On The Existence of The Adversarial Bayes Classifier

Pranjal Awasthi
Google Research
New York, NY 10011, USA
pranjalawasthi@google.com

Natalie S. Frank
Courant Institute
New York, NY 10012
nf1066@nyu.edu

Mehryar Mohri
Google Research & Courant Institute
New York, NY 10011, USA
mohri@google.com

Abstract

Adversarial robustness is a critical property in a variety of modern machine learning applications. While it has been the subject of several recent theoretical studies, many important questions related to adversarial robustness are still open. In this work, we study a fundamental question regarding Bayes optimality for adversarial robustness. We provide general sufficient conditions under which the existence of a Bayes optimal classifier can be guaranteed for adversarial robustness. Our results can provide a useful tool for a subsequent study of surrogate losses in adversarial robustness and their consistency properties.

1 Introduction

A key problem with using neural networks is their susceptibility to small perturbations: imperceptible changes to the input at test time may result in an incorrect classification by the network (Szegedy et al., 2013). A slightly perturbed picture of a dog could be misclassified as a hand-blower. The same phenomenon appears with other types of data such as biosequences, text, or speech. This problem has motivated a series of research publications studying the design of *adversarially robust* algorithms, both from an empirical and a theoretical perspective (Szegedy et al., 2013; Biggio et al., 2013; Madry et al., 2017; Schmidt et al., 2018; Athalye et al., 2018; Bubeck et al., 2018b; Montasser et al., 2019).

In the context of classification problems, instead of the standard zero-one loss, an *adversarial zero-one loss* has been adopted which penalizes a classifier not only if it misclassifies an input x but also if it does not maintain the correct x-label in a ϵ-neighborhood around x (Goodfellow et al., 2014; Madry et al., 2017; Tsipras et al., 2018; Carlini and Wagner, 2017). Since optimizing the adversarial zero-one loss is computationally intractable, a common approach for adversarial learning is to use a surrogate loss instead. However, optimizing a surrogate loss over a class of functions may not always lead to a minimizer of the true underlying loss over that class. In the case of the standard zero-one loss, there is a large body of literature identifying conditions under which surrogate losses are *consistent*, that is, minimizing them over the family of all measurable functions leads to minimizers of the true loss (Zhang, 2004; Bartlett et al., 2006; Steinwart, 2005; Lin, 2004). More precisely, as argued by Long and Servedio (2013), it is in fact $\mathcal{H}$*-consistency* that is needed, which is consistency restricted to the hypothesis set under consideration. A surrogate loss may be consistent for the family of all measurable functions but not for the specific family of functions $\mathcal{H}$, and a surrogate loss can be $\mathcal{H}$-consistent for a particular family $\mathcal{H}$, without being consistent for all measurable functions.

When are adversarial surrogate losses $\mathcal{H}$-consistent? This problem is already non-trivial for the standard zero-one loss: while there are well-known results for the consistency of losses for the zero-

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

one loss such as (Bartlett et al., 2006; Steinwart, 2005), these results do not hold for $\mathcal{H}$-consistency. Existing theoretical results for $\mathcal{H}$-consistency assume that the Bayes risk is zero (Long and Servedio, 2013; Zhang and Agarwal, 2020). A similar situation seems to hold for the more complex case of the adversarial loss. Recently, Awasthi et al. (2021) gave a detailed study of $\mathcal{H}$-calibration and $\mathcal{H}$-consistency of surrogates to the adversarial loss and also pointed out some technical issues with some $\mathcal{H}$-consistency claims made in prior work (Bao et al., 2020). These authors presented a number of negative results for adversarial $\mathcal{H}$-consistency and positive results for some surrogate losses which assume realizability. For these positive results, the zero Bayes adversarial loss seems necessary. In fact, the authors show empirically that without the realizability assumption, $\mathcal{H}$-consistency does not hold for a variety of surrogate losses, even when they are $\mathcal{H}$-calibrated.

But when is the Bayes adversarial loss zero? Clearly, the adversarial risk can only be zero if it admits a minimizer, which we call the *adversarial Bayes classifier*. However, it is unclear under what conditions such a classifier exists. This is the primary theoretical question that we study in this work.

We now describe the challenges involved in finding minimizers of the adversarial zero-one loss. Most of the existing work on the study of Bayes optimal classifiers focuses on loss functions such as the zero-one loss that admit the *pointwise optimality* property (Steinwart, 2005; Steinwart et al., 2006). To illustrate this better, consider the case of binary classification where on a given input x, $\eta(x)$ denotes the conditional class probability, that is, $\eta(x) := \mathbb{P}(y = 1 \mid x)$. In this case, it is well-known that the Bayes optimal classifier can be obtained by making optimal predictions per point in the domain: at a point x predict 1 if $\eta(x) \geq \frac{1}{2}$, -1 otherwise. Similar to the notion of a Bayes optimal classifier, an adversarial Bayes optimal classifier is the one that minimizes the adversarial loss. However, an immediate obstacle is that the pointwise optimality property does not hold for adversarial losses.

As an example, consider the case of binary classification and perturbations measured in the ℓ_2 norm. Then, for a given labeled point (x, y) and a perturbation radius ϵ, the adversarial zero-one loss of a classifier f is defined as $\max_{x' \colon \|x'-x\|_2 \leq \epsilon} \mathbb{1}(f(x') \neq y)$. Thus, the loss at a point x cannot be measured simply by inspecting the prediction of the classifier at x. In other words, the construction of an adversarial Bayes optimal classifier necessarily involves arguing about the global patterns in the predictions of the classifier across the entire input domain. As a result, most of the technical tools developed for the study of Bayes optimal classifiers for traditional loss functions are not applicable to the analysis of adversarial loss functions, and new mathematical techniques are required.

The above discussion leads to our second motivation for studying the question of existence of the adversarial Bayes classifier. Insights regarding the structure of the adversarial Bayes optimal classifier could have algorithmic implications. For example, in the case of the standard zero-one loss, many popular learning algorithms seek to approximate the conditional probability of a class at a point because the conditional probability defines the Bayes optimal classifier in this case. Analogously, one could hope to develop new algorithmic techniques for adversarial learning with a better understanding of the properties of adversarial Bayes classifiers. In fact, two recent publications propose this approach (Yang et al., 2020; Bhattacharjee and Chaudhuri, 2020). Although their results do not rely on the existence of the adversarial Bayes classifier, they implicitly make this assumption to make their arguments clearer. Our work provides a rigorous basis for this premise.

A second related concept is *certified robustness*. A point x is certifiably robust for a classifier f and a perturbation radius ϵ if *every* perturbation of radius at most ϵ leaves the class of x unchanged. In this paper, we further study a property which we refer to as *pseudo-certified robustness*, which is necessary for certified robustness. We show that there always exists an adversarial Bayes classifier which satisfies the pseudo-certified robustness condition for a fixed radius at every point. However, a non-trivial classifier cannot be certifiably robust for a fixed radius at every point – specifically, a classifier is not certifiably robust at points within ϵ of the decision boundary. Furthermore, we argue that a classifier that is not pseudo-certifiably robust is typically not optimal. Lastly, Lewicka and Peres (2020) prove that for 2-norm perturbations, the boundary of a pseudo-certifiably robust set is differentiable and has Lipschitz normals.

The concept of certified robustness has algorithmic implications. Cohen et al. (2019) recently showed that after training a classifier, a process called *randomized smoothing* makes the classifier certifiably robust at a point x in the ℓ_2 norm with a radius that depends on the point x. As the adversarial Bayes classifier is pseudo-certifiably robust but not certifiably robust with a fixed radius at every point, one could try to design algorithms which ensure pseudo-certifiable robustness during or after training. Recent works have explored constructing certificates of robustness as well (Raghunathan

et al., 2018; Weng et al., 2018; Zhang et al., 2018; Wong and Kolter, 2018). A better understanding of the adversarial Bayes classifier could help find additional learning algorithms. By studying the existence of the adversarial Bayes classifier, we take a first step towards this broader goal.

We now describe the organization of the paper. Section 2 summarizes related work and Section 3 presents the mathematical formulation of our problem. Section 4 discusses our main result and the proof. Next, Section 5 addresses the measurability issues relating to this problem. Section 6 demonstrates how our techniques might apply to other models of perturbations. Subsequently, in Appendix A, we prove the measurability results stated in Section 5 and, in Appendix B, we prove a variant of Prokhorov's theorem that is essential for our proofs. Next, in Appendix C, we prove one of our key lemmas. Appendicies A, B and C present stand-alone results which do not depend on material elsewhere in the appendix. In Appendix D, we subsequently provide some background material for the results in Appendicies E-G. Next, we prove the rest of our key lemmas in Appendicies E and F. Lastly, Appendix G states and proves two generalizations of our main result.

2 Related Work

Existing theoretical work on adversarial robustness focuses on questions such as adversarial counterparts of VC-dimension and Rademacher complexity (Cullina et al., 2018; Khim and Loh, 2018; Yin et al., 2019; Awasthi et al., 2020), evidence of computational barriers (Bubeck et al., 2018b,a; Nakkiran, 2019; Degwekar et al., 2019) and statistical barriers towards ensuring low adversarial test error (Tsipras et al., 2018).

Cullina et al. (2018) formulate a notion of adversarial VC-dimension, aimed at capturing uniform convergence of robust empirical risk minimization. The authors show that, for linear models, adversarial VC-dimension coincides with the VC-dimension. However, in general, the two could be arbitrarily separate. In a similar vein, Khim and Loh (2018), Yin et al. (2019) and Awasthi et al. (2020) study the Rademacher complexity of adversarially robust losses for binary and multi-class classification. Schmidt et al. (2018) provide an instance of a learning problem where one can provably demonstrate a gap between the sample complexity of (standard) learning and adversarial learning.

Tsipras et al. (2018) points out a problem where any learning algorithm that achieves low (standard) test error must necessarily admit high adversarial test error, that is close to 1. This highlights a fundamental tension between ensuring low test error and low adversarial error. There are also studies of the conditions on the data distribution that lead to the presence of adversarial examples and the design of adversaries that can exploit them (Diochnos et al., 2018; Bartlett et al., 2021). The recent work of Montasser et al. (2019) shows that any function class with finite VC-dimension d can be adversarially robustly learned (in a PAC-style model) using $\exp(d)$ many samples.

Bubeck et al. (2018b,a) provide evidence of computational barriers in adversarial learning by constructing learning tasks that are easy in the PAC model, but that become intractable when adversarial robustness is required. Several recent publications have studied the question of characterizing the Bayes adversarial risk (Pydi and Jog, 2019; Bhagoji et al., 2019) for binary classification and relate it to the optimal transportation cost between the two class conditional distributions. While these studies aim to establish a lower bound on the Bayes adversarial risk, we study a more fundamental question of when the Bayes adversarial classifier exists. There have also been publications studying robustness beyond ℓ_p norm perturbations (Feige et al., 2015, 2018; Attias et al., 2018).

Finally, there are studies in the mathematical community of various properties regarding the direct sum of a set and an ϵ-ball, which we use to model adversarial perturbations. Similar, but not identical mathematical constructions have also appeared in the PDE literature. Cesaroni and Matteo (2017) and Cesaroni et al. (2018) consider perturbations to the measure-theoretic boundary of a set. However, the measure-theoretic boundary and the topological boundary behave quite differently. Chambolle et al. (2012) consider problems involving integrals of indicator functions of perturbed sets A^ϵ divided by the size of the perturbation. Additionally, Bellettini (2004) and Chambolle et al. (2015) assume some set properties that are satisfied by sets perturbed by ℓ_p balls, and then use these to show regularity and the curvature of the boundary. Lastly, Bertsekas and Shreve (1996) study the universal σ-algebra in detail, however they did not show that the sets we use in this paper are universally measurable. We prove a new measurability result in Section 5.

3 Problem Setup

We study binary classification with class labels in $\{-1, +1\}$. We consider a probability distribution $\mathcal{D}$ over $\mathbb{R}^d \times \{-1, +1\}$. For convenience, we denote by η the conditional distribution, $\eta(\mathbf{x}) = \mathcal{D}(Y = +1|\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$, and by $\mathbb{P}$ the marginal, $\mathbb{P}(A) = \mathcal{D}(A \times \{-1, +1\})$ for any measurable set $A \subseteq \mathbb{R}^d$. Let $f \colon \mathbb{R}^d \to \mathbb{R}$ be a function whose sign defines a classifier. Then, for a perturbation set B, the *adversarial loss* of f is defined as

$$R^\epsilon(f) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\sup_{\mathbf{h}\in B} \mathbb{1}_{y \operatorname{sign}(f(\mathbf{x}+\mathbf{h}))<0} \right] \quad \text{where} \quad \operatorname{sign}(z) = \begin{cases} +1 & \text{if } z > 0 \\ -1 & \text{otherwise} \end{cases}.$$

The adversarial loss has been extensively studied in recent years (Montasser et al., 2019; Tsipras et al., 2018; Bubeck et al., 2018b; Khim and Loh, 2018; Yin et al., 2019), motivated by the empirical phenomenon of adversarial examples (Szegedy et al., 2013). In the rest of the paper, we will find it more convenient to work with an alternative set-based definition of classifiers (and adversarial losses), which we describe below. The function f induces two complementary sets $A = \{\mathbf{x} \colon f(\mathbf{x}) > 0\}$ and $A^C = \{\mathbf{x} \colon f(\mathbf{x}) \le 0\}$. Conversely, specifying the set A is equivalent to specifying a function f since we could choose $f(\mathbf{x}) = \mathbb{1}_A(\mathbf{x})$. In the rest of the paper, we will specify the set of points A classified as $+1$ rather than the function f. The classification risk of a set A is then expressed as

$$R(A) = \int (1 - \eta(\mathbf{x}))\, \mathbb{1}_A(\mathbf{x}) + \eta(\mathbf{x})\, \mathbb{1}_{A^C}(\mathbf{x})\, d\mathbb{P}. \tag{1}$$

In the above formulation, it is easy to see that the Bayes optimal classifier is the set $A = \{\mathbf{x} \colon \eta(\mathbf{x}) > \frac{1}{2}\}$. We now extend this viewpoint to adversarial losses. We assume that the adversary knows the classification set A and that the adversary seeks to perturb each point in $\mathbb{R}^d$ outside of A, via an additive perturbation in a set B. In typical applications, B is a ball in some norm, and in the rest of the paper we will assume that $B = \overline{B_\epsilon(\mathbf{0})}$ is a closed ball with radius ϵ centered at the origin. Next, we define A^ϵ to be the set of points that can fall inside A after an additive perturbation of magnitude at most ϵ. Formally, $A^\epsilon = \{\mathbf{x} \in \mathbb{R}^d \colon \exists \mathbf{h} \in \overline{B_\epsilon(\mathbf{0})} \text{ for which } \mathbf{x} + \mathbf{h} \in A\}$. Therefore, we can define the adversarial risk as

$$R^\epsilon(A) = \int (1 - \eta(\mathbf{x}))\, \mathbb{1}_{A^\epsilon}(\mathbf{x}) + \eta(\mathbf{x})\, \mathbb{1}_{(A^C)^\epsilon}(\mathbf{x})\, d\mathbb{P}. \tag{2}$$

Pydi and Jog (2019); Bhagoji et al. (2019) also studied the adversarial Bayes classifiers using the $^\epsilon$ operation. We will now re-write A^ϵ in a form more amenable to analysis:

$$\begin{aligned} A^\epsilon &= \{\mathbf{x} \in \mathbb{R}^d \colon \exists \mathbf{h} \in \overline{B_\epsilon(\mathbf{0})} | \mathbf{x} + \mathbf{h} \in A\} = \{\mathbf{x} \in \mathbb{R}^d \colon \exists \mathbf{h} \in \overline{B_\epsilon(\mathbf{0})} \text{ and } \mathbf{a} \in A | \mathbf{x} + \mathbf{h} = \mathbf{a}\} \\ &= \left\{\mathbf{x} \colon \exists \mathbf{h} \in \overline{B_\epsilon(\mathbf{0})} \text{ and } \mathbf{a} \in A \mid \mathbf{a} - \mathbf{h} = \mathbf{x}\right\} = \{\mathbf{a} - \mathbf{h} \colon \mathbf{a} \in A, \mathbf{h} \in \overline{B_\epsilon(\mathbf{0})}\} = A \oplus \overline{B_\epsilon(\mathbf{0})}, \end{aligned}$$

where the last equality follows from the symmetry of the ball $\overline{B_\epsilon(\mathbf{0})}$. From these relations, we can recover a more typical expression of the adversarial loss. Note that $\mathbb{1}_{A^\epsilon}(\mathbf{x}) = \mathbb{1}_{A \oplus \overline{B_\epsilon(\mathbf{0})}}(\mathbf{x}) = \sup_{\mathbf{h}\in\overline{B_\epsilon(\mathbf{0})}} \mathbb{1}_A(\mathbf{x} + \mathbf{h})$, which implies

$$R^\epsilon(A) = \int (1 - \eta(\mathbf{x})) \sup_{\mathbf{h}\in\overline{B_\epsilon(\mathbf{0})}} \mathbb{1}_A(\mathbf{x} + \mathbf{h}) + \eta(\mathbf{x}) \sup_{\mathbf{h}\in\overline{B_\epsilon(\mathbf{0})}} \mathbb{1}_{A^C}(\mathbf{x} + \mathbf{h})\, d\mathbb{P}. \tag{3}$$

The papers (Szegedy et al., 2013; Biggio et al., 2013; Madry et al., 2017) (and many others) use the multi-class version of this loss to define adversarial risk. More specifically, they evaluate the risk on the set $A = \{f(\mathbf{x}) \ge 0\}$, where f is a function in their model class.

We define the *adversarial Bayes risk* R^ϵ_* as the infimum of (2) over all measurable sets, and we say that the set A is an adversarial Bayes classifier if $R^\epsilon(A) = R^\epsilon_*$. Note that the integral above is defined only if the sets $A^\epsilon, (A^C)^\epsilon$ are measurable. This consideration is nontrivial as there do exist measurable sets whose direct sum is not measurable, see (Erdös and Stone, 1970; Ciesielski et al., 2001/2002) for examples.

To address this issue, in Section 5, we discuss a σ-algebra called the *universal σ-algebra* which is denoted $\mathscr{U}(\mathbb{R}^d)$. Specifically, we show that if $A \in \mathscr{U}(\mathbb{R}^d)$, then $A^\epsilon \in \mathscr{U}(\mathbb{R}^d)$ as well. Thus, working in the universal σ-algebra $\mathscr{U}(\mathbb{R}^d)$ allows us to define the integral in (2) and then optimize R^ϵ over sets in $\mathscr{U}(\mathbb{R}^d)$. In particular, throughout this paper, we adopt the convention that $\mathbb{P}$ is the completion of a Borel measure restricted to $\mathscr{U}(\mathbb{R}^d)$. (We elaborate on this construction in Section 5.) We call a set *universally measurable* if it is in the universal σ-algebra $\mathscr{U}(\mathbb{R}^d)$.

We now introduce another important notation: we define $A^{-\epsilon} := ((A^C)^\epsilon)^C$. The set $A^{-\epsilon}$ contains the points that cannot be perturbed to fall outside of A. Figure 1 depicts the sets A, A^ϵ and $A^{-\epsilon}$.

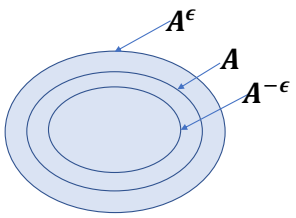

Figure 1: Sets A^ϵ and $A^{-\epsilon}$ with $B = \overline{B_\epsilon^2(\mathbf{0})}$, the closed ℓ_2 ball.

4 Main Results

In this section, we prove our main result establishing the existence of the optimal adversarial classifier. We first discuss challenges in establishing this theorem. In the case of the standard 0-1 loss, the risk is defined in (1) where the sets A and A^C are disjoint. As a result, the integrand equals either $\eta(\mathbf{x})$ or $(1-\eta(\mathbf{x}))$ at each point. Thus the set for which $1-\eta(\mathbf{x}) < \eta(\mathbf{x})$ minimizes R. In other words, the Bayes classifier minimizes the objective $\min(\eta(\mathbf{x}), 1-\eta(\mathbf{x}))$ at each point.

On the other hand, the same reasoning does not apply to the adversarial risk. The adversarial risk at a single point $\mathbf{x}$ depends on all the points in $\overline{B_\epsilon(\mathbf{x})}$. Hence, one cannot hope to find the adversarial Bayes classifier by studying the risk in a pointwise manner.

Next, we introduce the concepts of certifiable robustness and pseudo-certifiable robustness.

Definition 1. *Fix a perturbation radius ϵ. We say that a classifier A is* certifiably robust *at a point $\mathbf{x}$ with radius ϵ if either $\mathbf{x} \in A$ and $B_\epsilon(\mathbf{x}) \subset A$, or $\mathbf{x} \in A^C$ and $B_\epsilon(\mathbf{x}) \subset A^C$. We say that a classifier A is* pseudo-certifiably robust *at a point $\mathbf{x}$ with radius ϵ if either $\mathbf{x} \in A$ and there exists a ball $B_\epsilon(\mathbf{y})$ with $\mathbf{x} \in \overline{B_\epsilon(\mathbf{y})}$ and $B_\epsilon(\mathbf{y}) \subset A$ or $\mathbf{x} \in A^C$ and there exists a ball $B_\epsilon(\mathbf{y})$ with $\mathbf{x} \in \overline{B_\epsilon(\mathbf{y})}$ and $B_\epsilon(\mathbf{y}) \subset A^C$. We say a classifier A is* pseudo-certifiably robust *if it is pseudo-certifiably robust with radius ϵ at every point.*

In other words, a classifier is certifiably robust at a point $\mathbf{x}$ with radius ϵ if the entire ϵ-ball around $\mathbf{x}$ is classified the same as $\mathbf{x}$, and a classifier is pseudo-certifiably robust at a point $\mathbf{x}$ with radius ϵ if *some* ball radius ϵ whose closure contains $\mathbf{x}$ is classified the same as $\mathbf{x}$. Pseudo-certifiable robustness is a necessary condition for certifiable robustness.

We now discuss potential algorithmic applications of pseudo-certifiable robustness. To begin, we start by defining the set of points at which a classifier is not pseudo-certifiably robust. If we define

$$F(A) = \{\mathbf{x} \in A : \text{ every closed } \epsilon\text{-ball containing } \mathbf{x} \text{ also intersects } A^C\}. \tag{4}$$

Then, the set of points where a classifier is not pseudo-certifiably robust is $F(A) \cup F(A^C)$. In Appendix E, we show that if we "subtract" from a classifier the points at which it is not pseudo-certifiably robust, then we get a classifier with lower risk. Formally, we show that $R^\epsilon(A - F(A)) \le R^\epsilon(A)$ and $R^\epsilon(A \cup F(A^C) \le R^\epsilon(A)$ (Lemma 27). Furthermore, Lemma 27 suggests that near the adversarial Bayes classifier, these inequalities are typically strict. As illustrated in Figure 2, $F(A), F(A^C)$ are adjacent to the boundary ∂A. Furthermore, $F(A)$ is not very "large" – $F(A)^{-\epsilon} = \emptyset$. These observations suggest that, typically, if A is not pseudo-certifiably robust, then there is another classifier with lower risk that can be found by making local changes to A.

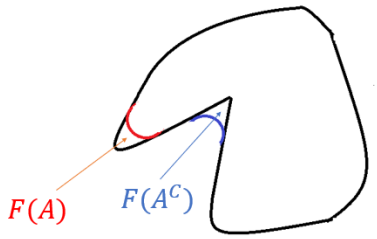

Figure 2: The figure illustrates a set A with the sets $F(A)$ and $F(A^C)$ roughly indicated. For a point $\mathbf{a} \in F(A)$, every closed ϵ-ball containing $\mathbf{a}$ also intersects A^C while for $\mathbf{a} \in F(A^C)$ every closed ϵ-ball containing $\mathbf{a}$ also intersects A.

We now state our main existence result.

Theorem 1. *Let $\mathbb{P}$ be the completion of a Borel measure on $\mathcal{B}(\mathbb{R}^d)$ restricted to $\mathscr{U}(\mathbb{R}^d)$. Let $B_\epsilon(\mathbf{0})$ be a ball for a norm for which the unit ball is strictly convex or a polytope. Define $A^\epsilon = A \oplus \overline{B_\epsilon(\mathbf{0})}$. Then, there exists a minimizer of (2) when minimizing over $\mathscr{U}(\mathbb{R}^d)$. Furthermore, this minimizer is pseudo-certifiably robust.*

For perturbations in many common norms, such as the ℓ_1, ℓ_2, and ℓ_∞ norms, the theorem provides a positive guarantee: for any distribution, the adversarial Bayes classifier exists. In fact, an even stronger result holds: if $\mathbb{P}$ is absolutely continuous with respect to the Lebesgue measure, we can show a statement analogous to every minimizing sequence of R^ϵ has a convergent subsequence. We formally state and prove this stronger version of our theorem as Theorem 12 in Appendix G.

This result actually holds for all norms. However, the extension of one of our lemmas (Lemma 2) to all norms is not straightforward, and thus we are leaving this result to an extended version of the paper. We also expect an existence result for perturbations by open balls.

Next, we briefly discuss two ways in which our results relate to the consistency of adversarial losses. First, Awasthi et al. (2021) show that when H is the class of linear functions, if the surrogate risk R^ϵ_Ψ of the adversarial surrogate loss Ψ is zero for a given distribution, then Ψ is H-consistent for that distribution. Furthermore, (Awasthi et al., 2021) give an example of a distribution for which the adversarial loss is nonzero and no continuous surrogate losses can be consistent. The existence of the adversarial Bayes classifier is required for this condition to hold. Next, a surrogate loss Ψ is consistent if a minimizing sequence of functions f_i also minimizes 0-1 adversarial loss. However, it may be easier to study minimizing sequences of the Ψ loss when we have information about the adversarial Bayes classifier. Theorem 12 in the appendix lists a variety of conditions under which a minimizing sequence of the adversarial loss approaches an adversarial Bayes classifier in a meaningful sense. Thus, we can find conditions under which $\{\mathbf{x}\colon f_i(\mathbf{x}) \geq 0\}$ approaches a set A. In other words: If Ψ is consistent and f_i is a sequence that minimizes the adversarial ϕ loss, then $f_i \geq 0$ must approach an adversarial Bayes classifier in the sense described by Theorem 12.

4.1 Proof strategy

We first outline the main ideas behind the proof of Theorem 1, which is presented in the next subsection. The proof applies the direct method of the calculus of variations, with an additional step (2a below). Specifically, we apply the following procedure:

1) Choose a sequence of sets $\{A_n\} \subset \mathscr{U}(\mathbb{R}^d)$ along which $R^\epsilon(A_n)$ approaches its infimum;
2a) Using the sequence $\{A_n\}$, we find a decreasing minimizing sequence $\{B_n\}$ with nice properties
2b) Extract a subsequence $\{B_{n_k}\}$ of $\{B_n\}$ that is convergent in some topology;
3) Show that R^ϵ is sequentially lower semi-continuous: for a convergent subsequence $\{A_n\}$,

$$\liminf_{n\to\infty} R^\epsilon(A_n) \geq R^\epsilon(\lim_{n\to\infty} A_n).$$

Classically, the direct method of the calculus of variations consists of 1), 2b) and 3). In typical applications of the direct method, step 2) is almost immediate as it is achieved by working in the appropriate Sobolev space. However, showing step 3) is usually quite difficult. See Dacorogna (2008) for more on the direct method in PDEs. In contrast, in our scenario, the situation is the opposite: finding the right topology for step 2) is quite difficult but the lower semi-continuity is a direct implication of the dominated convergence theorem.

As described above, one of the main considerations in the proof of Theorem 1 is the convergence of set sequences. In order to apply the dominated convergence theorem, we need the indicator functions $\mathbb{1}_{(A_n)^\epsilon}, \mathbb{1}_{(A_n^C)^\epsilon}$ to converge. With that in mind, we adopt the following standard set-theoretic definitions for a sequence of sets $\{A_n\}$:

$$\limsup A_n = \bigcap_{N\geq 1} \bigcup_{n\geq N} A_n \text{ and } \liminf A_n = \bigcup_{N\geq 1} \bigcap_{n\geq N} A_n.$$

See (Rockafellar and Wets, 1998) for further discussion of the intuition behind these definitions. As with $\limsup$ and $\liminf$ for a sequences of numbers, $\liminf A_n \subset \limsup A_n$ or in other words $\mathbb{1}_{\liminf A_n} \leq \mathbb{1}_{\limsup A_n}$. With the above definitions, the following holds:

$$\liminf_{n\to\infty} \mathbb{1}_{A_n} = \mathbb{1}_{\liminf A_n} \text{ and } \limsup_{n\to\infty} \mathbb{1}_{A_n} = \mathbb{1}_{\limsup A_n}.$$

Specifically, these relations imply that the limit $\lim_{n\to\infty} \mathbb{1}_{A_n} d\mathbb{P}$ exists a.e. if and only if the $\limsup$ and the $\liminf$ of the sequence $\{A_n\}$ match up to sets of measure zero under $\mathbb{P}$. We denote equality up to sets of measure zero by $\doteq$. In order to find a sequence for which $\liminf A_n \doteq \limsup A_n$, we first define the measures $\{\mathbb{P}_n\}$ by $\mathbb{P}_n(B) = \mathbb{P}(A_n \cap B)$. The hope is that if $\mathbb{P}_n$ converges to a measure $\mathbb{Q}$, this would imply that $\liminf A_n \doteq \limsup A_n$.

To this end, we apply Prokhorov's theorem to obtain a subsequence $\{\mathbb{P}_{n_k}\}$ of $\{\mathbb{P}_n\}$ that converges to a measure $\mathbb{Q}$ in some sense. The notion of convergence for probability measures discussed in Prokhorov's theorem is that of *weak convergence*. In order to extract a sequence A_n for which the $\liminf$ and the $\limsup$ match, we apply the following lemma to the sequence of measures $\mathbb{P}_{n_k}$.

Lemma 1. *Let $\{\mathbb{P}_n\}, \mathbb{Q}$ be measures on $\mathbb{R}^d$. Assume that $\mathbb{P}_n$ weakly converges to $\mathbb{Q}$ with $\mathbb{P}_n$ given by $\mathbb{P}_n(B) = \mathbb{P}(B \cap A_n)$ for a sequence of sets A_n. Then $\mathbb{Q}(B) = \mathbb{P}(A \cap B)$ for a set A given by*

$$A \doteq \limsup A_{n_j} \doteq \liminf A_{n_j},$$

where $\{A_{n_j}\}$ is some subsequence of A_n. Furthermore, $\mathbb{1}_{A_{n_j}} \to \mathbb{1}_A$ $\mathbb{P}$-a.e.

The lemma above is proved in Appendix C. The next challenge is that $\liminf A_n^\epsilon / \limsup A_n^\epsilon$ do not necessarily equal A^ϵ for some set A. However, finding a sequence satisfying this property is not too difficult if the sequence A_n is in fact decreasing, and $\overline{B^\epsilon(\mathbf{0})}$ is either strictly convex or a polytope.

Lemma 2. *Let B_n be a decreasing sequence $(B_{n+1} \subset B_n)$. Let $A^\epsilon = A \oplus \overline{B_\epsilon(\mathbf{0})}$, where $\overline{B_\epsilon(\mathbf{0})}$ is a strictly convex set or a polytope. Then, there exists another decreasing sequence C_n with $R^\epsilon(C_n) \leq R^\epsilon(B_n)$ for which $\bigcap_{n=1}^\infty C_n$ is pseudo-certifiably robust at every point and satisfies*

$$\bigcap_{n=1}^\infty C_n^\epsilon = \left(\bigcap_{n=1}^\infty C_n\right)^\epsilon, \bigcap_{n=1}^\infty C_n^{-\epsilon} = \left(\bigcap_{n=1}^\infty C_n\right)^{-\epsilon}.$$

The lemma is proved in Appendix E. Note that for decreasing sequences of sets, $\liminf C_n = \limsup C_n = \bigcap_{n\geq 1} C_n$. Thus, using the sequence of sets given by Lemma 2, one can swap the order of the $\lim$, $^\epsilon$, and $^{-\epsilon}$ operations to conclude

$$\liminf C_n^\epsilon = (\liminf C_n)^\epsilon \text{ and } \liminf C_n^{-\epsilon} = (\liminf C_n)^{\dot{-}\epsilon}$$

Finally, it remains to show that we can actually apply Lemma 2. This step requires proving that one can find a decreasing minimizing sequence B_n. Subsequently, the inequality $R^\epsilon(C_n) \leq R^\epsilon(B_n)$ of Lemma 2 implies that C_n must be a minimizing sequence when B_n is a minimizing sequence.

Lemma 3. *Let A_n be a minimizing sequence of R^ϵ for which $\liminf A_n^\epsilon \doteq \limsup A_n^\epsilon$ and $\liminf A_n^{-\epsilon} = \limsup A_n^{-\epsilon}$. Then, there is a decreasing minimizing sequence B_n, i.e., $B_{n+1} \subset B_n$.*

We prove the above Lemma in Appendix F. Lemma 1 is used to satisfy the conditions of the lemma.

4.2 Formal Proof of Theorem 1

We now formally prove Theorem 1. We start by introducing three basic tools: weak convergence of probability measures, Prokhorov's theorem, and inner regularity. We start with weak convergence.

Definition 2. *A sequence of measures $\mathbb{Q}_n$* converges weakly *to a measure $\mathbb{Q}$ if for all continuous and bounded functions f, $\lim_{n\to+\infty} \int f d\mathbb{Q}_n = \int f d\mathbb{Q}$.*

Given a sequence $\{\mathbb{P}_n\}$, Prokohrov's theorem allows one to extract a weakly convergent subsequence.

Theorem 2 (Prokohrov's Theorem)**.** *Consider $\mathbb{R}^d$ with the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$. A sequence of probability measures $\{\mathbb{P}_n\}$ on the $\mathcal{B}(\mathbb{R}^d)$ admits a weakly convergent subsequence iff for all $\epsilon > 0$, there exists a compact set K for which the condition $\mathbb{P}_n(\mathbb{R}^d \setminus K) < \epsilon$ holds uniformly for all n.*

If for the sequence $\{\mathbb{P}_n\}$, for all $\epsilon > 0$, there exists a compact set K for which $\mathbb{P}_n(\mathbb{R}^d \setminus K) < \epsilon$ uniformly for all n, then the sequence $\mathbb{P}_n$ is referred to as *tight*. However, as discussed in Section 3, the σ-algebra $\mathscr{U}(\mathbb{R}^d)$ which we work with is larger than $\mathcal{B}(\mathbb{R}^d)$, the σ-algebra present in Theorem 2. To address this technicality, we state a variant of Prokhorov's Theorem which we prove in Appendix B.

Corollary 1 (Prokhorov's Theorem)**.** *Let $(\mathbb{P}_n, \mathbb{R}^d, \mathscr{U}(\mathbb{R}^d))$ be a sequence of probability measure spaces for which each $\mathbb{P}_n$ is the completion of a Borel measure restricted to $\mathscr{U}(\mathbb{R}^d)$. Then the sequence of measures $\{\mathbb{P}_n\}$ admits a weakly convergent subsequence iff the sequence is tight.*

In order to demonstrate that Prokhorov's theorem applies, we use the concept of inner regularity.

Definition 3. *Let $\mathbb{P}$ be a Borel measure on $\mathbb{R}^d$ or its completion. We say that $\mathbb{P}$ on $\mathbb{R}^d$ is inner regular if $\mathbb{P}(E) = \sup\{\mathbb{P}(K) \colon K \subset E, K \text{ compact}\}$.*

The following Lemma states that all probability measures on $\mathbb{R}^d$ are inner regular.

Lemma 4. *Every Borel measure ν on $\mathbb{R}^d$ with $\nu(X) < \infty$ is inner regular.*

The above lemma is a consequence of Theorem 7.8 and Proposition 7.5 of Folland (1999) and further implies that the completion of every Borel measure on $\mathbb{R}^d$ restricted to $\mathscr{U}(\mathbb{R}^d)$ is inner regular.

Proof of Theorem 1. Let A_n be universally measurable minimizing sequence. Consider two sequences of measures given by $\mathbb{P}_n^1(B) = \mathbb{P}(A_n^\epsilon \cap B)$ and $\mathbb{P}_n^2(B) = \mathbb{P}(A_n^{-\epsilon} \cap B)$. Since $\mathbb{P}$ is inner regular, by the comment after Lemma 4, both of these sequences are tight. Furthermore, each $\mathbb{P}_n$ is defined on the universal σ-algebra. Thus, we can apply Prokhorov's Theorem in the form of Corollary 1 to extract weakly convergent subsequences of $\mathbb{P}_n^i$. In fact, by diagonalization, we can choose the same subsequence for both measures. Specifically, using Prohkorov's Theorem, we choose a weakly convergent subsequence $\{\mathbb{P}^1_{n_k}\}$. Note that the subsequence $\{\mathbb{P}^2_{n_k}\}$ is also tight. This means that we can choose another weakly convergent subsequence $\{\mathbb{P}^2_{n_{k_m}}\}$. Therefore both $\{\mathbb{P}^1_{n_{k_m}}\}$ and $\{\mathbb{P}^2_{n_{k_m}}\}$ are weakly convergent.

To simplify notation, we drop the triple subscript and let A_n denote a sequence of sets for which $\mathbb{P}_n^1$ weakly converges to $\mathbb{Q}^1$ and $\mathbb{P}_n^2$ weakly converges to $\mathbb{Q}^2$. Next we use another diagonalization argument. By Lemma 1, we have

$$\mathbb{Q}^1(B) = \mathbb{P}(C \cap B) \quad \text{with} \quad C \doteq \limsup A^\epsilon_{n_j} \doteq \liminf A^\epsilon_{n_j}$$

for a subsequence A_{n_j} of $\{A_n\}$. Note that for any subsequence $A_{n_{j_k}}$, it still holds that

$$\mathbb{Q}^1(B) = \mathbb{P}(C \cap B) \quad \text{with} \quad C \doteq \limsup A^\epsilon_{n_{j_k}} \doteq \liminf A^\epsilon_{n_{j_k}}.$$

This statement holds because for any sequence of functions $\{f_j\}$ and any subsequence $\{f_{j_k}\}$, $\limsup_{k\to\infty} f_{j_k} \leq \limsup_{j\to\infty} f_j$ and $\liminf_{k\to\infty} f_{j_k} \geq \liminf_{k\to\infty} f_{j_k}$. Thus we can apply Lemma 1 to the sequence $\mathbb{P}^2_{n_j}$ to extract a subsequence of indices $\{n_{j_k}\}$ for which

$$\begin{aligned}&\mathbb{Q}^1(B) = \mathbb{P}(C \cap B) \text{ and } \mathbb{Q}^2(B) = \mathbb{P}(D \cap B)\\ &\text{with} \quad C \doteq \limsup A^\epsilon_{n_{j_k}} \doteq \liminf A^\epsilon_{n_{j_k}} \text{ and } \quad D \doteq \limsup A^{-\epsilon}_{n_{j_k}} \doteq \liminf A^{-\epsilon}_{n_{j_k}}.\end{aligned}$$

Note that Lemma 1 further implies that the convergence is $\mathbb{P}$-a.e., not just weak convergence. Again, for clarity, we drop the triple subscript and refer to $A_{n_{j_k}}$ as A_n. Subsequently, Lemma 3 gives a decreasing minimizing sequence B_n. Next, Lemma 2 produces a decreasing sequence C_n for which

$$\bigcap_{n=1}^\infty C_n^\epsilon = \left(\bigcap_{n=1}^\infty C_n\right)^\epsilon \text{ and } \bigcap_{n=1}^\infty C_n^{-\epsilon} = \left(\bigcap_{n=1}^\infty C_n\right)^{-\epsilon}$$

and $R^\epsilon(C_n) \leq R^\epsilon(B_n)$. Since B_n is a minimizing sequence, C_n must be a minimizing sequence as well. Now, pick $A = \bigcap_{n=1}^\infty C_n$. An application of the dominated convergence theorem then gives

$$\begin{aligned}\inf_{S\in\mathscr{U}(\mathbb{R}^d)} R^\epsilon(S) &= \lim_{n\to\infty} \int (1-\eta(\mathbf{x}))\mathbb{1}_{C_n^\epsilon} + (1-\eta(\mathbf{x}))\mathbb{1}_{(C_n^{-\epsilon})^C}\\ &= \int \lim_{n\to\infty} \left((1-\eta(\mathbf{x}))\mathbb{1}_{C_n^\epsilon} + \eta(\mathbf{x})(1-\mathbb{1}_{C_n^{-\epsilon}})\right) d\mathbb{P}\\ &= \int (1-\eta(\mathbf{x}))\mathbb{1}_{\bigcap_{n=1}^\infty C_n^\epsilon} + \eta(\mathbf{x})\left(1-\mathbb{1}_{\bigcap_{n=1}^\infty C_n^{-\epsilon}}\right) d\mathbb{P}\\ &= \int (1-\eta(\mathbf{x}))\mathbb{1}_{(\bigcap_{n=1}^\infty C_n)^\epsilon} + \eta(\mathbf{x})\left(1-\mathbb{1}_{(\bigcap_{n=1}^\infty C_n)^{-\epsilon}}\right) d\mathbb{P}\\ &= \int (1-\eta(\mathbf{x}))\mathbb{1}_{A^\epsilon} + \eta(\mathbf{x})\mathbb{1}_{(A^{-\epsilon})^C} d\mathbb{P} = R^\epsilon(A).\end{aligned}$$

Thus, we have found a minimizer of R^ϵ. Lastly, by Lemma 2 A is pseudo-certifiably robust. $\square$

4.3 Proof Strategy for Lemma 2

In this section, we explain the intuition for the proof of Lemma 2. Appendix E presents the formal proofs. We begin by studying sets of the form A^ϵ. From $A^\epsilon = \bigcup_{\mathbf{a}\in A} \overline{B_\epsilon(\mathbf{a})}$, one can show

$$\bigcap_{n\geq 1} A_n^{-\epsilon} = \left(\bigcap_{n\geq 1} A_n\right)^{-\epsilon} \tag{5}$$

for any sequence of sets $\{A_n\}$. We prove (5) formally in Appendix D. Thus, it remains to find a sequence C_n for which $\bigcap_{n=1}^\infty C_n^\epsilon = (\bigcap_{n=1}^\infty C_n)^\epsilon$. With this observation in mind, we first study properties of sets of the form A^ϵ. Notably, as $A^\epsilon = \bigcup_{\mathbf{a}\in A} \overline{B_\epsilon(\mathbf{a})}$, every point $\mathbf{x} \in A^\epsilon$ is contained in some ball $\overline{B_\epsilon(\mathbf{a})}$ which is completely included in A^ϵ. Thus, a necessary condition for $\bigcap_{n=1}^\infty C_n^\epsilon = (\bigcap_{n=1}^\infty C_n)^\epsilon$ is that every point in $\bigcap_{n=1}^\infty C_n^\epsilon$ must be contained in some ball $\overline{B_\epsilon(\mathbf{a})}$ which is completely included in $\bigcap_{n=1}^\infty C_n^\epsilon$. In the proof of Lemma 2, we show this condition is also sufficient when one chooses $C_n = ((B_n^{-\epsilon})^{2\epsilon})^{-\epsilon}$. We start by showing $(A^{-\epsilon})^\epsilon = A - F(A)$. Subsequently, applying (5),

$$\begin{aligned}\left(\bigcap_{n=1}^\infty C_n\right)^\epsilon &= \left(\bigcap_{n=1}^\infty \left(\left(B_n^{-\epsilon}\right)^{2\epsilon}\right)^{-\epsilon}\right)^\epsilon \\ &= \left(\left(\bigcap_{n=1}^\infty \left(B_n^{-\epsilon}\right)^{2\epsilon}\right)^{-\epsilon}\right)^\epsilon = \bigcap_{n=1}^\infty \left(B_n^{-\epsilon}\right)^{2\epsilon} - F\left(\bigcap_{n=1}^\infty \left(B_n^{-\epsilon}\right)^{2\epsilon}\right)\end{aligned}$$

and then one can argue that $C_n^\epsilon = (B_n^{-\epsilon})^{2\epsilon}$. Lastly, we demonstrate that $F(\bigcap_{n=1}^\infty (B_n^{-\epsilon})^{2\epsilon}) = \emptyset$.

5 Addressing Measurability

As mentioned earlier, defining the adversarial loss requires integrating over A^ϵ. However, one must ensure that A^ϵ is measurable. Furthermore, in the proof of Lemma 2, we apply the $^\epsilon, ^{-\epsilon}$ operations multiple times in succession. In particular, we consider sets of the form $((A^{-\epsilon})^{2\epsilon})^{-\epsilon}$. Hence we would like to work in a σ-algebra Σ for which if $A \in \Sigma$, $A^\epsilon \in \Sigma$ as well. Below, we explain that a σ-algebra called the *universal σ-algebra* satisfies this property.

We follow the treatment of Nishiura (2010) for our definitions. Let $\mathcal{B}(\mathbb{R}^d)$ be the Borel σ-algebra on $\mathbb{R}^d$ and let ν be a measure on this σ-algebra. We will denote the completion of the measure space $(\nu, \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ by $(\overline{\nu}, \mathbb{R}^d, \mathcal{L}_\nu(\mathbb{R}^d))$ where $\mathcal{L}_\nu(\mathbb{R}^d)$ is the σ-algebra of Lebesgue measurable sets. Let $\mathscr{M}(\mathbb{R}^d)$ be the set of all σ-finite Borel measures on $\mathbb{R}^d$. Then we define the *universal σ-algebra* as $\mathscr{U}(\mathbb{R}^d) = \bigcap_{\nu\in\mathscr{M}(\mathbb{R}^d)} \mathcal{L}_\nu(\mathbb{R}^d)$. In other words, $\mathscr{U}(\mathbb{R}^d)$ is the sets which are measurable under *every* complete σ-finite Borel measure. One can verify that an arbitrary intersection of σ-algebras is indeed a σ-algebra, so that $\mathscr{U}(\mathbb{R}^d)$ is in fact a σ-algebra. For the universal σ-algebra, we have the following theorem proved in Appendix A:

Theorem 3. *If $A \in \mathscr{U}(\mathbb{R}^d)$, then $A^\epsilon \in \mathscr{U}(\mathbb{R}^d)$ as well.*

Specifically, Theorem 3 allows us to define the adversarial risk in Equation (2). Recall that for a probability measure $\mathbb{Q}$, by definition $\mathscr{U}(\mathbb{R}^d) \subset \mathcal{L}_\mathbb{Q}(\mathbb{R}^d)$. Therefore, if $A \in \mathscr{U}(\mathbb{R}^d)$, then A^ϵ is measurable with respect to $(\overline{\mathbb{Q}}, \mathbb{R}^d, \mathcal{L}_\mathbb{Q}(\mathbb{R}^d))$. However, as this only holds for $A \in \mathscr{U}(\mathbb{R}^d)$ and not all of $\mathcal{L}_\mathbb{Q}(\mathbb{R}^d)$, throughout this paper, *we implicitly assume that our measure space is* $(\overline{\mathbb{Q}}, \mathbb{R}^d, \mathscr{U}(\mathbb{R}^d))$. In other words, we assume that the probability measure $\mathbb{P}$ is the complete measure $\overline{\mathbb{Q}}$ restricted to the σ-algebra $\mathscr{U}(\mathbb{R}^d)$. As $\mathscr{U}(\mathbb{R}^d)$ is closed under the $^\epsilon, ^{-\epsilon}$ operations, this convention allows us to mostly ignore measurability considerations.

Results similar to Theorem 3 appear in the literature, but are inadequate for our construction. For instance, Proposition 7.50 of Bertsekas and Shreve (1996) implies that if A is Borel measurable, then A^ϵ is universally measurable. However, as discussed earlier in this section, this result does not suffice because we need to show that for a σ-algebra Σ, $A \in \Sigma$ implies that $A^\epsilon \in \Sigma$ as well.

6 Alternative Models of Perturbations

In this paper, we developed techniques for proving the existence of the adversarial Bayes classifier on $\mathbb{R}^d$ with additive perturbations. Our techniques could be applied to other natural models of perturbations. In Appendix G, we state a general theorem that summarizes the part of our theory that is applicable beyond additive perturbations. Below, we discuss three notable examples.

Example 1 (Elementwise Scaling)**.** *For $\mathbf{x} \in \mathbb{R}^d$, we perturb each coordinate by multiplying it by a number in $[1-\epsilon, 1+\epsilon]$. Thus, to perturb $\mathbf{x}$, we multiply it elementwise by another vector in $B_\epsilon^\infty(\mathbf{1})$.*

(Engstrom et al., 2019) studied the following perturbation empirically in image classification tasks.

Example 2 (Rotations)**.** *Let $\mathbf{x} \in \mathbb{R}^d$. We perturb $\mathbf{x}$ by multiplying it by a "small" rotation matrix $\mathbf{R}$. We define our perturbation set this time as the set of matrices with*

$$B = \Big\{\mathbf{R} \colon \sup_{\|\mathbf{x}\|_2=1} \mathbf{x} \cdot \mathbf{R}\mathbf{x} \geq 1-\epsilon\Big\}.$$

Our final example is inspired from applications in natural language processing (Ebrahimi et al., 2018).

Example 3 (Discrete Perturbations)**.** *Let $\mathcal{A}$ be an alphabet. For an input string x, consider perturbations that replace a character of x at a given index with another character in $\mathcal{A}$.*

The above perturbation models have a lot in common with additive perturbations in $\mathbb{R}^d$. All three are examples of *semigroup actions*, and in fact the first two are group actions. Furthermore, all three involve metric spaces. Lastly, denoting a perturbed set as A^ϵ, we still have the containments

$$\left(\bigcup_{i=1}^\infty A_i\right)^\epsilon = \bigcup_{i=1}^\infty A_i^\epsilon \quad \text{and} \quad \left(\bigcap_{i=1}^\infty A_i\right)^\epsilon \subset \bigcap_{i=1}^\infty A_i^\epsilon. \tag{6}$$

Many aspects of the theory developed in this work are applicable in more general scenarios. In Appendix G.1, we prove the existence of the adversarial Bayes classifier for a simpler version of Example 3 using the techniques we developed in this paper. Proving the existence of the adversarial Bayes classifier for the other two examples remains an open problem.

Note that the proof of Theorem 1 only depends on Lemmas 1, 2, and 3, and not on the properties of $\mathbb{R}^d$. Thus in order to generalize our main theorem, one needs to generalize the three lemmas. Lemma 3 follows from the containments in (6) and Lemma 1 can be extended to separable metric metric spaces. Thus it remains to generalize both the measurability considerations and Lemma 2 on a case-by-case basis. Regarding measurability, we prove a more general version of Theorem 3 in Appendix A (Theorem 4) which applies to perturbations given by a metric ball in a metric space. Lastly, our tools may be useful for proving Lemma 2 in other scenarios and we discuss in Appendix G.

7 Conclusion

We initiated the study of fundamental questions regarding the existence of adversarial Bayes optimal classifiers. We provided sufficient conditions that ensure the existence of such classifiers when perturbing by an ϵ-ball. More importantly, our work highlights the need for new tools to understand Bayes optimality under adversarial perturbations, as one cannot simply rely on constructing pointwise optimal classifiers. Our paper also introduces several theorems which could be useful tools in further theoretical work. Specifically, Appendices D and E study properties of adversarially perturbed sets, and Appendix A gives some conditions under which adversarially perturbed sets are universally measurable. Both of these results may be useful in other contexts.

Similar to the case of standard loss functions, the most interesting extension of our work is to formulate and study questions related to the consistency of surrogate loss functions for adversarial robustness. We hope that this line of study will lead to new practically useful surrogate losses for designing adversarially robust classifiers.

Acknowledgements

This work was partly funded by NSF CCF-1535987 and NSF IIS-1618662. Special thanks to Professor James A. Morrow for some useful pointers on measurability.

References

A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

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

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

P. Awasthi, N. Frank, A. Mao, M. Mohri, and Y. Zhong. Calibration and consistency of adversarial surrogate losses. *NeurIPS*, 2021.

H. Bao, C. Scott, and M. Sugiyama. Calibrated surrogate losses for adversarially robust classification. *arXiv preprint arXiv:2005.13748*, 2020.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473), 2006.

P. L. Bartlett, S. Bubeck, and Y. Cherapanamjeri. Adversarial examples in multi-layer random relu networks. *CoRR*, 2021.

G. Bellettini. Anisotropic and crystalline mean curvature flow. *MSRI publications*, 2004.

D. P. Bertsekas and S. E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, 1996.

A. N. Bhagoji, D. Cullina, and P. Mittal. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems*, pages 7498–7510, 2019.

R. Bhattacharjee and K. Chaudhuri. When are non-parametric methods robust? *ICML*, 2020.

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*, pages 387–402. Springer, 2013.

P. Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. Wiley, 2nd edition, 1999. ISBN 0-471-19745-9.

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

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

N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

A. Cesaroni and N. Matteo. Isoperimetric problems for a nonlocal perimeter of minkowski type. *arXiv preprint arXiv:1709.05284*, 2017.

A. Cesaroni, D. Serena, and N. Matteo. Minimizers for nonlocal perimeters of minkowski type. *SpringerLink*, 2018.

A. Chambolle, M. Morini, and M. Ponsiglione. A non-local mean curvature flow and its semi-implicit time-discrete approximation. *SIAM Journal on Mathematical Analysis*, 2012.

A. Chambolle, M. Morini, and M. Ponsiglione. Nonlocal curvature flows. *Archive for Rational Mechanics and Analysis*, 218(3):1263–1329, 2015.

K. Ciesielski, H. Fejzić, and C. Freiling. Measure zero sets with non-measurable sum. *Real Analysis Exchange*, 2001/2002.

J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. *CoRR*, 2019.

D. Cullina, A. N. Bhagoji, and P. Mittal. Pac-learning in the presence of evasion adversaries. *arXiv preprint arXiv:1806.01471*, 2018.

B. Dacorogna. *Direct Methods in the Calculus of Variations*. Springer, 2008.

A. Degwekar, P. Nakkiran, and V. Vaikuntanathan. Computational limitations in robust classification and win-win results. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 994–1028, 2019.

D. Diochnos, S. Mahloujifar, and M. Mahmoody. Adversarial risk and robustness: General definitions and implications for the uniform distribution. In *Advances in Neural Information Processing Systems*, pages 10359–10368, 2018.

J. Ebrahimi, A. Rao, D. Lowd, and D. Dou. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, jul 2018.

L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, pages 1802–1811. PMLR, 2019.

P. Erdös and A. H. Stone. On the sum of two borel sets. *Proceedings of the American Mathematical Society*, 1970.

U. Feige, Y. Mansour, and R. Schapire. Learning and inference in the presence of corrupted inputs. In *Conference on Learning Theory*, pages 637–657, 2015.

U. Feige, Y. Mansour, and R. E. Schapire. Robust inference for multiclass classification. In *Algorithmic Learning Theory*, pages 368–386, 2018.

G. B. Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

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

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

M. Lewicka and Y. Peres. Which domains have two-sided supporting unit spheres at every boundary point? *Expositiones Mathematicae*, 38(4):548–558, 2020.

Y. Lin. A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68 (1):73–82, 2004. ISSN 0167-7152.

P. Long and R. Servedio. Consistency versus realizable H-consistency for multiclass classification. In *International Conference on Machine Learning*, pages 801–809, 2013.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

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

P. Nakkiran. Adversarial robustness may be at odds with simplicity. *arXiv preprint arXiv:1901.00532*, 2019.

T. Nishiura. *Absolute Measurable Spaces*. Cambridge University Press, 2010.

M. S. Pydi and V. Jog. Adversarial risk via optimal transport and optimal couplings. *arXiv preprint arXiv:1912.02794*, 2019.

A. Raghunathan, J. Steinhardt, and P. Liang. Semidefinite relaxations for certifying robustness to adversarial examples. *NeurIPS*, 2018. URL http://arxiv.org/abs/1811.01057.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 2015.

R. T. Rockafellar and R. J. Wets. *Variational Analysis*. Springer, 1998.

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

P. Stein. A note on the volume of a simplex. *The American Mathematical Monthly*, 1966.

I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE transactions on information theory*, 51(1):128–142, 2005.

I. Steinwart, D. Hush, and C. Scovel. Function classes that approximate the bayes risk. In *International Conference on Computational Learning Theory*, pages 79–93. Springer, 2006.

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

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

L. Weng, H. Zhang, H. Chen, Z. Song, C.-J. Hsieh, L. Daniel, D. Boning, and I. Dhillon. Towards fast computation of certified robustness for ReLU networks. In *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2018.

E. Wong and J. Z. Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Y.-Y. Yang, C. Rashtchian, Y. Wang, and K. Chaudhuri. Robustness for non-parametric classification: A generic attack and defense. *mlr*, 2020.

D. Yin, K. Ramchandran, and P. L. Bartlett. Rademacher complexity for adversarially robust generalization. In *Proceedings of ICML*, pages 7085–7094, 2019.

L. Zajìček. Porosity and σ-porosity. *Real Analysis Exchange*, 13(2):314–350, 1987.

H. Zhang, T. Weng, P. Chen, C. Hsieh, and L. Daniel. Efficient neural network robustness certification with general activation functions. *NeurIPS*, 2018.

M. Zhang and S. Agarwal. Bayes consistency vs. h-consistency: The interplay between surrogate loss functions and the scoring function class. In *Advances in Neural Information Processing Systems*, 2020.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals for Statistics*, 32, 2004.