European Journal of Applied Mathematics (2026), **37**, pp. 43–72
doi:10.1017/S0956792525100223

CAMBRIDGE UNIVERSITY PRESS

PAPER

Uniform convergence of adversarially robust classifiers

Rachel Morris and Ryan Murray

North Carolina State University, Raleigh, NC, USA
Corresponding author: Rachel Morris; Email: rachel.morris@mail.concordia.ca

Received: 21 June 2024; **Revised:** 08 October 2025; **Accepted:** 08 October 2025

Keywords: adversarial training; geometric measure theory; Hausdorff convergence; robust classification

2020 Mathematics Subject Classification: 28A75 (Primary), 68Q32,62G35, 35B25 (Secondary)

Abstract

In recent years, there has been significant interest in the effect of different types of adversarial perturbations in data classification problems. Many of these models incorporate the adversarial power, which is an important parameter with an associated trade-off between accuracy and robustness. This work considers a general framework for adversarially perturbed classification problems, in a large data or population-level limit. In such a regime, we demonstrate that as adversarial strength goes to zero that optimal classifiers converge to the Bayes classifier in the Hausdorff distance. This significantly strengthens previous results, which generally focus on L^1-type convergence. The main argument relies upon direct geometric comparisons and is inspired by techniques from geometric measure theory.

1. Introduction

In recent years, neural networks have achieved remarkable success in a wide range of classification and learning tasks. However, it is now well-known that these networks do not learn in the same ways as humans and will fail in specific settings. In particular, a wide range of recent work has shown that they fail to be robust to specially designed adversarial attacks [11, 12, 18, 23, 27].

One general approach for mitigating this problem is to include an adversary in the training process. A simple mathematical formulation of this method for 0–1 loss [19], in the large-data or population limit, is to consider the optimization problem

$$\min_A J_\varepsilon(A), \qquad J_\varepsilon(A) := \mathbb{E}\left[\max_{\tilde{x}\in B(x,\varepsilon)} |\mathbb{1}_A(\tilde{x}) - y|\right],$$

where the y variables represent observed classification labels and x variables represent features (we give a more precise description of our setting in the next section). This can be seen as a robust optimization problem, where an adversary is allowed to modify the inputs to our classifier up to some distance ε. When $\varepsilon = 0$, this corresponds to the standard Bayes risk.

Recent work has significantly expanded our mathematical understanding of this problem. Our work directly builds upon [5], which rewrites the previous functional as

$$J_\varepsilon(A) = \mathbb{E}[|\mathbb{1}_A(x) - y|] + \varepsilon \mathrm{Per}_\varepsilon(A),$$

where Per_ε is a special data-adapted perimeter, whose definition is given in (4). This is related to a growing body of recent work, for example, showing that Per_ε converges to the (weighted) classical perimeter [7], and demonstrating links between the adversarially robust training problem and mean curvature flow [6, 16]. This literature seeks to provide a more complete description of the effect of ε on adversarially robust classifiers in a geometric sense. This relates to the study of nonlocal perimeter minimization and flows [8, 9, 10], where the *unweighted* ε-perimeter is considered. As training these robust classifiers is generally a challenging task, one overarching goal of this type of work is to provide

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

a more precise understanding of the effect of ε, practical means for approximating that effect and the impact on classifier complexity: each of these has the potential to improve more efficient solvers for these problems.

Various modifications of the robust classification energy J_ε have been proposed. For example, some authors relax either the criteria for an adversarial attack or the loss function to interpolate between the accurate yet brittle Bayes classifier and the robust yet costly minimizers of the adversarial training problem [4, 17, 24, 25]. Still others employ optimal transport techniques to study distributionally robust optimization, where instead of perturbing data points, the adversary perturbs the underlying data distribution [13, 14, 15, 21, 22].

The main goal of this paper is to study the convergence of solutions to the adversarially robust classification problem towards the original Bayes classification task for data-perturbing models. We build a framework that allows us to consider a wide range of adversarial settings at the same time. In doing so, we obtain Hausdorff convergence results, which are generally much stronger than the L^1-type results previously obtained [4]. These results parallel many of the basic results in the study of variational problems involving perimeters, wherein one first proves stability in L^∞ spaces, and then subsequently proves stronger regularity results for minimizers. In a similar way, we see our results as a building block towards stronger regularity results for the adversarially robust classification problem, which have received significant attention in the literature. We begin by concretely describing the setup of our problem and then giving an informal statement of our results along with some discussion.

1.1. Setup

Let the Euclidean space $\mathbb{R}^d$ equipped with the metric $\mathrm{d}(\cdot,\cdot)$ represent the space of features for a data point, and let $\mathcal{B}(\mathbb{R}^d)$ be the set of all Borel measurable subsets of $\mathbb{R}^d$. We will let $\mathcal{L}^d$ be the d-dimensional Lebesgue measure. We are considering a supervised binary classification setting, in which training pairs (x, y) are distributed according to a probability measure μ over $\mathbb{R}^d \times \{0, 1\}$. Here y represents the class associated with a given data point, and the fact that $y \in \{0, 1\}$ corresponds to the binary classification setting. Let ρ denote the $\mathbb{R}^d$ marginal of μ, namely $\rho(A) = \mu(A \times \{0, 1\})$. We decompose $\rho \in \mathcal{P}(\mathbb{R}^d)$ into $\rho = w_0\rho_0 + w_1\rho_1$ where $w_i = \mu(\mathbb{R}^d \times \{i\})$, and the conditional probability measure $\rho_i \in \mathcal{P}(\mathbb{R}^d)$ for a set $A \in \mathcal{B}(\mathbb{R}^d)$ is

$$\rho_i(A) = \frac{\mu(A \times \{i\})}{w_i}$$

for $i = 0, 1$. All of these measures are assumed to be Radon measures.

In binary classification, we associate a set $A \in \mathcal{B}(\mathbb{R}^d)$ with a classifier, meaning that $x \in \mathbb{R}^d$ is assigned label 1 when $x \in A$ and x is assigned the label 0 when $x \in A^c$. Unless otherwise stated, we will assume all classifiers $A \in \mathcal{B}(\mathbb{R}^d)$. The Bayes classification problem for the 0–1 loss function is given by

$$\inf_{A \in \mathcal{B}(\mathbb{R}^d)} \mathbb{E}_{(x,y)\sim\mu}[|\mathbb{1}_A(x) - y|]. \tag{1}$$

In this work, we will only consider the 0–1 loss function, which allows us to restrict our attention to indicator functions for minimizers of (1). We refer to minimizes of the Bayes risk as *Bayes classifiers*.

Remark 1.1 (Uniqueness of Bayes Classifiers). *If we assume that ρ has a density everywhere on $\mathbb{R}^d$ and identify the measures ρ_i with the density at $x \in \mathbb{R}^d$ given by $\rho_i(x)$, then we can describe the uniqueness, or lack thereof, of Bayes classifiers in terms of those densities. Specifically, Bayes classifiers are unique up to the set $\{w_1\rho_1 = w_0\rho_0\}$, which may be a set of positive measure depending on μ. We define maximal and minimal Bayes classifiers (in the sense of set inclusion) by*

$$A_0^{\max} := \{x \in \mathbb{R}^d : w_1\rho_1(x) \geq w_0\rho_0(x)\}, \qquad A_0^{\min} := \{x \in \mathbb{R}^d : w_1\rho_1(x) > w_0\rho_0(x)\}. \tag{2}$$

When $w_0\rho_0 - w_1\rho_1 \in C^1$ and $|w_0\nabla\rho_0 - w_1\nabla\rho_1| > \alpha > 0$ on the set $\{w_0\rho_0 = w_1\rho_1\}$, the Bayes classifier is unique up to sets of ρ measure zero. In the case where ρ is supported everywhere, Bayes classifiers

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

are unique up to sets of $\mathcal{L}^d$ measure zero. Whenever we refer to the Bayes classifier as unique, we mean unique in this measure-theoretic sense. Later on in Assumption 3.11, we will refer to such uniqueness as the 'non-degeneracy' of the Bayes classifier and represent the unique classifier by A_0.

Throughout this paper, we will consider and seek to unify two optimization problems from the literature that aim to train robust classifiers. First, we consider the *adversarial training problem*, which trains classifiers to mitigate the effect of worst-case perturbations [19]. The adversarial training problem is

$$\inf_{A\in\mathcal{B}(\mathbb{R}^d)} \mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B_{\mathrm{d}}(x,\varepsilon)} |\mathbb{1}_A(\tilde{x}) - y|\right],$$

where $B_{\mathrm{d}}(x,\varepsilon)$ is the *open* metric ball of radius $\varepsilon > 0$. The existence of solutions in this setting was previously established [5]. The parameter ε is called the *adversarial budget*, and it represents the strength of the adversary. By using the open ball, we are following the conventions set in the previous work on convergence of optimal adversarial classifiers [7]. Other works have utilized the closed ball due to consistency with the standard classification problem when $\varepsilon = 0$, but that comes at the price of added measurability concerns: see Remark 1.2 for more details.

An equivalent form of this variational problem (see [5]), wherein the problem is rewritten using a nonlocal perimeter, is

$$\inf_{A\in\mathcal{B}(\mathbb{R}^d)} \mathbb{E}_{(x,y)\sim\mu}[|\mathbb{1}_A(x) - y|] + \varepsilon \mathrm{Per}_\varepsilon(A) \tag{3}$$

with the *ε-perimeter* defined by

$$\mathrm{Per}_\varepsilon(A) := \frac{w_0}{\varepsilon}\int_{\mathbb{R}^d}\left[\sup_{\tilde{x}\in B_{\mathrm{d}}(x,\varepsilon)} \mathbb{1}_A(\tilde{x}) - \mathbb{1}_A(x)\right] d\rho_0(x) + \frac{w_1}{\varepsilon}\int_{\mathbb{R}^d}\left[\mathbb{1}_A(x) - \inf_{\tilde{x}\in B_{\mathrm{d}}(x,\varepsilon)} \mathbb{1}_A(\tilde{x})\right] d\rho_1(x).$$

This normalization with ε in the denominator is chosen so that we recover the (weighted) classical perimeter as $\varepsilon \to 0^+$. In this sense, we consider the nonlocal ε-perimeter a data-adapted approximation of the classical perimeter. From the variational problem given by (3), we define the *adversarial classification risk* for a classifier $A \in \mathcal{B}(\mathbb{R}^d)$ as

$$J_\varepsilon(A) := \mathbb{E}_{(x,y)\sim\mu}[|\mathbb{1}_A(x) - y|] + \varepsilon \mathrm{Per}_\varepsilon(A).$$

When considering the ε-perimeter, the region affected by adversarial perturbations must be within distance ε of the decision boundary of the classifier. As such, it will be helpful to be able to discuss sets that either include or exclude the ε-perimeter region. From mathematical morphology [26], for a set $A \in \mathbb{R}^d$ and $\varepsilon > 0$, we define the

- *ε-dilation* of A as $A^\varepsilon := \{x \in \mathbb{R}^d : \mathrm{d}(x, A) < \varepsilon\}$,
- *ε-erosion* of A as $A^{-\varepsilon} := \{x \in \mathbb{R}^d : \mathrm{d}(x, A^{\mathrm{c}}) \geq \varepsilon\}$.

Using this notation, one can equivalently express the ε-perimeter as

$$\mathrm{Per}_\varepsilon(A) = \frac{w_0}{\varepsilon}\rho_0(A^\varepsilon \setminus A) + \frac{w_1}{\varepsilon}\rho_1(A \setminus A^{-\varepsilon}). \tag{4}$$

Inspired by the notation in geometric measure theory, we also define the *relative ε-perimeter* for a classifier $A \in \mathcal{B}(\mathbb{R}^d)$ with respect to a set $E \in \mathcal{B}(\mathbb{R}^d)$ by

$$\mathrm{Per}_\varepsilon(A; E) := \frac{w_0}{\varepsilon}\rho_0((A^\varepsilon \setminus A) \cap E) + \frac{w_1}{\varepsilon}\rho_1((A \setminus A^{-\varepsilon}) \cap E).$$

Remark 1.2 (Previous work for the adversarial training problem (3)). *The worst-case adversarial training model was initially proposed for general loss functions by [19]. When the loss function is specified to be the 0–1 loss function, previous work has established the existence and considered the equivalence of minimizers to (3) for the open and closed ball models [2, 5, 14, 22]. Although the open and closed*

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

ball models are similar, there are some subtle differences that must be considered. While measurability of $\sup_{\tilde{x}\in B_{\mathrm{d}}(x,\varepsilon)} \mathbb{1}_A(\tilde{x})$ *for a Borel set A in the open ball model is trivial, the same cannot be said for the closed ball model; to address these measurability concerns in the closed ball model, one must employ the universal* σ*-algebra instead of the Borel* σ*-algebra. We emphasize that we choose to study the open ball model as this simplifies the analysis and measurability concerns associated with the closed ball model, and the open ball model was used for prior convergence results [7].*

Some papers consider a surrogate adversarial risk which is more computationally tractable [1, 3, 13, 20]; others explore necessary conditions and geometric properties of minimizers [6, 7, 16]. Of particular note to the present work is the study of the limit of minimizers of J_ε*. Theorem 2.5 states [7].*

Theorem (Conditional convergence of adversarial training). *Under the conditions of Theorems 2.1 and 2.3 from [7] and assuming the source condition, any sequence of solutions to*

$$\inf_{A\in\mathcal{B}(\Omega)} \mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B(x,\varepsilon)\cap\Omega} |\mathbb{1}_A(\tilde{x}) - y|\right]$$

possesses a subsequence converging to a minimizer of

$$\min\{\mathrm{Per}(A;\rho) : A \in \arg\min_{B\in\mathcal{B}(\Omega)} \mathbb{E}_{(x,y)\sim\mu}[|\mathbb{1}_B(x) - y|]\}.$$

The convergence is proven in the $L^1(\Omega)$ *topology for some open, bounded Lipschitz domain* $\Omega \subset \mathbb{R}^d$*. Here,* $\mathrm{Per}(\,\cdot\,;\rho)$ *is a weighted version of the classical perimeter. The source condition mentioned provides minor regularity assumptions on the Bayes classifier. Note that in the referenced theorem, there are additional assumptions on the underlying data distribution* ρ*. In our work, we strengthen this convergence result by proving Hausdorff convergence of minimizers of (3) to the Bayes classifier with similar assumptions on* ρ*.*

The second optimization problem, which serves as an important model case, interpolates between the accuracy on clean data of the Bayes classifier and the robustness of the adversarial training problem minimizers. The *probabilistic adversarial training problem* for $p \in [0, 1)$ and probability measures $\mathfrak{p}_x \in \mathcal{P}(\mathbb{R}^d)$ for each $x \in \mathbb{R}^d$ is

$$\inf_{A\in\mathcal{B}(\mathbb{R}^d)} \mathbb{E}_{(x,y)\sim\mu}[|\mathbb{1}_A(x) - y|] + \mathrm{ProbPer}_p(A), \tag{5}$$

with the *probabilistic perimeter* defined by

$$\mathrm{ProbPer}_p(A) := w_0\rho_0(\Lambda_p^0(A)) + w_1\rho_1(\Lambda_p^1(A)), \tag{6}$$

and the set functions Λ_p^i for $i = 0, 1$ defined by

$$\Lambda_p^0(A) := \{x \in A^{\mathsf{c}} : \mathbb{P}(x' \in A : x' \sim \mathfrak{p}_x) > p\},$$

$$\Lambda_p^1(A) := \{x \in A : \mathbb{P}(x' \in A^{\mathsf{c}} : x' \sim \mathfrak{p}_x) > p\}.$$

Here, $\mathbb{P}(x' \in A : x' \sim \mathfrak{p}_x)$ is the probability that a point x' sampled from the probability distribution $\mathfrak{p}_x$ belongs to the set A. We notice that (6) takes the same form as (4) where we replace the metric boundary fattening by a probabilistic fattening. We define the *probabilistic adversarial classification risk* for a classifier $A \in \mathcal{B}(\mathbb{R}^d)$ as

$$J_p(A) := \mathbb{E}_{(x,y)\sim\mu}[|\mathbb{1}_A(x) - y|] + \mathrm{ProbPer}_p(A).$$

The *relative probabilistic perimeter* for a classifier $A \in \mathcal{B}(\mathbb{R}^d)$ with respect to a set $E \in \mathcal{B}(\mathbb{R}^d)$ is given by

$$\mathrm{ProbPer}_p(A; E) := w_0\rho_0\left(\Lambda_p^0(A) \cap E\right) + w_1\rho_1\left(\Lambda_p^1(A) \cap E\right).$$

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

To make the connection with the ε-perimeter more concrete, we will restrict our attention to certain families of probability measures that scale appropriately with ε for the remainder of this work.

Assumption 1.3. *Let* $\xi : \mathbb{R}^d \to [0, \infty)$ *such that* $\xi \in L^1(\mathbb{R}^d)$, $\int_{\mathbb{R}^d} \xi(z)\, dz = 1$, $\xi(z) = 0$ *if* $|z| > 1$, *and* $\xi(z) > c$ *for some constant* $c > 0$ *and for* $|z| \leq 1$. *For* $x, x' \in \mathbb{R}^d$, *we assume that*

$$\mathfrak{p}_{x,\varepsilon}(x') = \varepsilon^{-d} \xi\left(\frac{x' - x}{\varepsilon}\right).$$

□

We will now write $\mathrm{ProbPer}_{\varepsilon,p}$ and refer to it as the probabilistic ε-perimeter to emphasize the dependence on the adversarial budget. Unlike with the Per_ε, we do not normalize $\mathrm{ProbPer}_{\varepsilon,p}$ with respect to ε. We also write $J_{\varepsilon,p}$ instead of J_p and $\Lambda^i_{\varepsilon,p}$ instead of Λ^i_p for $i = 0, 1$. Under Assumption 1.3, $\Lambda^0_{\varepsilon,p}(A)$ and $\Lambda^1_{\varepsilon,p}(A)$ are subsets of the ε-perimeter regions $A^\varepsilon \setminus A$ and $A \setminus A^{-\varepsilon}$, respectively. Specifically, this means that $J_{\varepsilon,p}(A) \leq J_\varepsilon(A)$ for all $A \in \mathcal{B}(\mathbb{R}^d)$ when the underlying data distribution μ is the same. We note that probabilistic ε-perimeter that most closely coincides with the ε-perimeter when $p = 0$ and $\mathfrak{p}_{x,\varepsilon} = \mathrm{Unif}(B_{\mathrm{d}}(x, \varepsilon))$ for each $x \in \mathbb{R}^d$.

Remark 1.4 (Previous work for the probabilistic adversarial training problem (5)). *This form of the problem was proposed by [4] as a revision of probabilistically robust learning [25]. Although* $\mathrm{ProbPer}_p$ *is not a perimeter in the sense that it has not been shown to be submodular and it does not admit a coarea formula, we follow the convention from [4] and refer to* $\mathrm{ProbPer}_p$ *as the probabilistic perimeter. Importantly, existence of minimizers has not been proved for either the original or modified probabilistic adversarial training problem. There have also been no results pertaining to the convergence of minimizers, provided they exist, to the Bayes classifier for either version.*

However, [4] proposes and proves the existence of minimizers for a related probabilistically robust Ψ *risk*

$$J_\Psi(A) := \mathbb{E}_{(x,y)\sim\mu}[|\mathbb{1}_A(x) - y|] + \mathrm{ProbPer}_\Psi(A)$$

for suitable functions $\Psi : [0, 1] \to [0, 1]$ *where the* Ψ*-perimeter takes the form*

$$\mathrm{ProbPer}_\Psi(A) := \int_{A^c} \Psi(\mathbb{P}(x' \in A : x' \sim \mathfrak{p}_x)) d\rho_0(x) + \int_A \Psi(\mathbb{P}(x' \in A^c : x' \sim \mathfrak{p}_x)) d\rho_1(x).$$

However, the convergence results proved in this paper do not currently extend to the Ψ*-perimeter case. The details will be further discussed in Remark 4.17.*

If we juxtapose the variational problem for the adversarial training problem (3) and the probabilistic adversarial training problem (5), both risks are of the form

$$J(A) = \text{ Bayes risk } + \text{ data-adapted perimeter},$$

where the data-adapted perimeters can be expressed as

$$\text{data-adapted perimeter} = w_0 \rho_0(\text{subset of } A^{\mathsf{c}}) + w_1 \rho_1(\text{subset of } A).$$

We seek to develop a unifying framework for various adversarial models, including, but not limited to, (3) and (5). These types of attacks are designed to flexibly capture a range of adversarial behaviours, not just the idealized ones given in the original adversarial training problem. Under the proper assumptions, which will be discussed in Sections 2 and 4, we can extend the convergence result to a broad class of adversarial attacks. We begin by giving some concrete definitions.

Definition 1.5. *For a classifier* $A \in \mathcal{B}(\mathbb{R}^d)$, *we define the Lebesgue measurable function* $\phi : \mathbb{R}^d \to \{0, 1\}$ *by*

$$\phi(x; A) := \begin{cases} 1, & \text{if the adversary can perturb a data point } x \text{ from } A \text{ to } A^{\mathsf{c}} \text{ or vice versa}, \\ 0, & \text{otherwise}. \end{cases}$$

We refer to ϕ *as the deterministic attack function with respect to the classifier* A.

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

Deterministic refers to the fact that the classification risk is completely determined at any point $x \in \mathbb{R}^d$ by the choice of classifier and the associated attack function. We emphasize that this attack function does not consider the true label y associated with x.

In order to generalize the classification risk, it will be essential to isolate the sets where classification loss occurs. We can define the following set operators based on the values of ϕ.

Definition 1.6. *Let $A \in \mathcal{B}(\mathbb{R}^d)$. For a deterministic attack function ϕ, we define the set operators $\Lambda_\phi^i : \mathcal{B}(\mathbb{R}^d) \to \mathcal{B}(\mathbb{R}^d)$ and $\tilde{\Lambda}_\phi^i : \mathcal{B}(\mathbb{R}^d) \to \mathcal{B}(\mathbb{R}^d)$ for $i = 0, 1$ by*

$$\Lambda_\phi^0(A) := \{x \in A^c : \phi(x; A) = 1\}, \qquad \Lambda_\phi^1(A) := \{x \in A : \phi(x; A) = 1\},$$

$$\tilde{\Lambda}_\phi^0(A) := \{x \in A^c : \phi(x; A) = 0\}, \qquad \tilde{\Lambda}_\phi^1(A) := \{x \in A : \phi(x; A) = 0\}.$$

We refer to these four sets collectively as Λ-sets. For convenience, we also define $\Lambda_\phi(A) = \Lambda_\phi^0(A) \cup \Lambda_\phi^1(A)$ and $\tilde{\Lambda}_\phi(A) = \tilde{\Lambda}_\phi^0(A) \cup \tilde{\Lambda}_\phi^1(A)$. Note the 0 and 1 superscripts indicate the label assigned by the classifier A and not the value of the deterministic attack function (i.e. 0 corresponds to points in A^c and 1 corresponds to points in A).

The set $\Lambda_\phi(A)$ contains points that meet the attack criteria for the deterministic attack function ϕ, whereas the set $\tilde{\Lambda}_\phi(A)$ contains points that do not meet the attack criteria. The Λ-sets are mutually disjoint with $A = \Lambda_\phi^1(A) \cup \tilde{\Lambda}_\phi^1(A)$ and $A^c = \Lambda_\phi^0(A) \cup \tilde{\Lambda}_\phi^0(A)$.

We can express the classification risk for a set $A \in \mathcal{B}(\mathbb{R}^d)$ by the loss on the attacked sets, given by $\Lambda_\phi(A)$, and by the loss inherent to the choice of classifier. More formally, we define the generalized classification risk as follows.

Definition 1.7. *The generalized classification risk for a deterministic attack function ϕ and classifier $A \in \mathcal{B}(\mathbb{R}^d)$ is given by*

$$J_\phi(A) := w_0\rho_0(\Lambda_\phi^0(A) \cup A) + w_1\rho_1(\Lambda_\phi^1(A) \cup A^c). \tag{7}$$

As in [28], we seek to separate the total classification risk J_ϕ into the standard Bayes risk (natural error) and the risk attributed to the adversary's attack.

Definition 1.8. *The adversarial deficit for a classifier $A \in \mathcal{B}(\mathbb{R}^d)$ and a deterministic attack function ϕ is defined to be*

$$D_\phi(A) := J_\phi(A) - \mathbb{E}_{(x,y)\sim\mu}[|\mathbb{1}_A(x) - y|],$$

where $\mathbb{E}_{(x,y)\sim\mu}[|\mathbb{1}_A(x) - y|]$ is the standard Bayes risk.

As one can express the standard Bayes risk as

$$\mathbb{E}_{(x,y)\sim\mu}[|\mathbb{1}_A(x) - y|] = w_0\rho_0(A) + w_1\rho_1(A^c),$$

we can derive a more useful equation for the adversarial deficit that mirrors the formulas for the data-adapted perimeters (4) and (6), namely,

$$D_\phi(A) = w_0\rho_0\left(\Lambda_\phi^0(A)\right) + w_1\rho_1\left(\Lambda_\phi^1(A)\right).$$

Unlike the data-adapted perimeters we described above, at this stage $\Lambda_\phi(A)$ is *not* necessarily in some neighbourhood of the decision boundary. We define the relative adversarial deficit for a classifier $A \in \mathcal{B}(\mathbb{R}^d)$ with respect to a set $E \in \mathcal{B}(\mathbb{R}^d)$ to be

$$D_\phi(A; E) := w_0\rho_0\left(\Lambda_\phi^0(A) \cap E\right) + w_1\rho_1\left(\Lambda_\phi^1(A) \cap E\right).$$

With the appropriate definitions in place, we now present the generalized adversarial training problem for the deterministic attack function ϕ.

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

Definition 1.9. *For a deterministic attack function ϕ, the generalized adversarial training problem is given by*

$$\inf_{A\in\mathcal{B}(\mathbb{R}^d)} \mathbb{E}_{(x,y)\sim\mu}[|\mathbb{1}_A(x)-y|]+D_\phi(A). \tag{8}$$

In the previous equation, the adversarial deficit, D_ϕ, takes the place of the data-adapted perimeter terms from (3) and (5).

Remark 1.10. *By construction, the adversarial training problem (3) and the probabilistic adversarial training problem (5) are two examples that fall under this generalized attack function framework. For (3), the ε-deterministic attack function with respect to a classifier $A\in\mathcal{B}(\mathbb{R}^d)$ for $\varepsilon>0$ is*

$$\phi_\varepsilon(x;A):=\begin{cases}1, & \text{if } \mathrm{d}(x,\partial A)<\varepsilon,\\ 0, & \text{otherwise.}\end{cases}$$

For ϕ_ε, we will let $\Lambda^0_\varepsilon(A):=A^\varepsilon\setminus A$, $\Lambda^1_\varepsilon(A):=A\setminus A^{-\varepsilon}$, $\tilde{\Lambda}^0_\varepsilon(A):=A^{\mathsf{c}}\setminus A^\varepsilon$, and $\tilde{\Lambda}^1_\varepsilon(A):=A^{-\varepsilon}$ denote the Λ-sets for convenience.

On the other hand for (5), the (ε,p)-deterministic attack function with respect to a classifier $A\in\mathcal{B}(\mathbb{R}^d)$ for $\varepsilon>0$ and $p\in[0,1)$ is

$$\phi_{\varepsilon,p}(x;A)=\begin{cases}1, & \text{if } \mathbb{P}(\mathbb{1}_A(x')\neq\mathbb{1}_A(x)):x'\sim\mathfrak{p}_{x,\varepsilon})>p,\\ 0, & \text{otherwise.}\end{cases}$$

1.2. Informal main results and discussion

We will focus the main results and discussion on the generalized adversarial training problem (8) and comment on the application to the adversarial training problem (3) and the probabilistic adversarial training problem (5) when appropriate. By Remark 1.10, all statements pertaining to (8) automatically apply to (3) and (5). However, because (3) is sensitive to measure zero changes, results for (3) are stronger than what can be stated in the generalized or probabilistic cases. On the other hand, the results for (5) are identical to those for (8) up to notation.

The first crucial result for (8) provides an estimate on the relative adversarial deficit.

Proposition ((Informal) Energy Exchange Inequality for (8)). *Under mild assumptions on ϕ (see Assumption 2.1), for a classifier $A\in\mathcal{B}(\mathbb{R}^d)$ and a set $E\in\mathcal{B}(\mathbb{R}^d)$ such that $w_0\rho_0-w_1\rho_1>\delta>0$ on E, if $J_\phi(A\setminus E)-J_\phi(A)\geq 0$, then*

$$D_\phi(A;E)\leq D_\phi(E^{\mathsf{c}};A)-\delta\mathcal{L}^d(A\cap E)+w_0\rho_0(\widehat{U}_{11})+w_1\rho_1(\widehat{U}_1)$$

where $\widehat{U}_1\subset\tilde{\Lambda}^1_\phi(A)\cap\tilde{\Lambda}^0_\phi(E)$ and $\widehat{U}_{11}\subset\Lambda^0_\phi(A)\cap\Lambda^1_\phi(E)$.

The energy exchange inequality asserts that if it favourable according to the densities to be labelled 0 on E but adversarial training labels it 1, then the 'perimeter' (more generally, the adversarial deficit) of the original set A must be quantifiably better in the sense of (1.2). In spirit, the energy exchange inequality is connected to relative isoperimetric comparisons as it seeks to relate the relative adversarial deficits (or for (3) the relative ε-perimeters) of two sets to the volume of their intersection. However, the energy exchange inequality has additional error terms that must be accounted for. In the case of the stronger ε-perimeter, $\widehat{U}_1=\emptyset$ so the energy exchange inequality simplifies and can be expressed as follows.

Proposition ((Informal) Energy Exchange Inequality for (3)). *For a classifier $A\in\mathcal{B}(\mathbb{R}^d)$ and a set $E\in\mathcal{B}(\mathbb{R}^d)$ such that $w_0\rho_0-w_1\rho_1>\delta>0$ on E, if $J_\varepsilon(A\setminus E)-J_\varepsilon(A)\geq 0$, then*

$$\varepsilon\mathrm{Per}_\varepsilon(A;E)\leq\varepsilon\mathrm{Per}_\varepsilon(E^{\mathsf{c}};A)-\delta\mathcal{L}^d(A\cap E)+w_0\rho_0(\widehat{U}_{11})$$

where $\widehat{U}_{11}\subset(A^\varepsilon\setminus A)\cap(E\setminus E^{-\varepsilon})$ (see Figure 1).

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

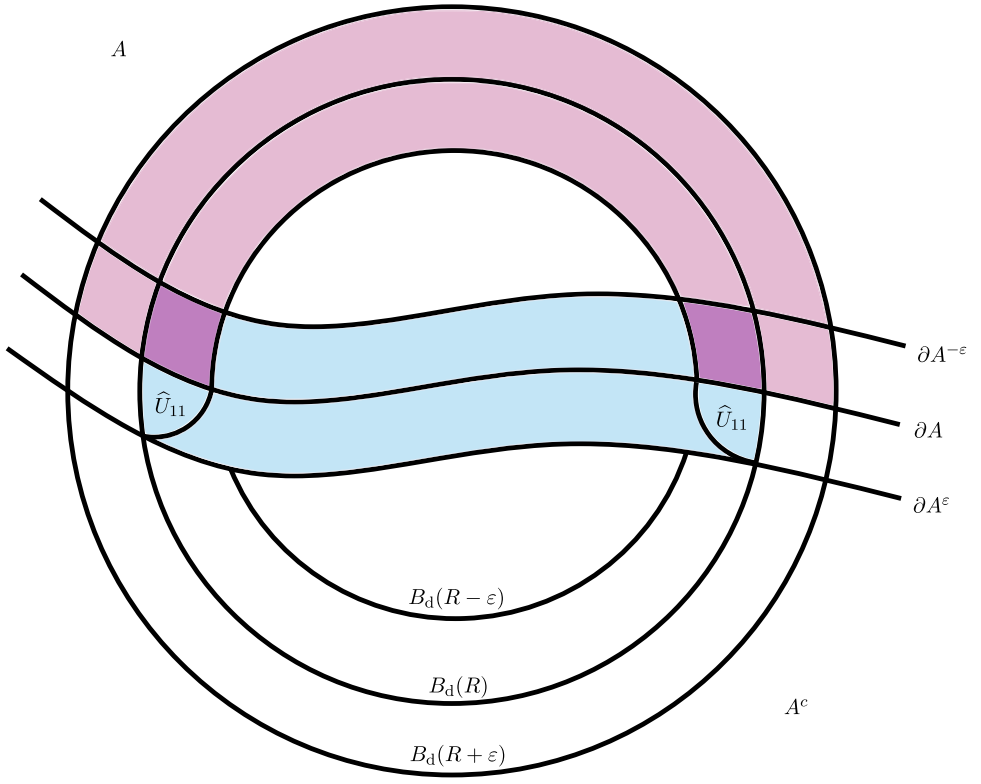


Figure 1. *This diagram illustrates the sets present in the energy exchange inequality for the adversarial training problem (3) when $E = B_{\mathrm{d}}(R)$. The sets comprising $\varepsilon\mathrm{Per}_{\varepsilon}(A; B_{\mathrm{d}}(R))$ are shaded blue and purple, whereas the sets comprising $\varepsilon\mathrm{Per}_{\varepsilon}(B_{\mathrm{d}}(R)^{\mathrm{c}}; A)$ are shaded pink and purple.*

As for the relative probabilistic perimeter $\mathrm{ProbPer}_{\varepsilon,p}$, the energy exchange inequality is the same as that for (8) up to notation.

Proposition ((Informal) Energy Exchange Inequality for (5)). *For a classifier $A \in \mathcal{B}(\mathbb{R}^d)$ and a set $E \in \mathcal{B}(\mathbb{R}^d)$ such that $w_0\rho_0 - w_1\rho_1 > \delta > 0$ on E, if $J_{\varepsilon,p}(A \setminus E) - J_{\varepsilon,p}(A) \geq 0$, then*

$$\mathrm{ProbPer}_{\varepsilon,p}(A; E) \leq \mathrm{ProbPer}_{\varepsilon,p}\left(E^{\mathrm{c}}; A\right) - \delta\mathcal{L}^d(A \cap E) + w_0\rho_0(\widehat{U}_{11}) + w_1\rho_1(\widehat{U}_1)$$

where $\widehat{U}_1 \subset (A \setminus A^{-\varepsilon}) \cap (E^{\varepsilon} \setminus E)$ and $\widehat{U}_{11} \subset (A^{\varepsilon} \setminus A) \cap (E \setminus E^{-\varepsilon})$.

The energy exchange inequality allows us to argue that classifiers which are minimizers of the generalized adversarial training problem (8), if they exist, can be made disjoint from sets where it is energetically preferable to be labelled 0 when the adversarial budget ε is small enough. As we will see, in the generalized setting we can only guarantee the uniqueness of minimizers of (5) and (8) up to sets of measure zero; however, we can show that the intersection of such sets with an energetic preference for the label zero with minimizers must have $\mathcal{L}^d$ measure zero. For the adversarial training problem (3), we can improve the result to show that any minimizer must be disjoint from these sets when ε is small enough. This result builds towards proving uniform convergence of minimizers of (8) to the Bayes classifier, which is the next main result. In order to prove the convergence rate, we must include a non-degeneracy assumption to ensure $\mathrm{d}_H(A_0^{\max}, A_0^{\min}) = 0$ and the Bayes classifier is unique in the sense of Remark 1.1.

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

Theorem (Informal). *With mild assumptions on ϕ, let K be compact and let $\{A_{\varepsilon,\phi}\}_{\varepsilon>0}$ be any sequence of minimizers to the generalized adversarial training problem (8). Assuming that $w_0\rho_0 - w_1\rho_1$ is non-degenerate, then*

$$\mathrm{d}_H((A_{\varepsilon,\phi} \cup N_1 \setminus N_2) \cap K, A_0 \cap K) \to 0$$

as $\varepsilon \to 0^+$, where N_1, N_2 are sets of $\mathcal{L}^d$ measure zero, d_H is the Hausdorff distance and A_0 is the Bayes classifier.

However, the theorem actually proved is more general and does not require a unique Bayes classifier. Under these relaxed assumptions, we prove a corralling result for the sequence $\{A_{\varepsilon,\phi}\}_{\varepsilon>0}$ with respect to the Hausdorff distance from the maximal Bayes classifier, $A_0^{\max}$, and the minimal Bayes classifier, $A_0^{\min}$. In essence, the corralling result states that the boundary of $\lim_{\varepsilon\to 0^+} A_{\varepsilon,\phi} \cup N_1 \setminus N_2$ must lie between the boundaries of $A_0^{\max}$ and $A_0^{\min}$. When we specify this result to the adversarial training problem (3), we no longer have to remove a $\mathcal{L}^d$ measure zero set and instead prove the following.

Theorem (Informal). *Let K be compact and let $\{A_\varepsilon\}_{\varepsilon>0}$ be any sequence of minimizers to the adversarial training problem (3). Assuming that $w_0\rho_0 - w_1\rho_1$ is non-degenerate, then*

$$\mathrm{d}_H(A_\varepsilon \cap K, A_0 \cap K) \to 0$$

as $\varepsilon \to 0^+$, where d_H is the Hausdorff distance and A_0 is the Bayes classifier.

For the probabilistic adversarial training problem, the uniform convergence result states,

Theorem (Informal). *Let K be compact and let $\{A_{\varepsilon,p}\}_{\varepsilon>0}$ be any sequence of minimizers to the probabilistic adversarial training problem (5) for some fixed $p \in [0,1)$. Assuming that $w_0\rho_0 - w_1\rho_1$ is non-degenerate, then*

$$\mathrm{d}_H((A_{\varepsilon,p} \cup N_1 \setminus N_2) \cap K, A_0 \cap K) \to 0$$

as $\varepsilon \to 0^+$, where N_1, N_2 are sets of $\mathcal{L}^d$ measure zero, d_H is the Hausdorff distance and A_0 is the Bayes classifier.

As with (8), if we relax the assumption that the Bayes classifier is unique, we can instead prove an analogous corralling result with respect to $A_0^{\max}$ and $A_0^{\min}$ for (3) and (5).

With the non-degeneracy condition in place, we can also consider the rate of convergence and show that it is at most $O(\varepsilon^{\frac{1}{d+2}})$ for all three adversarial training problems. However, we do not expect this result to be optimal and would expect that the convergence rate to be $O(\varepsilon)$, which we discuss further in Remark 3.13.

2. Energy exchange inequality

In this section, we will prove a quantitative result for the adversarial deficit, which can then be applied to the ε-perimeter and the probabilistic ε-perimeter. In order to do so, we will require the deterministic attack function ϕ and the corresponding Λ-sets to have the following structural properties.

Assumption 2.1. *Recall Definition 1.6. Let $A, E \in \mathcal{B}(\mathbb{R}^d)$. We will make the following two assumptions to ensure consistency with respect to complements and set difference:*

1. *Complement Property (CP): $\phi(x; A) = \phi(x; A^c)$, or in terms of Λ-sets, $\Lambda_\phi^0(A) = \Lambda_\phi^1(A^c)$ and $\tilde{\Lambda}_\phi^0(A) = \tilde{\Lambda}_\phi^1(A^c)$.*
2. *Λ-Monotonicity (ΛM):*
   (i) If $x \in \tilde{\Lambda}_\phi^0(A)$, then $x \in \tilde{\Lambda}_\phi^0(A \setminus E)$.
   (ii) If $x \in \tilde{\Lambda}_\phi^1(E)$, then $x \in \tilde{\Lambda}_\phi^0(A \setminus E)$.

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

(iii) *If* $x \in \Lambda^0_\phi(E) \cap A$, *then* $x \in \Lambda^1_\phi(A \setminus E)$.
(iv) *If* $x \in \Lambda^1_\phi(A) \cap E^c$, *then* $x \in \Lambda^1_\phi(A \setminus E)$.

□

In the following series of remarks, we seek to better understand these two properties generally and as they apply to the adversarial and probabilistic adversarial settings.

Remark 2.2 (On Monotonicity). *We note that the deterministic attack functions* ϕ *that satisfy Assumption 2.1 are not monotonic with respect to set inclusion unless* ϕ *is the trivial attack function (i.e.* $\phi \equiv 0$ *or* $\phi \equiv 1$*). To illustrate this, suppose* ϕ *is monotonic. By the monotonicity of* ϕ *with respect to set inclusion coupled with the complement property,*

$$\phi(x; A) \overset{(CP)}{=} \phi(x; A^c) \leq \phi(x; (A \setminus E)^c) \overset{(CP)}{=} \phi(x; A \setminus E) \leq \phi(x; A).$$

This implies $\phi(x; A) \equiv \phi(x; A \setminus E)$ *and, if we let* $E = A$, *that* $\phi(x; A) \equiv \phi(x; \emptyset)$. *Hence, the attack is independent of A, which can only be satisfied by a trivial attack function.*

Although ϕ *itself is not monotonic, if you have a function* ψ *which is monotonic in terms of set inclusion, then setting* ϕ *via its level set yields an attack function which satisfies* Λ*-monotonicity. In particular, both the distance function and the probability function are monotonic.*

Remark 2.3. *We will verify that the adversarial training problem (3) and the probabilistic adversarial training problem (5) satisfy Assumption 2.1. Recall from Remark 1.10, the attack for (3) is denoted* ϕ_ε *and the attack for (5) is denoted* $\phi_{\varepsilon,p}$ *for some* $\varepsilon > 0$ *and* $p \in [0, 1)$.

We will first show that ϕ_ε *satisfies Assumption 2.1. For the complement property, recognize that since* $\partial A = \partial(A^c)$, $\Lambda^0_\varepsilon(A) = \Lambda^1_\varepsilon(A^c)$ *and* $\tilde{\Lambda}^0_\varepsilon(A) = \tilde{\Lambda}^1_\varepsilon(A^c)$ *by definition. As for* Λ*-monotonicity, we can verify these four statements directly.*

(i) *If* $x \in \tilde{\Lambda}^0_\varepsilon(A)$, *then* $d(x, A \setminus E) \geq d(x, A) \geq \varepsilon$ *so* $x \in \tilde{\Lambda}^0_\varepsilon(A \setminus E)$.
(ii) *If* $x \in \tilde{\Lambda}^1_\varepsilon(E)$, *then* $d(x, A \setminus E) \geq d(x, E^c) \geq \varepsilon$ *so* $x \in \tilde{\Lambda}^0_\varepsilon(A \setminus E)$.
(iii) *If* $x \in \Lambda^0_\varepsilon(E) \cap A$, *then* $d(x, (A \setminus E)^c) \leq d(x, E) < \varepsilon$ *so* $x \in \Lambda^1_\varepsilon(A \setminus E)$.
(iv) *If* $x \in \Lambda^1_\varepsilon(A) \cap E^c$, *then* $d(x, (A \setminus E)^c) \leq d(x, A^c) < \varepsilon$ *so* $x \in \Lambda^1_\varepsilon(A \setminus E)$.

Now, we consider $\phi_{\varepsilon,p}$. *By definition,*

$$\Lambda^1_{\varepsilon,p}(A^c) = \{x \in A^c : \mathbb{P}(x' \in (A^c)^c : x' \sim \mathfrak{p}_{x,\varepsilon}) > p\} = \Lambda^0_{\varepsilon,p}(A).$$

Similarly, one can show $\tilde{\Lambda}^0_{\varepsilon,p}(A) = \tilde{\Lambda}^1_{\varepsilon,p}(A^c)$. *Hence, the complement property holds for* $\phi_{\varepsilon,p}$. *Now we consider* Λ*-monotonicity. To simplify notation, we let* $\mathbb{P}(x; A) := \mathbb{P}(x' \in A : x' \sim \mathfrak{p}_{x,\varepsilon})$. *Examining each of the* Λ*-monotonicity properties, we find the monotonicity with respect to set inclusion of the probability function*

(i) *If* $x \in \tilde{\Lambda}^0_{\varepsilon,p}(A)$, *then* $\mathbb{P}(x; A \setminus E) \leq \mathbb{P}(x; A) \leq p$ *so* $x \in \tilde{\Lambda}^0_{\varepsilon,p}(A \setminus E)$.
(ii) *If* $x \in \tilde{\Lambda}^1_{\varepsilon,p}(E)$, *then* $\mathbb{P}(x; A \setminus E) \leq \mathbb{P}(x; E^c) \leq p$ *so* $x \in \tilde{\Lambda}^0_{\varepsilon,p}(A \setminus E)$.
(iii) *If* $x \in \Lambda^0_{\varepsilon,p}(E) \cap A$, *then* $\mathbb{P}(x; (A \setminus E)^c) \geq \mathbb{P}(x; E) > p$ *so* $x \in \Lambda^1_{\varepsilon,p}(A \setminus E)$.
(iv) *If* $x \in \Lambda^1_{\varepsilon,p}(A) \cap E^c$, *then* $\mathbb{P}(x; (A \setminus E)^c) \geq \mathbb{P}(x; A^c) > p$ *so* $x \in \Lambda^1_{\varepsilon,p}(A \setminus E)$.

Thus, $\phi_{\varepsilon,p}$ *satisfies* Λ*-monotonicity and Assumption 2.1.*

Remark 2.4 (Λ-set Decompositions). *Under Assumption 2.1, we may decompose* $\mathbb{R}^d$ *in terms of the* Λ*-sets for* $A, E \in \mathcal{B}(\mathbb{R}^d)$ *according to* Λ*-monotonicity. In doing so, we define the sets* $U_1, \ldots, U_{13}$, *which partition* $\mathbb{R}^d$ *(see Table 1 and Figure 2).*

For the sets U_i *where no conclusion can be made about* $\phi(x; A \setminus E)$, *we will further decompose them into two subsets based on the* ϕ *values, i.e.*

$$\widetilde{U}_i = \{x \in U_i : \phi(x; A \setminus E) = 0\}, \quad \widehat{U}_i = \{x \in U_i : \phi(x; A \setminus E) = 1\}, \tag{9}$$

for $i = 1, 3, 6, 9, 10,$ *and* 11.

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

Table 1. *This table defines the 13 U_i sets and exhibits all possible conclusions about the Λ-sets for $A \setminus E$ based on the Λ-sets for A and E from Λ-monotonicity. This set decomposition, along with the further refinement in (9), will be key in proving the energy exchange inequality*

If $x \in U_i$	Then for $\Lambda(A \setminus E)$ we have
$U_1 := \tilde{\Lambda}^1_\phi(A) \cap \tilde{\Lambda}^0_\phi(E)$	N/A
$U_2 := \tilde{\Lambda}^1_\phi(A) \cap \Lambda^0_\phi(E)$	$x \in \Lambda^1_\phi(A \setminus E)$
$U_3 := \tilde{\Lambda}^1_\phi(A) \cap \Lambda^1_\phi(E)$	N/A
$U_4 := \tilde{\Lambda}^1_\phi(A) \cap \tilde{\Lambda}^1_\phi(E)$	$x \in \tilde{\Lambda}^0_\phi(A \setminus E)$
$U_5 := \Lambda^1_\phi(A) \cap \tilde{\Lambda}^1_\phi(E)$	$x \in \tilde{\Lambda}^0_\phi(A \setminus E)$
$U_6 := \Lambda^1_\phi(A) \cap \Lambda^1_\phi(E)$	N/A
$U_7 := \Lambda^1_\phi(A) \cap \Lambda^0_\phi(E)$	$x \in \Lambda^1_\phi(A \setminus E)$
$U_8 := \Lambda^1_\phi(A) \cap \tilde{\Lambda}^0_\phi(E)$	$x \in \Lambda^1_\phi(A \setminus E)$
$U_9 := \Lambda^0_\phi(A) \cap \Lambda^0_\phi(E)$	N/A
$U_{10} := \Lambda^0_\phi(A) \cap \tilde{\Lambda}^0_\phi(E)$	N/A
$U_{11} := \Lambda^0_\phi(A) \cap \Lambda^1_\phi(E)$	N/A
$U_{12} := \Lambda^0_\phi(A) \cap \tilde{\Lambda}^1_\phi(E)$	$x \in \tilde{\Lambda}^0_\phi(A \setminus E)$
$U_{13} := \tilde{\Lambda}^0_\phi(A)$	$x \in \tilde{\Lambda}^0_\phi(A \setminus E)$

The auxiliary symbols are meant to help the reader group the terms. Notice that the $\widetilde{U}_i$ sets contain points that cannot be perturbed by the adversary into the other class for the classifier $A \setminus E$ in accordance with all $\tilde{\Lambda}$ sets also containing points that are unable to be attacked by the adversary. On the other hand, the $\widehat{U}_i$ sets contain only points that can be perturbed into the opposite class.

With this decomposition, we can express the Λ-sets for $A \setminus E$ using the U sets as follows:

$$\Lambda^0_\phi(A \setminus E) = \widehat{U}_3 \cup \widehat{U}_6 \cup \widehat{U}_9 \cup \widehat{U}_{10} \cup \widehat{U}_{11},$$

$$\Lambda^1_\phi(A \setminus E) = \widehat{U}_1 \cup U_2 \cup U_7 \cup U_8,$$

$$\tilde{\Lambda}^0_\phi(A \setminus E) = \widetilde{U}_3 \cup U_4 \cup U_5 \cup \widetilde{U}_6 \cup \widetilde{U}_9 \cup \widetilde{U}_{10} \cup \widetilde{U}_{11} \cup U_{12} \cup U_{13},$$

$$\tilde{\Lambda}^1_\phi(A \setminus E) = \widetilde{U}_1.$$

Depending on extra structure imposed by the choice of ϕ, sometimes we can conclude certain sets are empty. For example, when $\phi = \phi_\varepsilon$ (see Remark 1.10), we have $\widehat{U}_1 = \emptyset$, $\widetilde{U}_3 = \emptyset$, and $\widetilde{U}_{10} = \emptyset$. In the case where such sets are unambiguous in terms of the values of $\phi(x; A \setminus E)$, we drop the hat or tilde notation. However, U_6, U_9 and U_{11} still require a finer decomposition. Note that generally $\widetilde{U}_6, \widetilde{U}_9 = \emptyset$, but when boundaries of A and E intersect at more than discrete points, then these sets can be non-empty. When $\widetilde{U}_6, \widetilde{U}_9 = \emptyset$ (such as in Figure 2), we also drop the tilde notation and let $U_6 = \widehat{U}_6$ and $U_9 = \widehat{U}_9$. The claims made here are verified in Appendix A.1.

Having stated our assumptions on ϕ, we now turn to proving the first main result. In the following proposition, we examine the difference in energy between classifiers A and $A \setminus E$ for $A, E \in \mathcal{B}(\mathbb{R}^d)$ when E belongs to a region where the label 0 is energetically preferable according to the Bayes risk. We refer to the resulting inequality as the *energy exchange inequality* because it quantifies the effect of removing the set E from a classifier A by examining the difference in risks.

Proposition 2.5 (Energy Exchange Inequality). *Let ϕ be a deterministic attack function that satisfies Assumption 2.1, let $A, E \in \mathcal{B}(\mathbb{R}^d)$, and assume that $w_0\rho_0 - w_1\rho_1 > \delta > 0$ on E. If $J_\phi(A \setminus E) - J_\phi(A) \geq 0$, then*

$$D_\phi(A; E) \leq D_\phi(E^c; A) - \delta \mathcal{L}^d(A \cap E) + w_0\rho_0(\widehat{U}_{11}) + w_1\rho_1(\widehat{U}_1),$$

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

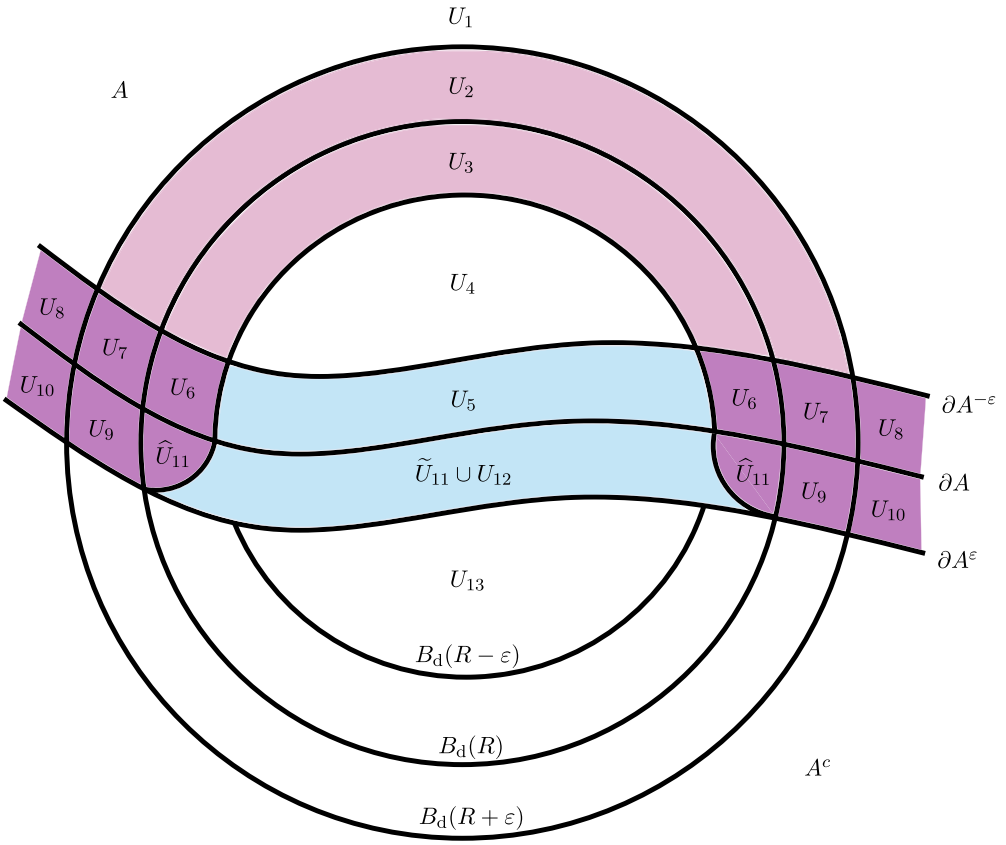


Figure 2. *This diagram depicts the U_i regions for the attack function ϕ_ε associated with adversarial training problem (3). The ε-perimeter regions of A are shaded blue and purple, whereas ε-perimeter regions of $A \setminus B_{\mathrm{d}}(R)$ are shaded pink and purple. Note that some sets, such as $\widehat{U}_1$, are null sets for the ε-perimeter, and so do not appear in this figure.*

where U_1^1 and $\widehat{U}_{11}$ are defined in Table 1, namely $\widehat{U}_1 = \{x \in \tilde{\Lambda}_\phi^1(A) \cap \tilde{\Lambda}_\phi^0(E) : \phi(x; A \setminus E) = 1\}$ and $\widehat{U}_{11} = \{x \in \Lambda_\phi^0(A) \cap \Lambda_\phi^1(E) : \phi(x; A \setminus E) = 1\}$.

Proof. By (7), we have

$$J_\phi(A) = w_0\rho_0\left(A \cup \Lambda_\phi^0(A)\right) + w_1\rho_1\left(A^{\mathsf{c}} \cup \Lambda_\phi^1(A)\right),$$
$$J_\phi(A \setminus E) = w_0\rho_0\left((A \setminus E) \cup \Lambda_\phi^0(A \setminus E)\right) + w_1\rho_1\left((A \setminus E)^{\mathsf{c}} \cup \Lambda_\phi^1(A \setminus E)\right).$$

Based on Remark 2.4 with further details shown in Appendix A.2, we can express $A \cap E$ and the sets comprising $J_\phi(A \setminus E)$ as

$$\begin{aligned} A \cap E &= U_3 \cup U_4 \cup U_5 \cup U_6, \\ \Lambda_\phi^0(A \setminus E) &= \widehat{U}_3 \cup \widehat{U}_6 \cup \widehat{U}_9 \cup \widehat{U}_{10} \cup \widehat{U}_{11}, \\ \Lambda_\phi^1(A \setminus E) &= \widehat{U}_1 \cup U_2 \cup U_7 \cup U_8, \\ A \setminus E &= U_1 \cup U_2 \cup U_7 \cup U_8, \\ (A \setminus E)^{\mathsf{c}} &= U_3 \cup U_4 \cup U_5 \cup U_6 \cup U_9 \cup U_{10} \cup U_{11} \cup U_{12} \cup U_{13}. \end{aligned} \tag{10}$$

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

We can write the adversarial deficit terms as

$$D_\phi(A;E) = w_0\rho_0(U_{11}\cup U_{12}) + w_1\rho_1(U_5\cup U_6), \tag{11}$$
$$D_\phi(E^{\mathsf{c}};A) = w_0\rho_0(U_3\cup U_6) + w_1\rho_1(U_2\cup U_7).$$

Then we estimate,

$$\begin{aligned}
J_\phi(A\setminus E) - J_\phi(A) &= w_0\rho_0(U_1\cup U_2\cup \widehat{U}_3\cup \widehat{U}_6\cup U_7\cup U_8\cup \widehat{U}_9\cup \widehat{U}_{10}\cup \widehat{U}_{11})\\
&\quad + w_1\rho_1(\widehat{U}_1\cup U_2\cup U_3\cup U_4\cup U_5\cup U_6\cup U_7\cup U_8\cup U_9\cup U_{10}\cup U_{11}\cup U_{12}\cup U_{13})\\
&\quad - w_0\rho_0(U_1\cup U_2\cup U_3\cup U_4\cup U_5\cup U_6\cup U_7\cup U_8\cup U_9\cup U_{10}\cup U_{11}\cup U_{12})\\
&\quad - w_1\rho_1(U_5\cup U_6\cup U_7\cup U_8\cup U_9\cup U_{10}\cup U_{11}\cup U_{12}\cup U_{13})\\
&\le w_0\rho_0(\cancel{U_1}\cup \cancel{U_2}\cup \underline{\underline{U_3}}\cup \underline{\underline{U_6}}\cup \cancel{U_7}\cup \cancel{U_8}\cup \cancel{\widehat{U}_9}\cup \cancel{\widehat{U}_{10}}\cup \widehat{U}_{11})\\
&\quad + w_1\rho_1(\widehat{U}_1\cup \underline{U_2}\cup \underline{U_3}\cup \underline{U_4}\cup \underline{U_5}\cup \underline{U_6}\cup \underline{U_7}\cup \cancel{U_8}\cup \cancel{U_9}\cup \cancel{U_{10}}\cup \cancel{U_{11}}\cup \cancel{U_{12}}\cup \cancel{U_{13}})\\
&\quad - w_0\rho_0(\cancel{U_1}\cup \cancel{U_2}\cup \underline{U_3}\cup \underline{U_4}\cup \underline{U_5}\cup \underline{U_6}\cup \cancel{U_7}\cup \cancel{U_8}\cup (\widetilde{U}_9\cup \cancel{\widehat{U}_9})\cup (\widetilde{U}_{10}\cup \cancel{\widehat{U}_{10}})\cup \underline{U_{11}}\cup \underline{U_{12}})\\
&\quad - w_1\rho_1(\underline{U_5}\cup \underline{U_6}\cup U_7\cup \cancel{U_8}\cup \cancel{U_9}\cup \cancel{U_{10}}\cup \cancel{U_{11}}\cup \cancel{U_{12}}\cup \cancel{U_{13}})\\
&\le \underline{\underline{D_\phi(E^{\mathsf{c}};A)}} - \underline{D_\phi(A;E)} - (w_0\rho_0 - w_1\rho_1)(\underline{A\cap E}) + w_0\rho_0(\widehat{U}_{11}) + w_1\rho_1(\widehat{U}_1).
\end{aligned}$$

In the last line, the inequality results from neglecting all remaining terms with a negative sign. As $J_\phi(A\setminus E) - J_\phi(A) \ge 0$ and $w_0\rho_0 - w_1\rho_1 > \delta > 0$ on E, we estimate

$$\begin{aligned}
D_\phi(A;E) &\le D_\phi(E^{\mathsf{c}};A) - (w_0\rho_0 - w_1\rho_1)(A\cap E) + w_0\rho_0(\widehat{U}_{11}) + w_1\rho_1(\widehat{U}_1)\\
&< D_\phi(E^{\mathsf{c}};A) - \delta\mathcal{L}^d(A\cap E) + w_0\rho_0(\widehat{U}_{11}) + w_1\rho_1(\widehat{U}_1).
\end{aligned}$$

□

Observe that if $A\in\mathcal{B}(\mathbb{R}^d)$ is a minimizer of J_ϕ for some deterministic attack function ϕ, then $J_\phi(A\setminus E) - J_\phi(A)\ge 0$ for any $E\in\mathcal{B}(\mathbb{R}^d)$ and Proposition 2.5 applies. This will be the setting for our results, although we state the result in its most general form here.

In later energy arguments, it will be helpful to express the difference in classification risks exactly instead of combining terms to form $D_\phi(A;E)$, $D_\phi(E^{\mathsf{c}};A)$, and $\mathcal{L}^d(A\cap E)$. In Corollary 2.6, we consider the same computation for $J_\phi(A\setminus E) - J_\phi(A)$ but now aim to simplify the difference as much as possible.

Corollary 2.6. *Let $A\in\mathcal{B}(\mathbb{R}^d)$ be a classifier for the generalized adversarial training problem and let $E\in\mathcal{B}(\mathbb{R}^d)$. Then, using the same notation as in Proposition 2.5 and under the same assumptions,*

$$\begin{aligned}
J_\phi(A\setminus E) - J_\phi(A) &= w_1\rho_1(\widehat{U}_1\cup U_2\cup \widehat{U}_3) - (w_0\rho_0 - w_1\rho_1)(\widetilde{U}_3\cup U_4)\\
&\quad - w_0\rho_0(U_5\cup \widetilde{U}_6\cup \widetilde{U}_9\cup \widetilde{U}_{10}\cup \widetilde{U}_{11}\cup U_{12}).
\end{aligned}$$

Proof. Let all sets U_i, $\widehat{U}_i$, $\widetilde{U}_i$ be as defined in Table 1 and (9). We compute the exact difference in energies as follows:

$$\begin{aligned}
J_\phi(A\setminus E) - J_\phi(A) &= w_0\rho_0(\cancel{U_1}\cup \cancel{U_2}\cup \cancel{\widehat{U}_3}\cup \cancel{\widehat{U}_6}\cup \cancel{U_7}\cup \cancel{U_8}\cup \cancel{\widehat{U}_9}\cup \cancel{\widehat{U}_{10}}\cup \cancel{\widehat{U}_{11}})\\
&\quad + w_1\rho_1(\widehat{U}_1\cup U_2\cup U_3\cup U_4\cup \cancel{U_5}\cup \cancel{U_6}\cup \cancel{U_7}\cup \cancel{U_8}\cup \cancel{U_9}\cup \cancel{U_{10}}\cup \cancel{U_{11}}\cup \cancel{U_{12}}\cup \cancel{U_{13}})\\
&\quad - w_0\rho_0(\cancel{U_1}\cup \cancel{U_2}\cup (\widetilde{U}_3\cup \cancel{\widehat{U}_3})\cup U_4\cup U_5\cup (\widetilde{U}_6\cup \cancel{\widehat{U}_6})\cup \cancel{U_7}\cup \cancel{U_8}\cup (\widetilde{U}_9\cup \cancel{\widehat{U}_9})\ldots\\
&\qquad \ldots\cup (\widetilde{U}_{10}\cup \cancel{\widehat{U}_{10}})\cup (\widetilde{U}_{11}\cup \cancel{\widehat{U}_{11}})\cup U_{12})\\
&\quad - w_1\rho_1(\cancel{U_5}\cup \cancel{U_6}\cup \cancel{U_7}\cup \cancel{U_8}\cup \cancel{U_9}\cup \cancel{U_{10}}\cup \cancel{U_{11}}\cup \cancel{U_{12}}\cup \cancel{U_{13}})
\end{aligned}$$

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

$$
\begin{aligned}
&= w_1\rho_1(\widehat{U}_1 \cup U_2 \cup \widetilde{U}_3 \cup \widehat{U}_3 \cup U_4) \\
&- w_0\rho_0(\widetilde{U}_3 \cup U_4 \cup U_5 \cup \widetilde{U}_6 \cup \widetilde{U}_9 \cup \widetilde{U}_{10} \cup \widetilde{U}_{11} \cup U_{12}) \\
&= w_1\rho_1(\widehat{U}_1 \cup U_2 \cup \widehat{U}_3) - (w_0\rho_0 - w_1\rho_1)(\widetilde{U}_3 \cup U_4) \\
&- w_0\rho_0(U_5 \cup \widetilde{U}_6 \cup \widetilde{U}_9 \cup \widetilde{U}_{10} \cup \widetilde{U}_{11} \cup U_{12}).
\end{aligned}
$$

□

In the following pair of corollaries, we will apply Proposition 2.5 to the adversarial training problem (3) and the probabilistic adversarial training problem (5).

Corollary 2.7. *Let $\varepsilon > 0$ and $\phi = \phi_\varepsilon$. Let $A, E \in \mathcal{B}(\mathbb{R}^d)$ such that $w_0\rho_0 - w_1\rho_1 > \delta > 0$ on E and $J_\varepsilon(A \setminus E) - J_\varepsilon(A) \geq 0$. Then*

$$
\varepsilon \mathrm{Per}_\varepsilon(A; E) \leq \varepsilon \mathrm{Per}_\varepsilon(E^{\mathsf{c}}; A) - \delta \mathcal{L}^d(A \cap E) + w_0\rho_0(\widehat{U}_{11}), \tag{12}
$$

where $\widehat{U}_{11} = \{x \in A^{\mathsf{c}} \cap E : \mathrm{d}(x, A \setminus E) < \varepsilon\}$.

Proof. To prove the corollary, we only need to check that ϕ_ε satisfies Assumption 2.1 (which is done in Remark 2.3) and to verify that $\widehat{U}_1$ is empty. To that end, if $x \in \widehat{U}_1$, then

$$
x \in A \cap E^c \text{ such that } d(x, A^c) > \varepsilon \text{ and } d(x, E) > \varepsilon,
$$

which in turn implies that $d(x, A^{\mathsf{c}} \cup E) > \varepsilon$. Hence for such x, $\phi_\varepsilon(x; A \setminus E) = 0$ and accordingly $\widehat{U}_1 = \emptyset$. □

Corollary 2.8. *Let $\varepsilon > 0$, $p \in [0, 1)$, $\{\mathfrak{p}_{x,\varepsilon}\}_{x \in \mathbb{R}^d}$ be a family of probability measures, and $\phi = \phi_{\varepsilon,p}$. Let $A, E \in \mathcal{B}(\mathbb{R}^d)$ such that $w_0\rho_0 - w_1\rho_1 > \delta > 0$ on E and $J_{\varepsilon,p}(A \setminus E) - J_{\varepsilon,p}(A) \geq 0$. Then,*

$$
\mathrm{ProbPer}_{\varepsilon,p}(A; E) \leq \mathrm{ProbPer}_{\varepsilon,p}(E^{\mathsf{c}}; A) - \delta \mathcal{L}^d(A \cap E) + w_0\rho_0(\widehat{U}_{11}) + w_1\rho_1(\widehat{U}_1),
$$

where $\widehat{U}_{11} = \{x \in A^{\mathsf{c}} \cap E : \mathbb{P}(x' \in A \setminus E : x' \sim \mathfrak{p}_{x,\varepsilon}) > p\}$ and $\widehat{U}_1 = \{x \in \tilde{\Lambda}^1_{\varepsilon,p}(A) \cap \tilde{\Lambda}^0_{\varepsilon,p}(E) : \mathbb{P}(x' \in (A \setminus E)^{\mathsf{c}} : x' \sim \mathfrak{p}_{x,\varepsilon}) > p\}$.

Proof. We verified that $\phi_{\varepsilon,p}$ satisfies Assumption 2.1 in Remark 2.3. Thus, we can apply Proposition 2.5 to conclude that the energy exchange inequality holds for the probabilistic adversarial training problem (5). □

3. Uniform convergence for the adversarial training problem

Before tackling convergence for the generalized adversarial problem (8), we first consider the convergence for the adversarial training problem (3) to understand the results in a more concrete setting. The results for (3) are also stronger than those for (8) and allow for more straightforward proofs that provide the basis for our approach in the subsequent section. We will return to (8) in Section 4 equipped with better intuition and understanding.

In this section, we establish uniform convergence in the Hausdorff metric of minimizers of the adversarial training problem (3) to Bayes classifiers on compact sets as the parameter $\varepsilon \to 0^+$. As previously stated in Remark 1.2, current convergence results are in the (weaker) L^1 topology. We begin by stating a modest assumption we make about the underlying metric space.

Assumption 3.1. *For the remainder of the paper, we assume that the metric d is induced by a norm. Then, $\mathcal{L}^d(B_{\mathrm{d}}(r)) := \omega_{\mathrm{d}} r^d$ for the constant $\omega_{\mathrm{d}} = \mathcal{L}^d(B_{\mathrm{d}}(1))$. Naturally, ω_{d} will also depend on the dimension d, but we suppress this in the notation. Additionally, we will identify the conditional measures in (4) with their densities, meaning that we can express $d\rho_i = \rho_i(x)\,dx$.*

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

For these norm balls, it will be useful to estimate their ε-perimeter. When $\varepsilon \leq R$ and ρ_0, ρ_1 are bounded from above, this amounts to estimating the volume between two norm balls that are distance 2ε apart.

Lemma 3.2. *Let $0 < \varepsilon \leq R$ for some fixed $R > 0$. Suppose $\rho_0, \rho_1 \leq M$ on $\mathbb{R}^d$. Then, there exists a constant $\alpha > 0$ independent of R, ε, and x such that*

$$\varepsilon \mathrm{Per}_\varepsilon(B_{\mathrm{d}}(x,R)) \leq \alpha R^{d-1}\varepsilon.$$

Proof. Recall that (4) for $A = B_{\mathrm{d}}(x,R)$ gives

$$\varepsilon \mathrm{Per}_\varepsilon(B_{\mathrm{d}}(x,R)) = w_0\rho_0(B_{\mathrm{d}}(x,R+\varepsilon) \setminus B_{\mathrm{d}}(x,R)) + w_1\rho_1(B_{\mathrm{d}}(x,R) \setminus B_{\mathrm{d}}(x,R-\varepsilon)).$$

As ρ_0, ρ_1 are bounded from above by M,

$$\varepsilon \mathrm{Per}_\varepsilon(B_{\mathrm{d}}(x,R)) \leq M\mathcal{L}^d(B_{\mathrm{d}}(x,R+\varepsilon) \setminus B_{\mathrm{d}}(x,R-\varepsilon)) = M(\mathcal{L}^d(B_{\mathrm{d}}(x,R+\varepsilon)) - \mathcal{L}^d(B_{\mathrm{d}}(x,R-\varepsilon))).$$

By the scaling properties of the norm ball, $\mathcal{L}^d(B_{\mathrm{d}}(x,r)) = \omega_{\mathrm{d}}(r)^d$ for all $r \geq 0$. By convexity, we estimate

$$(R+\varepsilon)^d - (R-\varepsilon)^d \leq d(R+\varepsilon)^{d-1}2\varepsilon.$$

As $\varepsilon \leq R$, we conclude

$$\varepsilon \mathrm{Per}_\varepsilon(B_{\mathrm{d}}(x,R)) \leq M\omega_{\mathrm{d}}d(2R)^{d-1}2\varepsilon \leq \alpha R^{d-1}\varepsilon.$$

□

Throughout the paper, we will require an upper bound on the ε-perimeter of the *complement* of $B_{\mathrm{d}}(x,R)$. By the complement property from Assumption 2.1 (verified to hold for the ε-perimeter in Remark 2.3), the bound given by Lemma 3.2 still holds for $\varepsilon\mathrm{Per}_\varepsilon(B_{\mathrm{d}}(x,R)^{\mathrm{c}})$ since the same upper bound is true for ρ_0 and ρ_1, namely,

$$\varepsilon \mathrm{Per}_\varepsilon(B_{\mathrm{d}}(x,R)^{\mathrm{c}}) \leq \alpha R^{d-1}\varepsilon. \tag{13}$$

With our normed setting clear, we begin the process of proving uniform Hausdorff convergence for minimizers of the adversarial training problem (3). The first step involves proving a technical lemma about the interaction between minimizers and sets $B_{\mathrm{d}}(x,R) \subset \{w_0\rho_0 - w_1\rho_1 > \delta > 0\}$. Importantly, this means $B_{\mathrm{d}}(x,R) \cap A_0 = \emptyset$ for a Bayes classifier A_0, which can help us relate minimizers of the adversarial training problem to Bayes classifiers. By applying a slicing argument, we will show that minimizers are disjoint from $B_{\mathrm{d}}(x, R/2^{d+1})$.

Lemma 3.3. *Let $A \in \mathcal{B}(\mathbb{R}^d)$ be a minimizer of the adversarial training problem (3) for $\varepsilon > 0$. Suppose there exists $x \in \mathbb{R}^d$ and $R > 0$ such that $w_0\rho_0 - w_1\rho_1 > \delta > 0$ on $B_{\mathrm{d}}(x,2R)$ with $\rho_0, \rho_1 \leq M$ on $\mathbb{R}^d$. Then, there exists a $C > 0$ independent of R, δ, ε, and x such that if $\varepsilon \leq \min\left\{R/2^{d+2}, CR\delta^{d+1}\right\}$, then $A \cap B_{\mathrm{d}}(x, R/2^{d+1}) = \emptyset$.*

Proof. Fix $\varepsilon > 0$. Choose a coordinate system such that $x = 0$ and write $B_{\mathrm{d}}(0,R) = B_{\mathrm{d}}(R)$. For the sake of contradiction, suppose there exists $z \in A \cap B_{\mathrm{d}}(R/2^{d+1})$. Then,

$$\mathcal{L}^d(A^\varepsilon \cap B_{\mathrm{d}}(R/2^d)) \geq \mathcal{L}^d(B_{\mathrm{d}}(\varepsilon)) = \omega_{\mathrm{d}}\varepsilon^d. \tag{14}$$

Corollary 2.7 shows for $r \leq R$,

$$\varepsilon \mathrm{Per}_\varepsilon(A; B_{\mathrm{d}}(r)) \leq \varepsilon \mathrm{Per}_\varepsilon(B_{\mathrm{d}}(r)^{\mathrm{c}}; A) - \delta\mathcal{L}^d(A \cap B_{\mathrm{d}}(r)) + w_0\rho_0(\widehat{U}_{11})$$

with $\widehat{U}_{11} \subset \Lambda_\varepsilon(B_{\mathrm{d}}(r)) \cap A^\varepsilon$ and $w_0\rho_0(\widehat{U}_{11}) \leq \varepsilon \mathrm{Per}_\varepsilon(B_{\mathrm{d}}(r)^{\mathrm{c}}; A^\varepsilon)$.

In particular, using the fact that $w_0\rho_0 > \delta$ in $B_{\mathrm{d}}(R)$, we obtain

$$\begin{aligned}\mathcal{L}^d((A^\varepsilon \setminus A) \cap B_{\mathrm{d}}(R)) &\leq \frac{w_0}{\delta}\rho_0((A^\varepsilon \setminus A) \cap B_{\mathrm{d}}(R)) \\ &\leq \frac{\varepsilon}{\delta}\mathrm{Per}_\varepsilon(A; B_{\mathrm{d}}(R)) \leq \frac{\varepsilon}{\delta}\mathrm{Per}_\varepsilon(B_{\mathrm{d}}(R)^{\mathrm{c}}; A) - \mathcal{L}^d(A \cap B_{\mathrm{d}}(R)) + \frac{w_0}{\delta}\rho_0(\widehat{U}_{11}).\end{aligned}$$

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

Rearranging and applying the bound $w_0\rho_0(\widehat{U}_{11}) \le \varepsilon \mathrm{Per}_\varepsilon(B_\mathrm{d}(R)^\mathsf{c}; A^\varepsilon)$, we estimate

$$\begin{aligned}\mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(R)) &\le \frac{\varepsilon}{\delta}\mathrm{Per}_\varepsilon(B_\mathrm{d}(R)^\mathsf{c}; A) + \frac{w_0}{\delta}\rho_0(\widehat{U}_{11}) \\ &\le \frac{2\varepsilon}{\delta}\mathrm{Per}_\varepsilon(B_\mathrm{d}(R)^\mathsf{c}; A^\varepsilon) \\ &\le 2\alpha\frac{R^{d-1}}{\delta}\varepsilon \end{aligned} \tag{15}$$

with the last inequality due to (13). Note $\mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(r)) \le 2\alpha\frac{r^{d-1}}{\delta}\varepsilon$ for $0 < r \le R$.

Using that ρ_0, ρ_1 are bounded from above by M, we estimate

$$\begin{aligned}\sum_{k=0}^{\lfloor \frac{R}{4\varepsilon}\rfloor -1} \mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(R/2+2k\varepsilon)) &\le \sum_{k=0}^{\lfloor \frac{R}{4\varepsilon}\rfloor -1} \frac{2\varepsilon}{\delta}\mathrm{Per}_\varepsilon(B_\mathrm{d}(R/2+2k\varepsilon)^\mathsf{c}; A^\varepsilon) \\ &\le \frac{2M}{\delta}\sum_{k=0}^{\lfloor \frac{R}{4\varepsilon}\rfloor -1} \mathcal{L}^d(A^\varepsilon \cap (B_\mathrm{d}(R/2+(2k+1)\varepsilon) \setminus B_\mathrm{d}(R/2+(2k-1)\varepsilon)) \\ &\le \frac{2M}{\delta}\mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(R)) \le 4\alpha M\frac{R^{d-1}}{\delta^2}\varepsilon \end{aligned}$$

thanks to (15).

In particular,

$$\left\lfloor \frac{R}{4\varepsilon}\right\rfloor \min_k \mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(R/2+2k\varepsilon)) \le \frac{2M}{\delta}\mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(R)) \le 4\alpha M\frac{R^{d-1}}{\delta^2}\varepsilon.$$

If $\varepsilon \le R/8$ so that $\lfloor \frac{R}{4\varepsilon}\rfloor \ge \frac{R}{4\varepsilon} - 1 \ge \frac{R}{8\varepsilon}$, then by letting $s_1 = R/2 + 2k\varepsilon$ achieve $\min_k \mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(R/2 + 2k\varepsilon))$, we then obtain

$$\mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(s_1)) \le 32\alpha M\frac{R^{d-2}}{\delta^2}\varepsilon^2.$$

Then, repeating the same construction at the scale $R/2^i$, $i \ge 2$, we find

$$\mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(s_i)) \le \frac{8\varepsilon}{R/2^{i-1}}\frac{2M}{\delta}\mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(s_{i-1})) \le \frac{2^{i+3}M}{R\delta\varepsilon}\mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(s_{i-1}))$$

as long as $\varepsilon \le \frac{R}{2^{i+2}}$ (that is, $i \le \log_2\left(\frac{R}{4\varepsilon}\right)$).

For $i = d$, it follows

$$\begin{aligned}\mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(s_d)) &\le 2^{\sum_{i=2}^d i}\left(\frac{8M\varepsilon}{R\delta}\right)^{d-1}\mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(s_1)) \\ &\le 2^{\frac{d(d+1)}{2}+3d-4}\left(\frac{M\varepsilon}{R\delta}\right)^{d-1} 32\alpha M\frac{R^{d-2}}{\delta^2}\varepsilon^2.\end{aligned}$$

Hence,

$$\mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(s_d)) \le 2^{\frac{d(d+1)}{2}+3d+1}\alpha M^d\frac{\varepsilon^{d+1}}{R\delta^{d+1}}.$$

Letting $C_{d+1} := 2^{\frac{d(d+1)}{2}+3d+1}\alpha M^d$, we conclude if $\varepsilon < \min\{R/2^{d+2}, \omega_d C_{d+1}^{-1} R\delta^{d+1}\}$, then

$$\mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(R/2^d)) \le \mathcal{L}^d(A^\varepsilon \cap B_\mathrm{d}(s_d)) \le \frac{C_{d+1}}{R\delta^{d+1}}\varepsilon^{d+1} < \omega_d\varepsilon^d$$

which implies that $A \cap B_\mathrm{d}(R/2^{d+1}) = \emptyset$ by (14). □

Remark 3.4. *In Lemma 3.3, we can slightly relax the assumption that A is a minimizer as follows: Recall that we assume $w_0\rho_0 - w_1\rho_1 > \delta > 0$ on $B_\mathrm{d}(x, 2R)$. If we have that $J_\varepsilon(A \setminus B_\mathrm{d}(x, r)) - J_\varepsilon(A) \ge 0$ for*

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

all r such that $R/2^{d+2} \le r \le R$, then the energy exchange inequality (12) still holds and the same proof for Lemma 3.3 shows that $A \cap B_{\mathrm{d}}(x, R/2^{d+1}) = \emptyset$.

We now aim to directly relate minimizers of the adversarial training problem (3) to Bayes classifiers. Recall that the maximal and minimal Bayes classifiers (2) are given by

$$A_0^{\max} = \{x \in \mathbb{R}^d : w_0\rho_0(x) \le w_1\rho_1(x)\}, \qquad A_0^{\min} = \{x \in \mathbb{R}^d : w_0\rho_0(x) < w_1\rho_1(x)\}.$$

We will *not* be assuming that $A_0^{\max}$ and $A_0^{\min}$ coincide up to a set of ρ measure zero unless explicitly stated.

We will now show that on a compact set, we can 'corral' the minimizer of the adversarial training problem (3) by any distance $\eta > 0$, in the sense that it must lie between the η-dilation of $A_0^{\max}$ and the η-erosion of $A_0^{\min}$ when ε is small enough.

Lemma 3.5. *Let $A_0^{\max}$ be the maximal Bayes classifier. Suppose that ρ_0, ρ_1 are continuous and bounded from above on $\mathbb{R}^d$, and let $\eta > 0$. Then for any compact set $K \subset \mathbb{R}^d$, there exists an $\varepsilon_0 > 0$ such that for all $0 < \varepsilon < \varepsilon_0$,*

$$\left[A_\varepsilon \cap K\right] \subset \left[(A_0^{\max})^\eta \cap K\right]$$

where $A_\varepsilon \subset \mathbb{R}^d$ is an arbitrary minimizer of the adversarial training problem.

Proof. For convenience, we abuse notation and let $A_0 = A_0^{\max}$. Assume that $\left(A_0^\eta\right)^{\mathsf{c}} \cap K \neq \emptyset$ as otherwise the result is trivial. The conditions are also trivially satisfied if $w_0\rho_0 - w_1\rho_1$ never changes sign. This is because, for all $\varepsilon > 0$, either $A_0 = A_\varepsilon = \emptyset$ if $w_0\rho_0 - w_1\rho_1 > 0$ on $\mathbb{R}^d$, or $A_0 = A_\varepsilon = \mathbb{R}^d$ otherwise.

Fix $\eta > 0$. Let $R = \frac{\eta}{3}$. Observe that $\overline{A_0^R} \cap \overline{K^{2R}}$ is compact and $A_0 \subset \overline{A_0^R}$. Then, by the continuity of $w_0\rho_0 - w_1\rho_1$ on $\overline{A_0^R} \cap \overline{K^{2R}}$, there exists a $\delta > 0$ such that

$$\left[E_\delta \cap \overline{K^{2R}}\right] \subset \left[\overline{A_0^R} \cap \overline{K^{2R}}\right]$$

where $E_\delta = \{x \in \mathbb{R}^d : w_0\rho_0(x) - w_1\rho_1(x) \le \delta\}$. This implies $\left[E_\delta^{\mathsf{c}} \cap \overline{K^{2R}}\right] \supset \left[\left(\overline{A_0^R}\right)^{\mathsf{c}} \cap \overline{K^{2R}}\right]$, so $w_0\rho_0 - w_1\rho_1 > \delta > 0$ on $\left(\overline{A_0^R}\right)^{\mathsf{c}} \cap \overline{K^{2R}}$. In particular, as $(A_0^\eta)^{\mathsf{c}} \cap K \subset \left[\left(\overline{A_0^R}\right)^{\mathsf{c}} \cap \overline{K^{2R}}\right]$, the difference in densities $w_0\rho_0 - w_1\rho_1 > \delta$ on $(A_0^\eta)^{\mathsf{c}} \cap K$.

Take $x \in (A_0^\eta)^{\mathsf{c}} \cap K$. Observe that $B_{\mathrm{d}}(x, 2R)$ satisfies the conditions of Lemma 3.3 for δ as determined previously. Take $\varepsilon_0 = \min\left\{R/2^{d+2}, CR\delta^{d+1}\right\}$ for C is independent of R, δ, ε, and x. Let $\varepsilon \le \varepsilon_0$ and let A_ε be a minimizer of the adversarial training problem (3). Then, $A_\varepsilon \cap B(x, R/2^{d+1}) = \emptyset$ for all $x \in (A_0^\eta)^{\mathsf{c}} \cap K$, which implies that $A_\varepsilon \cap (A_0^\eta)^{\mathsf{c}} \cap K = \emptyset$. Thus, we conclude

$$\left[A_\varepsilon \cap K\right] \subset \left[A_0^\eta \cap K\right].$$

□

Remark 3.6. *The only place where we use the compactness assumption in Lemma 3.5 is to determine δ from η by the continuity of $w_0\rho_0 - w_1\rho_1$ a compact set.*

The proof established that minimizers A_ε of the adversarial training problem (3) can be corralled by the maximal Bayes classifier. We can also corral A_ε by the minimal Bayes classifier as follows: Consider interchanging the densities so data points x are distributed according to $\widetilde{\rho}_0 = \rho_1$ and $\widetilde{\rho}_1 = \rho_0$. We can apply Lemma 3.5 to the minimizer $\widetilde{A}_\varepsilon = (A_\varepsilon)^{\mathsf{c}}$ of the interchanged problem. We can conclude that for all compact sets $K \subset \mathbb{R}^d$ and $\eta > 0$, there exists an $\varepsilon_0 > 0$, such that

$$\left[(A_0^{\min})^{-\eta} \cap K\right] \subset \left[A_\varepsilon \cap K\right]$$

for all $\varepsilon \le \varepsilon_0$. This means that we have a two-sided, or corralling', bound on our minimizer for ε small enough, namely

$$\left[(A_0^{\min})^{-\eta} \cap K\right] \subset \left[A_\varepsilon \cap K\right] \subset \left[(A_0^{\max})^\eta \cap K\right].$$

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

The corralling argument will allow us to examine the Hausdorff distance between Bayes classifiers and minimizers of the adversarial training problem (3) as the adversarial budget decreases to zero. To begin, we recall the definition of the Hausdorff distance.

Definition 3.7. *The Hausdorff distance between two sets $A, E \subset \mathbb{R}^d$ is given by*

$$\mathrm{d}_H(A, E) := \max\left\{\sup_{x\in A} \mathrm{d}(x, E), \sup_{x\in E} \mathrm{d}(x, A)\right\}$$

for a metric d *on* $\mathbb{R}^d$. *Furthermore,* d_H *is a pseudometric on* $\mathcal{B}(\mathbb{R}^d)$.

Remark 3.8. *If* $\mathrm{d}_H(A_0^{\max}, A_0^{\min}) = 0$, *then for any* $\eta > 0$ *and compact set* $K \subset \mathbb{R}^d$, *there exists an* $\varepsilon_0 > 0$ *such that*

$$\left[(A_0)^{-\eta} \cap K\right] \subset \left[A_\varepsilon \cap K\right] \subset \left[(A_0)^{\eta} \cap K\right]$$

for all $\varepsilon \leq \varepsilon_0$ *and for* A_0 *the unique Bayes classifier.*

We now have the tools to show the uniform convergence of minimizers A_ε of the adversarial training problem (3) to the Bayes classifier A_0. To begin, we prove the more general version of the result when the Bayes classifier is not unique up to a set of ρ measure zero. In this case, we can only show that $\lim_{\varepsilon\to 0^+} A_\varepsilon$ must be corralled by the maximal and minimal Bayes classifiers.

Theorem 3.9. *Suppose* ρ_0, ρ_1 *are continuous and bounded from above on* $\mathbb{R}^d$. *Let* $\{A_\varepsilon\}_{\varepsilon>0}$ *be a sequence of minimizers of the adversarial training problem (3) for* $\varepsilon \to 0^+$. *Then, for any compact set* $K \subset \mathbb{R}^d$,

$$\lim_{\varepsilon\to 0^+} \mathrm{d}_H((A_\varepsilon \cup A_0^{\max}) \cap K, A_0^{\max} \cap K) = 0 \quad \textit{and} \quad \lim_{\varepsilon\to 0^+} \mathrm{d}_H((A_\varepsilon \cap A_0^{\min}) \cap K, A_0^{\min} \cap K) = 0.$$

Proof. Let K be a compact set. Observe that $A_0^{\max} \subset A_\varepsilon \cup A_0^{\max}$, so

$$\mathrm{d}_H((A_\varepsilon \cup A_0^{\max}) \cap K, A_0^{\max} \cap K) = \sup_{x\in(A_\varepsilon\cup A_0^{\max})\cap K} \mathrm{d}(x, A_0^{\max} \cap K).$$

For the sake of contradiction, suppose this quantity does not go to zero as $\varepsilon \to 0^+$. Then, there exists an $\eta > 0$ such that for all $\varepsilon_0 > 0$, there exists an $0 < \varepsilon \leq \varepsilon_0$ such that

$$\sup_{x\in(A_\varepsilon\cup A_0^{\max})\cap K} \mathrm{d}(x, A_0^{\max} \cap K) > \eta.$$

However, this contradicts Lemma 3.5. Thus, we conclude

$$\lim_{\varepsilon\to 0^+} \mathrm{d}_H((A_\varepsilon \cup A_0^{\max}) \cap K, A_0^{\max} \cap K) = 0.$$

As $A_\varepsilon \cap A_0^{\min} \subset A_0^{\min}$, an analogous argument proves that

$$\lim_{\varepsilon\to 0^+} \mathrm{d}_H((A_\varepsilon \cap A_0^{\min}) \cap K, A_0^{\min} \cap K) = 0.$$

□

Corollary 3.10. *Suppose that* $\mathrm{d}_H(A_0^{\max}, A_0^{\min}) = 0$. *Then under the same assumptions as Theorem 3.9,*

$$\lim_{\varepsilon\to 0^+} \mathrm{d}_H(A_\varepsilon \cap K, A_0 \cap K) = 0$$

for A_0 *the unique Bayes classifier.*

Proof. This follows from Theorem 3.9 as the result of Lemma 3.5 simplifies when $\mathrm{d}_H(A_0^{\max}, A_0^{\min}) = 0$ as described in Remark 3.8. □

In the case where $\mathrm{d}_H(A_0^{\max}, A_0^{\min}) = 0$, it is natural to consider rates of convergence. In order to obtain such rates, we introduce the following assumption:

Assumption 3.11. *The level set* $\{w_0\rho_0 = w_1\rho_1\}$ *is non-degenerate, meaning that* $w_0\rho_1 - w_1\rho_1 \in C^1(\mathbb{R}^d)$ *and* $|w_0\nabla\rho_0 - w_1\nabla\rho_1| > \alpha > 0$ *on* $\{w_0\rho_0 = w_1\rho_1\}$ *for some constant* α. *In this case, Bayes classifiers are unique up to a set of* $\mathcal{L}^d$ *measure zero and* $\mathrm{d}_H(A_0^{\max}, A_0^{\min}) = 0$.

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

Now, we establish the convergence rate for minimizers of the adversarial training problem (3) to Bayes classifiers under this non-degeneracy assumption.

Corollary 3.12. *Suppose Assumption 3.11 holds and that ρ_0, ρ_1 are continuous and bounded from above on $\mathbb{R}^d$. For any compact set $K \subset \mathbb{R}^d$, there exists a constant $C > 0$ such that*

$$\limsup_{\varepsilon \to 0^+} \frac{\mathrm{d}_H(A_\varepsilon \cap K, A_0 \cap K)}{\varepsilon^{\frac{1}{d+2}}} \leq C$$

where A_0 is the Bayes classifier.

Proof. Consider a sequence $\{\eta_i\}_{i \in \mathbb{N}}$ where $\eta_i > 0$ and $\eta_i \to 0^+$. Define $\varepsilon_i = \min\{C\eta_i, C\eta_i \delta_i^{d+1}\}$ based on the requirements on ε from Lemma 3.3 with $R = \eta_i$ and the continuity bound $\delta = \delta_i$ from Lemma 3.5. In this proof, C is a constant always independent of η_i, ε_i, and δ_i that we will allow to vary throughout this proof.

As $w_0\rho_0 - w_1\rho_1 \in C^1(\mathbb{R}^d)$ and its gradient is bounded away from 0, the boundary $\partial A_0 = \{w_0\rho_0 = w_1\rho_1\}$ is a C^1 surface by the implicit function theorem, and hence the Hausdorff distance between the minimal and maximal sets is zero. Furthermore for $\eta_i \ll 1$, δ_i is the same order as η_i, which implies $\varepsilon_i = C\eta_i^{d+2}$.

For each ε_i, let A_{ε_i} be the associated minimizer of the adversarial training problem (3). By Theorem 3.9 along with Remark 3.10, for any compact set $K \subset \mathbb{R}^d$ we have that

$$\mathrm{d}_H(A_{\varepsilon_i} \cap K, A_0 \cap K) < \eta_i = C\varepsilon_i^{\frac{1}{d+2}}.$$

Thus, we conclude that

$$\limsup_{\varepsilon_i \to 0^+} \frac{\mathrm{d}_H(A_{\varepsilon_i} \cap K, A_0 \cap K)}{\varepsilon_i^{\frac{1}{d+2}}} \leq C.$$

□

Remark 3.13. *Although we have shown the convergence rate to be at most $O(\varepsilon^{\frac{1}{d+2}})$, we expect that the convergence rate is actually $O(\varepsilon)$ (see the formal asymptotics near $\varepsilon = 0$ derived by [16]). The reason we get the convergence rate $\varepsilon^{\frac{1}{d+2}}$ is from the δ^{d+1} that appears in our bounds for ε. In Lemma 3.3, this term comes from the iterative argument that often employs crude volume bounds. More precise estimates would be required to improve the convergence rate.*

4. Uniform convergence for other deterministic attacks

Now, we will turn our focus to the generalized adversarial training problem (8). At the end, we will present the results for the probabilistic adversarial training problem (5) as an example of our results for (8). Unlike the case of the adversarial training problem (3), existence of minimizers to (8) is an open question, and in this case, our convergence result can be understood in the spirit of 'a priori' estimates in partial differential equations. First, we will make it precise which deterministic attack functions we consider.

Definition 4.1. *A deterministic attack function ϕ is metric if an adversary's attack on x only depends upon points within distance ε of x for some adversarial budget $\varepsilon > 0$. More precisely for two classifiers $A, \widetilde{A} \in \mathcal{B}(\mathbb{R}^d)$,*

$$A \cap B_{\mathrm{d}}(x, \varepsilon) = \widetilde{A} \cap B_{\mathrm{d}}(x, \varepsilon) \implies \phi(x; A) = \phi(x; \widetilde{A}).$$

To avoid a trivial situation where x is always attacked independent of the choice of A, we assume the adversary has no power, meaning $\phi(x; A) \equiv 0$, if $A = \emptyset$ or $A = \mathbb{R}^d$ when ϕ is a metric attack function.

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

In the following pair of lemmas, we will show two important properties of metric attack functions. The first will allow us to relate D_ϕ with $\varepsilon\mathrm{Per}_\varepsilon$ and provides an upper bound on D_ϕ by $\varepsilon\mathrm{Per}_\varepsilon$. This will allow us to employ many of the estimates of $\varepsilon\mathrm{Per}_\varepsilon$ from Lemma 3.3 in Lemma 4.6.

Lemma 4.2. *Let ϕ be a metric deterministic attack function. For any set $A, E \in \mathcal{B}(\mathbb{R}^d)$, we have that*

$$D_\phi(A) \leq \varepsilon Per_\varepsilon(A) \;\; \text{and} \;\; D_\phi(A; E) \leq \varepsilon Per_\varepsilon(A; E).$$

Proof. It will be sufficient to show that $\Lambda^i_\phi(A) \subset \Lambda^i_\varepsilon(A)$. Take $x \in \Lambda^0_\phi(A)$. If we consider $\widetilde{A} = \emptyset$, the metric property states

$$A \cap B_{\mathrm{d}}(x, \varepsilon) = \emptyset \implies \phi(x; A) = \phi(x; \emptyset) = 0.$$

For $x \in A^{\mathsf{c}}$, $A \cap B_{\mathrm{d}}(x, \varepsilon) \neq \emptyset$ implies $x \in A^\varepsilon \setminus A = \Lambda^0_\varepsilon(A)$. Thus, we conclude $\Lambda^0_\phi(A) \subset \Lambda^0_\varepsilon(A)$. A similar argument with $\widetilde{A} = \mathbb{R}^d$ shows that $\Lambda^1_\phi(A) \subset \Lambda^1_\varepsilon(A)$. □

We now prove a second property of metric attack functions, which isolates where the values of $\phi(x; A)$ and $\phi(x; A \setminus E)$ may differ.

Lemma 4.3. *Let ϕ be a metric deterministic attack function. For sets $A, E \in \mathcal{B}(\mathbb{R}^d)$, if $x \in (E^\varepsilon)^{\mathsf{c}}$, then $\phi(x; A) = \phi(x; A \setminus E)$.*

Proof. Suppose $x \in (E^\varepsilon)^{\mathsf{c}}$. Then, $B(x, \varepsilon) \subset E^{\mathsf{c}}$ and so $A \cap B(x, \varepsilon) = (A \setminus E) \cap B(x, \varepsilon)$. Hence, the metric property then implies that $\phi(x; A) = \phi(x; A \setminus E)$. □

We require one additional assumption on a metric attack function ϕ in order to prove the generalized version of Lemma 3.3. Namely, if the size of the intersection of $B_{\mathrm{d}}(x, \varepsilon)$ with the opposite class of x satisfies a lower bound, then $x \in \Lambda_\phi(A)$.

Assumption 4.4. *Let ϕ be a metric deterministic attack function with budget $\varepsilon > 0$. For a classifier $A \in \mathcal{B}(\mathbb{R}^d)$, we assume:*

$$x \in A^{\mathsf{c}} \text{ and } \mathcal{L}^d(A \cap B_{\mathrm{d}}(x, \varepsilon)) > \beta\varepsilon^d \implies x \in \Lambda^0_\phi(A),$$

$$x \in A \text{ and } \mathcal{L}^d(A^{\mathsf{c}} \cap B_{\mathrm{d}}(x, \varepsilon)) > \beta\varepsilon^d \implies x \in \Lambda^1_\phi(A),$$

for some constant $0 < \beta < \omega_{\mathrm{d}}$ independent of x, ε, and A.

As a consequence of this assumption, we have if $x \in \tilde{\Lambda}^0_\phi(A)$, then $\mathcal{L}^d(A \cap B_{\mathrm{d}}(x, \varepsilon)) \leq \beta\varepsilon^d$. Likewise, if $x \in \tilde{\Lambda}^1_\phi(A)$, then $\mathcal{L}^d(A^{\mathsf{c}} \cap B_{\mathrm{d}}(x, \varepsilon)) \leq \beta\varepsilon^d$. Furthermore, if Assumption 2.1 also holds for ϕ, then only one of the two lower bounds needs to be assumed, as the other follows by the complement property.

Remark 4.5. *This assumption states that a point $x \in \mathbb{R}^d$ is attacked if the portion of its ε-neighbours with the opposite label is on the order of ε^d. In this way, the deterministic attack function depends on the adversarial budget ε and the metric.*

Observe that the adversarial training problem (3) satisfies Assumption 4.4. In fact, it satisfies the statements

$$x \in \Lambda^0_\varepsilon(A) \iff x \in A^{\mathsf{c}} \text{ and } A \cap B_{\mathrm{d}}(x, \varepsilon) \neq \emptyset,$$

$$x \in \Lambda^1_\varepsilon(A) \iff x \in A \text{ and } A^{\mathsf{c}} \cap B_{\mathrm{d}}(x, \varepsilon) \neq \emptyset.$$

In Proposition 4.15, we will verify that the probabilistic adversarial training problem (5) also satisfies Assumption 4.4.

In order to show uniform convergence for the generalized adversarial training problem (8), we first prove the analogue of Lemma 3.3 by a similar slicing argument. We leverage the relationship between the adversarial deficit and the ε-perimeter established in Lemma 4.2. However, there are a few key differences in both the results and the proof. Whereas in Lemma 3.3, we show that minimizers of the adversarial training problem (3) are disjoint from certain norm balls that are misclassified, we show that

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

the intersection of minimizers of (8) with a misclassified norm ball must have $\mathcal{L}^d$ measure zero. In this sense, we establish a necessary condition for minimizers of (8). As for the proof of the statement, the final step differs significantly between Lemmas 3.3 and 4.6. In final step of Lemma 3.3, we are able to use the fact that a single point causes misclassification on the order of ε^d. For the general case, the lower bound on the $\mathcal{L}^d$ measure condition for misclassification from Assumption 4.4 requires a more delicate energy argument that examines the exact difference in energies.

Lemma 4.6. *Let ϕ be a metric deterministic attack function for $\varepsilon > 0$, satisfying Assumptions 2.1 and 4.4. Suppose ρ_0, ρ_1 are continuous and bounded from above on $\mathbb{R}^d$. Furthermore, suppose $A \in \mathcal{B}(\mathbb{R}^d)$ is a minimizer of the generalized training problem (8) and there exists $x \in \mathbb{R}^d$ and $R > 0$ such that $w_0\rho_0 - w_1\rho_1 > \delta > 0$ on $B_{\mathrm{d}}(x, 2R)$. Then, there exists a constant $C > 0$ independent of R, δ, ε, and x such that if $\varepsilon \leq \min\left\{R/2^{d+2}, CR\delta^{d+1}\right\}$, then $\mathcal{L}^d(A \cap B(x, R/2^{d+2})) = 0$.*

Proof. Choose a coordinate system such that $x = 0$ and write $B_{\mathrm{d}}(0, R) = B_{\mathrm{d}}(R)$ with x as in the statement above.

We will first find an initial estimate for $\mathcal{L}^d(A \cap B_{\mathrm{d}}(R))$. As A is a minimizer and $w_0\rho_0 - w_1\rho_1 > \delta > 0$ on $B_{\mathrm{d}}(R)$, we can apply Proposition 2.5 to find that

$$D_\phi(A; B_{\mathrm{d}}(R)) \leq D_\phi(B_{\mathrm{d}}(R)^{\mathrm{c}}; A) - \delta\mathcal{L}^d(A \cap B_{\mathrm{d}}(R)) + w_0\rho_0(\widehat{U}_{11}) + w_1\rho_1(\widehat{U}_1) \tag{16}$$

where

$$\widehat{U}_1 = \{x \in \tilde{\Lambda}^1_\phi(A) \cap \tilde{\Lambda}^0_\phi(B_{\mathrm{d}}(R)) : \phi(x; A \setminus B_{\mathrm{d}}(R)) = 1\},$$

$$\widehat{U}_{11} = \{x \in \Lambda^0_\phi(A) \cap \Lambda^1_\phi(B_{\mathrm{d}}(R)) : \phi(x; A \setminus B_{\mathrm{d}}(R)) = 1\}.$$

By (11), we have $w_0\rho_0(\widehat{U}_{11}) \leq D_\phi(A; B_{\mathrm{d}}(R))$. Combining the upper bound on $w_0\rho_0(\widehat{U}_{11})$ with (16) and simplifying, we find

$$\mathcal{L}^d(A \cap B_{\mathrm{d}}(R)) \leq \frac{1}{\delta} D_\phi(B_{\mathrm{d}}(R)^{\mathrm{c}}; A) + \frac{w_1}{\delta}\rho_1(\widehat{U}_1).$$

Recall that by definition, $\widehat{U}_1 \subset A \cap B_{\mathrm{d}}(R)^{\mathrm{c}}$. Additionally by Lemma 4.3, $\widehat{U}_1 \subset B_{\mathrm{d}}(R + \varepsilon)$ as $\phi(x; A \setminus B_{\mathrm{d}}(R)) = 1$, while $\phi(x; A) = 0$. Thus, $\widehat{U}_1 \subset A \cap (B_{\mathrm{d}}(R + \varepsilon) \setminus B_{\mathrm{d}}(R))$. In particular,

$$w_1\rho_1(\widehat{U}_1) \leq w_1\rho_1(A \cap (B_{\mathrm{d}}(R + \varepsilon) \setminus B_{\mathrm{d}}(R))) \leq \varepsilon \mathrm{Per}_\varepsilon(B_{\mathrm{d}}(R)^{\mathrm{c}}; A).$$

Additionally, by Lemma 4.2, we have $D_\phi(B_{\mathrm{d}}(R)^{\mathrm{c}}; A) \leq \varepsilon \mathrm{Per}_\varepsilon(B_{\mathrm{d}}(R)^{\mathrm{c}}; A)$. Applying (15) from Lemma 3.2,

$$\mathcal{L}^d(A \cap B_{\mathrm{d}}(R)) \leq \frac{1}{\delta} D_\phi(B_{\mathrm{d}}(R)^{\mathrm{c}}; A) + \frac{w_1}{\delta}\rho_1(\widehat{U}_1) \leq \frac{2}{\delta}\varepsilon \mathrm{Per}_\varepsilon(B_{\mathrm{d}}(R)^{\mathrm{c}}; A) \leq \frac{2\alpha R^{d-1}}{\delta}\varepsilon$$

for α independent of R, δ, ε, and x as in Lemma 3.2.

Next, we want to find a radius $s_1 \in (R/2, R)$ that will give an order ε^2 estimate for $\mathcal{L}^d(A \cap B_{\mathrm{d}}(s_1))$. For $r \leq R$, one has

$$\mathcal{L}^d(A \cap B_{\mathrm{d}}(r)) \leq \frac{2}{\delta}\varepsilon \mathrm{Per}_\varepsilon(B_{\mathrm{d}}(r)^{\mathrm{c}}; A).$$

We can argue by a discrete slicing argument like in Lemma 3.3 to show that there exists an $s_1 \in (R/2, R)$ such that

$$\mathcal{L}^d(A \cap B_{\mathrm{d}}(s_1)) \leq \frac{2}{\delta}\varepsilon \mathrm{Per}_\varepsilon(B_{\mathrm{d}}(s_1)^{\mathrm{c}}; A) \leq 32\alpha M \frac{R^{d-2}}{\delta^2}\varepsilon^2.$$

Iterating the argument as in Lemma 3.3 yields an order ε^{i+1} estimate of $\mathcal{L}^d(A \cap B_{\mathrm{d}}(s_i))$ for $s_i \in (R/2^i, R/2^{i-1})$ and $2 \leq i \leq \log_2(\frac{R}{4\varepsilon})$ (i.e. $\varepsilon \leq \frac{R}{2^{i+2}}$). After d iterations, we find

$$\mathcal{L}^d(A \cap B_{\mathrm{d}}(R/2^d)) \leq \left(\frac{C_{d+1}}{R\delta^{d+1}}\right)\varepsilon^{d+1},$$

where $C_{d+1} := 2^{\frac{d(d+1)}{2}+3d+1}\alpha M^d$.

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

Finally, we must show that $\mathcal{L}^d(A \cap B(R/2^{d+2})) = 0$. Let $z \in A \cap B_{\mathrm{d}}(\frac{R}{2^{d+1}} + \varepsilon)$. We must consider a region slightly outside of $B_{\mathrm{d}}(R/2^{d+1})$ as the following argument needs to apply all points in the ε-dilation of $B_{\mathrm{d}}(R/2^{d+1})$. We want to show that $z \in \Lambda^1_\phi(A)$. To do so, by Assumption 5, it will suffice to show that if $z \in A$, then $\mathcal{L}^d(A^{\mathrm{c}} \cap B(z, \varepsilon)) > \beta\varepsilon^d$.

Recall the estimate from the previous steps, $\mathcal{L}^d(A \cap B_{\mathrm{d}}(R/2^d)) \leq \left(\frac{C_{d+1}}{R\delta^{d+1}}\right)\varepsilon^{d+1}$. As long as $\left(\frac{C_{d+1}}{R\delta^{d+1}}\right)\varepsilon < (\omega_{\mathrm{d}} - \beta)$, or in other words $\varepsilon < (\omega_{\mathrm{d}} - \beta)\left(\frac{R\delta^{d+1}}{C_{d+1}}\right) := C$, then we have

$$\mathcal{L}^d(A \cap B_{\mathrm{d}}(R/2^d)) < (\omega_{\mathrm{d}} - \beta)\varepsilon^d$$

where β is the constant from Assumption 4.4. Then, as $B_{\mathrm{d}}(z, \varepsilon) \subset B_{\mathrm{d}}(R/2^d)$, we estimate

$$\mathcal{L}^d(A \cap B_{\mathrm{d}}(z, \varepsilon)) + \mathcal{L}^d(A^{\mathrm{c}} \cap B_{\mathrm{d}}(z, \varepsilon)) = \omega_{\mathrm{d}}\varepsilon^d \implies \mathcal{L}^d(A^{\mathrm{c}} \cap B_{\mathrm{d}}(z, \varepsilon)) > \beta\varepsilon^d.$$

Hence, $z \in \Lambda^1_\phi(A)$. In particular, this means that $\tilde{\Lambda}^1_\phi(A) \cap B_{\mathrm{d}}(\frac{R}{2^{d+1}} + \varepsilon) = \emptyset$.

We will now examine the difference in energies after removing $B_{\mathrm{d}}(R/2^{d+1})$ in order to show that we must actually remove $B_{\mathrm{d}}(R/2^{d+2})$ in order to achieve $\mathcal{L}^d(A \cap B_{\mathrm{d}}(R/2^{d+2})) = 0$. By Corollary 2.6, the difference in energy after removing the set $E = B_{\mathrm{d}}(R/2^{d+1})$ from A is

$$\begin{aligned} J_\phi(A \setminus B_{\mathrm{d}}(R/2^{d+1})) - J_\phi(A) &= w_1\rho_1(\widehat{U}_1 \cup U_2 \cup \widehat{U}_3) - (w_0\rho_0 - w_1\rho_1)(\widetilde{U}_3 \cup U_4) \\ &\quad - w_0\rho_0(U_5 \cup \widetilde{U}_6 \cup \widetilde{U}_9 \cup \widetilde{U}_{10} \cup \widetilde{U}_{11} \cup U_{12}), \end{aligned}$$

where all sets are as defined in Table 1 and (9). By construction,

$$[\widehat{U}_1 \cup U_2 \cup \widehat{U}_3] \subset \left[\tilde{\Lambda}^1_\phi(A) \cap B_{\mathrm{d}}\left(\frac{R}{2^{d+1}} + \varepsilon\right)\right].$$

However, we have just shown that $\tilde{\Lambda}^1_\phi(A) \cap B_{\mathrm{d}}(\frac{R}{2^{d+1}} + \varepsilon) = \emptyset$. Thus, we conclude that $\widehat{U}_1 = U_2 = \widehat{U}_3 = \emptyset$.

As $w_0\rho_0 - w_1\rho_1 > \delta > 0$ on $B_{\mathrm{d}}(R/2^{d+1})$, the difference in energies becomes

$$\begin{aligned} J_\phi(A \setminus B_{\mathrm{d}}(R/2^{d+1})) - J_\phi(A) &= -(w_0\rho_0 - w_1\rho_1)(\widetilde{U}_3 \cup U_4) - w_0\rho_0(U_5 \cup \widetilde{U}_6 \cup \widetilde{U}_9 \cup \widetilde{U}_{10} \cup \widetilde{U}_{11} \cup U_{12}) \\ &\leq -\delta\mathcal{L}^d(\widetilde{U}_3 \cup U_4) - \delta\mathcal{L}^d(U_5 \cup \widetilde{U}_6 \cup \widetilde{U}_9 \cup \widetilde{U}_{10} \cup \widetilde{U}_{11} \cup U_{12}) \\ &\leq 0. \end{aligned}$$

By our assumption, A is a minimizer, so $J_\phi(A \setminus B_{\mathrm{d}}(R/2^{d+1})) - J_\phi(A) = 0$. This means all remaining sets must have measure zero, i.e.

$$\mathcal{L}^d(\widetilde{U}_3) = \mathcal{L}^d(U_4) = \mathcal{L}^d(U_5) = \mathcal{L}^d(\widetilde{U}_6) = \mathcal{L}^d(\widetilde{U}_9) = \mathcal{L}^d(\widetilde{U}_{10}) = \mathcal{L}^d(\widetilde{U}_{11}) = \mathcal{L}^d(U_{12}) = 0.$$

Recall from (10) that $A \cap B_{\mathrm{d}}(R/2^{d+1}) = U_3 \cup U_4 \cup U_5 \cup U_6$. We have already shown that $U_3 = \widetilde{U}_3 \cup \widehat{U}_3$, U_4 and U_5 all have measure zero. However, we notice that $\widehat{U}_6 \subset B(\frac{R}{2^{d+1}} + \varepsilon) \setminus B(\frac{R}{2^{d+1}} - \varepsilon)$, and so we can conclude that $\mathcal{L}^d(A \cap B(\frac{R}{2^{d+1}} - \varepsilon)) = 0$.

Then combining with the facts about U_1, U_2, and U_3, we then get that for any $s < \frac{R}{2^{d+1}} - \varepsilon$ we have that $A \setminus B_{\mathrm{d}}(s)$ is a minimizer of (8) and that $A \cap B_{\mathrm{d}}(s)$ has measure zero. □

Remark 4.7. *As stated at the end of the proof, $A \setminus B_{\mathrm{d}}(x, R/2^{d+2})$ is also a minimizer of (8). In addition to providing a necessary condition for minimizers, Lemma 4.6 also gives a construction for a minimizer that is disjoint from $B_{\mathrm{d}}(x, R/2^{d+2})$.*

Considering the assumptions, we cannot relax the assumption that A is a minimizer to $J_\phi(A \setminus B_{\mathrm{d}}(x, r)) - J_\phi(A) \geq 0$ as we could for Lemma 3.3 (see Remark 3.4). Although the energy exchange inequality will still hold, we require that A is a minimizer of (8) to show $\mathcal{L}^d(A \cap B_{\mathrm{d}}(R/2^{d+2})) = 0$.

Assuming a minimizer to (8) exists, Lemma 4.6 allows us to show that minimizers are (a.e.) disjoint from certain sets where it is energetically advantageous to be assigned label 0 by the classifier. In Lemma 4.8, we will use this result to show that for a prescribed distance $\eta > 0$, there exists a minimizer of (8) that can be corralled to be within distance η of any Bayes classifier for all ε smaller than some threshold. This is the generalized version of Lemma 3.5. As we cannot expect minimizers of (8) to be sensitive

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

to modification by a $\mathcal{L}^d$ measure zero set, we do not expect arbitrary minimizers to have this property. However, from an arbitrary minimizer, Lemma 4.8 provides a method to construct a $\mathcal{L}^d$-a.e. equivalent minimizer that does satisfy this distance condition.

Lemma 4.8. *Let $A_0^{\max}$ be the maximal Bayes classifier, i.e. $A_0^{\max} = \{x \in \mathbb{R}^d : w_0\rho_0(x) \leq w_1\rho_1(x)\}$. Suppose $\rho_0, \rho_1 > 0$ and continuous and bounded from above on $\mathbb{R}^d$. Let K be a compact set and fix $\eta > 0$. Then, there exists an $\varepsilon_0 > 0$ such that for any $0 < \varepsilon \leq \varepsilon_0$ and deterministic attack function ϕ satisfying Assumptions 2.1 and 4.4 for adversarial budget ε such that for any minimizer $A_{\varepsilon,\phi}$ of the generalized adversarial training problem (8) there exist a $\mathcal{L}^d$ measure zero set $N^{\max} \in \mathcal{B}(\mathbb{R}^d)$ such that*

$$\left[(A_{\varepsilon,\phi} \setminus N^{\max}) \cap K\right] \subset \left[(A_0^{\max})^{\eta} \cap K\right].$$

Furthermore, $A_{\varepsilon,\phi} \setminus N^{\max}$ is also a minimizer of (8).

Proof. We will follow the proof of Lemma 3.5. We again abuse notation and let $A_0 = A_0^{\max}$. Assume that $(A_0^{\eta})^{\mathsf{c}} \cap K \neq \emptyset$ as otherwise the result is trivial. The conditions are also trivially satisfied if $w_0\rho_0 - w_1\rho_1$ never changes sign.

Let $R = \frac{\eta}{3}$. By the same argument as in Lemma 3.5, the continuity of $w_0\rho_0 - w_1\rho_1$ on the compact set $\overline{A_0^R} \cap \overline{K^{2R}}$ allows us to conclude that there exists a $\delta > 0$ such that $w_0\rho_0 - w_1\rho_1 > \delta$ on $\left(\overline{A_0^R}\right)^{\mathsf{c}} \cap \overline{K^{2R}}$ and $(A_0^{\eta})^{\mathsf{c}} \cap K$.

As $(A_0^{\eta})^{\mathsf{c}} \cap K$ is compact, there exists a finite covering of $(A_0^{\eta})^{\mathsf{c}} \cap K$ by $\{B_{\mathrm{d}}(x_i, R/2^{d+2})\}_{1 \leq i \leq n}$ for some $n \in \mathbb{N}$ such that

$$\bigcup_{i=1}^{n} B_{\mathrm{d}}(x_i, 2R) \subset \left[E_{\delta}^{\mathsf{c}} \cap \overline{K^{2R}}\right]$$

where $E_{\delta} = \{x \in \mathbb{R}^d : w_0\rho_0(x) - w_1\rho_1(x) \leq \delta\}$. Hence, each $B_{\mathrm{d}}(x_i, 2R)$ satisfies the conditions of Lemma 4.6 for δ from the continuity bound. As the constant C from Lemma 4.6 is independent of x, we can let $\varepsilon_0 = \min\left\{R/2^{d+2}, CR\delta^{d+1}\right\}$.

Suppose $A_{\varepsilon,\phi}$ is a minimizer of the generalized adversarial training problem (8) for some $0 < \varepsilon \leq \varepsilon_0$ and let $N^{\max} = \bigcup_{i=1}^{n} [A_{\varepsilon,\phi} \cap B_{\mathrm{d}}(x_i, R/2^{d+2})]$. Then,

$$\mathcal{L}^d\left(\bigcup_{i=1}^{n}\left[A_{\varepsilon,\phi} \cap B_{\mathrm{d}}(x_i, R/2^{d+2})\right]\right) \leq \sum_{i=1}^{n} \mathcal{L}^d(A_{\varepsilon,\phi} \cap B_{\mathrm{d}}(x_i, R/2^{d+2})) = 0,$$

so $N^{\max}$ is a $\mathcal{L}^d$ measure zero set. By Remark 4.7, an iterative application of Lemma 4.6, removing one norm ball at a time, ensures that $A_{\varepsilon,\phi} \setminus N^{\max}$ is a minimizer of (8). Furthermore, $[A_{\varepsilon,\phi} \setminus N^{\max}] \cap [(A_0^{\eta})^{\mathsf{c}} \cap K] = \emptyset$ by construction which implies

$$\left[(A_{\varepsilon,\phi} \setminus N^{\max}) \cap K\right] \subset \left[A_0^{\eta} \cap K\right].$$

□

Remark 4.9. *In Lemma 4.6, we require compactness both for the continuity argument and for the finite covering argument to ensure that we are removing a set of $\mathcal{L}^d$ measure zero. Compare this with Lemma 3.5 and Remark 3.6.*

We can analogously show that (up to a set of $\mathcal{L}^d$ measure zero) we can corral $A_{\varepsilon,\phi}$ by an η-erosion of the minimal Bayes classifier $A_0^{\min}$ by considering the flipped density problem. We can apply the result from Lemma 4.8 to conclude that on a compact set $K \subset \mathbb{R}^d$ for a fixed $\eta > 0$, there exists a $\varepsilon_0 > 0$ such that for $0 < \varepsilon \leq \varepsilon_0$ and a deterministic attack function ϕ with adversarial budget ε satisfying the appropriate assumptions, then for any minimizer $A_{\varepsilon,\phi}$ of (8) there exist a $\mathcal{L}^d$ measure zero set $N^{\min}$ such that

$$\left[(A_0^{\min})^{-\eta} \cap K\right] \subset \left[(A_{\varepsilon,\phi} \cup N^{\min}) \cap K\right].$$

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

Observe that by construction $N^{\max} \subset A_0^{\mathsf{c}}$ and $N^{\min} \subset A_0$ so the two sets are disjoint. Like in the previous case, this establishes a two-sided, 'corralling' bound on any minimizer for ε small enough, namely,

$$\left[(A_0^{\min})^{-\eta} \cap K\right] \subset \left[(A_{\varepsilon,\phi} \cup N^{\min} \setminus N^{\max}) \cap K\right] \subset \left[(A_0^{\max})^{\eta} \cap K\right].$$

Remark 4.10. *If the Bayes classifier is unique in the sense of Remark 1.1, then for any $\eta > 0$ and compact set $K \subset \mathbb{R}^d$, there exists an $\varepsilon_0 > 0$ such that for all $0 < \varepsilon \leq \varepsilon_0$ and ϕ satisfying Assumptions 2-1 and 4.4 for adversarial budget ε,*

$$\left[(A_0)^{-\eta} \cap K\right] \subset \left[(A_{\varepsilon,\phi} \cup N^{\min} \setminus N^{\max}) \cap K\right] \subset \left[(A_0)^{\eta} \cap K\right],$$

provided that $A_{\varepsilon,\phi}$ exists.

Following the sequence of proofs in Section 3, we will now use the corralling result from Lemma 4.8 to examine the distance between minimizers of the generalized adversarial training problem (8) and Bayes classifiers. The next theorem is the generalization of Theorem 3.9 and establishes uniform convergence in the Hausdorff distance. As previously stated, there is currently no proof of existence for minimizers of (8), so this result should be seen as a type of a priori uniform convergence estimate.

Theorem 4.11. *Let ϕ be a deterministic attack function satisfying Assumptions 2.1 and 4.4. Suppose ρ_0, ρ_1 are continuous and bounded from above on $\mathbb{R}^d$. Additionally, suppose $\{A_{\varepsilon_i,\phi}\}_{i\in\mathbb{N}}$ is a sequence of minimizers of the generalized adversarial training problem (8) with $\varepsilon_i \to 0^+$ as $i \to \infty$. For any compact set $K \subset \mathbb{R}^d$, there exist sequences $\{N_i^{\min}\}_{i\in\mathbb{N}}$ and $\{N_i^{\max}\}_{i\in\mathbb{N}}$ of $\mathcal{L}^d$ measure zero sets such that*

$$\lim_{i\to\infty} \mathrm{d}_H\big(((A_{\varepsilon_i,\phi} \setminus N_i^{\max}) \cup A_0^{\max}) \cap K, A_0^{\max} \cap K\big) = 0$$

and

$$\lim_{i\to\infty} \mathrm{d}_H\big(((A_{\varepsilon_i,\phi} \cup N_i^{\min}) \cap A_0^{\min}) \cap K, A_0^{\min} \cap K\big) = 0.$$

Proof. The proof is identical to that of Theorem 3.9, where $N_i^{\max}$ and $N_i^{\min}$ are as defined in the proof of Lemma 4.8. □

Corollary 4.12. *If the Bayes classifier A_0 is unique in the sense of Remark 1.1, then under the same assumptions as Theorem 3.9,*

$$\lim_{i\to\infty} \mathrm{d}_H((A_{\varepsilon_i,\phi} \cup N_i^{\min} \setminus N_i^{\max}) \cap K, A_0 \cap K) = 0.$$

Proof. This follows directly from Theorem 4.11. □

Recall that Assumption 3.11 is a non-degeneracy assumption on the Bayes classifier A_0 that ensures that $\mathrm{d}_H(A_0^{\max}, A_0^{\min}) = 0$ and that A_0 is unique up to a set of $\mathcal{L}^d$ measure zero. If we assume that the Bayes classifier is non-degenerate, then it becomes natural to examine the rates of convergence.

Corollary 4.13. *Let ϕ be a deterministic attack function satisfying Assumptions 2.1 and 4.4. Suppose Assumption 3.11 holds and that for every $\varepsilon > 0$, there exists a minimizer $A_{\varepsilon,\phi}$ to the generalized adversarial training problem (8). Additionally, suppose ρ_0, ρ_1 are continuous and bounded from above on $\mathbb{R}^d$. For any compact set $K \subset \mathbb{R}^d$, there exist sequences $\{N_i^{\min}\}_{i\in\mathbb{N}}$ and $\{N_i^{\max}\}_{i\in\mathbb{N}}$ of $\mathcal{L}^d$ measure zero sets and a constant $C > 0$ such that*

$$\limsup_{i\to\infty} \frac{\mathrm{d}_H((A_{\varepsilon_i,\phi} \cup N_i^{\min} \setminus N_i^{\max}) \cap K, A_0 \cap K)}{\varepsilon_i^{\frac{1}{d+2}}} \leq C$$

where A_0 is the Bayes classifier.

Proof. This proof is identical to that of Corollary 3.12. □

Remark 4.14. *As in Lemma 3.12, we expect that the convergence rate for minimizers of the generalized adversarial training problem (8) should be improved to $O(\varepsilon)$, but this would require more refined estimates than those available in Lemma 4.6.*

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

4.1. Application to the probabilistic adversarial training problem

We now turn our attention to the probabilistic adversarial training problem (5), which we will view as an instance of the generalized adversarial training problem (8). In order to apply the results for (8) to (5), we must verify that Assumptions 2.1 and 4.4 hold. In Remark 2.3, we established that (5) satisfies Assumption 2.1; thus, it only remains to show in the following proposition that (5) satisfies Assumption 4.4.

Proposition 4.15. *Let $\varepsilon > 0$, $p \in [0,1)$ and $\{\mathfrak{p}_{x,\varepsilon}\}_{x\in\mathbb{R}^d}$ be a family of probability measures satisfying Assumption 1.3. The deterministic attack function $\phi_{\varepsilon,p}$ associated with the probabilistic adversarial training problem (5) satisfies Assumption 4.4.*

Proof. Suppose $x \in A^{\mathsf{c}}$ such that $\mathcal{L}^d(A \cap B_{\mathrm{d}}(x,\varepsilon)) > \beta\varepsilon^d$ for some $\beta > 0$ to be determined. It will be sufficient to show that $\mathbb{P}(x' \in A : x' \sim \mathfrak{p}_{x,\varepsilon}) > p$. Recall that we can express

$$\begin{aligned}\mathbb{P}(x' \in A : x' \sim \mathfrak{p}_{x,\varepsilon}) &= \int_{\mathbb{R}^d} \varepsilon^{-d} \mathbb{1}_A(x') \xi\left(\frac{x'-x}{\varepsilon}\right) dx' \\ &= \int_{A\cap B_{\mathrm{d}}(x,\varepsilon)} \varepsilon^{-d} \xi\left(\frac{x'-x}{\varepsilon}\right) dx' \\ &> c\varepsilon^{-d}\mathcal{L}^d(A \cap B_{\mathrm{d}}(x,\varepsilon))\end{aligned}$$

where $c > 0$ is the lower bound on ξ from Assumption 1.3. If $\beta = \frac{p}{c}$, then $\mathbb{P}(x' \in A : x' \sim \mathfrak{p}_{x,\varepsilon}) > p$ as desired. As the probabilistic adversarial training problem (5) satisfies the complement property, this is sufficient to conclude that Assumption 4.4 holds for $\beta = \frac{p}{c}$. □

Since the probabilistic adversarial training problem (5) satisfies the requisite assumptions, we can state the following convergence result.

Theorem 4.16. *Suppose ρ_0, ρ_1 are continuous and bounded from above on $\mathbb{R}^d$ and fix $p \in [0,1)$. Additionally, suppose $\{A_{\varepsilon_i,p}\}_{i\in\mathbb{N}}$ is a sequence of minimizers of the probabilistic adversarial training problem (5) with $\varepsilon_i \to 0^+$ as $i \to \infty$. For any compact set $K \subset \mathbb{R}^d$, there exist sequences $\{N_i^{\min}\}_{i\in\mathbb{N}}$ and $\{N_i^{\max}\}_{i\in\mathbb{N}}$ of measure zero sets such that*

$$\lim_{i\to\infty} \mathrm{d}_H\big(((A_{\varepsilon_i,p} \setminus N_i^{\max}) \cup A_0^{\max}) \cap K, A_0^{\max} \cap K\big) = 0$$

and

$$\lim_{i\to\infty} \mathrm{d}_H\big(((A_{\varepsilon_i,p} \cup N_i^{\min}) \cap A_0^{\min}) \cap K, A_0^{\min} \cap K\big) = 0.$$

When Assumption 3.11 holds, Theorem 4.16 asserts that

$$\lim_{i\to\infty} (((A_{\varepsilon_i,p} \cup N_i^{\min} \setminus N_i^{\max}) \cup A_0) \cap K, A_0 \cap K) = 0$$

where A_0 is the unique Bayes classifier. Applying Corollary 4.13 in this case, we find that the minimizers for the probabilistic training problem (5) converge to the Bayes classifier at the rate $O(\varepsilon^{\frac{1}{d+2}})$.

We conclude the discussion of the probabilistic adversarial training problem (5) by commenting on why this result fails to extend to the Ψ-perimeter problem mentioned in Remark 1.4.

Remark 4.17. *Recall that [4] considers the Ψ adversarial training problem*

$$\inf_{A\in\mathcal{B}(\mathbb{R}^d)} \mathbb{E}_{(x,y)\sim\mu}[|\mathbb{1}_A(x) - y|] + \mathrm{ProbPer}_\Psi(A) \tag{17}$$

where the Ψ-perimeter is given by

$$\mathrm{ProbPer}_\Psi(A) := \int_{A^c} \Psi(\mathbb{P}(x' \in A : x' \in \mathfrak{p}_x)) d\rho_0(x) + \int_A \Psi(\mathbb{P}(x' \in A^{\mathsf{c}} : x' \in \mathfrak{p}_x)) d\rho_1(x)$$

where $\Psi : [0,1] \to [0,1]$ is concave and non-decreasing. As opposed to the probabilistic adversarial training problem (5), the existence of minimizers to (17) has been established.

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

However, notice that the Ψ-perimeter cannot be expressed as $w_0\rho_0(\Lambda_\Psi^0(A)) + w_1\rho_1(\Lambda_\Psi^1(A))$ if Ψ is concave and non-decreasing as indicator functions are not concave. The Ψ-perimeter is an example of a data-adapted perimeter from the literature that cannot be represented via the deterministic attack framework. At present, whether the energy exchange inequality holds for the Ψ-perimeter remains an open question and proving this inequality for the Ψ-perimeter would be a promising first step towards showing uniform convergence of minimizers of (17).

5. Conclusion

In this paper, we developed a unifying framework for the adversarial and probabilistic adversarial training problems to define more generalized adversarial attacks. Under natural set-algebraic assumptions, we derived the energy exchange inequality to quantify the effect of removing a set where a given label was energetically preferable from a minimizer. Utilizing the energy exchange inequality to show that there exist minimizers disjoint from sets where the label 0 is strongly preferred energetically, we then proved uniform convergence in the Hausdorff distance for various adversarial attacks. This significantly strengthens the type of convergence established via Γ-convergence techniques [7], as well as generalizing it to a broader class of adversarial attacks. Finally, we derived the rate of convergence based on our proof techniques.

There are various future directions of research suggested by our results in this paper. First, the uniform convergence results increase the information that we have about minimizers and sequences of approximate minimizers. That information may be useful in establishing regularity results about minimizers, for example, in the case of the adversarial training problem (3), or may provide helpful information for proving existence in the generalized case. A different avenue of research to pursue would be to sharpen the convergence rates found in this paper by improving estimates from Lemmas 3.3 and 4.6 to determine whether the formally derived rate of $O(\varepsilon)$ can be achieved. Finally, one could consider how to expand the theoretical deterministic attack function framework to encapsulate other types of adversarial training problems, such as Ψ adversarial training problem (17).

Acknowledgements. The authors would also like to thank the reviewers for their insightful comments, in particular, one reviewer who offered a simplified proof of Lemma 3.3.

Funding statement. The authors gratefully acknowledge the support of the NSF DMS 2307971 and the Simons Foundation TSM.

Competing interest. The authors declare none.

References

[1] Awasthi, P., Frank, N., Mao, A., Mohri, M. & Zhong, Y. (2021). Calibration and consistency of adversarial surrogate losses, *Advances in Neural Information Processing Systems*, Vol. **34**, Curran Associates, Inc, pp. 9804–9815, https://proceedings.neurips.cc/paper.

[2] Awasthi, P., Frank, N. & Mohri, M. (2023) *On the existence of the adversarial bayes classifier (extended version)*, arXiv: 2112.01694.

[3] Bao, H., Scott, C. & Sugiyama, M. (2020). Calibrated Surrogate Losses for Adversarially Robust Classification, *Proceedings of Machine Learning Research*, *Thirty Third Annual Conference on Learning Theory*, Vol. **125**, PMLR, pp. 1–43, http://proceedings.mlr.press/v125/bao20a/bao20a.pdf.

[4] Bungert, L., Trillos, N. G., Jacobs, M., McKenzie, D., Nikolić,Đ. & Wang, Q. (2023) It begins with a boundary: A geometric view on probabilistically robust learning, arXiv: 2305.18779.

[5] Bungert, L., Trillos, N. García & Murray, R. (2023) The geometry of adversarial training in binary classification. *Inf. Inference* **12**(2), 921–968. DOI: 10.1093/imaiai/iaac029.

[6] Bungert, L., Laux, T. & Stinson, K. (2024) *A mean curvature flow arising in adversarial training*, arXiv: 2404.14402.

[7] Bungert, L. & Stinson, K. (2024) Gamma-convergence of a nonlocal perimeter arising in adversarial machine learning. *Calc. Var.* **63**(5), 114. DOI: 10.1007/s00526-024-02721-9.

[8] Cesaroni, A., Dipierro, S., Novaga, M. & Valdinoci, E. (2018) Minimizers for nonlocal perimeters of Minkowski type. *Calc. Var.* **57**(2), 64. DOI: 10.1007/s00526-018-1335-9.

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

[9] Cesaroni, A. & Novaga, M. (2017) Isoperimetric problems for a nonlocal perimeter of Minkowski type. *Geom. Flows* **2**(1), 86–93. DOI: 10.1515/geofl-2017-0003.
[10] Chambolle, A., Morini, M. & Ponsiglione, M. (2015) Nonlocal curvature flows. *Arch. Rational Mech. Anal.* **218**(3), 1263–1329. DOI: 10.1007/s00205-015-0880-z.
[11] Engstrom, L., Tran, B., Tsipras, D., Schmidt, L. & Madry, A. (2019). Exploring the Landscape of Spatial Robustness, Proceedings of Machine Learning Research, *Proceedings of the 36th International Conference on Machine Learning*, Vol. **97**, PMLR, pp. 1802–1811, https://proceedings.mlr.press/v97/engstrom19a.html.
[12] Eykholt, K., Evtimov, I., Fernandes, E., et al. (2018) Robust physical-world attacks on deep learning visual classification, In: *IEEE/CVF Conference On Computer Vision and Pattern Recognition*, pp. 1625–1634. DOI: 10.1109/CVPR.2018.00175.
[13] Frank, N. S. & Niles-Weed, J. (2024) Existence and minimax theorems for adversarial surrogate risks in binary classification. *J. Mach. Learn. Res.* **25**(58), 1–41, http://jmlr.org/papers/v25/23-0456.html.
[14] Trillos, N. García, Jacobs, M. & Kim, J. (2023) *On the existence of solutions to adversarial training in multiclass classification*, arXiv: 2305.00075.
[15] Trillos, N. García, Kim, J. & Jacobs, M. (2023) The multimarginal optimal transport formulation of adversarial multiclass classification. *J. Mach. Learn. Res.* **24**, 1533–7928, issn: 1532-4435, https://www.jmlr.org/papers/volume24/22-0698/22-0698.pdf.
[16] Trillos, N. García & Murray, R. (2022) Adversarial classification: Necessary conditions and geometric flows. *J. Mach. Learn. Res.* **23**, 1–38, https://www.jmlr.org/papers/volume23/21-0222/21-0222.pdf.
[17] Heredia, L. Gnecco, Pydi, M. S., Meunier, L., Negrevergne, B. & Chevaleyre, Y. (2023). On the Role of Randomization in Adversarially Robust Classification, In: *Thirty Seventh Conference on Neural Information Processing Systems*, https://proceedings.neurips.cc/paper.
[18] Goodfellow, I. J., Shlens, J. & Szegedy, C. (2015) Explaining and harnessing adversarial examples, arXiv: 1412.6572.
[19] Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks, In: *Sixth International Conference on Learning Representations*, https://dspace.mit.edu/handle/1721.1/137496.
[20] Meunier, L., Ettedgui, R., Pinot, R., Chevaleyre, Y. & Atif, J. (2022). Towards Consistency in Adversarial Classification, In: *Thirty Sixth Conference on Neural Information Processing Systems*, https://proceedings.neurips.cc/paper.
[21] Pydi, M. S. & Jog, V. (2021) Adversarial risk via optimal transport and optimal couplings. *IEEE Trans. Inform. Theory* **67**(9), 6031–6052. DOI: 10.1109/TIT.2021.3100107.
[22] Pydi, M. S. & Jog, V. (2024) The many faces of adversarial risk: An expanded study. *IEEE Trans. Inform. Theory* **70**(1), 550–570. DOI: 10.1109/TIT.2023.3303221.
[23] Qin, Y., Carlini, N., Goodfellow, I., Cottrell, G. & Raffel, C. (2019). Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition, In: Proceedings of Machine Learning Research, *Proceedings of the 36th International Conference on Machine Learning*, Vol. **97**, PMLR, https://proceedings.mlr.press/v97/qin19a/qin19a.pdf.
[24] Raman, V., Subedi, U. & Tewari, A. (2023). On Proper Learnability Between Average- and Worst-case Robustness, In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*, https://proceedings.neurips.cc/paper.
[25] Robey, A., Chamon, L., Pappas, G. J. & Hassani, H. (2022) Probabilistically Robust Learning: Balancing Average and Worst-case Performance, Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G. & Sabato, S. (editors), *Proceedings of the 39th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, Vol. **162**, PMLR, pp. 18667–18686, https://proceedings.mlr.press/v162/robey22a.html.
[26] Serra, J. (1983). Image analysis and mathematical morphology. USA, Academic Press, Inc., isbn: 0126372403.
[27] Szegedy, C., Zaremba, W., Sutskever, I., et al. (2014) Intriguing properties of neural networks, arXiv: 1312.6199.
[28] Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. El & Jordan, M. I. (2019). Theoretically Principled Trade-off between Robustness and Accuracy, In: *Proceedings of the 36th International Conference on Machine Learning*, Vol. **97**, http://proceedings.mlr.press/v97/zhang19p/zhang19p.pdf.

Appendix A

A.1. The U sets for ϕ_ε

In Remark 2.4, we claim that further conclusions about the U sets may be drawn when $\phi = \phi_\varepsilon$. We will now verify these claims. We consider only the cases where whether the entire set U_i is attacked cannot be unambiguously determined by Λ-monotonicity (see Table 1). In all of the following cases, we assume

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

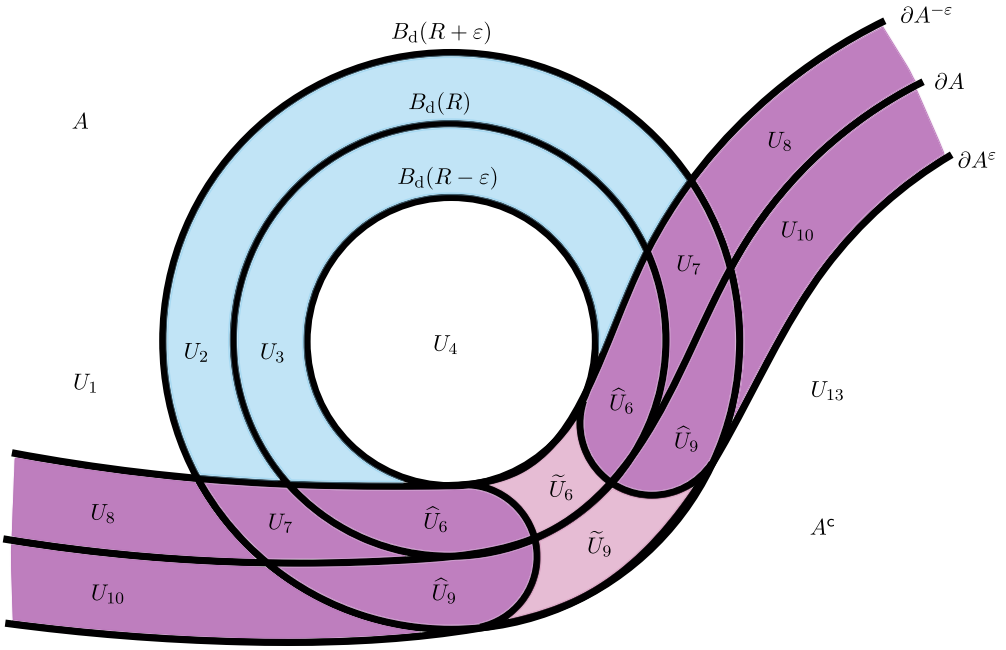


Figure A1. *A degenerate example where U_6 and U_9 are neither solely attacked nor unattacked sets. The example arises because the boundaries of A and $B_{\mathrm{d}}(R)$ coincide. The pink and purple sets represent the ε-perimeter regions of A, whereas the blue and purple regions represent the ε-perimeter regions for $A \setminus \overline{B_{\mathrm{d}}(R)}$.*

that the interaction of A, E is nontrivial in the sense that $A \cap E$ and $A^{\mathrm{c}} \cup E$ are both nonempty. Otherwise, the following sets will either be empty themselves or we trivially find $\mathrm{d}(x, \emptyset) = \infty$.

Proposition A.1. *Let $\phi = \phi_\varepsilon$ and $A, E \in \mathcal{B}(\mathbb{R}^d)$. Then, $\widehat{U}_1 = \emptyset$.*

Proof. *Suppose $x \in U_1 \subset A \cap E^{\mathrm{c}}$. By construction, we have $\mathrm{d}(x, A^{\mathrm{c}}) \geq \varepsilon$ and $\mathrm{d}(x, E) \geq \varepsilon$. This implies that $\mathrm{d}(x, A^{\mathrm{c}} \cup E) = \mathrm{d}(x, (A \setminus E)^{\mathrm{c}}) \geq \varepsilon$ as well. Thus, $\widehat{U}_1 = \emptyset$.* □

Proposition A.2. *Let $\phi = \phi_\varepsilon$ and $A, E \in \mathcal{B}(\mathbb{R}^d)$. Then, $\widetilde{U}_3 = \emptyset$.*

Proof. *Suppose $x \in U_3 \subset A \cap E$. By construction, we have $\mathrm{d}(x, A^{\mathrm{c}}) \geq \varepsilon$ and $\mathrm{d}(x, E^{\mathrm{c}}) < \varepsilon$. As $\mathrm{d}(x, A^{\mathrm{c}}) \geq \varepsilon$, $B(x, \varepsilon) \subset A$. Furthermore, as $\mathrm{d}(x, E^{\mathrm{c}}) < \varepsilon$, $B(x, \varepsilon) \cap E^{\mathrm{c}} \neq \emptyset$. Thus, there exists some $y \in B(x, \varepsilon) \cap E^{\mathrm{c}} \subset A \cap E^{\mathrm{c}}$. Hence, $\mathrm{d}(x, A \setminus E) < \varepsilon$, so $\widetilde{U}_3 = \emptyset$.* □

Proposition A.3. *Let $\phi = \phi_\varepsilon$ and $A, E \in \mathcal{B}(\mathbb{R}^d)$. Then, $\widetilde{U}_{10} = \emptyset$.*

Proof. *Suppose $x \in U_{10}$. By construction, we have $\mathrm{d}(x, A) < \varepsilon$ and $\mathrm{d}(x, E) \geq \varepsilon$. As $\mathrm{d}(x, E) \geq \varepsilon$, $B(x, \varepsilon) \subset E^{\mathrm{c}}$. Furthermore, as $\mathrm{d}(x, A) < \varepsilon$, $B(x, \varepsilon) \cap A \neq \emptyset$. Thus, there exists some $y \in B(x, \varepsilon) \cap A \subset A \cap E^{\mathrm{c}}$. Hence, $\mathrm{d}(x, A \setminus E) < \varepsilon$, so $\widetilde{U}_{10} = \emptyset$.* □

As for U_6, U_9 and U_{11}, we can make no determinations about whether all points in these sets must be attacked or not. Figure 2 shows an example where U_{11} must be split into attacked and unattacked subsets. In special cases where the boundaries of the sets A and E coincide, U_6 and U_9 may also need to be split into attacked and unattacked subsets (see Figure A1).

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

A.2. Λ*-set decompositions*

For completeness, we give further details about the decompositions by U sets in Proposition 2.5, namely $\Lambda_\phi^0(A \setminus E)$, $\Lambda_\phi^1(A \setminus E)$, $A \cap E$, $A \setminus E$, $(A \setminus E)^{\mathsf{c}}$, $D_\phi(A; E)$, and $D_\phi(E^{\mathsf{c}}; A)$. Table 1 is reproduced for ease of reference.

If $x \in U_i$	Then for $\Lambda(A \setminus E)$ we have
$U_1 := \tilde{\Lambda}_\phi^1(A) \cap \tilde{\Lambda}_\phi^0(E)$	N/A
$U_2 := \tilde{\Lambda}_\phi^1(A) \cap \Lambda_\phi^0(E)$	$x \in \Lambda_\phi^1(A \setminus E)$
$U_3 := \tilde{\Lambda}_\phi^1(A) \cap \Lambda_\phi^1(E)$	N/A
$U_4 := \tilde{\Lambda}_\phi^1(A) \cap \tilde{\Lambda}_\phi^1(E)$	$x \in \tilde{\Lambda}_\phi^0(A \setminus E)$
$U_5 := \Lambda_\phi^1(A) \cap \tilde{\Lambda}_\phi^1(E)$	$x \in \tilde{\Lambda}_\phi^0(A \setminus E)$
$U_6 := \Lambda_\phi^1(A) \cap \Lambda_\phi^1(E)$	N/A
$U_7 := \Lambda_\phi^1(A) \cap \Lambda_\phi^0(E)$	$x \in \Lambda_\phi^1(A \setminus E)$
$U_8 := \Lambda_\phi^1(A) \cap \tilde{\Lambda}_\phi^0(E)$	$x \in \Lambda_\phi^1(A \setminus E)$
$U_9 := \Lambda_\phi^0(A) \cap \Lambda_\phi^0(E)$	N/A
$U_{10} := \Lambda_\phi^0(A) \cap \tilde{\Lambda}_\phi^0(E)$	N/A
$U_{11} := \Lambda_\phi^0(A) \cap \Lambda_\phi^1(E)$	N/A
$U_{12} := \Lambda_\phi^0(A) \cap \tilde{\Lambda}_\phi^1(E)$	$x \in \tilde{\Lambda}_\phi^0(A \setminus E)$
$U_{13} := \tilde{\Lambda}_\phi^0(A)$	$x \in \tilde{\Lambda}_\phi^0(A \setminus E)$

• $\Lambda_\phi^0(A \setminus E)$ is comprised of all U sets such that $U_i \notin A \setminus E$ and the points can be attacked by the adversary for the classifier $A \setminus E$. The $U_i \notin A \setminus E$ are all U sets such that the Λ-set for A has the superscript 0 or the Λ-set for E has the superscript 1, that is, U_3, U_4, U_5, U_6, U_9, U_{10}, U_{11}, U_{12} and U_{13}. However, U_4, U_5, U_{12}, and U_{13} are all unattacked by Table 1. Thus, $\Lambda_\phi^0(A \setminus E)$ contains the attacked subsets of U_3, U_6, U_9, U_{10}, and U_{11}. Hence,

$$\Lambda_\phi^0(A \setminus E) = \widehat{U}_3 \cup \widehat{U}_6 \cup \widehat{U}_9 \cup \widehat{U}_{10} \cup \widehat{U}_{11}.$$

• $\Lambda_\phi^1(A \setminus E)$ is comprised of all U sets such that $U_i \in A \setminus E$ and the points can be attacked by the adversary for the classifier $A \setminus E$. The $U_i \notin A \setminus E$ are all U sets such that the Λ-set for A has the superscript 1 and the Λ-set for E has the superscript 0, that is, U_1, U_2, U_7, and U_8. By Table 1, the sets U_2, U_7 and U_8 are belong entirely to $\Lambda_\phi^1(A \setminus E)$, so $\Lambda_\phi^1(A \setminus E)$ contains those sets and the attacked subset of U_1. Hence,

$$\Lambda_\phi^1(A \setminus E) = \widehat{U}_1 \cup U_2 \cup U_7 \cup U_8.$$

• $A \cap E$ is comprised of all U sets such that the Λ-sets for A and E both have the superscript 1. Hence,

$$A \cap E = U_3 \cup U_4 \cup U_5 \cup U_6.$$

• $A \setminus E$ is comprised of all U sets such that the Λ-set for A has the superscript 1 and the Λ-set for E has the superscript 0. Hence,

$$A \setminus E = U_1 \cup U_2 \cup U_7 \cup U_8.$$

• $(A \setminus E)^{\mathsf{c}}$ is comprised of all U sets not in $A \setminus E$, or alternatively, all U sets such that either the Λ-set for A has the superscript 0 or the Λ-set for E has the superscript 1. Hence,

$$(A \setminus E)^{\mathsf{c}} = U_3 \cup U_4 \cup U_5 \cup U_6 \cup U_9 \cup U_{10} \cup U_{11} \cup U_{12} \cup U_{13}.$$

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press

• Recall $D_\phi(A;E) = w_0\rho_0(\Lambda^0_\phi(A)\cap E) + w_1\rho_1(\Lambda^1_\phi(A)\cap E)$. The set E can be expressed in terms of Λ-sets by $E = \Lambda^1_\phi(E)\cup\tilde{\Lambda}^1_\phi(E)$. By Table 1,

$$D_\phi(A;E) = w_0\rho_0(U_{11}\cup U_{12}) + w_1\rho_1(U_5\cup U_6).$$

• Recall $D_\phi(E^{\mathsf{c}};A) = w_0\rho_0(\Lambda^1_\phi(E)\cap A) + w_1\rho_1(\Lambda^0_\phi(E)\cap A)$ by the Complement Property of Λ-sets. The set A can be expressed in terms of Λ-sets by $A = \Lambda^1_\phi(A)\cup\tilde{\Lambda}^1_\phi(A)$. By Table 1,

$$D_\phi(E^{\mathsf{c}};A) = w_0\rho_0(U_3\cup U_6) + w_1\rho_1(U_2\cup U_7).$$

Cite this article: Morris R. and Murray R. (2026). Uniform convergence of adversarially robust classifiers. *European Journal of Applied Mathematics*, **37**, 43–72. https://doi.org/10.1017/S0956792525100223

https://doi.org/10.1017/S0956792525100223 Published online by Cambridge University Press