On The Existence of The Adversarial Bayes Classifier (Extended Version)

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. This manuscript is the extended and corrected version of the paper *On the Existence of the Adversarial Bayes Classifier* published in NeurIPS 2021. There were two errors in theorem statements in the original paper— one in the definition of pseudo-certifiable robustness and the other in the measurability of A^ϵ for arbitrary metric spaces. In this version we correct the errors. Furthermore, the results of the original paper did not apply to some non-strictly convex norms and here we extend our results to all possible norms.

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

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

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-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. (2021a) 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 $f\colon \mathbb{R}^d \to \{-1, +1\}$ 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.

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 an adversarial Bayes classifier can be 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 describe a similar result for metric spaces. Next, in Appendix B, we prove one of our key lemmas about convergence of sets. These appendices present stand-alone results which do depend on material elsewhere in the appendix. In Appendix C, we subsequently provide some background material for the results in Appendicies D-F. Next, we prove the rest of our key lemmas in Appendicies D and E. Lastly, Appendix F 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, η will denote the conditional distribution, $\eta(\mathbf{x}) = \mathcal{D}(Y = +1|\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$, and $\mathbb{P}$ will denote 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

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

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 $A^C = \{\mathbf{x}\colon f(\mathbf{x}) \leq 0\}$. Conversely, specifying the set A is equivalent to specifying a function f since one could choose $f(\mathbf{x}) = \mathbb{1}_A(\mathbf{x})$. In the remainder 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 a 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 $A = \{f(\mathbf{x}) \geq 0\}$, 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 $\mathscr{U}(\mathbb{R}^d)$.

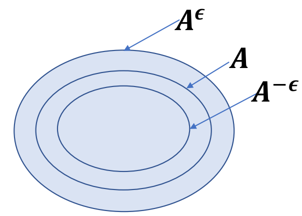

Figure 1: Sets A^ϵ and $A^{-\epsilon}$ with $B = \overline{B^2_\epsilon(\mathbf{0})}$, the closed ℓ_2 ball.

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 (see Lemma 14 in Appendix C for a formal proof). Figure 1 depicts the sets A, A^ϵ and $A^{-\epsilon}$.

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) with the sets A and A^C 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. [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 $\overline{B_\epsilon(\mathbf{x})} \subset A$, or $\mathbf{x} \in A^C$ and $\overline{B_\epsilon(\mathbf{x})} \subset A^C$. We say that a classifier A is* pseudo-certifiably robust at a point $\mathbf{x} \in A$ with radius ϵ *if there exists a ball $\overline{B_\epsilon(\mathbf{y})}$ with $\mathbf{x} \in \overline{B_\epsilon(\mathbf{y})}$ and $\overline{B_\epsilon(\mathbf{y})} \subset A$. 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 $\mathbf{x} \in A$ with radius ϵ if the entire ϵ-ball around $\mathbf{x}$ is classified the same as A, and a classifier is pseudo-certifiably robust at a point $\mathbf{x} \in A$ with radius ϵ if *some* closed ϵ-ball radius containing $\mathbf{x}$ is included in A. 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 $F(A)$. In Appendix D, we show that "subtracting" from a classifier the points at which it is not pseudo-certifiably robust can only reduce the risk. Similarly, "adding" to a classifier A the points at which A^C is not pseudo-certifiably robust can only reduce the risk as well. Formally, we show that $R^\epsilon(A - F(A)) \leq R^\epsilon(A)$ and $R^\epsilon(A \cup F(A^C)) \leq R^\epsilon(A)$ (Lemma 19). As illustrated in Figure 2, $F(A), F(A^C)$ are adjacent to the boundary ∂A. Furthermore, $F(A)$ is not very "large"— in fact, $F(A)^{-\epsilon} = \emptyset$. These observations suggest that,

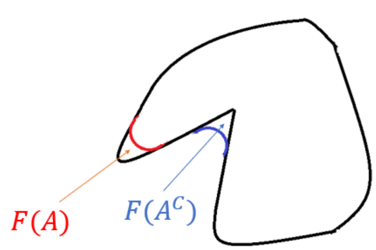

Figure 2: The figure illustrates a set A with the sets $F(A)$ and $F(A^C)$ roughly indicated. For a point $\mathbf{a} \in F(A)$, every closed ϵ-ball containing $\mathbf{a}$ also intersects A^C while for $\mathbf{a} \in F(A^C)$ every closed ϵ-ball containing $\mathbf{a}$ also intersects A.

[1] Pseudo-certifiable robustness was defined differently in the original version of this paper. We thank Ryan Murray for pointing out an error in Theorem 1 stemming from the earlier version of this definition.

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.

We now state our main existence result. We define the measures $\mathbb{P}_0$, $\mathbb{P}_1$ as in (Pydi and Jog, 2019) as

$$\mathbb{P}_1(A) = \int_A \eta d\mathbb{P}, \quad \mathbb{P}_0(A) = \int_A (1-\eta) d\mathbb{P}$$

Theorem 1. [2] *Let $\mathbb{P}$ be the completion of a Borel measure on $\mathcal{B}(\mathbb{R}^d)$ restricted to $\mathscr{U}(\mathbb{R}^d)$ and assume that either $\mathbb{P}_0, \mathbb{P}_1$ is absolutely continuous with respect to Lebesgue measure. Define $A^\epsilon = A \oplus \overline{B_\epsilon(\mathbf{0})}$, where $B_\epsilon(\mathbf{0})$ is a norm ball. Then, there exists a minimizer of (2) when minimizing over $\mathscr{U}(\mathbb{R}^d)$. Furthermore, there exists a minimizer that is pseudo-certifiably robust and a minimizer whose complement is pseudo-certifiably robust.*

The original version of this paper published in NeurIps (Awasthi et al., 2021b) proves this result for a restricted class of norms. For perturbations in an arbitrary norm, the theorem provides a positive guarantee: for any distribution, the adversarial Bayes classifier exists. In fact the proof of Theorem 1 shows that an even stronger result holds: under the hypotheses of Theorem 1, every minimizing sequence A_n has a subsequece A_{n_j} for which $\limsup_j A_{n_j}$ is the adversarial Bayes classifier. This conclusion is analogous to saying that every minimizing sequence must have a convergent subsequence.

To understand the significance of this statement, we compare to minimizing a function over $\mathbb{R}$. Consider the three functions $f(x) = (x^2-1)^2$, $g(x) = \sin(x)^2$, and $h(x) = 1/x^2$. The infimum of all three functions is 0. We can find minimizing sequences for f, g, and h which don't converge. For instance, the sequence given by

$$x_k = \begin{cases} +1 & k \text{ even} \\ -1 & k \text{ odd} \end{cases}$$

is a minimizing subsequence of f because $f(x_k) = 0$ for all k, but x_k is not a convergent subsequence. Intuitively, this phenomenon occurs because x_i is actually comprised of two subsequences each of which converges to a minimizer of f. In this case, every minimizing sequence of f has a convergent subsequence. On the other hand, minimizing sequences of g have very different behavior. For instance, consider the sequence given by $y_k = k\pi$. Then y_k is a minimizing sequence of g because $g(y_k) = 0$ for all k. However, the sequence y_k diverges to infinity, so $\{y_k\}$ does not have any convergent subsequence. Lastly, the sequence y_k also minimizes $h(x)$. Notably, h does not have a minimizer and thus all minimizing sequences diverge.

We also expect an analogous 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. (2021a) 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. 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. The proof of Theorem 1 shows that under reasonable assumptions on η and $\mathbb{P}$, every sequence A_n has a subsequence A_{n_j} for which $\limsup A_{n_j}$ is an adversarial Bayes classifier. 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 have a subsequence that approaches an adversarial Bayes classifier.

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. 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;

[2] The original version of this paper did not assume that either $\mathbb{P}_0$ or $\mathbb{P}_1$ was absolutely continuous with respect to Lebesgue measure.

2) Extract a subsequence $\{A_{n_j}\}$ of $\{A_n\}$ that is convergent in some topology;

3) Show that R^ϵ is sequentially lower semi-continuous: for a convergent subsequence $\{A_n\}$,

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

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 Fatou's lemma.

As described above, one of the main considerations in the proof of Theorem 1 is the convergence of set sequences. In order to find a minimizer, 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 $\{A_n\}$:

$$\limsup A_n = \bigcap_{N\geq 1} \bigcup_{n\geq N} A_n \text{ and } \liminf A_n = \bigcup_{N\geq 1} \bigcap_{n\geq N} A_n. \tag{5}$$

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 $\lim_{n\to\infty} \mathbb{1}_{A_n}$ exists $\mathbb{P}$-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 Lebesgue measure zero by $\doteq$. In order to find a sequence for which $\liminf A_n^\epsilon \doteq \limsup A_n^\epsilon$, we apply a theorem from variational analysis in (Rockafellar and Wets, 1998). Specifically, we show

Lemma 1. *Let $\mathbb{Q}$ be a finite positive measure and assume that $\mathbb{Q}$ is absolutely continuous with respect to Lebesgue measure. For any sequence of sets A_n, there is a sub-sequence A_{n_j} for which*

$$\limsup A_{n_j}^\epsilon \doteq \liminf A_{n_j}^\epsilon$$

The lemma above is proved in Appendix B. The next challenge is that $\liminf A_n^\epsilon / \limsup A_n^\epsilon$ do not necessarily equal A^ϵ for some set A. However, moving the $^\epsilon$ operation inside the $\liminf / \limsup$ decreases the risk.

Lemma 2. *Let A_n be any sequence of sets. Then*

$$\limsup A_n^\epsilon \supset (\limsup A_n)^\epsilon \quad \textit{and} \quad \liminf A_n^\epsilon \supset (\liminf A_n)^\epsilon$$

The lemma is proved in Appendix D.

Finally, it remains to show the claim about pseudo-certifiable robustness. We prove that for any set A there are sets B,E for which B and E^C are pseudo-certifiably robust and have lower robust risk than A.

Lemma 3. *Let A be any set. Then there exist sets B, E for which B and E^C are pseudo-certifiably robust and $R^\epsilon(B) \leq R^\epsilon(A)$, $R^\epsilon(E) \leq R^\epsilon(A)$.*

To prove this result, we show that applying $^{-\epsilon}$, $^\epsilon$ in succession to a set removes $F(A)$ as defined in (4). Analogously, applying $^{-\epsilon}$,$^\epsilon$ in succession to a set adds $F(A^C)$. We prove the above Lemma in Appendix E.

4.2 Formal Proof of Theorem 1

We now formally prove Theorem 1 using these three lemmas.

Proof of Theorem 1. WLOG assume that $\mathbb{P}_1$ is absolutely continuous with respect to Lebesgue measure.

Let A_n be a minimizing sequence of R^ϵ. By Lemma 1, there is a subsequence A_{n_j} for which $\limsup_j A_{n_j}^\epsilon \doteq \liminf_j A_{n_j}^\epsilon$ and thus

$$\int \eta \mathbb{1}_{\limsup_j A_{n_j}^\epsilon} d\mathbb{P} = \int \eta \mathbb{1}_{\liminf_j A_{n_j}^\epsilon} d\mathbb{P} \tag{6}$$

Fatou's lemma then implies that

$$\begin{aligned}
\inf_A R^\epsilon(A) = \liminf_{j\to\infty} R^\epsilon(A_{n_j}) &\geq \int \liminf_{j\to\infty} \left(\eta \mathbb{1}_{A_{n_j}^\epsilon} + (1-\eta)\mathbb{1}_{(A_{n_j}^C)^\epsilon}\right) d\mathbb{P} \\
&\geq \int \liminf_{j\to\infty} \eta \mathbb{1}_{A_{n_j}^\epsilon} + \liminf_{j\to\infty} (1-\eta)\mathbb{1}_{(A_{n_j}^C)^\epsilon} d\mathbb{P} \\
&= \int \eta \mathbb{1}_{\limsup_j A_{n_j}^\epsilon} + (1-\eta)\mathbb{1}_{\liminf_j (A_{n_j}^C)^\epsilon} d\mathbb{P} && \text{(Equation 6)} \\
&\geq \int \eta \mathbb{1}_{(\limsup_j A_{n_j})^\epsilon} + (1-\eta)\mathbb{1}_{(\liminf_j A_{n_j}^C)^\epsilon} d\mathbb{P} && \text{(Lemma 2)} \\
&= \int \eta \mathbb{1}_{(\limsup_j A_{n_j})^\epsilon} + (1-\eta)\mathbb{1}_{((\limsup_j A_{n_j})^C)^\epsilon} d\mathbb{P}
\end{aligned}$$

Therefore, $A = \limsup_j A_{n_j}$ is a minimizer of R^ϵ. Lemma 3 then implies that there are sets B, E for which B, E^C are pseudo-certifiably robust and $R^\epsilon(B) \leq R^\epsilon(A)$ and $R^\epsilon(E) \leq R^\epsilon(A)$. Therefore, B, E are minimizers as well. ☐

4.3 Proof Outline for Lemmas 1, 2, and 3

In this section, we explain the intuition for the proofs of Lemmas 1, 2, and 3. Lemmas 1 and 2 follow directly from properties of the $^\epsilon$ operation. Specifically, in Appendix C we show that

$$\left(\bigcup_{i=1}^\infty A_i\right)^\epsilon = \bigcup_{i=1}^\infty A_i^\epsilon \quad \text{and} \quad \left(\bigcap_{i=1}^\infty A_i\right)^\epsilon \subset \bigcap_{i=1}^\infty A_i^\epsilon. \tag{7}$$

As the $\liminf$ and $\limsup$ operations of (5) are defined by unions and intersections, this result immediately implies Lemma 2. Next, one can use the relations of (7) to argue that if $B = (A^{-\epsilon})^\epsilon$, then $(B^C)^\epsilon = (A^C)^\epsilon$ and $B^\epsilon \subset A^\epsilon$ so therefore $R^\epsilon(B) \leq R^\epsilon(A)$. One can make an analogous statement with $E = (A^\epsilon)^{-\epsilon}$, see Appendix D for the formal statement and proof.

The proof of Lemma 1 combines the analysis of the $^\epsilon$ operation with measure theoretic considerations. Rockafellar and Wets (1998) prove a set convergence result for a different notion of the $\liminf$ and $\limsup$ of a sequence of sets S_n. This notion of set convergence includes points that are arbitrarily close to S_n for infinitely many n. The standard $\liminf$ / $\limsup$ operations have a similar interpretation in terms of sequences. Recall that

$$\liminf S_n = \{\mathbf{x}\colon \text{ there exists an } N \text{ for which } \mathbf{x} \in S_n \text{ for all } n > N\} \tag{8}$$

$$\limsup S_n = \{\mathbf{x}\colon \text{ there exists a sequence } n_j \text{ for which } \mathbf{x} \in S_{n_j} \text{ for all } j\} \tag{9}$$

On the other hand, the Rockafellar and Wets (1998) defines $\widetilde{\liminf}, \widetilde{\limsup}$ in terms of convergent sequences $\{\mathbf{x}_n\}$ with $\mathbf{x}_n \in S_n$:

$$\widetilde{\liminf}\, S_n = \{\mathbf{x}\colon \text{there exists a sequence with } \mathbf{x}_n \in S_n \text{ and } \lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}\}$$

$$\widetilde{\limsup}\, S_n = \{\mathbf{x}\colon \text{there exists a subsequence } \{n_j\} \text{ with } \mathbf{x}_{n_j} \in S_{n_j} \text{ and } \lim_{i\to\infty} \mathbf{x}_{n_j} = \mathbf{x}\} \tag{10}$$

In other words, a point $\mathbf{x}$ is in $\liminf S_n$ if $\mathbf{x} \in S_n$ for sufficiently large n while $\mathbf{x} \in \widetilde{\liminf}\, S_n$ if the distance between $\mathbf{x}$ and S_{n_j} approaches zero. Similarly, a point is in $\limsup S_n$ if $\mathbf{x} \in S_{n_j}$ for some subsequence n_j while $\mathbf{x} \in \widetilde{\limsup} S_{n_j}$ if there is a subsequence n_j for which the distance between $\mathbf{x}$ and S_n approaches zero. This characterization immediately implies $\liminf S_n \subset \widetilde{\liminf} S_n$ and $\limsup S_n \subset \widetilde{\limsup} S_n$. For the notions of set limit $\widetilde{\liminf}, \widetilde{\limsup}$ every subsequence has a convergent subsequence:

Theorem 2 ((Rockafellar and Wets, 1998)). *Let S_n be any sequence of sets in $\mathbb{R}^d$. Then there is a subsequence S_{n_j} of S_n for which $\widetilde{\liminf} S_{n_j} = \widetilde{\limsup} S_{n_j}$.*

This statement is a consequence of Theorem 4.18 of (Rockafellar and Wets, 1998). In other words, one can always choose a subsequence S_{n_k} of S_n for which the $\widetilde{\liminf}$ and the $\widetilde{\limsup}$ match. This result is *false* for the standard definitions of $\liminf, \limsup$.

However, pseudo-certifiably robust sets are fairly well-behaved, so one would hope such sets would also interact well with the standard definition of $\liminf$ and $\limsup$. Furthermore, a standard argument from geometric measure theory implies that pseudo-certifiably robust sets have a measure zero boundary.

Lemma 4. *Let μ be Lebesgue measure and let $S \subset \mathbb{R}^d$. If for each $\mathbf{s} \in \partial S$ there exists a ball $B_\epsilon(\mathbf{a})$ with $B_\epsilon(\mathbf{a}) \subset S$ and $\mathbf{s} \in \partial B_\epsilon(\mathbf{a})$, then $\mu(\partial S) = 0$.*

See Appendix B.2 for a proof.

One can show that for a subsequence A_{n_j} with $\widetilde{\liminf} A_{n_j} = \widetilde{\limsup} A_{n_j}$, the set $\liminf A^\epsilon_{n_j}$ satisfies a property similar to pseudo-certifiable robustness: for all $\mathbf{x} \in \liminf A^\epsilon_{n_j}$, there is a ball $B_\epsilon(\mathbf{a})$ for which $\mathbf{x} \in \overline{B_\epsilon(\mathbf{a})}$ and $B_\epsilon(\mathbf{a}) \subset \liminf A^\epsilon_{n_j}$ (See Lemma 10 in Appendix B). In other words, the condition of Lemma 4 is satisfied at every point in $\liminf A^\epsilon_{n_j}$. By taking limits, one can then argue that this property also holds for all $\mathbf{x} \in \partial \liminf A^\epsilon_{n_j}$. Lemma 4 then implies Lemma 1.

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 3, we apply the $^\epsilon, ^{-\epsilon}$ operations multiple times in succession. In particular, we consider sets of the form $(A^{-\epsilon})^\epsilon$. 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.

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 completion of $\mathcal{B}(\mathbb{R}^d)$ under ν. Let $\mathscr{M}(\mathbb{R}^d)$ be the set of all 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 finite Borel measure. For the universal σ-algebra, we have the following theorem proved in Appendix A.2:

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). Appendix A.1 proves a similar measurability theorem for metric spaces. 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 a 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 7.36 of Bertsekas and Shreve (1996) implies that if A is Borel measurable, then A^ϵ is universally measurable (See Appendix A.1 for more details). 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. However, as we detail in Appendix A.1, this approach shows that for an arbitrary metric space, one can still define the adversarial risk R^ϵ.

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 attacks. In Appendix F, we state a general theorem that summarizes the part of our theory that is applicable beyond additive perturbations. 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 in (7).

Many aspects of the theory developed in this work are applicable in more general scenarios. In Appendix F.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. Lemmas 2 and 3 follow directly from the containments in (7).

Thus it remains to generalize both the measurability considerations and Lemma 1 on a case-by-case basis. Regarding measurability, we prove a statement similar to Theorem 3 in Appendix A (Theorem 4) which applies to perturbations given by a metric ball in a metric space. Specifically, this theorem states that if A is Borel in a metric space, then A^ϵ is universally measurable. Lastly, our tools may be useful for proving Lemma 1 in other scenarios.

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.

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. Natalie Frank was supported in part by the Research Training Group in Modeling and Simulation funded by the National Science Foundation via grant RTG/DMS – 1646339. The authors would like to thank Laurent Meunier for pointing out a missing step in the proof of Theorem 1 in the earlier version of the manuscript, Ryan Murray for finding an error in one of our proofs. The authors would also like to thank Professor James A. Morrow for a helpful citation (Nishiura, 2010) 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*, 2021a.

P. Awasthi, N. S. Frank, and M. Mohri. On the existence of the adversarial bayes classifier. *Advances in Neural Information Processing Systems*, 2021b.

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.

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.

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 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.

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.

Contents of Appendix

A The Measurability of A^ϵ

In this section, we prove two versions of Theorem 3. The first applies to general metric spaces and the second to abstract vector spaces. We discuss the theorem in high generality for two reasons. First, discussing this result in terms of abstract concepts actually clarifies main idea underlying these results. In fact, the proof of the statement we show for metric spaces is simpler than the one we show for vector spaces. Second, we suspect that our framework will be useful in discussing other models of perturbations.

Throughout this section, we denote elements of the vector space $\mathbb{R}^d$ in bold ($\mathbf{x}$) and elements of a general metric space X as non-bold (x).

A.1 Measurability for Metric Spaces

For a measure space $(\nu, X, \mathcal{B}(X))$ equipped with the Borel σ-algebra $\mathcal{B}(X)$, we will denote its completion as $(\overline{\nu}, X, \mathcal{L}_\nu(X))$. Furthermore, throughout this section, we assume that X is a metric space and that the Borel σ-algebra $\mathcal{B}(X)$ is generated by the sets open in the metric on X.

Recall that in Section 4, we defined A^ϵ as $A^\epsilon = A \oplus \overline{B_\epsilon(\mathbf{0})}$. Another way to write this relation is

$$A^\epsilon = \bigcup_{\mathbf{a} \in A} \overline{B_\epsilon(\mathbf{a})}. \tag{11}$$

This form for A^ϵ is helpful because it allows us to define A^ϵ for general metric spaces. Notably, one can define the adversairal risk in a general metric space as

$$\begin{aligned} R^\epsilon(A) &= \int (1-\eta(x))\mathbb{1}_{A^\epsilon}(x) + \eta(\mathbf{x})\mathbb{1}_{(A^C)^\epsilon}(x)\, d\mathbb{P} \\ &= \int (1-\eta(x)) \sup_{a\in A} \mathbb{1}_{\overline{B_\epsilon(a)}}(x) + \eta(x) \sup_{a \in A^C} \mathbb{1}_{\overline{B_\epsilon(a)}}(\mathbf{x})\, d\mathbb{P} \end{aligned}$$

holds for general metric spaces when we define A^ϵ as in (11). On $\mathbb{R}^d$, the second line is equivalent to the expression (3). We will use this observation later to prove a generalized version of our theorem for alternative models of perturbations.

We start by defining the universal σ-algebra for a measure space X.

Definition 2. *Let X be a Borel space and let $\mathscr{M}(X)$ be the set of all finite positive Borel measures on X. We define the* universal σ-algebra *to be*

$$\mathscr{U}(X) = \bigcap_{\nu \in \mathscr{M}(X)} \mathcal{L}_\nu(X). \tag{12}$$

If $A \in \mathscr{U}(X)$, then we say that A is universally measurable. [3]

In this section we prove the following theorem:

Theorem 4. [4] *Let (X, d) be complete separable metric space. Define A^ϵ as in (11). If $A \subset X$ is Borel measurable, then A^ϵ is universally measurable.*

That the metric on our space X generates the topology on X which in turn generates $\mathcal{B}(X)$ is implicit in this theorem statement.

This subtlety is crucial when applying Theorem 4. For norms $\mathbb{R}^d$ however, the situation simplifies– all norms generate the standard topology. In contrast, a general seminorm does not generate the standard topology on $\mathbb{R}^d$, so Theorem 4 in this case would not apply to $\mathbb{R}^d$ with the usual Borel σ-algebra.

We now describe the basic idea behind the proof of Theorem 4. Consider a Borel set $A \subset X$. Then $X \times A$ is Borel in $X \times X$. The set $\Delta_\epsilon = \{(x, y) \in X \times X : d(x, y) \leq \epsilon\}$ is closed, and therefore

[3] Alternatively, one could compute the intersection in (12) over all σ-finite measures. These two approaches are equivalent because for every σ-finite measure λ and compact set K, the restriction $\lambda \llcorner K$ is a finite measure with $\mathcal{L}_{\lambda\llcorner K}(X) \supset \mathcal{L}_\lambda(X)$. See Theorem 1.5 of and Proposition 2.5 (Nishiura, 2010).

[4] The original version of this paper stated that in an arbitrary metric space, if A is universally measurable, then A^ϵ is universally measurable as well, which was an error.

Borel. Thus $X \times A \cap \Delta_\epsilon$ is Borel in $X \times X$. Notice that A^ϵ is the projection of this set onto the first coordinate. Such a projection is universally measurable.

In (Bertsekas and Shreve, 1996), Propositions 7.41 and the statement that $\mathscr{U}(X)$ contains the analytic σ-algebra implies the following theorem:

Theorem 5. *Let S be a Borel set in $X \times Y$ and let $\Pi_1 \colon X \times Y \to X$ be projection onto the first coordinate: $\Pi_1(x, y) = x$. Then $\Pi_1(S)$ is universally measurable*

In fact, this analysis implies that A^ϵ is measurable with respect to a smaller σ-algebra called the *analytic σ-algebra*. See Chapter 7 of (Bertsekas and Shreve, 1996) for details. We formally perform this calculation below.

Proof of Theorem 4. Let $\Delta_\epsilon = \{(x, y) \colon d(x, y) \leq \epsilon\}$. We will show that $A^\epsilon = \Pi_1(X \times A \cap \Delta_\epsilon)$. Theorem 5 will then imply the result.

$$\begin{aligned}\Pi_1(X \times A \cap \Delta_\epsilon) &= \{\mathbf{x} \colon \text{for some } \mathbf{a} \in A, (\mathbf{x}, \mathbf{a}) \in \Delta_\epsilon\} = \{\mathbf{x} \colon \text{for some } \mathbf{a} \in A, d(\mathbf{a}, \mathbf{x}) \leq \epsilon\} \\ &= \bigcup_{\mathbf{a} \in A} \overline{B_\epsilon(\mathbf{a})} = A^\epsilon\end{aligned}$$

□

A.2 Measurability for Vector Spaces

In this section we show the following measurability result:

Theorem 6. *Let $(X, \|\cdot\|)$ be a separable vector space. Define A^ϵ as $A^\epsilon = A \oplus \overline{B_\epsilon(\mathbf{0})}$, where B_ϵ is an ϵ-ball in the norm $\|\cdot\|$. If $A \subset X$ is universally measurable, then A^ϵ is universally measurable as well.*

As $\mathbb{R}^d$ with the standard topology is separable and all norms on $\mathbb{R}^d$ generate the standard topology, Theorem 3 immediately follows from Theorem 6.

Again, that the norm on our space X generates the topology on X which in turn generates $\mathcal{B}(X)$ is implicit in this theorem statement.

Before proving Theorem 6, we define another useful concept.

Definition 3. *Let X, Y be a separable metric spaces and let $(\overline{\nu}, Y, \mathcal{L}_\nu(Y))$ be a complete σ-finite measure space. Then X is* absolute measurable *if for every injective continuous map $h \colon X \to Y$, $h(X)$ is an element of $\mathcal{L}_\nu(Y)$.*

This definition is useful due to the following theorem:

Theorem 7. *Let X be an absolute measurable Borel space and Y a separable metrizable space. Let $f \colon X \to Y$ be a homeomorphism.*

Then

$$f[\mathscr{U}(X)] \subset \mathscr{U}(Y).$$

This theorem is the implication (1) ⇒ (4) of the Purves-Darst-Grzegorek Theorem, stated on page 33 in Chapter 2.1 of (Nishiura, 2010). A separable vector space is σ-compact. This fact implies that that Theorem 7 applies.

Lemma 5. *A σ-compact space is absolute measurable.*

We now describe the basic idea behind the proof of Theorem 6 for $\mathbb{R}^d$. Consider the homeomorphism $w \colon \overline{B_\epsilon(\mathbf{0})} \times \mathbb{R}^d \to \overline{B_\epsilon(\mathbf{0})} \times \mathbb{R}^d$ given by $w(\mathbf{v}, \mathbf{x}) = (\mathbf{v}, \mathbf{x} + \mathbf{v})$. Then for any set A, $w(\overline{B_\epsilon(\mathbf{0})}, A) = \overline{B_\epsilon(\mathbf{0})} \times A^\epsilon$. Therefore, if $\overline{B_\epsilon(\mathbf{0})}$ is universally measurable in $\overline{B_\epsilon(\mathbf{0})} \times \mathbb{R}^d$, then $\overline{B_\epsilon(\mathbf{0})} \times A^\epsilon$ is also universally measurable. To conclude that A^ϵ is universally measurable in $\mathbb{R}^d$, it remains to show that that $\overline{B_\epsilon(\mathbf{0})} \times S$ is universally measurable in $\overline{B_\epsilon(\mathbf{0})} \times \mathbb{R}^d$ iff S is universally measurable in $\mathbb{R}^d$.

Lemma 6. *Let X, Y be Borel spaces. If $S \in \mathscr{U}(Y)$, then $X \times S \in \mathscr{U}(X \times Y)$.*

Lemma 7. *Let X, Y be second countable, locally compact Hausdorff spaces. Assume that X is compact and Y is σ-compact. If $X \times S \in \mathscr{U}(X \times Y)$, then $S \in \mathscr{U}(Y)$.*

We prove Theorem 6 using these results.

Proof of Theorem 6. Consider the function $w\colon \overline{B_\epsilon(\mathbf{0})} \times X \to \overline{B_\epsilon(\mathbf{0})} \times X$ given by $w(\mathbf{h}, \mathbf{x}) = (\mathbf{h}, \mathbf{x} + \mathbf{h})$. Then w is continuous and invertible, so it is a homeomorphism. Furthermore, it maps the set $\overline{B_\epsilon(\mathbf{0})} \times A$ to $\overline{B_\epsilon(\mathbf{0})} \times (A \oplus \overline{B_\epsilon(\mathbf{0})})$.

Let A be a universally measurable subset of X. Then by Lemma 6, $\overline{B_\epsilon(\mathbf{0})} \times A$ is universally measurable in $\overline{B_\epsilon(\mathbf{0})} \times X$. Thus, Theorem 7 implies that $w(\overline{B_\epsilon(\mathbf{0})}, A) = \overline{B_\epsilon(\mathbf{0})} \times A^\epsilon$ is universally measurable in $\overline{B_\epsilon(\mathbf{0})} \times X$. Lastly, Lemma 7 implies that A^ϵ is measurable.

□

A.3 Proofs of Lemmas 5, 6, and 7

Lemma 5. *A σ-compact space is absolute measurable.*

Proof of Lemma 5. We will start by showing that a compact space is absolute measurable. Let H be a compact topological space and let Y be a separable metric space. If $f\colon H \to Y$ is continuous, a well-known theorem from topology implies that $f(H)$ is compact as well. A compact subset of a metric space is always closed, and therefore $f(H)$ is a Borel set.

Next, consider a σ-compact space X. Write

$$X = \bigcup_{n \in \mathbb{N}} H_n$$

where each H_n is compact. Then if $f\colon X \to Y$ is a continuous map, then $f(X)$ is a countable union of Borel sets;

$$f(X) = \bigcup_{n \in \mathbb{N}} f(H_n)$$

and is therefore Borel as well. □

Lemma 6. *Let X, Y be Borel spaces. If $S \in \mathscr{U}(Y)$, then $X \times S \in \mathscr{U}(X \times Y)$.*

Proof of Lemma 6. Let $(\nu, X \times Y, \mathcal{B}(X \times Y))$ be an arbitrary finite Borel measure on $X \times Y$ and let $S \in \mathscr{U}(Y)$. We will show that $X \times S \in \mathcal{L}_\nu(X \times Y)$. As the universal σ-algebra is the intersection of all $\mathcal{L}_\nu(X \times Y)$ for all finite Borel measures ν, this inclusion will imply that $X \times S \in \mathscr{U}(X \times Y)$.

Let λ be the marginal distribution on Y given by $\lambda(B) = \nu(X \times B)$ with σ-algebra $\mathcal{B}(Y)$. Now consider the completion $(\overline{\lambda}, Y, \mathcal{L}_\lambda(Y))$. Because S is in the universal σ-algebra for Y, we know that $S \in \mathcal{L}_\lambda(Y)$. Therefore, $S = B \cup N'$ where B is a Borel set and N' is a subset of a null Borel set N. Because N is Borel, $X \times N$ is as well and $\nu(X \times N) = \lambda(N) = 0$. Therefore, $X \times N$ is a null Borel set for the measure space $(\nu, X \times Y, \mathcal{B}(X \times Y))$. Thus both $X \times N'$ and $X \times B$ are in the complete measure space $(\overline{\nu}, X \times Y, \mathcal{L}_\nu(X \times Y))$. Therefore, $X \times S = X \times B \cup X \times N'$ is in $\mathcal{L}_\nu(X \times Y)$ as well. □

However, to prove the converse to Lemma 6, one must apply the concept of regularity of measures.

Definition 4. *Let τ be a topology on a set X and $\mathcal{B}(X)$ the Borel σ-algebra generated by τ. Let $\mathbb{P}$ be a Borel measure on $(X, \mathcal{B}(X))$.Then $\mathbb{P}$ is* inner regular *if for all measurable sets E, $\mathbb{P}(E) = \sup\{\mathbb{P}(K)\colon K \subset E, K \text{ compact}\}$. A space is* regular *if all finite Borel measures on X are inner regular.*

Theorem 7.8 of (Folland, 1999) implies that most measure spaces encountered in applications are regular:

Theorem 8. *Let X be a second-countable and locally compact Haudsorff space. Then every finite Borel measure is inner regular.*

The notion of regularity extends to complete measures.

Lemma 8. *Let ν be a finite positive Borel measure on a regular space X and let $\overline{\nu}$ be the completion of ν. Let $A \in \mathcal{L}_\nu(X)$. Then*

$$\overline{\nu}(A) = \sup_{\substack{K \subset A \\ K \text{ compact}}} \nu(K).$$

Proof. If $A \in \mathcal{L}_\nu(X)$, then there is a Borel set B with $B \subset A$ and $\overline{\nu}(A) = \nu(B)$. The result then follows from the definition of inner regularity for Borel measures. □

Lemma 7. *Let X, Y be second countable, locally compact Hausdorff spaces. Assume that X is compact and Y is σ-compact. If $X \times S \in \mathscr{U}(X \times Y)$, then $S \in \mathscr{U}(Y)$.*

Proof of Lemma 7. As X, Y are both σ-compact spaces, Theorem 8 implies that $X \times Y$ is regular. Fix a Borel probability measure λ on X, and let ν be any finite Borel measure on Y. Then $\lambda \times \nu$ is a Borel probability measure on $X \times Y$, so it is inner regular. Let $\overline{\lambda \times \nu}$ be the completion of $\lambda \times \nu$. Then

$$\overline{\lambda \times \nu}(X \times S) = \sup_{\substack{K \text{ compact} \\ K \subset X \times S}} \lambda \times \nu(K)$$

We will now argue that

$$\sup_{\substack{K \text{ compact} \\ K \subset X \times S}} \lambda \times \nu(K) = \sup_{\substack{K \text{ compact} \\ K \subset S}} \nu(K) \tag{13}$$

Let $K \subset X \times S$ and let $\Pi_2 \colon X \times X \to X$ be projection onto the second coordinate. Because the continuous image of a compact set is compact, $K' = \Pi_2(K)$ is compact and contained in S. Thus $X \times S \supset X \times K' \supset X \times K$, which implies (13). Now (13) applied to S^C implies that

$$\overline{\lambda \times \nu}(X \times S) = \inf_{\substack{U^C \text{ compact} \\ U \supset X \times S}} \lambda \times \nu(U) = \inf_{\substack{U^C \text{ compact} \\ U \supset S}} \nu(U)$$

Thus

$$\sup_{\substack{K \text{ compact} \\ K \subset S}} \nu(K) = \inf_{\substack{U^C \text{ compact} \\ U \supset S}} \nu(U) := m$$

Let K_n be a sequence of compact sets contained in A for which $\lim_{n\to\infty} \nu(K_n) = m$ and U_n a sequence of sets containing A for which U_n^C is compact and $\lim_{n\to\infty} \nu(U_n) = m$. Because a finite union of compact sets is compact, one can choose such sequences that satisfy $K_{n+1} \supset K_n$ and $U_{n+1} \subset U_n$. Then $D = \bigcup K_n$, $V = \bigcap U_n$ are Borel sets that satisfy $D \subset S \subset V$ and $\nu(D) = \nu(V)$, so $V - D$ a null set. Thus $S - D$ is a subset of the null Borel set $V - D$, so $S \in \mathcal{L}(\mathbb{R}^d)$. As ν was arbitrary, it follows that S is universally measurable.

□

B Proof of Lemma 1

In the next subsection we show how Lemma 4 implies Lemma 1. The proof of Lemma 4 is delayed to B.2.

B.1 Main Argument

The following lemma assists in understanding the convergence of pseudo-certifiably robust sets.

Lemma 9. *Let $\{\mathbf{a}_n\}$ be a sequence converging to $\mathbf{a}$. Then if $\mathbf{x} \in B_\epsilon(\mathbf{a})$, then for sufficiently large n, $\mathbf{x} \in B_\epsilon(\mathbf{a}_n)$.*

Proof. Set $r = \|\mathbf{x} - \mathbf{a}\|$. Then if we choose n large enough so that $\|\mathbf{a} - \mathbf{a}_n\| < \epsilon - r$, then

$$\|\mathbf{x} - \mathbf{a}_n\| \leq \|\mathbf{x} - \mathbf{a}\| + \|\mathbf{a} - \mathbf{a}_n\| < r + (\epsilon - r) = \epsilon$$

Therefore, for sufficiently large n, $\mathbf{x} \in B_\epsilon(\mathbf{a}_n)$. □

The following result is central to the proof of Lemma 1.

Lemma 4. *Let μ be Lebesgue measure and let $S \subset \mathbb{R}^d$. If for each $\mathbf{s} \in \partial S$ there exists a ball $B_\epsilon(\mathbf{a})$ with $B_\epsilon(\mathbf{a}) \subset S$ and $\mathbf{s} \in \partial B_\epsilon(\mathbf{a})$, then $\mu(\partial S) = 0$.*

The next lemma shows that if $\widetilde{\limsup} A_n = \widetilde{\liminf} A_n$, then for every $\mathbf{x} \in \liminf A_n^\epsilon$ there is a ball for which $\mathbf{x} \in \overline{B_\epsilon(\mathbf{a})}$ but $B_\epsilon(\mathbf{a}) \subset \liminf A_n^\epsilon$. Thus the set $\liminf A_n^\epsilon$ satisfies the property required by Lemma 4 at all points. It remains to show that this property also holds on the boundary $\partial \liminf A_n^\epsilon$, which is later accomplished by taking limits.

Lemma 10. *Let A_n be a sequence for which $\widetilde{\liminf} A_n = \widetilde{\limsup} A_n$. Then $\liminf A_n^\epsilon$ has the following property: for every $\mathbf{x} \in \liminf A_n^\epsilon$, there is a ball $B_\epsilon(\mathbf{a})$ for which $\mathbf{x} \in \overline{B_\epsilon(\mathbf{a})}$ and $B_\epsilon(\mathbf{a}) \subset \liminf A_n^\epsilon$.*

Proof. Let $\mathbf{x} \in \liminf A_n^\epsilon$. The expression for the $\liminf$ in (8) implies that there is a J for which $\mathbf{x} \in A_n^\epsilon$ for all $n > N$. Hence one can write $\mathbf{x} = \mathbf{a}_n + \mathbf{h}_n$ with $\mathbf{a}_n \in A_n$ and $\mathbf{h}_n \in \overline{B_\epsilon(\mathbf{0})}$ for all $n > N$. Now pick a subsequence n_j for which $\mathbf{h}_{n_j}$ converges and set $\mathbf{h} = \lim_{j\to\infty} \mathbf{h}_{n_j}$. Then let

$$\mathbf{a} = \lim_{j\to\infty} \mathbf{a}_{n_j} = \mathbf{x} - \mathbf{h}. \tag{14}$$

Due to the definition of $\widetilde{\limsup}$ in (10), $\mathbf{a} \in \widetilde{\limsup} A_n$ and by the assumption on our sequence A_n, $\widetilde{\limsup} A_n = \widetilde{\liminf} A_n$. Thus there is a sequence $\tilde{\mathbf{a}}_n$ for which $\tilde{\mathbf{a}}_n \in A_n$ and $\lim_{n\to\infty} \tilde{\mathbf{a}}_n = \mathbf{a}$. Then $B_\epsilon(\tilde{\mathbf{a}}_n) \subset A_n^\epsilon$ and Lemma 9 then implies that $B_\epsilon(\mathbf{a}) \subset \liminf_j A_n^\epsilon$.

Lastly, one can conclude that $\|\mathbf{x} - \mathbf{a}\| \leq \epsilon$ from the definition of $\mathbf{a}$ in (14). ☐

Finally, Lemma 1 is a consequence of Lemma 4, Lemma 10, and Theorem 2.

Lemma 1. *Let $\mathbb{Q}$ be a finite positive measure and assume that $\mathbb{Q}$ is absolutely continuous with respect to Lebesgue measure. For any sequence of sets A_n, there is a sub-sequence A_{n_j} for which*

$$\limsup A_{n_j}^\epsilon \doteq \liminf A_{n_j}^\epsilon$$

Proof. Let μ denote Lebesgue measure. We will find a subsequence A_{n_j} of A_n for which $\mu(\widetilde{\limsup} A_{n_j} - \widetilde{\liminf} A_{n_j}) = 0$. By Theorem 2, one can find a subsequence n_j for which $\widetilde{\liminf} A_{n_j} = \widetilde{\limsup} A_{n_j}$. Then for this subsequence, Lemma 10 then applies to this subsequence. We will argue that $\liminf A_n^\epsilon$ in fact satisfies the property of Lemma 4.

Let $\mathbf{x} \in \partial \liminf A_{n_j}^\epsilon$. We will find a ball $B_\epsilon(\mathbf{a}) \subset \liminf A_{n_j}^\epsilon$ for which $\mathbf{x} \in \overline{B_\epsilon(\mathbf{a})}$. If $\mathbf{x} \in \liminf A_{n_j}^\epsilon$, then there is a sequence $\mathbf{x}^k \in \liminf A_{n_j}^\epsilon$ converging to $\mathbf{x}$. By Lemma 10, for each $\mathbf{x}^k$, there is a $\mathbf{a}^k$ with $\mathbf{x} \in \overline{B_\epsilon(\mathbf{a}^k)}$ and $B_\epsilon(\mathbf{a}^k) \subset \liminf A_{n_j}^\epsilon$. Furthermore, because $\mathbf{x}^k \to \mathbf{x}$, the set $\{\mathbf{a}^k\}$ is bounded. Let k_m be a subsequence for which $\mathbf{a}^{k_m}$ converges and set $\mathbf{a} = \lim_{m\to\infty} \mathbf{a}^{k_m}$. Then because $B_\epsilon(\mathbf{a}^{k_m}) \subset \liminf A_{n_j}^\epsilon$, Lemma 9 implies that $B_\epsilon(\mathbf{a}) \subset \liminf A_{n_j}^\epsilon$ as well. Next,

$$\|\mathbf{x} - \mathbf{a}\| \leq \|\mathbf{x} - \mathbf{x}^{k_m}\| + \|\mathbf{x}^{k_m} - \mathbf{a}^{k_m}\| + \|\mathbf{a}^{k_m} - \mathbf{a}\| \leq \|\mathbf{x} - \mathbf{x}^{k_m}\| + \epsilon + \|\mathbf{a}^{k_m} - \mathbf{a}\|$$

As $\|\mathbf{x} - \mathbf{x}^{k_m}\|$, $\|\mathbf{a} - \mathbf{a}^{k_m}\|$ both approach zero, it follows that $\mathbf{x} \in \overline{B_\epsilon(\mathbf{a})}$. Therefore, Lemma 4 applies.

☐

B.2 Proof of Lemma 4

To prove Lemma 4 we take an approach that is standard in geometric measure theory. The strategy is to apply the Lebesgue differentiation theorem.

Theorem 9 (Lebesgue Differentiation Theorem)**.** *Assume that $f\colon \mathbb{R}^d \to \mathbb{R}$ is bounded. Then the following holds for $\mathbf{x}$ μ-a.e.:*

$$\lim_{r\to 0} \frac{1}{\mu(B_r^2(\mathbf{x}))} \int_{B_r^2(\mathbf{x})} f\, d\mu = f(\mathbf{x})$$

See Folland (1999) for a proof. We prove a slightly more general version of Lemma 4:

Lemma 11. *Let μ be Lebesgue measure and let $S \subset \mathbb{R}^d$. If for each $\mathbf{s} \in \partial S$ there exists an open convex C with $C \subset S$ and $\mathbf{s} \in \partial C$, then $\mu(\partial S) = 0$.*

Proof of Lemma 11. We will apply the Lebesgue differentiation theorem to the function $\mathbb{1}_{\partial S}$.

Let $B_r^2(\mathbf{x})$ denote the radius r ball centered at $\mathbf{x}$ given by the 2-norm. The Lebesgue differentiation theorem implies that the set defined by

$$E = \left\{ \mathbf{x} \colon \lim_{r \to 0} \frac{1}{\mu(B_r^2(\mathbf{x}))} \int_{B_r^2(\mathbf{x})} \mathbb{1}_{\partial S}(\mathbf{y}) dy \neq \mathbb{1}_{\partial S}(\mathbf{x}) \right\}$$

has measure zero. We will show $\partial S \subset E$, which will imply $\mu(\partial S) = 0$.

This amounts to showing that for $\mathbf{x} \in \partial S$,

$$\lim_{r \to 0} \frac{1}{\mu(B_r^2(\mathbf{x}))} \int_{B_r^2(\mathbf{x})} \mathbb{1}_{\partial S}(\mathbf{y}) dy = \lim_{r \to 0} \frac{\mu(\partial S \cap B_r^2(\mathbf{x}))}{\mu(B_r^2(\mathbf{x}))} \neq 1$$

Specifically, we will show that for sufficiently small r, there exists a constant $K > 0$ independent of r for which

$$\frac{\mu(\operatorname{int} S \cap B_r^2(\mathbf{x}))}{\mu(B_r^2(\mathbf{x}))} \geq K > 0.$$

This inequality will imply the result as

$$\lim_{r \to 0} \frac{\mu(\partial S \cap B_r^2(\mathbf{x}))}{\mu(B_r^2(\mathbf{x}))} \leq 1 - \liminf_{r \to 0} \frac{\mu(\operatorname{int} S \cap B_r^2(\mathbf{x}))}{\mu(B_r^2(\mathbf{x}))} - \liminf_{r \to 0} \frac{\mu(\operatorname{int} S^C \cap B_r^2(\mathbf{x}))}{\mu(B_r^2(\mathbf{x}))}.$$

Pick $\mathbf{x}_0 \in \partial S$. Then by assumption, there is an open convex set C for which $\mathbf{x}_0 \in \partial C$ and $C \subset S$. As C is open, $C \subset \operatorname{int} S$. Furthermore, $B_1^2(\mathbf{x}_0) \cap C$ is non-empty, open, and convex. Thus we can pick d points $\mathbf{x}_1 \ldots \mathbf{x}_d \in C$ for which the vectors $\{\mathbf{x}_1 - \mathbf{x}_0 \ldots \mathbf{x}_d - \mathbf{x}_0\}$ are linearly independent. By the convexity of $C \subset \operatorname{int} S$, the interior of the convex hull of $\{\mathbf{x}_0 \ldots \mathbf{x}_d\}$ is contained in $\operatorname{int} S$. We will call the interior of this convex hull T. By construction, for any r, $T \cap B_r^2(\mathbf{x}_0)$ is disjoint from ∂S and contained in S. This implies

$$\frac{\mu(\operatorname{int} S \cap B_r^2(\mathbf{x}))}{\mu(B_r^2(\mathbf{x}))} \geq \frac{\mu(T \cap B_r^2(\mathbf{x}))}{\mu(B_r^2(\mathbf{x}))}$$

We we will show that for $r < \min_{i \in [1,n]} \|\mathbf{x}_i - \mathbf{x}_0\|$,

$$\frac{\mu(T \cap B_r^2(\mathbf{x}))}{\mu(B_r^2(\mathbf{x}))} \geq K > 0$$

for some constant K.

Specifically, if $r < \min_{i \in [1,n]} \|\mathbf{x}_i - \mathbf{x}_0\|$, $B_r^2(\mathbf{x})$ contains $\mathbf{x}_0 + r \frac{\mathbf{x}_i - \mathbf{x}_0}{\|\mathbf{x}_i - \mathbf{x}_0\|}$ for each i. Then because $B_r^2(\mathbf{x})$ is convex, it must contain the simplex defined by these vectors which we will call W. See Figure B.2 for an illustration. A standard calculation shows that $\mu(W) = \frac{r^d}{n!} |\det(M)|$, where

$$M = \left[\frac{\mathbf{x}_1 - \mathbf{x}_0}{\|\mathbf{x}_1 - \mathbf{x}_0\|} \cdots \frac{\mathbf{x}_n - \mathbf{x}_0}{\|\mathbf{x}_n - \mathbf{x}_0\|} \right]$$

see for instance Stein (1966).

Therefore, as the volume of a ball radius r is $\frac{\pi^{\frac{d}{2}} r^d}{\Gamma(\frac{d}{2}+1)}$, we have shown that for $r < \min_{i \in [1,n]} \|\mathbf{x}_i - \mathbf{x}_0\|$,

$$\frac{\mu(T \cap B_r^2(\mathbf{x}))}{\mu(B_r^2(\mathbf{x}))} \geq \frac{\mu(W)}{\mu(B_r^2(\mathbf{x}))} = \frac{\frac{r^d}{d!} |\det M|}{\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} r^d} = \frac{\Gamma(\frac{d}{2}+1) |\det M|}{d! \pi^{\frac{d}{2}}} > 0$$

$\square$

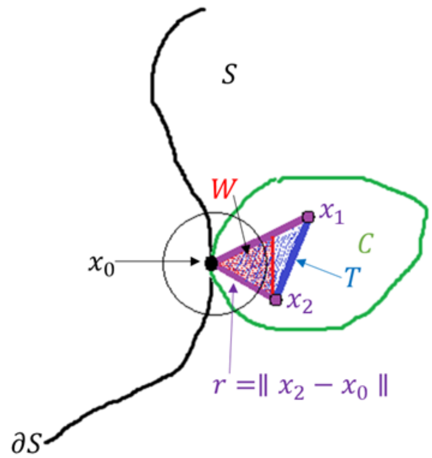

Figure 3: The convex set C, the simplex T, and the simplex W in two dimensions. The illustrated ball has radius less than r. In this Figure, $r = \|\mathbf{x}_2 - \mathbf{x}_0\|$.

C Properties of the $^{\epsilon}$,$^{-\epsilon}$ Operations

In this section, we will discuss some basic properties of the $^{\epsilon}$,$^{-\epsilon}$ operations. We will apply these results throughout the rest of the appendix. Furthermore, this section should highlight some of the intuition for working with these set operations.

We will adopt (11) as our definition for A^{ϵ}.

This convention will allow us to generalize much of the results in this section to arbitrary metric spaces. After defining A^{ϵ} we can then define $A^{-\epsilon}$ as

$$A^{-\epsilon} = ((A^C)^{\epsilon})^C. \tag{15}$$

The following lemma details how the $^{\epsilon}$ and $^{-\epsilon}$ operations interact with unions and intersections.

Lemma 12. *Define A^{ϵ} as in (11) and $A^{-\epsilon}$ as (15). Then for any sequence of sets $\{A_i\}$, the following set containments hold:*

$$\bigcup_{i=1}^{\infty} A_i^{\epsilon} = \left[\bigcup_{i=1}^{\infty} A_i\right]^{\epsilon} \tag{16}$$

$$\bigcap_{n\geq 1} A_n^{-\epsilon} = \left(\bigcap_{n\geq 1} A_n\right)^{-\epsilon} \tag{17}$$

$$\bigcap_{i=1}^{\infty} A_i^{\epsilon} \supset \left[\bigcap_{i=1}^{\infty} A_i\right]^{\epsilon} \tag{18}$$

$$\bigcup_{i=1}^{\infty} A_i^{-\epsilon} \subset \left[\bigcup_{i=1}^{\infty} A_i\right]^{-\epsilon} \tag{19}$$

Proof. **Showing (16):**
For any set A, one can write

$$A^{\epsilon} = \bigcup_{\mathbf{a}\in A} \overline{B_{\epsilon}(\mathbf{a})}.$$

Thus

$$\bigcup_{i=1}^{\infty} A_i^{\epsilon} = \bigcup_{i=1}^{\infty} \bigcup_{\mathbf{a}\in A_i} \overline{B_{\epsilon}(\mathbf{a})} = \bigcup_{\mathbf{a}\in \bigcup_{i=1}^{\infty} A_i} \overline{B_{\epsilon}(\mathbf{a})} = \left(\bigcup_{i=1}^{\infty} A_i\right)^{\epsilon}.$$

Showing (18):
First note that if $C \supset B$, then $C^{\epsilon} \supset B^{\epsilon}$. Next, since $A_i \supset \bigcap_{i=1}^{\infty} A_i$,

$$A_i^{\epsilon} \supset \left(\bigcap_{j=1}^{\infty} A_j\right)^{\epsilon}$$

for all i. Thus (18) holds.

Showing (17):
Recall that $A^{-\epsilon} = ((A^C)^\epsilon)^C$. If we apply (16) to $(A_i^C)^\epsilon$, we get that

$$\bigcup_{i=1}^{\infty} (A_i^C)^\epsilon = \left(\bigcup_{i=1}^{\infty} A_i^C\right)^\epsilon = \left(\left(\bigcap_{i=1}^{\infty} A_i\right)^C\right)^\epsilon.$$

Now upon taking complements,

$$\left(\bigcap_{i=1}^{\infty} A_i\right)^{-\epsilon} = \left(\bigcup_{i=1}^{\infty} (A_i^C)^\epsilon\right)^C = \bigcap_{i=1}^{\infty} \left((A_i^C)^\epsilon\right)^C = \bigcap_{i=1}^{\infty} A_i^{-\epsilon}.$$

Showing (19): If we apply (18) to A_i^C, then

$$\bigcap_{i=1}^{\infty} (A_i^C)^\epsilon \supset \left(\bigcap_{i=1}^{\infty} A_i^C\right)^\epsilon = \left(\left(\bigcup_{i=1}^{\infty} A_i\right)^C\right)^\epsilon.$$

Taking complements gives (19). □

Next, we use the previous representations to show that $F(A^\epsilon) = \emptyset$ and $F((A^{-\epsilon})^C) = \emptyset$, where we define $F(\cdot)$ in (4).

Lemma 13. *For a set A, define*

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

Then

$$F(A^\epsilon) = \emptyset \tag{20}$$

$$F((A^{-\epsilon})^C) = \emptyset \tag{21}$$

This lemma is an important stepping stone towards showing that there exists a pseudo-certifiably robust adversarial Bayes classifier.

Proof of Lemma 13. Equation 11 implies that each point $\mathbf{x}$ in A^ϵ is included in some closed ϵ-ball that is contained in A^ϵ. Subsequently, the definition of F in (4) implies (20). Lastly, (21) follows by applying (20) to $(A^{-\epsilon})^C$. □

The next lemma provides an alternative interpretation of the $^\epsilon, ^{-\epsilon}$ operations.

Lemma 14. *Define $A^\epsilon, A^{-\epsilon}$ as in (11),(15). Then alternative characterizations of $A^\epsilon, A^{-\epsilon}$ are given by*

$$A^\epsilon = \{\mathbf{x} \in X : \overline{B_\epsilon(\mathbf{x})} \cap A \neq \emptyset\} \tag{22}$$

$$A^{-\epsilon} = \{\mathbf{a} : \overline{B_\epsilon(\mathbf{a})} \subset A\} \tag{23}$$

Notice that in $\mathbb{R}^d$ (23) reduces to

$$A^{-\epsilon} = \{\mathbf{a} \in A : \mathbf{a} + \mathbf{h} \in A \text{ for all } \mathbf{h} \text{ with } \|\mathbf{h}\| \leq \epsilon\}$$

Proof of Lemma 14. **Showing (22):**
Recall that $\mathbf{z} \in A^\epsilon$ iff for some $\mathbf{a} \in A$, $\mathbf{z} \in \overline{B_\epsilon(\mathbf{a})}$. However,

$$\mathbf{z} \in \overline{B_\epsilon(\mathbf{a})} \Leftrightarrow \mathbf{a} \in \overline{B_\epsilon(\mathbf{z})} \Leftrightarrow \overline{B_\epsilon(\mathbf{z})} \text{ intersects } A$$

Showing (23):
Recall the definition $A^{-\epsilon} = ((A^C)^\epsilon)^C$. Then

$$\begin{aligned}
&\mathbf{a} \in A^{-\epsilon} \\
&\Leftrightarrow \mathbf{a} \notin (A^C)^\epsilon \\
&\Leftrightarrow \overline{B_\epsilon(\mathbf{a})} \text{ does not intersect } A^C \text{ (by (22))} \\
&\Leftrightarrow \overline{B_\epsilon(\mathbf{a})} \subset A
\end{aligned}$$

□

D Proof of Lemma 3

In some of our proofs, we apply the $^\epsilon$ and $^{-\epsilon}$ operations to sets multiple times in succession. In this section, we describe how the $^\epsilon$ and the $^{-\epsilon}$ operations interact. These considerations turn out to be important because applying $^\epsilon$ followed by $^{-\epsilon}$ to a set (or vice versa) decreases the adversarial loss. We prove this statement in Lemmas 3 and 19, which are the central conclusions of this Appendix.

Our first result states that applying $^{-\epsilon}$ an then $^\epsilon$ to a set A makes the set smaller while applying $^\epsilon$ and then $^{-\epsilon}$ makes the set larger.

Lemma 15. *Define the* $^\epsilon$,$^{-\epsilon}$ *operations as in* (11), (15). *Then*

$$(A^\epsilon)^{-\epsilon} \supset A \tag{24}$$

$$(A^{-\epsilon})^\epsilon \subset A \tag{25}$$

Proof. To start, note that (25) follows from applying (24) to A^C and then taking complements.

In order to show (24), we make use of Equation 23. Equation 23 implies that if $\mathbf{x} \in A^{-\epsilon}$, then $\overline{B_\epsilon(\mathbf{x})} \subset A$. As

$$(A^{-\epsilon})^\epsilon = \bigcup_{\mathbf{x}\in A^{-\epsilon}} \overline{B_\epsilon(\mathbf{x})}$$

and each $\overline{B_\epsilon(\mathbf{x})}$ is entirely contained in A, the entire set $(A^{-\epsilon})^\epsilon$ is contained in A as well. □

Lemma 16. *Define* A^ϵ, $A^{-\epsilon}$ *as in* (11),(15). *Then the following hold:*

$$A = (A^{-\epsilon})^\epsilon \sqcup F(A) \tag{26}$$
$$(A^\epsilon)^{-\epsilon} = A \sqcup F(A^C). \tag{27}$$

Specifically, (26) implies that $(A^{-\epsilon})^\epsilon = A - F(A)$ and (27) implies that $(A^\epsilon)^{-\epsilon} = A \cup F(A^C)$. Figure 2 illustrates the sets $F(A)$ and $F(A^C)$.

Proof of Lemma 16.
Showing $\supset$ for (26):
It's clear that $F(A) \subset A$ and Lemma 15 implies that $(A^{-\epsilon})^\epsilon \subset A$ as well.
Showing $\subset$ for (26):
We will prove that $A - F(A) \subset (A^{-\epsilon})^\epsilon$. Assume that $\mathbf{x} \in A - F(A)$. Then there is a closed ϵ-ball containing $\mathbf{x}$ that does not intersect A^C, which means that this ball is completely contained in A. Thus for some $\mathbf{a} \in A$, $\mathbf{x} \in \overline{B_\epsilon(\mathbf{a})} \subset A$. Thus by (23), $\mathbf{a} \in A^{-\epsilon}$. Furthermore, $\mathbf{x} \in \overline{B_\epsilon(\mathbf{a})}$ implies that $\mathbf{x} \in (A^{-\epsilon})^\epsilon$.

Showing disjoint union for (26):

Lemma 15 states that $(A^{-\epsilon})^\epsilon \subset A$. Specifically, every point in $(A^{-\epsilon})^\epsilon$ is contained in a closed ϵ-ball that is contained in A. As no point in $F(A)$ satisfies this property, $(A^{-\epsilon})^\epsilon$ and $F(A)$ are disjoint.

Showing (27):

Applying (26) to A^C results in

$$A^C = ((A^C)^{-\epsilon})^{\epsilon} \sqcup F(A^C) = ((A^{\epsilon})^C)^{\epsilon} \sqcup F(A^C).$$

Taking complements of both sides of this equation produces

$$A = (A^{\epsilon})^{-\epsilon} \cap F(A^C)^C$$

and therefore

$$A \cup \big((A^{\epsilon})^{-\epsilon} \cap F(A^C)\big) = (A^{\epsilon})^{-\epsilon}.$$

The union is actually a disjoint union because $F(A^C) \subset A^C$ which is disjoint from A. It remains to show that $F(A^C) \subset (A^{\epsilon})^{-\epsilon}$, so that $F(A^C) \cap (A^{\epsilon})^{-\epsilon} = F(A^C)$.

We now show that $F(A^C) \subset (A^{\epsilon})^{-\epsilon}$. Pick $\mathbf{x} \in F(A^C)$. We will show that for every $\mathbf{y} \in \overline{B_{\epsilon}(\mathbf{x})}$, $\mathbf{y} \in A^{\epsilon}$. This statement will imply that $\overline{B_{\epsilon}(\mathbf{x})} \subset A^{\epsilon}$ and then (23) will then imply that $\mathbf{x} \in (A^{\epsilon})^{-\epsilon}$.

If $\mathbf{y} \in \overline{B_{\epsilon}(\mathbf{x})}$, then $\overline{B_{\epsilon}(\mathbf{y})}$ contains $\mathbf{x}$. By definition, because $\mathbf{x} \in F(A^C)$, every ball containing $\mathbf{x}$ intersects A. Therefore $\overline{B_{\epsilon}(\mathbf{y})}$ intersects A and then (22) then implies that $\mathbf{y} \in A^{\epsilon}$.

□

In the previous lemma, we characterized $(A^{-\epsilon})^{\epsilon}$ and $(A^{\epsilon})^{-\epsilon}$, in terms of A and $F(\cdot)$ but this characterization is a little complicated. Here, we show that if in fact $A = B^{-\epsilon}$ some set B, then $(A^{\epsilon})^{-\epsilon}$ simplifies. Similarly, $(A^{-\epsilon})^{\epsilon}$ simplifies if in fact $A = B^{\epsilon}$ for some set B.

Lemma 17. *For any set A, the following hold:*

$$\Big((A^{\epsilon})^{-\epsilon}\Big)^{\epsilon} = A^{\epsilon}, \qquad \Big((A^{-\epsilon})^{\epsilon}\Big)^{-\epsilon} = A^{-\epsilon}.$$

Proof of Lemma 17. By Lemmas 13 and 16,

$$\Big((A^{\epsilon})^{-\epsilon}\Big)^{\epsilon} = (\Big(A^{\epsilon}\Big)^{-\epsilon})^{\epsilon} = A^{\epsilon} - F(A^{\epsilon}) = A^{\epsilon}.$$

Similarly,

$$\Big((A^{-\epsilon})^{\epsilon}\Big)^{-\epsilon} = (\Big(A^{-\epsilon}\Big)^{\epsilon})^{-\epsilon} = A^{-\epsilon} \cup F((A^{-\epsilon})^C) = A^{-\epsilon}.$$

□

We next prove a short lemma that will help us understand how the $-\epsilon, \epsilon$ operations reduce the adversarial loss.

Lemma 18. *Let $^{\epsilon}, ^{-\epsilon}$ be as in (11) and (15). Consider a set $B \subset X$. Then if $D = (B^{-\epsilon})^{\epsilon}$ and $C = (B^{\epsilon})^{-\epsilon}$, then $C^{\epsilon} \subset B^{\epsilon}, C^{-\epsilon} \supset B^{-\epsilon}$ and $D^{\epsilon} \subset B^{\epsilon}, D^{-\epsilon} \supset B^{-\epsilon}$.*

Proof. First consider the set D. Then by Lemma 17, $D^{-\epsilon} = B^{-\epsilon}$. Furthermore, according to Lemma 15, $D \subset B$, so that $D^{\epsilon} \subset B^{\epsilon}$.

Next, according to Lemma 17, $C^{\epsilon} = B^{\epsilon}$. Furthermore, according to Lemma 15, $C \supset B$, so that $C^{-\epsilon} \supset B^{-\epsilon}$. □

Lastly, we prove a lemma which states that applying the $^{\epsilon}, ^{-\epsilon}$ operations in succession decreases the adversarial loss. Observe that R^{ϵ} incurs a penalty of 1 on both $F(A)$ and $F(A^C)$ because points in these sets are always within ϵ of a point with the opposite class label.

Lemma 19. *For any set A, the following hold:*

$$R^{\epsilon}(A) \geq R((A^{\epsilon})^{-\epsilon}) \tag{28}$$

$$R^{\epsilon}(A) \geq R((A^{-\epsilon})^{\epsilon}). \tag{29}$$

Proof of Lemma 19. The basic idea here is that the maximum penalty is incurred on $F(A)$, so removing $F(A)$ from A and adding it to A^C will not increase the loss. (Compare the statement of this lemma with Lemma 16 and Figure 2.) The same holds for $F(A^C)$ and A^C.

Let $B = (A^{-\epsilon})^\epsilon$ or $B = (A^\epsilon)^{-\epsilon}$. Lemma 18 implies that $B^\epsilon \subset A^\epsilon$ and $B^{-\epsilon} \supset A^{-\epsilon}$. These containments imply the result because if $B^\epsilon \subset A^\epsilon$ and $B^{-\epsilon} \supset A^{-\epsilon}$ then

$$\eta(\mathbf{x})\mathbb{1}_{A^\epsilon} + (1-\eta(\mathbf{x}))\mathbb{1}_{(A^C)^\epsilon} \geq \eta(\mathbf{x})\mathbb{1}_{B^\epsilon} + (1-\eta(\mathbf{x}))\mathbb{1}_{(B^C)^\epsilon}$$

holds pointwise, so

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

□

By taking $B = (A^{-\epsilon})^\epsilon$, $E = (A^\epsilon)^{-\epsilon}$, Lemma 3 immediately follows from Lemma 19 and the definition of the $^\epsilon$ operation.

Lemma 3. *Let A be any set. Then there exist sets B, E for which B and E^C are pseudo-certifiably robust and $R^\epsilon(B) \leq R^\epsilon(A)$, $R^\epsilon(E) \leq R^\epsilon(A)$.*

E Proof of Lemma 2 and a Generalization (Lemma 20)

We begin by reviewing some results of Appendix C. To start, recall that the operation $A^\epsilon = A \oplus \overline{B_\epsilon(\mathbf{0})}$, satisfies the relations of (7):

$$\left(\bigcup_{i=1}^\infty A_i\right)^\epsilon = \bigcup_{i=1}^\infty A_i^\epsilon \quad , \quad \left(\bigcap_{i=1}^\infty A_i\right)^\epsilon \subset \bigcap_{i=1}^\infty A_i^\epsilon \tag{7}$$

In the next section, we will prove a version of Lemma 2 for other models of perturbations. Thus, in the rest of this appendix, rather than focusing on $\mathbb{R}^d$, we will assume that $^\epsilon$ is a set operation that satisfies (7). This formulation will allow us to prove an existence theorem for other models of perturbations. As elements of our space X are not necessarily vectors, we write them in non-bold font (x). We now state a generalized version of Lemma 2.

Lemma 20. *Let A^ϵ be any set operation that satisfies (7). Then*

$$\limsup A_n^\epsilon \supset (\limsup A_n)^\epsilon \quad \textit{and} \quad \liminf A_n^\epsilon \supset (\liminf A_n)^\epsilon$$

Note that Lemma 2 is simply Lemma 20 combined with the fact that A^ϵ defined as $A \oplus \overline{B_\epsilon(\mathbf{0})}$ satisfies (7) (shown in Lemma 12).

Proof of Lemma 20. We start by proving the statement for $\limsup$. By (7),

$$\limsup A_n^\epsilon = \bigcap_{N=1}^\infty \bigcup_{n=N}^\infty A_n^\epsilon = \bigcap_{N=1}^\infty \left(\bigcup_{n=N}^\infty A_n\right)^\epsilon \supset \left(\bigcap_{N=1}^\infty \bigcup_{n=N}^\infty A_n\right)^\epsilon$$

The statement for $\liminf$ follows from a similar argument:

$$\liminf A_n^\epsilon = \bigcup_{N=1}^\infty \bigcap_{n=N}^\infty A_n^\epsilon \supset \bigcup_{N=1}^\infty \left(\bigcap_{n=N}^\infty A_n\right)^\epsilon = \left(\bigcup_{N=1}^\infty \bigcap_{n=N}^\infty A_n\right)^\epsilon$$

□

F More General Results

In this Appendix, we present a generalization of our main result. This generalization concerns other models of perturbations. As discussed in Section 6, there are many other possible models of perturbations in adversarial learning. A more general result would help address the existence of the adversarial Bayes classifier in these scenarios as well. We provide a motivating example in the next subsection.

Theorem 10. *Let X be a separable metric space and let $\mathcal{B}(X), \mathscr{U}(X)$ be the corresponding Borel and universal σ-algebras respectively. Let $\mathbb{P}$ be the completion of a measure on $\mathcal{B}(X)$ restricted to $\mathscr{U}(X)$. For $A \subset X$, let $^\epsilon : A \to A^\epsilon$ be a set operation for which A^ϵ is universally measurable for all sets $A \in \mathscr{U}(X)$. Furthermore, assume that $^\epsilon$ satisfies the properties*

$$\bigcup_{n\in\mathbb{N}} A_n^\epsilon = \left(\bigcup_{n\in\mathbb{N}} A_n\right)^\epsilon \tag{30}$$

$$\bigcap_{n\in\mathbb{N}} A_n^\epsilon \supset \left(\bigcap_{n\in\mathbb{N}} A_n\right)^\epsilon \tag{31}$$

for every sequence of sets $\{A_n\}$. Define the loss

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

Assume that for some minimizing sequence A_n, one can always find a subsequence A_{n_j} for which $\limsup A_{n_j}^\epsilon \doteq \liminf A_{n_j}^\epsilon$, where $\doteq$ is with respect to the measure $\mathbb{P}$. Then there exists a minimizer to R^ϵ in the σ-algebra $\mathscr{U}(X)$.

If A^ϵ is defined by perturbations in a metric space, Theorem 4 could be used to conclude that A^ϵ is universally measurable.

Just like Theorem 1, one could also define the $^{-\epsilon}$ operation as $A^{-\epsilon} = ((A^C)^\epsilon)^C$, and then argue that there exists a pseudo-certifiably robust minimizer, if A^ϵ is universally measurable for universally measurable A. This statement follows from the same argument as Lemma 19.

F.1 A Motivating Example–Applying Theorem 10

To show the utility of Theorem 10, we present an application inspired by NLP. For clarity, we choose a model of discrete perturbations somewhat simpler than Example 3. Let X be all strings of finite length with a finite alphabet $\mathcal{A}$. This space is countable and therefore separable. Furthermore, this space is discrete. Recall that in a discrete space, every set is measurable. Hence, the Borel σ-algebra consists of all subsets of X, which implies that $\mathscr{U}(X)$ and $\mathcal{B}(X)$ are equal.

We will define our perturbations as swapping two letters in a string at specified positions. Formally, for $w \in X$, let $|w|$ denote the length of the string. Furthermore, let T be the set of functions defined by

$$T = \Big\{ b^{i,j} \colon X \to X \Big| b^{i,j}(w)_k = w_k \text{ if } k \neq i,j \text{ or } \max(i,j) > |w|,$$
$$b^{i,j}(w)_i = w_j, b^{i,j}(w)_j = w_i \text{ otherwise} \Big\}.$$

In other words, $b^{i,j}$ will swap the letters at i and j in w if w has length at least $\max(i,j)$ and will keep the string fixed otherwise. Now let B be a finite subset of T.

If A is a set of strings, we define

$$A^\epsilon = \{b(a) \colon a \in A, b \in B\}.$$

To start, note that for this definition of the $^\epsilon$ operation, we still have that for every sequence of sets A_n, the relations (30), (31) hold. The proofs are the same as (16), (18) of Lemma 12, so we do not reproduce it here.

Next, we argue that for every sequence A_n, there is a subsequence A_{n_j} for which $\liminf A_{n_j} = \limsup A_{n_j}$. The proof is similar to that of Theorem 2 presented in (Rockafellar and Wets, 1998).

Let $A = \limsup A_n$. Because the space X is countable, one can enumerate $A = \limsup A_n = \{a_n\}_{n=1}^N$, with $N \in \mathbb{N} \cup \{\infty\}$. Now we inductively define N nested subsequences of sets $\{A_n^k\}_{n=1}^\infty$ indexed by k as follows: Because $a_1 \in \limsup A_n$, by the characterization of $\limsup$ in (9), one can find a subsequence A_{n_m} for which $a_1 \in A_{n_m}$ for all m. Let $A_m^1 = A_{n_m}$.

Now given the sequence $\{A_n^k\}$, we inductively define $\{A_n^{k+1}\}$. Consider the element a_{k+1} of the sequence $A = \{a_n\}_{n=1}^N$. If $a_{k+1} \in \limsup_n A_n^k$, then one can find a subsequence $A_{n_m}^k$ for which

$a_{k+1} \in A^k_{n_m}$ for all m. Similarly, if $a_{k+1} \notin \limsup_n A^k_n$, then one can find a subsequence $A^k_{n_m}$ for which $a_{k+1} \notin A^k_{n_m}$ for all m. Thus we choose the subsequence $A^k_{n_m}$ so that either $a_{k+1} \in A^k_{n_m}$ for all m or $a_{k+1} \notin A^k_{n_m}$ for all m. Set $A^{k+1}_n = A^k_{n_m}$.

If $N = |\limsup A_n|$ is finite, consider the sequence A^N_n. Then $\limsup A^N_n \subset \limsup A_n$. Furthermore, by the construction of the sequences $\{A^k_n\}_{n=1}^N$, for each $a \in \limsup A_n$, either $a \in A^N_n$ for all n or $a \notin A^N_n$ for all n. This observation implies that if $a \in \limsup_n A^N_n$, then a is eventually in the tail of the sequence A^N_n and thus $a \in \liminf_n A^N_n$. Therefore $\liminf_n A^N_n = \limsup_N A^N_n$.

We now analyze the case $N = \infty$. Consider now the diagonal sequence $A_{n_k} = A^k_k$. Again, we have the containment $\limsup_k A^K_k \subset \limsup_n A_n$. Let a_j be an element of $A = \limsup A_n$. Then by construction of the sequences $\{A^k_n\}_{n=1}^\infty$, either $a_j \in A^k_k$ for all $k \geq j$ or $a_j \notin A^k_k$ for all $k \geq j$. Thus every $a \in \limsup A_{n_k}$ is also in A_{n_j} for all sufficiently large j so $a \in \liminf_k A_{n_k}$. Therefore, $\limsup_k A_{n_k} = \liminf_k A_{n_k}$.

F.2 Proving Theorem 10

The proof of Theorem 10 follows the same steps as the proof of Theorem 1. As mentioned in Section 6, the big picture motivation is that Theorem 1 followed directly from Lemmas 1, 2, and 3 – we did not use properties of $^\epsilon$ or the space $\mathbb{R}^d$ outside of these three Lemmas. The main challenge is generalizing these concepts. With the proper definitions, the proof of Theorem 10 is exactly the same as the proof of Theorem 1, except that we replace Lemma 1 with the assumption that for every minimizing sequence A_n there exists a subsequence A_{n_j} for which $\limsup A^\epsilon_{n_j} - \liminf A^\epsilon_{n_j}$ is a null set.

Proof of Theorem 10. Let A_n be a minimizing sequence of R^ϵ. Pick a subsequence A_{n_j} for which

$$\mathbb{P}(\limsup A^\epsilon_{n_j} - \liminf A^\epsilon_{n_j}) = 0 \tag{32}$$

and set $A = \limsup A_{n_j}$. Then

$$\begin{aligned}
\inf_{A \text{ Borel}} R^\epsilon(A) = \lim_{j\to\infty} R^\epsilon(A_j) &\geq \int \liminf_{j\to\infty} \left(\eta \mathbb{1}_{A^\epsilon_{n_j}} + (1-\eta)\mathbb{1}_{(A^C_{n_j})^\epsilon}\right) d\mathbb{P} && \text{(Fatou's Lemma)}\\
&\geq \int \eta \liminf_{j\to\infty} \mathbb{1}_{A^\epsilon_{n_j}} + (1-\eta) \liminf_{j\to\infty} \mathbb{1}_{(A^C_{n_j})^\epsilon} d\mathbb{P} \\
&= \int \eta \mathbb{1}_{\limsup_j A^\epsilon_{n_j}} + (1-\eta)\mathbb{1}_{\liminf_j (A^C_{n_j})^\epsilon} d\mathbb{P} && \text{(Equation 32)}\\
&\geq \int \eta \mathbb{1}_{\left(\limsup_j A_{n_j}\right)^\epsilon} + (1-\eta)\mathbb{1}_{\left(\liminf_j A^C_{n_j}\right)^\epsilon} d\mathbb{P} && \text{(Lemma 20)}\\
&= \int \eta \mathbb{1}_{\left(\limsup_j A_{n_j}\right)^\epsilon} + (1-\eta)\mathbb{1}_{\left((\limsup_j A_{n_j})^C\right)^\epsilon} d\mathbb{P}\\
&= R^\epsilon(A)
\end{aligned}$$

Therefore, A is a minimizer of R^ϵ. ∎