

# ADVERSARIAL CLASSIFICATION: NECESSARY CONDITIONS AND GEOMETRIC FLOWS

NICOLÁS GARCÍA TRILLOS AND RYAN MURRAY

ABSTRACT. We study a version of adversarial classification where an adversary is empowered to corrupt data inputs up to some distance  $\varepsilon$ , using tools from variational analysis. In particular, we describe necessary conditions associated with the optimal classifier subject to such an adversary. Using the necessary conditions, we derive a geometric evolution equation which can be used to track the change in classification boundaries as  $\varepsilon$  varies. This evolution equation may be described as an uncoupled system of differential equations in one dimension, or as a mean curvature type equation in higher dimension. In one dimension we rigorously prove that one can use the initial value problem starting from  $\varepsilon = 0$ , which is simply the Bayes classifier, in order to solve for the global minimizer of the adversarial problem. Numerical examples illustrating these ideas are also presented.

## 1. INTRODUCTION

In many learning settings, and in particular in the setting of deep learning, classifiers are known to behave poorly when exposed to adversarial examples. This has led to a significant body of work studying both the construction of specific adversaries and possible algorithms defending against them. Furthermore, the notion of pitting learners versus adversaries has stimulated significant new algorithms such as generative adversarial networks. One may view such adversarial frameworks as one possible notion of robustness of a learning algorithm, a critical concern in many applications.

In this work we consider the problem of optimal adversarial learning and aim at connecting it with a family of geometric evolution equations. The evolution equations that we derive answer the question: how would the decision boundary of a robust classifier change infinitesimally, if the adversary was to infinitesimally increase its power to perturb the data? Besides establishing new theoretical understanding for adversarial classification linking it with a set of geometric equations of surface diffusion type (similar to the ones describing the dynamics of interfaces of droplets of viscous fluids), our aim is also to explore computational alternatives to solve adversarial classification problems. At the theoretical level, a standard un-robust classification problem admits an explicit solution (i.e. the Bayes classifier), while adversarial problems typically do not have explicit solutions and in general are quite challenging from a numerical point of view.

While the general perspective that we have described above can be studied in a variety of settings, here we will study a concrete model for adversarial binary classification. In particular, we assume that a binary classifier is subject to a data perturbing adversary: namely, that for any future input  $x \in \mathbb{R}^d$  and associated output  $y \in \{0, 1\}$ , the adversary may select a new associated input  $\tilde{x} = x + \eta$  in order to disrupt a classifier. The adversary is assumed to possess limited power, namely that  $\|\eta\|_2 < \varepsilon$ , but is assumed to have knowledge of the classifier that has been chosen. A basic question is how such an adversary affects optimal classifiers. Various works have posited that adversaries do have an effect on classifiers, and that they can induce regular decision boundaries. Heuristically, from a geometric perspective this is natural, as boundaries with more surface area offer more opportunity for adversaries to disrupt classifiers. However, rigorous justification of this assertion is, to this point, unavailable. Several recent works have derived sufficient conditions for the adversarial learning problem with such an adversary. In

---

NGT was supported by NSF grant DMS 2005797.

particular, [3, 31] both derive a duality principle related to the optimal adversarial classifier. They use this to derive bounds on the effect on the loss of such an adversary. Such a duality principle provides an embedding of the optimal adversarial classification problem as an optimal transportation problem. However, even in light of such a duality principle, the solution of these adversarial problems for a particular  $\varepsilon$  requires the solution of a challenging optimization problem.

As mentioned earlier, despite the potential difficulty of solving the optimal adversarial classification problem for a fixed  $\varepsilon > 0$  via optimization, we notice that the solution of the problem for  $\varepsilon = 0$  is well-known and does not require optimization: the optimizer is the classical Bayes classifier. Namely, if we define

$$w_0\rho_0(x) = \mathbb{P}(X = x, Y = 0), \quad w_1\rho_1(x) = \mathbb{P}(X = x, Y = 1),$$

then the Bayes classifier given by

$$u_0(x) = \begin{cases} 1 & \text{if } w_1\rho_1(x) > w_0\rho_0(x) \\ 0 & \text{otherwise} \end{cases}$$

is known to be a minimizer of the un-robust risk. In the one dimensional case we expect to be able to write  $u_0(x) = \mathbb{1}_E$  for a set of the form  $E = \cup_{i=1}^K [a_i(0), b_i(0)]$ , where the “0” indicates that  $\varepsilon = 0$ . The central idea of this work is to derive evolution equations for the decision boundary of an optimal classifier as  $\varepsilon$  increases from zero, in the regime where we may construct optimal classifiers as a perturbation of the explicit Bayes classifier. This is achieved by deriving local necessary conditions (i.e. Euler-Lagrange type equations) for optimal adversarial classifiers for any fixed  $\varepsilon$  (4.1). In particular, in the one dimensional case, these necessary conditions take the form of the *algebraic equation*

$$w_1\rho_1(b_i(\varepsilon) - \varepsilon) = w_0\rho_0(b_i(\varepsilon) + \varepsilon).$$

Analogous necessary conditions are derived for the  $a_i$ . These necessary conditions are then used to derive evolution equations (4.2),(4.3). In particular, in one dimension this necessary condition takes the form of a decoupled, *ordinary differential equation* (ODE)

$$\frac{db_i}{d\varepsilon} = -\frac{w_0\rho'_0(b_i(\varepsilon) + \varepsilon) + w_1\rho'_1(b_i(\varepsilon) - \varepsilon)}{w_0\rho'_0(b_i(\varepsilon) + \varepsilon) - w_1\rho'_1(b_i(\varepsilon) - \varepsilon)},$$

with an analogous equation for the  $a_i$ . We remark that the resulting equation involves a sort of weak non-local algebraic condition, which in turn means the evolution equation includes a weak non-local forcing term. The evolution equation is ultimately a relatively simple decoupled ODE, which may then be solved directly using numerical solvers, with very modest computational effort and *no optimization*. This gives an easily computed candidate solution to the optimal adversarial classification problem for  $\varepsilon$  sufficiently close to zero.

As the equations that we derive are based upon necessary conditions, a natural question is whether solutions to the ODE indeed correspond to global minimizers of the optimal adversarial classification problems. Following the duality principle derived in [3][31] (which we extend here to include unbalanced classes), we derive the following theorem (stated informally):

**Theorem 1.1.** *In one dimension, under mild technical assumptions on  $w_0\rho_0, w_1\rho_1$  and the associated Bayes classifier, there exists an interval  $[0, \varepsilon_0]$  such that the solution of the optimal adversarial classification problem is given by the solution to the decoupled differential equations (4.2),(4.3) with initial values given by the decision boundary of the Bayes classifier (when  $\varepsilon = 0$ ).*

Subsequently, we turn our attention to studying the problem in higher dimensions, where decision boundaries are now expressed as hyper-surfaces. After deriving necessary conditions, which again take the form of weakly non-local algebraic equations (6.1), we derive an evolution equation for the decision boundary as  $\varepsilon$  varies (6.2). This equation is necessarily more complicated than in one dimension, but for  $\varepsilon$  small one can use a Taylor approximation to formally

reduce it to an evolution equation (6.3), which may be written as follows:

$$v(x) = -\frac{\nabla\rho \cdot \nu + \rho \sum_i \kappa_i}{(\nabla w_1 \rho_1 - \nabla w_0 \rho_0) \cdot \nu},$$

where here  $v$  represents the normal velocity (with respect to  $\varepsilon$ ) of a point on the decision boundary,  $\nu$  is the normal vector to the boundary,  $\kappa_i$  denote the principal curvatures of the boundary, and  $\rho = w_0 \rho_0 + w_1 \rho_1 = \mathbb{P}(X = x)$ . This evolution equation takes the form of a weighted *mean curvature flow* plus a biasing term (the biasing term is driven by the gradient of the distribution  $\rho$ ). Mean curvature flow is an important geometric flow with many nice properties, including a comparison principle, and is known in many instances to induce significant regularity to surfaces. In particular, mean curvature flow may be seen, within an appropriate function space, as a gradient flow of the perimeter functional (in particular a flow that aims at minimizing surface area). In addition, at least for the unweighted case, there are powerful and efficient numerical algorithms to compute mean curvature flows (i.e. the MBO scheme [28]). The geometric evolution equation that we derive suggest that as  $\varepsilon$  increases, the optimal decision boundaries will become shorter and smoother, supporting previous work on the topic. While the work relating to higher dimension in this work is largely formal, we believe that our derivations here raise many interesting and important theoretical and methodological questions, some of which are discussed in the conclusions section.

In summary, in this paper we view an adversarial problem as an ensemble of problems indexed by a parameter controlling the ability of an adversary to perturb the data. Then, starting from an un-robust optimal classifier one can evolve the corresponding decision boundary following a geometric equation to obtain a solution to the ensemble of adversarial problems. For the specific adversarial model that we study here the adversarial problem and its corresponding geometric evolution equations can be connected to a dual optimal transport problem. This connection is useful to certify global optimality of the decision boundaries generated by the geometric flow.

The remainder of this work is organized as follows: In section 2 we review some relevant literature. In section 3 we describe concretely the model that we consider. In section 3.1 we review and extend the duality principle related to the model. In sections 4 and 5 we derive the main results in one dimension. Subsequently, section 6 formally studies the higher-dimensional case. Finally, section 7 concludes by summarizing our work and describing a number of promising future directions.

## 2. RELATED LITERATURE

**2.1. Adversarial learning.** A significant body of recent work considers the problem of adversarial learning; we only aim to provide a review of the most relevant references. Early works focused on the existence of adversarial examples in deep learning [33, 19]. These examples typically involved adding carefully structured noise to images in ways that was imperceptible to humans, but which led to gross classification errors for fitted neural networks. A number of different algorithms were then developed for both constructing adversarial attacks and defending against them; these models are distinct from but related to the one we consider in this work. Several works advocate for attempting to differentiate between “natural” and “adversarial” inputs [17, 20, 29], while other works describe the ability of adversaries to circumvent such a defense [8, 2]. A parallel line of work posed a construction of improved classifiers by posing a game in which adversaries and classifiers iteratively try to best one another: this is the underlying framework for generative adversarial networks [18].

One work along this vein which relates closely with our work is [30]. That work observes that many boundaries obtained via robust classification are empirically observed to have smaller curvature. They then propose including a regularization term in classification that penalizes boundaries with higher curvature. Our work complements theirs in that we directly obtain a mean curvature in our d-dimensional evolution equation, indicating that the curvature indeed plays an explicit role in how decision boundaries change upon introducing stronger adversaries. While we do not explicitly prove that lower curvature is induced in our adversarial setting, the

evolution equation implicitly suggests that such is the case, and a rigorous connection between these notions is a topic of current work.

The fact that simple defenses were often insufficient against adversaries led to a number of theoretical works regarding the inherent difficulty of finding classifiers that are robust to adversaries. For example, [5] suggests that in some settings computation is the primary bottleneck in constructing adversarially robust classifiers. [16, 25, 32] all highlight how high dimensional geometry induces inherent limitations in the ability to avoid adversarial examples. [22] argues that adversarial examples are often based upon human derived notions of similarity that are incompatible with the geometry and training that occurs in deep learning.

While the above works highlight the difficulty of completely avoiding adversarial examples, they do not study the ability of classifiers to mitigate the effects of adversarial examples. One such framework for mitigating, on average, these effects is the optimal adversarial classification problem that we study here. Several variants of this problem have been previously studied. One variant permits the adversary to perturb the distribution of  $(x, y)$ 's that are inputted [4, 12]; in [4] a family of robust regression and classification problems are seen to be equivalent to a series of regularized risk minimization problems. A second variant, considered in both [3, 31], studies the data perturbing adversary. In particular, those works derive a duality principle relating the optimal classification problem for balanced classes to a optimal coupling or transportation problem. [31] uses Strassen's theorem from the theory of optimal transportation [35] to derive a duality principle, and demonstrates that minimizers of the adversarial problem may be taken to be closed sets. This may be seen as an initial step towards proving that optimal adversarial classifiers are indeed smoother than ones without adversaries. Finally, it is worth mentioning that other notions of classification robustness have been introduced in the literature [36]. Similar questions to the ones explored in this paper can also be studied under the setting proposed in that work.

**2.2. Geometric flows and PDE methods in learning.** Our work also draws upon ideas from geometric evolutions, and more generally variational problems. Mean curvature flow is well-studied from a theoretical standpoint, in particular as a gradient flow of the perimeter. Desirable properties of this flow, such as comparison principles, and local regularity theorems, are available in [11]. High fidelity numerical approximations are also available [28]. Our evolution equation is also not unrelated to non-local versions of curvature flow, which also are a topic of significant current interest [9].

In recent years, there has also been a growing interest in using the ideas and techniques from the analysis of interfacial flows to construct new algorithms in data analysis. These algorithms arise as iterative schemes to solve optimization problems closely related to graph-based supervised, unsupervised, and semi-supervised learning; see [6, 10, 21, 23, 26, 27, 34] and references within.

We also note that this work fits into a larger body of work that utilizes variational and PDE methods to better understand learning problems. For example, the duality principle that we derive in section 3.1 (which extends previous work by considering unbalanced cases), is analogous to the use of  $TL^p$  distances used in studying clustering problems [15]. The use of differential equation methods has received significant recent attention, for example in the study of empirical risk minimization [7, 13] and clustering problems [14], and has served to inspire many aspects of this work.

### 3. PROBLEM SETUP

Let  $\nu$  be a Borel probability measure on  $\mathbb{R}^d \times \{0, 1\}$  representing a data distribution for pairs  $(x, y)$  where  $x$  is a feature vector and  $y$  an associated label. Let  $(X, Y) \sim \nu$ . We assume that the conditional distribution of  $X$  given  $Y = 0$  takes the form  $\rho_0 dx$ , while the conditional distribution of  $X$  given  $Y = 1$  equals  $\rho_1 dx$ , for two density functions  $\rho_0, \rho_1$  that are assumed to satisfy certain regularity and non-degeneracy properties that we will make precise later on (for example see Assumptions 5.1 for the one dimensional setting). We use  $\rho dx$  to denote the

marginal distribution of  $X$ . Notice that  $\rho$  can be expressed as

$$\rho = w_0\rho_0 + w_1\rho_1,$$

where  $w_0 = \mathbb{P}(Y = 0)$  and  $w_1 = \mathbb{P}(Y = 1)$ . We let

$$\mu(x) := \mathbb{P}(Y = 1|X = x)$$

represent the conditional probability (or mean) of the label variable  $Y$  given  $X$ .

**Remark 3.1.** *Another way of writing  $\mu$  is:*

$$\mu(x) = \mathbb{P}(Y = 1) \cdot \frac{\rho_1(x)}{\rho(x)} = \frac{w_1\rho_1(x)}{\rho(x)}.$$

*This is a consequence of Bayes theorem.*

The classical classification problem seeks to minimize the functional

$$R(f) = \mathbb{E}(\ell(f(x), y)) = \int \ell(f(x), y) d\nu(x, y)$$

over some class of functions  $f \in \mathcal{F}$ . Usually, one is required to select  $f = \mathbb{1}_A$  for some Borel set  $A$ . Of particular importance is the case when  $\ell(f(x), y) = \mathbb{1}_{f(x) \neq y}$  (known as the 0-1 loss), where one may actually minimize over the class of  $L^1$  functions, and where minimizers of the form  $\mathbb{1}_A$  always exist. In particular, the function

$$u_B(x) = \begin{cases} 1 & \text{if } \mu(x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

known as the *Bayes classifier*, is a minimizer to the 0-1 loss problem. In short, at least from a theoretical perspective, the optimization of the risk functional  $R$  relative to 0-1 loss admits a closed form solution.

In the adversarial classification problem, one supposes an adversary that is able to modify incoming data points. In particular, in this paper we imagine that the adversary is allowed to shift any data point  $x$  with label  $y$  to a nearby point  $g(x, y)$  so that  $|x - g(x, y)| \leq \varepsilon$ . Here  $\varepsilon$  is a parameter that describes the power of the adversary: the larger the value of  $\varepsilon$ , the more the adversary can perturb the data. In this setting, one seeks to build a classifier that minimizes the robust risk

$$R_\varepsilon(f) := \sup_{g: |g(x, y) - x|_\infty \leq \varepsilon} \int \ell(f(g(x, y)), y) d\nu(x, y),$$

which factors in the action of the adversary. Notice that in the above model, the adversary can use information of a feature vector  $x$  as well as of its corresponding label  $y$  in order to decide on the new features for that data point. This model has been studied previously in [3, 31] where interesting connections with optimal transport problems have been established. In this paper we revisit these connections and extend them.

In order to analyze the minimization of the above robust risk, we first must characterize the  $g$  which achieves the maximum risk for a given  $f = \mathbb{1}_A$ . We begin by defining the distance between a point and a set  $A \in \mathcal{M}(\mathbb{R}^d)$  via

$$d(x, A) := \inf_{y \in A} |x - y|,$$

where  $\mathcal{M}(\mathbb{R}^d)$  denotes the Borel sets of  $\mathbb{R}^d$ . For convenience, we also define a signed distance via

$$\tilde{d}_A(x) = \begin{cases} d(x, A) & \text{if } x \notin A \\ -d(x, A^c) & \text{if } x \in A. \end{cases}$$

The maximization problem for the adversary admits a direct representation in terms of this signed distance. In particular, we notice that for  $f = \mathbb{1}_A$ , if  $|\tilde{d}_A(x)| \leq \varepsilon$ , then the adversary is free to select an arbitrary response at the point  $(x, y)$  regardless of the value of  $y$ . On the other hand, if  $|\tilde{d}_A(x)| > \varepsilon$  the adversary is unable to modify the label  $f(x)$  by moving the inputted

point by distance  $\varepsilon$ . This information may be encoded by rewriting our objective functional  $R_\varepsilon$  in the form:

$$R_\varepsilon(\mathbb{1}_A) = \int_{\tilde{d}_A(x) < -\varepsilon} \ell(1, y) d\nu(x, y) + \int_{\tilde{d}_A(x) > \varepsilon} \ell(0, y) d\nu(x, y) + \int_{|\tilde{d}_A(x)| < \varepsilon} \max_{z \in \{0,1\}} \ell(z, y) d\nu(x, y).$$

We notice that when  $\varepsilon = 0$  this functional reduces to the standard, non-adversarial, loss.

In order to simplify notation, we define, for any  $s \in \mathbb{R}$ , the set  $A^s := \{x \in \mathbb{R}^d : \tilde{d}_A(x) \leq s\}$ . Furthermore, in what follows we will always consider the 0-1 loss function. In that case, we may rewrite our objective function as follows:

$$\begin{aligned} R_\varepsilon(\mathbb{1}_A) &= \int_{A^{-\varepsilon}} w_0 \rho_0 dx + \int_{(A^\varepsilon)^c} w_1 \rho_1 dx + \int_{|\tilde{d}_A(x)| \leq \varepsilon} \rho(x) dx \\ &= \int_{A^\varepsilon} w_0 \rho_0 dx + w_1 - \int_{A^{-\varepsilon}} w_1 \rho_1 dx. \end{aligned}$$

We are interested in the robust classification problem:

$$(3.1) \quad \inf_{A \in \mathcal{M}(\mathbb{R}^d)} R_\varepsilon(\mathbb{1}_A).$$

**3.1. Duality principle and connection to an optimal transport problem.** Problem (3.1) admits a strong duality theorem. To illustrate, we recall previous results in [3, 31]. In those works, they consider  $w_0 = w_1 = 1/2$ , in which case the robust risk minimization problem becomes

$$\inf_{A \in \mathcal{M}(\mathbb{R}^d)} R_\varepsilon(\mathbb{1}_A) = \frac{1}{2} \left( 1 - \sup_{A \in \mathcal{M}(\mathbb{R}^d)} \left\{ \int_{A^{-\varepsilon}} \rho_1 dx - \int_{A^\varepsilon} \rho_0 dx \right\} \right).$$

It is then shown that

$$\sup_{A \in \mathcal{M}(\mathbb{R}^d)} \left\{ \int_{A^{-\varepsilon}} \rho_1 dx - \int_{A^\varepsilon} \rho_0 dx \right\} = \inf_{\pi \in \Gamma(\rho_1, \rho_0)} \int \mathbb{1}_{|x_1 - x_2| > 2\varepsilon} d\pi(x_1, x_2) =: d_\varepsilon(\rho_1, \rho_0),$$

where here  $\Gamma(\rho_1, \rho_0)$  denotes the set of probability measures on  $\mathbb{R}^d \times \mathbb{R}^d$  with marginals  $\rho_1$  and  $\rho_0$  (i.e. the set of couplings or transportation plans between  $\rho_1$  and  $\rho_0$ ); the above result is closely connected to Strassen's theorem (see Corollary 1.28 in [35]). This result may be restated in the following way

$$(3.2) \quad \inf_{A \in \mathcal{M}(\mathbb{R}^d)} R_\varepsilon(\mathbb{1}_A) = \sup_{\pi \in \Gamma(\rho_1, \rho_0)} \frac{1}{2} \left( 1 - \int \mathbb{1}_{|x_1 - x_2| > 2\varepsilon} d\pi(x_1, x_2) \right).$$

In order to state a duality principle for more general  $w_i$ , it will be convenient to define the probability measure on  $\mathbb{R}^d \times \{0, 1\}$  given by

$$\nu^S(E \times \{1\}) = \nu(E \times \{0\}), \quad \nu^S(F \times \{0\}) = \nu(F \times \{1\}).$$

In words,  $\nu^S$  is simply the data distribution after swapping the  $y$  labels. Using the measures  $\nu$  and  $\nu^S$ , we now state a more general duality principle that applies for arbitrary  $w_0, w_1$  and not just for  $w_0 = w_1 = 1/2$ .

**Proposition 3.2.** *Let  $c_\varepsilon : (\mathbb{R}^d \times \{0, 1\})^2 \rightarrow \mathbb{R}$  be the cost defined by*

$$c_\varepsilon(z_1, z_2) := \mathbb{1}_{\{|x_1 - x_2| > 2\varepsilon\} \cup \{y_1 \neq y_2\}},$$

where we write  $z_i = (x_i, y_i)$ . Then,

$$2 \sup_{B \in \mathcal{M}(\mathbb{R}^d)} \left\{ \int_{B^{-\varepsilon}} w_1 \rho_1 dx - \int_{B^\varepsilon} w_0 \rho_0 dx \right\} - w_1 + w_0 = \inf_{\pi \in \Gamma(\nu, \nu^S)} \int c_\varepsilon(z_1, z_2) d\pi(z_1, z_2),$$

which is also equal to

$$2 \sup_{A \in \mathcal{M}(\mathbb{R}^d)} \left\{ \int_{A^{-\varepsilon}} w_0 \rho_0 dx - \int_{A^\varepsilon} w_1 \rho_1 dx \right\} - w_0 + w_1.$$

*Proof.* We follow Theorem 1.27 in [35]. First, by the Kantorovich duality theorem (see Theorem 1.3 in [35]) we have

$$(3.3) \quad \sup_{\phi(z_1) + \psi(z_2) \leq c_\varepsilon(z_1, z_2)} \int \phi(z_1) d\nu(z_1) + \int \psi(z_2) d\nu^S(z_2) = \inf_{\pi \in \Gamma(\nu, \nu^S)} \int c_\varepsilon(z_1, z_2) d\pi(z_1, z_2).$$

where the sup is over all  $\phi \in L^1(\nu)$  and  $\psi \in L^1(\mu)$  (known as Kantorovich potentials), and the inequality constraint must be interpreted for  $\nu$  almost every  $z_1$  and for  $\nu^S$  almost every  $z_2$ .

Let  $\phi$  and  $\psi$  be two arbitrary Kantorovich potentials. Notice that if  $\phi(z) + \psi(\tilde{z}) \leq c_\varepsilon(z, \tilde{z})$  then necessarily  $\phi$  is (essentially) bounded above. By subtracting a constant from  $\phi$  and adding this same constant to  $\psi$ , we can assume without the loss of generality that  $\sup_z \phi(z) = 1$ . Now, for a given such  $\phi$  the best corresponding  $\psi$ , i.e. its dual conjugate potential, is given by

$$\phi^{c_\varepsilon}(\tilde{z}) := \inf_z \{c_\varepsilon(z, \tilde{z}) - \phi(z)\}.$$

Notice that  $\phi^{c_\varepsilon}$  can be written as:

$$\begin{aligned} \phi^{c_\varepsilon}(\tilde{x}, 0) &= \min \left\{ 1 - \sup_{x: |\tilde{x}-x| > 2\varepsilon} \phi(x, 0), - \sup_{x: |\tilde{x}-x| \leq 2\varepsilon} \phi(x, 0), 1 - \sup_x \phi(x, 1) \right\}, \\ \phi^{c_\varepsilon}(\tilde{x}, 1) &= \min \left\{ 1 - \sup_{x: |\tilde{x}-x| > 2\varepsilon} \phi(x, 1), - \sup_{x: |\tilde{x}-x| \leq 2\varepsilon} \phi(x, 1), 1 - \sup_x \phi(x, 0) \right\}. \end{aligned}$$

Since we have assumed that  $\sup_z \phi(z) = 1$  we can deduce from the above that  $\phi^{c_\varepsilon}(\tilde{z}) \in [-1, 0]$ . In particular, the supremum in (3.3) can be restricted to pairs  $\phi, \psi$  satisfying the cost constraint and  $\psi \in [-1, 0]$ .

Let us now consider an arbitrary  $\psi \in [-1, 0]$  with its best associated  $\phi$ :

$$\begin{aligned} \psi^{c_\varepsilon}(x, 0) &:= \min \left\{ 1 - \sup_{\tilde{x}: |\tilde{x}-x| > 2\varepsilon} \psi(\tilde{x}, 0), - \sup_{\tilde{x}: |\tilde{x}-x| \leq 2\varepsilon} \psi(\tilde{x}, 0), 1 - \sup_{\tilde{x}} \psi(\tilde{x}, 1) \right\}, \\ \psi^{c_\varepsilon}(x, 1) &:= \min \left\{ 1 - \sup_{\tilde{x}: |\tilde{x}-x| > 2\varepsilon} \psi(\tilde{x}, 1), - \sup_{\tilde{x}: |\tilde{x}-x| \leq 2\varepsilon} \psi(\tilde{x}, 1), 1 - \sup_{\tilde{x}} \psi(\tilde{x}, 0) \right\}. \end{aligned}$$

Since  $\psi$  is negative, it follows that

$$\psi^{c_\varepsilon}(x, 0) = - \sup_{\tilde{x}: |\tilde{x}-x| \leq 2\varepsilon} \psi(\tilde{x}, 0), \quad \psi^{c_\varepsilon}(x, 1) = - \sup_{\tilde{x}: |\tilde{x}-x| \leq 2\varepsilon} \psi(\tilde{x}, 1),$$

which in particular implies that  $\psi^{c_\varepsilon} \in [0, 1]$ . Finally, computing the conjugate of  $\phi := \psi^{c_\varepsilon}$  we get

$$\phi^{c_\varepsilon}(\tilde{x}, 0) = - \sup_{x: |\tilde{x}-x| \leq 2\varepsilon} \phi(x, 0), \quad \phi^{c_\varepsilon}(\tilde{x}, 1) = - \sup_{x: |\tilde{x}-x| \leq 2\varepsilon} \phi(x, 1)$$

which is then seen to take values on  $[-1, 0]$ . Since  $\phi^{c_\varepsilon}$  is the best  $\psi$  for a given  $\phi \in [0, 1]$ , it follows that the supremum in (3.3) is equal to

$$\sup_{\phi \in [0, 1]} \int \phi(z) d\nu(z) + \int \phi^{c_\varepsilon}(\tilde{z}) d\nu^S(z).$$

From the fact that for arbitrary  $\phi \in [0, 1]$  we have  $\phi^{c_\varepsilon} \in [-1, 0]$ , we deduce, using the ‘‘layer cake’’ representation,

$$\begin{aligned} \int \phi(z) d\nu(z) + \int \phi^{c_\varepsilon}(\tilde{z}) d\nu^S(\tilde{z}) &= \int_0^1 \int \mathbf{1}_{\phi(z) > s} d\nu(z) ds - \int_0^1 \int \mathbf{1}_{-\phi^{c_\varepsilon}(\tilde{z}) > s} d\nu^S(\tilde{z}) ds, \\ (3.4) \quad &= \int_0^1 \left( \int \mathbf{1}_{\phi(z) > s} d\nu(z) ds - \int \mathbf{1}_{-\phi^{c_\varepsilon}(\tilde{z}) > s} d\nu^S(\tilde{z}) \right) ds. \end{aligned}$$

We now rewrite the indicator function  $\mathbb{1}_{-\phi^{c_\varepsilon}(\tilde{z}) \geq s}$  in terms of a  $2\varepsilon$ -expansion of a set. Indeed, for  $\tilde{z} = (\tilde{x}, 0)$  we have:

$$\begin{aligned}\mathbb{1}_{\{-\phi^{c_\varepsilon}(\cdot) > s\}}(\tilde{z}) &= 1 \Leftrightarrow -\phi^{c_\varepsilon}(\tilde{x}, 0) > s \\ &\Leftrightarrow \exists x \text{ s.t. } |x - \tilde{x}| \leq 2\varepsilon \text{ and } \phi(x, 0) > s \\ &\Leftrightarrow \tilde{x} \in \{x : \phi(x, 0) > s\}^{2\varepsilon}.\end{aligned}$$

Thus,  $\mathbb{1}_{\{-\phi^{c_\varepsilon}(\cdot) > s\}}(\tilde{x}, 0) = \mathbb{1}_{\{\phi(\cdot, 0) > s\}^{2\varepsilon}}(\tilde{x})$ . In the exact same way we see that  $\mathbb{1}_{\{-\phi^{c_\varepsilon}(\cdot) > s\}}(\tilde{x}, 1) = \mathbb{1}_{\{\phi(\cdot, 1) > s\}^{2\varepsilon}}(\tilde{x})$ . Since we are integrating over  $s \in [0, 1]$ , we may infer that there exists  $s \in [0, 1]$  such that

$$\begin{aligned}\int_0^1 \left( \int \mathbb{1}_{\phi(z) > s} d\nu(z) ds - \int \mathbb{1}_{-\phi^{c_\varepsilon}(\tilde{z}) > s} d\nu^S(\tilde{z}) \right) ds &\leq \int \mathbb{1}_{\phi(z) > s} d\nu(z) ds - \int \mathbb{1}_{-\phi^{c_\varepsilon}(\tilde{z}) > s} d\nu^S(\tilde{z}) \\ &= \int \mathbb{1}_{\{\phi(x, 0) > s\}} w_0 \rho_0(x) dx + \int \mathbb{1}_{\{\phi(x, 1) > s\}} w_1 \rho_1(x) dx \\ &\quad - \int \mathbb{1}_{\{\phi(x, 0) > s\}^{2\varepsilon}} w_1 \rho_1(x) dx - \int \mathbb{1}_{\{\phi(x, 1) > s\}^{2\varepsilon}} w_0 \rho_0(x) dx,\end{aligned}$$

where we have used the definitions of  $\nu$  and  $\nu^S$ . The above computations, along with (3.4), allow us to conclude that:

$$\begin{aligned}&\inf_{\pi \in \Gamma(\nu, \nu^S)} \int c_\varepsilon(z_1, z_2) d\pi(z_1, z_2) \\ &= \sup_{A \in \mathcal{M}(\mathbb{R}^d)} \left\{ \int_A w_0 \rho_0 dx - \int_{A^{2\varepsilon}} w_1 \rho_1 dx \right\} + \sup_{B \in \mathcal{M}(\mathbb{R}^d)} \left\{ \int_B w_1 \rho_1 dx - \int_{B^{2\varepsilon}} w_0 \rho_0 dx \right\} \\ &= \sup_{A \in \mathcal{M}(\mathbb{R}^d)} \left\{ \int_{A^{-\varepsilon}} w_0 \rho_0 dx - \int_{A^\varepsilon} w_1 \rho_1 dx \right\} + \sup_{B \in \mathcal{M}(\mathbb{R}^d)} \left\{ \int_{B^{-\varepsilon}} w_1 \rho_1 dx - \int_{B^\varepsilon} w_0 \rho_0 dx \right\} \\ &= \sup_{A \in \mathcal{M}(\mathbb{R}^d)} \left\{ \int_{A^{c-\varepsilon}} w_1 \rho_1 dx - \int_{A^{c\varepsilon}} w_0 \rho_0 dx \right\} + \sup_{B \in \mathcal{M}(\mathbb{R}^d)} \left\{ \int_{B^{-\varepsilon}} w_1 \rho_1 dx - \int_{B^\varepsilon} w_0 \rho_0 dx \right\} - w_1 + w_0 \\ &= 2 \sup_{B \in \mathcal{M}(\mathbb{R}^d)} \left\{ \int_{B^{-\varepsilon}} w_1 \rho_1 dx - \int_{B^\varepsilon} w_0 \rho_0 dx \right\} - w_1 + w_0.\end{aligned}$$

Notice that we also obtain:

$$= 2 \sup_{A \in \mathcal{M}(\mathbb{R}^d)} \left\{ \int_{A^{-\varepsilon}} w_0 \rho_0 dx - \int_{A^\varepsilon} w_1 \rho_1 dx \right\} - w_0 + w_1.$$

This shows our desired result. □

**Corollary 3.3.**  $\mathbb{1}_A$  for some  $A \in \mathcal{M}(\mathbb{R}^d)$  minimizes  $R_\varepsilon$  if and only if  $A$  maximizes

$$\sup_{A \in \mathcal{M}(\mathbb{R}^d)} \left\{ w_1 \int_{A^{-\varepsilon}} \rho_1 dx - w_0 \int_{A^\varepsilon} \rho_0 dx \right\}.$$

Moreover,

$$\inf_{A \in \mathcal{M}(\mathbb{R}^d)} R_\varepsilon(\mathbb{1}_A) = \frac{1}{2} - \frac{1}{2} \inf_{\pi \in \Gamma(\nu, \nu^S)} \int c_\varepsilon(z_1, z_2) d\pi(z_1, z_2).$$

*Proof.* Recall that

$$R_\varepsilon(\mathbb{1}_A) = \int_{A^\varepsilon} w_0 \rho_0 dx + w_1 - \int_{A^{-\varepsilon}} w_1 \rho_1 dx$$

so

$$\inf_{A \in \mathcal{M}(\mathbb{R}^d)} R_\varepsilon(\mathbb{1}_A) = w_1 - \sup_A \left\{ \int_{A^{-\varepsilon}} w_1 \rho_1 dx - \int_{A^\varepsilon} w_0 \rho_0 dx \right\}$$

$$= w_1 - \frac{1}{2}(w_1 - w_0) - \frac{1}{2} \inf_{\pi \in \Gamma(\nu, \nu^S)} c_\varepsilon(z_1, z_2) d\pi(z_1, z_2)$$

□

**Remark 3.4.** Let us consider the balanced case  $w_0 = w_1 = 1/2$ . Since

$$\inf_{A \in \mathcal{M}(\mathbb{R}^d)} R_\varepsilon(A) = \frac{1}{2} \left( 1 - \inf_{\pi \in \Gamma(\nu, \nu^S)} c_\varepsilon(z_1, z_2) d\pi(z_1, z_2) \right),$$

it follows that

$$\inf_{\pi \in \Gamma(\nu, \nu^S)} \int c_\varepsilon(z_1, z_2) d\pi(z_1, z_2) = \inf_{\gamma \in \Gamma(\rho_0, \rho_1)} \int \mathbf{1}_{|x_1 - x_2| > 2\varepsilon} d\gamma(x_1, x_2).$$

#### 4. NECESSARY CONDITIONS AND CORRESPONDING EVOLUTION EQUATION IN ONE DIMENSION

We now describe, in detail, the necessary conditions for minimizing  $R_\varepsilon$ , and the evolution equation that they induce. For clarity, we begin by describing this evolution equation in the simple case where  $x \in \mathbb{R}$ . In this case we will be able to prove that the resulting evolution equation completely characterizes the global minimizer of  $R_\varepsilon$  for small  $\varepsilon$  under mild assumptions; the formal statement and proof of this result is given in section 5. Subsequently, in section 6 we will turn our attention to the case where  $x \in \mathbb{R}^d$ .

To begin, let us assume that we may represent the boundary of the optimal set  $A_\varepsilon^*$  in terms of two collections of points  $a_i$  and  $b_i$ , so that  $A_\varepsilon^* = \cup_{i=1}^K [a_i, b_i]$ . Here we allow  $a_1 = -\infty$  and  $b_K = +\infty$  if necessary, and we notice that, as  $w_0\rho_0, w_1\rho_1$  are both absolutely continuous (see Assumptions 5.1), it makes no difference whether the sub-intervals are open or closed. We note that this assumption will hold for  $\varepsilon = 0$  as long as the set where  $w_0\rho_0 = w_1\rho_1$  is a discrete set, a mild assumption. Finally, in the remainder we may suppress the dependence of  $a_i, b_i$  on  $\varepsilon$ , in order to decrease the notational burden. We use the convention  $a_1 < b_1 < a_2 < b_2 < \dots < a_K < b_K$ .

As  $A_\varepsilon^*$  is a minimizer of  $R_\varepsilon$ , we may freely perturb the boundary points (i.e.  $a_i, b_i$ ) without increasing the energy. In particular, for  $|\delta|$  small enough, if we consider the set  $A(\delta) = (a_1, b_1 + \delta) \cup (\cup_{i=2}^K (a_i, b_i))$ , then since  $A_\varepsilon^*$  is a minimizer we have that  $R_\varepsilon(A(\delta)) - R_\varepsilon(A_\varepsilon^*) \geq 0$ . Taking  $\delta \rightarrow 0$  and using the fundamental theorem of Calculus then allows us to write

$$\begin{aligned} 0 &= \lim_{\delta \rightarrow 0} \frac{R_\varepsilon(A(\delta)) - R_\varepsilon(A_\varepsilon^*)}{\delta} \\ &= w_0\rho_0(b_1 + \varepsilon) - w_1\rho_1(b_1 - \varepsilon). \end{aligned}$$

An analogous argument for the  $a_i$  and for the rest of the  $b_i$  then allows us to write the necessary conditions:

$$(4.1) \quad w_1\rho_1(b_i - \varepsilon) = w_0\rho_0(b_i + \varepsilon), \quad w_1\rho_1(a_i + \varepsilon) = w_0\rho_0(a_i - \varepsilon),$$

which hold for all  $a_i$  and  $b_i$  that are not  $-\infty$  or  $+\infty$ . In the remainder, if  $a_1(0) = -\infty$  we set  $a_1(\varepsilon) = -\infty$  for  $\varepsilon > 0$  and likewise if  $b_K(0) = +\infty$ , then  $b_K(\varepsilon) = +\infty$ . This relates to the fact that our differential equation approach does not track potential “topological changes” in the decision boundaries, and is mostly focused on the case where  $\varepsilon$  is small. We remark that when  $\varepsilon = 0$ , the above necessary condition gives precisely  $w_0\rho_0 = w_1\rho_1$ , which characterizes the boundary points of the Bayes classifier. In a sense, we may view the necessary condition above as a *non-local algebraic* condition: namely, that the condition that  $w_0\rho_0(b_i) = w_1\rho_1(b_i)$  (for  $\varepsilon = 0$ ) has been replaced by the non-local algebraic condition  $w_0\rho_0(b_i + \varepsilon) = w_1\rho_1(b_i - \varepsilon)$  (for  $\varepsilon > 0$ ).

Using the necessary conditions (4.1), we can exactly describe the local evolution of the boundary of the set  $A_\varepsilon^*$  for small changes in  $\varepsilon$ . In particular, let us suppose that each boundary point

varies smoothly in  $\varepsilon$ , namely that we express  $a_i(\varepsilon)$  and  $b_i(\varepsilon)$  as smooth functions in  $\varepsilon$ . Differentiating the necessary condition and using the chain rule, we find that

$$w_0\rho'_0(b_i + \varepsilon) \left( \frac{db_i}{d\varepsilon} + 1 \right) - w_1\rho'_1(b_i - \varepsilon) \left( \frac{db_i}{d\varepsilon} - 1 \right) = 0.$$

We may then solve this equation for  $\frac{db_i}{d\varepsilon}$ ,

$$(4.2) \quad \frac{db_i}{d\varepsilon} = - \frac{w_0\rho'_0(b_i + \varepsilon) + w_1\rho'_1(b_i - \varepsilon)}{w_0\rho'_0(b_i + \varepsilon) - w_1\rho'_1(b_i - \varepsilon)}.$$

The necessary condition for the  $a_i$  is analogous:

$$(4.3) \quad \frac{da_i}{d\varepsilon} = - \frac{w_1\rho'_1(a_i + \varepsilon) + w_0\rho'_0(a_i - \varepsilon)}{w_1\rho'_1(a_i + \varepsilon) - w_0\rho'_0(a_i - \varepsilon)}.$$

We continue to use the convention that  $a_1 = -\infty$  when  $a_1(0) = -\infty$  and similarly  $b_K = +\infty$  if  $b_K(0) = +\infty$ .

The previous equations allow us to precisely describe (locally) the evolution of the decision boundaries using ordinary differential equations. In particular, beginning at  $\varepsilon = 0$  with the decision boundary of the Bayes classifier, we may directly solve for the optimizer of  $R_\varepsilon$  by solving a system of at most  $2K$  decoupled differential equations. High fidelity approximations of these equations may be obtained using standard software packages.

We remark that the differential equations at  $\varepsilon = 0$  are much simpler, for example:

$$(4.4) \quad \frac{db_i}{d\varepsilon}[\varepsilon = 0] = - \frac{\rho'(b_i)}{w_0\rho'_0(b_i) - w_1\rho'_1(b_i)}.$$

This indicates that the  $b_i$  initially moves downhill in  $\rho$ , with speed dictated by the inverse of the derivative of the difference between the probability of the different classes. To determine the sign of the denominator, we notice that since  $w_1\rho_1$  is assumed to be larger than  $w_0\rho_0$  inside  $(a_i(0), b_i(0))$ , it is natural to assume that  $w_1\rho'_1(b_i(0)) < w_0\rho'_0(b_i(0))$ . This assumption is made explicit in Assumption 5.1). A similar conclusion holds for the left endpoints  $a_i$ . Although the above non-local formulas are not too complicated here, the analogous approximation near  $\varepsilon = 0$  will be more important in understanding the geometric flow induced in dimension higher than one as we will see in section 6.

**4.1. Simple example.** Suppose that  $\mathbb{P}(X = x|Y = 1) = \phi(x)dx$ , where  $\phi$  is the standard normal density  $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ , and let  $\mathbb{P}(X = x|Y = 0) = \frac{\phi((x-2)/2)dx}{2}$ . Assume also that  $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 0) = 1/2$ . Since the variances of the two Gaussians  $\mathbb{P}(X = x|Y = 0)$  and  $\mathbb{P}(X = x|Y = 1)$  are different, their densities must intersect at exactly two points, in this case at

$$a_1(0) = -\frac{2}{3} \left( 1 + \sqrt{2(2 + 3 \log(2))} \right) \approx -2.57, \quad b_1(0) = \frac{2}{3} \left( \sqrt{2(2 + 3 \log(2))} - 1 \right) \approx 1.23.$$

The corresponding Bayes classifier for this problem is the indicator function of the set  $(a_1(0), b_1(0))$ .

Since the solutions  $a_1$  and  $b_1$  of the ODEs (4.2) and (4.3) satisfy the necessary conditions (4.1), it follows from Theorem 2 in [31] (which characterizes optimality in the Gaussian setting) that the set  $A_\varepsilon^* := (a_1(\varepsilon), b_1(\varepsilon))$  is a global solution to the adversarial robust problem (3.1) for all  $\varepsilon$  small enough.

In order to provide concrete numerical values for the decision boundary as a function of  $\varepsilon$  we use a standard ODE solver in Python. The decision boundary, as well as the associated densities, are given in Figure 1.

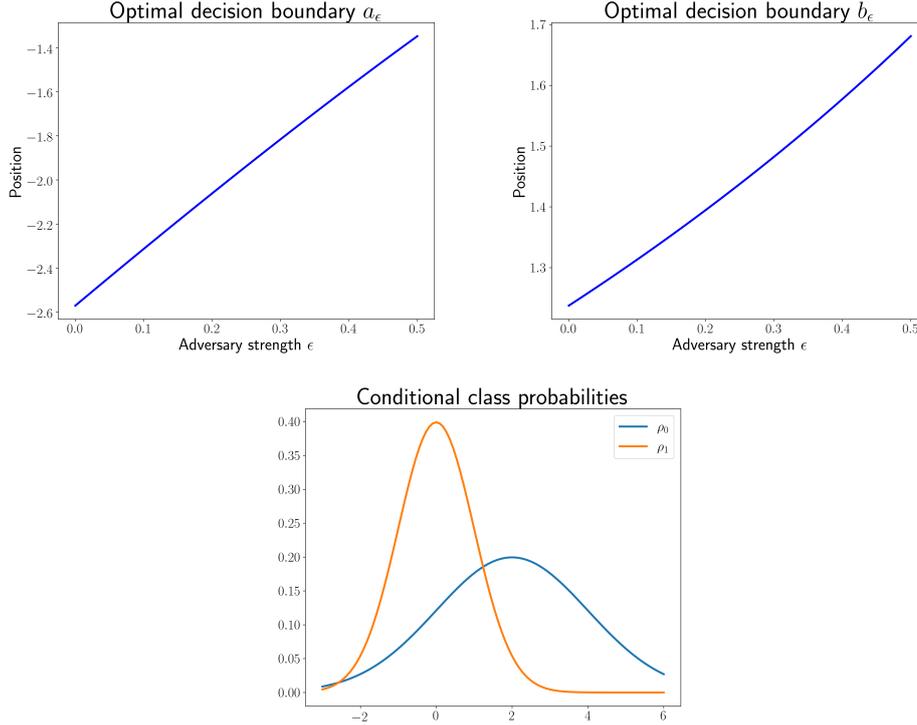


FIGURE 1. Plot of decision boundaries as  $\varepsilon$  varies for example in section 4.1 , as well as the underlying probabilities.

## 5. GLOBAL MINIMIZERS IN ONE DIMENSION

The evolution equations of the previous section are based upon necessary conditions for the adversarial classification problem. Since they are based upon necessary conditions, it is not immediately obvious whether or not these solutions are global minimizers of the adversarial variational problem (3.1). The goal of this section is to prove that solutions of the evolution equation are indeed global minimizers for all small enough  $\varepsilon$ , or in other words that the evolution equation locally characterizes the minimizers of the adversarial problem. In order to do so, we will require the following mild assumptions on the densities  $w_0\rho_0, w_1\rho_1$ :

**Assumption 5.1.** *We make the following assumptions on the densities  $\rho_0$  and  $\rho_1$ .*

- i) *Regularity condition:*  $\rho_0, \rho_1 \in C^1(\mathbb{R})$ .
- ii) *Non-degeneracy condition I:* there are only finitely many  $t \in \mathbb{R}$  for which  $w_0\rho_0(t) = w_1\rho_1(t) > 0$ .
- iii) *Non-degeneracy condition II:* for every  $t \in \mathbb{R}$  for which  $w_0\rho_0(t) = w_1\rho_1(t) > 0$  we have  $w_0\rho_0'(t) \neq w_1\rho_1'(t)$ .

Before stating the main result of this section, we begin with a few remarks which will be important in our proof strategy.

**Remark 5.2** (Global optimality via duality). *Suppose that  $A$  is a measurable subset of  $\mathbb{R}^d$  that satisfies*

$$(5.1) \quad \frac{1}{2} \int c_\varepsilon(z_1, z_2) d\pi_\varepsilon(z_1, z_2) + \frac{1}{2}(w_1 - w_0) \leq \int_{A^{-\varepsilon}} w_1\rho_1 dx - \int_{A^\varepsilon} w_0\rho_0 dx,$$

for some  $\pi_\varepsilon \in \Gamma(\nu, \nu^S)$ . Then, it follows from Proposition 3.2 that  $A$  and  $\pi_\varepsilon$  are solutions to the optimization problems in that same proposition, and by Corollary 3.3,  $A$  is also a minimizer of (3.1).

**Remark 5.3** (Knott-Smith optimality criterion). *According to Remark (5.2), to show that a given measurable set  $A$  is an optimizer for (3.1), we would need to construct a coupling  $\pi_\varepsilon$  for which (5.1) holds. Now, let  $A$  be a measurable subset of  $\mathbb{R}^d$ , and suppose that  $\pi_\varepsilon \in \Gamma(\nu, \nu^S)$  is concentrated on the set:*

$$(5.2) \quad \{(z_1, z_2) \in (\mathbb{R}^d \times \{0, 1\})^2 : \mathbb{1}_{A^{-\varepsilon} \times \{0\}}(z_1) - \mathbb{1}_{A^\varepsilon \times \{0\}}(z_2) + \mathbb{1}_{(A^c)^{-\varepsilon} \times \{1\}}(z_1) - \mathbb{1}_{(A^c)^\varepsilon \times \{1\}}(z_2) = c_\varepsilon(z_1, z_2)\}.$$

*Then, it is straightforward to check that  $A$  and  $\pi_\varepsilon$  satisfy (5.1) (with equality). The above condition for  $\pi_\varepsilon$  suggests then how mass must be exchanged between the measures  $\nu$  and  $\nu^S$  in order to get an optimal coupling. This insight is used to build the coupling from Theorem 5.4 below.*

We are now ready to state the main result of this section, which states that the evolution equations (4.2) and (4.3) locally characterize minimizers of the adversarial classification problem.

**Theorem 5.4.** *Under Assumptions 5.1 on  $\rho_0, \rho_1, w_0, w_1$ , there exists  $\varepsilon_0 > 0$  such that for every  $\varepsilon \in [0, \varepsilon_0]$  there exists a coupling  $\pi_\varepsilon \in \Gamma(\nu, \nu^S)$  satisfying:*

$$w_1 \int_{(A^*)^{-\varepsilon}} \rho_1 dx - w_0 \int_{(A^*)^\varepsilon} \rho_0 dx = \frac{1}{2} \int c_\varepsilon(z_1, z_2) d\pi_\varepsilon(z_1, z_2) + \frac{1}{2}(w_1 - w_0),$$

*where  $A^* = A_\varepsilon^* := \bigcup_{i=1}^K (a_i(\varepsilon), b_i(\varepsilon))$  and the functions  $a_i, b_i$  solve the equations (4.2) and (4.3) with initial conditions  $a_i(0), b_i(0)$ ; here, the points  $a_i(0), b_i(0)$  form the decision boundary for the Bayes classifier. In particular, according to Remark 5.2 the set  $A_\varepsilon^*$  induces an optimal robust classifier for  $\varepsilon$ , i.e. it is a solution to the Problem (3.1).*

*Proof.* Let us recall that by convention we have ordered the endpoints as  $a_1(0) < b_1(0) < a_2(0) < b_2(0) < \dots < a_K(0) < b_K(0)$ . We can pick  $\delta > 0$  small enough so that for all  $i > 1$  we have

$$b_{i-1}(0) + \delta < a_i(0) - \delta < a_i(0) + \delta < b_i(0) - \delta.$$

For  $i = 1$ , if  $a_1(0)$  is finite the same inequality applies, interpreting  $b_0(0) := -\infty$ . Similarly,

$$a_i(0) + \delta < b_i(0) - \delta < b_i(0) + \delta < a_{i+1}(0) - \delta$$

for  $i < K$ , and if  $b_K(0) < +\infty$  the same inequality applies, interpreting  $a_{K+1}(0) := +\infty$ . We notice that the solutions of the evolution equations (4.2) and (4.3), which will possess local solutions under Assumption 5.1, are guaranteed to satisfy the necessary conditions (4.1). This fact will be used repeatedly below.

**Step 1:** Let us fix a finite right endpoint  $b_i$  and notice that for small enough  $\varepsilon > 0$  we have

$$b_i(0) - \delta/2 \leq b_i - \varepsilon < b_i + \varepsilon < b_i(0) + \delta/2 < a_{i+1}(0) - \delta,$$

where we recall that  $b_i = b_i(\varepsilon)$  (we have dropped the dependence on  $\varepsilon$  to ease the notation). Now, by Assumption 5.1 we know that  $w_0 \rho'_0(b_i(0)) \neq w_1 \rho'_1(b_i(0))$ , and from the fact that  $w_0 \rho_0 < w_1 \rho_1$  inside  $(a_i(0), b_i(0))$ , we deduce that  $w_0 \rho'_0(b_i(0)) > w_1 \rho'_1(b_i(0))$ . Moreover, the fact that  $\rho_0, \rho_1$  are  $C^1(\mathbb{R})$  allows us to deduce that

$$(5.3) \quad w_0 \rho'_0(t_0) > w_1 \rho'_1(t_1)$$

for every  $t_0, t_1$  in  $[b_i(0) - \delta, b_i(0) + \delta]$  (by making  $\delta$  smaller if needed). In particular, for all  $\varepsilon > 0$  small enough we have

$$(5.4) \quad \frac{d}{ds} (w_0 \rho_0(b_i - \varepsilon - s)) < \frac{d}{ds} (w_1 \rho_1(b_i + \varepsilon - s)), \quad \forall s \in (0, \delta/2).$$

The above condition can be combined with the necessary condition for  $b_i$  in (4.1) and the fundamental theorem of Calculus to obtain

$$(5.5) \quad w_0 \rho_0(b_i - \varepsilon - s) \leq w_1 \rho_1(b_i + \varepsilon - s), \quad \forall s \in (0, \delta/2),$$

for all small enough  $\varepsilon > 0$ .

Let  $r_i^+$  (which depends on  $\varepsilon$ ) be the largest number smaller than  $b_i - \varepsilon$  satisfying:

$$(5.6) \quad \int_{r_i^+}^{b_i - \varepsilon} w_1 \rho_1(x) dx = \int_{r_i^+}^{b_i + \varepsilon} w_0 \rho_0(x) dx.$$

The existence of  $r_i^+$  (at least for small enough  $\varepsilon$ ) follows from (5.4) and condition (4.1), which combined also imply that  $r_i^+$  satisfies  $b_i - \delta/2 \leq r_i^+$ . On the other hand, we can see that  $r_i^+$  also satisfies  $r_i^+ \leq b_i(0)$ . Indeed, if  $b_i - \varepsilon \leq b_i(0)$  this is immediate. If on the other hand,  $b_i - \varepsilon > b_i(0)$  we see that for all  $t \in [b_i(0), b_i - \varepsilon]$

$$\int_t^{b_i - \varepsilon} w_1 \rho_1(x) dx < \int_t^{b_i + \varepsilon} w_0 \rho_0(x) dx,$$

because in the interval  $(b_i(0), b_i + \varepsilon)$  we have  $w_0 \rho_0 > w_1 \rho_1$ . Therefore,  $r_i^+ \leq b_i(0)$  in this case too. In summary,

$$(5.7) \quad r_i^+ \in [b_i - \delta/2, b_i(0)].$$

Now we define the function  $\phi_{b_i} : [r_i^+, b_i - \varepsilon] \rightarrow [r_i^+, b_i + \varepsilon]$  as  $t \mapsto \phi_{b_i}(t)$  where  $\phi_{b_i}(t)$  is the largest number in  $[r_i^+, b_i + \varepsilon]$  which satisfies:

$$\int_t^{b_i - \varepsilon} w_1 \rho_1(x) dx = \int_{\phi_{b_i}(t)}^{b_i + \varepsilon} w_0 \rho_0(x) dx.$$

Due to inequality (5.5),  $\phi_{b_i}$  satisfies:

$$(5.8) \quad |t - \phi_{b_i}(t)| \leq 2\varepsilon, \quad \forall t \in [r_i^+, b_i - \varepsilon].$$

The map  $\phi_{b_i}$  induces a measure  $\gamma_{b_i}$  on  $\mathbb{R} \times \mathbb{R}$  given by

$$\gamma_{b_i} := (Id \times \phi_{b_i})_{\#} (w_1 \rho_1 \llcorner [r_i^+, b_i - \varepsilon]),$$

whose first and second marginals are the measures  $w_1 \rho_1 \llcorner [r_i^+, b_i - \varepsilon]$  and  $w_0 \rho_0 \llcorner [r_i^+, b_i + \varepsilon]$  respectively; in the above  $\#$  denotes the push-forward operation and  $\llcorner$  the restriction of a measure to a given set. We also consider the inverse coupling  $\gamma_{b_i}^{-1}$  defined according to the identity

$$\gamma_{b_i}^{-1}(D \times D') := \gamma_{b_i}(D' \times D),$$

for all  $D, D'$  measurable subsets of  $\mathbb{R}$ .

**Step 2:** We now consider a symmetric construction to the one from Step 1. Using again (5.3) and combining with the necessary condition for  $b_i$  in (4.1) we obtain:

$$(5.9) \quad w_1 \rho_1(b_i - \varepsilon + s) \leq w_0 \rho_0(b_i + \varepsilon + s), \quad \forall s \in (0, \delta/2).$$

We let  $\tilde{r}_i^+$  be the smallest number larger than  $b_i + \varepsilon$  that satisfies:

$$\int_{b_i - \varepsilon}^{\tilde{r}_i^+} w_1 \rho_1(x) dx = \int_{b_i + \varepsilon}^{\tilde{r}_i^+} w_0 \rho_0(x) dx.$$

This quantity can be shown to exist and to satisfy

$$(5.10) \quad \tilde{r}_i^+ \in [b_i(0), b_i + \delta/2]$$

using similar arguments to the ones employed in Step 1.

We let  $\tilde{\phi}_{b_i} : [b_i - \varepsilon, \tilde{r}_i^+] \rightarrow [b_i + \varepsilon, \tilde{r}_i^+]$  be the function defined as  $t \mapsto \tilde{\phi}_{b_i}(t)$  where  $\tilde{\phi}_{b_i}(t)$  is the smallest number in  $[b_i + \varepsilon, \tilde{r}_i^+]$  which satisfies:

$$\int_{b_i - \varepsilon}^t w_1 \rho_1(x) dx = \int_{b_i + \varepsilon}^{\tilde{\phi}_{b_i}(t)} w_0 \rho_0(x) dx.$$

Inequality (5.9) implies

$$(5.11) \quad |t - \tilde{\phi}_{b_i}(t)| \leq 2\varepsilon, \quad \forall t \in [b_i - \varepsilon, \tilde{r}_i^+].$$

The map  $\tilde{\phi}_{b_i}$  induces a measure  $\tilde{\gamma}_{b_i}$  on  $\mathbb{R} \times \mathbb{R}$  given by

$$\tilde{\gamma}_{b_i} := (Id \times \tilde{\phi}_{b_i})_{\#} (w_1 \rho_1 \llcorner [b_i - \varepsilon, \tilde{r}_i^+]),$$

whose first and second marginals are the measures  $w_1 \rho_1 \llcorner [b_i - \varepsilon, \tilde{r}_i^+]$  and  $w_0 \rho_0 \llcorner [b_i + \varepsilon, \tilde{r}_i^+]$  respectively. We also consider the inverse coupling  $\tilde{\gamma}_{b_i}^{-1}$ .

**Step 3:** So far we have constructed measures  $\gamma_{b_i}, \gamma_{b_i}^{-1}, \tilde{\gamma}_{b_i}, \tilde{\gamma}_{b_i}^{-1}$  relative to a finite right endpoint  $b_i$ , but following a completely analogous scheme we can introduce measures  $\gamma_{a_i}, \gamma_{a_i}^{-1}, \tilde{\gamma}_{a_i}, \tilde{\gamma}_{a_i}^{-1}$  satisfying completely equivalent properties to their  $a_i$  counterparts. In particular, for a finite left endpoint  $a_i$  we introduce two quantities  $r_i^-$  and  $\tilde{r}_i^-$  that satisfy

$$r_i^- \in [a_i(0), a_i + \delta/2], \quad \tilde{r}_i^- \in [a_i - \delta/2, a_i(0)],$$

$$\int_{\tilde{r}_i^-}^{a_i + \varepsilon} w_1 \rho_1(x) dx = \int_{\tilde{r}_i^-}^{a_i - \varepsilon} w_0 \rho_0(x) dx, \quad \int_{a_i + \varepsilon}^{r_i^-} w_1 \rho_1(x) dx = \int_{a_i - \varepsilon}^{r_i^-} w_0 \rho_0(x) dx.$$

Two maps  $\phi_{a_i} : [a_i + \varepsilon, r_i^-] \rightarrow [a_i - \varepsilon, r_i^-]$  and  $\tilde{\phi}_{a_i} : [\tilde{r}_i^-, a_i + \varepsilon] \rightarrow [\tilde{r}_i^-, a_i - \varepsilon]$  satisfying

$$|t - \phi_{a_i}(t)| \leq 2\varepsilon, \quad \forall t \in [a_i + \varepsilon, r_i^-], \quad |t - \tilde{\phi}_{a_i}(t)| \leq 2\varepsilon, \quad \forall t \in [\tilde{r}_i^-, a_i + \varepsilon].$$

can be constructed. These maps induce the couplings

$$\gamma_{a_i} = (Id \times \phi_{a_i})_{\#} (w_1 \rho_1 \llcorner [a_i + \varepsilon, r_i^-]) \quad \text{and} \quad \tilde{\gamma}_{a_i} = (Id \times \tilde{\phi}_{a_i})_{\#} (w_1 \rho_1 \llcorner [\tilde{r}_i^-, a_i + \varepsilon]).$$

**Step 4:** In addition to the constructions in Steps 1-3 we introduce  $\tilde{r}_0^+ = -\infty$  and  $\tilde{r}_{K+1}^- = +\infty$ . Also, we set  $\tilde{r}_1^- = r_1^- = -\infty$  in case  $a_1(0) = -\infty$  and  $r_K^+ = \tilde{r}_K^+ = +\infty$  in case  $b_K(0) = +\infty$ . We now define the desired transport plan  $\pi_\varepsilon$ .

Let  $\nu_0^R, \nu_1^R$  be the measures on  $\mathbb{R}$  given by:

$$\nu_0^R := \sum_{i=1}^K (w_1 \rho_1 - w_0 \rho_0) \llcorner [r_i^-, r_i^+]$$

$$\nu_1^R := \left( \sum_{i=1}^K (w_0 \rho_0 - w_1 \rho_1) \llcorner [\tilde{r}_{i-1}^+, \tilde{r}_i^-] \right) + (w_0 \rho_0 - w_1 \rho_1) \llcorner [\tilde{r}_K^+, \tilde{r}_{K+1}^-],$$

we notice that since  $[r_i^-, r_i^+] \subseteq [a_i(0), b_i(0)]$ ,  $\nu_0$  is indeed a positive measure. Similarly, we can see that  $\nu_1^R$  is a positive measure too. From our construction it follows that

$$(5.12) \quad \int_{r_i^+}^{\tilde{r}_i^+} w_0 \rho_0 dx = \int_{r_i^+}^{\tilde{r}_i^+} w_1 \rho_1 dx, \quad \int_{\tilde{r}_i^-}^{r_i^-} w_0 \rho_0 dx = \int_{\tilde{r}_i^-}^{r_i^-} w_1 \rho_1 dx,$$

for all  $i$ , and hence

$$\nu_0^R(\mathbb{R}) = \nu_1^R(\mathbb{R}) + w_1 - w_0.$$

Let  $\pi^R$  be *any* coupling between the measures

$$(\nu_0^R \otimes \delta_0 + \nu_1^R \otimes \delta_1), \quad \text{and} \quad (\nu_1^R \otimes \delta_0 + \nu_0^R \otimes \delta_1);$$

notice that  $\pi^R$  is a measure on  $(\mathbb{R} \times \{0, 1\})^2$ . This is always possible using a product coupling. In the above  $\otimes$  is used to denote the product of two measures.

Let  $\pi^0$  be the measure on  $(\mathbb{R} \times \{0, 1\})^2$  given by:

$$\pi^0(dz_1, dz_2) := \sum_{i=1}^K (\gamma_{b_i}(dx_1, dx_2) \otimes \delta_{\{0\} \times \{0\}}(dy_1, dy_2) + \gamma_{b_i}^{-1}(dx_1, dx_2) \otimes \delta_{\{1\} \times \{1\}}(dy_1, dy_2))$$

$$+ \tilde{\gamma}_{b_i}(dx_1, dx_2) \otimes \delta_{\{0\} \times \{0\}}(dy_1, dy_2) + \tilde{\gamma}_{b_i}^{-1}(dx_1, dx_2) \otimes \delta_{\{1\} \times \{1\}}(dy_1, dy_2)$$

$$+ \gamma_{a_i}(dx_1, dx_2) \otimes \delta_{\{0\} \times \{0\}}(dy_1, dy_2) + \gamma_{a_i}^{-1}(dx_1, dx_2) \otimes \delta_{\{1\} \times \{1\}}(dy_1, dy_2)$$

$$+ \tilde{\gamma}_{a_i}(dx_1, dx_2) \otimes \delta_{\{0\} \times \{0\}}(dy_1, dy_2) + \tilde{\gamma}_{a_i}^{-1}(dx_1, dx_2) \otimes \delta_{\{1\} \times \{1\}}(dy_1, dy_2)).$$

The first and fourth terms in this expression with eight terms are the mass exchanges illustrated in Figure 2. The other terms have similar interpretations. Finally, we let  $\pi^F$  be the measure on

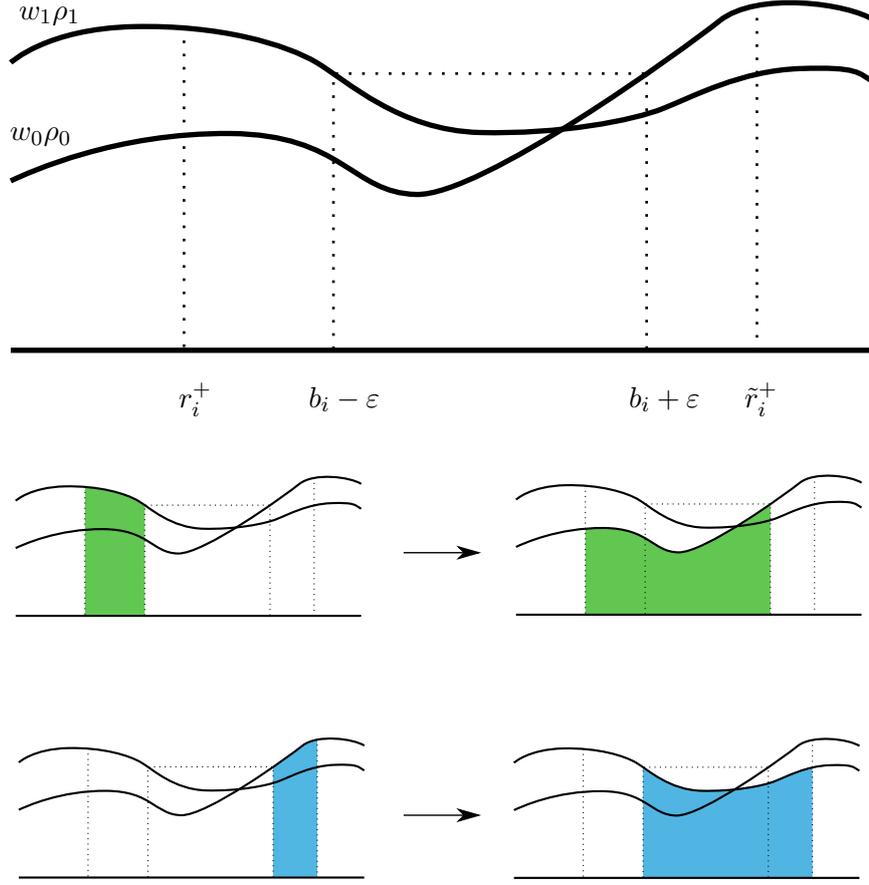


FIGURE 2. Illustration of mass exchange defined by  $\gamma_{b_i}$  (middle) and by  $\tilde{\gamma}_{b_i}^{-1}$  (bottom).

$(\mathbb{R} \times \{0, 1\})^2$  given by  $\pi^F := (Id \times Id)_\#(\nu - (\pi^0 + \pi^R)_1)$ , where  $(\pi^0 + \pi^R)_1$  is the first marginal of  $\pi^0 + \pi^R$ .

With all the above definitions in hand, we can now introduce:

$$\pi_\varepsilon := \pi^0 + \pi^R + \pi^F.$$

Here,  $\pi^0$  satisfies the property that for all the points in its support  $c_\varepsilon = 0$ .  $\pi^F$  corresponds to the mass that is fixed and thus does not contribute to the cost of  $\pi_\varepsilon$ . Finally,  $\pi^R$  corresponds to the remaining mass. Our construction then guarantees that

$$\begin{aligned} \int c_\varepsilon(z_1, z_2) d\pi_\varepsilon(z_1, z_2) &= \int c_\varepsilon(z_1, z_2) d\pi^R(z_1, z_2) \leq \pi^R((\mathbb{R} \times \{0, 1\})^2) \\ &= \nu_0^R(\mathbb{R}) + \nu_1^R(\mathbb{R}) = 2\nu_0^R(\mathbb{R}) + w_0 - w_1 = 2 \sum_{i=1}^K \int_{r_i^-}^{r_i^+} (w_1 \rho_1 - w_0 \rho_0) dx + w_0 - w_1. \end{aligned}$$

In turn,

$$\begin{aligned} \sum_{i=1}^K \int_{r_i^-}^{r_i^+} (w_1 \rho_1 - w_0 \rho_0) dx &= \sum_{i=1}^K \left( \int_{a_i+\varepsilon}^{b_i-\varepsilon} w_1 \rho_1 dx - \int_{a_i-\varepsilon}^{b_i+\varepsilon} w_0 \rho_0 dx \right) \\ &= \int_{(A^*)^{-\varepsilon}} w_1 \rho_1 dx - \int_{(A^*)^\varepsilon} w_0 \rho_0 dx, \end{aligned}$$

thanks to equation (5.6) and the analogues for the  $a_i$ . Thus,

$$\frac{1}{2} \int c_\varepsilon(z_1, z_2) d\pi_\varepsilon(z_1, z_2) + \frac{1}{2}(w_1 - w_0) \leq \int_{(A^*)^{-\varepsilon}} w_1 \rho_1 dx - \int_{(A^*)^\varepsilon} w_0 \rho_0 dx,$$

which thanks to Remark 5.2 implies that  $A^*$  solves the optimization problem and that the above inequality is actually an equality.  $\square$

**Remark 5.5.** *The construction of transportation plans in the previous proof is possible due to the necessary conditions (4.1) that are maintained by the evolution equations (4.3) and (4.2). Indeed, the necessary conditions are crucially used to prove the equalities in (5.12), which factored prominently in the construction of the certifying transportation plan  $\pi_\varepsilon$ .*

**Remark 5.6.** *The construction in the previous proof is local, in the sense that we can only show that solutions to our evolution equations are global minimizers for  $\varepsilon$  sufficiently small. However, the proof of the previous proposition indicates some situations where one can detect that these solutions cease to be global minimizers. For example, if at some point  $r_i^+ = \tilde{r}_{i+1}^-$  then one expects that the construction may not be continued for larger  $\varepsilon$ . This should correspond to a change in topology of the global optimizer. Understanding the type of degeneracies that may arise when solving the geometric evolution equations, as well as their implications to the adversarial risk minimization problem are topics of current investigation.*

## 6. NECESSARY CONDITIONS AND GEOMETRIC EVOLUTION EQUATIONS IN HIGHER DIMENSION

In one dimension, the necessary condition allowed us to derive an ordinary differential equation that described the motion of decision boundaries as we increased the adversarial power  $\varepsilon$ . This evolution equation was driven, for small  $\varepsilon$ , by the gradient of  $\rho$ . In higher dimension the optimality conditions and their associated geometric evolution equations are necessarily more complex. In particular, the presence of curvature in higher dimensions introduces a greater degree of complexity. In order to gain some intuition about the problem, we begin with an explicit, radial example.

**Example 6.1.** *Let us consider the case where  $\rho$  is a uniform distribution on a ball of radius 1 in  $\mathbb{R}^d$ , and  $w_0 \rho_0(x) = \frac{|x|}{\omega_d}$ , with  $\omega_d$  the  $\mathcal{L}^d$  measure of the unit ball. Here the Bayes classifier is given by  $u_B(x) = \mathbb{1}_{|x| \leq 1/2}$ . We then consider a classifier, parameterized in  $\varepsilon$ , which (by way of ansatz) is given by  $\mathbb{1}_{|x| \leq r(\varepsilon)}$ , which minimizes the adversarial cost. If one takes variations in  $r$ , it is straightforward to deduce the necessary condition*

$$\begin{aligned} (r(\varepsilon) + \varepsilon)(r(\varepsilon) + \varepsilon)^{d-1} &= \omega_d w_0 \rho_0(r(\varepsilon) + \varepsilon)(r(\varepsilon) + \varepsilon)^{d-1} \\ &= \omega_d w_1 \rho_1(r(\varepsilon) - \varepsilon)(r(\varepsilon) - \varepsilon)^{d-1} \\ &= (1 - (r(\varepsilon) - \varepsilon))(r(\varepsilon) - \varepsilon)^{d-1}, \end{aligned}$$

where here we are abusing notation slightly and writing  $\rho_1(t)$  and  $\rho_0(t)$  to represent  $\rho_1(x)$  and  $\rho_0(x)$  for all  $x$  such that  $|x| = t$ . Taking a derivative in  $\varepsilon$  we obtain

$$d(r(\varepsilon) + \varepsilon)^{d-1} \left( \frac{d}{d\varepsilon} r + 1 \right) = \left( (d-1)(r(\varepsilon) - \varepsilon)^{d-2} - d(r(\varepsilon) - \varepsilon)^{d-1} \right) \left( \frac{d}{d\varepsilon} r - 1 \right),$$

which may be written

$$\frac{dr}{d\varepsilon} = - \frac{d(r(\varepsilon) + \varepsilon)^{d-1} + ((d-1)(r(\varepsilon) - \varepsilon)^{d-2} - d(r(\varepsilon) - \varepsilon)^{d-1})}{d(r(\varepsilon) + \varepsilon)^{d-1} - ((d-1)(r(\varepsilon) - \varepsilon)^{d-2} - d(r(\varepsilon) - \varepsilon)^{d-1})}.$$

At  $\varepsilon = 0$  this becomes

$$\frac{dr}{d\varepsilon}(\varepsilon = 0) = - \frac{(d-1)r^{d-2}}{2dr^{d-1} - (d-1)r^{d-2}} = - \frac{(d-1)r^{-1}}{2d - (d-1)r^{-1}}$$

Recalling that  $r^{-1} = \kappa$  is the mean curvature of a sphere of radius  $r$  in  $\mathbb{R}^d$ , we immediately see the effect of curvature, namely that this evolution corresponds, at  $\varepsilon = 0$  to a rescaled, mean curvature flow. We notice that here,  $\nabla \rho \equiv 0$ , which in the one dimensional case dominated the evolution for small  $\varepsilon$  regimes. This example was specifically chosen in order to highlight the effect of curvature, but we will subsequently see that both curvature and  $\nabla \rho$  play a role in the surface evolution.

With the previous example in mind, we derive the necessary condition, assuming that the decision boundary is sufficiently smooth.

**Proposition 6.2.** *Suppose that  $A_\varepsilon$  is a critical point (with respect to normal variations [24], further description in the proof) of the problem  $R_\varepsilon$  and that the signed distance function  $\tilde{d}_{A_\varepsilon}$  is  $C^3$  on the set  $|\tilde{d}_{A_\varepsilon}| < 2\varepsilon$ . For  $x \in \partial A_\varepsilon$ , let  $\nu$  denote the outward unit normal and  $\kappa_i$  denote the principal curvatures. Then the following necessary condition holds for almost every  $x \in \partial A_\varepsilon$ :*

$$(6.1) \quad w_1 \rho_1(x - \varepsilon \nu(x)) \prod_{i=1}^{d-1} |1 - \kappa_i \varepsilon| - w_0 \rho_0(x + \varepsilon \nu(x)) \prod_{i=1}^{d-1} |1 + \kappa_i \varepsilon| = 0.$$

*Proof.* We again recall

$$\begin{aligned} R_\varepsilon(\mathbb{1}_A) &= \int_{\tilde{d}_A(x) < -\varepsilon} w_0(x) \rho_0(x) dx + \int_{\tilde{d}_A(x) > \varepsilon} w_1 \rho_1(x) dx \\ &\quad + \int_{|\tilde{d}_A(x)| < \varepsilon} \rho(x) dx. \end{aligned}$$

We consider the class of normal variations [24] of the set  $A = A_\varepsilon$ : that is, we consider a one parameter family of sets  $A^t$  of the form  $A^t = \phi(t, A)$  for some diffeomorphism  $\phi(t, x)$  which satisfies  $\phi(0, A) = A$  and  $\frac{d\phi}{dt}(t=0) = F(x)$ , where  $F$  satisfies  $F(x) = \nu(x)\psi(x)$  for  $x \in \partial A$  and for some scalar valued function  $\psi$ . Taking the derivative of  $R_\varepsilon(\mathbb{1}_{A^t})$  and evaluating it at  $t = 0$ , we obtain that

$$0 = \int_{\tilde{d}_A(y) = \varepsilon} w_0 \rho_0(y) \psi(P_{\partial A}(y)) d\mathcal{H}^{d-1}(y) - \int_{\tilde{d}_A(y) = -\varepsilon} w_1 \rho_1(y) \psi(P_{\partial A}(y)) d\mathcal{H}^{d-1}(y),$$

where here  $P_{\partial A}(x)$  is the projection of  $x$  onto the boundary of  $A$ .

Noting that  $y = x \pm \varepsilon \nu(x)$  in the previous two integrals, we then use a change of variables as in Corollary A.2 to convert to

$$0 = \int_{\tilde{d}_A(x)=0} \left( w_0 \rho_0(x + \varepsilon \nu(x)) \prod_{i=1}^{d-1} |1 + \kappa_i(x) \varepsilon| - w_1 \rho_1(x - \varepsilon \nu(x)) \prod_{i=1}^{d-1} |1 - \kappa_i(x) \varepsilon| \right) \psi(x) d\mathcal{H}^{d-1}(x).$$

Since this holds for all smooth  $\psi$ , we then have that, for  $\mathcal{H}^{d-1}$  almost every  $x \in \partial A$

$$0 = \left( w_0 \rho_0(x + \varepsilon \nu(x)) \prod_{i=1}^{d-1} |1 + \kappa_i(x) \varepsilon| - w_1 \rho_1(x - \varepsilon \nu(x)) \prod_{i=1}^{d-1} |1 - \kappa_i(x) \varepsilon| \right)$$

□

We remark that assuming that the conditional densities  $\rho_0, \rho_1$  are smooth is not sufficient to guarantee that the set of  $x$ s for which  $w_0 \rho_0 - w_1 \rho_1 = 0$  is smooth, as evidenced by the following basic example:

**Example 6.3.** *Suppose in  $\mathbb{R}^2$  that one places normals associated with  $y = +1$  at  $(1, 1)$  and  $(-1, -1)$ , and then places normals associated with  $y = -1$  at  $(1, -1)$  and  $(-1, 1)$ , with  $w_0 = w_1 = 1/2$ . In this case the set where  $w_0 \rho_0 = w_1 \rho_1$  is given by the set  $\{x = 0\} \cup \{y = 0\}$ , which is not smooth at  $(0, 0)$ .*

**6.1. Geometric flow.** In this section we seek to formally derive a geometric flow which characterizes the evolution of the boundary of the  $A_\varepsilon$ . As in the one-dimensional case, we can Taylor expand for  $\varepsilon$  small to derive an approximating geometric flow which is more transparent and easier to interpret.

To begin, let us suppose that  $\phi(\varepsilon, x)$  be a diffeomorphism so that  $\phi(\varepsilon, A) = A_\varepsilon$ . We shall utilize the necessary condition (6.1) to characterize this diffeomorphism for points  $x \in \partial A_0$ .

We now use a chain rule on the necessary condition as follows (suppressing the dependence on  $x, \varepsilon$ , and always assuming that  $x \in \partial A_0$ ):

$$\begin{aligned} 0 &= \frac{d}{d\varepsilon} \left( w_0 \rho_0(\phi + \varepsilon \nu(\phi)) \prod_{i=1}^{d-1} (1 + \varepsilon \kappa_i(\phi)) - w_1 \rho_1(\phi - \varepsilon \nu(\phi)) \prod_{i=1}^{d-1} (1 - \varepsilon \kappa_i(\phi)) \right) \\ &= \prod_{i=1}^{d-1} (1 + \varepsilon \kappa_i(\phi)) \left( \nabla w_0 \rho_0(\phi + \varepsilon \nu(\phi)) \left( \frac{d}{d\varepsilon} \phi + \nu(\phi) + \varepsilon \frac{d}{d\varepsilon} (\nu(\phi)) \right) + w_0 \rho_0(\phi + \varepsilon \nu(\phi)) \sum_i \frac{\kappa_i(\phi) + \varepsilon \frac{d}{d\varepsilon} \kappa(\phi)}{1 + \varepsilon \kappa_i(\phi)} \right) \\ &\quad - \prod_{i=1}^{d-1} (1 - \varepsilon \kappa_i(\phi)) \left( \nabla w_1 \rho_1(\phi - \varepsilon \nu(\phi)) \left( \frac{d}{d\varepsilon} \phi - \nu(\phi) - \varepsilon \frac{d}{d\varepsilon} (\nu(\phi)) \right) + w_1 \rho_1(\phi - \varepsilon \nu(\phi)) \sum_i \frac{-\kappa_i(\phi) - \varepsilon \frac{d}{d\varepsilon} \kappa(\phi)}{1 - \varepsilon \kappa_i(\phi)} \right) \end{aligned}$$

One major challenge here is that  $\frac{d}{d\varepsilon} \nu(\phi)$  and  $\frac{d}{d\varepsilon} \kappa$  will involve mixed derivatives, i.e. derivatives in both  $\varepsilon$  and  $x$ . Indeed, we recall that, in terms of  $\phi$  and for  $x \in \partial A_0$ , one may express the geometric quantity  $\nu$  (the outward surface normal) as

$$\nu(\phi(\varepsilon, x)) = \frac{\frac{\partial}{\partial x} \phi(\varepsilon, x) \cdot \nu_0(x)}{\left\| \frac{\partial}{\partial x} \phi(\varepsilon, x) \cdot \nu_0(x) \right\|}$$

Similarly, the curvatures  $\kappa_i$  may be expressed as appropriate spatial derivatives of  $\nu$ . Thus for  $\varepsilon > 0$  this evolution equation is a non-local, mixed-type partial differential equation, which appears difficult to solve.

However, each of the terms involving mixed derivatives is pre-multiplied by  $t$ , and hence may plausibly be ignored for  $\varepsilon$  sufficiently small. To this end, we rearrange the previous equation

$$\begin{aligned} (6.2) \quad & \left( \prod_{i=1}^{d-1} (1 + \varepsilon \kappa_i(\phi)) \nabla w_1 \rho_1(\phi + \varepsilon \nu(\phi)) - \prod_{i=1}^{d-1} (1 - \varepsilon \kappa_i(\phi)) \nabla w_0 \rho_0(\phi - \varepsilon \nu(\phi)) \right) \frac{d}{d\varepsilon} \phi \\ &= - \prod_{i=1}^{d-1} (1 + \varepsilon \kappa_i(\phi)) \left( \nabla w_1 \rho_1(\phi + \varepsilon \nu(\phi)) (\nu(\phi) + \varepsilon \frac{d}{d\varepsilon} \nu(\phi)) + w_1 \rho_1(\phi + \varepsilon \nu(\phi)) \sum_i \frac{\kappa_i(\phi) + \varepsilon \frac{d}{d\varepsilon} \kappa(\phi)}{1 + \varepsilon \kappa_i(\phi)} \right) \\ &\quad - \prod_{i=1}^{d-1} (1 - \varepsilon \kappa_i(\phi)) \left( \nabla w_0 \rho_0(\phi - \varepsilon \nu(\phi)) (\nu(\phi) + \varepsilon \frac{d}{d\varepsilon} \nu(\phi)) + w_0 \rho_0(\phi - \varepsilon \nu(\phi)) \sum_i \frac{\kappa_i(\phi) + \varepsilon \frac{d}{d\varepsilon} \kappa(\phi)}{1 - \varepsilon \kappa_i(\phi)} \right) \end{aligned}$$

Evaluating at  $\varepsilon = 0$ , we find that

$$(w_1 \nabla \rho_1 - w_0 \nabla \rho_0) \frac{d\phi}{d\varepsilon} = -(\nabla \rho \cdot \nu + \rho \sum_i \kappa_i).$$

If we express  $\frac{d\phi}{d\varepsilon} = v\nu$ , namely we consider the normal speed  $v$ , then we may write

$$(6.3) \quad v(x, \varepsilon = 0) = -\frac{\nabla \rho \cdot \nu + \rho \sum_i \kappa_i}{(w_1 \nabla \rho_1 - w_0 \nabla \rho_0) \cdot \nu}$$

Here we observe two terms: one which induces motion “downhill” in  $\rho$  and a second which is a positively weighted mean curvature term. As we have used  $\nu$  as an outwardly pointing normal vector, the  $-\sum \kappa$  will correspond to the standard mean curvature flow. This indicates that heuristically, near  $\varepsilon = 0$ , the optimal adversarial classifier seeks to i) go downhill in  $\rho$ , and ii) decrease the perimeter of the decision boundary (since mean curvature flow is a type of gradient flow of perimeter). While the reweighting in the denominator is not homogeneous, and

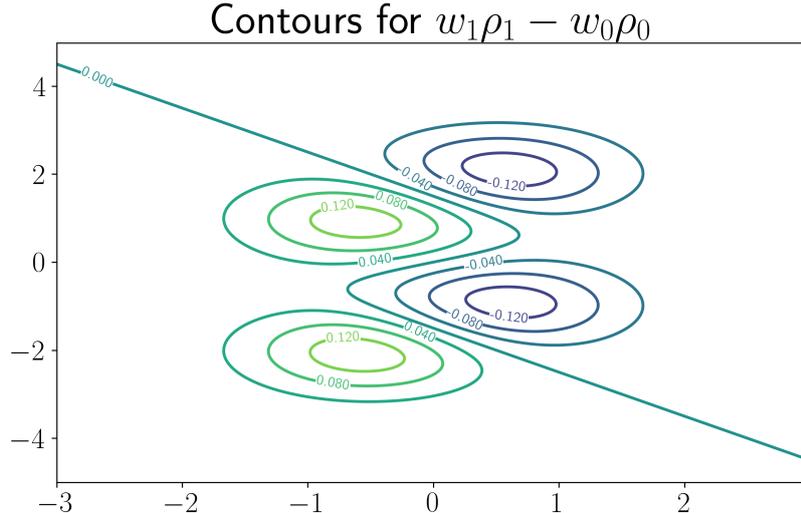


FIGURE 3. The contours of the function  $w_1\rho_1 - w_0\rho_0$  for the example in section 6.2. The contour corresponding to  $w_1\rho_1 - w_0\rho_0 = 0$  is the s-shaped curve, and represents the decision boundary for the Bayes classifier.

indeed makes this heuristic description imprecise, we believe this heuristic picture is helpful for understanding the local effects induced by adversarial robustness.

**6.2. Illustration in two dimensions.** Here we show a basic numerical example of the geometric evolution (6.3) in two dimensions. This example is intended to be an illustration, rather than a detailed computational study. Such a study would require careful numerical analysis, which lies outside of the scope of this work.

We consider two different classes  $\rho_1 \sim N((-0.5, -2), \Sigma) + N((-0.5, 0.5), \Sigma)$  and  $\rho_2 \sim N((0.5, -0.5), \Sigma) + N((0.5, 2), \Sigma)$ , where  $\Sigma = .2I$ , and  $w_0 = w_1 = .5$ . The Bayes classifier boundary, along with contours of the misclassification error  $w_1\rho_1 - w_0\rho_0$  are shown in Figure 3.

We then use a modified version of the scheme from [28] to track the evolution of the decision boundary under the evolution equation (6.3) for different values of  $\varepsilon$ . These curves are displayed in Figure 4, next to the curves evolved via standard mean curvature flow as a point of reference.

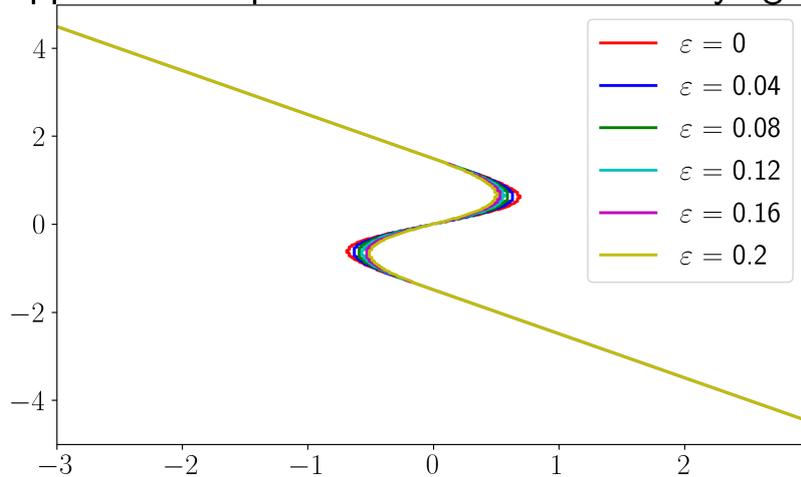
## 7. CONCLUSION

This work provides a first analysis of the evolution equations associated with an ensemble of adversarial classification problems. In particular, we have shown that for the model considered here, the evolution equations in one dimension are completely able to characterize the global minimizer for small enough  $\varepsilon$  (the power level of the adversary) without needing to conduct any optimization. In higher dimension the same evolution equations are linked with mean curvature flow and allude to implicit curvature regularization.

This work suggests many promising future directions, both in terms of analysis and implementation. We list a few here, some of which are the topic of current investigation.

- i) The question of efficient numerical methods is important in implementing these evolution equations. In one dimension, detecting global optimality via primal-dual methods, as well as topological changes in decision boundaries, is a promising method. In higher dimensions, efficient solvers for both the approximate and exact equations are important considerations.
- ii) In higher dimension, a natural question is whether the evolution equation is i) well-posed and ii) determines global minimizers (locally in  $\varepsilon$ ).

### Approximate optimal robust classifiers varying in $\varepsilon$



### Evolution of Bayes classifier via mean curvature flow

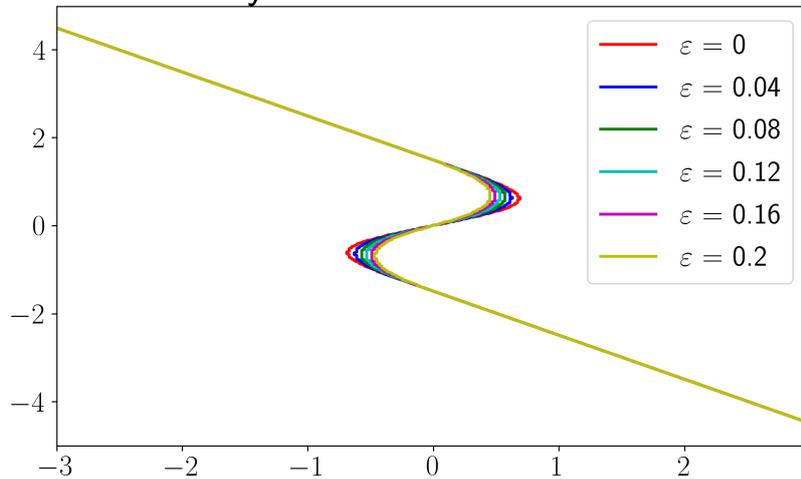


FIGURE 4. The first set of curves represent the evolution of the decision boundary according to the geometric evolution (6.3), which includes a weighted curvature flow and a drift term. The second set of curves is the geometric evolution following standard mean curvature flow. In this case the curves are largely the same, with only a very small damping of the curvature flow in the first case (which makes sense since  $\nabla\rho$  is of modest size in this example).

- iii) The smoothness of minimizers, and whether curvature is implicitly bounded, is a natural question. This is not obvious, as the objective functional of the adversarial problem does not impose a priori regularity. Similarly, the evolution of singularities (and whether they may disappear or appear) is completely unclear.
- iv) Various notions of distance have been used in studying adversarial examples. Notable examples include the  $\ell_\infty$  distance. The effect of such a distance on the evolution equations that we describe in this work is an interesting question to study.
- v) The problem of a data perturbing adversary for multiple labels, and the resulting evolution equations, is also a compelling, open problem.

vi) Finally, here we have considered one specific example of adversarial classification model, but many others are possible. Likewise, we have restricted our attention to the classification problem with 0-1 loss, while one may also study other settings like regression under different loss functions. Exploring other settings and studying their connection to other geometric flows is a promising direction of research that we hope to explore. Our hope is to provide deeper insights into the properties of different robust learning methodologies.

## APPENDIX A. PROPERTIES OF THE SIGNED DISTANCE FUNCTION

We recall the definition of the signed distance function

$$\tilde{d}_E(x) = \begin{cases} d(x, E) & \text{if } x \notin E \\ -d(x, E^c) & \text{for } x \in E \end{cases}$$

The following properties are classical and may be found in, e.g. [1]:

**Proposition A.1.** *Let  $E$  be an open set with  $C^2$  boundary. Then on some neighborhood  $U$  of  $\partial E$  we have the following:*

- $\tilde{d} \in C^2(U)$ .
- Each  $y$  in  $U$  has a unique closest point  $P(y)$  in  $\partial E$ , and  $P$  is a continuous function in  $y$ .
- We have, for  $y \notin \partial E$ , that  $\nabla \tilde{d} = \frac{P(y)-y}{\tilde{d}(y)}$ . For  $y \in \partial E$  the outward unit normal is given by  $\nu(y) = \nabla \tilde{d}(y)$ .
- For  $y \in \partial E$ , the matrix  $D^2 \tilde{d}$  has 1 eigenvalue that is equal to zero (with eigenvector in the normal direction  $\nu$ ), and  $d - 1$  eigenvalues with eigenvectors spanning the tangent directions. These eigenvalues are called the principal curvatures of the surface, and are denoted  $\kappa_i$ .

The principal curvatures of a surface may be viewed as inverses of the principal radii. The principal radii grow (or shrink depending on their sign) linearly in their distance from  $\partial E$ . By using these facts and applying a classical change of variables to the transformation  $T(x) = x + \varepsilon \nu(x)$ , we obtain the following formula:

**Corollary A.2.** *If  $\partial E$  is a  $C^2$  surface then for  $\varepsilon$  sufficiently small*

$$\int_{\tilde{d}_A(y)=\varepsilon} g(y) d\mathcal{H}^{d-1}(y) = \int_{\tilde{d}_A(x)=0} g(x + \varepsilon \nu(x)) \prod_{i=1}^{d-1} |1 + \varepsilon \kappa_i(x)| d\mathcal{H}^{d-1}(x).$$

## REFERENCES

- [1] Luigi Ambrosio and Carlo Mantegazza. Curvature and distance function from a manifold. *The Journal of Geometric Analysis*, 8(5):723–748, 1998.
- [2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.
- [3] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems*, pages 7498–7510, 2019.
- [4] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.
- [5] Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, pages 831–840, 2019.
- [6] Luca Calatroni, Yves van Gennip, Carola-Bibiane Schönlieb, Hannah M. Rowland, and Arjuna Flenner. Graph clustering, variational image segmentation methods and hough transform scale detection for object measurement in images. *Journal of Mathematical Imaging and Vision*, 57(2):269–291, 2017.
- [7] J. Calder. The game theoretic  $p$ -Laplacian and semi-supervised learning with few labels. *Nonlinearity*, 32(1):301, 2018.
- [8] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017.

- [9] Antonin Chambolle, Massimiliano Morini, and Marcello Ponsiglione. Nonlocal curvature flows. *Archive for Rational Mechanics and Analysis*, 218(3):1263–1329, 2015.
- [10] Mihai Cucuringu, Andrea Pizzoferrato, and Yves van Gennip. An mbo scheme for clustering and semi-supervised clustering of signed networks, 2019.
- [11] Klaus Ecker. *Regularity theory for mean curvature flow*, volume 57. Springer Science & Business Media, 2012.
- [12] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.
- [13] Nicolás García Trillos and Ryan Murray. A maximum principle argument for the uniform convergence of graph Laplacian regressors. *Preprint*, 2019.
- [14] Nicolás García Trillos, Ryan Murray, and Matthew Thorpe. From graph cuts to isoperimetric inequalities: Convergence rates of cheeger cuts on data clouds. *arXiv preprint arXiv:2004.09304*, 2020.
- [15] Nicolás García Trillos and Dejan Slepčev. Continuum limit of total variation on point clouds. *Archive for Rational Mechanics and Analysis*, pages 1–49, 2015.
- [16] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.
- [17] Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. Adversarial and clean data are not twins. *arXiv preprint arXiv:1704.04960*, 2017.
- [18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [19] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [20] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.
- [21] Huiyi Hu, Thomas Laurent, Mason A. Porter, and Andrea L. Bertozzi. A method based on total variation for network modularity optimization using the mbo scheme. *SIAM Journal on Applied Mathematics*, 73(6):2224–2246, 2013.
- [22] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019.
- [23] Matt Jacobs, Ekaterina Merkurjev, and Selim Eshedoglu. Auction dynamics: A volume constrained mbo scheme. *Journal of Computational Physics*, 354:288 – 310, 2018.
- [24] Francesco Maggi. *Sets of finite perimeter and geometric variational problems*, volume 135 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2012. An introduction to geometric measure theory.
- [25] Saeed Mahloujifar, Dimitrios I Diochnos, and Mohammad Mahmoody. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4536–4543, 2019.
- [26] Ekaterina Merkurjev, Tijana Kostić, and Andrea L. Bertozzi. An mbo scheme on graphs for classification and image processing. *SIAM Journal on Imaging Sciences*, 6(4):1903–1930, 2013.
- [27] Ekatherina Merkurjev, A. Bertozzi, and F. Chung. A semi-supervised heat kernel pagerank mbo algorithm for data classification. *Communications in Mathematical Sciences*, 16:1241–1265, 2018.
- [28] Barry Merriman, James Kenyard Bence, and Stanley Osher. *Diffusion generated motion by mean curvature*. Department of Mathematics, University of California, Los Angeles, 1992.
- [29] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017.
- [30] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9078–9086, 2019.
- [31] Muni Sreenivas Pydi and Varun Jog. Adversarial risk via optimal transport and optimal couplings. *arXiv preprint arXiv:1912.02794*, 2019.
- [32] Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018.
- [33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [34] Yves van Gennip, Nestor Guillen, Braxton Osting, and Andrea L. Bertozzi. Mean curvature, threshold dynamics, and phase field theory on finite graphs. *Milan Journal of Mathematics*, 82(1):3–65, 2014.
- [35] C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics. American Mathematical Society, 2003.

- [36] Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. Analyzing the robustness of nearest neighbors to adversarial examples. volume 80 of *Proceedings of Machine Learning Research*, pages 5133–5142, Stockholmström, Stockholm Sweden, 10–15 Jul 2018. PMLR.

DEPARTMENT OF STATISTICS, UNIVERSITY OF WISCONSIN, MADISON, WISCONSIN, USA  
*Email address:* `garciatrillo@wisc.edu`

DEPARTMENT OF MATHEMATICS, NORTH CAROLINA STATE UNIVERSITY, RALEIGH, NC, USA  
*Email address:* `rwmurray@ncsu.edu`