An Analytical and Geometric Perspective on Adversarial Robustness

Nicolás García Trillos and Matt Jacobs

. . . y luego se fueron el uno para el otro, como si fueran dos mortales enemigos.

Don Quixote; Chapter 8, Part 1.

1. Introduction

In the last ten years, neural networks have made incredible strides in classifying large data sets, to the point that they can now outperform humans in raw accuracy. However, the robustness of these systems is a completely different story. Suppose you were asked to identify whether a photo contained an image of a cat or a dog. You probably would have no difficulty at all; at worst, maybe you would only be tripped up by a particularly small or unusual Shiba Inu. In contrast, it has been widely documented that an adversary can convince an otherwise well-performing neural network that a dog is actually a cat (or vice-versa) by making tiny human-imperceptible changes to an image at the pixel level. These small perturbations are known as *adversarial attacks* and they are a significant obstacle to the deployment of machine learning systems in security-critical applications [GSS14]. The susceptibility to adversarial attacks is not exclusive to neural network models, and many other learning systems have also been observed to be brittle when facing adversarial perturbations of data.

Nicolás García Trillos is an assistant professor in the department of statistics at the University of Wisconsin-Madison, Madison, Wisconsin. His email address is garciatrillo@wisc.edu.

Matt Jacobs is an assistant professor in the Department of Mathematics at the University of California Santa Barbara, Santa Barbara, California. His email address is majaco@ucsb.edu.

Communicated by Notices *Associate Editor Daniela De Silva.*

DOI: https://doi.org/10.1090/noti2758

The business of defending against adversarial attacks is known as *adversarial robustness*, *robust training*, or simply *adversarial training* (although we will mostly reserve the latter name for a specific optimization objective). There are many methods in the literature that can be used to build defenses against adversarial attacks, but here we will be particularly interested in methods that enforce robustness *during model training*. In these types of methods, the standard training process — driven primarily by accuracy maximization — is substituted by a training process that promotes robustness, typically through the use of a different optimization objective that factors in the actions of a well-defined adversary.

In this article, we give a brief overview of adversarial attacks and adversarial robustness and summarize some recent attempts to mathematically understand the process of robust training. Adversarial training and its mathematical foundations are active areas of research and a thorough review of its extant literature is beyond the scope of this article.[1] For this reason, we will focus our discussion around some important lines of research in the theory of adversarial robustness, some of which are based on our own research work in the field, which takes a distinctive analytic and geometric perspective. One of our goals is to convey the mathematical richness of the field and discuss some of the many opportunities that are available for mathematicians to contribute to the development and understanding of this important applied problem.

1.1. Basics of training learning models.

Data classification or regression typically occurs over a product space of the form $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Here $\mathcal{X}$ is the *data space* or *feature space*, an abstract metric space containing the data points, while $\mathcal{Y}$ is the set of *labels*, usually a finite set for classification tasks or the real line for regression. In the remainder, we mostly focus our discussion on the classification problem. There, the goal is to construct a function that accurately partitions the data space into the possible classes contained in $\mathcal{Y}$. The learner does this by searching for a function $f : \mathcal{X} \to \mathcal{Y}$ or $f : \mathcal{X} \to S_{\mathcal{Y}}$ where $S_{\mathcal{Y}}$ is the probability simplex with $|\mathcal{Y}|$ vertices $S_{\mathcal{Y}} = \{p \in [0,1]^{\mathcal{Y}} : \sum_{y \in \mathcal{Y}} p_y = 1\}$. If we write $f(x) = (f_y(x))_{y \in \mathcal{Y}}$, then each f_y represents the learner's confidence that the data point $x \in \mathcal{X}$ belongs to class $y \in \mathcal{Y}$. While a probabilistic classifier is typically the desired output of a learning task, note that one can always obtain a deterministic classifier $f : \mathcal{X} \to \mathcal{Y}$ by selecting the largest entry of any tuple.

To train a machine learning system, one typically needs a finite training set $Z \subset \mathcal{Z}$ consisting of data pairs (x_i, y_i), i.e., feature vectors with their associated ground truth classification. One then minimizes a *loss function* over some chosen function space $\mathcal{F} \subset \{f : \mathcal{X} \to S_{\mathcal{Y}}\}$ with the goal of finding a function f^* that produces an accurate classification of the data. For instance, $\mathcal{F}$ may be the space of all neural networks with a certain architecture, while the loss function typically has the form

$$\frac{1}{|Z|} \sum_{(x,y) \in Z} \ell(f(x), y), \tag{1}$$

where ℓ is a function that is small when $f(x)$ gives high probability to the ground truth label y, and large otherwise. In practice, it may only be possible to find a classifier f that is a local minimizer of (1) over $\mathcal{F}$, though it is often possible to drive the loss function to nearly zero during training via stochastic gradient descent. Either way, well-trained classifiers typically perform well—at least in the absence of adversarial attacks.

1.2. Adversarial attacks.

Given a trained classifier f and a data point $x \in \mathcal{X}$, what is the best way for an adversary to perturb x to produce an incorrect classification? In order for this to be a nontrivial question, we must assume that there are some restrictions on how far the adversary can perturb x. This restriction is known as the *adversarial budget*, and it plays a crucial role in both adversarial attacks and robust training. For our purposes, we will formulate the adversarial budget through a parameter $\varepsilon > 0$ and assume that the adversary can only produce a perturbed data point that lies in $B_\varepsilon(x) \subset \mathcal{X}$, the ball of radius ε centered at the original data point x.

Returning to our question, if the adversary has full access to the function f and knows that the correct label for x is y, then the most powerful attack for a given budget ε is found by replacing x with any point $\tilde{x}$ satisfying

$$\tilde{x} \in \operatorname*{argmax}_{x' \in B_\varepsilon(x)} \ell(f(x'), y). \tag{2}$$

In practice, the adversary may not be able to find a point $\tilde{x}$ that exactly satisfies (2). However, when $\mathcal{X}$ is a subspace of Euclidean space, a simpler approach that produces highly effective attacks is to perturb the data in the direction of steepest ascent for the loss function by choosing

$$\tilde{x} = x + \varepsilon \frac{\nabla_x \ell(f(x), y)}{\|\nabla_x \ell(f(x), y)\|}, \tag{3}$$

or by considering the popular *PGD* attack

$$\tilde{x} = x + \varepsilon \operatorname{sign}(\nabla_x \ell(f(x), y)), \tag{4}$$

where $\operatorname{sign}(\cdot)$ denotes the coordinatewise sign of its input; see [GSS14, MMS+18], and [TT22] for a motivation for the PGD attack.

Regardless of how the adversary chooses its attack, there are two key takeaways from formulas (2), (3), and (4) that we would like to highlight. Firstly, we see that adversarial attacks are found by attempting to maximize the loss function with respect to data perturbations. In contrast,

[1] *AMS Notices limits to 20 the references per article; we refer to the references cited here for further pointers to the literature.*

the learner trains the classifier by attempting to minimize the loss function among classifiers belonging to a chosen function space $\mathcal{F}$ (typically a parametric family). Hence, the learner and adversary can be viewed as playing a two-player game where they compete to set the value of the loss function using the tools at their disposal (the learner first gets to choose f, the adversary then gets to modify data); this connection to game theory will become more important shortly. Secondly, it should be clear from formulas (2), (3), and (4) that the effectiveness of adversarial attacks must stem from a certain lack of smoothness in the trained classifier f. Indeed, if f were say 1-Lipschitz, then an adversary with budget ε could not change the classification probabilities at any point by more than ε. Thus, attacks that can fool an image classifier by making human-imperceptible changes to pixel values must be exploiting a significant lack of regularity.

1.3. **Adversarial training.** In light of the above considerations, to stave off adversarial attacks one must find a way to construct classifiers with better regularity properties. The most classical (and perhaps most obvious) way to do this would be to replace the training objective (1) with a new objective

$$\frac{1}{|Z|}\sum_{(x,y)\in Z}\ell(f(x),y)+R(f) \tag{5}$$

where $R:\mathcal{F}\to\mathbb{R}$ is a term that promotes regularity. For instance, R could constrain the Lipschitz constant of f or could be some other gradient penalty term. This approach has a long history of success in inverse problems (in that setting one typically adds a regularizing term to a data-fitting term to help mitigate the effect of noise), however, in the context of machine learning, it is often too difficult to efficiently minimize (5). For instance, it is very difficult to train a neural network with a Lipschitz constant constraint. On the other hand, popular and computationally feasible regularization terms for neural network training, for instance, weight regularization, do not seem to provide any defense against adversarial attacks.

Adversarial training is a different approach to regularization/robustification that has become very popular in the machine learning community. In adversarial training, rather than modifying the training objective with a regularizing term, one instead incorporates the adversary into the training process [SZS+14, MMS+18]. More precisely, adversarial training replaces the objective (1) with

$$\frac{1}{|Z|}\sum_{(x,y)\in Z}\sup_{\tilde{x}\in B_\varepsilon(x)}\ell(f(\tilde{x},y), \tag{6}$$

where the adversarial budget ε is chosen by the user. When training using (6), the learner is forced to find a function f that cannot be easily attacked by an adversary with budget ε. The advantages of training using (6) compared to (5) are conceptual and computational. Conceptual, because in the formulation (6) one explicitly trains to defend against a well-defined adversary (although in practice this requires "understanding the enemy"). Computational, because (6) is in essence a min-max problem (the adversary maximizes over ε-perturbations while the learner minimizes by altering f) for which many implementable algorithms exist (for instance alternating gradient descent and ascent steps). Furthermore, the regularizing effect of (6) is data-dependent in contrast to the regularization induced by a standard gradient penalty term which has very little connection to the structure of the data.

On the other hand, compared to (5), it is harder to understand (analytically and geometrically) how exactly (6) is regularizing/robustifying the classifier f; see the discussion in [TT22] and references therein. Furthermore, it is not so clear how the user should choose the budget parameter ε and be mindful of the tradeoff between accuracy (on clean data) and robustness that problem (6) introduces. Answering these questions in full generality is challenging and remains an open problem in the field.

1.4. **Outline.** To fix some ideas, we write (6) in the general form:

$$\min_{f\in\mathcal{F}}\mathbb{E}_{(x,y)\sim\mu}[\sup_{\tilde{x}\in B_\varepsilon(x)}\ell(f(\tilde{x}),y)], \tag{AT}$$

where μ is an arbitrary probability measure over $\mathcal{Z}$, the "clean data distribution," and not just an empirical measure as in (6). We work with arbitrary μ to avoid distinguishing, whenever unnecessary, between population and finite data settings. Only in some parts of Section 2 will it be important to make assumptions on μ.

In this paper we explore the following general questions:

1. What type of regularization is enforced on learning models by the presence of adversaries?
2. What are the tradeoffs between accuracy and robustness when training models robustly?
3. How can one actually train models to be robust to specific adversarial attacks?
4. How can one compute meaningful lower bounds for the (AT) problem?

The above questions are too broad to be answered in complete generality, and in the remainder, we will focus on specific settings where we can reveal interesting geometric and analytic structures. In particular, in Section 2, we explore the type of regularization enforced on binary classifiers, revealing a connection between adversarial training and perimeter minimization problems. This connection will allow us to interpret geometrically the tradeoff between accuracy and robustness. In Section 3, we discuss a concrete game-theoretic interpretation of adversarial training and discuss a more general framework for adversarial

robustness based on *distributionally robust optimization* (DRO). We use this connection with game theory to discuss some potential strategies for training robust learning models and highlight the significance of the concept of Nash equilibrium for adversarial training. Finally, in Section 4, we discuss how an *agnostic learner* setting can be used to derive lower bounds for more general (AT) problems. We show that in the agnostic learner setting for multiclass classification the adversarial robustness objective can be equivalently rewritten as the geometric problem of finding a (generalized) barycenter of a collection of measures and then discuss the computational implications of this equivalence. We wrap up the paper in Section 5 by discussing some research directions connected to the topics presented throughout the paper.

1.4.1. *Additional notation.* When working on $\mathcal{X} = \mathbb{R}^d$ we will often consider balls $B_\varepsilon(x)$ associated to a given norm $\|\cdot\|$ on $\mathbb{R}^d$. We use $\|\cdot\|_*$ to denote the dual norm of $\|\cdot\|$, which is defined according to

$$\|v\|_* = \sup_{u\in\mathbb{R}^d\,:\,\|u\|\le 1} \langle u, v\rangle.$$

We will denote by $\mathcal{P}(\mathcal{Z})$ the space of Borel probability measures over the set $\mathcal{Z}$. Given two measures $\mu, \tilde{\mu} \in \mathcal{P}(\mathcal{Z})$ we denote by $\Gamma(\mu, \tilde{\mu})$ the space of couplings between μ and $\tilde{\mu}$, i.e., probability measures over $\mathcal{Z}\times\mathcal{Z}$ whose first and second marginals are, respectively, μ and $\tilde{\mu}$. We will also use the notion of a *pushforward* of a measure by a map. Precisely, if $T : A \mapsto B$ is a measurable map between two measurable spaces, and μ is a probability measure over A, we define $T_\sharp\mu$, the pushforward of μ by T, to be the measure on B for which $T_\sharp\mu(C) = \mu(T^{-1}(C))$ for all measurable subsets C of B.

2. Adversarial Robustness: Regularization and Perimeter

In this section, we discuss the connection between adversarial training and explicit regularization methods. To motivate this connection, let us first consider a simple robust linear regression setting. In this setting, models in the family $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ take the form

$$f_\theta(x) = \langle \theta, x\rangle, \quad x \in \mathcal{X},$$

where Θ is some subset of $\mathbb{R}^d$ and $\mathcal{X} = \mathbb{R}^d$. Here, the learner's goal is to select a linear regression function relating inputs $x \in \mathbb{R}^d$ to real-valued outputs y. We show the following equivalence between problem (AT) and an explicit regularization problem taking the form of a Lasso-type linear regression:

$$\begin{aligned}\min_{\theta\in\Theta} &\mathbb{E}_{(x,y)\sim\mu}[\sup_{\tilde{x}\in B_\varepsilon(x)} |\langle\theta,\tilde{x}\rangle - y|] \\ &= \min_{\theta\in\Theta} \mathbb{E}_{(x,y)\sim\mu}[|\langle\theta,x\rangle - y|] + \varepsilon\|\theta\|_*;\end{aligned} \tag{7}$$

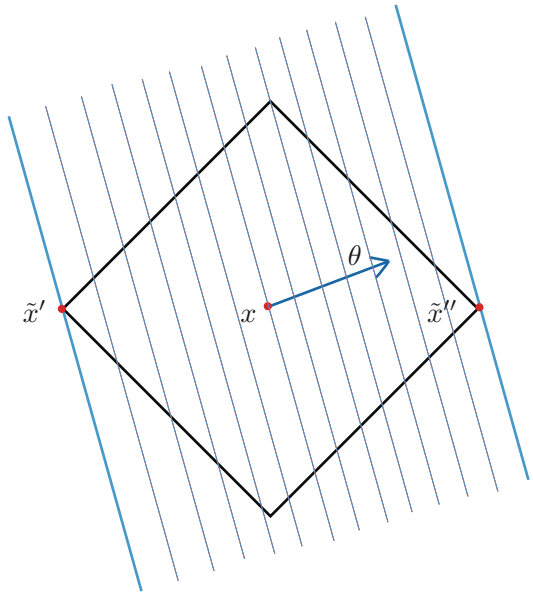

Figure 1. ℓ^1 ball around x of radius ε crossed by level sets of function $x \mapsto \langle\theta, x\rangle$. The value $\sup_{\tilde{x}\in B_\varepsilon(x)} |\langle\theta,\tilde{x}\rangle - y|$ is realized at either $\tilde{x}'$ or $\tilde{x}''$.

note that in order to be consistent with the standard definition of the Lasso regularization, we would require $\|\cdot\|$ to be the l^∞-norm to get $\|\cdot\|_*$ to be the l^1-norm. Identity (7) is only one of many similar identities relating regularization methods and adversarially robust learning problems for classical families of statistical models. The extent of this type of equivalences is more apparent when considering DRO versions (see Section 3 for a definition) of the adversarial training problem, e.g., see [BKM19], where in addition some statistical inference methodologies, motivated by these equivalences, are proposed.

To deduce (7), it is enough to consider the optimization problem $\sup_{\tilde{x}\in B_\varepsilon(x)} |f_\theta(x) - y|$ at every fixed (x, y) and realize that the sup can be written as either $\sup_{\tilde{x}\in B_\varepsilon(x)}\langle\tilde{x},\theta\rangle - y$ when $\langle x,\theta\rangle \ge y$, or as $\sup_{\tilde{x}\in B_\varepsilon(x)} y - \langle\tilde{x},\theta\rangle$ when $\langle x,\theta\rangle \le y$; see Figure 1 for an illustration. Using the definition of the dual norm $\|\cdot\|_*$ one can deduce that in all cases this expression is equal to $|\langle\theta, x\rangle - y| + \varepsilon\|\theta\|_*$, from which (7) follows.

Equivalence (7), although limited to the linear regression setting, motivates exploring the regularization effect of adversaries on more general families of learning models. In the remainder of this section we discuss how in the *binary classification setting* this regularization effect can be related to geometric properties of decision boundaries, in particular to their size or curvature. By presenting this analysis we hope to convey that the connection between adversarial robustness and regularization methods goes beyond simple classical statistical settings, in turn revealing a variety of interesting geometric problems motivated by machine learning.

2.1. **Perimeter regularization.** Let us consider a binary classification version of (AT) where $y \in \{0,1\}$, ℓ is the 0-1 loss (i.e., 0 if the two inputs of ℓ are the same, and 1 otherwise), and $\mathcal{F}$ is a family of binary classifiers $\mathcal{F} = \{\mathbb{1}_A : A \in \mathcal{A}\}$ for $\mathcal{A}$ a family of measurable subsets of $\mathcal{X}$. Here $\mathbb{1}_A$ denotes the indicator function of a subset A of $\mathcal{X}$, defined according to

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

Notice that we can "parameterize" a family of binary classifiers with a family of subsets in $\mathcal{X}$ without losing any generality, due to the fact that binary classifiers output one of two values, 0 or 1, and thus can be characterized by the regions of points in $\mathcal{X}$ that they classify as a 1.

In general, as shown in [BGTM23], Problem (AT) is equivalent to a regularization problem, with a *non-local perimeter* regularizer, of the form:

$$\inf_{A \in \mathcal{A}} \mathbb{E}_{(x,y)\sim\mu}\left[|\mathbb{1}_A(x) - y|\right] + \varepsilon \mathrm{Per}_\varepsilon(A;\mu), \tag{8}$$

where

$$\begin{aligned} \mathrm{Per}_\varepsilon(A) &:= \frac{1}{\varepsilon}\mu_0(\partial_\varepsilon A^c) + \frac{1}{\varepsilon}\mu_1(\partial_\varepsilon A), \\ \partial_\varepsilon A^c &:= \{x \in A^c : \operatorname{dist}(x, A) < \varepsilon\}, \\ \partial_\varepsilon A &:= \{x \in A : \operatorname{dist}(x, A^c) < \varepsilon\}; \end{aligned} \tag{9}$$

Figure 2 illustrates the sets $\partial_\varepsilon A^c$ and $\partial_\varepsilon A$. In the above, the measures μ_0 and μ_1 are the measures over $\mathcal{X}$ defined according to $\mu_0(\cdot) := \mu(\cdot \times \{0\})$ and $\mu_1(\cdot) := \mu(\cdot \times \{1\})$, i.e., up to scaling factors they are the conditional distributions of the variable x given the possible values that y may take. The equivalence between (AT) and (8) can be deduced, at least at a formal level, by adding and subtracting the term $\mathbb{E}_{(x,y)\sim\mu}[|\mathbb{1}_A(x) - y|]$ from the term $\mathbb{E}_{(x,y)\sim\mu}[\sup_{\tilde{x}\in B_\varepsilon(x)} \ell(\mathbb{1}_A(x), y)]$ and then identifying the resulting terms with those in (9). To make these computations rigorous and to show existence of solutions to problem (9), there are several technical challenges, beginning with the measurability of the operations involved in defining the problem (AT), that must be overcome; the first part of the work [BGTM23] discusses some of these challenges.

Now, let us motivate the use of the word *perimeter* when describing the functional $\mathrm{Per}_\varepsilon(A)$. Suppose that $\mathcal{X} = \mathbb{R}^d$ and that the measures μ_0 and μ_1 are absolutely continuous with respect to the Lebesgue measure so that we can write them as $\mathrm{d}\mu_0 = \rho_0\,\mathrm{d}x$ and $\mathrm{d}\mu_1 = \rho_1\,\mathrm{d}x$ for two non-negative Lebesgue-integrable functions ρ_0 and ρ_1 that for simplicity will be assumed to be smooth. In this case, (8) can be rewritten as

$$\mathrm{Per}_\varepsilon(A) = \frac{1}{\varepsilon}\int_{\partial_\varepsilon A^c} \rho_0(x)\,\mathrm{d}x + \frac{1}{\varepsilon}\int_{\partial_\varepsilon A} \rho_1(x)\,\mathrm{d}x.$$

Notice that the sets $\partial_\varepsilon A, \partial_\varepsilon A^c$ in the volume integrals shrink toward ∂A, the boundary of A, as we send $\varepsilon \to 0$. Moreover, due to the rescaling factor ε in front of these integrals, one may anticipate a connection between $\mathrm{Per}_\varepsilon(A)$ and the more classical notion of (weighted) perimeter:

$$\mathrm{Per}(A) := \int_{\partial A} (\rho_0(x) + \rho_1(x))\,\mathrm{d}x = \int_{\partial A} \rho(x)\,\mathrm{d}\mathcal{H}^{d-1}(x),$$

where $\rho(x) := \rho_0(x) + \rho_1(x)$. Note that ∂A is precisely the decision boundary between classes 1 and 0 according to the classifier $\mathbb{1}_A$ and that $\rho(x)$ is the density, with respect to the Lebesgue measure, of the marginal of the data distribution μ on the x variable. In the definition of $\mathrm{Per}(A)$ we have used $\mathcal{H}^{d-1}$, the $d-1$ dimensional Hausdorff measure, which can be used to measure the size of a hypersurface of codimension one. In what follows we discuss two different ways to understand the relationship between Per_ε and Per.

One first possible way to relate Per_ε and Per is through a *pointwise* convergence analysis: fix a set A with regular enough boundary and then study the behavior of $\mathrm{Per}_\varepsilon(A)$ as $\varepsilon \to 0$. This is the type of analysis discussed in [GTM22], which was used by the authors to motivate the connection between adversarial robustness in binary classification and geometric variational problems involving perimeter. However, pointwise convergence of Per_ε toward Per is not sufficient to ensure that Per_ε induces a perimeter regularization type effect on its minimizers. For that, we need a different type of convergence.

A different way to compare the functionals $\mathrm{Per}_\varepsilon(\cdot)$ and Per is through a *variational* analysis; we refer the interested reader to the recent paper [BS22], which explains in detail this type of convergence and shows that Per_ε converges variationally toward Per, at least when balls are induced by the Euclidean distance. Here we restrict ourselves to discussing some of the mathematical implications of the analysis in [BS22], which considers problem (9) when $\mathcal{A}$ is the set $\mathfrak{B}(\mathbb{R}^d)$ of all (Borel) measurable subsets of $\mathcal{X} = \mathbb{R}^d$; this setting corresponds to an *agnostic* learner setting.

First, [BS22] shows that minimizers A_ε of (8) converge, as $\varepsilon \to 0$, toward minimizers of the problem

$$\min_{A \in \mathcal{A}_{\mathrm{Opt}}} \mathrm{Per}(A), \tag{10}$$

where $\mathcal{A}_{\mathrm{Opt}} := \operatorname{argmin}_{A' \in \mathfrak{B}(\mathbb{R}^d)} \mathbb{E}_{(x,y)\sim\mu}[|\mathbb{1}_{A'}(x) - y|]$. This means that, as $\varepsilon \to 0$, solutions to the adversarial training problem select among minimizers to the unrobust risk the ones with minimal perimeter. In particular, this result helps capture the idea that when ε is small, the presence of an adversary has the same effect as imposing a perimeter penalization term on the classifiers; c.f. [GTM22] for more discussion on this idea and on the relation with *mean curvature flows*.

A second consequence of the results in [BS22] is an expansion in ε for the adversarial risk R^*_ε, i.e., the minimum

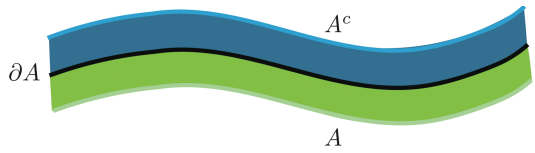

Figure 2. In green, the set of points in A within distance ε from the boundary of A; in blue, the set of points in A^c within distance ε from the boundary. The union of $\partial_\varepsilon A^c$ and $\partial_\varepsilon A$ is the region where the adversary may attack and guarantee a mismatch between predicted and true labels.

value of (AT). Precisely,

$$R^*_\varepsilon = R^*_0 + \varepsilon \mathrm{Per}^* + o(\varepsilon), \tag{11}$$

where Per^* denotes the minimum value in (10). The above result is reminiscent to Danskin's theorem for functions over Euclidean space, a theorem that is used to characterize the first "derivative" of a function defined as the infimum over a family of functions. Naturally, the difficulty in proving a result like (11) lies in the fact that the functionals of interest take subsets of $\mathcal{X}$ as input, and thus Danskin's theorem cannot be applied. Formula (11) can be used to give a geometric interpretation to the rate at which accuracy is lost when building robust classifiers as a function of the adversarial budget ε. Indeed, this result says that for small ϵ, accuracy is lost at the rate given by the Bayes classifier with minimal perimeter. The trade-off between accuracy and robustness has been investigated from statistical perspectives in [ZYJ+1909], for example. In contrast, the tools and concepts discussed here have a geometric and analytic flavor, with the caveat that they are only meaningful for a population-level analysis of adversarial robustness in binary classification. To gain an even better understanding of the situation, higher-order expansions of the adversarial risk in ε would be desirable and are a current topic of investigation.

2.2. **Certifiability and regularity.** In this section, we discuss notions of certifiability and regularity of robust binary classifiers. We begin with a definition.

Definition 2.1. Let $\mathbb{1}_A$ be a binary classifier. We say that $x \in \mathcal{X}$ is ε-certifiable (for $\mathbb{1}_A$) if $B_\varepsilon(x) \subseteq A$ or if $B_\varepsilon(x) \subseteq A^c$.

In simple terms, the certifiable points of a given classifier are the points in $\mathcal{X}$ for which the classification rule stays constant within the ball of radius ε around them: they are the points that are insensitive to the adversarial attacks in the adversarial problem (AT).

While it is not possible to build nontrivial sets A for which all points in $\mathcal{X}$ are certifiable, we can still ask whether it is possible to find robust classifiers that are fully characterized by their certifiable points. This motivates the following definitions.

Definition 2.2. We say that a measurable set A is ε inner regular if for all $x \in \partial A$ there exists $x' \in A$ such that $B_\varepsilon(x') \subset A$ and $x \in \partial B_\varepsilon(x')$. Likewise, we say that A is ε outer regular, if A^c is ε inner regular. Sets that are both inner and outer ε regular will be referred to as ε pseudo-certifiable.

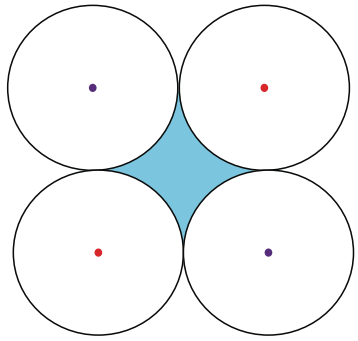

Figure 3. A data set for which no optimal robust classifier is ε pseudo-certifiable. If x is in the blue region and $x \in \partial B_\varepsilon$ for some ball of radius ε, then B_ε must intersect both a red and purple circle.

Notice that a classifier that is ε pseudo-certifiable is completely determined by its outputs on its certifiable points. Pseudo-certifiability is thus a desirable property. It is then natural to wonder whether it is always possible to construct an ε pseudo-certifiable classifier $\mathbb{1}_A$ minimizing the adversarial risk, i.e., a set A solving (AT) when the class of sets $\mathcal{A}$ is $\mathfrak{B}(\mathbb{R}^d)$. As it turns out, the notion of pseudo-certifiability is very strong, and in the case of the Euclidean distance, for example, it implies that decision boundaries are locally the graph of a $C^{1,1}$ function; see [BGTM23] and references therein. An example of a setting where no optimal robust classifier is ε pseudo-certifiable is given in Figure 3. There, μ is the sum of four delta measures in $\mathbb{R}^2$ at the points $(\pm\varepsilon, \pm\varepsilon)$, two red and two purple. Any optimal classifier must stay constant within the ε-balls centered at each point. The color choice in the shaded blue region does not affect optimality, however, there is no way to color this region and maintain ϵ-inner regularity for both sets (c.f. Figure 3).

While we cannot guarantee pseudo-certifiability for robust classifiers in general, we can still guarantee existence of ε-inner regular solutions, ε-outer regular solutions, and sometimes solutions with other forms of regularity. This is the content of a series of results in [BGTM23] stated informally below.

Theorem 2.3 (Informal from [BGTM23]). *Let μ be an arbitrary probability measure over $\mathcal{X} \times \{0,1\}$, and let $\mathcal{F}$ be the class of all Borel measurable classifiers. Let A be any solution to the (AT) problem. Then there exist two solutions A_I, A_O to* (AT) *such that $A_I \subseteq A \subseteq A_O$, and*

1. *A_I is ε inner-regular and A_O is ε outer-regular.*
2. *Any measurable set A' satisfying $A_I \subseteq A' \subseteq A_O$ is a solution to (AT).*

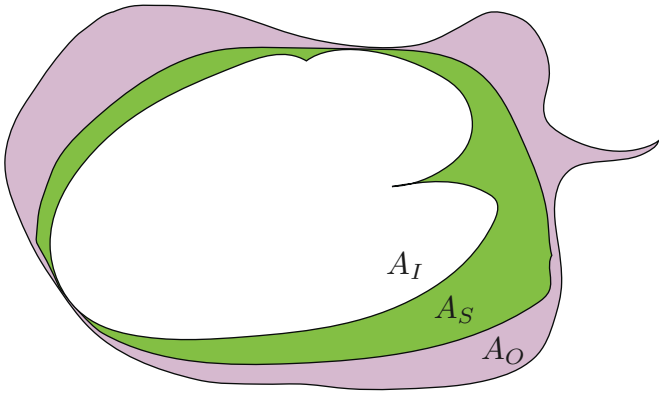

Figure 4. A set A_I that is ε-inner regular (some inwards cusps allowed), a set A_O that is ε-outer regular (some outward cusps allowed), and a smooth set A_S in between. In the context of Theorem 2.3 for Euclidean balls, the set A_S would be a solution to (AT).

Moreover, if $\mathcal{X}$ is an Euclidean space and the balls B_ε are induced by the Euclidean distance, then there exists a solution A to (AT) *such that the boundary of A is locally the graph of a $C^{1,1/3}$ function.*

The analysis in [BGTM23] also provides quantitative estimates for the regularity of the decision boundary of the classifier $\mathbb{1}_A$ in the last part of Theorem 2.3. In general, these estimates blow up when $\varepsilon \to 0$. It is however expected that one could get finer regularity estimates under additional assumptions on μ, e.g., assuming that $\mu_0 = \rho_0\,\mathrm{d}x$, $\mu_1 = \rho_1\,\mathrm{d}x$, and the set $\{x \in \mathbb{R}^d : \rho_1(x) = \rho_0(x)\}$ is sufficiently regular. Obtaining these finer estimates and characterizing the needed regularity for these finer estimates to apply are topics of current investigation.

3. Connections to Game Theory: DRO Formulations of AT

In this section, we introduce a framework for adversarial training that encompasses (AT) and that can be cast more precisely within game theory. In particular, in this larger framework we will be able to discuss the notion of Nash equilibrium in adversarial training and consider its implications on the robust training of learning models.

The idea is as follows. Instead of considering pointwise attacks as in (AT), where for every single data point (x, y) the adversary proposes an attack, we allow the adversary to modify μ by producing an entirely new data distribution $\tilde{\mu}$. Naturally, as in model (AT), the adversary must pay a price (or use a budget) for carrying out this modification. In precise terms, we consider the following families of problems:

$$\min_{f\in\mathcal{F}} \max_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})\ \text{s.t.}\ D(\mu,\tilde{\mu})\le\varepsilon} \mathbb{E}_{(\tilde{x},\tilde{y})}[\ell(f(\tilde{x}),\tilde{y})] \tag{12}$$

and

$$\min_{f\in\mathcal{F}} \max_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \mathbb{E}_{(\tilde{x},\tilde{y})}[\ell(f(\tilde{x}),\tilde{y})] - C(\mu,\tilde{\mu}). \tag{13}$$

The above are two instances of *distributionally robust optimization* problems given their inner maximization over probability measures. Notice that, in general, problem (12) can be written as (13) by defining the cost $C(\mu,\tilde{\mu})$ to be 0 if the explicit constraint $D(\mu,\tilde{\mu}) \le \varepsilon$ is satisfied and infinity otherwise. Both C and D can be interpreted as "distances" between probability distributions.

In the remainder of the paper, we will restrict our attention to problem (13) with a cost function C taking the form of an optimal transport problem:

$$C(\mu,\tilde{\mu}) = \inf_{\pi\in\Gamma(\mu,\tilde{\mu})} \int c_{\mathcal{Z}}(z,\tilde{z})\,d\pi(z,\tilde{z}), \tag{14}$$

for a cost function $c_{\mathcal{Z}} : \mathcal{Z}\times\mathcal{Z} \mapsto [0,\infty]$ that describes the marginal cost that the adversary must pay in order to move a clean data point z to a new location $\tilde{z}$ (recall that $\Gamma(\mu,\tilde{\mu})$ is the space of probability measures on $\mathcal{Z}\times\mathcal{Z}$ with first marginal μ and second marginal $\tilde{\mu}$). Notice that, in this generality, the adversary has the ability to modify both the feature vector x and the label y.

Some natural examples of cost functions $c_{\mathcal{Z}}$ are $c_{\mathcal{Z}}(z,\tilde{z}) := c_a|z-\tilde{z}|^2$, a choice that is particularly meaningful when $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$; here, c_a is a positive constant that can be interpreted as reciprocal to an adversarial budget. Another example of cost function $c_{\mathcal{Z}}$ of interest is (15) below, which can be used to rewrite problem (AT) in the form (13). Problem (13) is thus a rather general mathematical formulation for adversarial training.

Proposition 3.1 (Informal). *Problem* (AT) *is equivalent to problem* (13) *for a cost function C of the form* (14) *with marginal cost*

$$c_{\mathcal{Z}}(z,\tilde{z}) = \begin{cases} 0, & \text{if } d(x,\tilde{x}) \le \varepsilon \text{ and } y = \tilde{y} \\ \infty, & \text{otherwise}. \end{cases} \tag{15}$$

Proof. For a given f, let $\tilde{\mu}$ be a solution of the inner maximization problem in (13). Notice that, without the loss of generality, we can assume that $C(\mu,\tilde{\mu}) < \infty$, which means that there exists a coupling $\pi \in \Gamma(\mu,\tilde{\mu})$ whose support is contained in the set $\{(z,\tilde{z}) : y = \tilde{y}, \quad d(x,\tilde{x}) \le \varepsilon\}$. Due to this, we can write

$$\begin{aligned}\mathbb{E}_{\tilde{z}\sim\tilde{\mu}}[\ell(f(\tilde{x}),\tilde{y})] &= \mathbb{E}_{(z,\tilde{z})\sim\pi}[\ell(f(\tilde{x}),y)] \\ &\le \mathbb{E}_{(z,\tilde{z})\sim\pi}[\sup_{\tilde{x}'\in B_\varepsilon(x)} \ell(f(\tilde{x}'),y)] \\ &\le \mathbb{E}_{z\sim\mu}[\sup_{\tilde{x}'\in B_\varepsilon(x)} \ell(f(\tilde{x}'),y)].\end{aligned}$$

This shows that (13) $\le$ (AT).

On the other hand, for an arbitrary f and (x,y) in the support of μ, let $T_1(x,y) \in \operatorname{argmax}_{\tilde{x}\in B_\varepsilon(x)} \ell(f(\tilde{x}),y)$ (assuming, for simplicity, that the sup is indeed reached and that this operation can be defined in a measurable

way). We then define $T(x,y) = (T_1(x,y), y)$ and consider $\tilde{\mu} := T_\sharp \mu$. Notice that by construction we have $C(\mu, \tilde{\mu}) = 0$, from where it follows that

$$\mathbb{E}_{z\sim\mu}[\sup_{x'\in B_\varepsilon(x)} \ell(f(\tilde{x}'), y)] = \mathbb{E}_{\tilde{z}\sim\tilde{\mu}}[\ell(f(\tilde{x}), \tilde{y})] \leq \max_{\tilde{\mu}'\in\mathcal{P}(\mathcal{Z})} \mathbb{E}_{\tilde{z}\sim\tilde{\mu}'}[\ell(f(\tilde{x}), \tilde{y})] - C(\mu, \tilde{\mu}').$$

From this we can deduce (AT) $\leq$ (13).

□

Remark 3.2. Problem (AT) can also be written in the form (12). To see this, it is sufficient to define $D(\mu, \tilde{\mu})$ as the following ∞-Wasserstein distance in $\mathcal{Z}$:

$$W_\infty(\mu, \tilde{\mu}) = \inf_{\pi\in\Gamma(\mu,\tilde{\mu})} \operatorname{ess\,sup}_{(z,\tilde{z})\sim\pi} \delta(z, \tilde{z}),$$

where $\delta(z, \tilde{z}) = d(x, \tilde{x})$ if $y = \tilde{y}$, and $\delta(z, \tilde{z}) = \infty$ if $y \neq \tilde{y}$.

3.1. Nash equilibria in DRO.

One of the merits of writing adversarial training problems in the form (13) (or (12)) is that it allows us to explicitly interpret the process of robust training as a zero-sum game between two players, a learner and an adversary. In this interpretation, the learner's strategies consist of learning models $f \in \mathcal{F}$ (regression functions/classifiers), while the adversary's consist of data perturbations $\tilde{\mu}$. The payoff function for the adversary is set to be

$$\mathcal{U}(\tilde{\mu}, f) := \mathbb{E}_{\tilde{z}\sim\tilde{\mu}}[\ell(f(x), y)] - C(\mu, \tilde{\mu}), \tag{16}$$

and the adversary's goal is to maximize it, while the learner's goal is to minimize it.

We now recall the notion of a Nash equilibrium of a game, one of the central notions in game theory.

Definition 3.3. We say that $(\tilde{\mu}^*, f^*) \in \mathcal{P}(\mathcal{Z}) \times \mathcal{F}$ is a Nash equilibrium for the adversarial training game

$$\min_{f\in\mathcal{F}} \max_{\tilde{\mu}\in\mathcal{P}(Z)} \mathcal{U}(\tilde{\mu}, f), \tag{17}$$

if $\mathcal{U}(\tilde{\mu}, f^*) \leq \mathcal{U}(\tilde{\mu}^*, f)$ for all $f \in \mathcal{F}$ and all $\tilde{\mu} \in \mathcal{P}(\mathcal{Z})$.

For adversarial training, the theoretical existence of a Nash equilibrium $(\tilde{\mu}^*, f^*)$ means that if the learner were to choose model f^*, then its worst outcome would occur precisely if the adversary played $\tilde{\mu}^*$. This means that the learner would have no incentive to use a model different from f^* regardless of the adversary's attack. In addition, when Nash equilibria exist, the min and the max in (17) can be swapped and the apparent advantage that the adversary has over the learner in the formulation (17) (the adversary plays after observing the classifier chosen by the learner) is in fact only apparent. Existence of Nash equilibria for (17) thus means good news for the robust training of models provided one could actually compute one of them. Before we move on, it is important to highlight that for f^* to be useful in applications, one would need to make sure that the cost function C indeed restricts the adversary to consider only small perturbations of clean data points, but the exact meaning of "small perturbation" may be application dependent. In what follows, we put aside the challenges of modelling the cost function C and instead discuss the existence of Nash equilibria for (17) assuming C has been fixed (i.e., we have already determined how to model the adversary).

A well-known meta-result in game theory states that Nash equilibria for a game typically exist in the players' spaces of mixed strategies. In mathematical terms, this means that to prove existence of Nash equilibria of a given game one typically needs a convexification of the original space of strategies. For the adversarial training problem (17), since the adversary takes strategies in the space of probability measures $\mathcal{P}(\mathcal{Z})$, no convexification is needed for the adversary because $\mathcal{P}(\mathcal{Z})$ is already a convex space. On the other hand, the space $\mathcal{F}$ of classification/regression models may not be convex in general. One way to convexify $\mathcal{F}$ when $\mathcal{F}$ is a parametric family of models $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, the most standard setting in practice, is to consider a randomization of the classifiers/regression functions in the original $\mathcal{F}$; this is the approach taken in [MSP+2118]. Precisely, for the setting described in Proposition 3.1, and given a parametric family $\mathcal{F}$, the authors of [MSP+2118] show that the problem

$$\min_{\nu\in\mathcal{P}(\Theta)} \max_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \int_\Theta \mathcal{U}(f_\theta, \tilde{\mu})\, \mathrm{d}\nu(\theta) \tag{18}$$

admits Nash equilibria. Here ν can be interpreted as a mixed strategy for the original game and induces a regression function/classification rule as follows: given an input x, sample θ from ν and then evaluate $f_\theta(x)$.

Another approach to convexify the set $\mathcal{F}$, useful in the regression setting or when considering probabilistic classifiers, is to work with the space of aggregate models $\hat{\mathcal{F}} := \{\int_\Theta f_\theta(\cdot)\, \mathrm{d}\nu(\theta) : \nu \in \mathcal{P}(\Theta)\}$; notice that elements in this family can be directly interpreted as deterministic regression functions/probabilistic classifiers. In this setting, one considers the game

$$\min_{\nu\in\mathcal{P}(\Theta)} \max_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \mathcal{U}(\nu, \tilde{\mu}), \tag{19}$$

where we abuse notation slightly and write $\mathcal{U}(\nu, \tilde{\mu})$ to denote $\mathcal{U}(\int_\Theta f_\theta(\cdot)\, \mathrm{d}\nu(\theta), \tilde{\mu})$. The above setting is the one motivating the work [GG23].

While the convexification of the spaces of strategies is important, to guarantee the existence of Nash equlibria one also needs to make assumptions on the payoff function $\mathcal{U}$. Sion's theorem [Sio58], for example, a very general result that can be used to guarantee existence of Nash equilibria for rather general games, requires lower and upper semicontinuity of the payoff function $\mathcal{U}$ with respect to some topology, as well as some weaker form of

convexity/concavity of the payoff (a compactness property is required as well). It is actually not so difficult to check these assumptions for problems (18) and (19) when the spaces $\mathcal{P}(\mathcal{Z})$ and $\mathcal{P}(\Theta)$ are endowed with the topology of weak convergence of probability measures and the loss function ℓ in (16) is convex in its first argument; see [MSP+2118] and [GG23] for more details on these assumptions.

Works discussing the existence of Nash equilibria for a different variety of games go at least as far back as the work by von Neumann [vN59] (English translation from the original paper from 1928) and include other classical papers such as [Gli52, Sio58]. There are plenty of results in the literature that hold under a variety of assumptions that are worth discussing, and we will yet see another minmax result in Section 4, but discussing the extent of this topic is certainly beyond the scope of this paper.

3.2. **Greedy algorithms for DRO.** Existence results are statements made by an optimist: "there exists at least one Nash equilibrium, and thus there must be a way to find one…" the realist would immediately inquire how. In this section, we review a classical, perhaps the most popular, greedy algorithm that has been introduced in the literature to attempt solving minmax problems in Euclidean spaces. After that, we provide some pointers to recent literature where tools from optimal transport theory are used to adapt those greedy methods to solve minmax games in spaces of measures over continuum domains, examples of which are the DRO adversarial training problems (13) when $\mathcal{X}$ is a domain of $\mathbb{R}^d$, and Θ is, for example, the space of parameters of a neural network.

If the minmax problem that we were interested in was one of the form $\min_{p\in D}\max_{q\in E}\Phi(p,q)$, where D, E are subsets of two Euclidean spaces, a natural greedy strategy to alternate gradient ascent steps in the q coordinate and gradient descent steps in the p coordinate. This greedy algorithm is to minmax games what the gradient descent algorithm is for minimization problems. In continuous time, this *descent-ascent* approach can be interpreted as a system of ODEs of the form:

$$\begin{cases} \dot{p}_t = -\nabla_p \Phi(p_t, q_t) \\ \dot{q}_t = \nabla_q \Phi(p_t, q_t), \end{cases} \tag{20}$$

or projected versions thereof to guarantee that the dynamics stay within the feasible sets D and E. As can be expected, convergence of this scheme, especially toward Nash equilibria for the problem, depends on properties of the payoff function Φ (like for example strong convexity-concavity) or on whether one is interested in the behavior of (p_t, q_t) as $t \to \infty$ or in the behavior of average iterates $(\frac{1}{t}\int_0^t p_s\,\mathrm{d}s, \frac{1}{t}\int_0^t q_s\,\mathrm{d}s)$ as $t \to \infty$. We refer the reader to the recent works [LJJ2013, DP18], which discuss some of the existing literature on the topic and discuss drawbacks of and alternatives to gradient descent-ascent dynamics to solve minmax games.

While in general these potential issues about convergence may play against the use of ascent-descent schemes, they remain to be the simplest methods to consider for solving minmax games. Due to this, it is of interest to adapt them to the setting of problems (18) and (19) — the difficulty lies in the fact that now the dynamics must be defined in spaces of probability measures. Fortunately, the theory of optimal transport, which has experienced tremendous growth in the past two decades and has made its way into a variety of applications in a variety of fields (including machine learning), provides some useful avenues for carrying out this adaptation. Some of these ideas can be found in the works [GG23, WC22, Lu22]. For example, the recent work [GG23] discusses the use of optimal transport based dynamics to solve convex-concave adversarial training problems like (19). [WC22, Lu22], on the other hand, use optimal transport based dynamics to solve minmax games on spaces of measures with bilinear payoff structure and thus are more suited for problems such as (18). All works [GG23, WC22, Lu22] present some promising results on the convergence properties of their schemes, but, as it is discussed there, their theories remain far from complete. Designing schemes that can efficiently find Nash equilibria is an important question for adversarial training and for game theory at large.

4. Adversaries and Barycenters

In Section 2, we discussed how adversarial training can be seen as a perimeter minimization problem from the perspective of the learner in the case of binary classification. In this section, we will instead consider the *non-parametric problem* from the perspective of the adversary and show that the optimal adversarial strategy is given by solving a generalized barycenter problem among data distributions—an interpretation that holds regardless of the number of classes. This constitutes yet another piece of evidence that there is a rich geometric structure to adversarial learning. The discussion here will be a brief sketch of the main result from [GTKJ23].

Our starting point is the non-parametric, agnostic-classifier version of the DRO training problem introduced in Section 3. Here we will suppose that $y \in \mathcal{Y} = \{1, \dots, K\}$, ℓ is the 0-1 loss, and we allow the learner to choose *any* possible probabilistic classifier. The resulting adversarial training problem takes the form

$$\min_{f:\mathcal{X}\to S_y} \max_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}}[1 - f_{\tilde{y}}(\tilde{x})] - C(\mu, \tilde{\mu}). \tag{21}$$

While focusing on the non-parametric case and the 0,1 loss simplifies the problem, it is still a challenge to give an interpretation for (21) in its current form. To make progress,

we will eliminate the learner from the problem and obtain a pure maximization problem that determines the optimal strategy for the adversary. To do so, we will need to interchange the order of the min and max operations.

4.1. **Interchanging min and max.** As we discussed in Section 3, there are various well-known theorems, such as Sion's minimax theorem, that guarantee that the interchange of min and max does not affect the value of the problem or the optimal strategies for the players. The space of all probabilistic classifiers $\{f : \mathcal{X} \to S_y\}$ is convex and the 0-1 loss is linear with respect to both f and $\tilde{\mu}$. As a result, Sion's minimax theorem applies and the interchange is valid. From a mathematical perspective, switching the order allows us to simplify the objective as the minimization problem $\min_{f:\mathcal{X}\to S_y} \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}}[1-f_{\tilde{y}}(\tilde{x})]$ has a simple explicit solution.

To see this more clearly, we decompose the measure $\tilde{\mu}$ over the label space writing $\tilde{\mu} = (\tilde{\mu}_1, \dots, \tilde{\mu}_K)$. The previous line is then equivalent to

$$\min_{f:\mathcal{X}\to S_y} \sum_{y\in\{1,\dots,K\}} \mathbb{E}_{\tilde{x}\sim\tilde{\mu}_y}[1 - f_y(\tilde{x})]. \tag{22}$$

Obviously, the learner would like to choose f such that $f_y(\tilde{x}) = 1$. However, this may not always be possible. For instance, if $\tilde{x}$ belongs to the support of $\tilde{\mu}_1$ and $\tilde{\mu}_2$, then the learner must make a choice in order to respect the constraint $\sum_{y=1}^{K} f_y(\tilde{x}) = 1$. If $\tilde{\mu}_1$ gives more mass to $\tilde{x}$ than $\tilde{\mu}_2$, then it is best to choose $f_1(\tilde{x}) = 1$ (and vice versa in the other case); however, either way, the learner will have no choice but to classify some of the data incorrectly. In general, if a point $\tilde{x}$ belongs to the support of multiple measures, then the learner achieves the smallest value at $\tilde{x}$ by choosing $f(\tilde{x})$ to concentrate on the label y_* such that $\tilde{\mu}_{y_*}$ gives more mass to $\tilde{x}$ than any of the other measures (i.e., the mass from y_* is classified correctly and the rest of the data at $\tilde{x}$ is misclassified). This reveals an extremely important facet of the adversary's strategy: if the adversary can manipulate the data so that points from different classes are on top of one another, then the learner is forced to misclassify some of the data; furthermore, this effect gets stronger as the number of overlapping classes increases.

From the above considerations, it turns out that (22) is equal to

$$\max_{\lambda\in\mathcal{M}_+(\mathcal{X})} -\lambda(\mathcal{X}) + \sum_{y\in\{1,\dots K\}} \tilde{\mu}_y(\mathcal{X}) \text{ s.t. } \tilde{\mu}_y \le \lambda,$$

i.e., λ will be the smallest possible measure that lies above each of the $\tilde{\mu}_y$ (note that $\mathcal{M}_+(\mathcal{X})$ represents the space of all nonnegative Borel measures on $\mathcal{X}$). Note that since the adversary cannot change the number of data points (equivalently the total mass of the data) we must have $\sum_{y\in\{1,\dots,K\}} \tilde{\mu}_y(\mathcal{X}) = \sum_{y\in\{1,\dots,K\}} \mu_y(\mathcal{X})$, which is a positive constant that we will denote as M. Hence, after interchanging the min and the max, we can eliminate the learner and replace (25) with a problem that only considers the action of the adversary

$$\max_{\lambda\in\mathcal{M}_+(\mathcal{X}),\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} M - \lambda(\mathcal{X}) - C(\mu,\tilde{\mu}) \text{ s.t. } \tilde{\mu}_y \le \lambda. \tag{23}$$

Let us note that the quantity $M-\lambda(\mathcal{X})$ is precisely the adversarial risk. Hence, the adversary would like to maximize the risk, while respecting the constraints and not paying too much in the transportation cost $C(\mu,\tilde{\mu})$. In what follows, we will show that this problem can be viewed as a generalization of a barycenter problem with respect to the Wasserstein distance.

4.2. **Generalized barycenters.** Given K probability measures $\varrho_1, \dots, \varrho_K \in \mathcal{P}(\mathcal{X})$ and a cost $c : \mathcal{X}\times\mathcal{X} \to [0,\infty]$, the *Wasserstein barycenter problem* tries to find a measure ϱ_* such that the summed cost of transporting each of the ϱ_i onto ϱ_* (with respect to the optimal transport cost induced by c) is as small as possible. We now claim that problem (23) is a generalization of this barycenter problem when the adversary is not allowed to change class labels. In that case, the cost $C(\mu,\tilde{\mu})$ decomposes into a sum over each of the possible class labels $C(\mu,\tilde{\mu}) = \sum_{y\in\{1,\dots,K\}} C_{\mathcal{X}}(\mu_y,\tilde{\mu}_y)$ (where $C_{\mathcal{X}}$ is the optimal transport cost for measures defined over $\mathcal{X}$ rather than $\mathcal{Z}$). For λ fixed, let us write

$$\bar{C}(\mu_y,\lambda) \coloneqq \min_{\tilde{\mu}_y} C(\mu_y,\tilde{\mu}_y) \text{ s.t. } \tilde{\mu}_y \le \lambda,$$

and note that $\bar{C}(\mu_y,\lambda)$ represents the cheapest possible way to transport μ_y onto *some part* of λ. Using this notation, problem (23) becomes

$$\max_{\lambda\in\mathcal{M}_+(\mathcal{X})} M - \lambda(\mathcal{X}) - \sum_{y\in\{1,\dots,K\}} \bar{C}(\mu_y,\lambda), \tag{24}$$

which we will refer to as the generalized Wasserstein barycenter problem (GBP) and an optimal solution λ^* as a generalized Wasserstein barycenter.

In GBP, we try to find a nonnegative measure λ (no longer necessarily a probability measure) such that the total mass of λ plus the summed cost of transporting each μ_y onto some part of λ is as small as possible. To understand this in the context of adversarial training, let us consider two extreme choices for λ. In the first extreme case, let us choose $\lambda_1 = \sum_{y\in\{1,\dots,K\}} \mu_y$. With this choice, $\bar{C}(\mu_y,\lambda_1) = 0$ for all y, since μ_y is already part of λ_1 and hence we do not need to transport any mass. On the other hand, $M - \lambda_1(\mathcal{X}) = M - \sum_{y\in\{1,\dots,K\}} \mu_y(\mathcal{X}) = 0$. In other words, this choice produces 0 adversarial risk, meaning that the learner will be able to classify everything correctly (i.e., the adversary has not created any confusion between the classes). Clearly, this is a bad choice for the adversary even though the transportation cost is 0. In the second extreme case, λ_2, we try to make the adversarial risk $M-\lambda_2(\mathcal{X})$

as large as possible. Since each of the μ_y must be transported on to λ_2, we must have $\lambda_2(\mathcal{X}) \geq \max_y \mu_y(\mathcal{X})$. In order to avoid paying a large transportation cost, we want λ_2 to satisfy

$$\lambda_2 \in \operatorname*{argmin}_{\lambda \in \mathcal{M}_+(\mathcal{X}), \lambda(X) = \max_y \mu_y(\mathcal{X})} \sum_{y \in \{1,\dots,K\}} \bar{C}(\mu_y, \lambda).$$

When all of the μ_y have the same total mass, λ_2 will be a solution to the Wasserstein barycenter problem, since the condition $\lambda(X) = \mu_1(\mathcal{X}) = \cdots = \mu_K(\mathcal{X})$ means that each μ_y must transport all of its mass onto λ. In other words, to maximize the adversarial risk, the best thing the adversary can do is rearrange the data distributions for each class so that they all fully overlap on the same measure (note that this lines up with our insights from the previous subsection). In this case, the learner must necessarily misclassify $100 * \frac{M-1}{M}\%$ of the data as every location in $\mathcal{X}$ containing a data point will have an equally mixed fraction of each class. Note however, that this may not be the best overall adversarial strategy, as the costs $\bar{C}(\mu_y, \lambda_2)$ may outweigh the maximization of the adversarial risk. This is particularly the case when we consider the most relevant cost (15), where the adversarial budget parameter ε may make it literally impossible for the adversary to move each μ_y onto a single common distribution. In general, the optimal choice of λ is something between the two extremes offered by λ_1, λ_2 (c.f. Figure 5), in other words the adversary must balance the desire to maximize the risk against the imposition of the adversarial budget.

Now one may ask, what can we gain from understanding the optimal adversarial strategy through the lens of GBP? Furthermore, one might also wonder does this shed any light on adversarial learning outside of the non-parametric setting? First, let us highlight that the GBP connection allows us to use powerful tools from computational optimal transport to compute the optimal λ and hence the optimal adversarial strategy. Furthermore, because the adversary can only combine points that are distance at most ε away from one another, GBP appears to be computationally easier than the classical barycenter problem. Development of efficient algorithms that take advantage of the special structure of GBP is an ongoing work. Next, GBP reveals that the optimal adversarial strategy is strongly tied to the geometry of the data. Indeed, if λ is an optimal solution, then one can show that every point $\tilde{x} \in \operatorname{spt}(\lambda)$ there exists a set $A \subset \{1, \dots, K\}$ such that

$$\tilde{x} \in \operatorname*{argmin}_{x \in \mathcal{X}} \sum_{y \in A} c(x, x_y) \quad \text{for some } x_y \in \operatorname{spt}(\mu_y),$$

i.e., every point in $\operatorname{spt}(\lambda)$ is itself a barycenter (with respect to the distance c) of K or fewer points drawn from each of the μ_y. Hence the λ encodes local data (combining nearby points in different classes to get pointwise barycenters) as well as global data (choosing which points from different classes to combine). Finally, there are two ways in the non-parametric problem is still meaningful for the above model-specific problem where the learner is forced to choose from a parametric family of classifiers $\mathcal{F} \subset \{f : \mathcal{X} \to S_K\}$. The key insight is the fact that we always have the inequality

$$\min_{f : \mathcal{X} \to S_y} \max_{\mu \in \mathcal{P}(\mathcal{Z})} \mathbb{E}_{(\tilde{x}, \tilde{y}) \sim \mu}[1 - f_{\tilde{y}}(\tilde{x})] - C(\mu, \tilde{\mu}) \leq \min_{f \in \mathcal{F}} \max_{\tilde{\mu} \in \mathcal{P}(\mathcal{Z})} \mathbb{E}_{(\tilde{x}, \tilde{y}) \sim \mu}[1 - f_{\tilde{y}}(\tilde{x})] - C(\mu, \tilde{\mu}). \quad (25)$$

Hence, the non-parametric setting provides a universal lower bound on the adversarial risk and the perturbations $\tilde{\mu}_1, \dots, \tilde{\mu}_K, \lambda$ found in GBP are universally powerful attacks against any classifier. As a result, 1) the computable optimal $\tilde{\mu}_y$ can be used as a way to generate strong adversarial examples that could be used during training of any desired model; 2) the optimal value of (23) can serve as a benchmark for robust training within *any* family of models and help provide insight on how to choose the budget parameter ε properly.

5. Conclusions

In this paper we have discussed some recent analytic and geometric perspectives on adversarial training. Three key takeaways that we would like to highlight are: (1) learners respond to adversaries by choosing more regular decision boundaries, in particular boundaries with smaller perimeter (at least in the binary case), (2) adversarial training can be formulated as a game between two players, and (3) in the agnostic learner setting the optimal adversarial strategy to perturb the data is given by solving a generalized version of the Wasserstein barycenter problem. This can be summarized more glibly as learners minimize perimeter, adversaries find barycenters, together they arrive at a Nash equilibrium for their zero sum game. Extending these results to more general settings is an important open question. It would also be desirable to give finer estimates for the smoothness of decision boundaries beyond the $C^{1,\frac{1}{3}}$ result from Theorem 2.3. While we did not discuss it in this paper, lurking behind many of these variational problems are interesting PDEs whose analysis may shed further light on these problems. We hope that the discussion here will pique the interest of readers to continue adding results in this direction. There is much to still understand about these problems.

ACKNOWLEDGMENT. NGT is supported by the NSF grants DMS-2005797 and DMS-2236447.

References

[BKM19] Jose Blanchet, Yang Kang, and Karthyek Murthy, *Robust Wasserstein profile inference and applications to machine*

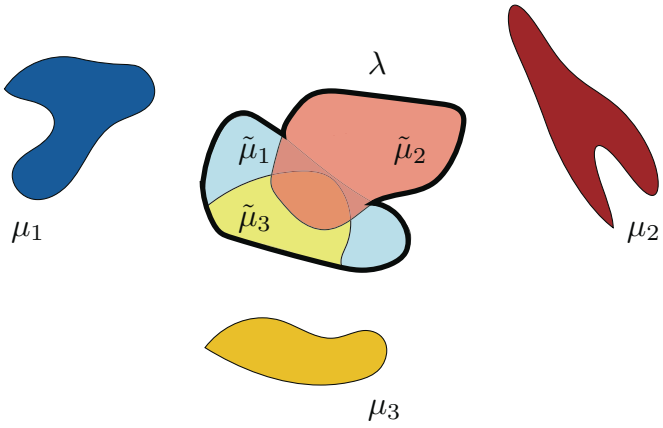

Figure 5. Illustration of a generalized barycenter λ for the measures μ_1, μ_2, μ_3 and the associated perturbations $\tilde{\mu}_y$. The smaller the total mass of λ, the better for the adversary. Since λ must lie above the μ_y, the only way to reduce the mass of λ is to make the $\tilde{\mu}_y$ overlap.

learning, J. Appl. Probab. **56** (2019), no. 3, 830–857, DOI 10.1017/jpr.2019.49. MR4015639

[BGTM23] Leon Bungert, Nicolás García Trillos, and Ryan Murray, *The geometry of adversarial training in binary classification*, Inf. Inference **12** (2023), no. 2, 921–968, DOI 10.1093/imaiai/iaac029. MR4565755

[BS22] Leon Bungert and Kerrek Stinson, *Gamma-convergence of a nonlocal perimeter arising in adversarial machine learning*, ArXiv Preprint (2022), available at arXiv:2211.15223.

[DP18] Constantinos Daskalakis and Ioannis Panageas, *The limit points of (optimistic) gradient descent in min-max optimization*, Advances in neural information processing systems, 2018.

[TT22] Camilo Andrés García Trillos and Nicolás García Trillos, *On the regularized risk of distributionally robust learning over deep neural networks*, Res. Math. Sci. **9** (2022), no. 3, Paper No. 54, 32, DOI 10.1007/s40687-022-00349-9. MR4468594

[GG23] Camilo Andrés García Trillos and Nicolás García Trillos, *On adversarial robustness and the use of wasserstein ascent-descent dynamics to enforce it*, ArXiv Preprint (2023), available at arXiv:2301.03662.

[GTKJ23] Nicolás García Trillos, Jakwang Kim, and Matt Jacobs, *The multimarginal optimal transport formulation of adversarial multiclass classification*, J. Mach. Learn. Res. **24** (2023), Paper No. 45, 56, DOI 10.4995/agt.2023.17046. MR4582467

[GTM22] Nicolás García Trillos and Ryan Murray, *Adversarial classification: necessary conditions and geometric flows*, J. Mach. Learn. Res. **23** (2022), Paper No. [187], 38. MR4577140

[Gli52] I. L. Glicksberg, *A further generalization of the Kakutani fixed theorem, with application to Nash equilibrium points*, Proc. Amer. Math. Soc. **3** (1952), 170–174, DOI 10.2307/2032478. MR46638

[GSS14] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, *Explaining and harnessing adversarial examples*, arXiv, 2014.

[LJJ2013] Tianyi Lin, Chi Jin, and Michael Jordan, *On gradient descent ascent for nonconvex-concave minimax problems*, Proceedings of the 37th international conference on machine learning, 202013, pp. 6083–6093.

[Lu22] Yulong Lu, *Two-scale gradient descent ascent dynamics finds mixed nash equilibria of continuous games: A mean-field perspective*, ArXiv Preprint (2022), available at arXiv:2212.08791.

[MMS+18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, *Towards deep learning models resistant to adversarial attacks*, 6th international conference on learning representations, ICLR 2018, proceedings, 2018.

[MSP+2118] Laurent Meunier, Meyer Scetbon, Rafael B Pinot, Jamal Atif, and Yann Chevaleyre, *Mixed nash equilibria in the adversarial examples game*, Proceedings of the 38th international conference on machine learning, 202118, pp. 7677–7687.

[Sio58] Maurice Sion, *On general minimax theorems*, Pacific J. Math. **8** (1958), 171–176. MR97026

[SZS+14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus, *Intriguing properties of neural networks*, 2nd international conference on learning representations, ICLR 2014, 2014.

[vN59] John von Neumann, *On the theory of games of strategy*, Contributions to the theory of games, Vol. IV, Annals of Mathematics Studies, no. 40, Princeton University Press, Princeton, N.J., 1959, pp. 13–42. MR0101828

[WC22] Guillaume Wang and Lénaïc Chizat, *An exponentially converging particle method for the mixed nash equilibrium of continuous games*, ArXiv Preprint (2022), available at arXiv:2211.01280.

[ZYJ+1909] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan, *Theoretically principled trade-off between robustness and accuracy*, Proceedings of the 36th international conference on machine learning, 201909, pp. 7472–7482.

Nicolás García Trillos

Matt Jacobs

Credits

The opening image is courtesy of peterschreiber.media via Getty.

Figures 1–5 are courtesy of the authors.

Photo of Nicolás García Trillos is courtesy of Nicolás García Trillos.

Photo of Matt Jacobs is courtesy of Matt Jacobs.