Robust Adversarial Classification via Abstaining

Abed AlRahman Al Makdah, Vaibhav Katewa, and Fabio Pasqualetti

Abstract—In this work, we consider a binary classification problem and cast it into a binary hypothesis testing framework, where the observations can be perturbed by an adversary. To improve the adversarial robustness of a classifier, we include an abstain option, where the classifier abstains from making a decision when it has low confidence about the prediction. We propose metrics to quantify the nominal performance of a classifier with an abstain option and its robustness against adversarial perturbations. We show that there exist a tradeoff between the two metrics regardless of what method is used to choose the abstain region. Our results imply that the robustness of a classifier with an abstain option can only be improved at the expense of its nominal performance. Further, we provide necessary conditions to design the abstain region for a 1-dimensional binary classification problem. We validate our theoretical results on the MNIST dataset, where we numerically show that the tradeoff between performance and robustness also exist for the general multi-class classification problems.

I. INTRODUCTION

Data-driven and machine learning models are shown to be vulnerable to adversarial examples, which are small, targeted, and malicious perturbations of the inputs that induce unwanted, and possibly dangerous model behavior [1]. For instance, placing stickers at specific locations on a stop sign can fool a state-of-the-art model into classifying it as a speed limit sign [2]. This vulnerability is one of the main limitations that hurdle the deployment of data-driven systems in safety-critical applications, such as medical diagnosis [3], robotic surgery [4], and self-driving cars [5]. In control applications, classifiers play an important role in decision making, in particular for autonomous systems [5]–[7]. Unlike data-driven models that help with writing an email, classify images of cats and dogs, or recommend movies, small error in safety-critical applications can result in catastrophic consequences [8].

A substantial body of literature addresses adversarial robustness of data-driven models [9]–[12]. Despite all these contributions to guarantee robustness against adversarial perturbations, robust models still fail to achieve optimal robustness. In fact, improving the robustness of these models comes at the expense of their nominal performance [12]–[14]. Thus, unwanted behavior will still exist for robust models on nominal inputs, therefore, safety remains at risk.

Several frameworks are developed to improve the adversarial robustness in classification [9]–[12]. However, in all these frameworks, robustness of a classifier is mainly improved via tuning the position of its decision boundaries.

This material is based upon work supported in part by awards ONR ONR-N00014-19-1-2264, AFOSR FA9550-19-1-0235 and FA9550-20-1-0140. A. A. Al Makdah and F. Pasqualetti are with the Department of Electrical and Computer Engineering and the Department of Mechanical Engineering at the University of California, Riverside, respectively. {aalmakdah,fabiopas}@engr.ucr.edu. V. Katewa is with the Department of Electrical Communication Engineering at the Indian Institute of Science, Bangalore, India, vKatewa@iisc.ac.in.

In this work, we take a different route for addressing adversarial robustness in classification problems. We consider an abstain option, where a classifier with fixed classification boundaries may abstain from giving an output over some region in the input space that the classifier is uncertain about. Mainly the inputs in such a region are the most prone to adversarial attacks. Thus, abstaining over such a region helps the classifier to avoid misclassifying perturbed inputs, and hence improve its adversarial robustness. In particular, under a perturbed input, instead of giving a wrong output (or possibly a correct output with low confidence), the model decides to abstain from giving one. For instance, if a self-driving car detects an object that it is uncertain about (it could be a shadow or maybe sensor measurements are perturbed by an adversary), it could abstain from giving an output that might lead to a car accident, and ask a human to take control.

In safety critical applications, abstaining on low confidence output might be better than making a wrong decision.

Motivated by this, we study the problem of classification with an abstain option by casting it into a binary hypothesis testing framework, where we add a third region in the observation space that corresponds to the observations on which the classifier abstains on (Fig. 1). Particularly, we study the relation between the accuracy and the adversarial robustness of a binary classifier upon varying the abstain region, where we show that improving the adversarial robustness of a classifier via abstaining comes at the expense of its accuracy.

Contributions. This paper features three main contributions. First, we propose metrics to quantify the performance of a classifier with an abstain option and its adversarial robustness. Second, we show that for a binary classification
problem with an abstain option, a tradeoff between performance and adversarial robustness always exist regardless of which region of the input space is abstained on. Thus, the robustness of a classifier with an abstain option can only be improved at the expense of its nominal performance. Further, we numerically show that such a tradeoff exist for the general multi-class classification problems. The type of the tradeoff we present in this paper is different than the one studied in the literature [12]–[14], degrading the nominal performance implies that the classifier abstains more often on nominal inputs, and it does not imply an increase in the misclassification rate. Third, we provide necessary conditions to optimally design the abstain region for a given classifier for the 1-dimensional binary classification problem.

Related work. The literature on classification with an abstain option (also referred to as reject option or selective classification) mainly discusses methods on how to abstain on uncertain inputs. [15], [16] augmented the output class set with a reject class in a binary classification problem, where inputs with probability below a certain threshold are abstained on. Further, [17] used abstaining in multi-class classification problems, where abstaining was used in deep neural networks. In [18], abstaining was used in a regression learning problem. While little work has been done on using abstaining in the context of adversarial robustness, recent work has developed algorithms that guarantee robustness against adversarial attacks via abstaining [19], [20], where a tradeoff between nominal performance and adversarial robustness has been observed upon tuning their algorithms. In this work, we formally prove the existence of such a tradeoff between performance and adversarial robustness, where we show that this tradeoff exist regardless of what algorithm is used to select the abstain region.

Paper’s organization. The rest of the paper is organized as follows. Section II contains our mathematical setup. Section III contains the tradeoff between performance and robustness, design of optimal abstain region, and an illustrative example. Section IV contains our numerical experiment on the MNIST dataset, and Section V concludes the paper.

II. PROBLEM SETUP AND PRELIMINARY NOTIONS

We consider a d-dimensional binary classification problem formulated as hypothesis testing problem as in [13]. The objective is to decide whether an observation ̂x ∈ ℜd belongs to class ̃H0 or class ̃H1. We assume that the distribution of the observations under class ̃H0 and class ̃H1 satisfy

\[ \begin{align*}
    \mathcal{H}_0 : x \sim f_0(x), \quad \mathcal{H}_1 : x \sim f_1(x),
\end{align*} \]

where \( f_0(x) \) and \( f_1(x) \) are known arbitrary probability density functions. For notational convenience, in the rest of this paper we denote \( f_0(x) \) and \( f_1(x) \) by \( f_0 \) and \( f_1 \), respectively. We denote the prior probabilities of the observations under \( f_0 \) and \( f_1 \) by \( p_0 \) and \( p_1 \), respectively. In this setup, any classifier can be represented by a partition of the ℜd space by placing decision boundaries at suitable positions (see Fig. 1).

We consider adversarial manipulations of the observations, where an attacker is capable of adding perturbations to the observations in order to degrade the performance of the classifier. We model\(^1\) such manipulations as a change of the probability density functions in (1). We refer to the perturbed \( f_0 \) and \( f_1 \) in (1) as \( \tilde{f}_0 \) and \( \tilde{f}_1 \), respectively. In this work, we aim to improve the adversarial robustness of any classifier by abstaining from making a decision for low confidence outputs. A classifier with an abstain option can be written as

\[ \mathcal{C}(x; g(x), g_1(x), g_2(x)) = \begin{cases}
    \mathcal{H}_0, & x \in \mathcal{R}_0 \cap \mathcal{R}_a, \\
    \mathcal{H}_1, & x \in \mathcal{R}_1 \cap \mathcal{R}_a, \\
    \mathcal{H}_a, & x \in \mathcal{R}_a,
\end{cases} \]  

where \( g(x) \)\(^2\) gives the hyperspace decision boundary for the non-abstain case, \( g_1(x) \) and \( g_2(x) \) give the hyperspace boundaries for the abstain region, specifically,

\[ \begin{align*}
    \mathcal{R}_0 &= \{ z : g(z) \leq 0, \forall z \in \mathbb{R}^d \}, \\
    \mathcal{R}_1 &= \{ z : g(z) > 0, \forall z \in \mathbb{R}^d \}, \\
    \mathcal{R}_a &= \{ z : (g_1(z) \geq 0) \cap (g_2(z) \leq 0), \forall z \in \mathbb{R}^d \},
\end{align*} \]

and \( \mathcal{R}_a \) is the complement set of \( \mathcal{R}_a \). We define two metrics to measure the performance and robustness of classifier (2).

Definition 1: (Nominal error) The nominal error of a classifier with an abstain option is the proportion of the (un-perturbed) observations that are misclassified or abstained on,

\[ e_{\text{nom}}(\mathcal{R}_0, \mathcal{R}_1, \mathcal{R}_a) = p_0 \mathbb{P}[x \in \mathcal{R}_1 | \mathcal{H}_0] + p_1 \mathbb{P}[x \in \mathcal{R}_0 | \mathcal{H}_1] + p_0 \mathbb{P}[x \in (\mathcal{R}_0 \cap \mathcal{R}_a) | \mathcal{H}_0] + p_1 \mathbb{P}[x \in (\mathcal{R}_1 \cap \mathcal{R}_a) | \mathcal{H}_1], \]

The first two terms in (4) correspond to the error without abstaining, therefore, they do not depend on the abstain region \( \mathcal{R}_a \). The last two terms correspond to the abstain error, thus, they depend on \( \mathcal{R}_a \). Using Definition 1 and the distributions in (1), the nominal error for classifier (2) is written as

\[ e_{\text{nom}}(\mathcal{R}_0, \mathcal{R}_1, \mathcal{R}_a) = p_0 \int_{\mathcal{R}_1} f_0 dx + p_1 \int_{\mathcal{R}_0} f_1 dx \]

\[ + p_0 \int_{\mathcal{R}_0 \cap \mathcal{R}_a} f_0 dx + p_1 \int_{\mathcal{R}_1 \cap \mathcal{R}_a} f_1 dx. \]  

As can be seen in (5), the nominal classification error depends on \( \mathcal{R}_0, \mathcal{R}_1, \) and \( \mathcal{R}_a \), and thus on the position of the boundaries, \( g(x) \), \( g_1(x) \), and \( g_2(x) \), as described in (3). Lower nominal error implies higher classification performance. Note that, if there is no abstain option (\( \mathcal{R}_a = \mathcal{D} \)), then the nominal error is equal to the error computed in the classic hypothesis testing framework [21].

Definition 2: (Adversarial error) The adversarial error of a classifier with an abstain option is the proportion of the perturbed observations that are misclassified and not abstained on,

\[ e_{\text{adv}}(\mathcal{R}_0, \mathcal{R}_1, \mathcal{R}_a) = p_0 \mathbb{P}[\tilde{x} \in (\mathcal{R}_1 \cap \mathcal{R}_a) | \mathcal{H}_0] + p_1 \mathbb{P}[\tilde{x} \in (\mathcal{R}_0 \cap \mathcal{R}_a) | \mathcal{H}_1], \]

\(^1\)In this work, we do not specify a model for the adversary, our analysis holds independently of the adversary model.

\(^2\)Technically, \( g(x) \) is not the boundary, \( g(x) = 0 \) provides the boundary, but for the notational convenience we use \( g(x) \) to refer to the boundary. Similarly, we use \( g_1(x) \) and \( g_2(x) \) instead of \( g_1(x) = 0 \) and \( g_2(x) = 0 \).
where \( \tilde{x} \in \mathbb{R}^d \) is a perturbed observation that follows distributions \( f_0 \) and \( f_1 \) under classes \( \mathcal{H}_0 \) and \( \mathcal{H}_1 \), respectively.

Using Definition 2 and the distributions in (1), we can write the adversarial error for classifier (2) as

\[
e_{\text{adv}}(R_0, R_1, \tilde{R}_a) = p_0 \int_{R_1 \cap \tilde{R}_a} f_0 dx + p_1 \int_{\tilde{R}_a \setminus R_0} f_1 dx,
\]

Similar to the nominal error, the adversarial error depends on \( R_0, R_1, \) and \( \tilde{R}_a \) defined in (3). Further, the adversarial error depends on the perturbed distributions \( f_0 \) and \( f_1 \). The adversarial error is related to the classifier’s robustness to adversarial attacks, where low adversarial error implies higher robustness. Note that, if a classifier abstains over the whole input space (\( \tilde{R}_a = \mathbb{R}^d \)), then the adversarial error converges to zero, and the classifier achieves maximum possible robustness. Yet, such classifier achieves maximum nominal error.

**Remark 1:** (Intuition behind Definition 1 and 2) Abstaining from making a decision can be better than making a wrong one, yet worse than making a correct one. \( e_{\text{nom}} \) penalizes abstaining (along with misclassification) since the classifier is not performing the required task, which is to make a decision. On the other hand, \( e_{\text{adv}} \) does not penalize abstaining since by abstaining from making a decision on perturbed inputs, the classifier is avoiding an adversarial attack that can lead to classification. Each of these two definitions is a different performance metric, where \( e_{\text{nom}} \) measures the classifier’s nominal performance, while \( e_{\text{adv}} \) measures the classifier’s robustness against adversarial perturbations of the input. Further, these definitions guarantee that abstaining does not yield a unilateral advantage or disadvantage, where the classifier would abstain always or never. We remark that different definitions are possible.

### III. Tradeoff Between Nominal and Adversarial Errors

Ideally, we would like both the nominal error and the adversarial error to be small. However, in this section we show that these errors cannot be minimized simultaneously.

**Theorem 3.1:** (Nominal-adversarial error tradeoff) For classifier (2), let \( R_{a0} = R_0 \cap \tilde{R}_a \) and \( R_{a1} = R_1 \cap \tilde{R}_a \), and let \( \tilde{R}_a \supset R_a \) be another abstain region that is partitioned as \( R_a = R_{a0} \cup R_{a1} \), with \( R_{a0} \supset \tilde{R}_a \) and \( R_{a1} \supset \tilde{R}_a \). Then,

\[
e_{\text{nom}}(R_0, R_1, \tilde{R}_a) < e_{\text{nom}}(R_0, R_1, \tilde{R}_a),
\]

\[
e_{\text{adv}}(R_0, R_1, \tilde{R}_a) > e_{\text{adv}}(R_0, R_1, \tilde{R}_a).
\]

**Proof:** For notational convenience, we denote \( e_{\text{nom}}(R_0, R_1, \tilde{R}_a) \), \( e_{\text{adv}}(R_0, R_1, \tilde{R}_a) \), \( e_{\text{nom}}(R_0, R_1, \tilde{R}_a) \), and \( e_{\text{adv}}(R_0, R_1, \tilde{R}_a) \) by \( e_{\text{nom}}, e_{\text{adv}}, e_{\text{nom}}, \) and \( e_{\text{adv}} \), respectively.

For a classifier as in (2) with abstain region \( \tilde{R}_a \), we can write

\[
\bar{e}_{\text{nom}} = p_0 \int_{R_1 \cap \tilde{R}_a} f_0 dx + p_1 \int_{\tilde{R}_a \setminus R_0} f_1 dx.
\]

Then, we can write

\[
\bar{e}_{\text{nom}} - e_{\text{nom}} = p_0 \int_{R_{a0} \setminus R_0} f_0 dx + p_1 \int_{R_{a1} \setminus R_a} f_1 dx > 0.
\]

Similarly, we can write

\[
\bar{e}_{\text{adv}} - e_{\text{adv}} = -p_0 \int_{R_{a0} \setminus R_0} f_0 dx + p_1 \int_{R_{a1} \setminus R_a} f_1 dx < 0.
\]

As we increase the abstain region from \( R_a \) to \( \tilde{R}_a \), \( e_{\text{nom}} \) strictly increases, while \( e_{\text{adv}} \) strictly decreases, which indicates a tradeoff relation between both errors as we vary the abstain region. Theorem 3.1 implies that there exist a tradeoff between \( e_{\text{nom}} \) and \( e_{\text{adv}} \). Therefore, the classifier’s adversarial robustness can be improved only at the expense of its classification performance. In practice, the classifier’s robustness can be improved by increasing \( R_a \), while the nominal classification performance can be improved by decreasing \( R_a \).

**Remark 2:** (Comparing our tradeoff with the literature) [12]–[14] showed that a tradeoff relation exists between a classifier’s nominal performance and its adversarial robustness. Despite using different frameworks, their performance-robustness tradeoff relation is obtained via tuning the classifier’s boundaries in a way that improves its robustness. In our result, we fix the classifier’s decision boundaries, and include an abstain region that can be tuned to obtain our performance-robustness tradeoff. It is possible that a classifier with an abstain option and a classifier without an abstain option but with different decision boundaries achieve the same \( e_{\text{nom}} \) and \( e_{\text{adv}} \). Although both classifiers achieve same metrics, they are different, where the latter gives an output for all the time, while the former abstains on some inputs.

Next we provide our analysis on how to select the abstain region for the 1-dimensional binary classification problem.

Consider the same binary hypothesis testing problem introduced in Section II, but with a scalar observation space where the observation \( x \in \mathbb{R} \) is distributed under classes \( \mathcal{H}_0 \) and \( \mathcal{H}_1 \) as in (1). In this setup, any classifier can be represented by a partition of the real line by placing decision boundaries at suitable positions (see Fig. 2). Let \( -\infty = y_0 \leq y_1 \leq \cdots \leq y_{n+1} = \infty \) denote \( n \) decision boundaries with \( y = [y_i] \). Then, the classifier regions are

\[
R_0 = \{ z : y_i < z < y_{i+1}, \text{ for } i = 0, 2, ..., n \},
\]

\[
R_1 = \{ z : y_i \leq z \leq y_{i+1}, \text{ for } i = 1, 3, ..., n - 1 \},
\]

\[
R_a = \{ z : y_i - \gamma_i \leq z \leq y_i + \gamma_i, \text{ for } i = 1, 2, ..., n \},
\]

where \( \gamma_i \in \mathbb{R}_{\geq 0} \) for \( i = 1, 2, ..., n \) and \( j = 1, 2 \). Let

\[
\gamma = [\gamma_{11}, \gamma_{12}, \cdots, \gamma_{1i}, \gamma_{2i}, \cdots, \gamma_{n1}, \gamma_{n2}]^T, \quad y_i = y_i - \gamma_{ii},
\]

\( ^3 \)For simplicity and without loss of generality, we assume that \( n \) is even. Further, an alternative configuration of the classifier (2) assigns \( \mathcal{H}_0 \) and \( \mathcal{H}_1 \) to \( R_1 \) and \( R_0 \), respectively. However, we consider only the configuration in (2) without affecting the generality of our analysis.
and \( y_{i2} = y_i + \gamma_{i2} \). Using (4) and (1), we have

\[
e_{\text{nom}}(y, \gamma) = p_0 \left( \sum_{l=1}^{n} (-1)^l \int_{-\infty}^{y_{il}} f_0 dx \right) + p_1 \left( \sum_{l=1}^{n} (-1)^{l+1} \int_{-\infty}^{y_{il}} f_1 dx + 1 \right),
\]

where \( j = \frac{(-1)^l+1}{2} + 1 \) and \( k = \frac{(-1)^{l+1}+1}{2} + 1 \) for \( l = 1, \ldots, n \). Using (6), the adversarial error becomes

\[
e_{\text{adv}}(y, \gamma) = p_0 \left( \sum_{l=1}^{n} (-1)^l \int_{-\infty}^{y_{il}} f_0 dx \right) + p_1 \left( \sum_{l=1}^{n} (-1)^{l+1} \int_{-\infty}^{y_{il}} f_1 dx + 1 \right),
\]

where \( j \) and \( k \) are the same as above. Given a classifier as in (2) with known boundaries \( y \), we are interested in how to select the abstain region, i.e., how to choose \( \gamma \) given \( y \). To this aim, we cast the following optimization problem:

\[
e^*(\gamma) = \min_{\gamma} \ e_{\text{adv}}(y, \gamma)
\]

s.t. \( e_{\text{nom}}(y, \gamma) \leq \zeta \),

where \( \zeta \in [e_{\text{nom}}(y, 0), 1] \). In what follows, we characterize the solution \( \gamma^* \) to (10). We begin by writing the derivative of the errors in (8) and (9) with respect to \( \gamma \):

\[
\frac{\partial e_{\text{nom}}}{\partial \gamma_{il}} = p_q \tilde{f}_q(y_i - \gamma_{i1}), \quad \frac{\partial e_{\text{nom}}}{\partial \gamma_{i2}} = p_r \tilde{f}_r(y_i + \gamma_{i2}),
\]

\[
\frac{\partial e_{\text{adv}}}{\partial \gamma_{i1}} = -p_r \tilde{f}_r(y_i - \gamma_{i1}), \quad \frac{\partial e_{\text{adv}}}{\partial \gamma_{i2}} = -p_q \tilde{f}_q(y_i + \gamma_{i2}),
\]

where \( q = \frac{(-1)^l+1}{2} \) and \( r = \frac{(-1)^{l+1}+1}{2} \) for \( i = 1, \ldots, n \). Note that the derivative of \( e_{\text{nom}} \) with respect to \( \gamma \) is strictly positive, while that of \( e_{\text{adv}} \) is strictly negative. Thus, \( e_{\text{nom}} \) increases while \( e_{\text{adv}} \) decreases as \( \gamma \) increases (i.e., as \( \mathcal{R}_a \) increases), which agrees with the result of Theorem 3.1. Problem (10) is not convex and it might not exhibit a unique solution. The following theorem characterizes a solution \( \gamma^* \) to (10).

**Theorem 3.2:** (Characterizing the solution to the minimization problem) Given classifier (2) with 1-dimensional input and known \( n \) boundaries \( y \), the solution \( \gamma^* \) to problem (10) satisfies the following necessary conditions

\[
\frac{\partial e_{\text{nom}}}{\partial \gamma_{il}} = \zeta, \quad \frac{\partial e_{\text{adv}}}{\partial \gamma_{i1}} = \frac{\partial e_{\text{nom}}}{\partial \gamma_{i1}}, \quad \frac{\partial e_{\text{adv}}}{\partial \gamma_{i2}} = \frac{\partial e_{\text{nom}}}{\partial \gamma_{i2}},
\]

for \( i, j = 1, \ldots, n, i \neq j \), and \( u, v = 1, 2 \), where the derivatives of \( e_{\text{nom}} \) and \( e_{\text{adv}} \) with respect to \( \gamma \) are as in (11).

**Proof:** Defining the Lagrange function of (10)

\[
\mathcal{L}(\gamma, \lambda) = e_{\text{adv}}(y, \gamma) + \lambda(e_{\text{nom}}(y, \gamma) - \zeta),
\]

where \( \lambda \) is the Karush-Kuhn-Tucker (KKT) multiplier. For notational convenience, we denote \( e_{\text{adv}}(y, \gamma) \) and \( e_{\text{nom}}(y, \gamma) \) by \( e_{\text{adv}} \) and \( e_{\text{nom}} \), respectively. The stationarity KKT condition implies \( \frac{\partial \mathcal{L}}{\partial \gamma} = -\lambda \frac{\partial e_{\text{nom}}}{\partial \gamma} \).

Using (15) we write

\[
-\lambda = \frac{\partial e_{\text{adv}}}{\partial \gamma_{i1}} / \frac{\partial e_{\text{nom}}}{\partial \gamma_{i1}} = \frac{\partial e_{\text{adv}}}{\partial \gamma_{i2}} / \frac{\partial e_{\text{nom}}}{\partial \gamma_{i2}},
\]

for \( i, j = 1, \ldots, n, i \neq j \), and \( u, v = 1, 2 \), which gives us (13). The KKT condition for dual feasibility implies that \( \lambda \geq 0 \). However, since we have \( \frac{\partial e_{\text{adv}}}{\partial \gamma_{i1}} \neq 0 \) and \( \frac{\partial e_{\text{nom}}}{\partial \gamma_{i1}} \neq 0 \) from (11), we get from (15) that \( \lambda > 0 \). Further, the KKT condition for complementary slackness implies \( \lambda(e_{\text{nom}} - \zeta) = 0 \). Since \( \lambda > 0 \), then \( e_{\text{nom}} - \zeta = 0 \), which gives us (12).

**Remark 3:** (Location of the abstain region in the observation space) The abstain region in Theorem 3.1 can be located anywhere in the observation space. However, in Theorem 3.2, we assume that the abstain region is located around the decision boundaries. This assumption is fair since the observations near the classifier’s boundaries tend to have low classification confidence and are prone to misclassification. □

We conclude this section with an illustrative example.

**Example 1:** (Classifier with an abstain option for exponential distributions) Consider a 1-D binary hypothesis testing problem, where the observation \( x \in \mathbb{R} \) under classes \( \mathcal{H}_0 \) and \( \mathcal{H}_1 \) follows exponential distributions, i.e., the probability density functions in (1) have the form \( f_i(x) = \rho_i \exp(-\rho_i x) \) over the domain \( x \in \mathbb{R}_{\geq 0} \) with parameter \( \rho_i > 0 \) for \( i = 0, 1 \). We consider a single boundary classifier with an abstain option as in (2), with boundary \( y_i \) and abstain parameters \( \gamma_{11} \) and \( \gamma_{12} \) (see Fig. 2). For simplicity, we model the adversarial manipulations of the observations as perturbation added to the distributions’ parameters. We refer to the perturbed parameters as \( \tilde{\rho}_0 \) and \( \tilde{\rho}_1 \). Using Theorem 3.2:

\[
p_0 \exp(-\tilde{\rho}_0 (y_{i1} - \gamma_{11})) - p_1 \exp(-\tilde{\rho}_1 (y_{i1} + \gamma_{12})) + p_1 = \zeta, \quad p_1^2 \tilde{\rho}_1 p_1 \exp(-\tilde{\rho}_1 (y_{i1} - \gamma_{11}) - p_1 (y_{i1} + \gamma_{12}))
\]

\[
= p_0^2 \tilde{\rho}_0 p_0 \exp(-\tilde{\rho}_0 (y_{i1} + \gamma_{12}) - p_0 (y_{i1} - \gamma_{11})).
\]
choose the optimal abstain region by solving (17) for \( \gamma_{11} \) and \( \gamma_{12} \). A solution of (17) corresponds to a local minima of (10). Note that the constraint (10) is active (see Theorem 3.2), hence we have \( e_{\text{nom}}(y_1, \gamma) = \zeta \). Fig. 3 shows the values of \( e_{\text{adv}} \) obtained by solving (17) for \( \gamma_{11} \) and \( \gamma_{12} \) over the range \( \zeta \in [e_{\text{nom}}(y_1, 0), 1] \) with \( \rho_0 = 1.5, \rho_1 = 0.5, \rho_0 = 1.2, \rho_1 = 0.7, \) and \( p_0 = 0.5, p_1 = 0.5 \). Moreover, Fig. 3 shows the values of \( e_{\text{adv}} \) as a function of \( e_{\text{nom}} \) as \( \gamma_{11} \) and \( \gamma_{12} \) are varied arbitrarily. Both curves show a tradeoff between \( e_{\text{nom}} \) and \( e_{\text{adv}} \) as predicted by Theorem 3.1. Further, at each value of \( e_{\text{nom}} \in (e_{\text{nom}}(y_1, 0), 1) \), we observe that \( e_{\text{adv}}(\zeta) < e_{\text{adv}}(y_1, \gamma) \). □

IV. NUMERICAL EXPERIMENT USING MNIST DATASET

In this section, we illustrate the implications of Theorem 3.1 using the classification of hand-written digits from the MNIST dataset [22]. First, we design and train a classifier with an abstain option. Then, we use Definition 1 and 2 to compute \( e_{\text{nom}} \) and \( e_{\text{adv}} \) for a classifier given the dataset. Finally, we present our numerical results on the MNIST dataset. Although our theoretical results are for binary classification, we show that a tradeoff between \( e_{\text{nom}} \) and \( e_{\text{adv}} \) exists for multi-class classification using the MNIST dataset.

A. Classifier design and training

We design a classifier \( h : \mathbb{X} \rightarrow \mathbb{Y} \) using the Lipschitz-constrained loss minimization scheme introduced in [11]:

\[
\min_{h \in \text{Lip}(\mathbb{X};\mathbb{Y})} \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} L(h(x_i), y_i), \quad (18)
\]

s.t. \lip(h) \leq \alpha,

where \( \mathbb{X} \subset \mathbb{R}^d \) and \( \mathbb{Y} \subset \mathbb{R}^m \) are the respective input and output space, \text{Lip}(\mathbb{X};\mathbb{Y}) \) denotes the space of the Lipschitz continuous maps from \( \mathbb{X} \) to \( \mathbb{Y} \), \( L \) is the loss function of the learning problem, the pair \( \{x_i, y_i\}_{i=1}^{N_{\text{test}}} \) denotes the training dataset of size \( N_{\text{train}} \), with input \( x \in \mathbb{X} \) and output \( y \in \mathbb{Y} \), \lip(h) \) is the Lipschitz constant of classifier \( h \), and \( \alpha \in \mathbb{R}_{\geq 0} \) is the upper bound constraint on the Lipschitz constant. The classifier takes an input image of \( d \) pixels and outputs a vector of probabilities of size \( m \), which is the number of classes. The classifier chooses the class with the highest probability: higher probability implies higher decision confidence. We incorporate an abstain option, where the classifier abstains if the maximum probability is less than a threshold probability \( p_a \). We consider adversarial examples, \( \tilde{x} = x + \delta \), computed as in [11], where \( \delta \in \mathbb{R}^d \) is a bounded perturbation (\( ||\delta||_\infty \leq \xi \) in the direction that induces misclassification.

B. Nominal and Adversarial error

Let \( \mathcal{Z} = \{0, 1, \ldots, m-1\} \) and \( \hat{\mathcal{Z}} = \{0, 1, \ldots, m-1, a\} \) be the sets containing all possible true labels and all possible predicted labels by classifier \( h \), respectively, where \( a \) corresponds to the abstain option. Let \( z_i \in \mathcal{Z} \) and \( \hat{z}_i \in \hat{\mathcal{Z}} \) be the true label and the label predicted by \( h \) for the input \( x_i \), respectively (i.e., \( \hat{z}_i \) is the label that corresponds to the maximum probability in the vector \( h(x_i) \), or label \( a \) if the maximum probability is less than \( p_a \)). Further, let \( \tilde{z}_i \in \hat{\mathcal{Z}} \) be the label predicted by \( h \) for the perturbed input image \( \tilde{x}_i \).

Using Definition 1 and 2 we compute \( e_{\text{nom}} \) and \( e_{\text{adv}} \) for \( h \) with threshold probability \( p_a \) on the testing dataset of size \( N_{\text{test}} \) as,

\[
e_{\text{nom}}(h, p_a) = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \mathbb{1}\{\hat{z}_i \neq z_i\},
\]

\[
e_{\text{adv}}(h, p_a) = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \mathbb{1}\{\hat{z}_i \neq z_i \cap \tilde{z}_i \neq a\}, \quad (19)
\]

where \( \mathbb{1}\{\cdot\} \) denotes the indicator function.

C. Nominal-Adversarial error tradeoff

To show the implications of Theorem 3.1, we train four classifiers on the MNIST dataset using (18) with \( \alpha = 5, 10, 20, \) and \( 300 \), respectively (refer to [11] for details about the training scheme). Then, we compute \( e_{\text{nom}} \) and \( e_{\text{adv}} \) for each classifier using (19) with different values of \( p_a \) and a bound on the perturbation \( \xi = 0.3 \). Fig. 4 shows the numerical results on the testing dataset. Fig. 4(a) shows the tradeoff between \( e_{\text{nom}} \) and \( e_{\text{adv}} \) for all the classifiers, which agrees with Theorem 3.1. Fig. 4(b)-(c) show \( e_{\text{nom}} \) and \( e_{\text{adv}} \) as a function of \( p_a \), respectively, while Fig. 4(d) shows the ratio of the abstain region to the input space, denoted by \( \mathcal{A} \), as a function of \( p_a \). As shown in Fig. 4(d), \( \mathcal{A} \) increases at a low rate from zero to 0.1 for \( p_a \in [0.5, 0.95] \) for the classifier with \( \alpha = 300 \), then it increases at a high rate till it reaches 1 for \( p_a \in (0.95, 1] \). The rate at which \( \mathcal{A} \) increases becomes more uniform as \( \alpha \) decreases, where for the classifier with \( \alpha = 5 \), \( \mathcal{A} \) increases with an almost uniform rate from zero at \( p_a = 0.5 \) to 1 at \( p_a = 1 \). This is because as we decrease \( \alpha \) in (18), the learned function becomes more smooth, and the change of the output probability vector over the input space becomes smoother. As observed in Fig. 4(b) and Fig. 4(c), \( e_{\text{nom}} \) increases, while \( e_{\text{adv}} \) decreases for \( p_a \in [0.5, 1] \).

V. CONCLUSION AND FUTURE WORK

In this work, we include an abstain option in a binary classification problem, to improve adversarial robustness. We propose metrics to quantify the nominal performance of a classifier with an abstain option and its adversarial
robustness. We formally prove that, for any classifier with an abstain option, there exist a tradeoff between its nominal performance and its robustness, thus, the classifier’s robustness can only be improved at the expense of its nominal performance. Further, we provide necessary conditions to design the abstain region that optimizes robustness for a desired nominal performance for 1-dimensional binary classification problem. Finally, we validate our theoretical results on the MNIST dataset, where we show that the tradeoff between performance and robustness also exist for the general multi-class classification problems. This research area contains several unexplored questions including comparing tradeoffs obtained with an abstain option and tradeoffs obtained via tuning the decision boundaries, as well as investigate whether it is possible to improve the tradeoff by tuning the boundaries and the abstain region simultaneously.

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In International Conference on Learning Representations, Banff, Canada, Apr. 2014.

[2] K. Eykholt, I.Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1625–1634, June 2018.

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639):115–118, 2017.

[4] A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, and P. C. W. Kim. Supervised autonomous robotic soft tissue surgery. Science Translational Medicine, 8(337):337ra64–337ra64, 2016.

[5] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba. End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316, 2016.

[6] K. D. Julian, J. Lopez, J. S. Brush, M. P. Owen, and M. J. Kochenderfer. Policy compression for aircraft collision avoidance systems. In Digital Avionics Systems Conference, pages 1–10. IEEE, Sept. 2016.

[7] P. Zhu, J. Isaacs, B. Fu, and S. Ferrari. Deep learning feature extraction for target recognition and classification in underwater sonar images. In IEEE Conference on Decision and Control, pages 2724–2731, Melbourne, Australia, Dec. 2017.

[8] S. Lohr. A lesson of Tesla crashes? Computer vision can’t do it all yet. The New York Times, Online, Sept. 2016.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations, Vancouver Convention Center, BC, Canada, May 2018.

[10] R. Anguluri, A. A. Al Makdah, V. Katewa, and F. Pasqualetti. On the robustness of data-driven controllers for linear systems. In Learning for Dynamics & Control, volume 120 of Proceedings of Machine Learning Research, pages 404–412, San Francisco, CA, USA, June 2020.

[11] V. Krishnan, A. A. Al Makdah, and F. Pasqualetti. Lipschitz bounds and provably robust training by laplacian smoothing. In Advances in Neural Information Processing Systems, volume 33, pages 10924–10935, Vancouver, Canada, Dec. 2020.

[12] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy. In International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 7472–7482, Long Beach, California, USA, June 2019. PMLR.

[13] A. A. Al Makdah, V. Katewa, and F. Pasqualetti. A fundamental performance limitation for adversarial classification. IEEE Control Systems Letters, 4(1):169–174, 2019.

[14] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. In International Conference on Learning Representations, Ernest N. Morial Convention Center, NO, USA, May 2019.

[15] R. Herbei and M. H. Wegkamp. Classification with reject option. The Canadian Journal of Statistics, pages 709–721, 2006.

[16] P. L. Bartlett and M. H. Wegkamp. Classification with a reject option using a hinge loss. Journal of Machine Learning Research, 9(59):1823–1840, 2008.

[17] Y. Geifman and R. E. Yaniv. Selective classification for deep neural networks. In Advances in Neural Information Processing Systems, volume 30, Long Beach Convention Center, CA, USA, Dec. 2017. Curran Associates, Inc.

[18] A. Zaoui, C. Denis, and M. Hebiri. Regression with reject option and application to knn. In Advances in Neural Information Processing Systems, volume 33, pages 20073–20082, Virtual, Dec 2020. Curran Associates, Inc.

[19] C. Laidlaw and S. Feizi. Playing it safe: Adversarial robustness with an abstain option. arXiv preprint arXiv:1911.11253, 2019.

[20] N. Balcan, A. Blum, D. Shamma, and H. Zhang. On the power of abstention and data-driven decision making for adversarial robustness. In International Conference on Learning Representations, Virtual, May 2021.

[21] T. A. Schonhoff and A. A. Giordano. Detection and estimation theory and its applications. Pearson College Division, 2006.

[22] Y. LeCun, C. Cortes, and C. J. C. Burges. The MNIST database of handwritten digits. URL: http://yann.lecun.com/exdb/mnist, 1998.