086*, 2021.

[9] R. Bhattacharjee, S. Jha, and K. Chaudhuri, "Sample complexity of adversarially robust linear classification on separated data," *arXiv preprint arXiv:2012.10794*, 2020.

[10] A. Shamir, I. Safran, E. Ronen, and O. Dunkelman, "A simple explanation for the existence of adversarial examples with small hamming distance," *arXiv preprint arXiv:1901.10861*, 2019.

[11] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is bert really robust? natural language attack on text classification and entailment," *arXiv preprint arXiv:1907.11932*, vol. 2, 2019.

[12] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, "Adversarial perturbations against deep neural networks for malware classification," *arXiv preprint arXiv:1606.04435*, 2016.

[13] J. Li, F. Schmidt, and Z. Kolter, "Adversarial camera stickers: A physical camera-based attack on deep learning systems," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3896–3904.

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

[15] A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard, "Sparsefool: a few pixels make a big difference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9087–9096.

[16] L. Schott, J. Rauber, M. Bethge, and W. Brendel, "Towards the first adversarially robust neural network model on mnist," *arXiv preprint arXiv:1805.09190*, 2018.

[17] F. Croce, M. Andriushchenko, N. D. Singh, N. Flammarion, and M. Hein, "Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks," *arXiv preprint arXiv:2006.12834*, 2020.

[18] A. Levine and S. Feizi, "Robustness certificates for sparse adversarial attacks by randomized ablation." in *AAAI*, 2020, pp. 4585–4593.

[19] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 582–597.

[20] P. Delgosha, H. Hassani, and R. Pedarsani, "Robust classification under ℓ_0 attack for the gaussian mixture model," *arXiv preprint arXiv:2104.02189, to appear in SIAM Journal on Mathematics of Data Science*, 2021.

[21] P. K. Varshney, *Distributed detection and data fusion*. Springer Science & Business Media, 2012.

[22] S. Marano, V. Matta, and L. Tong, "Distributed detection in the presence of byzantine attacks," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 16–29, 2008.

[23] Y. Ni, K. Ding, Y. Yang, and L. Shi, "On the performance analysis of binary hypothesis testing with byzantine sensors," in *2019 Chinese Control Conference (CCC)*. IEEE, 2019, pp. 8889–8894.

[24] Y. Chen, C. Caramanis, and S. Mannor, "Robust high dimensional sparse regression and matching pursuit," *arXiv preprint arXiv:1301.2725*, 2013.

[25] F. Zhang and V. Tan, "Robustifying algorithms of learning latent trees with vector variables," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[26] P. Delgosha, H. Hassani, and R. Pedarsani, "Binary classification under ℓ_0 attacks for general noise distribution," *arXiv preprint arXiv:2203.04855*, 2022.

[27] E. Lehmann and G. Casella, *Theory of Point Estimation*, ser. Springer Texts in Statistics. Springer New York, 2006. [Online]. Available: https://books.google.com/books?id=4f24CgAAQBAJ