Journal of Machine Learning Research 23 (2022) 1-38 Submitted 3/21; Published 7/22

Adversarial Classification: Necessary Conditions and Geometric Flows

Nicolás García Trillos GARCIATRILLO@WISC.EDU
Department of Statistics
University of Wisconsin
Madison, Wisconsin, USA

Ryan Murray RWMURRAY@NCSU.EDU
Department of Mathematics
North Carolina State University
Raleigh, NC, USA

Editor: Amos Storkey

Abstract

We study a version of adversarial classification where an adversary is empowered to corrupt data inputs up to some distance ε, 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 ε 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, and under mild assumptions on the data distribution, 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 for small values of ε. In higher dimensions we provide a similar result, albeit conditional to the existence of regular solutions of the initial value problem. In the process of proving our main results we obtain a result of independent interest connecting the original adversarial problem with an optimal transport problem under no assumptions on whether classes are balanced or not. Numerical examples illustrating these ideas are also presented.

Keywords: adversarial learning, classification, optimal transportation, geometric flow, differential equations, perimeter regularization

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 particular, (Bhagoji et al., 2019; Pydi and Jog, 2019) 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.

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

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

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 ε 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 ε (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 ε 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 (Bhagoji et al., 2019; Pydi and Jog, 2019) (which we extend here to include unbalanced classes, and which holds under arbitrary metrics constraining adversarial perturbations and in arbitrary dimension), we derive the following theorem (stated informally):

Theorem 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. For simplicity, we focus our attention on the setting where the adversarial constraints are given in terms of the standard Euclidean norm, which we denote by $|\cdot|$. 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 ε varies (6.4). The well-posedness of this geometric evolution equation is not immediately clear, but under the assumption that regular solutions do exist we can also prove that the solution of the evolution equation characterizes global minimizers on some interval $[0,\varepsilon_0]$, see Theorem 12. Using a Taylor expansion around $\varepsilon = 0$, we can also identify approximate geometric evolution dynamics which are more interpretable. In particular, we derive the evolution equation (6.5), which may be written as follows:

$$v(x) = -\frac{\nabla\rho\cdot\nu + \rho\sum_i \kappa_i}{(\nabla w_1\rho_1 - \nabla w_0\rho_0)\cdot\nu}, \tag{1.1}$$

where here v represents the normal velocity (with respect to ε) of a point on the decision boundary of the Bayes classifier, ν is the normal vector to the boundary, κ_i denote the principal curvatures (see the Appendix for a definition) of the boundary, and $\rho = w_0\rho_0 + w_1\rho_1 = \mathbb{P}(X \in dx)$. Conceptually, the vector field $v\nu$ describes the infinitesimal change of the Bayes classifier (i.e. the minimizer of the problem when $\varepsilon = 0$) as the adversary increases its power. Evolution equation (1.1) takes the form of a weighted *mean curvature flow* plus

a biasing term (the biasing term is driven by the gradient of the distribution ρ). Mean curvature flow is an important geometric flow with many convenient 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). Equation (1.1) thus suggests that as ε increases, the corresponding optimal decision boundaries become shorter and smoother, supporting previous work on the topic. 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 given in Merriman et al. 1992).

To be more concrete about the connection between equation (1.1) and perimeter minimization problems, let us consider the family of variational problems:

$$\min_{E \subseteq \mathbb{R}^d} \{R(\mathbb{1}_E) + \varepsilon \text{Per}_\rho(E)\} \tag{1.2}$$

indexed by $\varepsilon \geq 0$, where R denotes the standard average misclassification error and Per_ρ represents the weighted (by ρ) perimeter of the set E, which, for sets E with smooth boundary ∂E, can be written as:

$$\text{Per}_\rho(E) := \int_{\partial E} \rho(x) d\mathcal{H}^{d-1}(x);$$

in the above, $\mathcal{H}^{d-1}$ is the $d-1$ dimensional Hausdorff measure. Problem (1.2) can be interpreted as a regularized risk minimization problem over binary classifiers, where Per_ρ plays the role of an *explicit regularizer*, in this case penalizing binary classifiers when they have large decision boundaries. Problem (1.2) is relevant in the context of adversarial learning because, as we illustrate formally in Section 6.1.1, Equation (1.1) also describes the infinitesimal change of solutions to the family of problems (1.2) (indexed by ε) when starting at $\varepsilon = 0$ (i.e. when starting with the Bayes classifier, which is the minimizer of the risk R.) From this observation we can deduce that the instantaneous regularization effect that the adversary has on the Bayes classifier is the same as the infinitesimal regularization effect enforced by explicit perimeter regularization. This observation provides a novel geometric interpretation for the role of adversaries in binary classification: they are approximately equivalent to an explicit perimeter penalization. This line of research has been further explored by the authors in their work with Bungert, namely (Bungert et al., 2021), where they prove an equivalence between adversarial learning for binary classification and regularized risk minimization for all $\varepsilon > 0$ (and not just infinitesimally around $\varepsilon = 0$) at the expense of having to modify the notion of perimeter used to measure the size of the boundary of a set.

In summary, in this paper we take a novel approach and view an adversarial problem as an ensemble of problems indexed by a parameter controlling the ability of an adversary to perturb the data. The main motivation for doing this is to provide new theoretical insights into the role played by adversaries in the training of binary classifiers. In concrete terms, we discuss properties of the evolution equations that track solutions to an ensemble of adversarial problems, starting from an un-robust optimal classifier. These evolution equations take the form of geometric equations. 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, which is of interest on its own right and that extends earlier work in (Pydi and Jog, 2019) where the case of balanced labels ($w_0 = w_1$) was considered. In this paper, the connection to optimal transport is used to certify global optimality of the decision boundaries generated by the geometric flows.

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 (Szegedy et al., 2013; Goodfellow et al., 2014b). 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: one influential example from this literature is (Madry et al., 2017), which established important benchmarks for both adversarial attack and defense. Several works advocate for attempting to differentiate between "natural" and "adversarial" inputs (Gong et al., 2017; Grosse et al., 2017; Metzen et al., 2017), while other works describe the ability of adversaries to circumvent such a defense (Carlini and Wagner, 2017; Athalye et al., 2018). 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 (Goodfellow et al., 2014a).

One work along this vein which relates closely with our work is (Moosavi-Dezfooli et al., 2019). 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, (Bubeck et al., 2019) suggests that in some settings computation is the primary bottleneck in constructing adversarially robust classifiers. (Gilmer et al., 2018; Mahloujifar et al., 2019; Shafahi et al., 2018) all highlight how high dimensional

geometry induces inherent limitations in the ability to avoid adversarial examples. (Ilyas et al., 2019) 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. Finally, the interplay between the geometry of the types of perturbations used in measuring adversarial attacks was explored in (Khoury and Hadfield-Menell, 2018). That work demonstrated that adversarial robustness with respect to ℓ^∞ norm perturbations is not equivalent to ℓ^2 norm perturbations, and that under a manifold hypothesis adversarial examples may be a consequence of the the complicated nature of high-dimensional geometry.

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 (Blanchet et al., 2019; Gao et al., 2017); in (Blanchet et al., 2019) 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 (Bhagoji et al., 2019; Pydi and Jog, 2019), 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. (Pydi and Jog, 2019) uses Strassen's theorem from the theory of optimal transportation (Villani, 2003) 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. These works have focused on the sufficient conditions associated with duality principles, but to our knowledge there is no work deriving the necessary conditions associated with optimal decision boundaries of adversarially robust classifiers.

Tracking the effect of a regularization parameter on optimal solutions of a statistical problem has been studied in various contexts. For example, the evolution of optimal solutions of ℓ_1 regularized regression problems (i.e. Lasso) were studied in (Belloni et al., 2011). More recently, in the context of parametric adversarial learning, (Javanmard et al., 2020) studied the tradeoff between accuracy and adversarial robustness as a function of "ε". In that work the optimal solutions admit direct representation formulas, and hence one can directly describe the evolution of the optimal classifier. In contrast, our work focuses on non-parametric classifiers, and to our knowledge no other works attempt to describe the evolution, in terms of a differential equation, of classification boundaries as a function of the adversarial power.

Finally, it is worth mentioning that other notions of classification robustness have been introduced in the literature (Wang et al., 2018). 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 (Ecker, 2012). High fidelity numerical approximations are also available (Merriman et al., 1992). Our evolution equation is also not unrelated to non-local versions of curvature flow, which also are a topic of significant current interest (Chambolle et al., 2015).

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 (Calatroni et al., 2017; Cucuringu et al., 2019; Hu et al., 2013; Jacobs et al., 2018; Merkurjev et al., 2013, 2018; van Gennip et al., 2014) and references within.

3. Problem setup

Let ν 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 ρ_0, ρ_1 that are assumed to satisfy certain regularity and non-degeneracy properties that we will make precise later on (for example see Assumptions 5 for the one-dimensional setting). We use ρdx to denote the marginal distribution of X. Notice that ρ 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. Our conventional notation throughout the paper is that $\rho_i(z+w)$ is always meant to denote ρ_i evaluated at $z+w$, while any multiplication of ρ_i by $(z+w)$ will be denoted using $\rho_i \cdot (z+w)$. We notice that as a consequence of Bayes' theorem μ may be written using

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

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 ε is a parameter that describes the power of the adversary: the larger the value of ε, the more the adversary can perturb the data. In this setting, one seeks to build a classifier that minimizes the robust risk

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

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 (Bhagoji et al., 2019; Pydi and Jog, 2019) 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

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

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 ε. This information may be encoded by rewriting our objective functional R_ε 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:

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

where we have used the fact that ρ_1 is a probability distribution. We are interested in the robust classification problem:

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

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 (Bhagoji et al., 2019; Pydi and Jog, 2019). 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

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

where here $\Gamma(\rho_1,\rho_0)$ denotes the set of probability measures on $\mathbb{R}^d\times\mathbb{R}^d$ with marginals ρ_1 and ρ_0 (i.e. the set of couplings or transportation plans between ρ_1 and ρ_0); the above result is closely connected to Strassen's theorem (see Corollary 1.28 in (Villani, 2003)). This result may be restated in the following way

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

This duality principle provides a means of certifying the optimality of solutions to the functional R_ε, as is common in the context of convex optimization. Our later proofs establishing the global optimality of solutions that we construct using evolution equations will directly utilize this duality principle. Previous results in this vein focused only on the case with balanced classes: here we extend their results to the case of unbalanced classes. Indeed, the remainder of this section provides a direct generalization of the duality results given in (Bhagoji et al., 2019; Pydi and Jog, 2019).

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

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

In words, ν^S is simply the data distribution after swapping the y labels. Using the measures ν and ν^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 2 *Let $c_\varepsilon : (\mathbb{R}^d \times \{0,1\})^2 \to \mathbb{R}$ be the cost defined by*

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

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 (Villani, 2003). First, by the Kantorovich duality theorem (see Theorem 1.3 in (Villani, 2003)) we have

$$\sup_{\phi(z_1)+\psi(z_2)\leq c_\varepsilon(z_1,z_2)} \int \phi(z_1) d\nu(z_1) + \int \psi(z_2) d\nu^S(z_2) = \inf_{\pi\in\Gamma(\nu,\nu^S)} \int c_\varepsilon(z_1,z_2) d\pi(z_1,z_2). \quad (3.2)$$

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 ν almost every z_1 and for ν^S almost every z_2.

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

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

Notice that ϕ^{c_ε} can be written as:

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

Similarly we find that

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

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 (3.2) can be restricted to pairs ϕ, ψ satisfying the cost constraint and $\psi \in [-1, 0]$.

Let us now consider an arbitrary ψ taking values in $[-1, 0]$ with its best associated ϕ:

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

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

Since we are only considering ψ which take non-positive values, it follows that

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

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

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

which is then seen to take values on $[-1, 0]$. Since ϕ^{c_ε} is the best ψ for a given $\phi \in [0, 1]$, it follows that the supremum in (3.2) 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 (which we recall in Lemma 16 in the Appendix),

$$\begin{aligned}\int \phi(z)d\nu(z) + \int \phi^{c_\varepsilon}(\tilde{z})d\nu^S(\tilde{z}) &= \int_0^1 \int \mathbb{1}_{\phi(z)>s} d\nu(z)ds - \int_0^1 \int \mathbb{1}_{-\phi^{c_\varepsilon}(\tilde{z})>s} d\nu^S(\tilde{z})ds,\\ &= \int_0^1 \left(\int \mathbb{1}_{\phi(z)>s} d\nu(z)ds - \int \mathbb{1}_{-\phi^{c_\varepsilon}(\tilde{z})>s} d\nu^S(\tilde{z})\right) ds. \qquad (3.3)\end{aligned}$$

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

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

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

$$\int_0^1 \left(\int \mathbb{1}_{\phi(z)>s} d\nu(z)ds - \int \mathbb{1}_{-\phi^{c_\varepsilon}(\tilde{z})>s} d\nu^S(\tilde{z})\right) ds \leq \int \mathbb{1}_{\phi(z)>s} d\nu(z)ds - \int \mathbb{1}_{-\phi^{c_\varepsilon}(\tilde{z})>s} d\nu^S(\tilde{z})$$

$$= \int \mathbb{1}_{\{\phi(x,0)>s\}} w_0\rho_0(x)dx + \int \mathbb{1}_{\{\phi(x,1)>s\}} w_1\rho_1(x)dx$$
$$- \int \mathbb{1}_{\{\phi(x,0)>s\}^{2\varepsilon}} w_1\rho_1(x)dx - \int \mathbb{1}_{\{\phi(x,1)>s\}^{2\varepsilon}} w_0\rho_0(x)dx,$$

where we have used the definitions of ν and ν^S. The above computations, along with (3.3), allow us to conclude that:

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

In the previous computation the step from line two to line three is the only one which does not follow directly from definitions: its justification relies upon technical measure-theoretical arguments which can be found in the appendix of (Pydi and Jog, 2019), and which we omit here for the sake of brevity. 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.

■

We now translate the previous proposition, which mirrors the terminology used in describing duality in optimal transportation and linear programming, into a form which directly links the adversarial classification problem with the transportation problem from the previous proposition.

Corollary 3 *$\mathbb{1}_A$ for some $A \in \mathcal{M}(\mathbb{R}^d)$ minimizes R_ε 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,

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

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

$$\begin{aligned}\inf_{A\in\mathcal{M}(\mathbb{R}^d)} R_\varepsilon(\mathbb{1}_A) &= w_1 - \sup_A \left\{ \int_{A^{-\varepsilon}} w_1\rho_1 dx - \int_{A^\varepsilon} w_0\rho_0 dx \right\} \\ &= w_1 - \frac{1}{2}(w_1 - w_0) - \frac{1}{2}\inf_{\pi\in\Gamma(\nu,\nu^S)} c_\varepsilon(z_1,z_2)d\pi(z_1,z_2)\end{aligned}$$

■

Remark 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

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

4. Necessary conditions and corresponding evolution equation in one dimension

We now describe, in detail, the necessary conditions for minimizing R_ε, and the evolution equation that they induce. For clarity, we begin by describing this evolution equation in the simple case where $x \in \mathbb{R}$, under the standard metric. In this case we will be able to prove that the resulting evolution equation completely characterizes the global minimizer of R_ε for small ε 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_ε^* in terms of two parametrized collections of points $a_i(\varepsilon)$ and $b_i(\varepsilon)$, so that $A_\varepsilon^* = \cup_{i=1}^K [a_i(\varepsilon), b_i(\varepsilon)]$. 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), it makes no difference whether the subintervals 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 ε, in order to decrease the notational burden. Furthermore, and following our notational convention for the ρ_i, any time a_i, b_i are followed by parentheses we always mean that the parentheses denote function evaluation: cases where multiplication is implied will be denoted by $a_i \cdot z$. We use the convention $a_1 < b_1 < a_2 < b_2 < \cdots < a_K < b_K$.

As A_ε^* is a minimizer of R_ε, 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 $A(\delta) = (a_1, b_1 + \delta) \cup \left(\cup_{i=2}^K (a_i, b_i)\right)$, then since A_ε^* is a minimizer we have that $R_\varepsilon(A(\delta)) -$

$R_\varepsilon(A^*_\varepsilon) \geq 0$. Taking $\delta \to 0$ and using the fundamental theorem of Calculus then allows us to write

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

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

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

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 ε 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^*_ε for small changes in ε. In particular, let us suppose that each boundary point varies smoothly in ε, namely that we express $a_i(\varepsilon)$ and $b_i(\varepsilon)$ as smooth functions in ε. 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 $\frac{db_i}{d\varepsilon}$,

$$\frac{db_i}{d\varepsilon} = -\frac{w_0\rho_0'(b_i + \varepsilon) + w_1\rho_1'(b_i - \varepsilon)}{w_0\rho_0'(b_i + \varepsilon) - w_1\rho_1'(b_i - \varepsilon)}. \tag{4.2}$$

The necessary condition for the a_i is analogous:

$$\frac{da_i}{d\varepsilon} = -\frac{w_1\rho_1'(a_i + \varepsilon) + w_0\rho_0'(a_i - \varepsilon)}{w_1\rho_1'(a_i + \varepsilon) - w_0\rho_0'(a_i - \varepsilon)}. \tag{4.3}$$

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

$$\frac{db_i}{d\varepsilon}[\varepsilon = 0] = -\frac{\rho'(b_i)}{w_0\rho_0'(b_i) - w_1\rho_1'(b_i)}.$$

This indicates that the b_i initially moves downhill in ρ, 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). 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 $\mathbb{P}(X \in dx|Y = 1) = \phi(x)dx$, where ϕ is the standard normal density $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$, and let $\mathbb{P}(X \in dx|Y = 0) = \frac{\phi((x-2)/2)dx}{2}$. Assume also that $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 0) = 1/2$. Since the variances of the two Gaussians $\mathbb{P}(X \in dx|Y = 0)$ 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 (Pydi and Jog, 2019) (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 ε small enough.

In order to provide concrete numerical values for the decision boundary as a function of ε we use a standard ODE solver in Python. The decision boundary, as well as the associated densities, are given in Figure 1. We notice that a_ε moves to the left, which is consistent with the fact that ρ has positive slope at $a(0)$. Similarly, the b_ε moves to the right, which is consistent with the fact that ρ has negative slope at $b(0)$. These decision boundaries required *no optimization*, and are provably global minimizers for small ε.

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 ε, 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 *We make the following assumptions on the densities ρ_0 and ρ_1.*

i) *Regularity condition:* $\rho_0, \rho_1 \in C^1(\mathbb{R})$.

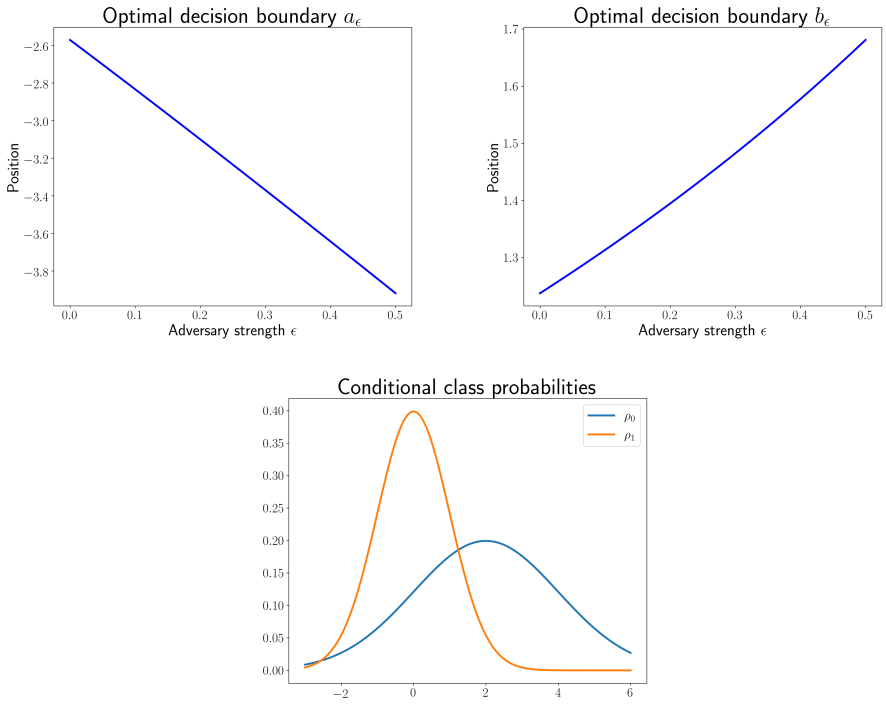

Figure 1: Plot of decision boundaries as ε varies for example in Section 4.1 , as well as the underlying probabilities.

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* $w_0\rho_0'(t) \neq w_1\rho_1'(t)$.

We pause to briefly discuss these assumptions. First, we notice that the points in the boundary of the Bayes' classifier will necessarily satisfy $w_0\rho_0(t) = w_1\rho_1(t)$. Condition ii) then can be restated as requiring that the Bayes classifier be composed of finitely many intervals on the support of ρ. Condition iii), which only makes sense if we assume enough regularity, i.e. Condition i), rules out degeneracies, implying that the Bayes' classifier is essentially unique. Condition iii) also implies that the Bayes' classifier is stable under C^1 perturbations of the ρ_i: in a sense, this is what is leveraged in the proof of our main result, namely Theorem 8. These conditions should be seen as relatively mild, and indeed Condition iii) ought to be generic within the class of C^1 functions. For example, if we require that the ρ_i be given by finite mixtures of Gaussians, then these assumptions will hold for almost every choice of parameters (i.e. means and variances). Before stating the main result, we begin with a few remarks which will be important in our proof strategy.

Remark 6 (Global optimality via duality) *Suppose that A is a measurable subset of $\mathbb{R}^d$ that satisfies*

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

for some $\pi_\varepsilon\in\Gamma(\nu,\nu^S)$. Then, it follows from Proposition 2 that A and π_ε are solutions to the optimization problems in that same proposition, and by Corollary 3, A is also a minimizer of (3.1).

Remark 7 (Knott-Smith optimality criterion) *According to Remark* (6)*, to show that a given measurable set A is an optimizer for* (3.1)*, we would need to construct a coupling π_ε 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:*

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

Then, it is straightforward to check that A and π_ε satisfy (5.1) *(with equality). The above condition for π_ε suggests then how mass must be exchanged between the measures ν and ν^S in order to get an optimal coupling. This insight is used to build the coupling from Theorem 8 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 8 *Under Assumptions 5 on ρ_0,ρ_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)*, which we recall to be*

$$\begin{aligned}\frac{db_i}{d\varepsilon}&=-\frac{w_0\rho_0'(b_i+\varepsilon)+w_1\rho_1'(b_i-\varepsilon)}{w_0\rho_0'(b_i+\varepsilon)-w_1\rho_1'(b_i-\varepsilon)},\\ \frac{da_i}{d\varepsilon}&=-\frac{w_1\rho_1'(a_i+\varepsilon)+w_0\rho_0'(a_i-\varepsilon)}{w_1\rho_1'(a_i+\varepsilon)-w_0\rho_0'(a_i-\varepsilon)},\end{aligned}$$

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 6 the set A^*_ε induces an optimal robust classifier for ε, i.e., it is a solution to the Problem (3.1)*.*

The proof of this result is somewhat technical, because one needs to explicitly construct the transportation plans in question. We construct these plans in several steps, which can be related to the different cases in the 0-1 cost function. The most crucial step of this construction, and the one which requires Assumption iii), involves how to construct the

part of the transportation plan close to the classification boundary. This is carried out in Steps 1-3 below, and is illustrated in Figure 2. Step 4 constructs the (relatively simple) remainder of the transportation plan and wraps up the proof.

Proof

Let us recall that by convention we have ordered the endpoints as $a_1(0) < b_1(0) < a_2(0) < b_2(0) < \cdots < a_K(0) < b_K(0)$. 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, are guaranteed to satisfy the necessary conditions (4.1). This fact will be used repeatedly below.

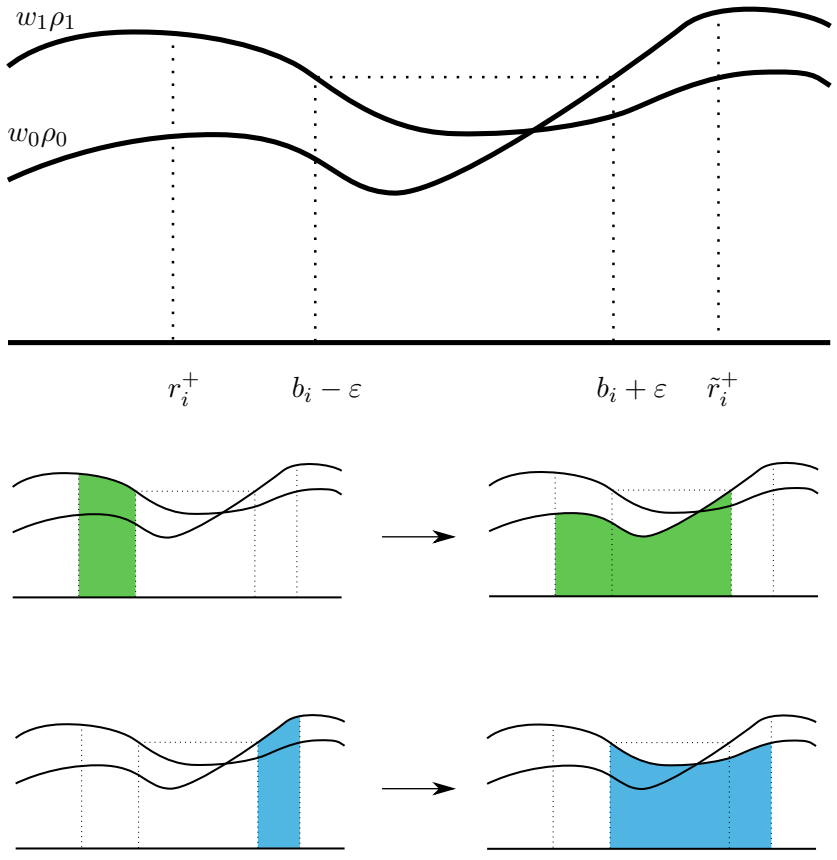

Figure 2: Illustration of mass exchange defined by γ_{b_i} (middle) and by $\tilde{\gamma}_{b_i}^{-1}$ (bottom).

Step 1: [Construct matching from the left of b_i] Figure 2 provides a visual illustration of the particular construction that we use in our transportation plan near the

boundary point $b_i(\varepsilon)$. In Step 1, we are focusing on constructing the mapping involving the green mass in the plot. This mapping is constructed in such a way that the total green mass on the left and the right are equal, and so that no mass needs to travel a distance greater than 2ε. We notice that the heights at $b_i(\varepsilon)-\varepsilon$ and $b_i(\varepsilon)+\varepsilon$ are equal, and so that mass travels distance exactly 2ε. The slope conditions assumed in Assumption 5, which we can show continues to hold locally near $b_i(0)$, is what allows us to infer that the rest of the green mass indeed is transported less than distance 2ε.

To begin the construction, let us fix a particular i corresponding to a finite right endpoint $b_i(\varepsilon)$ 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 ε to ease the notation). Now, by Assumption 5 we know that $w_0\rho_0'(b_i(0)) \neq w_1\rho_1'(b_i(0))$. Given that when $\varepsilon = 0$ we have that A_ε^* is the Bayes' classifier, we may deduce that $w_0\rho_0 < w_1\rho_1$ inside $(a_i(0), b_i(0))$. Hence $w_0\rho_0'(b_i(0)) > w_1\rho_1'(b_i(0))$. Moreover, the fact that ρ_0, ρ_1 are $C^1(\mathbb{R})$ allows us to deduce that

$$w_0\rho_0'(t_0) > w_1\rho_1'(t_1) \tag{5.4}$$

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

$$\frac{d}{ds}\left(w_0\rho_0(b_i + \varepsilon - s)\right) < \frac{d}{ds}\left(w_1\rho_1(b_i - \varepsilon - s)\right), \quad \forall s \in (0, \delta/2). \tag{5.5}$$

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

$$w_0\rho_0(b_i + \varepsilon - s) \leq w_1\rho_1(b_i - \varepsilon - s), \quad \forall s \in (0, \delta/2), \tag{5.6}$$

for all small enough $\varepsilon > 0$.

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

$$\int_{r_i^+}^{b_i-\varepsilon} w_1\rho_1(x)dx = \int_{r_i^+}^{b_i+\varepsilon} w_0\rho_0(x)dx. \tag{5.7}$$

The existence of r_i^+ (at least for small enough ε) follows from (5.5) 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,

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

Now we define the function $\phi_{b_i} : [r_i^+, b_i - \varepsilon] \to [r_i^+, b_i + \varepsilon]$ 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 (5.6), ϕ_{b_i} satisfies:

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

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

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

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 $\sharp$ denotes the push-forward operation and $\llcorner$ the restriction of a measure to a given set. Here we recall that the push-forward of a measure μ (defined over a space Ω_1) by a map $F : \Omega_1 \to \Omega_2$ is a measure on Ω_2 defined by $F_\sharp\mu(B) = \mu(F^{-1}(B))$, where by F^{-1} is the inverse image. 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: [Construct matching from the right of b_i] In Step 2 we are repeating the same type of construction that we did in Step 1, but to the other side of the boundary point. In terms of the illustration in Figure 2, we are now describing the transportation of the blue mass at the bottom of the figure. As in Step 1, we have to guarantee that such a mapping exists and transports mass at most distance 2ε.

To achieve this goal, we consider a symmetric construction to the one from Step 1. Using again (5.4) and combining with the necessary condition for b_i in (4.1) we obtain:

$$w_1\rho_1(b_i - \varepsilon + s) \leq w_0\rho_0(b_i + \varepsilon + s), \quad \forall s \in (0, \delta/2). \tag{5.8}$$

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

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

using similar arguments to the ones employed in Step 1 (including utilizing Assumption 5).

We let $\tilde{\phi}_{b_i} : [b_i - \varepsilon, \tilde{r}_i^+] \to [b_i + \varepsilon, \tilde{r}_i^+]$ 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 (5.8) implies

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

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

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

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: [Construct matchings for a_i] 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, and again utilizing Assumption 5, 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 $\phi_{a_i} : [a_i + \varepsilon, r_i^-] \to [a_i - \varepsilon, r_i^-]$ and $\tilde{\phi}_{a_i} : [\tilde{r}_i^-, a_i + \varepsilon] \to [\tilde{r}_i^-, a_i - \varepsilon]$ 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

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

Step 4:[Construct remainder of plan and compute cost] 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 π_ε.

Let ν_0^R, ν_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 $[r_i^-, r_i^+] \subseteq [a_i(0), b_i(0)]$, ν_0 is indeed a positive measure. Similarly, we can see that ν_1^R is a positive measure too . From our construction it follows that

$$\int_{r_i^+}^{\tilde{r}_i^+} w_0\rho_0 \, dx = \int_{r_i^+}^{\tilde{r}_i^+} w_1\rho_1 \, dx, \qquad \int_{\tilde{r}_i^-}^{r_i^-} w_0\rho_0 \, dx = \int_{\tilde{r}_i^-}^{r_i^-} w_1\rho_1 \, dx, \tag{5.9}$$

for all i, and hence

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

Let π^R be *any* coupling between the measures

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

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

$$\begin{aligned}\pi^0(dz_1, dz_2) := \sum_{i=1}^{K} (&\gamma_{b_i}(dx_1, dx_2) \otimes \delta_{\{0\}\times\{0\}}(dy_1, dy_2) + \gamma_{b_i}^{-1}(dx_1, dx_2) \otimes \delta_{\{1\}\times\{1\}}(dy_1, dy_2) \\ &+ \tilde{\gamma}_{b_i}(dx_1, dx_2) \otimes \delta_{\{0\}\times\{0\}}(dy_1, dy_2) + \tilde{\gamma}_{b_i}^{-1}(dx_1, dx_2) \otimes \delta_{\{1\}\times\{1\}}(dy_1, dy_2) \\ &+ \gamma_{a_i}(dx_1, dx_2) \otimes \delta_{\{0\}\times\{0\}}(dy_1, dy_2) + \gamma_{a_i}^{-1}(dx_1, dx_2) \otimes \delta_{\{1\}\times\{1\}}(dy_1, dy_2) \\ &+ \tilde{\gamma}_{a_i}(dx_1, dx_2) \otimes \delta_{\{0\}\times\{0\}}(dy_1, dy_2) + \tilde{\gamma}_{a_i}^{-1}(dx_1, dx_2) \otimes \delta_{\{1\}\times\{1\}}(dy_1, dy_2)).\end{aligned}$$

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 π^F be the measure on $(\mathbb{R} \times \{0,1\})^2$ given by $\pi^F := (Id \times Id)_\sharp(\nu - (\pi^0 + \pi^R)_1)$, 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, π^0 satisfies the property that for all the points in its support $c_\varepsilon = 0$. π^F corresponds to the mass that is fixed and thus does not contribute to the cost of π_ε. Finally, π^R corresponds to the remaining mass. Our construction then guarantees that

$$\int c_\varepsilon(z_1, z_2) d\pi_\varepsilon(z_1, z_2) = \int c_\varepsilon(z_1, z_2) d\pi^R(z_1, z_2) \leq \pi^R((\mathbb{R} \times \{0,1\})^2)$$

$$= \nu_0^R(\mathbb{R}) + \nu_1^R(\mathbb{R}) = 2\nu_0^R(\mathbb{R}) + w_0 - w_1 = 2\sum_{i=1}^{K} \int_{r_i^-}^{r_i^+} (w_1\rho_1 - w_0\rho_0)dx + w_0 - w_1.$$

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 (5.7) 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 6 implies that A^* solves the optimization problem and that the above inequality is actually an equality.

■

Remark 9 *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* (5.9)*, which factored prominently in the construction of the certifying transportation plan* π_ε*.*

Remark 10 *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* ε *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* ε*. This should correspond to a change in topology of the global optimizer. On the other hand, the fact that global minimizers are characterized by the differential equation implies that the topology of the minimizers does not change for small values of* ε*, and if one chose to track the range of values for which the constructions in the previous proof were valid (which would depend upon the difference between* $w_0\rho_0'$ *and* $w_1\rho_1'$ *and upon the* C^1 *norms of the densities) one could quantify the range of* ε *for which the topology does not change.Understanding the type of degeneracies that may arise when solving the geometric evolution equations, and the associated changes in topology of the optimizers,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 ε. This evolution equation was driven, for small ε, by the gradient of ρ. 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. For clarity, throughout our study of dimension $d > 1$ we will restrict our attention to the standard Euclidean metric, namely we let $d(x_1, x_2) = |x_1 - x_2|$.

To begin, we will develop some intuition about the problem by studying an explicit, radial example.

Example 1 *Let us consider the case where* ρ *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* ω_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* ε*, which (by way of ansatz) is given by* $\mathbb{1}_{|x|\leq r(\varepsilon)}$*, which minimizes the adversarial cost. Necessary*

conditions for optimality then take the form

$$\begin{aligned}0 &= \lim_{\delta\to 0}\frac{R_\varepsilon(r(\varepsilon)+\delta)-R_\varepsilon(r(\varepsilon))}{\delta}\\ &= \omega_d w_0\rho_0(r(\varepsilon)+\varepsilon)\cdot(r(\varepsilon)+\varepsilon)^{d-1}-\omega_d w_1\rho_1(r(\varepsilon)-\varepsilon)\cdot(r(\varepsilon)-\varepsilon)^{d-1}\\ &= (r(\varepsilon)+\varepsilon)\cdot(r(\varepsilon)+\varepsilon)^{d-1}-(1-(r(\varepsilon)-\varepsilon))\cdot(r(\varepsilon)-\varepsilon)^{d-1},\end{aligned}$$

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 $|x|=t$, and writing $R_\varepsilon(s)=R_\varepsilon(\mathbb{1}_{|x|\le s})$. Taking a derivative in ε we obtain

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

which may be written

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

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 mean curvature flow that has been reweighted in the denominator by a density-dependent factor. We notice that here, $\nabla\rho\equiv 0$, which in the one-dimensional case dominated the evolution for small ε 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.

We remark that, in order to make the formulas explicit, we choose to work with the uniform density on the ball, which has non-smooth density. We notice that the classification boundary occurs in the region where the density is smooth, and so mollifying the density near the boundary of the ball would not materially affect the behavior of the example besides complicating the formulas for the total density.

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

Proposition 11 *Suppose that A_ε is a critical point of the problem R_ε (with respect to normal variations (Maggi, 2012), further description given in the proof below) 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 ν denote the outward unit normal and κ_i denote the principal curvatures (see the Appendix for a definition). Then the following necessary condition holds for almost every $x\in\partial A_\varepsilon$:*

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

Proof We again recall

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

We consider the class of normal variations (Maggi, 2012) 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 smooth scalar valued function ψ that we assume, without loss of generality, satisfies $\nabla\psi(x) \cdot \nu(x) = 0$ for all $x \in \partial A$. 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, meaning the point in the boundary of A which is closest to x, whose uniqueness is guaranteed by the assumption upon the regularity of the signed distance function.

Noting that $y = x \pm \varepsilon\nu(x)$ in the previous two integrals, we then use a change of variables as in Corollary 15 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 ψ, we then have that, for $\mathcal{H}^{d-1}$ almost every $x \in \partial A$

$$0 = \left(w_0\rho_0(x + \varepsilon\nu(x)) \prod_{i=1}^{d-1} |1 + \kappa_i(x)\varepsilon| - w_1\rho_1(x - \varepsilon\nu(x)) \prod_{i=1}^{d-1} |1 - \kappa_i(x)\varepsilon| \right) \tag{6.2}$$

■

We notice that assuming that the conditional densities ρ_0, ρ_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 2 *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 $(0, 0)$.*

In Proposition 11 we notice that the necessary condition directly utilizes the normal vectors to the surface and the curvatures (which may be viewed as derivatives of the normal vectors). These notions are intimately tied with the classical Euclidean geometry: indeed if we did not have $d(x_1, x_2) = |x_1 = x_2|$ then the previous theorem would need to be modified in order to accommodate for normal vectors and their derivatives in the appropriate geometry. Such definitions have been pursued in the context of mean curvature flow, for example in (Bellettini, 2004). However, extending those definitions to apply to the present context is beyond the scope of this work.

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_ε. As in the one-dimensional case, we can Taylor expand for ε 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, ε, and always assuming that $x \in \partial A_0$):

$$
\begin{aligned}
0 &= \frac{d}{d\varepsilon}\left(w_0\rho_0(\phi + \varepsilon\nu(\phi)) \prod_{i=1}^{d-1}(1+\varepsilon\kappa_i(\phi)) - w_1\rho_1(\phi - \varepsilon\nu(\phi) \prod_{i=1}^{d-1}(1-\varepsilon\kappa_i(\phi))\right) \\
&= \prod_{i=1}^{d-1}(1+\varepsilon\kappa_i(\phi))\left(\nabla w_0\rho_0(\phi+\varepsilon\nu(\phi)\left(\frac{d}{d\varepsilon}\phi + \nu(\phi) + \varepsilon\frac{d}{d\varepsilon}(\nu(\phi))\right) + w_0\rho_0(\phi+\varepsilon\nu(\phi))\sum_i \frac{\kappa_i(\phi) + \varepsilon\frac{d}{d\varepsilon}\kappa(\phi)}{1+\varepsilon\kappa_i(\phi)}\right) \\
&- \prod_{i=1}^{d-1}(1-\varepsilon\kappa_i(\phi))\left(\nabla w_1\rho_1(\phi-\varepsilon\nu(\phi)\left(\frac{d}{d\varepsilon}\phi - \nu(\phi) - \varepsilon\frac{d}{d\varepsilon}(\nu(\phi))\right) + w_1\rho_1(\phi-\varepsilon\nu(\phi))\sum_i \frac{-\kappa_i(\phi) - \varepsilon\frac{d}{d\varepsilon}\kappa(\phi)}{1-\varepsilon\kappa_i(\phi)}\right)
\end{aligned}
\tag{6.3}
$$

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 ε and x. Indeed, we recall (see for example Section 17.1 in (Maggi, 2012)) that, in terms of ϕ and for $x \in \partial A_0$, one may express the geometric quantity ν (the outward surface normal) as

$$
\nu(\phi(\varepsilon, x)) = \frac{(\frac{\partial}{\partial x}\phi(\varepsilon, x))^{-T} \cdot \nu_0(x)}{|(\frac{\partial}{\partial x}\phi(\varepsilon, x))^{-T} \cdot \nu_0(x)|}
$$

Similarly, the curvatures κ_i may be expressed in terms of appropriate spatial derivatives of ν, more precisely, as the non-trivial eigenvalues of the matrix $\frac{\partial \nu}{\partial x}$; see Proposition 14 in the Appendix. 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 ε, and hence may plausibly be ignored for ε sufficiently small. To this end, we rearrange the previous equation

$$
\begin{aligned}
&\left(\prod_{i=1}^{d-1}(1+\varepsilon\kappa_i(\phi))\nabla w_1\rho_1(\phi+\varepsilon\nu(\phi)) - \prod_{i=1}^{d-1}(1-\varepsilon\kappa_i(\phi))\nabla w_0\rho_0(\phi-\varepsilon\nu(\phi))\right)\frac{d}{d\varepsilon}\phi \\
&= -\prod_{i=1}^{d-1}(1+\varepsilon\kappa_i(\phi))\left(\nabla w_1\rho_1(\phi+\varepsilon\nu(\phi)(\nu(\phi) + \varepsilon\frac{d}{d\varepsilon}\nu(\phi)) + w_1\rho_1(\phi+\varepsilon\nu(\phi))\sum_i \frac{\kappa_i(\phi) + \varepsilon\frac{d}{d\varepsilon}\kappa(\phi)}{1+\varepsilon\kappa_i(\phi)}\right) \\
&- \prod_{i=1}^{d-1}(1-\varepsilon\kappa_i(\phi))\left(\nabla w_0\rho_0(\phi-\varepsilon\nu(\phi)(\nu(\phi) + \varepsilon\frac{d}{d\varepsilon}\nu(\phi)) + w_0\rho_0(\phi-\varepsilon\nu(\phi))\sum_i \frac{\kappa_i(\phi) + \varepsilon\frac{d}{d\varepsilon}\kappa(\phi)}{1-\varepsilon\kappa_i(\phi)}\right).
\end{aligned}
\tag{6.4}
$$

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

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

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

$$v(x, \varepsilon = 0) = -\frac{\nabla \rho \cdot \nu + \rho \sum_i \kappa_i}{(w_1 \nabla \rho_1 - w_0 \nabla \rho_0) \cdot \nu} \tag{6.5}$$

Here we observe two terms: one which induces motion "downhill" in ρ and a second which is a positively weighted mean curvature term. As we have used ν 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 ρ, and ii) decrease the perimeter of the decision boundary (since mean curvature flow is a type of gradient flow of perimeter; see Section 6.1.1 below.) While the reweighing in the denominator is not homogeneous, and indeed makes this heuristic description imprecise, we believe this heuristic picture is helpful for understanding the local effects induced by adversarial robustness.

Mean curvature flow, namely the case where $v(x) = \sum_i \kappa_i$ without any density weighting, is a fundamental geometric flow. One reference text, among many, on the topic is (Mantegazza, 2011). This geometric flow is known to be the gradient flow of the perimeter or area functional with respect to a certain function space. It is also known to induce increased smoothness in surfaces, when measured in the correct function spaces. Finally, mean curvature flow obeys a comparison principle and admits efficient numerical methods. While the flow that we have derived for the adversarial problem does not match mean curvature flow exactly, one of the terms in the approximate surface evolution at $\varepsilon = 0$ amounts to a scalar function times mean curvature flow, suggesting that the flow induced by the adversarial problems also may enjoy similar useful characteristics. In the next section we take this analogy one step further by deriving a variant of perimeter regularization, which matches the adversarial evolution equation to higher order.

6.1.1 Connection with explicit perimeter regularization

As mentioned at the end of the Introduction, there is a close relationship between the evolution equation derived earlier and the one that one would obtain by tracking solutions to the family of problems (1.2) indexed with ε. In what follows we elaborate on this statement. It will be convenient to write R explicitly as:

$$R(\mathbb{1}_A) = \int_A w_0 \rho_0(x) dx + w_1 - \int_A w_1 \rho_1(x) dx.$$

First, let us derive necessary conditions for an optimal solution A to problem (1.2). We assume that A has at least C^2 boundary for simplicity. Let ϕ be a normal variation of A as considered in Section 6.1 with the same notation used there. Then the following two

relations hold:

$$\begin{aligned}\frac{d}{dt}\Big|_{t=0} R(\mathbb{1}_{A^t}) &= \frac{d}{dt}\Big|_{t=0}\left(\int_{A^t}(w_0\rho_0(x) - w_1\rho_1(x))dx\right)\\ &= \int_{\partial A}\psi(x)(w_0\rho_0(x) - w_1\rho_1(x))d\mathcal{H}^{d-1}(x),\end{aligned}$$

and

$$\begin{aligned}\frac{d}{dt}\Big|_{t=0}\mathrm{Per}_\rho(A^t) &= \frac{d}{dt}\Big|_{t=0}\left(\int_{\partial A^t}\rho(z)d\mathcal{H}^{d-1}(z)\right)\\ &= \frac{d}{dt}\Big|_{t=0}\int_{\partial A}\rho\left(x + t\psi(x)\nu(x)\right)\left|\det\left(I + t\psi(x)\frac{\partial\nu(x)}{\partial x}\right)\right|\mathcal{H}^{d-1}(x)\\ &= \int_{\partial A}\psi(x)(\nabla\rho(x)\cdot\nu(x) + \rho(x)\sum_i\kappa_i)d\mathcal{H}^{d-1}(x).\end{aligned}$$

We conclude that

$$\begin{aligned}0 &= \frac{d}{dt}\Big|_{t=0}\left(R(\mathbb{1}_{A^t}) + \varepsilon\mathrm{Per}_\rho(A^t)\right)\\ &= \int_{\partial A}\psi(x)(w_0\rho_0(x) - w_1\rho_1(x) + \varepsilon(\nabla\rho(x)\cdot\nu(x) + \rho(x)\sum_i\kappa_i))d\mathcal{H}^{d-1}(x).\end{aligned}$$

Since the normal variation was arbitrary, and thus the scalar function ψ was too, we deduce the necessary condition:

$$0 = w_0\rho_0(x) - w_1\rho_1(x) + \varepsilon(\nabla\rho(x)\cdot\nu(x) + \rho(x)\sum_i\kappa_i), \quad x\in\partial A. \tag{6.6}$$

Let A_ε be a solution to problem (1.2) for a given value of ε and let us assume that the family $\{A_\varepsilon\}_{\varepsilon>0}$ can be represented via a one-parameter family of diffeomorphisms $\{\phi(\varepsilon,\cdot)\}_{\varepsilon>0}$ so that $\phi(\varepsilon, A) = A_\varepsilon$, where A is the set induced by the Bayes classifier. As in Section 6.1, we now use the necessary conditions (6.6) at each point x to characterize the family of diffeomorphisms at $\varepsilon = 0$ and $x\in\partial A$. Indeed, differentiating the equation

$$\begin{aligned}0 &= w_0\rho_0(\phi(\varepsilon,x)) - w_1\rho_1(\phi(\varepsilon,x))\\ &\quad + \varepsilon(\nabla\rho(\phi(\varepsilon,x))\cdot\nu(\phi(\varepsilon,x)) + \rho(\phi(\varepsilon,x))\sum_i\kappa_i(\phi(\varepsilon,x))), \quad x\in\partial A, \quad \varepsilon\geq 0\end{aligned}$$

with respect to ε, and setting $\varepsilon = 0$, we deduce:

$$0 = (w_0\nabla\rho_0(x) - w_1\nabla\rho_1(x))\cdot\frac{d\phi}{d\varepsilon}(0,x) + (\nabla\rho(x)\cdot\nu(x) + \rho(x)\sum_i\kappa_i(x)), \quad x\in\partial A.$$

Expressing $\frac{d\phi}{d\varepsilon} = v\nu(x)$, we obtain (6.5), i.e., the same infinitesimal change as the one coming from the original adversarial problem.

6.2 Global minimizers in higher dimension

Theorem 12 *Suppose that the evolution equation* (6.3) *admits a classical solution for* $\varepsilon \in [0, \varepsilon_0]$*, in the sense that there exists a* C^2 *diffeomorphism* ϕ *which satisfies the equation at every point* $x \in \partial A_0$*. Suppose, furthermore, that* ρ_0, ρ_1 *are* C^2 *(i.e. bounded first and second derivatives) and that along the entire interface of the Bayes' classifier we have*

$$(w_0 \nabla \rho_0 - w_1 \nabla \rho_1) \cdot \nu \geq c_0 > 0, \tag{6.7}$$

for some constant c_0*, where* $\nu = \nu(x)$ *is the outer unit normal at* $x \in \partial A_0$*. Then for* ε *sufficiently small the solution to the evolution equation is also a global minimizer of the adversarial problem, where the metric determining the actions of the adversary is the Euclidean metric.*

Proof To begin, we notice that, by (6.7) along with the implicit function theorem, the boundary of A_0 is locally the graph of a C^2 function. Classical geometric results imply that for any point $\tilde{x}$ in a δ neighborhood of ∂A_0 there exists a unique closest point $P_{\partial A_0}(\tilde{x})$ in the boundary of ∂A_0, and that the signed distance function is C^3 in that same δ neighborhood of ∂A_0.

Under the assumption that ϕ is C^2, the chain rule computation shown in Section 6.1, along with the fact that A_0 is the Bayes' classifier, implies that the necessary condition (6.2) is satisfied at every point in the boundary of A_ε, namely for every $x \in \partial A_\varepsilon$

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

Next we notice that we may express points within the δ neighborhood of ∂A_0, using what is called a *normal coordinate system*. In particular, for any $\tilde{x}$ in that neighborhood, we may (uniquely) represent $\tilde{x} = P_{\partial A_0}(\tilde{x}) + \tilde{d}_{A_0}(\tilde{x})\nu(P(\tilde{x}))$. This essentially allows us to locally transform from points in a neighborhood of the boundary into flattened geometry. We note that, since ϕ is C^2 and starts as the identity mapping, the set A_ε also has boundary which is the graph of a C^2 function and admits local normal boundary coordinates, for ε sufficiently small, and for a δ neighborhood of the boundary of A_ε which is independent of ε. From this point on we will always assume that ε is small enough that this representation is possible.

Our goal now will be to construct a transportation plan which certifies the optimality of the set A_ε. In doing so, as in the one-dimensional proof, it suffices to construct mappings locally near the boundary which transfer mass from ρ_0 to ρ_1 (and vice versa), and which move mass at most 2ε distance. As we will see, our construction reduces the problem almost entirely to the one-dimensional setting.

In particular, fix $x_0 \in \partial A_0$ let $z_0 = \phi(\varepsilon, x_0)$, and consider points of the form $z_0 + t\nu(z_0) = z(t, z_0)$, for $t \in (-\delta, \delta)$. After changing variables (to be precise, in the boundary normal coordinates associated with ∂A_ε), the density associated with a particular t for fixed z_0 is given by

$$\tilde{\rho}_i(t|z_0) := \rho_i(z(t, z_0)) \prod_{i=1}^{d-1} |1 + t\kappa_i(z_0)|.$$

More precisely, using the coarea formula (Evans and Gariepy, 2015) and Corollary 15 in the Appendix, we can write

$$\begin{aligned}\int_{|\tilde{d}_{A_\varepsilon}(z)|\leq\delta} g(z)\rho_i(z)dz &= \int_{-\delta}^{\delta}\left(\int_{\tilde{d}_{A_\varepsilon}(z)=s} g(z)\rho_i(z)d\mathcal{H}^{d-1}(z)\right)dt \\ &= \int_{-\delta}^{\delta}\left(\int_{\partial A_\varepsilon} g(z(t,z_0))\rho_i(z(t,z_0))\prod_{i=1}^{d-1}|1+t\kappa_i(z_0)|d\mathcal{H}^{d-1}(z)\right)dt \\ &= \int_{\partial A_\varepsilon}\int_{-\delta}^{\delta} g(z(t,z_0))\tilde{\rho}_i(t|z_0)dtd\mathcal{H}^{d-1}(z),\end{aligned}$$

for arbitrary smooth and bounded test function g. This provides a representation of the distribution $\rho_i dx$ restricted to the set $\{|\tilde{d}_{A_\varepsilon}(z)|\leq\delta\}$ in normal coordinates, and in particular, up to rescaling, we can now interpret the function $\tilde{\rho}_i(\cdot|z_0)$ as the conditional distribution of $t\in[-\delta,\delta]$ given z_0.

Now for any fixed ε and x_0 (i.e. fixed $z_0=\phi(\varepsilon,x_0)$) we will construct a transportation plan using $\tilde{\rho}(\cdot|z_0)$: in the original high-dimensional problem this means that in the set $\{|\tilde{d}_{A_\varepsilon}(z)|\leq\delta\}$, our transportation plan only transports along rays normal to the boundary of ∂A_ε. Notice that outside of $\{|\tilde{d}_{A_\varepsilon}(z)|\leq\delta\}$, on the other hand, we may transport in any way we want so as to match the marginal constraints (just as in the 1d setting): this is why we focus on the set $\{|\tilde{d}_{A_\varepsilon}(z)|\leq\delta\}$ exclusively.

By the necessary condition (6.2) we have the necessary matching condition, namely that $w_0\tilde{\rho}_0(-\varepsilon|z_0)=w_1\tilde{\rho}_1(\varepsilon|z_0)$. All that remains is to verify that we can transport $w_0\tilde{\rho}_0(\cdot|z_0)$ on the interval $[-\varepsilon,\varepsilon]$ on to $w_1\tilde{\rho}_1(\cdot|z_0)$ without moving more than distance 2ε in t. To this end, we notice that, by the assumption (6.7), $|(w_1\nabla\rho_1(\tilde{x})-w_0\nabla\rho_0(\tilde{x}'))\cdot\nu(x_0)|>c_0/2>0$ for all $\tilde{x},\tilde{x}'$ in a neighborhood of $x_0\in\partial A_0$. Using the fact that ϕ is C^2 and that is the identity for $\varepsilon=0$, then implies that the same inequality, with constant $c_0/4$ holds in a neighborhood of $\phi(x_0,\varepsilon)$, for small enough ε, and for the size of the neighborhood independent of ε. We may then compute

$$\begin{aligned}w_1\tilde{\rho}_1'(t_1|z_0)-w_0\tilde{\rho}_0'(t_0|z_0) &= (w_1\nabla\rho_1(z(t_1,z_0))-w_0\nabla\rho_0(z(t_0,z_0)))\cdot\nu(z_0) \\ &\quad + w_1\rho_1(z(t_1,z_0))\sum_i\kappa_i(z_0)\prod_{j\neq i}|1+t_1\kappa_j(z_0)| \\ &\quad - w_0\rho_0(z(t_0,z_0))\sum_i\kappa_i(z_0)\prod_{j\neq i}|1+t_0\kappa_j(z_0)|.\end{aligned}$$

The first of these terms is bounded from below by $c_0/4$. Recalling that $w_0\rho_0(x_0)=w_1\rho_1(x_0)$ (as A_0 is a Bayes' classifier), we may bound the magnitude of the remaining terms by $|t_i|(1+\delta)$: in particular, if t_i is of order ε (notice that κ_i is uniformly bounded since ϕ was assumed C^2) then we may conclude that $w_1\tilde{\rho}_1'(t_1|z_0)-w_0\tilde{\rho}_0'(t_0|z_0)>c_0/8$ for small enough ε. This then allows us to directly use the one-dimensional construction from the proof of Theorem 8 in order to construct appropriate transportation plans for the $\tilde{\rho}(\cdot|z_0)$. By constructing such a plan along normal rays corresponding to every $z_0\in\partial A_\varepsilon$, we then have a candidate transportation plan, which transports points near the boundary at most distance 2ε. Using the same argument via the fundamental theorem of calculus and the

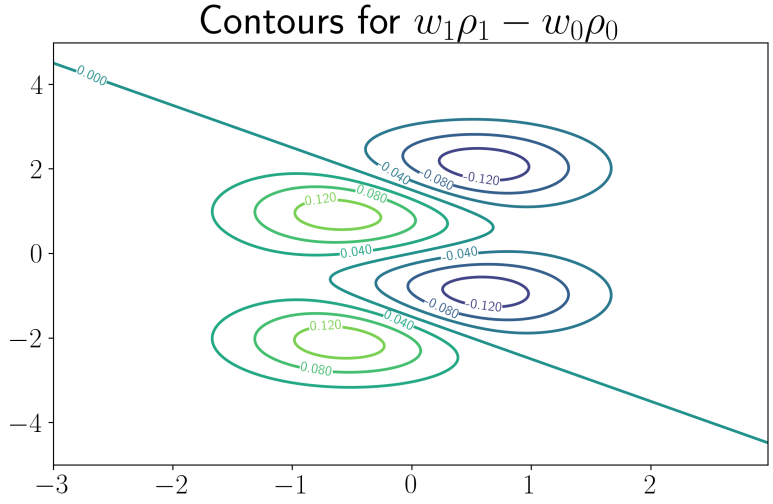

Figure 3: The contours of the function $w_1\rho_1 - w_0\rho_0$ for the example in Section 6.3. The contour corresponding to $w_1\rho_1 - w_0\rho_0 = 0$ is the s-shaped curve, and represents the decision boundary for the Bayes classifier.

duality principle as in the proof of Theorem 8, we obtain the desired result. ■

Remark 13 *In the previous proof we assumed that the solutions to the partial differential equation existed. We suspect that the local existence and uniqueness of solutions can be proved by an appropriate non-linear PDE argument under the assumption* (6.7)*, but the technical details of such a proof lie outside of the scope of this paper.*

We also notice that the same conclusions about topology which we indicated in one dimension would also hold in the setting of Theorem 12. Indeed, minimizers will not change their topology for sufficiently small ε, the size of which should depend upon the size of c_0 in (6.7) *and upon the smoothness of the underlying densities. Of course the topologies are potentially much more complicated in higher dimension, and the evolution of the topology of the minimizing classifiers is an intriguing potential future direction.*

6.3 Illustration in two dimensions

Here we show a basic numerical example of the geometric evolution (6.5) 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 $\rho_1 \sim N((-.5,-2),\Sigma) + N((-.5,.5),\Sigma)$ and $\rho_2 \sim N((.5,-.5),\Sigma) + N((.5,2),\Sigma$, 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 (Merriman et al., 1992) to track the evolution of the decision boundary under the evolution equation (6.5) for different values of ε. 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 ε (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 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) In this work, the connection between the evolution equations and global minimizers only holds for small ε, and we made no attempt to quantify the size of ε_0 in our theorem. This is partly unavoidable given the generality of the input distributions and the learned classifiers. We expect that the results in our paper should hold locally in ε, in the sense that if the adversarial problem admits a unique solution at $\tilde{\varepsilon}$ then the solution of the adversarial problem should be characterized by the evolution equation in a neighborhood of $\tilde{\varepsilon}$. We do not expect the theorem to generally hold for large ranges of ε, as topological changes can cause non-uniqueness of optimal solutions for certain (likely discrete) values of ε. It may be possible to use primal-dual methods to numerically detect when topological changes occur, or, in other words, in what ranges of ε the evolution equations describe optimal solution families. Investigating such methods could provide a lot more information about how large ε may be in our theorems.

ii) In higher dimensions, we made the additional assumption that the evolution equations admitted smooth solutions. Justifying this assumption would likely require careful analysis of the partial differential equation. Similarly, it would be interesting to develop more efficient numerical methods for the actual evolution equation in higher-dimensional settings.

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 ℓ_∞ 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.

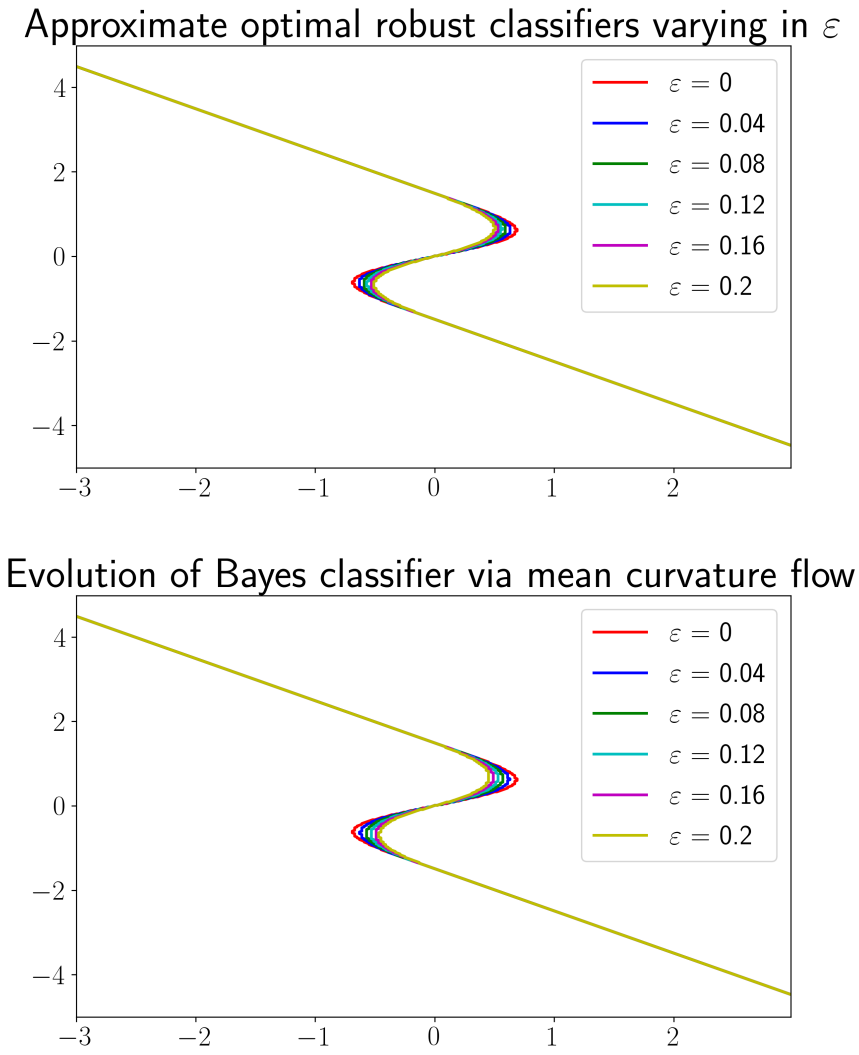

Figure 4: The first set of curves represent the evolution of the decision boundary according to the geometric evolution (6.5), which includes a weighted curvature flow and a drift term. The second set of curves is the geometric evolution following standard mean curvature flow. In this case the curves 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).

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.

Acknowledgments

NGT was supported by NSF grant DMS 2005797

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, for example, (Ambrosio and Mantegazza, 1998):

Proposition 14 *Let E be an open set with C^2 boundary. Then on some neighborhood U of ∂E we have the following:*

- $\tilde{d} \in C^2(U)$.
- *Each y in U has a unique closest point $P(y)$ in ∂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 $D^2 \tilde{d} = \frac{\partial \nu}{\partial x}$ has 1 eigenvalue that is equal to zero (with eigenvector in the normal direction ν), and $d-1$ eigenvalues with eigenvectors spanning the tangent directions. These eigenvalues are called the* principal curvatures *of the surface, and are denoted κ_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 ∂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 15 *If ∂E is a C^2 surface then for ε sufficiently small*

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

The following lemma, sometimes called the "layer cake representation", is a classical lemma from measure theory (see for example Chapter 1 in (Lieb and Loss, 2001)), and may be directly proved by using the Fubini-Tonelli theorem.

Lemma 16 *Given a non-negative function f on a measure space (X, μ) the following identity holds:*

$$\int_X f(x)\, d\mu(x) = \int_0^\infty \mu(\{x : f(x) > t\})\, dt.$$

References

Luigi Ambrosio and Carlo Mantegazza. Curvature and distance function from a manifold. *The Journal of Geometric Analysis*, 8(5):723–748, 1998.

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.

Giovanni Bellettini. Anisotropic and crystalline mean curvature flow. *A sampler of Riemann-Finsler geometry*, 50:49–82, 2004.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011. ISSN 00063444, 14643510. URL `http://www.jstor.org/stable/23076172`.

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.

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.

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.

Leon Bungert, Nicolás García Trillos, and Ryan Murray. The geometry of adversarial training in binary classification. *arXiv preprint arXiv:2111.13613*, 2021.

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. doi: 10.1007/s10851-016-0678-0. URL `https://doi.org/10.1007/s10851-016-0678-0`.

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.

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

Mihai Cucuringu, Andrea Pizzoferrato, and Yves van Gennip. An MBO scheme for clustering and semi-supervised clustering of signed networks, 2019.

Klaus Ecker. *Regularity theory for mean curvature flow*, volume 57. Springer Science & Business Media, 2012.

Lawrence Craig Evans and Ronald F Gariepy. *Measure theory and fine properties of functions*. Chapman and Hall/CRC, 2015.

Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.

Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. Adversarial and clean data are not twins. *arXiv preprint arXiv:1704.04960*, 2017.

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, 2014a.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.

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. doi: 10.1137/130917387. URL `https://doi.org/10.1137/130917387`.

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.

Matt Jacobs, Ekaterina Merkurjev, and Selim Esedoḡlu. Auction dynamics: A volume constrained MBO scheme. *Journal of Computational Physics*, 354:288 – 310, 2018. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2017.10.036. URL `http://www.sciencedirect.com/science/article/pii/S0021999117308033`.

Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pages 2034–2078. PMLR, 2020.

Marc Khoury and Dylan Hadfield-Menell. On the geometry of adversarial examples. *arXiv preprint arXiv:1811.00525*, 2018.

Elliott H Lieb and Michael Loss. *Analysis*, volume 14. American Mathematical Soc., 2001.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Francesco Maggi. *Sets of finite perimeter and geometric variational problems*, volume 135 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2012. ISBN 978-1-107-02103-7. doi: 10.1017/CBO9781139108133. URL `https://doi.org/10.1017/CBO9781139108133`. An introduction to geometric measure theory.

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.

Carlo Mantegazza. *Lecture notes on mean curvature flow*, volume 290. Springer Science & Business Media, 2011.

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. doi: 10.1137/120886935. URL `https://doi.org/10.1137/120886935`.

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.

Barry Merriman, James Kenyard Bence, and Stanley Osher. *Diffusion generated motion by mean curvature*. Department of Mathematics, University of California, Los Angeles, 1992.

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017.

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.

Muni Sreenivas Pydi and Varun Jog. Adversarial risk via optimal transport and optimal couplings. *arXiv preprint arXiv:1912.02794*, 2019.

Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018.

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.

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. doi: 10.1007/s00032-014-0216-8. URL `https://doi.org/10.1007/s00032-014-0216-8`.

Cedric Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics. American Mathematical Society, 2003. ISBN 9780821833124. URL `http://books.google.com/books?id=q6kyE2ZkxrcC`.

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, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL `http://proceedings.mlr.press/v80/wang18c.html`.