On The Existence of The Adversarial Bayes Classifier

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

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

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

Abstract

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

1 Introduction

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

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

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

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

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

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

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

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

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

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

The concept of certified robustness has algorithmic implications. Cohen et al. (2019) recently showed that after training a classifier, a process called *randomized smoothing* makes the classifier certifiably robust at a point x in the ℓ_2 norm with a radius that depends on the point x. As the adversarial Bayes classifier is pseudo-certifiably robust but not certifiably robust with a fixed radius at every point, one could try to design algorithms which ensure pseudo-certifiable robustness during or after training. Recent works have explored constructing certificates of robustness as well (Raghunathan et al., 2018; Weng et al., 2018; Zhang et al., 2018; Wong and Kolter, 2018). A better understanding of the adversarial Bayes classifier could help find additional learning algorithms. By studying the existence of the adversarial Bayes classifier, we take a first step towards this broader goal.

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

2 Related Work

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

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

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

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

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

3 Problem Setup

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

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

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 we could choose $f(\mathbf{x}) = \mathbb{1}_A(\mathbf{x})$. In the rest of the paper, we will specify the set of points A classified as $+1$ rather than the function f. The classification risk of a set A is then expressed as

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

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

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

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

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

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

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

The papers (Szegedy et al., 2013; Biggio et al., 2013; Madry et al., 2017) (and many others) use the multi-class version of this loss to define adversarial risk. More specifically, they evaluate the risk on the set $A = \{f(\mathbf{x}) \geq 0\}$, 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 $\mathscr{U}(\mathbb{R}^d)$.

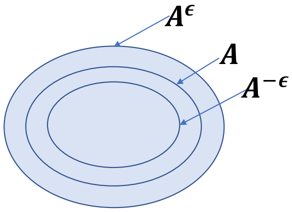

Figure 1: Sets A^ϵ and $A^{-\epsilon}$ with $B = \overline{B_\epsilon^2(\mathbf{0})}$, the closed ℓ_2 ball.

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. 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) where the sets A and A^C are disjoint. As a result, the integrand equals either $\eta(\mathbf{x})$ or $(1-\eta(\mathbf{x}))$ at each point. Thus the set for which $1-\eta(\mathbf{x}) < \eta(\mathbf{x})$ minimizes R. In other words, the Bayes classifier minimizes the objective $\min(\eta(\mathbf{x}), 1-\eta(\mathbf{x}))$ at each point.

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

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

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

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

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

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

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

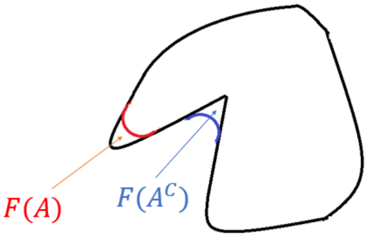

Figure 2: The figure illustrates a set A with the sets $F(A)$ and $F(A^C)$ roughly indicated. For a point $\mathbf{a} \in F(A)$, every closed ϵ-ball containing $\mathbf{a}$ also intersects A^C while for $\mathbf{a} \in F(A^C)$ every closed ϵ-ball containing $\mathbf{a}$ also intersects A.

We now state our main existence result.

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

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

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

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

4.1 Proof strategy

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.2 Formal Proof of Theorem 1

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.3 Proof Strategy for Lemma 2

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

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

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

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

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

5 Addressing Measurability

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

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

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

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

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

6 Alternative Models of Perturbations

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

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

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

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

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

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

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

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

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

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

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

7 Conclusion

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

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

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Contents of Appendix

A The Measurability of A^ϵ

In this section, we prove a version of Theorem 3 that applies to general metric spaces. We discuss the theorem in high generality for two reasons. First, the proof of Theorem 3 utilizes concepts from abstract metric spaces, measure spaces, and topology. Discussing this result in terms of these abstract concepts actually clarifies the intuition behind the proof of this theorem. Second, we suspect that our framework will be useful in discussing other models of perturbations. A general version of Theorem 3 could assist with the measurability considerations in alternative models of adversarial learning as well.

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 Proof of Measurability

For a measure space $(\mu, X, \mathcal{B}(X))$ equipped with the Borel σ-algebra $\mathcal{B}(X)$, we will denote its completion as $(\overline{\mu}, X, \mathcal{L}_\mu(X))$. 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

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

This form for A^ϵ is helpful because it allows us to define A^ϵ for general metric spaces. Notably,

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

holds for general metric spaces when we define A^ϵ in this way. The second line is similar to the expression (3) for the adversarial loss. We will use this idea later to prove a generalized version of our theorem for alternative models of perturbations.

Again, we follow the treatment of (Nishiura, 2010). We start by defining the universal σ-algebra for a measure space X.

Definition 4. *Let X be a Borel space and let $\mathscr{M}(X)$ be the set of all σ-finite Borel measures on X. We define the* universal σ-algebra *to be*

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

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

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 4.

In this section we prove the following theorem:

Theorem 4. *Let (X, d) be a separable metric space. Define A^ϵ as*

$$A^\epsilon = \bigcup_{a \in A} \overline{B_\epsilon(a)}. \tag{7}$$

If $A \subset X$ is universally measurable, then A^ϵ is universally measurable as well.

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.

Before proving Theorem 4, we define another useful concept.

Definition 5. *Let X, Y be a separable metric spaces and let $(\overline{\mu}, Y, \mathcal{L}_\mu(Y))$ be a complete σ-finite measure space. If every topological copy of X in Y is an element of $\mathcal{L}_\mu(Y)$, then X is called an* absolute measurable space.

This definition is useful due to the following theorem:

Theorem 5. *Let Z be an absolute measurable Borel space and Y a separable metrizable space. Let $f\colon Z \to Y$ be a continuous function.*

Then

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

This theorem is the implication (1) ⇒ (4) of the Purves-Darst-Grzegorek Theorem, stated on page 33 in Chapter 2.1 of (Nishiura, 2010), together with the observation that all continuous functions are Borel maps.

We now describe the basic idea behind the proof of Theorem 4 for $\mathbb{R}^d$. Let $A \subset \mathbb{R}^d$ One can write

$$\mathbb{1}_{A^\epsilon}(\mathbf{x}) = \sup_{\mathbf{a}\in A} \mathbb{1}_{\overline{B_\epsilon(\mathbf{a})}}(\mathbf{x}).$$

Let $f\colon \mathbb{R}^d \times A \to \mathbb{R}$ be given by

$$f(\mathbf{x}, \mathbf{a}) = \mathbb{1}_{\overline{B_\epsilon(\mathbf{a})}}(\mathbf{x}). \tag{8}$$

Then

$$\mathbb{1}_{A^\epsilon}(\mathbf{x}) = \sup_{\mathbf{a}\in A} f(\mathbf{x}, \mathbf{a}).$$

Set $F(\mathbf{x}) = \mathbb{1}_{A^\epsilon(\mathbf{x})}$. Then one can write A^ϵ as a level set of F.

$$A^\epsilon = F^{-1}(\{1\}) = F^{-1}([1,\infty)) = \{\mathbf{x} : f(\mathbf{x},\mathbf{a}) \geq 1 \text{ for some } \mathbf{a} \in A\} = \Pi_1(f^{-1}([1,\infty))) = \Pi_1(f^{-1}(\{1\}))$$

where $\Pi_1 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is the projection $\Pi_1(\mathbf{x}, \mathbf{y}) = \mathbf{x}$. The first and last equalities follow from the fact that the ranges of f and F are $\{0, 1\}$.

The hope is to apply Theorem 5 to the spaces $Z = \mathbb{R}^d \times \mathbb{R}^d$ and $Y = \mathbb{R}^d$ with the projection map Π_1, and then conclude that $A^\epsilon = \Pi_1(f^{-1}(\{1\})) \in \mathscr{U}(Y) = \mathscr{U}(\mathbb{R}^d)$. For this, we need to show that $\mathbb{R}^d \times \mathbb{R}^d$ is absolute measurable and that $f^{-1}(\{1\}) \in \mathscr{U}(\mathbb{R}^d \times \mathbb{R}^d)$.

We state three lemmas which imply these two facts. We give the proofs in the following subsection.

This lemma identifies a wide class of absolutely measurable spaces.

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

The next lemma computes $f^{-1}(\{1\})$ for f as in (8).

Lemma 6. *Let (X, d) be a metric space and $A \subset X$. Let $f\colon X \times A \to \mathbb{R}$ be defined by $f(x, y) = \mathbb{1}_{\overline{B_\epsilon(a)}}(x)$. Then*

$$f^{-1}(\{1\}) = (X \times A) \cap B,$$

where

$$B = \{(x, a) \in X \times A\colon d(x, a) \leq \epsilon\}.$$

Lastly, the following lemma allows us to show that $f^{-1}(\{1\})$ for f as in (8) is universally measurable.

Lemma 7. *Let X, Y be Borel spaces. If $S \in \mathscr{U}(Y)$, then $X \times S \in \mathscr{U}(X \times Y)$.*

We now formally prove Theorem 4.

Proof of Theorem 4. Let $\Pi_1 \colon X \times X \to X$ be defined by the projection $\Pi_1(x, y) = x$. This map is continuous. Recall that one can write

$$\mathbb{1}_{A^\epsilon}(x) = \sup_{a\in A} \mathbb{1}_{\overline{B_\epsilon(a)}}(x).$$

For convenience, define functions $F\colon X \to \mathbb{R}$ and $f\colon X \times A \to \mathbb{R}$ by $F(x) = \mathbb{1}_{A^\epsilon}(x)$ and $f(x, a) = \mathbb{1}_{\overline{B_\epsilon(a)}}(x)$.

Then

$$F(x) = \sup_{a \in A} f(x, a)$$

and

$$A^\epsilon = F^{-1}(\{1\}) = F^{-1}([1, \infty)) = \{x : f(x, a) \geq 1 \text{ for some } a \in A\} = \Pi_1(f^{-1}([1, \infty))) = \Pi_1(f^{-1}(\{1\})).$$

Now Lemma 6 implies that

$$\Pi_1(f^{-1}(\{1\})) = (X \times A) \cap B$$

where

$$B = \{(x, y) \in X \times X \colon d(x, y) \leq \epsilon\}.$$

Note that B is a closed set so it is Borel and thus also universally measurable. Furthermore, Lemma 7 implies that $X \times A$ is universally measurable. Therefore, $f^{-1}(\{1\})$ is universally measurable in $X \times X$.

Now note that all separable metric spaces are σ-compact. Thus Lemma 5 implies that $X \times X$ is absolute measurable. As Π_1 is continuous, Theorem 5 implies that $\Pi_1(f^{-1}(\{1\}) = A^\epsilon$ is universally measurable in X.

□

A.2 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. Let Y be a separable metric space. If $f \colon H \to Y$ is a homeomorphism, 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 homeomorphism, then $f(X)$ is the union of Borel sets as

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

because each $f(H_n)$ is closed. □

Lemma 6. *Let (X, d) be a metric space and $A \subset X$. Let $f \colon X \times A \to \mathbb{R}$ be defined by $f(x, y) = \mathbb{1}_{\overline{B_\epsilon(a)}}(x)$. Then*

$$f^{-1}(\{1\}) = (X \times A) \cap B,$$

where

$$B = \{(x, a) \in X \times A \colon d(x, a) \leq \epsilon\}.$$

Proof of Lemma 6. We will show $f^{-1}(\{1\}) \subset (X \times A) \cap B$, and $f^{-1}(\{1\}) \supset (X \times A) \cap B$ separately.

Showing $f^{-1}(\{1\}) \supset (X \times A) \cap B$: Assume that both $(x, a) \in X \times A$ and $(x, a) \in B$. First, $(x, a) \in X \times A$ implies that $a \in A$, so that (x, a) is in the domain of f. Next, if $(x, a) \in B$, then $d(x, a) \leq \epsilon$ so that $f(x, a) = \mathbb{1}_{\overline{B_\epsilon(a)}}(x) = 1$. (Note that (x, a) is in the domain of f is implicit in this statement). Therefore $(x, a) \in f^{-1}(\{1\})$.

Showing $f^{-1}(\{1\}) \subset (X \times A) \cap B$: Assume $(x, a) \in f^{-1}(\{1\})$. Then because $\mathbb{1}_{\overline{B_\epsilon(a)}}(x) = 1$, $d(x, a) \leq \epsilon$ so that $(x, a) \in B$. Next, as the domain of f is $X \times A$, it's clear that $(x, a) \in X \times A$. Therefore, $(x, a) \in (X \times A) \cap B$.

□

Lemma 7. *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 7. Let $(\nu, X \times Y, \mathcal{B}(X \times Y))$ be an arbitrary σ-finite Borel measure on $X \times Y$. 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 σ-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 $\lambda(X \times N) = \nu(N) = 0$. 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. □

B Proof of Corollaries 1 and 2–Variants of Prokhorov's Theorem

In this section, we prove Corollary 1 which we used in the proof of our main Theorem 1 in Section 4.2. However, we first state a more general version that will also be useful for proving an existence result for other models of perturbations. We begin with defining tightness for metric spaces and stating a version of Prokhorov's theorem for metric spaces as well.

In this case, for a metric space X, we consider measures on the Borel σ-algebra $\mathcal{B}(X)$. We say that a sequence of probability measures $\{\mathbb{P}_n\}$ on $(X, \mathcal{B}(X))$ is *tight* if for all ϵ there exists a compact set K for which $\mathbb{P}_n(X \setminus K) < \epsilon$ independently of n. Furthermore, a sequence of measures $\{\mathbb{P}_n\}$ *weakly converges* to a measure $\mathbb{P}$ if for all continuous bounded f

$$\lim_{n \to \infty} \int_X f d\mathbb{P}_n \to \int_X f d\mathbb{P}.$$

Prokhorov's theorem now relates these two concepts.

Theorem 6 (Prokhorov's Theorem)**.** *Let (X, d) be a separable metric space. A sequence of probability measures $\{\mathbb{P}_n\}$ on the σ-algebra $\mathcal{B}(X)$ admits a weakly convergent subsequence iff $\{\mathbb{P}_n\}$ is tight.*

See (Billingsley, 1999) for more details.

Next, we generalize this result to the measure space $(X, \mathscr{U}(X))$ with the universal σ-algebra $\mathscr{U}(X)$.

Corollary 2 (Prokhorov's Theorem)**.** *Let X be a separable metric space and let $(\mathbb{P}_n, X, \mathscr{U}(X))$ be a sequence of probability measure spaces, where each $\mathbb{P}_n$ is the completion of a Borel measure restricted to $\mathscr{U}(X)$. Then $\{\mathbb{P}_n\}$ admits a weakly convergent subsequence iff the sequence is tight.*

This statement is just like Corollary 1 but with $\mathbb{R}^d$ replaced by the separable metric space X.

We start by stating a lemma that relates the integral of function f with respect to a Borel measure $\mathbb{Q}$ and the integral of f with respect to the extension of $\mathbb{Q}$ to the universal σ-algebra.

Lemma 8. *Let $\mathbb{Q}$ be a Borel measure and let $\mathbb{P}$ be the completion of $\mathbb{Q}$ restricted to the universal σ-algebra. Then for every Borel measurable function f, $\int f d\mathbb{P} = \int f d\mathbb{Q}$*

The proof of this Lemma is presented at the end of this section. Next, the following Lemma allows one to canonically relate a sequence of measures on $\mathscr{U}(X)$ with a sequence of measures on $\mathcal{B}(X)$.

Lemma 9. *Let $\mathbb{P}$ be a measure on the universal σ-algebra $\mathscr{U}(X)$ which is the restriction to $\mathscr{U}(X)$ of the completion of a Borel measure. Let $\mathbb{Q}$ be $\mathbb{P}$ restricted to the Borel σ-algebra $\mathcal{B}(X)$. Then $\mathbb{P}$ is the completion of $\mathbb{Q}$ restricted to $\mathscr{U}(X)$.*

Proof. Let μ be the Borel measure for which $\mathbb{P}$ is $\overline{\mu}$ restricted to $\mathscr{U}(X)$. Then as $\overline{\mu}(B) = \mu(B)$ on Borel sets B, in fact $\mathbb{Q} = \mu$. Thus the completions of μ and $\mathbb{Q}$ are actually the same, $(\overline{\mathbb{Q}}, X, \overline{\mathcal{B}(X)})) = (\overline{\mu}, X, \overline{\mathcal{B}(X)}$. Therefore, the restriction of these two measures to the universal σ-algebra is the same. In other words,

$$(\mathbb{P}, X, \mathscr{U}(X)) = (\overline{\mu}, X, \mathscr{U}(X)) = (\overline{\mathbb{Q}}, X, \mathscr{U}(X)).$$

□

Using Lemmas 8 and 9, we now prove Corollary 2.

Proof of Corollary 2. Let $(\mathbb{P}_n, X, \mathscr{U}(X))$ be a sequence of probability measure spaces with $\{\mathbb{P}_n\}$ tight. As $\mathscr{U}(X) \supset \mathcal{B}(X)$, we can define measures $\mathbb{Q}_n$ on $\mathcal{B}(X)$ by

$$\mathbb{Q}_n = \mathbb{P}_n\big|_{\mathcal{B}(X)}.$$

As compact sets are in the Borel σ-algebra, the tightness of $\mathbb{P}_n$ implies the tightness of $\mathbb{Q}_n$. Thus Prokhorov's Theorem (Theorem 6) implies that $\mathbb{Q}_n$ has a weakly convergent subsequence. Let $\mathbb{Q}$ be the limit of this subsequence.

Now consider the completion of $\mathbb{Q}$ which is the measure space $(\overline{\mathbb{Q}}, X, \mathcal{L}_{\mathbb{Q}}(X))$. Let $\mathbb{P}$ be the measure $\overline{\mathbb{Q}}$ restricted to the universal σ-algebra. We will show that $\mathbb{P}_n$ converges weakly to $\mathbb{P}$, which means that for each continous bounded function f, $\lim_{n\to\infty}\int f d\mathbb{Q}_n = \int f d\mathbb{Q}$.

For this, we apply Lemmas 8 and 9. First, Lemma 9 implies that each $\mathbb{P}_n$ is the restriction to $\mathscr{U}(X)$ of the completion of $\mathbb{Q}_n$. Then Lemma 8 implies that for all Borel functions f,

$$\int f d\mathbb{P} = \int f d\mathbb{Q} \text{ and } \int f d\mathbb{P}_n = \int f d\mathbb{Q}_n.$$

Therefore, for all continuous bounded functions f,

$$\lim_{n\to\infty}\int f d\mathbb{P}_n = \lim_{n\to\infty}\int f d\mathbb{Q}_n = \int f d\mathbb{Q} = \int f d\mathbb{P}.$$

Therefore $\mathbb{P}_n$ weakly converges to $\mathbb{P}$.

□

Lastly, we present the proof of Lemma 8.

Proof of Lemma 8. To start, recall that the integral of a measurable function with respect to a measure μ is defined as

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu,$$

where $f^+ = \max(f, 0)$, $f^- = \max(-f, 0)$. Thus, it suffices to verify $\int f d\mathbb{P} = \int f d\mathbb{Q}$ for positive Borel measurable functions. Now recall that the integral of a positive function is defined as

$$\int f d\mu = \sup_{\substack{g\leq f \\ g \text{ measurable simple function}}} \int g\, d\mu.$$

Given a measure space (μ, Σ), let $\mathcal{F}(\Sigma)$ denote the set of simple functions for the σ-algebra Σ.

Formally,

$$\mathcal{F}(\Sigma) := \left\{\sum_{i=1}^n a_i \mathbb{1}_{E_i} \middle| a_i \in \mathbb{R}, E_i \in \Sigma\right\}$$

and the integral of a simple function is defined as

$$\int \sum_{i=1}^n a_i \mathbb{1}_{E_i} d\mu = \sum_{i=1}^n a_i \mu(E_i).$$

As $\mathbb{P}$ and $\mathbb{Q}$ match on Borel sets, if $g \in \mathcal{F}(\mathcal{B}(X))$, then $\int g d\mathbb{P} = \int g d\mathbb{Q}$. Therefore, as $\mathcal{B}(X) \subset \mathscr{U}(X)$, we have that $\mathcal{F}(\mathcal{B}(X)) \subset \mathcal{F}(\mathscr{U}(X))$. Thus if f is in fact Borel,

$$\int f d\mathbb{Q} = \sup_{\substack{g\leq f \\ g\in\mathcal{F}(\mathcal{B}(X))}} \int g d\mathbb{Q} = \sup_{\substack{g\leq f \\ g\in\mathcal{F}(\mathcal{B}(X))}} \int g d\mathbb{P} \leq \sup_{\substack{g\leq f \\ g\in\mathcal{F}(\mathscr{U}(X))}} g d\mathbb{Q} = \int f d\mathbb{P}.$$

We will now demonstrate the opposite inequality. Specifically, we will show that if g is a simple function in $\mathcal{F}(\mathscr{U}(X))$, then there is a simple function h in $\mathcal{F}(\mathcal{B}(X))$ for which $\int g d\mathbb{P} = \int h d\mathbb{P} = \int h d\mathbb{Q}$, and thus $\int f d\mathbb{Q} \geq \int f d\mathbb{P}$ for a Borel measurable function f.

To start, pick $g \in \mathcal{F}(\mathscr{U}(X))$. Then $g = \sum_{i=1}^n a_i \mathbb{1}_{E_i}$ where the E_is are universally measurable. Thus we can decompose each E_i as $E_i = B_i \cup N_i'$ where N_i' is a subset of a null Borel set N_i, and $\mathbb{P}(E_i) = \mathbb{P}(B_i)$. Now set

$$h = \sum_{i=1}^n a_i \mathbb{1}_{B_i}.$$

Then

$$\int g d\mathbb{P} = \sum_{i=1}^n a_i \mathbb{P}(E_i) = \sum_{i=1}^n a_i \mathbb{P}(B_i) = \sum_{i=1}^n a_i \mathbb{Q}(B_i) = \int h d\mathbb{Q}.$$

This completes the proof. □

C Proof of Lemma 1 and a Generalization (Lemma 10)

This appendix is the most technical, the reader may want to skim at a first reading. We will prove a generalized version of Lemma 1 that will be useful when considering other models of perturbations.

Lemma 10. *Let X be a separable metric space. Assume that $\mathbb{P}_n$ weakly converges to $\mathbb{Q}$ with $\mathbb{P}_n$ given by $\mathbb{P}_n(B) = \mathbb{P}(B \cap A_n)$, for a sequence of sets A_n. Then, for some subsequence A_{n_j} of A_n, $\mathbb{Q}(B) = \mathbb{P}(A \cap B)$ for a set A that is given by*

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

and further $\mathbb{1}_{A_{n_j}} \to \mathbb{1}_A$ $\mathbb{P}$-a.e., and $\mathbb{1}_{A_n} \to \mathbb{1}_A$ in measure.

As $\mathbb{R}^d$ is a separable metric space, Lemma 1 follows immediately from Lemma 10. Throughout this section, because we work in a general metric space X which may not be a vector space, we write elements of X as non-bold (x).

C.1 Main Proof

We prove Lemma 10 as Lemma 1 is a special case. To begin, we state two lemmas that will motivate the proof of Lemma 10. Furthermore, we use these lemmas as intermediate steps.

This lemma states that for our specific sequence of measures, weak convergence actually implies a stronger type of convergence called *set-wise convergence*.

Lemma 11. *Let X be a separable metric space endowed with the universal σ-algebra. Assume that $\mathbb{P}_n$ weakly converges to $\mathbb{Q}$ with $\mathbb{P}_n$ given by $\mathbb{P}_n(E) = \mathbb{P}(E \cap A_n)$ for a sequence of sets A_n. Then, in fact for every measurable set E,*

$$\lim_{n \to \infty} \mathbb{P}_n(E) = \mathbb{Q}(E).$$

The following Lemma describes convergence of sets.

Lemma 12. *Let $\{A_n\}$ be a sequence of sets and assume that $\mathbb{P}(A_n \cap E) \to \mathbb{P}(A \cap E)$ for some set A. Then, there is a subsequence of sets $\{A_{n_j}\}$ for which $A \doteq \liminf A_{n_j} \doteq \limsup A_{n_j}$. Furthermore, $\mathbb{1}_{A_n}$ converges to $\mathbb{1}_A$ in measure.*

Now, let $\mathbb{Q}$ be the measure of Lemma 11. In light of Lemmas 11 and 12 together, it remains to show that the measure $\mathbb{Q}$ is given by $\mathbb{Q}(E) = \mathbb{P}(A \cap E)$ for some set A. The proof of this statement is the focus of the proof of Lemma 10 presented below. Both Lemmas are proved in the next subsection, Appendix C.2.

We next discuss a real analysis theorem essential in the proof of Lemma 10 which characterizes absolute continuity.

We now define absolute continuity.

Definition 6. *A measure ν over a set X is* absolutely continuous *with respect to a measure μ over X if $\mu(A) = 0$ implies that $\nu(A) = 0$ for any measurable set $A \subseteq X$. This condition is denoted by $\nu \ll \mu$.*

The Radon-Nikodym theorem then gives the relationship between ν and μ if $\nu \ll \mu$:

Theorem 7 (Radon-Nikodym)**.** *Assume that ν, μ are σ-finite measures and $\nu \ll \mu$. Then there is a μ-integrable function f for which*

$$\nu(E) = \int_E f d\mu.$$

See for instance (Folland, 1999) for more details.

Lastly we present a rather technical lemma, which is very important for our proof.

Lemma 13. *Let $\mathbb{P}$ be a measure and let A_n be a sequence of measurable sets. Assume that there is a bounded function q for which*

$$\lim_{n\to\infty} \int_E (\mathbb{1}_{A_n} - q) d\mathbb{P} = 0 \tag{9}$$

for all measurable E. Then in fact

$$\lim_{n\to\infty} \int_E q(\mathbb{1}_{A_n} - q) d\mathbb{P} = 0 \tag{10}$$

for all measurable E as well.

Proof of Lemma 10. First, Lemma 11 implies that there is a Borel measure $\mathbb{Q}$ for which

$$\lim_{n\to\infty} \mathbb{P}_n(E) = \mathbb{Q}(E). \tag{11}$$

Next, we show that this equality implies that $\mathbb{Q} \ll \mathbb{P}$. First, note that for all E, $\mathbb{Q}(E) \geq 0$ since $\mathbb{Q}(E) = \lim_{n\to\infty} \mathbb{P}(A_n \cap E) \geq 0$. Furthermore, if $\mathbb{P}(E) = 0$, then

$$\mathbb{Q}(E) = \lim_{n\to\infty} \mathbb{P}(E \cap A_n) \leq \lim_{n\to\infty} \mathbb{P}(E) = 0.$$

Hence, we have $\mathbb{Q} \ll \mathbb{P}$. Thus, by the Radon-Nikodym theorem, we can write

$$\mathbb{Q}(E) = \int \mathbb{1}_E q(x) d\mathbb{P},$$

for some $\mathbb{P}$-integrable function q. Now, we aim to show that q is in fact an indicator function $\mathbb{P}$-a.e., because this would imply that $\mathbb{Q}(E) = \mathbb{P}(E \cap A)$ for some set A. Now, as

$$\mathbb{P}(A_n \cap E) = \int \mathbb{1}_{A_n \cap E} d\mathbb{P},$$

we can re-write the statement $\lim_{n\to\infty} \mathbb{P}(E \cap A_n) = \mathbb{Q}(E)$ as

$$\lim_{n\to\infty} \int \mathbb{1}_E (q - \mathbb{1}_{A_n}) d\mathbb{P} = 0. \tag{12}$$

We will use this equation to show that

$$\lim_{n\to\infty} \int \mathbb{1}_E (\mathbb{1}_{A_n} - q) d\mathbb{Q} = 0. \tag{13}$$

We will now discuss the motivation behind studying this equation. One can actually apply Fatou's lemma to argue that $\mathbb{1}_{\liminf A_n} \leq q \leq \mathbb{1}_{\limsup A_n}$ $\mathbb{P}$ a.e. The hope is to argue from (12) in fact $\limsup A_n = \liminf A_n$. However, this is surprisingly tricky. Equation 13 allows us to argue that q is an indicator function with out showing that $\liminf A_n \doteq \limsup A_n$ first.

We want to show that q is an indicator function. To start, we argue that q is in fact bounded. The inequality $\mathbb{Q}(E) \geq 0$ demonstrated above implies that $q \geq 0$. Next we show that $q \leq \mathbb{1}_{\limsup A_n}$.

Since indicator functions are bounded above by one, we can use the reverse Fatou lemma to conclude that for all sets E,

$$\begin{aligned} 0 &= \lim_{n\to\infty} \int \mathbb{1}_E(\mathbb{1}_{A_n} - q)d\mathbb{P} = \limsup_{n\to\infty} \int \mathbb{1}_E(\mathbb{1}_{A_n} - q)d\mathbb{P} \\ &\leq \int \mathbb{1}_E(\mathbb{1}_{\limsup A_n} - q)d\mathbb{P}. \qquad \text{(Reverse Fatou Lemma)} \end{aligned} \tag{14}$$

Now choose E as the set

$$E = \{x \colon \mathbb{1}_{\sup A_n}(x) - q(x) < 0\}\,.$$

As the integrand of (14) is strictly negative on E and yet the integral is at least zero, the set E satisfies $\mathbb{P}(E) = 0$. Therefore $\mathbb{1}_{\limsup A_n} \geq q(x)$ a.e.

Summarizing, we have shown

$$0 \leq q(x) \leq \mathbb{1}_{\limsup A_n}(x), \tag{15}$$

$\mathbb{P}$-a.e. In particular, q is always between 0 and 1 outside a set measure zero.

Now we apply Lemma 13 to conclude that

$$\lim_{n\to\infty} \int_E q(\mathbb{1}_{A_n} - q)d\mathbb{P} = 0,$$

for every measurable set E. In other words, by the Radon-Nikodym theorem, (13) holds for every measurable set E.

That q is actually an indicator function follows from the relation (13). We now discuss this argument.

Consider the set

$$\begin{aligned} E = \{x \colon \mathbb{1}_{\limsup A_n}(x) - q(x) > 0\} &= \{x \colon \mathbb{1}_{A_n}(x) - q(x) > 0 \text{ for infinitely many } n\} \\ &= \bigcup_{m\in\mathbb{N}} \left\{x \colon \mathbb{1}_{A_n}(x) - q(x) > \frac{1}{m} \text{ for infinitely many } n\right\}. \end{aligned}$$

We will show that $\mathbb{Q}(E) = 0$. For a fixed $\delta > 0$, define the set

$$E_\delta = \{x \colon \mathbb{1}_{A_n} - q > \delta \text{ for infinitely many n}\}.$$

We will show that $\mathbb{Q}(E_\delta) = 0$, and this will imply that $\mathbb{Q}(E) = 0$. Now note that the support of $\mathbb{Q}$ is the set $\overline{\{q = 0\}^C} = (\text{int}\{q = 0\})^C$. Therefore, the fact that $\mathbb{Q}(E) = 0$ implies that $E \dot{\subset} (\operatorname{supp}\mathbb{Q})^C \subset \{q = 0\}$, where $\dot{\subset}$ denotes containment up to a set measure zero. Hence $E^C \cup \{q = 0\} \doteq X$. In other words, $q = 0$ or $\limsup_{A_n} \leq q$ $\mathbb{P}$-a.e. As we have already shown the opposite inequality in (15), this will imply that $q = 0$ or $q = \mathbb{1}_{\limsup A_n}$ $\mathbb{P}$-a.e. Therefore, q is in fact an indicator function $\mathbb{P}$-a.e.

We now show that (13) implies $\mathbb{Q}(E_\delta) = 0$. For contradiction, assume that $\mathbb{Q}(E_\delta) = K > 0$. As $\mathbb{1}_{A_n} - q > \delta$ on E_δ, we can bound the integral in (13) below by

$$\int \mathbb{1}_{E_\delta}(\mathbb{1}_{A_n} - q)d\mathbb{Q} \geq \int \mathbb{1}_{E_\delta}\delta d\mathbb{Q} = \delta\mathbb{Q}(E_\delta) > \delta K > 0$$

and thus

$$\lim_{n\to\infty} \int \mathbb{1}_{E_\delta}(\mathbb{1}_{A_n} - q)d\mathbb{Q} \neq 0$$

which contradicts (13).

□

C.2 Proofs of Supporting Lemmas

In this Appendix we present the proofs of Lemmas 11, 12, and 13 which were used to prove Lemmas 1 and 10. We begin with the proof of Lemma 12 as the argument is most elementary and we end with the proof of Lemma 13 as the argument is the most technical.

Lemma 12. *Let $\{A_n\}$ be a sequence of sets and assume that $\mathbb{P}(A_n \cap E) \to \mathbb{P}(A \cap E)$ for some set A. Then, there is a subsequence of sets $\{A_{n_j}\}$ for which $A \doteq \liminf A_{n_j} \doteq \limsup A_{n_j}$. Furthermore, $\mathbb{1}_{A_n}$ converges to $\mathbb{1}_A$ in measure.*

Proof of Lemma 12. First, we show that $\mathbb{1}_{A_n}$ converges to $\mathbb{1}_A$ in measure. Subsequently, a basic fact of analysis implies that $\mathbb{1}_{A_n}$ contains a subsequence $\mathbb{1}_{A_{n_j}}$ that converges to $\mathbb{1}_A$ $\mathbb{P}$-a.e. Thus we will have that outside a set measure zero,

$$\mathbb{1}_A = \lim_{n\to\infty} \mathbb{1}_{A_n} = \liminf_{n\to\infty} \mathbb{1}_{A_n} = \mathbb{1}_{\liminf A_n}$$

and similarly

$$\mathbb{1}_A = \lim_{n\to\infty} \mathbb{1}_{A_n} = \limsup_{n\to\infty} \mathbb{1}_{A_n} = \mathbb{1}_{\limsup A_n}.$$

We now show that $\mathbb{1}_{A_n}$ converges to $\mathbb{1}_A$ in measure. First, if we pick $E = A^C$ and then $\mathbb{P}(A_n \cap E) \to \mathbb{P}(A \cap E)$ implies

$$\mathbb{P}(A_n \cap A^C) \to 0. \tag{16}$$

Next $\mathbb{P}(A_n \cap E) \to \mathbb{P}(A \cap E)$ also implies that

$$\lim_{n\to\infty} \mathbb{P}(E) - \mathbb{P}(A_n \cap E) = \lim_{n\to\infty} \mathbb{P}(E \cap (A_n \cap E)^C) = \lim_{n\to\infty} \mathbb{P}(E \cap A_n^C) = \mathbb{P}(E) - \mathbb{P}(A \cap E) = \mathbb{P}(A^C \cap E)$$

and thus

$$\lim_{n\to\infty} \mathbb{P}(E \cap A_n^C) = \mathbb{P}(A^C \cap E).$$

Evaluating at $E = A$ gives

$$\lim_{n\to\infty} \mathbb{P}(A \cap A_n^C) = \mathbb{P}(A^C \cap A) = 0 \tag{17}$$

Together, (16) and (17) imply that

$$\lim_{n\to\infty} \mathbb{P}(A_n \triangle A) = \lim_{n\to\infty} \mathbb{P}(|\mathbb{1}_{A_n} - \mathbb{1}_A| > 0) = 0$$

which means that $\mathbb{1}_{A_n}$ converges to $\mathbb{1}_A$ in measure. □

Next, we discuss the proof of Lemma 11. To start, we review three real analysis theorems essential for the proof. All can be found in (Folland, 1999).The first is Urysohn's lemma, see Chapter 4.2 of (Folland, 1999) for a discussion. (Our statement of this theorem is Folland's form of Urysohn's lemma together with the Urysohn metrization theorem, which can also be found in Folland (1999).) The second is about a concept called *inner regularity*, which we will further use in Appendix G as well. Lastly, we also discuss a concept called *outer regularity* which follows from inner regularity.

Theorem 8 (Urysohn's Lemma)**.** *Let X be a metrizable space. If A, F are disjoint closed sets in X, then there is a continuous function $f\colon X \to [0,1]$ for which $f = 0$ on A and $f = 1$ on F.*

We next discuss a theorem which implies that many common measures are inner regular. This statement is a generalization of Lemma 4 (compare the two). Again this result is a consequence of Theorem 7.8 and Proposition 7.5 of (Folland, 1999).

We begin with defining inner regularity for an arbitrary metric space.

Definition 7. *Let τ be a topology on a set X and Σ a σ-algebra on X. $\mathbb{P}$ be a measure on (X, Σ).Then $\mathbb{P}$ is* inner regular *if for all measurable sets E, $\mathbb{P}(E) = \sup\{\mathbb{P}(K)\colon K \subset E, K$ compact and measurable$\}$.*

We now present a theorem which implies that many common measure spaces are inner regular.

Theorem 9. *Let X be a second-countable and locally compact Haudsorff space. Then every Borel measure ν with $\nu(X) < \infty$ is inner regular.*

Lastly, we present a consequence of inner regularity which we will also use later in the appendix in the proof of Theorem 10.

Lemma 14. *If a measure μ is inner regular, then for every measurable set E,*

$$\mu(E) = \inf_{\substack{U\colon E \subset U \\ U \text{ open}}} \mu(U). \tag{18}$$

This concept is discussed in Folland (1999), where (18) is referred to as *outer regularity*.

The weak convergence of $\{\mathbb{P}_n\}$ to $\mathbb{P}$ implies that $\int f\mathbb{P}_n \to \int f d\mathbb{P}$ for continuous functions f. We use Urysohn's Lemma together with Lemma 14 to show that this relation also holds when f is an indicator function.

Lemma 11. *Let X be a separable metric space endowed with the universal σ-algebra. Assume that $\mathbb{P}_n$ weakly converges to $\mathbb{Q}$ with $\mathbb{P}_n$ given by $\mathbb{P}_n(E) = \mathbb{P}(E \cap A_n)$ for a sequence of sets A_n. Then, in fact for every measurable set E,*

$$\lim_{n\to\infty} \mathbb{P}_n(E) = \mathbb{Q}(E).$$

Proof of Lemma 11. We show that in fact for each universally measurable set E, we have in fact $\lim_{n\to\infty} \mathbb{P}_n(E) = \mathbb{Q}(E)$. This fact is strictly stronger than weak convergence.

To start, note that Theorem 9 implies that the completion of a Borel measure restricted to $\mathscr{U}(X)$ is inner regular as well.

Fix a set $E \in \mathscr{U}(X)$. Since $\mathbb{P}, \mathbb{Q}$ are both inner regular, there exist compact K_1, K_2 and open U_1, U_2 for which $K_1, K_2 \subset E \subset U_1, U_2$ satisfying $\mathbb{P}(U_1 - K_1) < \delta$ and $\mathbb{Q}(U_2 - K_2) < \delta$. (This statement involves an application of Lemma 14.) If we set $K = K_1 \cup K_2$ and $U = U_1 \cap U_2$ then $K \subset E \subset U$ and $\mathbb{P}(U - K) < \delta, \mathbb{Q}(U - K) < \delta$. Now K and U^C are disjoint closed sets, so by Urysohns's lemma, we can pick a function $f\colon X \to [0,1]$ that is 1 on K and zero outside U. Therefore, $|f - \mathbb{1}_E|$ is between 0 and 1 on $U \setminus K$ and zero everywhere else. Thus

$$\left|\int f - \mathbb{1}_E d\mathbb{Q}\right| \le \int |f - \mathbb{1}_E| d\mathbb{Q} \le \int_{U\setminus K} 1 d\mathbb{Q} < \delta.$$

Similarly,

$$\left|\int f - \mathbb{1}_E d\mathbb{P}_n\right| \le \int |f - \mathbb{1}_E| d\mathbb{P}_n \le \int_{U\setminus K} 1 d\mathbb{P}_n = \int_{U\setminus K} \mathbb{1}_{A_n} d\mathbb{P} \le \int_{U\setminus K} 1 d\mathbb{P} = \mathbb{P}(U \setminus K) < \delta.$$

Since f is continuous and bounded, weak convergence implies that that $\lim_{n\to\infty} \int f d\mathbb{P}_n = \int f d\mathbb{Q}$. Pick an n large enough so that

$$\left|\int f d\mathbb{P}_n - \int f d\mathbb{Q}\right| < \delta.$$

Then

$$|\mathbb{Q}(E) - \mathbb{P}_n(E)| \le \left|\int \mathbb{1}_E - f d\mathbb{Q}\right| + \left|\int f d\mathbb{Q} - \int f d\mathbb{P}_n\right| + \left|\int f - \mathbb{1}_E d\mathbb{P}_n\right| \le 3\delta.$$

Thus we have shown that in fact

$$\lim_{n\to\infty} \mathbb{P}_n(E) = \mathbb{Q}(E).$$

□

We now prove the following lemma which is instrumental in showing that in fact $\mathbb{Q}(E) = \mathbb{P}(A \cap E)$.

Lemma 13. *Let $(X, \Sigma, \mathbb{P})$ be a measure space and let A_n be a sequence of measurable sets. Assume that there is a bounded function q for which*

$$\lim_{n\to\infty} \int_E (\mathbb{1}_{A_n} - q) d\mathbb{P} = 0 \tag{9}$$

for all measurable E. Then in fact

$$\lim_{n\to\infty} \int_E q(\mathbb{1}_{A_n} - q) d\mathbb{P} = 0 \tag{10}$$

for all measurable E as well.

Proof of Lemma 13. Let $\mathcal{F}$ be the set of simple functions. Formally,

$$\mathcal{F} = \left\{ \sum_{i=1}^n k_i \mathbb{1}_{E_i} \colon E_i \in \Sigma \text{ measurable }, n \in \mathbb{N}, k_i \in \mathbb{R} \right\}.$$

Recall that $\mathcal{F}$ is dense in $L^1(\mathbb{P})$. Furthermore, (9) actually implies that for every $f \in \mathcal{F}$,

$$\lim_{n\to\infty} \int f(\mathbb{1}_{A_n} - q)d\mathbb{P} = 0. \tag{19}$$

Now the relation

$$f\mathbb{1}_{A_n} - fq = f\mathbb{1}_{A_n} - q\mathbb{1}_{A_n} + q\mathbb{1}_{A_n} - q^2 + q^2 - fq = (f-q)\mathbb{1}_{A_n} + q(\mathbb{1}_{A_n} - q) + q(q-f)$$

implies that

$$q(\mathbb{1}_{A_n} - q) = f(\mathbb{1}_{A_n} - q) - (f-q)\mathbb{1}_{A_n} - q(q-f). \tag{20}$$

Pick $\delta > 0$. By density, we can pick an $f^* \in \mathcal{F}$ for which $\int |f^* - q| d\mathbb{P} < \delta$, and (19) holds for this particular choice f^*. We integrate (20) over E to obtain

$$\int_E q(\mathbb{1}_{A_n} - q)d\mathbb{P} = \int_E f^*(\mathbb{1}_{A_n} - q)d\mathbb{P} - \int_E (f^* - q)\mathbb{1}_{A_n} d\mathbb{P} - \int_E q(q - f^*)d\mathbb{P}$$

and thus

$$\left|\int_E q(\mathbb{1}_{A_n} - q)d\mathbb{P}\right| \leq \left|\int_E f^*(\mathbb{1}_{A_n} - q)d\mathbb{P}\right| + \int_E |(f^* - q)|\,|\mathbb{1}_{A_n}|\,d\mathbb{P} + \int_E |q||(q - f^*)|\,d\mathbb{P}. \tag{21}$$

We assumed that the function q was bounded, so $|q| \leq K$ for some constant K. As $\mathbb{1}_{A_n}$ is between 0 and 1, we can bound the last two of the integrals of (21) as

$$\int_E |(f^* - q)||\mathbb{1}_{A_n}|\,d\mathbb{P} \leq \int |f^* - q| d\mathbb{P} \leq \delta, \int_E |q||(q - f^*)|d\mathbb{P} \leq \int K|f^* - q|d\mathbb{P} \leq K\delta.$$

Applying these inequalities, the bound on (21) becomes

$$\left|\int_E q(\mathbb{1}_{A_n} - q)d\mathbb{P}\right| \leq \left|\int_E f^*(\mathbb{1}_{A_n} - q)d\mathbb{P}\right| + (K+1)\delta.$$

Now we would like to take the limit $n \to \infty$ of both sides of this inequality. However, we don't yet know that the limit of the left hand side exists, so instead we evaluate the $\limsup$:

$$\begin{aligned}\limsup_{n\to\infty} \left|\int_E q(\mathbb{1}_{A_n} - q)d\mathbb{P}\right| &\leq \limsup_{n\to 0} \left|\int_E f^*(\mathbb{1}_{A_n} - q)d\mathbb{P}\right| + (K+1)\delta \\ &= \lim_{n\to\infty} \left|\int_E f^*(\mathbb{1}_{A_n} - q)d\mathbb{P}\right| + (K+1)\delta = (K+1)\delta.\end{aligned}$$

As $\delta > 0$ was arbitrary,

$$\limsup_{n\to\infty} \left|\int_E q(\mathbb{1}_{A_n} - q)d\mathbb{P}\right| = 0$$

and therefore (10) holds. □

D 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 (7) as our definition for A^ϵ:

$$A^\epsilon = \bigcup_{\mathbf{a}\in A} \overline{B_\epsilon(\mathbf{a})}. \tag{7}$$

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

$$A^{-\epsilon} = ((A^C)^\epsilon)^C. \tag{22}$$

Throughout this appendix we will work in $\mathbb{R}^d$, however, most of our proofs work for an arbitrary metric space where A^ϵ, $A^{-\epsilon}$ are defined as (7) and (22). The obstacle to generalizing these results to an arbitrary metric space is Lemma 21, which is false for general metric spaces. Proofs of the lemmas which do not depend on Lemma 21 were written so that they immediately generalize to arbitrary metric spaces.

D.1 Basic Properties and Proof of Equation 5

The following lemma details how the $^{\epsilon}$ and $^{-\epsilon}$ operations interact with unions and intersections. This lemma includes a proof of equation (5).

Lemma 15. *Define A^{ϵ} as in (7) and $A^{-\epsilon}$ as (22). Then for any sequence of sets $\{A_i\}$, the following set containments hold:*

$$\bigcup_{i=1}^{\infty} A_i^{\epsilon} = \left[\bigcup_{i=1}^{\infty} A_i\right]^{\epsilon} \quad (23) \qquad \bigcap_{i=1}^{\infty} A_i^{-\epsilon} = \left[\bigcap_{i=1}^{\infty} A_i\right]^{-\epsilon} \quad (5)$$

$$\bigcap_{i=1}^{\infty} A_i^{\epsilon} \supset \left[\bigcap_{i=1}^{\infty} A_i\right]^{\epsilon} \quad (24) \qquad \bigcup_{i=1}^{\infty} A_i^{-\epsilon} \subset \left[\bigcup_{i=1}^{\infty} A_i\right]^{-\epsilon} \quad (25)$$

Proof. **Showing (23):**
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 (24):
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

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

Showing (5):
Recall that $A^{-\epsilon} = ((A^C)^{\epsilon})^C$. If we apply (23) 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 (25): If we apply (24) 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 (25). □

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 16. *For a set A, define*

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

Then

$$F(A^\epsilon) = \emptyset \tag{26}$$

$$F((A^{-\epsilon})^C) = \emptyset \tag{27}$$

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

Proof of Lemma 16. Equation 7

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

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 (26). Lastly, (27) follows by applying (26) to A^C. □

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

Lemma 17. *Define $A^\epsilon, A^{-\epsilon}$ as in (7),(22). Then alternative characterizations of $A^\epsilon, A^{-\epsilon}$ are given by*

$$A^\epsilon = \{\mathbf{x} \in X : \overline{B_\epsilon(\mathbf{x})} \cap A \neq \emptyset\} \tag{28}$$

$$A^{-\epsilon} = \{\mathbf{a} : \overline{B_\epsilon(\mathbf{a})} \subset A\} \tag{29}$$

Notice that in $\mathbb{R}^d$ (29) 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 17. **Showing (28):**
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 (29):
Recall the definition $A^{-\epsilon} = ((A^C)^\epsilon)^C$. Then

$$\begin{aligned}
&\mathbf{a} \in A^{-\epsilon} \\
&\Leftrightarrow \mathbf{a} \notin (A^C)^\epsilon \\
&\Leftrightarrow \overline{B_\epsilon(\mathbf{a})} \text{ does not intersect } A^C \text{ (by (28))} \\
&\Leftrightarrow \overline{B_\epsilon(\mathbf{a})} \subset A
\end{aligned}$$

□

We now prove two basic results which follow from the characterizations of the $^\epsilon,^{-\epsilon}$ operations given by equations 28 and 29. The first lemma discusses how $^\epsilon,^{-\epsilon}$ affect disjoint sets.

Lemma 18. *If A, B are disjoint, then so are A^ϵ and $B^{-\epsilon}$*

Proof. We use the descriptions of the $^\epsilon,^{-\epsilon}$ operations given by Lemma 17. If $\mathbf{x} \in B^{-\epsilon}$, then $\overline{B_\epsilon(\mathbf{x})} \subset B$. Since A and B are disjoint, $\overline{B_\epsilon(\mathbf{x})}$ does not intersect A. Therefore, $\mathbf{x} \notin A^\epsilon$.

□

The next lemma describes how $^\epsilon,^{-\epsilon}$ affect intersecting sets.

Lemma 19. *Let $A, \overline{B_r(\mathbf{x})} \subset \mathbb{R}^d$. If $A, \overline{B_r(\mathbf{x})}$ intersect and $\epsilon \leq r$, then so do A^ϵ and $\overline{B_{r-\epsilon}(\mathbf{x})}$.*

Proof. Let $\mathbf{y} \in A \cap \overline{B_r(\mathbf{x})}$. We consider two cases, $\|\mathbf{y}-\mathbf{x}\| \leq \epsilon$, and $\|\mathbf{y}-\mathbf{x}\| > \epsilon$. If $\|\mathbf{y}-\mathbf{x}\| \leq \epsilon$, then $\mathbf{x} \in \overline{B_\epsilon(\mathbf{y})}$ and thus $\mathbf{x} \in A^\epsilon$. Hence $\overline{B_{r-\epsilon}(\mathbf{x})}$ and A^ϵ intersect. Now consider the case $\|\mathbf{y}-\mathbf{x}\| > \epsilon$ and set

$$\mathbf{z} = \left(1 - \frac{\epsilon}{\|\mathbf{x}-\mathbf{y}\|}\right)\mathbf{y} + \frac{\epsilon}{\|\mathbf{x}-\mathbf{y}\|}\mathbf{x}.$$

Then $\mathbf{z} \in \overline{B_\epsilon(\mathbf{y})}$. We now bound the distance from $\mathbf{z}$ to $\mathbf{x}$. We have

$$\|\mathbf{z}-\mathbf{x}\| = \left(1 - \frac{\epsilon}{\|\mathbf{y}-\mathbf{x}\|}\right)\|\mathbf{y}-\mathbf{x}\| = \|\mathbf{y}-\mathbf{x}\| - \epsilon \leq r - \epsilon$$

and therefore $\mathbf{z} \in \overline{B_{r-\epsilon}(\mathbf{x})}$. Thus $\mathbf{z} \in A^\epsilon \cap \overline{B_{r-\epsilon}(\mathbf{x})}$. □

Lastly, we study how the $^\epsilon$ operation interacts with set closures in $\mathbb{R}^d$.

Lemma 20. *Let $\overline{S}$ denote the closure of the set S and let $A \subset \mathbb{R}^d$. Then $\left(\overline{A}\right)^\epsilon = \overline{A^\epsilon}$.*

Proof. For convenience of notation, we will denote $\left(\overline{A}\right)^\epsilon$ as $\overline{A}^\epsilon$.

Showing $\overline{A}^\epsilon \subset \overline{A^\epsilon}$:
Let $\mathbf{x} \in \overline{A}^\epsilon$. Then $\mathbf{x} = \mathbf{a} + \mathbf{h}$ where $\mathbf{h} \in \overline{B_\epsilon(\mathbf{0})}$ and $\mathbf{a} \in \overline{A}$. Then there exists a sequence $\{\mathbf{a}_i\} \subset A$ for which $\mathbf{a}_i \to \mathbf{a}$. Then $\mathbf{a}_i + \mathbf{h}$ is a sequence in A^ϵ that converges to $\mathbf{x}$, so $\mathbf{x} \in \overline{A^\epsilon}$. Therefore, $\overline{A}^\epsilon \subset \overline{A^\epsilon}$.

Showing $\overline{A}^\epsilon \supset \overline{A^\epsilon}$:
Pick $\mathbf{x} \in \overline{A^\epsilon}$. Then there exists a sequence $\{\mathbf{a}_i\} \subset A$ and $\{\mathbf{h}_i\} \subset \overline{B_\epsilon(\mathbf{0})}$ for which $\mathbf{a}_i + \mathbf{h}_i \to \mathbf{x}$. By compactness, we can choose a convergent subsequence $\{\mathbf{h}_{i_j}\}$ of $\{\mathbf{h}_i\}$. Now we can define $\mathbf{a}$ as $\mathbf{a} = \lim_{j\to\infty} \mathbf{x} - \mathbf{h}_{i_j} = \mathbf{x} - \mathbf{h}$. Since $\mathbf{x} - \mathbf{h}_{i_j} = \mathbf{a}_{i_j} \in A$, $\mathbf{a}$ is a limit point of A. Therefore $\mathbf{x} = \mathbf{a} + \mathbf{h}$ with $\mathbf{a} \in \overline{A}$ and $\|\mathbf{h}\| \leq \epsilon$, so $\overline{A}^\epsilon \supset \overline{A^\epsilon}$.

□

D.2 Applying the $^\epsilon$,$^{-\epsilon}$ Operations in Succession

In some of our proofs, we apply the $^\epsilon$ and $^{-\epsilon}$ results to sets multiple times in succession. In this section, we describe the outcome of these operations and 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 Lemma 27, which is the most important conclusion of this subsection.

Our first result describes the result of applying the $^\epsilon$ operation twice, or the $^{-\epsilon}$ operation twice.

Lemma 21. *Let A be a set in $\mathbb{R}^d$ and define $A^\epsilon = A \oplus \overline{B_\epsilon(\mathbf{0})}$. Then, the following holds:*

$$(A^{\epsilon_1})^{\epsilon_2} = A^{\epsilon_1+\epsilon_2} \qquad \left(A^{-\epsilon_1}\right)^{-\epsilon_2} = A^{-\epsilon_1-\epsilon_2}. \tag{30}$$

The conclusion of this lemma is fairly intuitive. However, note that it does *not* hold if we mix the signs of the ϵ_i's: $(A^{\epsilon_1})^{-\epsilon_2} \neq A^{\epsilon_1-\epsilon_2}$ and $(A^{-\epsilon_2})^{\epsilon_1} = A^{\epsilon_1-\epsilon_2}$. We discuss applying $^\epsilon$ followed by $^{-\epsilon}$ and vice versa in subsequent lemmas.

Furthermore, the conclusion of this lemma is *false* for general metric spaces. For example, consider the space $\mathbb{Z}$ with the metric $d(z_1, z_2) = |z_1 - z_2|$. Then if we take the set $A = \{1\}$, then $A^{\frac{1}{2}} = \{1\} = A$, and therefore $(A^{\frac{1}{2}})^{\frac{1}{2}} = A$. However, $A^{\frac{1}{2}+\frac{1}{2}} = A^1 = \{1, 2, 3\}$.

Proof of Lemma 21. First, the second equality of (30) follows from applying the first equation of (30) to A^C and then taking complements. We now focus on showing the first equality of (30). To start,

$$(A^{\epsilon_1})^{\epsilon_2} = \left(\left(A \oplus \overline{B_{\epsilon_1}(\mathbf{0})}\right) \oplus \overline{B_{\epsilon_2}(\mathbf{0})}\right).$$

The associativity of addition implies the associativity of set addition $\oplus$. Therefore,

$$(A^{\epsilon_1})^{\epsilon_2} = \left(A \oplus \left(\overline{B_{\epsilon_1}(\mathbf{0})} \oplus \overline{B_{\epsilon_2}(\mathbf{0})}\right)\right) = A \oplus \overline{B_{\epsilon_1+\epsilon_2}(\mathbf{0})} = A^{\epsilon_1+\epsilon_2}.$$

□

Our next lemma 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 22. *Define the* ${}^{\epsilon}, {}^{-\epsilon}$ *operations as in* (7), (22). *Then*

$$(A^{\epsilon})^{-\epsilon} \supset A \tag{31}$$

$$(A^{-\epsilon})^{\epsilon} \subset A \tag{32}$$

Proof. To start, note that (32) follows from applying (31) to A^C and then taking complements. We focus on proving (31) in the remainder of the proof. In order to show (31), we make use of Equation 29. Equation 29 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 23. *Define* $A^\epsilon, A^{-\epsilon}$ *as in* (7),(22). *Then the following hold:*

$$A = (A^{-\epsilon})^{\epsilon} \sqcup F(A) \tag{33}$$

$$(A^{\epsilon})^{-\epsilon} = A \sqcup F(A^C). \tag{34}$$

Specifically, (33) implies that $(A^{-\epsilon})^{\epsilon} = A - F(A)$ and (34) implies that $(A^{\epsilon})^{-\epsilon} = A \cup F(A^C)$. Figure 2 illustrates the sets $F(A)$ and $F(A^C)$.

Proof of Lemma 23.
Showing $\supset$ for (33):
It's clear that $F(A) \subset A$ and Lemma 22 implies that $(A^{-\epsilon})^{\epsilon} \subset A$ as well.
Showing $\subset$ for (33):
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 (29), $\mathbf{a} \in A^{-\epsilon}$. Furthermore, $\mathbf{x} \in \overline{B_\epsilon(\mathbf{a})}$ implies that $\mathbf{x} \in (A^{-\epsilon})^{\epsilon}$.

Showing disjoint union for (33):

Lemma 22 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 (34):

Applying (33) 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

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

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 (29) 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 $B_\epsilon(\mathbf{y})$ intersects A and then (28) 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 24. *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 24. By Lemmas 16 and 23,

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

Similarly,

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

□

In fact, we can actually prove a generalization of Lemma 24:

Lemma 25. *Assume we are in a space for which Lemma 21 holds and let $\delta < \epsilon$. Then*

$$((A^{-\epsilon})^\epsilon)^{-\delta} = (A^{-\epsilon})^{\epsilon-\delta} \tag{35}$$

and

$$((A^\epsilon)^{-\epsilon})^\delta = (A^\epsilon)^{-(\epsilon-\delta)}. \tag{36}$$

Proof. We start by showing that (36) implies (35). If we apply (36) to A^C, then we get that $(((A^C)^\epsilon)^{-\epsilon})^\delta = ((A^C)^\epsilon)^{-(\epsilon-\delta)} = ((A^{-\epsilon})^{\epsilon-\delta})^C$. Upon taking complements, this equation becomes $((A^{-\epsilon})^\epsilon)^{-\delta} = (A^{-\epsilon})^{\epsilon-\delta}$

Next we prove Equation 36. By Lemma 21,

$$((A^\epsilon)^{-\epsilon})^\delta = (((A^\epsilon)^{-(\epsilon-\delta)})^{-\delta})^\delta$$

Therefore, by Lemma 24, $(A^\epsilon)^{-(\epsilon-\delta)} \subset ((A^\epsilon)^{-\epsilon})^\delta$. We will now show the opposite inclusion. Consider a $\mathbf{x}$ in $((A^{-\epsilon})^\epsilon)^\delta$. Then $\overline{B_\delta(\mathbf{x})}$ intersects $(A^\epsilon)^{-\epsilon}$. Thus there is a point $\mathbf{y} \in \overline{B_\delta(\mathbf{x})}$ for which $\overline{B_\epsilon(\mathbf{y})} \subset A^\epsilon$. Then by the triangle inequality, if $\mathbf{z} \in \overline{B_{\epsilon-\delta}(\mathbf{x})}$, then

$$d(\mathbf{z},\mathbf{y}) \leq d(\mathbf{z},\mathbf{x}) + d(\mathbf{x},\mathbf{y}) \leq \epsilon - \delta + \delta = \epsilon$$

Therefore, $\mathbf{z} \in \overline{B_\epsilon(\mathbf{y})}$ and hence $\overline{B_{\epsilon-\delta}(\mathbf{x})} \subset A^\epsilon$. By Lemma 17, $\overline{B_{\epsilon-\delta}(\mathbf{x})} \subset A^\epsilon$ implies that $\mathbf{x} \in (A^\epsilon)^{-(\epsilon-\delta)}$.

□

We next prove a short lemma that will help us reason about the adversarial loss.

Lemma 26. *Let $^\epsilon, ^{-\epsilon}$ be as in (7) and (22). 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 24, $D^{-\epsilon} = B^{-\epsilon}$. Furthermore, according to Lemma 22, $D \subset B$, so that $D^\epsilon \subset B^\epsilon$.

Next, according to Lemma 24, $C^\epsilon = D^\epsilon \subset B^\epsilon$. Furthermore, according to Lemma 22, $C \supset B$, 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 27. *For any set A, the following hold:*

$$R^\epsilon(A) \geq R((A^\epsilon)^{-\epsilon}) \tag{37}$$

$$R^\epsilon(A) \geq R((A^{-\epsilon})^\epsilon). \tag{38}$$

The proof of this lemma actually shows that unless $F(A) \subseteq \{\mathbf{x}\colon \eta(\mathbf{x}) = 0\}$ we expect the inequality of (38) to be a strict inequality. Furthermore, we also expect the set $A \cap \{\mathbf{x}\colon \eta(\mathbf{x}) = 0\}$ to have small measure because because if $\eta(\mathbf{x}) = 0$ and $\mathbf{x} \in A$ the adversarial loss necessarily incurs a penalty of 1 at $\mathbf{x}$. An analogous statement holds for A^C and the set $\{\mathbf{x}\colon \eta(\mathbf{x}) = 1\}$. Furthermore, this lemma implies that $R^\epsilon(((A^{-\epsilon})^{2\epsilon})^{-\epsilon}) \leq R^\epsilon(A)$ and Lemma 23 suggests that the set $((A^{-\epsilon})^{2\epsilon})^{-\epsilon}$ is pseudo-certifiably robust because we have removed $F(A)$ from A and then removed $F(((A^{-\epsilon})^\epsilon)^C)$ from A^C. We actually prove this formally in the next appendix. Lastly, notice that since the sets $F(A), F(A^C)$ cannot contain an ϵ-ball, every point in these sets is close to the boundary so removing $F(A)$ from A and $F(A^C)$ from A^C is a 'local' change.

Proof of Lemma 27. 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 23 and Figure 2.) The same holds for $F(A^C)$ and A^C.

Let $B = (A^{-\epsilon})^\epsilon$ or $B = (A^\epsilon)^{-\epsilon}$. Lemma 26 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

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

□

D.3 Pseudo-Certifiably Robust Sets

In this section, we prove that certain sets are pseudo-certifiably robust. Our utilize the the notion of the distance of a point to a set, which we define below.

Definition 8. *The distance of a point to a set is defined as*

$$\text{dist}(\mathbf{x}, A) = \inf_{a \in A} d(\mathbf{x}, \mathbf{a})$$

The next lemma describes how the ${}^\epsilon$ operation relates to the distance between a point and a set.

Lemma 28. *The distance between a point $\mathbf{x}$ and a set A can be expressed as*

$$\text{dist}(\mathbf{x}, A) = \sup\{\delta \geq 0\colon \overline{B_\delta(\mathbf{x})} \textit{ is disjoint from } A\} \tag{39}$$

$$= \sup\{\delta \geq 0\colon \mathbf{x} \textit{ is disjoint from } A^\delta\} \tag{40}$$

$$\tag{41}$$

Furthermore, if $\inf \emptyset$ *is interpreted as* ∞,

$$\text{dist}(\mathbf{x}, A) = \inf\{\delta\colon \overline{B_\delta(\mathbf{x})} \textit{ intersects } A\} \tag{42}$$

$$= \inf\{\delta\colon x \in A^\delta\} \tag{43}$$

Proof. To start, note that Equation 40 follows from (39) combined with Lemma 18 and similarly Equation 43 follows form (42) combined with Lemma 17.

Next we show (39). Clearly, if $\overline{B_\delta(\mathbf{x})}$ is disjoint from A, then every point of A must be distance greater than δ from $\mathbf{x}$. Hence we have the inequality

$$\text{dist}(\mathbf{x}, A) \geq \sup\{\delta > 0\colon \overline{B_\delta(\mathbf{x})} \text{ is disjoint from } A\}.$$

For the other inequality, consider c with

$$c > \sup\{\delta > 0\colon \overline{B_\delta(\mathbf{x})} \text{ is disjoint from } A\} \tag{44}$$

We will show that $c \geq \operatorname{dist}(\mathbf{x}, A)$, and this will imply that

$$\operatorname{dist}(\mathbf{x}, A) \leq \sup\{\delta > 0\colon \overline{B_\delta(\mathbf{x})} \text{ is disjoint from } A\}. \tag{45}$$

Consider c satisfying (44). Then $\overline{B_c(\mathbf{x})}$ intersects A, and thus there is a point $\mathbf{y} \in A$ with $\|\mathbf{x}-\mathbf{y}\| \leq c$. Hence $c \geq \operatorname{dist}(\mathbf{x}, A)$. Since c was arbitrary, this relation implies (45).

Now we show Equation 42. If $\operatorname{dist}(\mathbf{x}, A) = \infty$, then the equality holds because we defined $\inf \emptyset$ as ∞. Now assume that $\operatorname{dist}(\mathbf{x}, A) < \infty$. Consider a δ for which $\overline{B_\delta(\mathbf{x})}$ intersects A. Then

$$\delta \geq \sup\{c \geq 0\colon \overline{B_c(\mathbf{x})} \text{ is disjoint from } A\} = \operatorname{dist}(\mathbf{x}, A)$$

Therefore, if we take the infimum over the left-hand side, we get that

$$\inf\{\delta\colon \overline{B_\delta(\mathbf{x})} \text{ intersects } A\} \geq \operatorname{dist}(\mathbf{x}, A)$$

We will now show the other inequality. Pick a c strictly larger than $\operatorname{dist}(\mathbf{x}, A)$. Then there is a point $\mathbf{y}$ in A with $d(\mathbf{x}, \mathbf{y}) < c$, so $\overline{B_c(\mathbf{x})}$ intersects A. Therefore,

$$c \geq \inf\{\delta\colon \overline{B_\delta(\mathbf{x})} \text{ intersects } A\}$$

This inequality implies that in fact

$$\operatorname{dist}(\mathbf{x}, A) \geq \{\delta\colon \overline{B_\delta(\mathbf{x})} \text{ intersects } A\}$$

□

We next show how the distance from a point to a set A relates to ∂A.

Corollary 3. *Assume that* $\operatorname{dist}(\mathbf{x}, A) = \delta$. *Then* $\overline{B_\delta(\mathbf{x})}$ *intersects* ∂A.

Proof. If $\operatorname{dist}(\mathbf{x}, A) = \delta$, then (39) of Lemma 28 implies that for each n, $\overline{B_{\delta+\frac{1}{n}}(\mathbf{x})}$ intersects A. Thus define a sequence $\mathbf{a}_n$ by choosing $\mathbf{a}_n \in \overline{B_{\delta+\frac{1}{n}}(\mathbf{x})} \cap A$. Because the sequence $\mathbf{a}_n$ is contained in the compact set $\overline{B_{\delta+1}(\mathbf{x})}$, there must be a convergent subsequence. Call the limit point $\mathbf{a}$. Clearly, $\mathbf{a} \in \overline{A}$. Furthermore, because $\mathbf{a}$ is a limit point of $\{\mathbf{a}_n\}_{n=m}^\infty$ for all m, we have $\mathbf{a} \in \overline{B_{\delta+\frac{1}{m}}(\mathbf{x})}$ for all m and therefore $\mathbf{a} \in \overline{B_\delta(\mathbf{x})}$. Now by (39), $\overline{B_{\delta-1/n}(\mathbf{x})}$ is disjoint from A for all $\mathbf{x}$ so there is also a sequence $\mathbf{c}_i$ in A^C which approaches $\mathbf{a}$, for instance one can choose $\mathbf{c}_i = \mathbf{x} + (\delta - 1/n)(\mathbf{a} - \mathbf{x})$. Therefore $\mathbf{a} \in \partial A$. □

We next prove a key lemma which shows that the interplay between distances and the $^\epsilon$, $^{-\epsilon}$ operations is well-behaved.

Lemma 29. *Assume we are in a space for which Lemma 21 holds. Then for every* $\mathbf{x} \in (A^{-2\epsilon})^{2\epsilon} - A^{-2\epsilon}$,

$$\operatorname{dist}(\mathbf{x}, A^{-2\epsilon}) + \operatorname{dist}(\mathbf{x}, ((A^{-2\epsilon})^{2\epsilon})^C) = 2\epsilon \tag{46}$$

Proof. If $A^{-2\epsilon}$ is either empty or the whole space, then $(A^{-2\epsilon})^{2\epsilon} - A^{-2\epsilon}$ is empty. Hence we will assume that neither $A^{-2\epsilon}$ nor $(A^{-2\epsilon})^C$ is empty.

Furthermore, note that $A^{-2\epsilon} \neq \mathbb{R}^d, \emptyset$ and $((A^{-2\epsilon})^{2\epsilon})^{-2\epsilon} = A^{-2\epsilon}$ implies that $(A^{-2\epsilon})^{2\epsilon} \neq \emptyset, X$.

Now pick $\mathbf{x} \in (A^{-2\epsilon})^{2\epsilon} - A^{-2\epsilon}$. We first argue that $\operatorname{dist}(\mathbf{x}, ((A^{-2\epsilon})^{2\epsilon})^C) \leq 2\epsilon$. Because $\mathbf{x} \in (A^{-2\epsilon})^C = (((A^{-2\epsilon})^{2\epsilon})^C)^{2\epsilon}$, Lemma 28 implies that $\operatorname{dist}(\mathbf{x}, ((A^{-2\epsilon})^{2\epsilon})^C) \leq 2\epsilon$.

Next, note that by Lemma 28,

$$\operatorname{dist}(\mathbf{x}, A^{-2\epsilon}) = \inf\{\delta : \mathbf{x} \in (A^{-2\epsilon})^\delta\}. \tag{47}$$

We will now write $\operatorname{dist}(\mathbf{x}, ((A^{-2\epsilon})^{2\epsilon})^C)$ in terms of this quantity.

$$\begin{aligned}
&\operatorname{dist}(\mathbf{x}, ((A^{-2\epsilon})^{2\epsilon})^C) \\
&= \sup\{\delta : \mathbf{x} \text{ is disjoint from } (((A^{-2\epsilon})^{2\epsilon})^C)^\delta\} && \text{(Lemma 28)} \\
&= \sup\{\delta : \mathbf{x} \in ((((A^{-2\epsilon})^{2\epsilon})^C)^\delta)^C\} = \sup\{\delta : \mathbf{x} \in ((A^{-2\epsilon})^{2\epsilon})^{-\delta}\} \\
&= \sup\{\delta : \mathbf{x} \in (A^{-2\epsilon})^{2\epsilon-\delta}\} && (\operatorname{dist}(\mathbf{x}, ((A^{-2\epsilon})^{2\epsilon})^C) \leq 2\epsilon, \text{Lemma 25}) \\
&= \sup\{2\epsilon - (2\epsilon - \delta) : \mathbf{x} \in (A^{-2\epsilon})^{2\epsilon-\delta}\} \\
&= 2\epsilon - \inf\{2\epsilon - \delta \colon \mathbf{x} \in (A^{-2\epsilon})^{2\epsilon-\delta}\} \\
&= 2\epsilon - \inf\{\delta \colon \mathbf{x} \in (A^{-2\epsilon})^{\delta}\} && (\operatorname{dist}(\mathbf{x}, ((A^{-2\epsilon})^{2\epsilon})^C) \leq 2\epsilon \\
&= 2\epsilon - \operatorname{dist}(\mathbf{x}, A^{-2\epsilon}) && \text{(Equation 47)}
\end{aligned}$$

Therefore, (46) holds.

□

We next prove two Corollaries to Lemma 29 which state that if a set C is of the form $(A^{-2\epsilon})^\epsilon$, then C is very similar to $(C^\epsilon)^{-\epsilon}$.

Corollary 4. *If $B_\epsilon(\mathbf{x}) \subset (A^{-2\epsilon})^{2\epsilon} - A^{-2\epsilon}$, then $\overline{B_\epsilon(\mathbf{x})}$ intersects both $\partial(A^{-2\epsilon})^{2\epsilon}$ and $\partial A^{-2\epsilon}$.*

Proof. Assume that $\overline{B_\epsilon(\mathbf{x})} \subset (A^{-2\epsilon})^{2\epsilon} - A^{-2\epsilon}$. Then $\operatorname{dist}(\mathbf{x}, A^{-2\epsilon}) \geq \epsilon, \operatorname{dist}(\mathbf{x}, ((A^{-2\epsilon})^{2\epsilon})^C) \geq \epsilon$, and thus Lemma 29 implies that $\operatorname{dist}(\mathbf{x}, A^{-2\epsilon}) = \epsilon, \operatorname{dist}(\mathbf{x}, ((A^{-2\epsilon})^{2\epsilon})^C) = \epsilon$. Thus by Corollary 3, $\overline{B_\epsilon(\mathbf{x})}$ intersects both $\partial A^{-2\epsilon}$ and $\partial((A^{-2\epsilon})^{2\epsilon})^C$. □

Corollary 5. *Let $A \subset \mathbb{R}^d$. The sets $(A^{-2\epsilon})^\epsilon$, $((A^{-2\epsilon})^{2\epsilon})^{-\epsilon}$ have the same boundary and the same interior.*

Proof. To start, by Lemma 24, $(A^{-2\epsilon})^\epsilon \subset ((A^{-2\epsilon})^{2\epsilon})^{-\epsilon}$. Next, pick $\mathbf{x} \in ((A^{-2\epsilon})^{2\epsilon})^{-\epsilon} - (A^{-2\epsilon})^\epsilon$. Then $\mathbf{x} \in ((A^{-2\epsilon})^{2\epsilon})^{-\epsilon}$, so that by Lemma 17 $\overline{B_\epsilon(\mathbf{x})} \subset (A^{-2\epsilon})^{2\epsilon}$. Furthermore, $\overline{B_\epsilon(\mathbf{x})}$ is disjoint from $A^{-2\epsilon}$ by Lemma 18. By Corollary 4, $\overline{B_\epsilon(\mathbf{x})}$ intersects both $\overline{A^{-2\epsilon}}$ and $\overline{((A^{-2\epsilon})^{2\epsilon})^C}$. Then by Lemmas 20 and 19, $\{\mathbf{x}\}$ intersects both $\overline{(A^{-2\epsilon})^\epsilon}$ and $\overline{(((A^{-2\epsilon})^{2\epsilon})^C)^\epsilon}$. Therefore, since $\mathbf{x}$ is not contained in the sets $(A^{-2\epsilon})^\epsilon$, $(((A^{-2\epsilon})^{2\epsilon})^C)^\epsilon$, then $\mathbf{x}$ is in both $\partial(A^{-2\epsilon})^\epsilon$ and $\partial((A^{-2\epsilon})^{2\epsilon})^{-\epsilon}$. □

Lastly, we prove that sets of the form $(A^{-2\epsilon})^\epsilon$ and $(A^{2\epsilon})^{-\epsilon}$ are pseudo-certifiably robust.

Theorem 10. *Let A be any set in $\mathbb{R}^d$. Then both $(A^{2\epsilon})^{-\epsilon}$ and $(A^{-2\epsilon})^\epsilon$ are pseudo-certifiably robust.*

Proof. We first argue that it suffices to show that $(A^{-2\epsilon})^\epsilon$ is pseudo-certifiably robust for any set A.

If $(A^{-2\epsilon})^\epsilon$ is pseudo-certifiably robust for any set A, then $((A^C)^{-2\epsilon})^\epsilon$ is pseudo-certifiably robust. Furthermore, if a set is pseudo-certifiably robust, so is its complement. Thus $(((A^C)^{-2\epsilon})^\epsilon)^C = (A^{2\epsilon})^{-\epsilon}$ is pseudo-certifiably robust.

Now we will show that for any set A, $(A^{-2\epsilon})^\epsilon$ is pseudo-certifiably robust. To start, Lemma 16 implies that every point $\mathbf{y} \in (A^{-2\epsilon})^\epsilon$ is contained in some closed ϵ-ball $\overline{B_\epsilon(\mathbf{x})}$ which is completely contained in $(A^{-2\epsilon})^\epsilon$. Next, we consider points $\mathbf{x}$ outside of $(A^{-2\epsilon})^\epsilon$.

Because both $(A^{-2\epsilon})^\epsilon$, $((A^{-2\epsilon})^{2\epsilon})^{-\epsilon}$ have the same boundary and the same interior, $((A^{-2\epsilon})^\epsilon)^C$ and $(((A^{-2\epsilon})^{2\epsilon})^{-\epsilon})^C$ also have the same interior. However, Lemma 16 implies that for every $\mathbf{x} \in (((A^{-2\epsilon})^{2\epsilon})^{-\epsilon})^C$, there is a ball $\overline{B_\epsilon(\mathbf{y})}$ which contains $\mathbf{x}$ and is contained in $(((A^{-2\epsilon})^{2\epsilon})^{-\epsilon})^C$. Now because $(((A^{-2\epsilon})^{2\epsilon})^{-\epsilon})^C$ and $((A^{-2\epsilon})^\epsilon)^C$ have the same interior, $B_\epsilon(\mathbf{y}) \subset ((A^{-2\epsilon})^\epsilon)^C$. □

E Proof of Lemma 2 and a Generalization (Lemma 33)

E.1 Intermediate Results

We begin by presenting intermediate results that we use in the proof of Lemma 2 and we further discuss intuition about this lemma. Proofs of all intermediate results can be found either in this section, Appendix D.2, or Appendix E.3.

Our first result, proved in Appendix D.2 describes operations how the operations ${}^{\epsilon}$, ${}^{-\epsilon}$, and $F(\cdot)$ as defined in (4) interact.

Lemma 23. *Let A be a subset of $\mathbb{R}^d$. Then the following hold:*

$$A = (A^{-\epsilon})^{\epsilon} \sqcup F(A) \tag{33}$$

$$(A^{\epsilon})^{-\epsilon} = A \sqcup F(A^C). \tag{34}$$

Now recall 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 in the opposite class. This observation leads to the following lemma, which was also proved in Appendix D.2:

Lemma 27. *For any set A, the following hold:*

$$R^\epsilon(A) \geq R((A^{\epsilon})^{-\epsilon}) \tag{37}$$

$$R^\epsilon(A) \geq R((A^{-\epsilon})^{\epsilon}). \tag{38}$$

Start with a minimizing sequence A_n. These two results suggest that we should consider a minimizing sequence of the form $C_n = (((A_n^{-\epsilon})^{\epsilon})^{\epsilon})^{-\epsilon}$. Next, by applying Lemma 21, we show that $(((A_n^{-\epsilon})^{\epsilon})^{\epsilon})^{-\epsilon}$ simplifies to $((A_n^{-\epsilon})^{2\epsilon})^{-\epsilon}$.

Lemma 21. *Let A be a set and define $A^\epsilon = A \oplus \overline{B_\epsilon(\mathbf{0})}$. Then, the following hold:*

$$(A^{\epsilon_1})^{\epsilon_2} = A^{\epsilon_1+\epsilon_2} \qquad \left(A^{-\epsilon_1}\right)^{-\epsilon_2} = A^{-\epsilon_1-\epsilon_2}. \tag{30}$$

It is interesting to note that the proof of the lemma above heavily relies on the fact that the perturbation set is convex. Furthermore, recall that in Appendix D.1, we showed that this statement was actually false for general metric spaces. Next, we use the following lemma to compute C_n^ϵ.

Lemma 24. *For any set A, the following hold:*

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

Lemma 24 implies that $C_n^\epsilon = (A_n^{-\epsilon})^{2\epsilon}$. Thus, we actually have that $\bigcap_{n=1}^\infty C_n^\epsilon = \bigcap_{n=1}^\infty (A_n^{-\epsilon})^{2\epsilon}$ and furthermore every point in C_n^ϵ, is contained in a closed ball radius 2ϵ contained in C_n^ϵ. The goal now is to use this property to argue that in fact $F(\bigcap_{n=1}^\infty C_n^\epsilon) = \emptyset$. Consider a point $\mathbf{x}$ in $\bigcap_{n=1}^\infty C_n^\epsilon$. For each n, we can find a point $\mathbf{b}_n \in C_n^\epsilon$ for which $\mathbf{x} \in \overline{B_{2\epsilon}(\mathbf{b}_n)}$. This implies that each $\mathbf{b}_n$ is in $\overline{B_{2\epsilon}(\mathbf{x})}$, so we can find a convergent subsequence $\{\mathbf{b}_{n_j}\}$ of $\{\mathbf{b}_n\}$. If we can find a point $\mathbf{c}$ such that $\mathbf{x} \in \overline{B_\epsilon(\mathbf{c})}$ and for every $\mathbf{y} \in \overline{B_\epsilon(\mathbf{c})}$, we have $\mathbf{y} \in \overline{B_{2\epsilon}(\mathbf{b}_{n_j})}$ for all sufficiently large j, this will in fact imply that $\mathbf{y} \in \bigcap_{n=1}^\infty C_n^\epsilon$ and thus $\mathbf{x} \in \overline{B_\epsilon(\mathbf{c})} \subset \bigcap_{n=1}^\infty C_n^\epsilon$. Therefore, $F(\bigcap_{n=1}^\infty C_n^\epsilon) = \emptyset$. This observation motivates the following lemma:

Lemma 30. *Let $B_\delta(\mathbf{z})$ be the open ball radius δ centered at a point $\mathbf{z}$ in a norm $\|\cdot\|$. Let $\{\mathbf{b}_n\}$ be a sequence that converges to a point $\mathbf{b}$ and assume that $\mathbf{x} \in \overline{B_{2\epsilon}(\mathbf{b}_n)}$ for all $\mathbf{b}_n$. Let $\mathbf{c} = \frac{1}{2}(\mathbf{b}+\mathbf{x})$. Assume that one of the following three conditions holds:*

1. *The norm $\|\cdot\|$ is strictly convex*
2. *The unit ball in the $\|\cdot\|$ norm is a polytope*
3. *The function $\lambda_{\overline{B_{2\epsilon}(\mathbf{b})}} \colon \overline{B_{2\epsilon}(\mathbf{b})} \times \mathbb{R}^d - \{0\} \to \mathbb{R}$ given by*

   $$\lambda_{\overline{B_{2\epsilon}(\mathbf{b})}}(\mathbf{x}, \mathbf{v}) = \sup\{t \in \mathbb{R} \colon \mathbf{x} + t\mathbf{v} \in \overline{B_{2\epsilon}(\mathbf{b})}\} \tag{48}$$

   is lower semi-continuous on $\overline{B_{2\epsilon}(\mathbf{b})}$ for a fixed $\mathbf{v}$.

Then for all $\mathbf{y} \in \overline{B_\epsilon(\mathbf{c})}$ there is an N for which $n > N$ implies that

$$\mathbf{y} \in \overline{B_{2\epsilon}(\mathbf{b}_n)}.$$

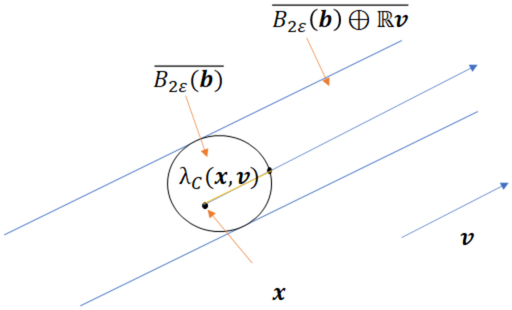
$\overline{B_{2\varepsilon}(\boldsymbol{b})} \oplus \mathbb{R}\boldsymbol{v}$
$\overline{B_{2\varepsilon}(\boldsymbol{b})}$
$\lambda_C(\boldsymbol{x},\boldsymbol{v})$
$\boldsymbol{v}$
$\boldsymbol{x}$

Figure 3: The function $\lambda_{\overline{B_{2\epsilon}(\mathbf{b})}}(\mathbf{x},\mathbf{v})$ is the distance from a point $\mathbf{x}$ to the farthest point for which the line $t \mapsto \mathbf{x} + t\mathbf{v}$ intersects the boundary of $\overline{B_{2\epsilon}(\mathbf{b})}$. The figure suggests that the function $\lambda_{\overline{B_{2\epsilon}(\mathbf{b})}}$ should be continuous for every norm.

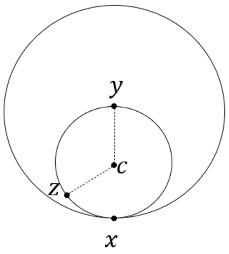

Figure 4: Illustration of the statement of Lemma 32 for the ℓ_2-norm ball, which is strictly convex. The point $\mathbf{x}$ is in the closed ϵ-ball around $\mathbf{c}$ and, excluding $\mathbf{x}$, that set is entirely within an open ball of radius 2ϵ around $\mathbf{y}$.

Specifically, this lemma allow us to conclude that $F(\bigcap_{n=1}^{\infty} C_n^\epsilon) = \emptyset$ for the ℓ_1 norm, the ℓ_∞ norm, and all strictly convex norms. In Appendix E.3, we prove part 2) by first proving 3) and then verifying this statement for norms whose unit ball is a polytope. In fact, more generally, one can prove that $\lambda_C(\cdot,\mathbf{v})$ is always continuous. However, the proof is much longer, thus we are leaving it to an extended version of this paper. See Figure 3 for further intuition regarding the third condition.

Here we will prove just 1) of Lemma 30, and we will delay the rest of the proof of this lemma to Section E.3. We will start with proving two short lemmas.

The first lemma proves our result for when $\|\mathbf{x} - \mathbf{b}\| < 2\epsilon$.

Lemma 31. *Let $\mathbf{b}_n \to \mathbf{b}$, and pick a point $\mathbf{y} \in B_{2\epsilon}(\mathbf{b})$. Then for sufficiently large n, $\mathbf{y} \in B_{2\epsilon}(\mathbf{b}_n)$.*

Proof of Lemma 31. Pick $\mathbf{y} \in B_{2\epsilon}(\mathbf{b})$. Set $t = \|\mathbf{y} - \mathbf{b}\| < 2\epsilon$. Pick n sufficiently large so that $\|\mathbf{b} - \mathbf{b}_n\| < 2\epsilon - t$. Then by the triangle inequality,

$$\|\mathbf{y} - \mathbf{b}_n\| \leq \|\mathbf{y} - \mathbf{b}\| + \|\mathbf{b} - \mathbf{b}_n\| < \|\mathbf{y} - \mathbf{b}\| + 2\epsilon - t = t + 2\epsilon - t = 2\epsilon.$$

□

The following lemma helps prove Lemma 30 when $\|\mathbf{b} - \mathbf{x}\| = 2\epsilon$.

Lemma 32. *Let $\|\cdot\|$ be a strictly convex norm and $B_\epsilon(\mathbf{z})$ denote an open ball in that norm. Let $B_{2\epsilon}(\mathbf{b})$ be a ball and let $\mathbf{x} \in \partial B_{2\epsilon}(\mathbf{b})$. Then if $\mathbf{c} = \frac{1}{2}(\mathbf{b} + \mathbf{x})$, $\overline{B_\epsilon(\mathbf{c})} - \{\mathbf{x}\} \subset B_{2\epsilon}(\mathbf{b})$.*

Proof of Lemma 32. Pick $\mathbf{c} = \frac{1}{2}(\mathbf{x} + \mathbf{b})$. Then $\|\mathbf{x} - \mathbf{c}\| = \frac{1}{2}\|\mathbf{b} - \mathbf{x}\| = \epsilon$, so that $\mathbf{x} \in \partial B_\epsilon(\mathbf{c})$. Next, if $\mathbf{y} \in B_\epsilon(\mathbf{c})$, then $\|\mathbf{y} - \mathbf{c}\| < \epsilon$ which implies that $\|\mathbf{y} - \mathbf{b}\| \leq \|\mathbf{y} - \mathbf{c}\| + \|\mathbf{c} - \mathbf{b}\| < 2\epsilon$. Thus $B_\epsilon(\mathbf{c}) \subset B_{2\epsilon}(\mathbf{b})$. Next, pick $\mathbf{y} \in \partial B_\epsilon(\mathbf{c})$ with $\mathbf{y} \neq \mathbf{x}$. Then $\mathbf{b} - \mathbf{c} \neq \mathbf{c} - \mathbf{y}$ and thus by strict convexity,

$$\|\mathbf{b} - \mathbf{y}\| = \|(\mathbf{b} - \mathbf{c}) + (\mathbf{c} - \mathbf{y})\| < 2\epsilon.$$

Therefore, $\overline{B_\epsilon(\mathbf{c})} - \{\mathbf{x}\} \subset B_{2\epsilon}(\mathbf{b})$. □

Proof of 1) of Lemma 30. Assume that the norm is strictly convex. First assume that $\|\mathbf{x}-\mathbf{b}\| < 2\epsilon$. Then every point $\mathbf{y}$ in $\overline{B_\epsilon(\mathbf{c})}$ is actually in $B_{2\epsilon}(\mathbf{b})$:

$$\|\mathbf{y}-\mathbf{b}\| \leq \|\mathbf{y}-\mathbf{c}\| + \|\mathbf{c}-\mathbf{b}\| = \|\mathbf{y}-\mathbf{c}\| + \frac{1}{2}\|\mathbf{x}-\mathbf{b}\| < \epsilon+\epsilon = 2\epsilon$$

Therefore, Lemma 31 implies that $\mathbf{y} \in \overline{B_{2\epsilon}(\mathbf{b}_n)}$ for sufficiently large n.

If in fact $\|\mathbf{b}-\mathbf{x}\| = 2\epsilon$ then Lemma 32 implies that $\overline{B_\epsilon(\mathbf{c})} - \{\mathbf{x}\} \subset B_{2\epsilon}(\mathbf{b})$. Therefore, Lemma 31 implies that if $\mathbf{y} \in B_\epsilon(\mathbf{c})$ and $\mathbf{y} \neq \mathbf{x}$, then for sufficiently large n, $\mathbf{y} \in \overline{B_{2\epsilon}(\mathbf{n})}$.

Furthermore, in the hypotheses in Lemma 30, we assumed that $\mathbf{x} \in \overline{B_{2\epsilon}(\mathbf{b}_n)}$ for all n. Therefore, the claim holds for every point in $\overline{B_\epsilon(\mathbf{c})}$. □

E.2 Proof of Lemmas 2 and 33

We state and prove more general version of Lemma 2.

Lemma 33. *Let B_n be a decreasing sequence ($B_{n+1} \subset B_n$),let $B_\epsilon(\mathbf{z})$ denote an ϵ-ball in a norm $\|\cdot\|$, and define $A^\epsilon = A \oplus \overline{B_\epsilon(\mathbf{0})}$. Assume that the norm $\|\cdot\|$ satisfies one of the three following conditions:*

1) The norm $\|\cdot\|$ is strictly convex

2) The unit ball in the $\|\cdot\|$ norm is a polytope

3) The function given by (48) is lower semi-continuous on $\overline{B_{2\epsilon}(\mathbf{b})}$ for a fixed $\mathbf{v}$.

Then if we define $C_n = ((B_n^{-\epsilon})^{2\epsilon})^{-\epsilon}$, this sequence satisfies $R^\epsilon(C_n) \leq R^\epsilon(B_n)$. Furthermore, $\bigcap_{n=1}^\infty C_n$ is pseudo-certifiably robust at every point and satisfies

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

Proof of Lemma 2. We will first we show that $R^\epsilon(C_n) \leq R^\epsilon(B_n)$. Lemma 21 gives

$$C_n = \left(\left(B_n^{-\epsilon}\right)^{2\epsilon}\right)^{-\epsilon} = \left(\left(\left(B_n^{-\epsilon}\right)^{\epsilon}\right)^{\epsilon}\right)^{-\epsilon}. \tag{49}$$

Lemma 27 then implies that $R^\epsilon(C_n) \leq R^\epsilon(B_n)$. We now discuss pseudo-certifiable robustness. Since $\bigcap_{n=1}^\infty C_n = \left(\bigcap_{n=1}^\infty (B_n^{-\epsilon})^{2\epsilon}\right)^{-\epsilon}$, Lemma 16 implies that if $\mathbf{x} \notin \bigcap_{n=1}^\infty C_n$, there is an ϵ-ball containing $\mathbf{x}$ that is contained in the complement of $\bigcap_{n=1}^\infty C_n$. We now consider a point $\mathbf{x} \in \bigcap_{n=1}^\infty C_n$. Then Theorem 10 implies that for each n, there is a $\mathbf{z}_n \in C_n$ for which $\mathbf{x} \in \overline{B_\epsilon(\mathbf{z}_n)}$ and $B_\epsilon(\mathbf{z}_n) \subset C_n$. Since each $\mathbf{z}_n \in \overline{B_\epsilon(\mathbf{x})}$, we can choose a limit point $\mathbf{z}$ of this sequence in $\overline{B_\epsilon(\mathbf{x})}$. Then clearly $\mathbf{x} \in \overline{B_\epsilon(\mathbf{z})}$ and Lemma 31 implies that for sufficiently large n, $\mathbf{y} \in \overline{B_\epsilon(\mathbf{z})}$ implies that $\mathbf{y} \in \overline{B_\epsilon(\mathbf{z}_n)}$. Therefore $\overline{B_\epsilon(\mathbf{z})} \subset \bigcap_{n=1}^\infty C_n$.

Next, recall that Equation 5 states that $\bigcap_{n=1}^\infty C_n^{-\epsilon} = (\bigcap_{n=1}^\infty C_n)^{-\epsilon}$, so it remains to prove $\bigcap_{n=1}^\infty C_n^\epsilon = (\bigcap_{n=1}^\infty C_n)^\epsilon$. The rest of the proof will be devoted to showing this fact. To start, we will argue that $C_n^\epsilon = (B_n^{-\epsilon})^{2\epsilon}$.

$$\begin{aligned}
C_n^\epsilon &= \left(\left(\left(\left(B_n^{-\epsilon}\right)^\epsilon\right)^\epsilon\right)^{-\epsilon}\right)^\epsilon && \text{(Equation (49))}\\
&= \left(\left(\left(\left(B_n^{-\epsilon}\right)^\epsilon\right)^\epsilon\right)^{-\epsilon}\right)^\epsilon && \text{(associativity of addition)}\\
&= \left(\left(B_n^{-\epsilon}\right)^\epsilon\right)^\epsilon && \text{(Lemma 24)}\\
&= \left(B_n^{-\epsilon}\right)^{2\epsilon} && \text{(Lemma 21)}
\end{aligned}$$

Therefore, we have $\bigcap_{n=1}^{\infty} C_n^{\epsilon} = \bigcap_{n=1}^{\infty}(B_n^{-\epsilon})^{2\epsilon}$. Next, using Equation 5 and Lemma 23, we obtain:

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

It remains to show that the second term involving F is empty. Pick $\mathbf{x} \in \bigcap_{n=1}^{\infty}(B_n^{-\epsilon})^{2\epsilon}$. Then for each n, $\mathbf{x} \in (B_n^{-\epsilon})^{2\epsilon}$ and thus there is a $\mathbf{b}_n$ for which $\mathbf{x} \in \overline{B_{2\epsilon}(\mathbf{b}_n)} \subset (B_n^{-\epsilon})^{2\epsilon}$. However, this relation implies that each $\mathbf{b}_n$ is in $\overline{B_{2\epsilon}(\mathbf{x})}$, so we can pick a convergent subsequence $\{\mathbf{b}_{n_j}\}$ of the $\{\mathbf{b}_n\}$s which converges to a point $\mathbf{b}$ in $\overline{B_{2\epsilon}(\mathbf{x})}$. Therefore, the subsequence $\{\mathbf{b}_{n_j}\}$ satisfies the hypotheses of Lemma 30. Hence if we let $\mathbf{c} = \frac{1}{2}(\mathbf{x} + \mathbf{b})$, then for every $\mathbf{y} \in \overline{B_\epsilon(\mathbf{c})}$, $\mathbf{y} \in \overline{B_{2\epsilon}(\mathbf{b}_{n_j})} \subset (B_{n_j}^{-\epsilon})^{2\epsilon}$ for sufficiently large j. However, because the sequence of sets $(B_{n_j}^{-\epsilon})^{2\epsilon}$ is decreasing, $\mathbf{y} \in (B_n^{-\epsilon})^{2\epsilon}$ for all n and therefore $\mathbf{y} \in \bigcap_{n=1}^{\infty} \left(B_n^{-\epsilon}\right)^{2\epsilon}$. This statement implies that $\mathbf{x} \in \overline{B_\epsilon(\mathbf{c})} \subset \bigcap_{n=1}^{\infty} \left(B_n^{-\epsilon}\right)^{2\epsilon}$ and hence $F(\bigcap_{n=1}^{\infty} \left(B_n^{-\epsilon}\right)^{2\epsilon}) = \emptyset$. □

E.3 Lemma 2 for Non-Strictly Convex Norms

Proving Lemma 30 for non-strictly convex norms is a lot trickier than for strictly convex norms. We will use a totally different approach which we begin below.

We will start by proving 3) of Lemma 30.

Proof of 3) of Lemma 30. Pick a point $\mathbf{y}$ different from $\mathbf{x}$ in $\overline{B_\epsilon(\mathbf{c})}$. Because $\mathbf{x} \in \overline{B_{2\epsilon}(\mathbf{b}_n)}$, we can define $\lambda_{\overline{B_{2\epsilon}(\mathbf{b}_n)}}(\mathbf{x}, \mathbf{y} - \mathbf{x})$. However,

$$\begin{aligned}&\lambda_{\overline{B_{2\epsilon}(\mathbf{b}_n)}}(\mathbf{x}, \mathbf{y} - \mathbf{x}) = \sup\{t \geq 0 \colon \mathbf{x} + t(\mathbf{y} - \mathbf{x}) \in B_{2\epsilon}(\mathbf{b}_n)\} = \\ &\sup\{t \geq 0 \colon \|\mathbf{x} + t(\mathbf{y} - \mathbf{x}) - \mathbf{b}_n\| \leq 2\epsilon\} = \sup\{t \geq 0 \colon \|\mathbf{x} + t(\mathbf{y} - \mathbf{x}) + (\mathbf{b} - \mathbf{b}_n) - \mathbf{b}\| \leq 2\epsilon\} \\ &= \sup\{t \geq 0 \colon \mathbf{x} + t(\mathbf{y} - \mathbf{x}) + \mathbf{b} - \mathbf{b}_n \in B_{2\epsilon}(\mathbf{b})\} \\ &= \lambda_{\overline{B_{2\epsilon}(\mathbf{b})}}(\mathbf{x} + \mathbf{b} - \mathbf{b}_n, \mathbf{y} - \mathbf{x})\end{aligned}$$

Furthermore, $\mathbf{x} \in \overline{B_{2\epsilon}(\mathbf{b}_n)}$ is equivalent to $\mathbf{x} + \mathbf{b} - \mathbf{b}_n \in \overline{B_{2\epsilon}(\mathbf{b})}$.

Next, note that the mapping $F(\mathbf{z}) = 2\mathbf{z} - \mathbf{x}$ maps $\overline{B_\epsilon(\mathbf{c})}$ to $\overline{B_{2\epsilon}(\mathbf{b})}$. Specifically, this implies that $2\mathbf{y} - \mathbf{x} = \mathbf{x} + 2(\mathbf{y} - \mathbf{x}) \in \overline{B_{2\epsilon}}(\mathbf{b})$ and thus $\lambda_{\overline{B_{2\epsilon}(\mathbf{b})}}(\mathbf{x}, \mathbf{y} - \mathbf{x}) \geq 2$. Therefore, because each $\mathbf{x} + \mathbf{b} - \mathbf{b}_n$ is in the convex set $\overline{B_{2\epsilon}(\mathbf{b})}$, the lower semi-continuity of λ implies that

$$\liminf_{n \to \infty} \lambda_{B_{2\epsilon}(\mathbf{b}_n)}(\mathbf{x}, \mathbf{y} - \mathbf{x}) = \liminf_{n \to \infty} \lambda_{B_{2\epsilon}(\mathbf{b})}(\mathbf{x} + \mathbf{b} - \mathbf{b}_n, \mathbf{y} - \mathbf{x}) \geq \lambda_{B_{2\epsilon}(\mathbf{b})}(\mathbf{x}, \mathbf{y} - \mathbf{x}) \geq 2$$

Specifically, for sufficiently large $\mathbf{b}_n$, $\mathbf{y} = \mathbf{x} + (\mathbf{y} - \mathbf{x}) \in \overline{B_{2\epsilon}(\mathbf{b}_n)}$. □

Next, one can argue that if C is a closed, convex and bounded set, then $\lambda_C(\cdot, \mathbf{v})$ defined by

$$\lambda_C(\mathbf{x}, \mathbf{v}) = \sup\{t \in \mathbb{R} \colon \mathbf{x} + t\mathbf{v} \in C\} \tag{50}$$

is concave and bounded on C. This fact is quite helpful because concave functions are 'almost' continuous, in fact they are continuous on every open set. However, showing that concave functions are continuous closed sets is quite tricky. We now formally state and prove a continuity result for λ_C on the closed set C.

Lemma 34. *The following two properties hold for any closed, convex, and bounded set* C*.*

1. $0 \leq \lambda_C(\mathbf{x}, \mathbf{v}) < \infty$ *for* $\mathbf{x} \in C$
2. $\lambda_C(\cdot, \mathbf{v})$ *is concave.*

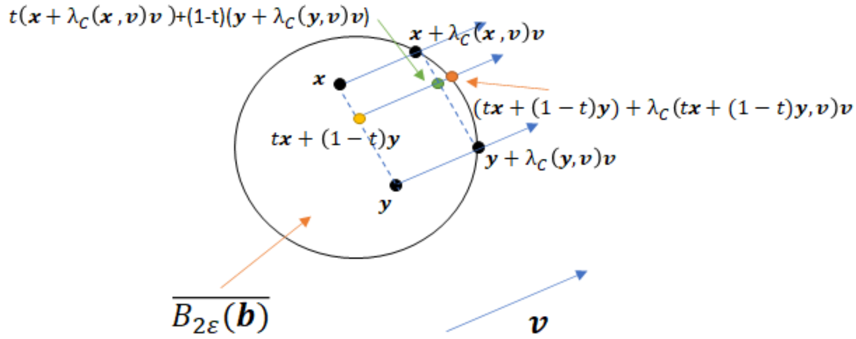
$t(x + \lambda_C(x, v)v) + (1-t)(y + \lambda_C(y, v)v)$
$x + \lambda_C(x, v)v$
x
$(tx + (1-t)y) + \lambda_C(tx + (1-t)y, v)v$
$tx + (1-t)y$
$y + \lambda_C(y, v)v$
y
$\overline{B_{2\varepsilon}(b)}$
v

Figure 5: In this picture, we consider $C = \overline{B_{2\epsilon}(\mathbf{b})}$. By convexity, if both $\mathbf{x}$ and $\mathbf{y}$ are in C, so is $t\mathbf{x} + (1-t)\mathbf{y}$. Similarly, as both $\mathbf{x} + \lambda_C(\mathbf{x}, \mathbf{v})\mathbf{v}$ and $\mathbf{y} + \lambda_C(\mathbf{y}, \mathbf{v})\mathbf{v}$ are in C, the convex combination $t(\mathbf{x} + \lambda_C(\mathbf{x}, \mathbf{v})\mathbf{v}) + (1-t)(\mathbf{y} + \lambda_C(\mathbf{y}, \mathbf{v})\mathbf{v})$ is in C as well. This implies that $\lambda_C(t\mathbf{x} + (1-t)\mathbf{y}, \mathbf{v}) \geq t\lambda_C(\mathbf{x}, \mathbf{v}) + (1-t)\lambda_C(\mathbf{y}, \mathbf{v})$.

See Figure 5 for an illustration of this proof.

Proof.
Showing 1): If $\mathbf{x} \in C$, then $\mathbf{x} + 0\mathbf{v} \in C$ so $\lambda_C(\mathbf{x}, \mathbf{v}) \geq 0$. Because C is bounded, $\mathbf{x} + s\mathbf{v}$ is not in C for some sufficiently large S. This implies $\lambda_C(\mathbf{x}, \mathbf{v}) < \infty$.

Showing 2): Pick $\mathbf{x}, \mathbf{y} \in C$ and $0 \leq t \leq 1$. Then by convexity, $t\mathbf{x} + (1-t)\mathbf{y} \in C$. Now note that because C is closed, both $\mathbf{x} + \lambda_C(\mathbf{x}, \mathbf{v})\mathbf{v}$ and $\mathbf{y} + \lambda_C(\mathbf{y}, \mathbf{v})\mathbf{v}$ are in C. Therefore, their convex combination $t(\mathbf{x} + \lambda_C(\mathbf{x}, \mathbf{v})\mathbf{v}) + (1-t)(\mathbf{y} + \lambda_C(\mathbf{y}, \mathbf{v})\mathbf{v}) = (t\mathbf{x} + (1-t)\mathbf{y}) + (t\lambda_C(\mathbf{x}, \mathbf{v}) + (1-t)\lambda_C(\mathbf{y}, \mathbf{v}))\mathbf{v}$ is also in C and thus $\lambda_C(t\mathbf{x} + (1-t)\mathbf{y}, \mathbf{v}) \geq t\lambda_C(\mathbf{x}, \mathbf{v}) + (1-t)\lambda_C(\mathbf{y}, \mathbf{v})$.

□

Lastly, a consequence of Corollary 3.10 of Rockafellar (2015) implies that $\lambda_C(\cdot, \mathbf{v})$ is lower semi-continuous if C is a polytope.

Lemma 35. *Let P be a polytope. Then any-real valued concave function defined on P is lower semi-continuous.*

This lemma together with 3) of Lemma 30 therefore proves Lemma 2 for any norm whose ball is a polytope, including the ℓ_1 and ℓ_∞ norms.

F Proof of Lemma 3 and a Generalization (Lemma 37)

We begin by reviewing some results of Appendix D. To start, recall that the operation $A^\epsilon = A \oplus \overline{B_\epsilon(\mathbf{0})}$, satisfies the relations of (6):

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

Next, recall that in Appendix D, we showed that the ${}^{-\epsilon}$ operation has analogous properties, and the proof followed only from the equalities in (6). Similarly, one can show that any set operation ${}^\epsilon$ satisfying the properties of (6) also satisfies analogous properties for the ${}^{-\epsilon}$ operation. In the next section, we will prove a version of our theorem for other models of perturbations. Instead of assuming perturbations in a ball in $\mathbb{R}^d$, we will let A^ϵ denote all perturbations of A and we will assume that the set operation A^ϵ satisfies (6). These two assumptions allow us to prove the relations of Lemma 15. We state this result without proof, as the proof is the same as the proofs of (25) and (24) of Lemma 15.

Lemma 36. *Let ${}^\epsilon$ be a set operation that satisfies the relations of (6), and define the operation ${}^{-\epsilon}$ via $A^{-\epsilon} = ((A^C)^\epsilon)^C$. Then*

For any sequence of sets $\{A_i\}$, the following set containments hold:

$$\bigcup_{i=1}^{\infty} A_i^{\epsilon} = \left[\bigcup_{i=1}^{\infty} A_i\right]^{\epsilon} \quad (51) \qquad \bigcap_{i=1}^{\infty} A_i^{-\epsilon} = \left[\bigcap_{i=1}^{\infty} A_i\right]^{-\epsilon} \quad (52)$$

$$\bigcap_{i=1}^{\infty} A_i^{\epsilon} \supset \left[\bigcap_{i=1}^{\infty} A_i\right]^{\epsilon} \quad (53) \qquad \bigcup_{i=1}^{\infty} A_i^{-\epsilon} \subset \left[\bigcup_{i=1}^{\infty} A_i\right]^{-\epsilon} \quad (54)$$

Parts (51) and (53) are just restating (6), we repeat them here as well for clarity in the exposition.

In the rest of this appendix, rather than focusing on $\mathbb{R}^d$, we will assume that $^{\epsilon}$ is a set operation that satisfies (6). This formulation will allow us to prove the 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 3.

Lemma 37. *Let $^{\epsilon}$ be a set operation for which A^{ϵ} is universally measurable if A is universally measurable. Further assume that $^{\epsilon}$ satisfies properties* (51) *and* (53)*.*

Let A_n be a minimizing sequence of R^{ϵ} for which $\liminf A_n^{\epsilon} \doteq \limsup A_n^{\epsilon}$ *and* $\liminf A_n^{-\epsilon} = \limsup A_n^{-\epsilon}$*. Then there is a decreasing minimizing sequence B_n (in other words $B_{n+1} \subset B_n$) for which* $\liminf B_n^{\epsilon} \doteq \limsup B_n^{\epsilon} = \limsup A_n^{\epsilon}$ *and* $\limsup B_n^{-\epsilon} \doteq \liminf B_n^{-\epsilon} \dot{\supset} \liminf A_n^{-\epsilon}$.

Note that Lemma 3 is simply Lemma 37 combined with the fact that A^{ϵ} defined as $A \oplus \overline{B_{\epsilon}(\mathbf{0})}$ satisfies (6) (shown in Lemma 15). Thus we will prove Lemma 37 in this section. To start, we state two lemmas that discuss the $\liminf$ and $\limsup$ of monotonic sequences of sets.

Lemma 38. *Let S_n be a decreasing sequence of sets. Then*

$$\limsup S_n = \liminf S_n = \bigcap_{n \geq 1} S_n.$$

Proof. First, note that for a decreasing sequence of sets S_n,

$$\bigcup_{n \geq N} S_n = S_N.$$

Therefore,

$$\limsup S_n = \bigcap_{N \geq 1} \bigcup_{n \geq N} S_n = \bigcap_{N \geq 1} S_N.$$

Furthermore, if S_n is decreasing, then

$$\bigcap_{n \geq N} S_n = \bigcap_{n \geq 1} S_n.$$

This observation implies that

$$\liminf S_n = \bigcup_{N \geq 1} \bigcap_{n \geq N} S_n = \bigcup_{N \geq 1} \bigcap_{n \geq 1} S_n = \bigcap_{n \geq 1} S_n.$$

□

Corollary 6. *Let S_n be an increasing sequence of sets. Then*

$$\limsup S_n = \liminf S_n = \bigcup_{n \geq 1} S_n.$$

Proof. Apply the previous lemma to the sequence S_n^C, and then take complements. □

With these Lemmas in hand, we prove Lemma 37.

Proof of Lemma 37. We will show that if A_k is a minimizing sequence, then

$$B_n = \bigcup_{k \geq n} A_k$$

is a minimizing sequence as well. In summary, we will show that $\limsup B_n^\epsilon = \liminf B_n^\epsilon = \limsup A_n^\epsilon \doteq \liminf A_n^\epsilon$, and that $\limsup B_n^{-\epsilon} = \liminf B_n^{-\epsilon} \supset \limsup A_n^{-\epsilon} \doteq \liminf A_n^{-\epsilon}$. Subsequently, we will prove that these two facts imply that $\lim_{n\to\infty} R^\epsilon(B_n) \leq \lim_{n\to\infty} R^\epsilon(A_n)$ and thus $\{B_n\}$ is a minimizing sequence.

First, note that since B_n is a decreasing sequence, Lemma 38 implies that it suffices to verify $\limsup B_n^\epsilon \doteq \limsup A_n^\epsilon$.

We start with showing that $\limsup A_n^\epsilon = \limsup B_n^\epsilon$

By equation (51),

$$B_n^\epsilon = \bigcup_{k \geq n} A_k^\epsilon.$$

Then because B_n is decreasing, Lemma 38 implies that

$$\limsup B_n^\epsilon = \bigcap_{N \geq 1} B_N^\epsilon = \bigcap_{N \geq 1} \bigcup_{n \geq N} A_n^\epsilon = \limsup A_n^\epsilon \tag{55}$$

We now consider the ${}^{-\epsilon}$ operation. By equation (54),

$$B_n^{-\epsilon} \supset \bigcup_{k \geq n} A_k^{-\epsilon} \tag{56}$$

Thus

$$\limsup B_n^{-\epsilon} \supset \limsup \left(\bigcup_{k \geq n} A_k^{-\epsilon} \right) = \bigcap_{N \geq 1} \bigcup_{n \geq N} \bigcup_{k \geq n} A_k^{-\epsilon} = \bigcap_{N \geq 1} \bigcup_{n \geq N} A_n^{-\epsilon} = \limsup A_n^{-\epsilon}.$$

By taking complements, we see that this relation implies

$$\liminf (B_n^{-\epsilon})^C = (\limsup B_n^{-\epsilon})^C \subset (\limsup A_n^{-\epsilon})^C = \liminf (A_n^{-\epsilon})^C. \tag{57}$$

As $B_n^{-\epsilon}$ is decreasing, $(B_n^{-\epsilon})^C$ is increasing, which implies that the limit $\lim_{n\to\infty} \mathbb{1}_{(B_n^{-\epsilon})^C}(x)$ exists at every x. Furthermore, Corollary 6 implies that $\lim_{n\to\infty} \mathbb{1}_{(B_n^{-\epsilon})^C} = \mathbb{1}_{\limsup (B_n^{-\epsilon})^C} = \mathbb{1}_{\liminf (B_n^{-\epsilon})^C}$. Moreover, we can evaluate $\lim_{n\to\infty} R^\epsilon(B_n)$ by using the dominated convergence theorem.

$$\begin{aligned}
&\lim_{n\to\infty} R^\epsilon(B_n) = \lim_{n\to\infty} \int (1 - \eta(x)) \mathbb{1}_{B_n^\epsilon}(x) + \eta(x) \mathbb{1}_{(B_n^{-\epsilon})^C}(x) d\mathbb{P} \\
&= \int (1 - \eta(x)) \mathbb{1}_{\limsup B_n^\epsilon}(x) + \eta(x) \mathbb{1}_{\liminf (B_n^{-\epsilon})^C}(x) d\mathbb{P} \\
&= \int (1 - \eta(x)) \mathbb{1}_{\limsup A_n^\epsilon}(x) + \eta(x) \mathbb{1}_{\liminf (B_n^{-\epsilon})^C}(x) d\mathbb{P} \qquad \text{(By (55))}
\end{aligned} \tag{58}$$

Similarly, because $\limsup A_n^\epsilon \doteq \liminf A_n^\epsilon$ and $\limsup A_n^{-\epsilon} \doteq \liminf A_n^{-\epsilon}$, the limits $\lim_{n\to\infty} \mathbb{1}_{A_n^\epsilon}(x)$, $\lim_{n\to\infty} \mathbb{1}_{(A_n^{-\epsilon})^C}(x)$ exist $\mathbb{P}$-a.e., so one can apply the dominated convergence theorem to $R^\epsilon(A_n)$.

$$\begin{aligned}
&\lim_{n\to\infty} R^\epsilon(A_n) = \lim_{n\to\infty} \int (1 - \eta(x)) \mathbb{1}_{A_n^\epsilon}(x) + \eta(x) \mathbb{1}_{(A_n^{-\epsilon})^C}(x) d\mathbb{P} \\
&= \int (1 - \eta(x)) \mathbb{1}_{\limsup A_n^\epsilon}(x) + \eta(x) \mathbb{1}_{\liminf (A_n^{-\epsilon})^C}(x) d\mathbb{P}
\end{aligned} \tag{59}$$

Therefore we can use (58) and (59) to evaluate the difference $\lim_{n\to\infty} R^\epsilon(B_n) - \lim_{n\to\infty} R^\epsilon(A_n)$:

$$\begin{aligned}
&\lim_{n\to\infty} R^\epsilon(B_n) - \lim_{n\to\infty} R^\epsilon(A_n)\\
&= \int \eta(x)\left(\mathbb{1}_{\liminf(B_n^{-e})^C} - \mathbb{1}_{\liminf(A_n^{-\epsilon})^C}\right) d\mathbb{P}\\
&= \int -\eta(x)\left(\mathbb{1}_{\liminf(A_n^{-\epsilon})^C} - \mathbb{1}_{\liminf(B_n^{-e})^C}\right) d\mathbb{P}\\
&= \int -\eta(x)\mathbb{1}_{\liminf((A_n)^{-\epsilon})^C - \liminf((B_n)^{-\epsilon})^C} d\mathbb{P} && \text{(By equation (57))}\\
&\leq 0 && (\eta(x) \geq 0)
\end{aligned}$$

Therefore we have shown that

$$\lim_{n\to\infty} R^\epsilon(B_n) \leq \lim_{n\to\infty} R^\epsilon(A_n).$$

However, A_n is a minimizing sequence for R^ϵ. We conclude that B_n is a minimizing sequence as well. □

G More General Results

In this Appendix, we present two generalizations of our main result. One generalization concerns other models of perturbations. As discussed in Section 6, there are many other possible models of perturbations for 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.

The second generalization discusses the question of when a minimizing sequence of R^ϵ corresponds to an adversarial Bayes classifier. The result we present in this Section (Theorem 12) states that if either $\eta \in (0,1)$ or $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure, *every* minimizing sequence of R^ϵ must have a subsequence that in some sense approaches an adversarial Bayes classifier. This result indicates that a minimizing sequence cannot 'diverge to infinity', so when studying consistency for adversarial learning, it suffices consider adversarial Bayes classifiers, instead of minimizing sequences of R^ϵ.

We now present our general theorem for alternative models of perturbations.

Theorem 11. *Let X be a separable metric space and let $\mathcal{B}(X), \mathscr{U}(X)$ be the corresponding Borel and universal σ-algebras respectively. Further 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*

$$\bigcup_{n\in\mathbb{N}} A_n^\epsilon = \left(\bigcup_{n\in\mathbb{N}} A_n\right)^\epsilon \tag{51}$$

$$\bigcap_{n\in\mathbb{N}} A_n^\epsilon \supset \left(\bigcap_{n\in\mathbb{N}} A_n\right)^\epsilon \tag{53}$$

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 given a decreasing minimizing sequence B_n of R^ϵ in $\mathscr{U}(X)$, one can find a decreasing minimizing sequence C_n in $\mathscr{U}(X)$ for which

$$\bigcap_{n=1}^\infty C_n^\epsilon \dot{\subset} \left(\bigcap_{n=1}^\infty C_n\right)^\epsilon \tag{60}$$

Where $\dot{\subset}$ denotes containment up to a set measure zero. Then there exists a minimizer to R^ϵ in the σ-algebra $\mathscr{U}(X)$.

Note that

$$\bigcap_{n=1}^{\infty} C_n^{-\epsilon} = \left(\bigcap_{n=1}^{\infty} C_n \right)^{-\epsilon} \tag{61}$$

follows by taking complements of (51), which is required for an analog of Lemma 2. In Section 4.3, this relation is Equation 5. Similarly, (60) and (53) together imply

$$\bigcap_{n=1}^{\infty} C_n^{\epsilon} \doteq \left(\bigcap_{n=1}^{\infty} C_n \right)^{\epsilon} \tag{62}$$

which the analog of Lemma 2.

We now discuss how some of the tools used to prove Lemma 2 extend to general metric spaces. The proof of Lemma 2 follows from Lemmas 21, 23, 24, 27, and 30–see Appendix E for an overview of these Lemmas and their role in proving Lemma 2. Lemmas 23, 24, 27, and Equation (5) only depend on the properties in (51) and (53). Lemma 21 does not hold for all metric spaces, but intuition suggests that this lemma should be true for metric spaces on a continuum. We now discuss generalizing Lemma 30. For strictly convex norms, this lemma was proved using Lemma 31 and Lemma 32. Lemma 31 holds when the perturbation set is a ball in a metric space and Lemma 32 can be proved when the perturbation set is a strictly convex ball in a metric space X and X satisfies a condition called the *midpoint property*.

We now state our second generalization of Theorem 1. In this paper, we have shown the existence of an adversarial Bayes classifier. However, Theorem 1 does not preclude the existence of a minimizing sequence A_n of R^ϵ which diverges.

We present a generalization of Theorem 1 which states that if $\eta \in (0,1)$ or $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure, then *every* minimizing sequence of sets $\{A_n\}$ has a subsequence $\{A_{n_k}\}$ for which $A_{n_k}^\epsilon, A_{n_k}^{-\epsilon}$ approach $A^\epsilon, A^{-\epsilon}$ for an adversarial Bayes classifier A. 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. Formally, we say 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 minimizing 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 in this case all minimizing sequences diverge.

Theorem 12. *Let $\mathbb{P}$ be the completion of a Borel measure on $\mathcal{B}(\mathbb{R}^d)$ restricted to $\mathscr{U}(\mathbb{R}^d)$. Let $B_\epsilon(\mathbf{0})$ be a ball in a norm that satisfies one of the following three conditions:*

1) *The norm $\|\cdot\|$ is strictly convex*

2) *The unit ball in the $\|\cdot\|$ norm is a polytope*

3) *The function given by (48) is lower semi-continuous on $\overline{B_{2\epsilon}(\mathbf{b})}$ for a fixed $\mathbf{v}$.*

Then if A_n is a minimizing sequence of R^ϵ, then the following hold:

a) There exists a subsequence A_{n_k} *for which* $\limsup A^\epsilon_{n_k} \doteq \liminf A^\epsilon_{n_k}$ *and* $\limsup A^{-\epsilon}_{n_k} \doteq \liminf A^{-\epsilon}_{n_k}$.

b) If A_{n_k} *is a sequence for which* $\limsup A^\epsilon_{n_k} \doteq \liminf A^\epsilon_{n_k}$ *and* $\limsup A^{-\epsilon}_{n_k} \doteq \liminf A^{-\epsilon}_{n_k}$, *then there exists a minimizer* A *of* R^ϵ *for which* $A^{-\epsilon}$ *equals* $\limsup A^{-\epsilon}_{n_k}$ *outside* $\{\eta = 0\}$ *and* A^ϵ *equals* $\limsup A^\epsilon_{n_k}$ *outside of the set* $\{\eta = 1\}$. *Formally,* $A^{-\epsilon} \cap \eta^{-1}((0,1]) \doteq \limsup A^{-\epsilon}_{n_k} \cap \eta^{-1}((0,1]) \doteq \liminf A^{-\epsilon}_{n_k} \cap \eta^{-1}((0,1])$ *and* $A^\epsilon \cap \eta^{-1}([0,1)) \doteq \limsup A^\epsilon_{n_k} \cap \eta^{-1}([0,1)) \doteq \liminf A^\epsilon_{n_k} \cap \eta^{-1}([0,1))$. *Furthermore,* A *is pseudo-ceritfiably robust.*

c) Let A_{n_k} *be the subsequence in a) and* A *the adversarial Bayes classifier in b). If either* $\eta \in (0,1)$ *a.e. or* $\mathbb{P}$ *is absolutely continuous with respect to Lebesgue measure, then there exists an adversarial Bayes classifier* E *for which* $E^\epsilon \doteq \limsup A^\epsilon_{n_k}$ *and* $E^{-\epsilon} \doteq \limsup A^{-\epsilon}_{n_k}$. *Furthermore,* $E = A \cup S_1 - S_0$, *where the sets* S_0, S_1 *satisfy* $S_0^\epsilon - (A^C)^\epsilon \dot\subset \{\eta = 0\}$ *and* $S_1^\epsilon - A^\epsilon \dot\subset \{\eta = 1\}$.

In Section 4, we claimed that in typical situations, if a set A does not satisfy the pseudo-certifiable robustness property, then A is not optimal for R^ϵ. We now further elaborate on this claim. Item c) suggests that in typical cases, $E = A$. If a subset S of $\{\eta = 1\}$ is in E^C, then the adversarial loss pays the maximum penalty of 1 on that set, while if $S \subset E^{-\epsilon}$, then the adversarial loss incurs zero penalty on S. Furthermore, one can show that $\{\eta = 1\}^{-\epsilon}$ is contained in every adversarial Bayes classifier. These heuristics suggest that typically $S_1 = \emptyset$ and similarly $S_0 = \emptyset$ and thus $F(E), F(E^C)$ would be empty as well. Hence, in typical situations, if for a classifier D either $F(D)$ or $F(D^C)$ are non-empty, then D is not optimal for R^ϵ. In fact, in Lemmas 24 and 27 of Appendix E, show that $R^\epsilon(D - F(D)) \leq R^\epsilon(D)$ and $R^\epsilon(D \cup F(D^C)) \leq R^\epsilon(D)$. As discussed in Section 4, adding or subtracting $F(D), F(D^C)$ from S is intuitively a 'local' change to the classifier. One could try and devise algorithms to take advantage of this property.

G.1 A Motivating Example–Applying Theorem 11

To show the utility of Theorem 11, 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 discrete spaces are metric spaces with the discrete metric. Moreover, 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

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

Recall that in a discrete space, every set is measurable. We will show that for every sequence of decreasing sets C_n, (51), (53), and (60) all hold. This will imply that Theorem 11 applies and that the adversarial Bayes classifier exists.

To start, note that for this definition of the $^\epsilon$ operation, we still have that for every sequence of sets A_n,

$$\bigcup_{n=1}^\infty A_n^\epsilon = \left(\bigcup_{n=1}^\infty A_n \right)^\epsilon.$$

We next show (60). To start, one can show that

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

This inclusion is not so hard, and in fact the proof follows the same steps as of (24) of Lemma 15, so we do not reproduce it here.

We now prove the opposite inclusion (60). Pick $y \in \bigcap_{n\in\mathbb{N}} C_n^{\epsilon}$. Then for each n, we have $y = b_n(c_n)$ for some $b_n \in B$ and $c_n \in C_n$. As B is finite, it is compact so that $\{b_n\}$ has a limit point b. Because we are in a discrete space, $b_n = b$ infinitely many times. Therefore, $b(c_n) = y$ for infinitely many n. Because each b is bijective, $c_n = b^{-1}(y)$. In other words, for infinitely many n, $b^{-1}(y) \in C_n$. Because C_n is decreasing, this implies that $b^{-1}(y) \in C_n$ *for all* n. Therefore

$$b^{-1}(y) \in \bigcap_{n=1}^{\infty} C_n \Rightarrow y \in \left(\bigcap_{n=1}^{\infty} C_n \right)^{\epsilon}.$$

G.2 Proving Theorem 11

We now discuss the proof of Theorem 11. As mentioned in Section 6, the big picture motivation is that Theorem 1 followed directly from Prohkorov's Theorem together with 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 11 is exactly the same as the proof of Theorem 1, except that we replace Lemma 2 with the assumption that there exists a sequence C_n satisfying (60). As $\bigcap C_n^{\epsilon} \doteq (\bigcap C_n)^{\epsilon}$ follows from (53) and (60), and $\bigcap C_n^{-\epsilon} = (\bigcap C_n)^{-\epsilon}$ follows from (51), the conclusions of Lemma 2 are a consequence of (51), (53), and (60). Because the proof of Theorem 11 is identical to that of Theorem 1 when using the appropriate definitions, we describe these new concepts but do not repeat the proof here.

Following the proof of Theorem 1, we begin with reviewing Prokhorov's theorem and the definition of inner regularity. Recall the analog of Corollary 1 presented in Appendix B:

Corollary 2. *Let $(\mathbb{P}_n, \mathbb{R}^d, \mathscr{U}(X))$ be a sequence of probability measure spaces, where each $\mathbb{P}_n$ is the completion of a Borel measure restricted to $\mathscr{U}(X)$. Then $\{\mathbb{P}_n\}$ admits a weakly convergent subsequence iff the sequence is tight.*

Again, to show that a sequence of measures is tight, we use inner regularity. We now review the definition of inner regularity for a general measure space.

Definition 7. *Let τ be a topology on a set X and Σ a σ-algebra on X. $\mathbb{P}$ be a measure on (X, Σ).Then $\mathbb{P}$ is* inner regular *if for all measurable sets E, $\mathbb{P}(E) = \sup\{\mathbb{P}(K)\colon K \subset E, K$ compact and measurable$\}$.*

Next we review a generalized version of Lemma 4, which we use to show that a sequence of measures $\mathbb{P}_n$ is tight. We first encountered this result in Appendix C.2.

Theorem 9. *Let X be a second-countable and locally compact Haudsorff space. Then every Borel measure ν with $\nu(X) < \infty$ is inner regular.*

Recall that on a metric space, second-countable is equivalent to separable. Again, this Lemma implies that on a separable metric space, the completion of every Borel measure restricted to the universal σ-algebra is inner regular.

As in the proof of Theorem 1, given a minimizing sequence A_n, one can define two sequences of measures $\{\mathbb{P}_n^i\}$ by $\mathbb{P}_n^1(E) = \mathbb{P}(A_n^{\epsilon} \cap E)$ and $\mathbb{P}_n^2(E) = \mathbb{P}(A_n^{-\epsilon} \cap E)$, and then find a weakly convergent subsequence using Corollary 2 and Theorem 9.

We now discuss the analogs of Lemmas 1, 2, and 3 and how they follow from the hypotheses of Theorem 11. First, the analog of Lemma 2 is just assumed. Next, Lemma 3 follows from the properties (51) and (53). Lastly, a generalization of Lemma 1 can be proven in an arbitrary separable metric spaces.

Following the proof of Theorem 1, we start by stating the analog of Lemma 1.

Lemma 10. *Let X be a separable metric space. Assume that $\mathbb{P}_n$ weakly converges to $\mathbb{Q}$ with $\mathbb{P}_n$ given by $\mathbb{P}_n(B) = \mathbb{P}(B \cap A_n)$, for a sequence of sets A_n. Then, for some subsequence A_{n_j} of A_n, $\mathbb{Q}(B) = \mathbb{P}(A \cap B)$ for a set A that is given by*

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

and further $\mathbb{1}_{A_{n_j}} \to \mathbb{1}_A$ $\mathbb{P}$-a.e.

We prove this result in Appendix C. That the metric space X is separable is the main assumption of this lemma. Metric spaces in applications are typically separable, so this requirement should not be limiting.

Again, just as in proof of Theorem 1, one uses Lemma 10 to prove that there is a subsequence $\{A_{n_k}\}$ of $\{A_n\}$ for which $\limsup A^\epsilon_{n_k} \doteq \liminf A^\epsilon_{n_k}$ and $\limsup A^{-\epsilon}_{n_k} \doteq \liminf A^{-\epsilon}_{n_k}$.

Lastly, we present a generalization of Lemma 3, which we prove in Appendix F.

Lemma 37. *Let $^\epsilon$ be a set operation for which A^ϵ is universally measurable if A is universally measurable. Further assume that $^\epsilon$ satisfies properties (51) and (53).*

Let A_n be a minimizing sequence of R^ϵ for which $\liminf A^\epsilon_n \doteq \limsup A^\epsilon_n$ and $\liminf A^{-\epsilon}_n = \limsup A^{-\epsilon}_n$. Then there is a decreasing minimizing sequence B_n (in other words $B_{n+1} \subset B_n$) for which $\liminf B^\epsilon_n \doteq \limsup B^\epsilon_n = \limsup A^\epsilon_n$ and $\limsup B^{-\epsilon}_n \doteq \liminf B^{-\epsilon}_n \dot\supset \liminf A^{-\epsilon}_n$.

The fact that $\liminf B^\epsilon_n \doteq \limsup B^\epsilon_n = \limsup A^\epsilon_n$ and $\limsup B^{-\epsilon}_n \doteq \liminf B^{-\epsilon}_n \dot\supset \liminf A^{-\epsilon}_n$ is used in the proof of Theorem 12 but not Theorems 1 or 11.

Again, this Lemma is used to produce a decreasing minimizing sequence B_n for which $\limsup B^\epsilon_{n_k} \doteq \liminf B^\epsilon_{n_k}$. Lastly, as discussed in the introduction to this appendix, the assumptions of Theorem 11 imply that given decreasing minimizing sequence B_n for which $\limsup B^\epsilon_{n_k} \doteq \liminf B^\epsilon_{n_k}$, one could find a minimizing sequence $\{C_n\}$ for which (62) and (61) hold. Just as in the proof of Theorem 1, this fact implies the existence of the adversarial Bayes classifier.

G.3 Proving Theorem 12

Before proving Theorem 12, we review three essential Lemmas.

The following Lemma gives some more information about how the sets B_n and C_n of Lemma 2 relate.

Lemma 39. *Let $B \subset \mathbb{R}^d$. Then if $C = ((B^{-\epsilon})^{2\epsilon})^{-\epsilon}$, then $C^\epsilon \subset B^\epsilon$ and $C^{-\epsilon} \supset B^{-\epsilon}$.*

Proof. Set $D = (B^{-\epsilon})^\epsilon$. Then Lemma 26 implies that $C^{-\epsilon} \supset D^{-\epsilon} \supset B^{-\epsilon}$ and $C^\epsilon \subset D^\epsilon \subset B^\epsilon$. □

In Section 4.3 we argued that a necessary condition for a set B to be equal to A^ϵ for some A is that $F(B) = 0$. The next Lemma gives conditions for when $\limsup A^\epsilon_n \doteq B$, where $F(B) = 0$.

Lemma 40. *Let $\mathbb{P}$ be absolutely continuous with respect to Lebesgue measure and let A_n be a sequence of universally measurable sets. Then $\overline{\limsup A^\epsilon_n} \doteq \limsup A^\epsilon_n$ and $F(\overline{\limsup A^\epsilon_n}) = \emptyset$.*

This lemma is a consequence of a concept in geometric measure theory called *porosity*, see for instance (Zajíček, 1987). We include a proof of Lemma 40 for completeness in the next subsection.

The last lemma implies that every point $\mathbf{x}$ in $\bigcap_{n=1}^\infty A^\epsilon_n$ is contained in a ball $B_\epsilon(\mathbf{x})$ which is completely contained in $\bigcap_{n=1}^\infty A^\epsilon_n$. This statement is very similar but not the same as $F(\bigcap_{n=1}^\infty A^\epsilon_n) = \emptyset$.

Lemma 41. *Let $\{S_n\}$ be a decreasing sequence of sets with $F(S_n) = 0$. Then if $\mathbf{x} \in \bigcap_{n=1}^\infty S_n$, then there exists a $\mathbf{b}$ with $\mathbf{x} \in \overline{B_\epsilon(\mathbf{b})}$ and $B_\epsilon(\mathbf{b}) \subset \bigcap_{n=1}^\infty S_n$.*

This lemma follows from Lemma 31, and this idea was discussed in Appendix E.1. We provide a formal proof here for completeness.

Proof. Pick $\mathbf{x} \in \bigcap_{n=1}^\infty S_n$. Then $\mathbf{x} \in S_n$ for each n. This means we can pick a $\mathbf{b}_n$ with $\mathbf{x} \in \overline{B_{2\epsilon}(\mathbf{b}_n)} \subset S_n$. Note that the sequence $\{\mathbf{b}_n\}$ is contained in the compact set $\overline{B_{2\epsilon}(\mathbf{x})}$. Thus $\{\mathbf{b}_n\}$

has a limit point $\mathbf{b}$, and $\mathbf{x} \in \overline{B_{2\epsilon}(\mathbf{b})}$. Now Lemma 41 implies that if $\mathbf{y} \in B_{2\epsilon}(\mathbf{b})$, then $\mathbf{y} \in B_{2\epsilon}(\mathbf{b}_n)$ for sufficiently large n. □

Proof of Theorem 12. Most of this proof follows the same outline as the proof of Theorem 1 so we will only give an outline here. To start, given a minimizing sequence A_n of R^ϵ, we apply Prokhorov's theorem and Lemma 10 to find a subsequence A_{n_k} for which $\limsup A_{n_k}^\epsilon \doteq \liminf A_{n_k}^\epsilon$ and $\limsup A_{n_k}^{-\epsilon} \doteq \liminf A_{n_k}^{-\epsilon}$. This proves a)).

Next, Lemma 37 implies that there is a decreasing minimizing sequence B_n with $\liminf B_n^\epsilon \doteq \limsup B_n^\epsilon \doteq \limsup A_n^\epsilon \doteq \liminf A_n^\epsilon$ and $\liminf B_n^{-\epsilon} \doteq \limsup B_n^{-\epsilon} \dot\supset \limsup A_n^{-\epsilon} \doteq \liminf A_n^{-\epsilon}$. We then Lemma 33 to find a well-behaved minimizing sequence. Namely, we pick $C_n = ((B_n^{-\epsilon})^{2\epsilon})^{-\epsilon}$. Again, we define the minimizer A by $A = \bigcap_{n=1}^\infty C_n$. Then Lemma 33 states that A is in fact a minimizer, and that

$$A^\epsilon = \bigcap_{n=1}^\infty C_n^\epsilon, A^{-\epsilon} = \bigcap_{n=1}^\infty C_n^{-\epsilon}$$

and further A is pseudo-certifiably robust.

Note that by Lemma 37, $\limsup A_n^\epsilon \doteq \limsup B_n^\epsilon$ and $\limsup B_n^{-\epsilon} \dot\supset \liminf A_n^{-\epsilon}$. According to Lemma 39, $C_n^\epsilon \subset B_n^\epsilon$ and $C_n^{-\epsilon} \supset B_n^{-\epsilon}$ and thus $\limsup C_n^\epsilon \subset \limsup A_n^\epsilon$, $\limsup C_n^{-\epsilon} \supset \limsup A_n^{-\epsilon}$. As both A_n and C_n are minimizing sequences,

$$\int (1-\eta(\mathbf{x}))\mathbb{1}_{\limsup A_n^\epsilon} + \eta(\mathbf{x})\mathbb{1}_{(\limsup(A_n^{-\epsilon}))^C} d\mathbb{P} = \int (1-\eta(\mathbf{x}))\mathbb{1}_{\limsup C_n^\epsilon} + \eta(\mathbf{x})\mathbb{1}_{(\limsup(C_n^{-\epsilon}))^C} d\mathbb{P}$$

which implies that

$$\int (1-\eta(\mathbf{x}))\mathbb{1}_{\limsup A_n^\epsilon - \limsup C_n^\epsilon} + \eta(\mathbf{x})\mathbb{1}_{(\limsup(A_n^{-\epsilon}))^C - (\limsup C_n^{-\epsilon}))^C} d\mathbb{P} = 0$$

As $0 \leq \eta(\mathbf{x}) \leq 1$, the integrand is non-negative everywhere, which implies that it must be zero a.e. Therefore, outside the set $\{\eta = 1\}$ $\limsup A_n^\epsilon$ must match with $\limsup C_n^\epsilon$ and outside the set $\{\eta = 0\}$, $\limsup A_n^{-\epsilon}$ must match with $\limsup C_n^{-\epsilon}$. This proves b).

It remains to show c). If in fact $\eta \in (0,1)$ a.e., then b) implies that $A^\epsilon \doteq \limsup A_{n_k}^\epsilon$ and $A^{-\epsilon} = \liminf A_{n_k}^{-\epsilon}$, a.e., so then c) holds trivially.

We now consider the case for which $\mathbb{P}$ is absolutely continuous with respect to Lebesgue measure μ. By Lemma 40, there are universally measurable sets T_1, T_2 for which $\limsup A_n^\epsilon \doteq T_1$, $\limsup(A_n^C)^\epsilon \doteq T_2$ and $F(T_1) = F(T_2) = \emptyset$.

Earlier we showed that

$$\limsup A_n^\epsilon - \limsup C_n^\epsilon \dot\subset \{\eta = 1\}, \limsup(A_n^C)^\epsilon - \limsup(C_n^C)^\epsilon \dot\subset \{\eta = 0\}.$$

Because $\limsup A_n^\epsilon, \limsup(A_n^C)^\epsilon$ equal T_1, T_2 and $\limsup C_n^\epsilon, \limsup(C_n^C)^\epsilon$ equal $A^\epsilon, (A^C)^\epsilon$ $\mathbb{P}$-a.e., we can in fact write

$$T_1 - A^\epsilon \dot\subset \{\eta = 1\}, T_2 - (A^C)^\epsilon \dot\subset \{\eta = 0\}.$$

Now I claim that we can in fact choose

$$S_1 = T_1^{-\epsilon} - A$$

and

$$S_0 = T_2^{-\epsilon} - A^C.$$

We will show the argument for S_1, the reasoning for S_0 is analogous. First we will show that $S_1^\epsilon - A^\epsilon \dot\subset \{\eta = 1\}$:

$$S_1^\epsilon - A^\epsilon = (T_1^{-\epsilon} \cap A^C)^\epsilon - A^\epsilon \subset (T_1^{-\epsilon})^\epsilon - A^\epsilon = T_1 - A^\epsilon \dot\subset \{\eta = 1\}$$

where the second-to-last containment follows from $F(T_1) = \emptyset$. Next we will show that $(A \cup S_1)^\epsilon \doteq T_1$. To start, $S_1^\epsilon = (T_1^{-\epsilon} \cap A^C)^\epsilon \subset (T_1^{-\epsilon})^\epsilon = T_1$, and $A^\epsilon \dot\subset T_1$ by construction, so $A^\epsilon \cup S_1^\epsilon \dot\subset T_1$. Next, note that $S_1 = T_1^{-\epsilon} \cap A^C$ is equivalent to $T_1^{-\epsilon} = S_1 \cup (T_1^{-\epsilon} \cap A)$. Thus $T_1^{-\epsilon} \subset S_1 \cup A$, and $T_1 = (T_1^{-\epsilon})^\epsilon \subset (S_1 \cup A)^\epsilon$.

□

G.4 Proof of Lemma 40

To prove Lemma 40, we first present another lemma, which we will prove at the end of the section.

Lemma 42. *Let μ be Lebesgue measure and let $S \subset \mathbb{R}^d$. If for each $s \in \partial S$ there exists an open convex C with $C \subset S$ and $s \in \partial C$, then $\mu(\partial S) = 0$.*

This lemma allows one to argue that in fact $\limsup A_n^\epsilon \doteq \overline{\limsup A_n^\epsilon}$.

Lemma 40. *Let $\mathbb{P}$ be absolutely continuous with respect to Lebesgue measure and let A_n be a sequence of universally measurable sets. Then $\overline{\limsup A_n^\epsilon} \doteq \limsup A_n^\epsilon$ and $F(\overline{\limsup A_n^\epsilon}) = \emptyset$.*

Proof of Lemma 40. By setting $D_n = \bigcup_{k \geq n} A_k$, one can write

$$\begin{aligned}\limsup A_n^\epsilon &= \bigcap_{n \geq 1} \bigcup_{k \geq n} A_n^\epsilon \\ &= \bigcap_{n \geq 1} \left(\bigcup_{k \geq n} A_n \right)^\epsilon \qquad\qquad \text{(Equation 6)} \\ &= \bigcap_{n \geq 1} D_n^\epsilon\end{aligned}$$

Now as $F(D_n^\epsilon) = 0$, Lemma 41 implies that for every $\mathbf{x} \in \limsup A_n^\epsilon$, there is a ball $B_\epsilon(\mathbf{b})$ with $\mathbf{x} \in \overline{B_\epsilon(\mathbf{b})}$ and $B_\epsilon(\mathbf{b}) \subset \bigcap_{n \geq 1} D_n^\epsilon = \limsup A_n^\epsilon$.

Next, Lemma 42 implies that $\partial \limsup A_n^\epsilon$ has Lebesgue measure zero. Thus $\overline{\limsup A_n^\epsilon} = \limsup A_n^\epsilon \cup \partial \limsup A_n^\epsilon \doteq \limsup A_n^\epsilon$. Furthermore, $\overline{\limsup A_n^\epsilon}$ is a closed set and therefore universally measurable.

□

To prove Lemma 42 we take an approach that is standard in geometric measure theory. The strategy is to apply the Lebesgue differentiation theorem. See Folland (1999) for a proof.

Theorem 13 (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.:*

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

We use this theorem in the proof of Lemma 42.

Lemma 42. *Let μ be Lebesgue measure and let $S \subset \mathbb{R}^d$. If for each $s \in \partial S$ there exists an open convex C with $C \subset S$ and $s \in \partial C$, then $\mu(\partial S) = 0$.*

Proof of Lemma 42. We will apply the Lebesgue differentiation theorem to the function $\mathbb{1}_{\partial S}$.

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

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

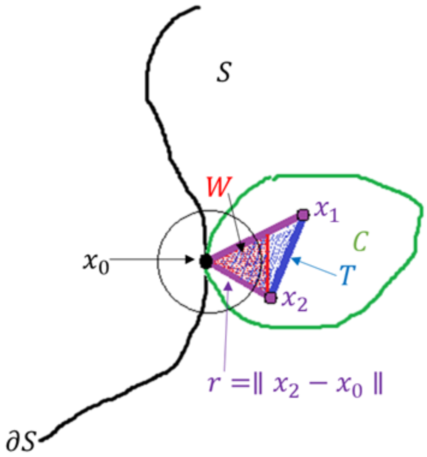

Figure 6: The convex set C, the simplex T, and the simplex W in two dimensions. The illustrated ball has radius less than r. In this Figure, $r = \|\mathbf{x}_2 - \mathbf{x}_0\|$.

This inequality will imply the result as

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

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, $B_1(\mathbf{x}_0) \cap C$ is non-empty, open, and convex. Thus we can pick d points $\mathbf{x}_1 \dots \mathbf{x}_d \in C$ for which the vectors $\{\mathbf{x}_1 - \mathbf{x}_0 \dots \mathbf{x}_d - \mathbf{x}_0\}$ are linearly independent. By the convexity of $C \subset \operatorname{int} S$, the simplex given by the open convex hull of $\{\mathbf{x}_0 \dots \mathbf{x}_d\}$ is contained in $\operatorname{int} S$. We will call this convex hull T. By construction $T \cap B_r(\mathbf{x}_0)$ 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 G.4 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 $\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$, we have shown that for $r < \min_{i\in[1,n]} \|\mathbf{x}_i - \mathbf{x}_0\|$,

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

□