Information and Inference: A Journal of the IMA (2023) **12**, 921–968
https://doi.org/10.1093/imaiai/iaac029
Advance Access publication on 13 January 2023

The geometry of adversarial training in binary classification

Leon Bungert†
Hausdorff Center for Mathematics, University of Bonn, Endenicher Allee 62, Villa Maria, 53115 Bonn, Germany
†Corresponding author. Email: leon.bungert@hcm.uni-bonn.de

Nicolás García Trillos
Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison 53706, WI, USA

and

Ryan Murray
Department of Mathematics, North Carolina State University, 2108 SAS Hall, Raleigh 27695, NC, USA

[Received on 6 December 2021; revised on 1 August 2022; accepted on 12 October 2022]

We establish an equivalence between a family of adversarial training problems for non-parametric binary classification and a family of regularized risk minimization problems where the regularizer is a nonlocal perimeter functional. The resulting regularized risk minimization problems admit exact convex relaxations of the type L^1 + (nonlocal) TV, a form frequently studied in image analysis and graph-based learning. A rich geometric structure is revealed by this reformulation which in turn allows us to establish a series of properties of optimal solutions of the original problem, including the existence of minimal and maximal solutions (interpreted in a suitable sense) and the existence of regular solutions (also interpreted in a suitable sense). In addition, we highlight how the connection between adversarial training and perimeter minimization problems provides a novel, directly interpretable, statistical motivation for a family of regularized risk minimization problems involving perimeter/total variation. The majority of our theoretical results are independent of the distance used to define adversarial attacks.

Keywords: adversarial training; nonlocal perimeter; nonlocal total variation; existence of solutions; regularity.

1. Introduction

In this paper, we investigate the connection between adversarial training and regularized risk minimization in the context of non-parametric binary classification. *Adversarial training* problems, in their distributionally robust optimization (DRO) version, can be written mathematically as min-max problems of the form

$$\inf_{\theta\in\Theta} \sup_{\tilde{\mu}\,:\,G(\mu,\tilde{\mu})\leq\varepsilon} J(\tilde{\mu},\theta), \tag{1.1}$$

where in general θ denotes the parameters of a statistical model (the parameters of a neural network, a binary classifier, the parameters of a linear statistical model, etc.), and μ denotes a data distribution to be fit by the model. To fully specify a DRO problem, one also needs to introduce a notion of 'distance' G between data distributions that is employed to define a region of uncertainty around the

original data distribution μ and that can be interpreted as the possible set of actions of an adversary who may perturb μ. The value of $\varepsilon \geq 0$ describes the 'power' of the adversary and is often referred to as *adversarial budget*. The function $J(\tilde{\mu}, \theta)$ is a risk relative to a data distribution $\tilde{\mu}$ and some loss function underlying the statistical model. Problem (1.1) is a transparent mathematical way to explicitly enforce the robustness of models to data perturbations (at least of a certain type). Although the origins of this type of problem are now classical [60], recent influential research [32] has shown that neural networks can be greatly improved by using the DRO framework, and as a result, a renewed interest in this class of problems has been generated, see, e.g. the monograph [22] and the paper [42]. In the context of the binary classification problem described in detail throughout this paper, the works [44] and [55] have explored the game theoretic interpretation of (1.1) and the existence of Nash equilibria in parametric and non-parametric settings, respectively. Other very recent works, e.g. [3, 36, 50, 59], have expanded our theoretical understanding about adversarial training problems, providing results on existence of robust classifiers and reformulating adversarial training problems in new ways that are amenable to further analysis and alternative computation schemes.

By *regularization*, on the other hand, we mean an optimization problem of the form

$$\inf_{\theta \in \Theta} \hat{J}(\mu, \theta) + \lambda R(\theta), \tag{1.2}$$

where $\hat{J}(\mu, \theta)$ is a risk functional that here is taken with respect to the single data distribution μ, R is the regularization functional and $\lambda > 0$ is a positive parameter describing the strength of regularization. Regularization problems are fundamental in inverse problems [5, 56], image analysis [12, 16], statistics [57] and machine learning [35]; the previous list of references is of course non-exhaustive. In contrast to problem (1.1), the effect of explicit regularization on the robustness of models is less direct, but this is compensated by a richer structure that can be used to study the theoretical properties of their solutions more directly.

The connections between adversarial training and regularization have been intensely explored in recent years in the context of classical parametric learning settings; see [7, 8, 22] and references within. For example, when $\theta \in \Theta = \mathbb{R}^d$ represents the parameters of a linear regression model and the loss function for the model is the squared loss, the following identity holds:

$$\min_{\theta \in \Theta} \max_{G_p(\mu, \tilde{\mu}) \leq \varepsilon} \mathbb{E}_{(x,y)\sim\tilde{\mu}} \left[(y - \langle \theta, x \rangle)^2 \right] = \min_{\theta \in \Theta} \left\{ \sqrt{\mathbb{E}_{(x,y)\sim\mu} \left[(y - \langle \theta, x \rangle)^2 \right]} + \sqrt{\varepsilon}\, |\theta|_q \right\}^2, \tag{1.3}$$

where G_p is an optimal transport distance of the form

$$G_p(\mu, \tilde{\mu}) := \min_{\pi \in \Gamma(\mu, \tilde{\mu})} \iint_{\mathbb{R}^{d+1} \times \mathbb{R}^{d+1}} c_p((x, y), (\tilde{x}, \tilde{y}))\, \mathrm{d}\pi((x, y), (\tilde{x}, \tilde{y})).$$

The cost function c_p is defined by

$$c_p\big((x, y), (\tilde{x}, \tilde{y})\big) := \begin{cases} |x - \tilde{x}|_p & \text{if } y = \tilde{y}, \\ +\infty & \text{if } y \neq \tilde{y}, \end{cases}$$

where $|\cdot|_p$ is the ℓ^p norm in $\mathbb{R}^d$ for p satisfying $\frac{1}{p} + \frac{1}{q} = 1$. In the definition of $G_p(\mu, \tilde{\mu})$, the set $\Gamma(\mu, \tilde{\mu})$ represents the set of transportation plans (a.k.a. couplings) between μ and $\tilde{\mu}$, namely, the

set of probability measures on $\mathbb{R}^{d+1} \times \mathbb{R}^{d+1}$ with marginals given by μ and $\tilde{\mu}$. Notice that equation (1.3) reveals a direct equivalence between a family of DRO problems (1.1) and a family of regularized risk minimization problems (1.2) which includes the popular squared-root Lasso model from [4]. In particular, in this setting, $|\cdot|_q$ becomes the regularization term R, the risk functional is $\hat{J} = \sqrt{J}$, where J is the mean squared error, and $\lambda = \sqrt{\varepsilon}$. Through an equivalence such as (1.3), it is possible to motivate new ways of calibrating regularization parameters in models with a convex loss function (where first-order optimality conditions guarantee global optimality) as has been done in [8]. Beyond linear regression, the equivalence between adversarial training and regularization problems has also been studied in parametric binary classification settings such as logistic regression and SVMs (see [8]), as well as in distributionally robust grouped variable selection, and distributionally robust multi-output learning (see [22]).

In more general learning settings, it is often unknown whether there is a direct equivalence between (1.1) and a problem of the form (1.2) that is somewhat tractable both from a computational perspective as well as from a theoretical one. In such cases, an illuminating strategy that can be followed in order to gain insights into the regularization counterpart of (1.1) is to analyze the max part of the problem for small ε and identify its leading order behavior to construct approximating regularization terms. This is a strategy that has been followed in many works that study the robust training of neural networks, e.g. [11, 25–27, 41, 45, 51, 53, 62]. The structure of the resulting approximate regularization problems can be exploited to motivate algorithms and provide a better theoretical understanding of the process of training robust deep learning models (see [27]).

Having discussed some of the literature exploring the connection between adversarial training and regularization, we move on to discussing, first in simple terms, the content of this work. Through our theoretical results, this paper continues the investigation started in [59], this time providing a deeper structural connection between adversarial training in the non-parametric binary classification setting and regularized risk minimization problems. In particular, we show that the equivalence between adversarial training and regularized risk minimization problems goes beyond the aforementioned parametric settings without relying on approximations. Here, θ is substituted with A which from now on will be interpreted as an arbitrary (measurable) subset of the data space $\mathcal{X}$ (i.e. A specifies a binary classifier), while J is the risk associated to the 0-1 loss; the other elements in problem (1.1) will be specified in more detail in Section 1.1. We show that perimeter functionals penalizing the 'boundary' of a set arise naturally as regularizers for binary classification problems regardless of the feature space or distance used to define the adversarial budget. This provides a more direct means of studying the evolution and regularity properties of minimizers of the adversarial problem than the ones that were implied by the evolution equations studied in [59]. This approach also provides tangible prospects for the design of new algorithms for the training of robust classifiers and suggests which algorithms are more suitable for enforcing robustness relative to specific adversary's actions. Finally, through the connection between adversarial training and regularization, we will deduce a variety of theoretical properties of robust classifiers, including the existence of 'regular' solutions, where regularity is understood in a suitable technical sense. Regularity results like the ones we obtain in Theorem 3.25 are, to the best of our knowledge, the first of their kind in the context of adversarial training.

In summary, our work reveals a rich geometric structure of adversarial training problems which is conceptually appealing and that at the same time opens up new avenues for the theoretical study of adversarial training for general binary classification models. In the next subsections, we provide a more detailed discussion of our theoretical results and some of its conceptual consequences right after introducing the specific mathematical setup that we follow throughout the paper.

1.1 *Setup*

Let $(\mathcal{X}, \mathsf{d})$ be a separable metric space representing the space of features of data points, and let $\mathfrak{B}(\mathcal{X})$ be its associated Borel σ-algebra. In most applications, $\mathcal{X}$ is a finite dimensional vector space, e.g. $\mathbb{R}^d$ for some $d \in \mathbb{N}$, and later, we will assume a certain, essentially finite dimensional, structure for some of our statements. We are given a probability measure $\mu \in \mathcal{P}(\mathcal{X} \times \{0,1\})$ describing the distribution of training pairs $(x, y) \in \mathcal{X} \times \{0,1\}$. Letting $\pi_1 : \mathcal{X} \times \{0,1\}$, $(x, y) \mapsto x$ be the projection onto the first factor of $\mathcal{X} \times \{0,1\}$, the first marginal of μ is denoted by $\rho := \pi_{1\sharp}\mu \in \mathcal{P}(\mathcal{X})$ and represents the distribution of input data. Here, $\pi_{1\sharp}\mu$ denotes the push-forward measure, whose definition we give in Appendix A. We decompose the data distribution as $\rho = w_0\rho_0 + w_1\rho_1$, where $w_i = \mu(\mathcal{X} \times \{i\})$ and $\rho_i \in \mathcal{P}(\mathcal{X})$ denote the conditional distributions

$$\rho_i(A) := \frac{\mu\,(A \times \{i\})}{w_i}, \quad i \in \{0,1\},\ A \in \mathfrak{B}(\mathcal{X}). \tag{1.4}$$

Throughout this paper, we make the assumption that all measures are Radon measures on $\mathcal{X}$. In the following example, we lay out two canonical situations which are highly relevant in machine learning.

Example 1.1. (Absolutely continuous and empirical data distribution). We let $\mathcal{X} = \mathbb{R}^d$, equipped with an arbitrary ℓ^p-metric for $p \in [1, \infty]$, i.e., $\mathsf{d}(x_1, x_2) := |x_1 - x_2|_p$. If we know the true distribution ρ of the data, and this distributions is assumed to be absolutely continuous with respect to the Lebesgue measure, we can work with ρ directly. If we are only given a finite number of data points $\{x_i\}_{i=1}^N$, we can work with the empirical measure $\rho = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$. Both measures are Radon measures, and the main results of the paper apply in both settings.

In binary classification, we seek a set $A \in \mathfrak{B}(\mathcal{X})$ and its induced classifier

$$\begin{aligned} x \in A \ &:\Longleftrightarrow\ x \text{ is assigned label 1,} \\ x \in A^c &:\Longleftrightarrow\ x \text{ is assigned label 0.} \end{aligned}$$

The most natural approach to constructing such a classifier is to minimize the empirical risk

$$\inf_{A \in \mathfrak{B}(\mathcal{X})} \mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A(x) - y\right|\right]. \tag{1.5}$$

A minimizer of this problem is known as a Bayes classifier relative to μ.

Remark 1.2. (0-1-loss). Note that introducing the 0-1-loss function $\ell(\hat{y}, y) = 0$ if $\hat{y} = y$ and $\ell(\hat{y}, y) = 1$ if $\hat{y} \neq y$, one can equivalently express (1.5) as $\inf_{A\in\mathfrak{B}(\mathcal{X})} \mathbb{E}_{(x,y)\sim\mu}[\ell(1_A(x), y)]$.

Applying the law of total expectation (or equivalently disintegrating the measure μ) one obtains that (1.5) coincides with the following geometric problem:

$$\inf_{A \in \mathfrak{B}(\mathcal{X})} \int_A w_0\,\mathrm{d}\rho_0 + \int_{A^c} w_1\,\mathrm{d}\rho_1. \tag{1.6}$$

Problem (1.6) forces the set A to be concentrated in places where the measure $w_1\rho_1$ is larger than $w_0\rho_0$. Defining the signed measure $\sigma := w_1\rho_1 - w_0\rho_0$, one can take a Hahn decomposition of $\mathcal{X}$ into

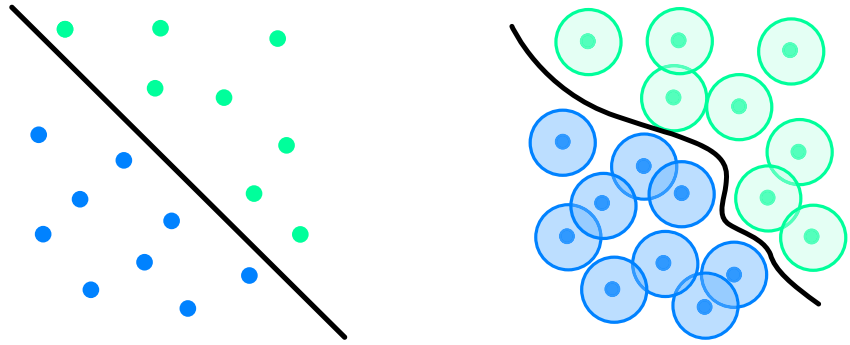


FIG. 1. Picture in the spirit of [42, Fig. 3]. An affine Bayes classifier on the left is not robust with respect to adversarial attacks. The robust classifier on the right has a smaller *nonlocal* perimeter than the Bayes classifier. The adversarial attacks play an important role in the geometry of the robust classifier and will define a particular form of nonlocal perimeter regularization.

$\mathcal{X} = P \uplus N$, where P is a positive set and N is a negative set under σ (see Appendix A for the definition of a Hahn decomposition). We then let $A := P$ and deduce that such a set A is a Bayes classifier, i.e. a minimizer of problem (1.6). Notably, the Hahn-decomposition is not unique and so neither is the Bayes classifier A. Furthermore, there is no control over the set, where $w_0\rho_0 = w_1\rho_1$. Those points might be arbitrarily assigned to either of the classes without affecting the objective functional; this is a potential source of non-robustness in classification.

Throughout the paper, we focus our attention on the following adversarial training problem for robust binary classification, see Fig. 1 for an illustration

$$\inf_{A\in\mathfrak{B}(\mathcal{X})} \mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B_\varepsilon(x)} \left|1_A(\tilde{x}) - y\right|\right]. \tag{1.7}$$

The model allows an adversary to choose the worst possible point in an open ε-ball $B_\varepsilon(x) := \{\tilde{x} \in \mathcal{X} : \mathsf{d}(x,\tilde{x}) < \varepsilon\}$ (relative to the metric d) around x to corrupt the classification. We emphasize that we *do not* use the essential supremum with respect to some measure but rather the actual supremum which potentially makes the adversarial attack much stronger. However, under mild assumptions on the space $\mathcal{X}$ and the measure ρ, it is possible to draw a connection between (1.7) and the following problem:

$$\inf_{A\in\mathfrak{B}(\mathcal{X})} \mathbb{E}_{(x,y)\sim\mu}\left[\nu\text{-ess}\sup_{\tilde{x}\in B_\varepsilon(x)} \left|1_A(\tilde{x}) - y\right|\right], \tag{1.8}$$

where ν is a suitably chosen reference measure. This problem has favorable functional analytic properties and we will use it as intermediate step to construct solutions of the original problem (1.7), as well as to analyze the structure of the set of solutions of (1.7).

Before proceeding to an informal presentation of our main results, we emphasize that, in contrast to some papers in the literature, here we consider *open* balls $B_\varepsilon(x)$ to describe the set of possible attacks available to the adversary around the point x. By making this modelling choice, we can simplify some technical steps in our analysis (e.g. see Remark 3.9 and Appendix B.1) and avoid measurability issues

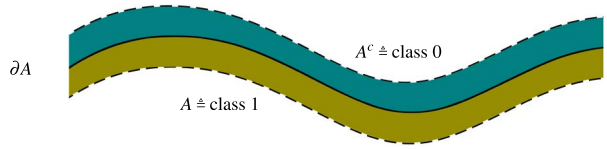


FIG. 2. Illustration of the 'perimeter' defined in (1.10). The blue strip outside A is measured with ρ_0. The olive strip inside A is measured with ρ_1. The sum of these two quantities being small means that the indicator of A is an adversarially robust classifier.

that may arise when working with closed balls (see [55] for a discussion on the measurability issue and contrast it with Remark 2.3).

1.2 *Informal main results and discussion*

Our first main result, at this stage stated informally, is a reformulation of the adversarial training problem (1.7) in terms of a variational regularization problem.

THEOREM. The objective in the adversarial training problem (1.7) can be rewritten as

$$\mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B_\varepsilon(x)} \left|1_A(\tilde{x}) - y\right|\right] = \mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A(x) - y\right|\right] + \varepsilon\, \widetilde{\mathrm{Per}}_\varepsilon(A;\mu), \tag{1.9}$$

where $\widetilde{\mathrm{Per}}_\varepsilon(A;\mu)$ is a *nonlocal* and *weighted* perimeter of A, defined as

$$\widetilde{\mathrm{Per}}_\varepsilon(A;\mu) = \frac{w_0}{\varepsilon}\rho_0(\{x \in A^c \,:\, \mathsf{dist}(x,A) < \varepsilon\}) + \frac{w_1}{\varepsilon}\rho_1(\{x \in A \,:\, \mathsf{dist}(x,A^c) < \varepsilon\}), \tag{1.10}$$

see Fig. 2 for a color-coded illustration.

The functional $\widetilde{\mathrm{Per}}_\varepsilon$ can be called a type of 'perimeter' since it is a non-negative functional over sets with the important submodularity property

$$\widetilde{\mathrm{Per}}_\varepsilon(A \cup B;\mu) + \widetilde{\mathrm{Per}}_\varepsilon(A \cap B;\mu) \le \widetilde{\mathrm{Per}}_\varepsilon(A;\mu) + \widetilde{\mathrm{Per}}_\varepsilon(B;\mu), \quad \forall A, B \in \mathfrak{B}(\mathcal{X}).$$

Submodular functionals over sets typically induce convex functionals over functions (referred to as a total variation), defined through the coarea formula

$$\widetilde{\mathrm{TV}}_\varepsilon(u;\mu) := \int_{-\infty}^{\infty} \widetilde{\mathrm{Per}}_\varepsilon(\{u \ge t\};\mu)\,\mathrm{d}t.$$

We will show that the so defined total variation takes the form

$$\widetilde{\mathrm{TV}}_\varepsilon(u;\mu) = \frac{w_0}{\varepsilon}\int_{\mathcal{X}} \sup_{\tilde{x}\in B_\varepsilon(x)} u(\tilde{x}) - u(x)\,\mathrm{d}\rho_0(x) + \frac{w_1}{\varepsilon}\int_{\mathcal{X}} u(x) - \inf_{\tilde{x}\in B_\varepsilon(x)} u(\tilde{x})\,\mathrm{d}\rho_1(x). \tag{1.11}$$

By our notation, we emphasize that both the perimeter and the total variation depend on the data distribution μ through w_i, ρ_i and not just through ρ, as will be detailed in the course of the paper. Hence, as opposed to standard (nonlocal) perimeters and total variations, they constitute a family of data-driven

regularizers. Such regularizers, typically learned in a supervised manner, have recently been shown to be superior over model-based regularizers for certain tasks in medical imaging, see [46].

It turns out that using the $\widetilde{\mathrm{TV}}_\varepsilon$ functional, we can define an *exact* convex relaxation for the problem (1.7)

THEOREM. The variational problem

$$\inf_{u:\mathcal{X}\to[0,1]} \mathbb{E}_{(x,y)\sim\mu}\left[|u(x)-y|\right] + \varepsilon\,\widetilde{\mathrm{TV}}_\varepsilon(u;\mu) \tag{1.12}$$

is an exact convex relaxation of problem (1.7). In particular, any solution to problem (1.7) is also a solution to (1.12), and conversely, for any solution u of problem (1.12), we can obtain a solution to problem (1.7) by considering level sets of u.

We move on to study the existence of solutions to problem (1.7).

THEOREM. (informal). Under some technical conditions on the metric space $\mathcal{X}$ and the measure ρ, the adversarial training problem (1.7) admits a solution $A \in \mathfrak{B}(\mathcal{X})$.

Our existence proof is technical and is based on the lifting of the variational problem (1.7) to a problem of the form (1.8). This problem admits an application of the direct method of the calculus of variations after establishing lower semicontinuity and compactness in a suitable weak-* Banach space topology, see Appendix A for a definition of the weak-* topology. In the course of this, we will introduce well-defined versions of $\widetilde{\mathrm{Per}}_\varepsilon$ and $\widetilde{\mathrm{TV}}_\varepsilon$, which will not carry the tilde anymore, and study their associated variational problems. We discuss how to build solutions to the original problem (1.7) from solutions to the modified problems.

REMARK 1.3. (Relation to previous results). Existence of solutions to other adversarial training problems has also been obtained recently in the work [3]. The existence results in [3] and ours are highly complementary to each other and in what follows we highlight the differences in their settings, which are apparent in at least three ways: first, the adversarial model in [3] is defined in terms of closed balls rather than open balls as done here; existence of solutions in the open ball model was left as an open question. Second, the collection of subsets A of $\mathcal{X}$ over which the optimization takes place in [3] is the so called universal σ-algebra, which is larger than the Borel σ-algebra considered here. For the adversarial model with closed balls, it is essential to use the universal σ-algebra in order to assure the measurability of the adversarial loss function, which we get for free in our open ball model. Naturally, lower semicontinuity is less of an issue in [3], whereas in our proof, we rely on relaxation methods and on the explicit construction of representatives. Lastly, we highlight that our setting is very general since we work on a metric measure space, whereas the results in [3] hold in the setting of norm balls in Euclidean space.

It is also worth highlighting that objective functions of type $L^1 + \mathrm{TV}$ and their relation to perimeter regularization have been extensively studied in the mathematical imaging community [16, 21, 23, 24, 63]. Further background on total variation methods in imaging is provided in [12, 16].

After establishing existence of solutions to (1.7), we proceed to studying their properties. In particular, we exploit the underlying convexity made manifest by our theorems and deduce a series of strong implications on the geometry and regularity of the family of solutions to the adversarial training problem (1.7). As a first step, we prove that solutions are closed under intersections and unions. From this, we will be able to prove the following.

THEOREM. (informal). There exist (unique) minimal and maximal solutions to (1.7) in the sense of set inclusion.

It is then possible to show that maximal and minimal solutions satisfy, respectively, inner and outer regularity conditions (in a suitable sense discussed in detail throughout the paper), providing in this way the first results on regularity of (certain) solutions to (1.7). We investigate the regularity of solutions further and establish Hölder regularity results like the following (see Appendix A for definitions).

THEOREM. (informal). Let $(\mathcal{X}, \mathsf{d})$ be $\mathbb{R}^d$ with the Euclidean distance. For any $\varepsilon > 0$, there exists a solution to the problem (1.7) whose boundary is locally the graph of a $C^{1,1/3}$ function.

Although stated for the Euclidean setting only, we highlight that similar results can be proved in more general settings provided that one adjusts the interpretation of regularity of solutions to non-Euclidean contexts. A more detailed investigation of this will be the topic of follow-up work. It is also important to reiterate that our results apply to a general measure μ regardless of whether it has densities with respect to Lebesgue measure or if it is an empirical measure. In particular, from our results, we can conclude that the presence of the adversary always enforces the regularization of decision boundaries, even when the original unrobust problem does not possess regular solutions (i.e. when the Bayes classifiers are not regular). We remark that we do not claim any sharpness in our regularity results. However, in general, one should not expect better regularity than $C^{1,1}$ (in the Euclidean setting) based on the discussion that we present in Section 3.6 and on the results from [39].

Finally, we remark that our results suggest that one should use algorithms for adversarial training that are based on training parametric models that are able to produce or approximate regular classifiers. Some examples of these models are suggested by recent results in the literature of approximation theory; these results state that it is possible to approximate characteristic functions of regular sets with neural networks whose size is determined by the level of regularity of the target set, see [49]. We believe that our regularity results can indeed inform new implementations of adversarial training, but there are still several points to be resolved before being able to carry out an actual algorithmic implementation. Moreover, since the notion of regularity depends on the distance function used to define the action space of the adversary, one should naturally adapt algorithms to produce robust classifiers of the specified type. The above discussion will be expanded in future work.

In addition to the adversarial model (1.7), we discuss other adversarial models that admit a representation of the form $L^1 +$ (nonlocal) TV. From this, we will be able to conclude that perimeter functionals penalizing the boundaries of sets indeed arise naturally as regularizers for binary classification problems. This fact can be interpreted conversely: it is possible to give a game theoretic interpretation for a class of variational problems that involve the use of (nonlocal) total variation (including those that have been used in graph-based learning for classification [28]). This work can then be naturally related to a collection of works that provide game theoretical interpretations of variational problems. For example, [13] and [48] connect fractional Dirichlet energies with a two-player game. Moreover, [37] connects mean curvature flow with a different two-player game. While our energies do not directly coincide with the ones in those papers, they are similar in form.

1.3 *Outline*

The rest of the paper is organized as follows. In Section 2, we discuss different reformulations of the adversarial training problem (1.7). First, in Section 2.1, we relax the problem in a suitable way in order to make it amenable to functional analytic treatment; this reformulation will be crucial for our latter exploration on existence of solutions to (1.7) and the study of some of their properties.

In Section 2.2, we discuss the reformulation of (1.7) as the regularized risk minimization that has already been introduced in Section 1.2, cf. (1.9).

Section 3 is devoted to the study of properties of the regularization reformulation of (1.7). We define suitable relaxations of the functionals $\widetilde{\mathrm{Per}}_\varepsilon$ and $\widetilde{\mathrm{TV}}_\varepsilon$ appearing in (1.9) and establish key properties including submodularity, convexity and lower semi-continuity with respect to suitable topologies. With these properties at hand, we show existence of solutions to problem (1.7) in Section 3.4.

In Section 3.5, we study maximal and minimal solutions, and in Section 3.6, we investigate regularity.

In Section 4, we explain how to generalize our insights to regression tasks and other adversarial training models that give rise to perimeter minimization problems with different perimeter functionals. In particular, we recover data-driven regularizers as well as statistically robust interpretations to regularization approaches used in graph-based learning.

We wrap up the paper in Section 5 where we present further discussion on the implications of our work and provide some directions for future research.

Technical definitions, some proofs and further remarks on the advantage of using open balls are given in the appendix.

2. Reformulations of adversarial training

2.1 *Relaxation in quotient σ-algebra*

To be able to prove existence of minimizers for (1.7), we have to relax it to make it amenable to functional analytic treatment. Note that, because of the presence of the non-essential supremum in (1.7), two sets A and A' whose symmetric difference

$$A \triangle A' := (A \setminus A') \cup (A' \setminus A) \tag{2.1}$$

satisfies $\nu(A \triangle A') = 0$ for some reference measure ν do not have to have the same value of the objective function, in general. This is a major difference to unregularized problem (1.5) and will cause problems, for instance, when proving existence of minimizers.

To fix this, we define the set

$$\mathfrak{N}_\nu := \{A \in \mathfrak{B}(\mathcal{X}) \,:\, \nu(A) = 0\}, \tag{2.2}$$

where ν is an arbitrary reference measure on $\mathcal{X}$, to be specified later. The set $\mathfrak{N}_\nu$ is a two-sided ideal in the σ-algebra $\mathfrak{B}(\mathcal{X})$, interpreted as ring with addition $\triangle$ and multiplication $\cap$. This allows us to define the quotient σ-algebra

$$\mathfrak{B}_\nu(\mathcal{X}) := \mathfrak{B}(\mathcal{X}) \Big/ \mathfrak{N}_\nu \tag{2.3}$$

with the equivalence relation $\sim_\nu$, defined by

$$A \sim_\nu B :\iff A \triangle B \in \mathfrak{N}_\nu \iff \nu(A \triangle B) = 0. \tag{2.4}$$

The function $d_\nu(A, B) := \nu(A \triangle B)$ is non-negative, symmetric and sub-additive and hence defines a pseudo-metric on $\mathfrak{B}(\mathcal{X})$. This function is also zero if and only if $A \sim_\nu B$, and hence, it is a metric on the

quotient σ-algebra $\mathfrak{B}_\nu(\mathcal{X})$. In some sources, this metric is called the Fréchet–Nikodým pseudo-metric, see, e.g. Section 1.12 in [9].

The following proposition states that the minimization in (1.7) can be rewritten as the minimization of some sort of quotient norm on the quotient σ-algebra $\mathfrak{B}_\nu(\mathcal{X})$. Interestingly, the choice of ν does not yet matter here.

PROPOSITION 2.1. For any Borel measure ν on $\mathcal{X}$, it holds that

$$(1.7) = \inf_{A\in\mathfrak{B}(\mathcal{X})} \inf_{\substack{B\in\mathfrak{B}(\mathcal{X})\\ A\sim_\nu B}} \mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B_\varepsilon(x)} \left|1_B(\tilde{x}) - y\right|\right]. \tag{2.5}$$

REMARK 2.2. (Similarity to quotient norms). The reason why we connect this reformulation with the quotient σ-algebra is that the objective function in (2.5) has strong similarities with the quotient norm on a quotient Banach space X/N, which is given by

$$\|x\|_{X/N} := \inf_{\substack{y\in X\\ y-x\in N}} \|y\|_X, \quad x \in X/N.$$

Proof. We have to prove the equality

$$\inf_{A\in\mathfrak{B}(\mathcal{X})} \mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B_\varepsilon(x)} \left|1_A(\tilde{x}) - y\right|\right] = \inf_{A\in\mathfrak{B}(\mathcal{X})} \inf_{\substack{B\in\mathfrak{B}(\mathcal{X})\\ A\sim_\nu B}} \mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B_\varepsilon(x)} \left|1_B(\tilde{x}) - y\right|\right].$$

First, choosing $B = A$, which obviously fulfills $A \sim_\nu B$, we obtain the inequality $\geq$. Second, omitting the constraint $A \sim_\nu B$ yields the inequality $\leq$. □

REMARK 2.3. In the definition of the adversarial problem (1.7) and throughout the rest of the paper, we will be working with quantities like $\sup_{\tilde{x}\in B_\varepsilon(x)} 1_A$ for a Borel measurable set A and $\sup_{\tilde{x}\in B_\varepsilon(x)} u$ for a Borel measurable function u. We remark that the resulting sets/functions are Borel measurable. Indeed, the function $x \mapsto \sup_{\tilde{x}\in B_\varepsilon(x)} 1_A$ is nothing but the indicator function of the set $\bigcup_{x\in A} B_\varepsilon(x)$ which is Borel measurable since it is an open set. Likewise, the function $x \mapsto \tilde{u}(x) := \sup_{\tilde{x}\in B_\varepsilon(x)} u$ is measurable because the sets $\{\tilde{u} > t\}$ are open sets.

An alternative way of avoiding ambiguities arising from equivalent sets with respect to ν is to consider the essential version of the adversarial problem given by (1.8), where the adversarial attack is performed using the essential supremum of the measure ν. This problem can be fundamentally different to our problem (1.7) or the relaxed one (2.5) since for example, in the case $\nu = \rho$, the attack can only be performed within the support of the given data distribution which is much weaker than (1.7). Still, in Section 3, we shall construct a measure ν such that the problems (1.7) and (1.8) *do* coincide, a property we will exploit later for proving existence of solutions to the original adversarial problem.

EXAMPLE 2.4. Consider the simple situation with the measure $\rho = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ on $\mathcal{X} = \mathbb{R}$ and $\nu = \rho$. The labels are set to be equal to zero on the left axis and one on the right one. Then, it holds

$$\mathbb{E}_{(x,y)\sim\mu}\left[\sup_{(x-\varepsilon,x+\varepsilon)} \left|1_A(x) - y\right|\right] = \frac{1}{2}\sup_{(-1-\varepsilon,-1+\varepsilon)} 1_A + \frac{1}{2}\sup_{(1-\varepsilon,1+\varepsilon)} 1_{A^c}.$$

Let us assume that $1 < \varepsilon < 2$. In this case, the intervals $(-1-\varepsilon, -1+\varepsilon)$ and $(1-\varepsilon, 1+\varepsilon)$ overlap. Therefore, for any choice of $A \in \mathfrak{B}(\mathbb{R})$, either A or A^c intersect both intervals. This implies that the optimal adversarial risk is $\geq \frac{1}{2}$. Furthermore, choosing $A = \{1\}$, we find the risk equals $\frac{1}{2}$.

For comparison, the objective of the quotient problem (2.5) is given by

$$\inf_{\substack{B\in\mathfrak{B}(\mathcal{X})\\ \rho(A\triangle B)=0}} \mathbb{E}_{(x,y)\sim\mu}\left[\sup_{(x-\varepsilon,x+\varepsilon)} \left|1_B - y\right|\right] = \inf_{\substack{B\in\mathfrak{B}(\mathcal{X})\\ \rho(A\triangle B)=0}} \frac{1}{2}\sup_{(-1-\varepsilon,-1+\varepsilon)} 1_B + \frac{1}{2}\sup_{(1-\varepsilon,1+\varepsilon)} 1_{B^c}$$

and, arguing as before, any choice of B leads to this term being $\geq \frac{1}{2}$ independently of A.

On the other hand, the objective in (1.8) for $\nu := \rho$ and $0 < \varepsilon < 2$ is

$$\mathbb{E}_{(x,y)\sim\mu}\left[\rho\text{-ess}\sup_{B_\varepsilon(x)} \left|1_A(x) - y\right|\right] = \frac{1}{2}1_A(-1) + \frac{1}{2}1_{A^c}(1),$$

which does not even depend on ε. This is due to the fact that the ρ-ess sup prevents the adversary from leaving the set of data points. For instance, the half axes $x \geq \alpha$ with $-1 < \alpha < 1$ have risk 0 and thus are optimal.

2.2 *Nonlocal variational regularization problem*

We now show how to express the adversarial training problem (1.7) as a variational regularization problem in the form of (1.2). More precisely, we show that it can be written as L^1 + TV-type problem. This class of problems has been intensively studied in the context of image processing, following the seminal paper [21]. The model there was related to a geometric problem involving the Lebesgue measure $\mathcal{L}^d(\cdot)$ and the standard perimeter functional $\mathrm{Per}(\cdot)$, namely

$$\min_{A\in\mathfrak{B}(\mathcal{X})} \mathcal{L}^d(A\triangle\Omega) + \lambda\,\mathrm{Per}(A). \tag{2.6}$$

This functional was shown to exhibit a range of different behaviors in terms of the regularization parameter λ.

In our context, we will show that the adversarial problem (1.7) can be interpreted analogously, with the modification that we use a weighted volume and a weighted and nonlocal perimeter, see Remark 2.9 below. Let us therefore first introduce the set function $\widetilde{\mathrm{Per}}_\varepsilon(\cdot;\mu) : \mathfrak{B}(\mathcal{X}) \to [0,+\infty]$ for $\varepsilon > 0$ which we refer to as nonlocal pre-perimeter and which is defined as

$$\widetilde{\mathrm{Per}}_\varepsilon(A;\mu) := \frac{w_0}{\varepsilon}\int_{\mathcal{X}} \sup_{\tilde{x}\in B_\varepsilon(x)} 1_A(\tilde{x}) - 1_A(x)\,\mathrm{d}\rho_0(x) + \frac{w_1}{\varepsilon}\int_{\mathcal{X}} 1_A(x) - \inf_{\tilde{x}\in B_\varepsilon(x)} 1_A(\tilde{x})\,\mathrm{d}\rho_1(x). \tag{2.7}$$

Here, the dependency on the data distribution μ is captured by the presence of the conditional distributions ρ_i and the class probabilities w_i for $i \in \{0,1\}$. Here, the tilde serves as a reminder that we are using supremum and infimum as opposed to their ν-essential forms. To see that $\widetilde{\mathrm{Per}}_\varepsilon(A;\mu)$ has units of a perimeter, we rewrite it as follows:

$$\widetilde{\mathrm{Per}}_\varepsilon(A;\mu) = \frac{w_0}{\varepsilon}\rho_0(\{x \in A^c \,:\, \mathsf{dist}(x,A) < \varepsilon\}) + \frac{w_1}{\varepsilon}\rho_1(\{x \in \mathsf{dist}(x,A^c) < \varepsilon\}), \tag{2.8}$$

where the distance of a point $x \in \mathcal{X}$ to a set $A \subseteq \mathcal{X}$ is defined as $\mathsf{dist}(x,A) := \inf_{\tilde{x}\in A} \mathsf{d}(x,\tilde{x})$. The quantity (2.8) is a weighted and nonlocal Minkowski content [14, 15] of the 'thickened boundary' $\partial^\varepsilon A := \{x \in \mathcal{X} \,:\, \mathsf{dist}(x,\partial A) < \varepsilon\}$, cf. Fig. 2 in Section 1.2. For sufficiently smooth sets and measures $\rho_{0/1}$, and for small ε, one expects [1] that $\widetilde{\mathrm{Per}}_\varepsilon(A;\mu)$ behaves like a weighted perimeter of A, see [17] for similar results.

Importantly, for two sets $A, B \in \mathfrak{B}(\mathcal{X})$ which differ only by a nullset with respect to some reference measure ν the associated pre-perimeters will generally be different. Therefore, using the technique from Section 2.1, we define the nonlocal perimeter with respect to ν as

$$\nu\text{-}\mathrm{Per}_\varepsilon(A;\mu) := \inf_{\substack{B\in\mathfrak{B}(\mathcal{X})\\ A\sim_\nu B}} \widetilde{\mathrm{Per}}_\varepsilon(B;\mu). \tag{2.9}$$

This way of defining a nonlocal and weighted perimeter generalizes approaches from [14, 15], which deal with the case of the Lebesgue measure.

REMARK 2.5. The nonlocal perimeter ν-$\mathrm{Per}_\varepsilon(\cdot;\mu)$ in (2.9) is a generalization of the nonlocal perimeter studied in [14; 15; 19; 20] which can be recovered by setting $\mathcal{X} = \mathbb{R}^d$ and by replacing $w_0\rho_0$ and $w_1\rho_1$ by the Lebesgue measure $\mathcal{L}^d$ and choosing $\nu := \mathcal{L}^d$. Our results from Section 3, in particular Proposition 3.7, show that (2.9) becomes

$$\mathrm{Per}_\varepsilon(A) := \frac{1}{2\varepsilon}\int_{\mathbb{R}^d} \mathrm{ess\,osc}_{B_\varepsilon(\cdot)}(1_A)\,\mathrm{d}x, \tag{2.10}$$

where ess osc = ess sup − ess inf is the essential oscillation with respect to the Lebesgue measure.

REMARK 2.6. (Asymmetry). It is obvious that the nonlocal perimeter (2.10) from [20] satisfies $\mathrm{Per}_\varepsilon(A^c) = \mathrm{Per}_\varepsilon(A)$ and the same is true for the usual local perimeter. For our perimeter (2.9), this is not the case if $w_0\rho_0 \neq w_1\rho_1$.

Let us now reformulate the adversarial training problem as a regularization problem with respect to the nonlocal perimeter (2.9). Our central observation is that the adversarial risk in (1.7) can be decomposed into an unregularized risk and the pre-perimeter. Then, using Proposition 2.1, we will rewrite (1.7) as a variational regularization problem for the perimeter.

PROPOSITION 2.7. For any Borel set $B \in \mathfrak{B}(\mathcal{X})$, it holds

$$\mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B_\varepsilon(x)} \left|1_B(\tilde{x}) - y\right|\right] = \mathbb{E}_{(x,y)\sim\mu}\left[\left|1_B(x) - y\right|\right] + \varepsilon\,\widetilde{\mathrm{Per}}_\varepsilon(B;\mu). \tag{2.11}$$

Proof. Disintegrating μ and doing elementary calculations yields

$$\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B_\varepsilon(x)}\left|1_B(\tilde{x})-y\right|\right]-\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_B(x)-y\right|\right]\\
&=\iint_{\mathcal{X}\times\{0,1\}}\sup_{\tilde{x}\in B_\varepsilon(x)}\left|1_B(\tilde{x})-y\right|\mathrm{d}\mu(x,y)-\iint_{\mathcal{X}\times\{0,1\}}\left|1_B(x)-y\right|\mathrm{d}\mu(x,y)\\
&=w_0\int_{\mathcal{X}}\sup_{B_\varepsilon(\cdot)}1_B\,\mathrm{d}\rho_0+w_1\int_{\mathcal{X}}\sup_{B_\varepsilon(\cdot)}1_{B^c}\,\mathrm{d}\rho_1-w_0\int_{\mathcal{X}}1_B\,\mathrm{d}\rho_0-w_1\int_{\mathcal{X}}1_{B^c}\,\mathrm{d}\rho_1\\
&=w_0\int_{\mathcal{X}}\sup_{B_\varepsilon(\cdot)}1_B-1_B\,\mathrm{d}\rho_0+w_1\int_{\mathcal{X}}1_B-\inf_{B_\varepsilon(\cdot)}1_B\,\mathrm{d}\rho_1\\
&=\varepsilon\,\widetilde{\mathrm{Per}}_\varepsilon(B;\mu).
\end{aligned}$$

□

Now we can finally state the equivalence of the adversarial training problem (1.7) and the variational regularization problem involving the nonlocal perimeter ν-$\mathrm{Per}_\varepsilon(\cdot;\mu)$. For this, we have to choose the measure ν in the definition of the perimeter (2.9) such that ρ is absolutely continuous with respect to the reference measure ν, written $\rho \ll \nu$.

Proposition 2.8. (Perimeter-regularized problem). Let ν be a measure on $\mathcal{X}$ such that $\rho \ll \nu$. Then, it holds that

$$(1.7)=\inf_{A\in\mathfrak{B}(\mathcal{X})}\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A(x)-y\right|\right]+\varepsilon\,\nu\text{-}\mathrm{Per}_\varepsilon(A;\mu). \tag{2.12}$$

Remark 2.9. (Geometric problem). Note that if the measures ρ_0 and ρ_1 have non-overlapping support, (2.12) can indeed be brought into the form of the geometric problem (2.6) which generalizes the problem studied in [21]. For this, we assume that there exists $\Omega \subseteq \mathcal{X}$ such that $\mathrm{supp}\rho_1 \subseteq \Omega \subseteq (\mathrm{supp}\rho_0)^c$. Then, the first term in (2.12) equals

$$\begin{aligned}
\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A(x)-y\right|\right]&=w_0\int_{\mathcal{X}}1_A\,\mathrm{d}\rho_0+w_1\int_{\mathcal{X}}1_{A^c}\,\mathrm{d}\rho_1=w_0\rho_0(A)+w_1\rho_1(A^c)\\
&=w_0\rho_0(A\cap\Omega^c)+w_1\rho_1(A^c\cap\Omega)=w_0\rho_0(A\setminus\Omega)+w_1\rho_1(\Omega\setminus A)\\
&=\rho(A\setminus\Omega)+\rho(\Omega\setminus A)-w_1\rho_1(A\setminus\Omega)-w_0\rho_0(\Omega\setminus A)\\
&=\rho(A\setminus\Omega)+\rho(\Omega\setminus A)-w_1\rho_1(\Omega^c\setminus A^c)-w_0\rho_0(\Omega\setminus A)\\
&=\rho((A\setminus\Omega)\cup(\Omega\setminus A))=\rho(A\triangle\Omega).
\end{aligned}$$

This implies that (2.12) equals the geometric problem

$$\inf_{A\in\mathfrak{B}(\mathcal{X})}\rho(A\triangle\Omega)+\varepsilon\,\nu\text{-}\mathrm{Per}_\varepsilon(A;\mu). \tag{2.13}$$

Proof. Fixing $A \in \mathfrak{B}(\mathcal{X})$ and taking the infimum over sets $B \in \mathfrak{B}(\mathcal{X})$ with $A \sim_\nu B$, we get from Proposition 2.7 that

$$\inf_{\substack{B\in\mathfrak{B}(\mathcal{X})\\ A\sim_\nu B}} \mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B_\varepsilon(x)} \left|1_B(\tilde{x}) - y\right|\right] = \inf_{\substack{B\in\mathfrak{B}(\mathcal{X})\\ A\sim_\nu B}} \mathbb{E}_{(x,y)\sim\mu}\left[\left|1_B(x) - y\right|\right] + \varepsilon\, \widetilde{\mathrm{Per}}_\varepsilon(B;\mu).$$

Now we note that for $A \sim_\nu B$, it holds

$$\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_B(x) - y\right|\right] = w_0 \int_B \mathrm{d}\rho_0 + w_1 \int_{B^c} \mathrm{d}\rho_1 = \mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A(x) - y\right|\right],$$

since $A \sim_\nu B$ implies $\nu(A\triangle B) = 0$ which by the absolute continuity implies $\rho(A\triangle B) = 0$. Hence, we obtain

$$\begin{aligned}\inf_{\substack{B\in\mathfrak{B}(\mathcal{X})\\ A\sim_\nu B}} \mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B_\varepsilon(x)} \left|1_B(\tilde{x}) - y\right|\right] &= \mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A(x) - y\right|\right] + \varepsilon \inf_{\substack{B\in\mathfrak{B}(\mathcal{X})\\ A\sim_\nu B}} \widetilde{\mathrm{Per}}_\varepsilon(B;\mu)\\ &= \mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A(x) - y\right|\right] + \varepsilon\, \nu\text{-}\mathrm{Per}_\varepsilon(A;\mu).\end{aligned}$$

Finally, using Proposition 2.1 concludes the proof. □

3. Analysis of the adversarial training problem

In the previous section, we have shown that the adversarial training problem (1.7) is equivalent to the variational regularization problem (2.12) involving a nonlocal perimeter term. Problems such as (2.12) are very well understood in the context of inverse problems [5]. We will use the structure of the objective in problem (2.12) to make strong mathematical statements about our original adversarial training problem under very general conditions on the space $(\mathcal{X}, \mathsf{d})$. The aim of this section is then to use the insights stemming from the reformulation in terms of perimeter in order to perform a rigorous analysis on the adversarial problem (1.7), focusing on proving existence of solutions and studying their properties. In particular, we will define convenient notions of uniqueness of solutions and show the existence of 'regular' solutions, at least in the Euclidean setting.

For this, we first introduce a nonlocal total variation which is associated with the perimeter (2.9) and that turns out to be useful for proving existence. Then, we prove important properties of the perimeter and the total variation related to convexity and lower semicontinuity. Here, the key ingredient is to construct suitable representatives which attain the infimum in the definition of the perimeter ν-$\mathrm{Per}_\varepsilon(\cdot;\mu)$. For this, we will have to focus on reference measures ν which satisfy a certain geometric assumption. Finally, we can use these insights to prove existence of solutions to (1.7) and study their geometric properties. Due to the lack of uniqueness of minimizers, we will investigate minimal and maximal solutions.

3.1 *The associated total variation*

Similar to the nonlocal pre-perimeter (2.7) and perimeter (2.9), we can also define an associated pre-total variation and total variation with respect to the measure ν of a measurable function $u : \mathcal{X} \to \mathbb{R}$ as

$$\widetilde{\mathrm{TV}}_\varepsilon(u;\mu) := \frac{w_0}{\varepsilon}\int_{\mathcal{X}} \sup_{\tilde{x}\in B_\varepsilon(x)} u(\tilde{x}) - u(x)\,\mathrm{d}\rho_0(x) + \frac{w_1}{\varepsilon}\int_{\mathcal{X}} u(x) - \inf_{\tilde{x}\in B_\varepsilon(x)} u(\tilde{x})\,\mathrm{d}\rho_1(x), \tag{3.1}$$

$$\nu\text{-}\mathrm{TV}_\varepsilon(u;\mu) := \inf_{\substack{v\in L^\infty(\mathcal{X};\nu)\\ v=u\ \nu\text{-a.e.}}} \widetilde{\mathrm{TV}}_\varepsilon(v;\mu). \tag{3.2}$$

Remark 3.1. If $\mathcal{X} = \mathbb{R}^d$ and $w_1\rho_1 = w_0\rho_0 = 1/2\mathcal{L}^d$ and $\nu = \mathcal{L}^d$, our results in this section, in particular Proposition 3.11, show that the total variation reduces to

$$\mathrm{TV}_\varepsilon(u;\mu) = \frac{1}{2\varepsilon}\int_{\mathbb{R}^d} \operatorname{ess\,osc}_{B_\varepsilon(x)}(u)\,\mathrm{d}x, \tag{3.3}$$

which is precisely the nonlocal total variation associated to (2.10) which was studied in [19].

Remark 3.2. We could have defined $\widetilde{\mathrm{TV}}_\varepsilon$ and ν-TV_ε using the coarea formula. For the sake of clarity we decided to define the functionals directly and prove the coarea formula later.

Having the total variation at hand, a natural convex relaxation of the perimeter-regularized variational problem (2.12) to functions instead of sets is

$$\inf_{\substack{u\in L^\infty(\mathcal{X};\nu)\\ 0\le u\le 1,\ \nu\text{-a.e.}}} \mathbb{E}_{(x,y)\sim\mu}\left[|u(x)-y|\right] + \varepsilon\,\nu\text{-}\mathrm{TV}_\varepsilon(u;\mu), \tag{3.4}$$

where we again assume $\rho \ll \nu$. Indeed, we will use this relaxation as an intermediate step in order to prove existence for minimizers of (2.12). Notably, since the first term in (3.4) involves integrals with respect to ρ_0 and ρ_1, as shown in the proof of Proposition 2.8, the condition $\rho \ll \nu$ implies that it makes sense to perform the optimization in (3.4) over $L^\infty(\mathcal{X};\nu) \subseteq L^\infty(\mathcal{X};\rho)$.

3.2 *Properties of the nonlocal perimeter*

The nonlocal perimeter ν-$\mathrm{Per}_\varepsilon(\cdot;\mu)$ satisfies many of the same properties as the classical perimeter, which will also ensure that the total variation (3.2) is well-defined and convex.

Proposition 3.3. The set function ν-$\mathrm{Per}_\varepsilon(\cdot;\mu)$ defined in (2.9) satisfies the following:

- $0 \le \nu\text{-}\mathrm{Per}_\varepsilon(A;\mu) < \infty$ for all sets $A \in \mathfrak{B}(\mathcal{X})$.
- $\nu\text{-}\mathrm{Per}_\varepsilon(\emptyset;\mu) = \nu\text{-}\mathrm{Per}_\varepsilon(\mathcal{X};\mu) = 0$.
- $\nu\text{-}\mathrm{Per}_\varepsilon(A;\mu) = \nu\text{-}\mathrm{Per}_\varepsilon(A';\mu)$ if $\nu(A\triangle A') = 0$.
- It is submodular, meaning that for all $A, A' \in \mathfrak{B}(\mathcal{X})$ it holds

$$\nu\text{-}\mathrm{Per}_\varepsilon(A\cup A';\mu) + \nu\text{-}\mathrm{Per}_\varepsilon(A\cap A';\mu) \le \nu\text{-}\mathrm{Per}_\varepsilon(A;\mu) + \nu\text{-}\mathrm{Per}_\varepsilon(A';\mu).$$

REMARK 3.4. (Properties of the pre-perimeter). If we choose ν to be the measure defined by $\nu(\emptyset) = 0$ and $\nu(A) = \infty$ for all $A \in \mathfrak{B}(\mathcal{X}) \setminus \{\emptyset\}$, it holds $\widetilde{\mathrm{Per}}_\varepsilon(A;\mu) = \nu\text{-Per}_\varepsilon(A;\mu)$ for all $A \in \mathfrak{B}(\mathcal{X})$. Hence, the pre-perimeter admits the same properties.

Proof. The first statement follows from the fact that $\mathrm{osc}_{B_\varepsilon(x)}(1_B) \leq 1$ for all sets $B \in \mathfrak{B}(\mathcal{X})$ and ρ is a probability measure. The second statement is obvious since $\mathrm{osc}_{B_\varepsilon(x)}(1_\mathcal{X}) = 0$ for all $x \in \mathcal{X}$. The third statement follows from the very definition of the perimeter, involving the infimum over sets $B \in \mathfrak{B}(\mathcal{X})$ with $\nu(A \triangle B) = 0$.

Let us now prove submodularity. Elementary properties of the symmetric difference show

$$(A \cup A') \triangle (B \cup B') \subseteq (A \triangle B) \cup (A' \triangle B'),$$
$$(A \cap A') \triangle (B \cap B') \subseteq (A \triangle B) \cup (A' \triangle B').$$

Using subadditivity of the measure ρ, this implies that for all $B, B' \in \mathfrak{B}(\mathcal{X})$ with $\nu(A \triangle B) = 0$ and $\nu(A' \triangle B') = 0$, we can estimate

$$\begin{aligned}
&\nu\text{-Per}_\varepsilon(A \cup A';\mu) + \nu\text{-Per}_\varepsilon(A \cap A';\mu) \\
&\leq \frac{w_0}{\varepsilon} \int_\mathcal{X} \sup_{B_\varepsilon(\cdot)} 1_{B\cup B'} - 1_{B\cup B'} \,\mathrm{d}\rho_0 + \frac{w_1}{\varepsilon} \int_\mathcal{X} 1_{B\cup B'} - \inf_{B_\varepsilon(\cdot)} 1_{B\cup B'} \,\mathrm{d}\rho_1 \\
&\quad + \frac{w_0}{\varepsilon} \int_\mathcal{X} \sup_{B_\varepsilon(\cdot)} 1_{B\cap B'} - 1_{B\cap B'} \,\mathrm{d}\rho_0 + \frac{w_1}{\varepsilon} \int_\mathcal{X} 1_{B\cap B'} - \inf_{B_\varepsilon(\cdot)} 1_{B\cap B'} \,\mathrm{d}\rho_1.
\end{aligned}$$

Since $1_{B\cup B'} + 1_{B\cap B'} = 1_B + 1_{B'}$ and $1_B - \inf_{B_\varepsilon(\cdot)} 1_B = \sup_{B_\varepsilon(\cdot)} 1_{B^c} - 1_{B^c}$ for all $B, B' \in \mathfrak{B}(\mathcal{X})$, it suffices to show

$$\sup_{B_\varepsilon(x)} 1_{B\cup B'} + \sup_{B_\varepsilon(x)} 1_{B\cap B'} \leq \sup_{B_\varepsilon(x)} 1_B + \sup_{B_\varepsilon(x)} 1_{B'}. \tag{3.5}$$

Case 0: If the left-hand side is zero, we are done.

Case 1: Let us therefore assume that the first term in (3.5) is equal to one and the second one equal to zero. This means that there exists $y \in B \cup B'$ such that $\mathsf{d}(x,y) \leq \varepsilon$. In particular, at least one of the two terms on the right-hand side in (3.5) is ≥ 1, which proves the inequality in this case.

Case 2: Now we assume that the second term is one. This implies that there exists $y \in B \cap B'$ such that $\mathsf{d}(x,y) \leq \varepsilon$. Hence, both terms on the right-hand side in (3.5) are $= 1$ which makes the inequality correct independent of the first term. □

Next, we prove that the infimum in the definition of the perimeter (2.9) is actually attained. In fact, for a suitable measure ν with $\rho \ll \nu$, we even construct a precise representative, i.e. a set $A^\star \in \mathfrak{B}(\mathcal{X})$ with $A \sim_\nu A^\star$ in the equivalence relation (2.4), which attains this minimal value. Even more, we show that the perimeter coincides with the essential perimeter with respect to the measure ν. The measure ν has to satisfy the following assumption.

ASSUMPTION 1. We assume that there exists a σ-finite measure ν on $\mathcal{X}$ such that

1. $\rho \ll \nu$,
2. $\{x \in \mathcal{X} : \mathsf{dist}(x, \mathrm{supp}\rho) < \varepsilon\} \subseteq \mathrm{supp}\nu$,

3. ν is locally doubling (a Vitali measure), i.e.

$$\limsup_{r\downarrow 0} \frac{\nu(B_{2r}(x))}{\nu(B_r(x))} < \infty, \quad \text{for } \nu\text{-a.e. } x \in \mathcal{X}. \tag{3.6}$$

REMARK 3.5. Let us comment on these assumptions

1. The absolute continuity $\rho \ll \nu$ is needed for proving the reformulation as variational regularization problem, cf. Proposition 2.8.
2. The condition on suppν makes sure that problem (2.12) detects the effect of the adversary on the balls around points in the support of ρ.
3. The doubling assumption (3.6) is a very weak assumption under which the Lebesgue differentiation theorem (Theorem A.4 in Appendix A) is valid.

REMARK 3.6. (Choice of the measure ν). If $\mathcal{X} = \mathbb{R}^d$, then one can utilize a full support Gaussian γ to define $\nu := \rho + \gamma$. In that case (1)–(3) are true by definition and it is straightforward to show that if ρ is locally doubling, then so is ν, see also [34, p.81]. In turn, notice that if ρ is supported on finitely many points (e.g. an empirical measure) or if ρ is absolutely continuous with respect to the Lebesgue measure, then the measure ρ is locally doubling.

More generally, if $(\mathcal{X}, \mathsf{d})$ is a finite-dimensional smooth Riemannian manifold (intrinsically defined without the need of an Euclidean ambient space) with a Riemannian volume form ω, and finite total volume, then ν can be taken to be of the form $\nu = \rho + \omega$.

Using such a measure ν, we can state the following proposition which says that (a) the infimum in the definition of the perimeter (2.9) is attained and (b) that the perimeter can be expressed as the essential perimeter with respect to ν.

PROPOSITION 3.7. Under Assumption 1 for any $A \in \mathfrak{B}(\mathcal{X})$, there exists $A^\star \in \mathfrak{B}(\mathcal{X})$ with $A \sim_\nu A^\star$ such that

$$\nu\text{-Per}_\varepsilon(A;\mu) = \widetilde{\text{Per}}_\varepsilon(A^\star;\mu). \tag{3.7}$$

Furthermore, the perimeter admits the characterization

$$\nu\text{-Per}_\varepsilon(A;\mu) = \frac{w_0}{\varepsilon}\int_{\mathcal{X}} \nu\text{-ess}\sup_{B_\varepsilon(\cdot)} 1_A - 1_A \,\mathrm{d}\rho_0 + \frac{w_1}{\varepsilon}\int_{\mathcal{X}} 1_A - \nu\text{-ess}\inf_{B_\varepsilon(\cdot)} 1_A \,\mathrm{d}\rho_1. \tag{3.8}$$

For the proof of the proposition, we need a preparatory lemma which deals with the construction of the representative set.

LEMMA 3.8. Under Assumption 1 for any $A \in \mathfrak{B}(\mathcal{X})$, there exists $A^\star \in \mathfrak{B}(\mathcal{X})$ with $A \sim_\nu A^\star$ such that

$$\sup_{B_\varepsilon(x)} 1_{A^\star} = \nu\text{-ess}\sup_{B_\varepsilon(x)} 1_{A^\star}, \quad \inf_{B_\varepsilon(x)} 1_{A^\star} = \nu\text{-ess}\inf_{B_\varepsilon(x)} 1_{A^\star}, \quad \forall x \in \text{supp}\rho. \tag{3.9}$$

Proof. Let $u = 1_A$ and let D_+, D_- be the sets defined by

$$D_+ := \left\{ x \in \operatorname{supp}\rho \ : \ \nu\text{-ess}\sup_{B_\varepsilon(x)} u = 1, \quad \nu\text{-ess}\inf_{B_\varepsilon(x)} u = 1 \right\},$$

$$D_- := \left\{ x \in \operatorname{supp}\rho \ : \ \nu\text{-ess}\sup_{B_\varepsilon(x)} u = 0, \quad \nu\text{-ess}\inf_{B_\varepsilon(x)} u = 0 \right\}.$$

Also, let D_+^ε and D_-^ε be the sets

$$D_\pm^\varepsilon := \left\{ x \in \mathbb{R}^d \ : \ \mathsf{dist}(x, D_\pm) < \varepsilon \right\}.$$

We claim that D_+^ε and D_-^ε are disjoint. Indeed, suppose for the sake of contradiction that there is a point $\tilde{x}$ in their intersection. Then, we would be able to find $x_1 \in D_+$ and $x_0 \in D_-$ such that $\tilde{x} \in B_\varepsilon(x_1)$ and $\tilde{x} \in B_\varepsilon(x_0)$. In particular, we could find $\delta > 0$ small enough such that

$$B_\delta(\tilde{x}) \subseteq B_\varepsilon(x_1) \cap B_\varepsilon(x_0).$$

In addition, since D_+^ε (or D_-^ε) is by Assumption 1 a subset of the support of ν, we would conclude that $\tilde{x}$ belongs to the support of ν and thus $\nu(B_\delta(\tilde{x})) > 0$. However, this would be a contradiction, because the above inclusion implies that, for example, $\nu\text{-ess}\inf_{B_\varepsilon(x_1)} u = 0$, contrary to the fact that $x_1 \in D_+$.

Since D_+^ε and D_-^ε are disjoint, we can now define the function $u^\star$ as

$$u^\star(x) := \begin{cases} 1 & \text{if } x \in D_+^\varepsilon \\ 0 & \text{if } x \in D_-^\varepsilon \\ u(x) & \text{if } x \in \mathbb{R}^d \setminus (D_+^\varepsilon \cup D_-^\varepsilon). \end{cases}$$

Notice that the function $u^\star$ is Borel measurable since the sets $D_\pm^\varepsilon$ are open sets. We claim that ν-a.e., it holds $u = u^\star$. To see this, notice that it suffices to show that $u(x) = 1$ for ν-a.e. $x \in D_+^\varepsilon$ and that $u(x) = 0$ for ν-a.e. $x \in D_-^\varepsilon$; we can focus on the first case as the second one is completely analogous. By definition of D_+^ε, it holds

$$\left\{ x \in D_+^\varepsilon \ : \ u(x) = 0 \right\} \subseteq \left\{ x \in D_+^\varepsilon \ : \ u(x) \neq \lim_{r \to 0} \frac{1}{\nu(B_r(x))} \int_{B_r(x)} u(\tilde{x}) \, \mathrm{d}\nu(\tilde{x}) \right\}.$$

Notice that this is the case since for $r > 0$ small enough ν-a.e., it holds $u = 1$ in $B_r(x)$ when $x \in D_+^\varepsilon$. However, by the Lebesgue differentiation theorem applied to the measure ν and the measurable function u (which is possible thanks to Assumption 1, see Theorem A.4), the latter set must have ν measure zero. This implies our claim.

On the other hand, for every $x \in D_+$, by definition of $u^\star$, we have

$$\sup_{B_\varepsilon(x)} u^\star = 1 = \nu\text{-ess}\sup_{B_\varepsilon(x)} u^\star, \quad \inf_{B_\varepsilon(x)} u^\star = 1 = \nu\text{-ess}\inf_{B_\varepsilon(x)} u^\star$$

and for every $x \in D_-$

$$\inf_{B_\varepsilon(x)} u^\star = 0 = \nu\text{-ess}\inf_{B_\varepsilon(x)} u^\star, \quad \sup_{B_\varepsilon(x)} u^\star = 0 = \nu\text{-ess}\sup_{B_\varepsilon(x)} u^\star.$$

Finally, if $x \in \operatorname{supp}(\rho) \setminus (D_+ \cup D_-)$, we have

$$\nu\text{-ess}\sup_{B_\varepsilon(x)} u^\star = 1, \quad \nu\text{-ess}\inf_{B_\varepsilon(x)} u^\star = 0.$$

In particular, we also have

$$\nu\text{-ess}\sup_{B_\varepsilon(x)} u^\star = 1 = \sup_{B_\varepsilon(x)} u^\star, \quad \nu\text{-ess}\inf_{B_\varepsilon(x)} u^\star = 0 = \inf_{B_\varepsilon(x)} u^\star.$$

The set $A^\star$ is now defined as $A^\star := (u^\star)^{-1}(\{1\})$. This concludes the proof. □

REMARK 3.9. Notice that in the previous proof, specifically when we state that there is a $\delta > 0$ such that $B_\delta(\tilde{x}) \subseteq B_\varepsilon(x_1) \cap B_\varepsilon(x_0)$, we implicitly use the fact that the adversarial model was defined in terms of *open* balls B_ε as opposed to closed balls. The bottom line is that the construction of u^* in the proof would not carry through if we replaced open with closed balls since in that case, the sets $D^\varepsilon_\pm$(appropriately modified) would not necessarily be disjoint.

Now we are ready to prove Proposition 3.7.

Proof. (Proof of Proposition 3.7). Using the construction from Lemma 3.8, the definition of the perimeter (2.9), and the fact that sup $\geq$ ess sup, we compute

$$\begin{aligned}
&\frac{w_0}{\varepsilon}\int_{\mathcal{X}} \nu\text{-ess}\sup_{B_\varepsilon(\cdot)} 1_A - 1_A \,\mathrm{d}\rho_0 + \frac{w_1}{\varepsilon}\int_{\mathcal{X}} 1_A - \nu\text{-ess}\inf_{B_\varepsilon(\cdot)} 1_A \,\mathrm{d}\rho_1 \\
&= \frac{w_0}{\varepsilon}\int_{\mathcal{X}} \nu\text{-ess}\sup_{B_\varepsilon(\cdot)} 1_{A^\star} - 1_{A^\star} \,\mathrm{d}\rho_0 + \frac{w_1}{\varepsilon}\int_{\mathcal{X}} 1_{A^\star} - \nu\text{-ess}\inf_{B_\varepsilon(\cdot)} 1_{A^\star} \,\mathrm{d}\rho_1 \\
&= \frac{w_0}{\varepsilon}\int_{\mathcal{X}} \sup_{B_\varepsilon(\cdot)} 1_{A^\star} - 1_{A^\star} \,\mathrm{d}\rho_0 + \frac{w_1}{\varepsilon}\int_{\mathcal{X}} 1_{A^\star} - \inf_{B_\varepsilon(\cdot)} 1_{A^\star} \,\mathrm{d}\rho_1 \\
&\geq \nu\text{-Per}_\varepsilon(A;\mu) \\
&= \inf_{\substack{B\in\mathfrak{B}(\mathcal{X}) \\ A\sim_\nu B}} \frac{w_0}{\varepsilon}\int_{\mathcal{X}} \sup_{B_\varepsilon(\cdot)} 1_B - 1_B \,\mathrm{d}\rho_0 + \frac{w_1}{\varepsilon}\int_{\mathcal{X}} 1_B - \inf_{B_\varepsilon(\cdot)} 1_B \,\mathrm{d}\rho_1 \\
&\geq \inf_{\substack{B\in\mathfrak{B}(\mathcal{X}) \\ A\sim_\nu B}} \frac{w_0}{\varepsilon}\int_{\mathcal{X}} \nu\text{-ess}\sup_{B_\varepsilon(\cdot)} 1_B - 1_B \,\mathrm{d}\rho_0 + \frac{w_1}{\varepsilon}\int_{\mathcal{X}} 1_B - \nu\text{-ess}\inf_{B_\varepsilon(\cdot)} 1_B \,\mathrm{d}\rho_1 \\
&= \frac{w_0}{\varepsilon}\int_{\mathcal{X}} \nu\text{-ess}\sup_{B_\varepsilon(\cdot)} 1_A - 1_A \,\mathrm{d}\rho_0 + \frac{w_1}{\varepsilon}\int_{\mathcal{X}} 1_A - \nu\text{-ess}\inf_{B_\varepsilon(\cdot)} 1_A \,\mathrm{d}\rho_1.
\end{aligned}$$

Hence, equality holds everywhere, which completes the proof. □

3.3 *Properties of the total variation*

We start with an elementary homogeneity property of the total variation ν-TV$(\cdot;\mu)$, which follow immediately from its definition.

PROPOSITION 3.10. The functional ν-TV$(\cdot;\mu)$ defined in (3.2) satisfies the following for all measurable functions $u : \mathcal{X} \to \mathbb{R}$, $c \in \mathbb{R}$ and $\alpha \geq 0$:

$$\nu\text{-TV}(\alpha u + c;\mu) = \alpha\, \nu\text{-TV}(u;\mu).$$

Proof. The proof is trivial and we omit it. □

Now we prove the analogous result of Proposition 3.7 for the total variation. We rely heavily on the construction from Lemma 3.8.

PROPOSITION 3.11. Under Assumption 1, for any $u \in L^\infty(\mathcal{X};\nu)$, there exists $u^\star \in L^\infty(\mathcal{X};\nu)$ such that $u = u^\star$ holds ν-almost everywhere and

$$\nu\text{-TV}_\varepsilon(u;\mu) = \widetilde{\text{TV}}_\varepsilon(u^\star;\mu). \tag{3.10}$$

Furthermore, the total variation admits the characterization

$$\nu\text{-TV}_\varepsilon(u;\mu) = \frac{w_0}{\varepsilon}\int_{\mathcal{X}} \nu\text{-ess}\sup_{B_\varepsilon(\cdot)} u - u \,\mathrm{d}\rho_0 + \frac{w_1}{\varepsilon}\int_{\mathcal{X}} u - \nu\text{-ess}\inf_{B_\varepsilon(\cdot)} u \,\mathrm{d}\rho_1. \tag{3.11}$$

Proof. The proof works just as the proof of Proposition 3.7, however, using Lemma 3.12 below. □

The following lemma extends the construction of Lemma 3.12 from sets to functions. The proof is given in Appendix C.

LEMMA 3.12. Under Assumption 1 for any Borel measurable function $u \in L^\infty(\mathcal{X};\nu)$, there exists $u^\star : \mathcal{X} \to \mathbb{R}$ such that $u = u^\star$ holds ν-almost everywhere and

$$\sup_{B_\varepsilon(x)} u^\star = \nu\text{-ess}\sup_{B_\varepsilon(x)} u^\star, \quad \inf_{B_\varepsilon(x)} u^\star = \nu\text{-ess}\inf_{B_\varepsilon(x)} u^\star, \quad \forall x \in \operatorname{supp}\rho. \tag{3.12}$$

In fact, the nonlocal perimeter and total variation are connected via a coarea formula, as it is the case for their local counterparts. Thanks to the characterizations as essential perimeter and total variation from Propositions 3.7 and 3.11, the proof becomes very simple.

PROPOSITION 3.13. (Coarea formula). Under Assumption 1, it holds for any $u \in L^\infty(\mathcal{X};\nu)$ that

$$\nu\text{-TV}_\varepsilon(u;\mu) = \int_{\mathbb{R}} \nu\text{-Per}_\varepsilon(\{u \geq t\};\mu)\,\mathrm{d}t. \tag{3.13}$$

Proof. Let us first assume that $u \geq 0$.

Using Propositions 3.7 and 3.11, the layer cake representation, monotone convergence to swap integrals and supremum/infima, and Tonelli's theorem to swap integrals, we can compute

$$\begin{aligned}
\nu\text{-TV}_\varepsilon(u;\mu) &= \frac{w_0}{\varepsilon}\int_{\mathcal{X}} \nu\text{-ess}\sup_{B_\varepsilon(\cdot)} u - u \,\mathrm{d}\rho_0 + \frac{w_1}{\varepsilon}\int_{\mathcal{X}} u - \nu\text{-ess}\inf_{B_\varepsilon(\cdot)} u \,\mathrm{d}\rho_1 \\
&= \frac{w_0}{\varepsilon}\int_{\mathcal{X}} \nu\text{-ess}\sup_{B_\varepsilon(\cdot)} \int_0^\infty 1_{\{u\geq t\}}\,\mathrm{d}t - \int_0^\infty 1_{\{u\geq t\}}\,\mathrm{d}t\,\mathrm{d}\rho_0 \\
&\quad + \frac{w_1}{\varepsilon}\int_{\mathcal{X}}\int_0^\infty 1_{\{u\geq t\}}\,\mathrm{d}t - \nu\text{-ess}\inf_{B_\varepsilon(\cdot)} \int_0^\infty 1_{\{u\geq t\}}\,\mathrm{d}t\,\mathrm{d}\rho_1 \\
&= \int_0^\infty \left(\frac{w_0}{\varepsilon}\int_{\mathcal{X}} \nu\text{-ess}\sup_{B_\varepsilon(\cdot)} 1_{\{u\geq t\}} - 1_{\{u\geq t\}}\,\mathrm{d}\rho_0 \right. \\
&\quad \left. + \frac{w_1}{\varepsilon}\int_{\mathcal{X}} 1_{\{u\geq t\}} - \nu\text{-ess}\inf_{B_\varepsilon(\cdot)} 1_{\{u\geq t\}}\,\mathrm{d}\rho_1 \right) \mathrm{d}t \\
&= \int_0^\infty \nu\text{-Per}_\varepsilon(\{u\geq t\})\,\mathrm{d}t.
\end{aligned}$$

In the general case, we have that ν-a.e. it holds $m \leq u \leq M$ for some $m, M \in \mathbb{R}$ with $m \leq M$. We can define the function $\tilde{u} := u - m$, which satisfies $\tilde{u} \geq 0$ and hence

$$\nu\text{-TV}(\tilde{u};\mu) = \int_0^\infty \nu\text{-Per}(\{\tilde{u}\geq t\};\mu)\,\mathrm{d}t. \tag{3.14}$$

Using Proposition 3.10, it holds

$$\nu\text{-TV}(\tilde{u};\mu) = \nu\text{-TV}(u-m;\mu) = \nu\text{-TV}(u;\mu).$$

Furthermore, the perimeter integral satisfies

$$\int_0^\infty \nu\text{-Per}(\{\tilde{u}\geq t\};\mu)\,\mathrm{d}t = \int_0^\infty \nu\text{-Per}(\{u\geq t+m\};\mu)\,\mathrm{d}t = \int_m^\infty \nu\text{-Per}(\{u\geq t\};\mu)\,\mathrm{d}t.$$

Plugging these two reformulations into (3.14) shows

$$\nu\text{-TV}(u;\mu) = \int_m^\infty \nu\text{-Per}(\{u\geq t\};\mu)\,\mathrm{d}t = \int_{\mathbb{R}} \nu\text{-Per}(\{u\geq t\};\mu)\,\mathrm{d}t.$$

□

The main consequence of the previous properties of the perimeter and the total variation is that the the latter constitutes a convex and weak-* lower semicontinuous functional on $L^\infty(\mathcal{X};\nu)$.

Showing the weak-* lower semicontinuity on $L^\infty(\mathcal{X};\nu)$ requires a little bit more work. For this, we need a couple of preparatory lemmas. These depend on the validity of the Lebesgue differentiation theorem which requires the doubling condition in Assumption 1.

LEMMA 3.14. Assume that $(\mathcal{X}, \mathsf{d}, \nu)$ is a Vitali metric measure space, meaning that ν satisfies (3.6), assume that ν is σ-finite and suppose that $u_k \rightharpoonup^* u$ in $L^\infty(\mathcal{X}; \nu)$. Then, for ν-almost every $x \in \mathcal{X}$ and all $\varepsilon > 0$

$$\limsup_{k\to\infty} \nu\text{-ess}\inf_{B_\varepsilon(x)} u_k \le u(x) \le \liminf_{k\to\infty} \nu\text{-ess}\sup_{B_\varepsilon(x)} u_k.$$

Proof. Since ν is σ-finite, $L^\infty(\mathcal{X}; \nu)$ is the dual of $L^1(\mathcal{X}; \nu)$ and hence by definition of weak-* convergence (see Appendix A), it holds

$$\int_{\mathcal{X}} u\,\phi \,\mathrm{d}\nu = \lim_{k\to\infty} \int_{\mathcal{X}} u_k\,\phi \,\mathrm{d}\nu, \quad \forall \phi \in L^1(\mathcal{X}; \nu).$$

Choosing $\phi = \frac{1}{\nu(B_r(x))} 1_{B_r(x)}$ for $r > 0$, it holds that

$$\frac{1}{\nu(B_r(x))} \int_{B_r(x)} u \,\mathrm{d}\nu = \lim_{k\to\infty} \frac{1}{\nu(B_r(x))} \int_{B_r(x)} u_k \,\mathrm{d}\nu.$$

Hence, using Theorem A.4, we obtain for ν-a.e. $x \in \mathcal{X}$ any $\varepsilon > 0$

$$\begin{aligned} u(x) &= \lim_{r\downarrow 0} \frac{1}{\nu(B_r(x))} \int_{B_r(x)} u \,\mathrm{d}\nu \\ &= \lim_{r\downarrow 0} \lim_{k\to\infty} \frac{1}{\nu(B_r(x))} \int_{B_r(x)} u_k \,\mathrm{d}\nu \\ &\le \limsup_{r\downarrow 0} \liminf_{k\to\infty} \nu\text{-ess}\sup_{B_r(x)} u_k \\ &\le \liminf_{k\to\infty} \nu\text{-ess}\sup_{B_\varepsilon(x)} u_k. \end{aligned}$$

Similarly, one establishes the inequality $u(x) \ge \limsup_{k\to\infty} \nu\text{-ess}\inf_{B_\varepsilon(x)} u_k$. □

LEMMA 3.15. Under the conditions of Lemma 3.14, it holds for ν-a.e. $x \in \mathcal{X}$ and all $\varepsilon > 0$

$$\begin{aligned} \nu\text{-ess}\sup_{B_\varepsilon(x)} u &\le \liminf_{k\to\infty} \nu\text{-ess}\sup_{B_\varepsilon(x)} u_k, \\ \nu\text{-ess}\inf_{B_\varepsilon(x)} u &\ge \limsup_{k\to\infty} \nu\text{-ess}\inf_{B_\varepsilon(x)} u_k. \end{aligned}$$

Proof. Let us choose $0 < \delta < \varepsilon$. For ν-almost every $y \in \mathcal{X}$, Lemma 3.14 implies

$$u(y) \le \liminf_{k\to\infty} \nu\text{-ess}\sup_{B_\delta(y)} u_k.$$

Taking the ν-ess sup over $y \in B_{\varepsilon-\delta}(x)$ yields

$$\begin{aligned}\nu\text{-ess}\sup_{B_{\varepsilon-\delta}(x)} u &\leq \nu\text{-ess}\sup_{y\in B_{\varepsilon-\delta}(x)} \liminf_{k\to\infty} \nu\text{-ess}\sup_{B_\delta(y)} u_k \\ &\leq \liminf_{k\to\infty} \nu\text{-ess}\sup_{y\in B_{\varepsilon-\delta}(x)} \nu\text{-ess}\sup_{B_\delta(y)} u_k \\ &\leq \liminf_{k\to\infty} \nu\text{-ess}\sup_{B_\varepsilon(x)} u_k.\end{aligned}$$

Choosing $\delta > 0$ arbitrarily small yields

$$\nu\text{-ess}\sup_{B_\varepsilon(x)} u \leq \liminf_{k\to\infty} \nu\text{-ess}\sup_{B_\varepsilon(x)} u_k.$$

Applying this reasoning to $-u_k$, one shows analogously that

$$\limsup_{k\to\infty} \nu\text{-ess}\inf_{B_\varepsilon(x)} u_k \leq \nu\text{-ess}\inf_{B_\varepsilon(x)} u.$$

□

Now we are ready to prove the following proposition which states important properties of the total variation.

PROPOSITION 3.16. Under Assumption 1, the functional $\nu\text{-TV}_\varepsilon(\cdot;\mu)$, defined in (3.2), is a positively homogeneous, weak-* lower semicontinuous and convex functional on $L^\infty(\mathcal{X};\nu)$. Lower semicontinuity is understood in the sense that

$$u_k \rightharpoonup^* u \text{ in } L^\infty(\mathcal{X};\nu) \implies \nu\text{-TV}_\varepsilon(u;\mu) \leq \liminf_{k\to\infty} \nu\text{-TV}_\varepsilon(u_k;\mu).$$

Proof. The positive homogeneity was already proved in Proposition 3.10. To prove lower semicontinuity, we use Proposition 3.11 to write

$$\begin{aligned}\nu\text{-TV}_\varepsilon(u;\mu) &= \frac{w_0}{\varepsilon}\left(\int_{\mathcal{X}} \nu\text{-ess}\sup_{B_\varepsilon(x)} u \,\mathrm{d}\rho_0 - \int_{\mathcal{X}} u\,\mathrm{d}\rho_0\right) \\ &\quad + \frac{w_1}{\varepsilon}\left(\int_{\mathcal{X}} u\,\mathrm{d}\rho_1 - \int_{\mathcal{X}} \nu\text{-ess}\inf_{B_\varepsilon(x)} u\,\mathrm{d}\rho_1\right) \\ &= \frac{w_0}{\varepsilon}\left(\int_{\mathcal{X}} \nu\text{-ess}\sup_{B_\varepsilon(x)} u\frac{\mathrm{d}\rho_0}{\mathrm{d}\nu}\,\mathrm{d}\nu - \int_{\mathcal{X}} u\frac{\mathrm{d}\rho_0}{\mathrm{d}\nu}\,\mathrm{d}\nu\right) \\ &\quad + \frac{w_1}{\varepsilon}\left(\int_{\mathcal{X}} u\frac{\mathrm{d}\rho_1}{\mathrm{d}\nu}\,\mathrm{d}\nu - \int_{\mathcal{X}} \nu\text{-ess}\inf_{B_\varepsilon(x)} u\frac{\mathrm{d}\rho_1}{\mathrm{d}\nu}\,\mathrm{d}\nu\right),\end{aligned}$$

where $\frac{\mathrm{d}\rho_i}{\mathrm{d}\nu}$ denotes the Radon–Nikodým derivative of ρ_i with respect to ν (note that $\rho \ll \nu$ and $\rho = w_0\rho_0 + w_1\rho_1$ implies $\rho_i \ll \nu$ for $i \in \{0,1\}$). Let $(u_k)_{k\in\mathbb{N}} \subseteq L^\infty(\mathcal{X};\nu)$ be a sequence such that $u_k \rightharpoonup^* u$

as $k \to \infty$ where $u \in L^\infty(\mathcal{X}; \nu)$. Then, it holds

$$\int_{\mathcal{X}} \nu\text{-ess sup}_{B_\varepsilon(x)} u \frac{\mathrm{d}\rho_0}{\mathrm{d}\nu}\,\mathrm{d}\nu \leq \int_{\mathcal{X}} \liminf_{k\to\infty} \nu\text{-ess sup}_{B_\varepsilon(x)} u_k \frac{\mathrm{d}\rho_0}{\mathrm{d}\nu}\,\mathrm{d}\nu \leq \liminf_{k\to\infty} \int_{\mathcal{X}} \nu\text{-ess sup}_{B_\varepsilon(x)} u_k \frac{\mathrm{d}\rho_0}{\mathrm{d}\nu}\,\mathrm{d}\nu.$$

Note that, being a weakly-* convergent sequence, $\{u_k\}_{k\in\mathbb{N}}$ is uniformly bounded in $L^\infty(\mathcal{X}; \nu)$ by some constant $C > 0$. Furthermore, $\int_{\mathcal{X}} C \frac{\mathrm{d}\rho_0}{\mathrm{d}\nu}\,\mathrm{d}\nu = C\rho_0(\mathcal{X}) < \infty$ which justifies an application of Fatou's lemma to the sequence $\nu\text{-ess sup}_{B_\varepsilon(x)} u_k + C$ for the second inequality. One argues analogously for the other integral containing the ν-ess inf, using the reverse Fatou lemma.

Furthermore, since $\frac{\mathrm{d}\rho_i}{\mathrm{d}\nu} \in L^1(\mathcal{X}; \nu)$ for $i \in \{0, 1\}$, weak-* convergence of u_k to u directly implies

$$\int_{\mathcal{X}} u \frac{\mathrm{d}\rho_i}{\mathrm{d}\nu}\,\mathrm{d}\nu = \lim_{k\to\infty} \int_{\mathcal{X}} u_k \frac{\mathrm{d}\rho_i}{\mathrm{d}\nu}\,\mathrm{d}\nu, \qquad i \in \{0, 1\}.$$

Hence, we have established weak-* lower semicontinuity of ν-TV.

Convexity is a direct consequence of the submodularity of the perimeter, the coarea formula from Proposition 3.13 and the lower semicontinuity; the proof works just as in [17, Prop. 3.4]. □

3.4 *Existence of solutions*

We have completed all preparations to finally state our existence result for the adversarial problem (1.7). The proof uses the direct method to establish existence of a minimizer of the variational problem (2.12). Then, we use the representative constructed in Proposition 3.7 to turn this minimizer into a minimizer of the original problem (1.7). This last step is shown in the following lemma.

LEMMA 3.17. Let $A \in \mathfrak{B}(\mathcal{X})$ be a solution of (2.12). Then $A^\star$, constructed in Proposition 3.7, is a solution of (1.7).

Proof. Using Proposition 2.8 and 3.7, we get

$$\begin{aligned}
\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_{A^\star} - y\right|\right] + \varepsilon\,\widetilde{\mathrm{Per}}_\varepsilon(A^\star; \mu) &= \mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A - y\right|\right] + \varepsilon\,\nu\text{-Per}_\varepsilon(A; \mu) \\
&= \inf_{A\in\mathfrak{B}(\mathcal{X})} \mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A - y\right|\right] + \varepsilon\,\nu\text{-Per}_\varepsilon(A; \mu) \\
&= \inf_{A\in\mathfrak{B}(\mathcal{X})} \mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A - y\right|\right] + \varepsilon\,\widetilde{\mathrm{Per}}_\varepsilon(A; \mu),
\end{aligned}$$

which, thanks to Proposition 2.7, is equivalent to $A^\star$ solving (1.7). □

THEOREM 3.18. (Existence of Minimizers). Under Assumption 1, there exists a solution $A \in \mathfrak{B}(\mathcal{X})$ of problem (1.7).

Proof. Let $(A_k)_{k\in\mathbb{N}} \subseteq \mathfrak{B}(\mathcal{X})$ be a minimizing sequence of (2.12) which is trivially bounded in $L^\infty(\mathcal{X}; \nu)$. Using weak-* precompactness of bounded subsets of $L^\infty(\mathcal{X}; \nu)$ (see Theorem A.6 in Appendix A), we know that there exists $u \in L^\infty(\mathcal{X}; \nu)$ such that a subsequence (which we don't relabel) satisfies $1_{A_k} \rightharpoonup^* u$ in $L^\infty(\mathcal{X}; \nu)$. Furthermore, from Lemma 3.14, we know that $0 \leq u(x) \leq 1$ for ν-a.e. $x \in \mathcal{X}$.

Let us first show that the empirical risk $\mathbb{E}_{(x,y)\sim\mu}\left[\left|u(x)-y\right|\right]$ is weak-* lower semicontinuous, in fact even continuous, along this sequence. For this, we compute is as

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mu}\left[|u(x)-y|\right] &= w_0\int_{\mathcal{X}}|u(x)|\,\mathrm{d}\rho_0(x)+w_1\int_{\mathcal{X}}|u(x)-1|\,\mathrm{d}\rho_1(x)\\ &= w_0\int_{\mathcal{X}}u(x)\frac{\mathrm{d}\rho_0}{\mathrm{d}\nu}\,\mathrm{d}\nu(x)+w_1\int_{\mathcal{X}}(1-u(x))\frac{\mathrm{d}\rho_1}{\mathrm{d}\nu}\,\mathrm{d}\nu(x)\\ &= \lim_{k\to\infty} w_0\int_{\mathcal{X}}1_{A_k}(x)\frac{\mathrm{d}\rho_0}{\mathrm{d}\nu}\,\mathrm{d}\nu(x)+w_1\int_{\mathcal{X}}(1-1_{A_k}(x))\frac{\mathrm{d}\rho_1}{\mathrm{d}\nu}\,\mathrm{d}\nu(x)\\ &= \lim_{k\to\infty}\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_{A_k}(x)-y\right|\right].\end{aligned}$$

Using Proposition 3.16 and the fact that $\nu\text{-TV}(1_A;\mu)=\nu\text{-Per}_\varepsilon(A;\mu)$ for all $A\in\mathfrak{B}(\mathcal{X})$, we infer that

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mu}\left[|u(x)-y|\right]+\varepsilon\,\nu\text{-TV}_\varepsilon(u;\mu) &\le \liminf_{k\to\infty}\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_{A_k}(x)-y\right|\right]+\varepsilon\,\nu\text{-Per}_\varepsilon(1_{A_k};\mu)\\ &= \inf_{A\subseteq\mathfrak{B}(\mathcal{X})}\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A(x)-y\right|\right]+\varepsilon\,\nu\text{-Per}_\varepsilon(A;\mu).\end{aligned}\tag{3.15}$$

For $t\in[0,1]$, define the set $A_t:=\{u\ge t\}$. It trivially holds

$$\inf_{A\subseteq\mathfrak{B}(\mathcal{X})}\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A(x)-y\right|\right]+\varepsilon\,\nu\text{-Per}_\varepsilon(A;\mu)\le\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_{A_t}(x)-y\right|\right]+\varepsilon\,\nu\text{-Per}_\varepsilon(A_t;\mu).$$

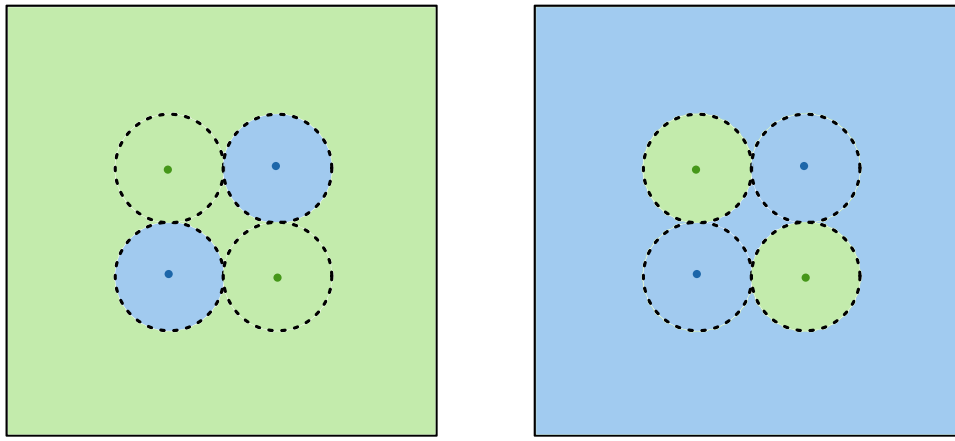
Aiming for a contradiction, we assume this inequality to be strict on a subset of $[0,1]$ with positive Lebesgue measure. Integrating over $t\in[0,1]$ and using Proposition 3.13, we get

$$\begin{aligned}&\inf_{A\subseteq\mathfrak{B}(\mathcal{X})}\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A(x)-y\right|\right]+\varepsilon\,\nu\text{-Per}_\varepsilon(A;\mu)\\ &<\int_0^1\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_{A_t}(x)-y\right|\right]+\varepsilon\,\nu\text{-Per}_\varepsilon(A_t;\mu)\,\mathrm{d}t\\ &=\mathbb{E}_{(x,y)\sim\mu}\left[|u(x)-y|\right]+\varepsilon\,\nu\text{-TV}_\varepsilon(u;\mu),\end{aligned}$$

which contradicts (3.15). Hence, the inequality is an equality which shows that also A_t is a minimizer of (2.12) for almost all $t\in[0,1]$. In particular, Lemma 3.17 shows that $A_t^\star$ solves (1.7) for almost every $t\in[0,1]$. □

The previous proposition establishes the existence of minimizers of the adversarial problem (1.7). However, it is not yet clear whether minimizers are unique or regular (of course considering equivalence classes modulo ν). However, since the problem is not strictly convex in nature—cf. the relaxation (3.4)—in general uniqueness cannot be expected. This can trivially arise due to a separation between the supports of ρ_0 and ρ_1, as evidenced by the following example.

EXAMPLE 3.19. Fixing $\varepsilon>0$, suppose that μ is given by four Dirac masses centered at $(\pm\varepsilon,\pm\varepsilon)$ in $\mathbb{R}^2$ and that opposing corners are (deterministically) given the same label, namely $\rho_0=\frac{1}{2}\delta_{(\varepsilon,-\varepsilon)}+\frac{1}{2}\delta_{(-\varepsilon,\varepsilon)}$



(a) Smallest adversarial minimizer (b) Largest adversarial minimizer

FIG. 3. Situation from Example 3.19 with non-unique, smooth minimizers. Here, the four Dirac masses are displayed as well as the balls of radius ε which surround them. Infinitely many minimizers of the adversarial risk exist, we display here the largest and smallest possible minimizers in blue color.

and $\rho_1 = \frac{1}{2}\delta_{(-\varepsilon,-\varepsilon)} + \frac{1}{2}\delta_{(\varepsilon,\varepsilon)}$, and $w_0 = w_1 = \frac{1}{2}$. Then, it is straightforward to check that any set A such that $B_\varepsilon((\varepsilon,\varepsilon)) \subseteq A$, $B_\varepsilon((-\varepsilon,-\varepsilon)) \subseteq A$, $B_\varepsilon((-\varepsilon,\varepsilon)) \cap A = \emptyset$ and $B_\varepsilon((\varepsilon,-\varepsilon)) \cap A = \emptyset$ will be minimizers of the adversarial risk: indeed any such set has zero risk. The largest and smallest such sets (in blue color) are demonstrated in Fig. 3.

The previous example demonstrates that one cannot hope for any type of uniqueness or even that *all* minimizers will necessarily be regular. Although the previous example utilized Dirac masses for simplicity, we suspect that many of the same issues can arise for distributions with smooth densities.

Despite the previous considerations, it is possible to obtain some positive results. In particular, we can define notions of maximal and minimal minimizers to (1.7) which are then shown to be unique. Moreover, we show that although there may be irregular minimizers, we can always find regular minimizers provided that we define an appropriate notion of regularity relative to the metric d.

Proving this will be the content of the following two sections.

3.5 *Extremal solutions*

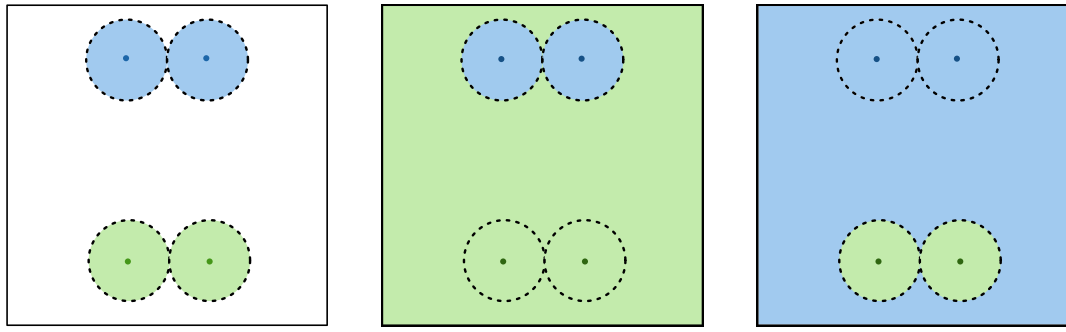
For notational convenience, we define the adversarial risk associated to (1.7) as

$$\widetilde{R}_\varepsilon(A) := \mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B_\varepsilon(x)} \left|1_A(\tilde{x}) - y\right|\right] = \mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A(x) - y\right|\right] + \varepsilon\,\widetilde{\mathrm{Per}}_\varepsilon(A;\mu). \tag{3.16}$$

To begin, we prove submodularity of the adversarial risk and show that the set of minimizers is closed under unions and intersections.

LEMMA 3.20. The adversarial risk is submodular, meaning that it satisfies

$$\widetilde{R}_\varepsilon(A\cup B) + \widetilde{R}_\varepsilon(A\cap B) \le \widetilde{R}_\varepsilon(A) + \widetilde{R}_\varepsilon(B), \quad \forall A, B \in \mathfrak{B}(\mathcal{X}). \tag{3.17}$$



(a) Data surrounded by ε-balls (b) Minimal set (c) Maximal set

FIG. 4. The maximal and minimal sets (in blue color) associated with a particular distribution of point masses. Here, the maximal and minimal sets have boundaries that cannot even be represented as graphs of a function at every point. In this case, any intermediate set, in the sense of inclusion, will also be a minimizer, and many smooth minimizers are possible.

Proof. We first notice that

$$\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A(x)-y\right|\right]+\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_B(x)-y\right|\right]=\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_{A\cap B}(x)-y\right|\right]+\mathbb{E}_{(x,y)\sim\mu}\left[\left|1_{A\cup B}(x)-y\right|\right].$$

This fact can be directly proved by decomposing $\mathcal{X}$ into $A\cap B$, $B\setminus A$, $A\setminus B$ and $\mathcal{X}\setminus(A\cup B)$, splitting the integrals and then reassembling. Together with the submodularity of the pre-perimeter (cf. Remark 3.4), this implies the assertion. □

PROPOSITION 3.21. Let A and B be minimizers of the adversarial problem (1.7) with parameter $\varepsilon\geq 0$. Then, both $A\cap B$ and $A\cup B$ are both also minimizers.

Proof. Using Lemma 3.20, it is immediate that, for any A,B, either $\widetilde{R}_\varepsilon(A\cup B)\leq\frac{\widetilde{R}_\varepsilon(A)+\widetilde{R}_\varepsilon(B)}{2}$ or $\widetilde{R}_\varepsilon(A\cap B)\leq\frac{\widetilde{R}_\varepsilon(A)+\widetilde{R}_\varepsilon(B)}{2}$: suppose that the former is true. Then, if A and B are both minimizers then we immediately obtain that $A\cup B$ is also a minimizer. Subtracting the minimal risk from both sides then also implies that $\widetilde{R}_\varepsilon(A\cap B)=\widetilde{R}_\varepsilon(A)$. The other case is completely analogous. □

We now proceed to introduce the setting under which we can make sense of maximal and minimal solutions to problem (1.7). In fact, we first work with the relaxed problem (2.12) and then use Lemma 3.17 to obtain statements about the original problem (1.7). We introduce the following notation for sets $A,A'\in\mathfrak{B}(\mathcal{X})$:

$$A\preceq_\nu A' :\Longleftrightarrow\ 1_A(x)\leq 1_{A'}(x)\quad\text{for }\nu\text{-a.e. }x\in\mathcal{X}. \tag{3.18}$$

Notice that the relation $\preceq_\nu$ above induces a partial order in the set of equivalence classes of $\sim_\nu$, in other words in the quotient σ-algebra $\mathfrak{B}_\nu(\mathcal{X})$. We now define maximal and minimal solutions and show their existence and uniqueness, see Fig. 4 for an example.

DEFINITION 3.22. (Maximal (minimal) solutions). We say that $A\in\mathfrak{B}(\mathcal{X})$ is a maximal (minimal) solution of (2.12) if A is a solution with the property that any other solution $A'\in\mathfrak{B}(\mathcal{X})$ to (2.12) satisfying $A\preceq_\nu A'$ ($A'\preceq_\nu A$) must satisfy $A\sim_\nu A'$.

PROPOSITION 3.23. Assume that ν is a finite measure on $\mathcal{X}$. Then, there exists a unique maximal (minimal) solution to problem (2.12) up to ν-equivalence. The maximal solution is denoted with $A_{\max}$, while the minimal solution is denoted with $A_{\min}$.

Proof. We follow an argument in [20] and proceed as follows. Since ν is a finite measure

$$m := \sup\{\nu(A) \ : \ A \text{ solution of (2.12)}\} < \infty.$$

Take a maximizing sequence $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathfrak{B}(\mathcal{X})$ in the definition of m so that $\lim_{n\to\infty} \nu(A_n) = m$. From Proposition 3.21, we know that for each $n \in \mathbb{N}$, the set $\bigcup_{k=1}^n A_k$ is also a solution to problem (2.12). Let $A := \bigcup_{k=1}^\infty A_k$, then it holds

$$1_{\bigcup_{k=1}^n A_k} \to 1_A, \quad \text{in } L^1(\mathcal{X};\nu) \text{ as } n \to \infty.$$

From the above strong convergence, it is immediate that A is also a solution to (2.12) and we have

$$m = \lim_{n\to\infty} \nu(A_n) \leq \lim_{n\to\infty} \nu\left(\bigcup_{k=1}^n A_k\right) = \nu(A) \leq m.$$

Now, notice that if there was a solution A' such that $A \preceq_\nu A'$ and $A \not\sim_\nu A'$, then we would have $m = \nu(A) < \nu(A')$ which would contradict the definition of m. Likewise, if there were two solutions A, A' with $\nu(A) = m = \nu(A')$ and the two sets were not equivalent, then by taking their union we would be able to obtain a solution with ν-volume strictly larger than m. This shows the existence and uniqueness of maximal solutions. A similar proof can be used to deduce the existence and uniqueness of minimal solutions. □

We can also introduce a notion of maximality and minimality of solutions for problem (1.7), at least when restricting to a class of solutions obtained by considering specific representatives of solutions $A \in \mathfrak{B}(\mathcal{X})$ to (1.7). In contrast to the definition of $A^\star$ in Lemma 3.8, which in general is representative dependent, the following notions are independent of the representative of A in the quotient σ-algebra $\mathfrak{B}_\nu(\mathcal{X})$. Given $A \in \mathfrak{B}(\mathcal{X})$, we define Borel sets A^+ and A^- through their indicators according to the formulas

$$1_{A^+}(x) := \begin{cases} 1 \text{ if } x \in \operatorname{supp}(\nu) \text{ and } \limsup_{r\downarrow 0} \frac{\nu(B_r(x)\cap A)}{\nu(B_r(x))} > 0, \\ 1 \text{ if } x \notin \operatorname{supp}(\nu), \\ 0 \text{ if } x \in \operatorname{supp}(\nu) \text{ and } \limsup_{r\downarrow 0} \frac{\nu(B_r(x)\cap A)}{\nu(B_r(x))} = 0, \end{cases} \qquad x \in \mathcal{X},$$

$$1_{A^-}(x) := \begin{cases} 1 \text{ if } x \in \operatorname{supp}(\nu) \text{ and } \liminf_{r\downarrow 0} \frac{\nu(B_r(x)\cap A)}{\nu(B_r(x))} = 1, \\ 0 \text{ if } x \notin \operatorname{supp}(\nu), \\ 0 \text{ if } x \in \operatorname{supp}(\nu) \text{ and } \liminf_{r\downarrow 0} \frac{\nu(B_r(x)\cap A)}{\nu(B_r(x))} < 1, \end{cases} \qquad x \in \mathcal{X}.$$

Notice that for any Borel set A, we have $A^- \subseteq A^+$. In addition, notice that $A^+ = (A^+)^\star$ as well as $A^- = (A^-)^\star$. In particular, if A is a solution to problem (2.12), then both A^+ and A^- are solutions to problem (1.7) according to Lemma 3.17.

The next is an immediate consequence of Proposition 3.23 and the above definitions.

COROLLARY 3.24. Assume that ν is a finite measure. Among the set of solutions to problem (1.7) of the form A^+ for some solution $A \in \mathfrak{B}(\mathcal{X})$ of problem (2.12), $A^+_{\max}$ is maximal in the sense of inclusions. Likewise, among the set of solutions to problem (1.7) of the form A^- for some solution A of problem (2.12), $A^-_{\min}$ is maximal in the sense of inclusions. In addition, $A^-_{\min} \subseteq A^+_{\max}$.

Proof. Notice that if $A \preceq_\nu A'$, we immediately have $A^+ \subseteq (A')^+$ and $A^- \subseteq (A')^-$. Since we have $A_{\min} \preceq_\nu A \preceq_\nu A_{\max}$ for any solution A of (2.12), the first part of the corollary follows. The inclusion $A^-_{\min} \subseteq A^+_{\max}$ follows from $A^-_{\min} \subseteq A^+_{\min} \subseteq A^+_{\max}$. □

3.6 *Regularity*

The goal of this section is to prove that, in an Euclidean setting, it is possible to construct a smooth minimizer of the adversarial problem. We offer a direct construction, under which normal vectors of the boundary are Hölder continuous and shall prove the following statement.

THEOREM 3.25. Consider the case where $\mathcal{X} = \mathbb{R}^d$, equipped with the standard Euclidean metric. Then, for any $\varepsilon > 0$, there exists a minimizer $B \in \mathfrak{B}(\mathbb{R}^d)$ to the adversarial problem (1.7) which is locally the graph of a $C^{1,1/3}$ function.

We will first deduce a series of regularity properties of minimizers that, although not as strong as those in Theorem 3.25, hold for general metric measure spaces $(\mathcal{X}, \mathrm{d}, \nu)$ satisfying Assumption 1, before proceeding to the proof of Theorem 3.25.

We start by introducing some fundamental concepts of mathematical morphology. In particular, we define the following important concepts from mathematical morphology (see, e.g. Chapter 2 in [54]).

DEFINITION 3.26. (Morphology). Let $A \subseteq \mathcal{X}$ be a set and $\varepsilon > 0$. We define its

- dilation as $A^\varepsilon := \{x \in \mathcal{X} : \mathsf{dist}(x, A) < \varepsilon\}$,
- erosion as $A^{-\varepsilon} := \{x \in \mathcal{X} : \mathsf{dist}(x, A^c) \geq \varepsilon\}$,
- closing as $\mathsf{cl}_\varepsilon(A) := (A^\varepsilon)^{-\varepsilon}$,
- opening as $\mathsf{op}_\varepsilon(A) := (A^{-\varepsilon})^\varepsilon$.

Notice that all these sets are measurable as they are open or closed sets. In the following proposition, we collect a couple of important properties of these operations, which can be proved in a straightforward way (see [33]).

PROPOSITION 3.27. The following statements hold true:

- $\mathsf{cl}_\varepsilon(A)$ is a closed set that contains A,
- $\mathsf{op}_\varepsilon(A)$ is an open set contained in A,
- $\mathsf{cl}_\varepsilon(A)^\varepsilon = A^\varepsilon$,
- $\mathsf{op}_\varepsilon(A)^{-\varepsilon} = A^{-\varepsilon}$,
- $A^{-\varepsilon} = ((A^c)^\varepsilon)^c$,
- $\mathsf{cl}_\varepsilon(A^c) = \mathsf{op}_\varepsilon(A)^c$.

The following definition of one-sided regularity of sets is strongly connected to the opening and closing procedures.

DEFINITION 3.28. (Inner and outer regularity). A set $A \subseteq \mathcal{X}$ is called ε inner regular relative to the metric d if, for any point, $x \in \partial A$, then there exists a point $y \in \mathcal{X}$ so that $\mathsf{d}(x, y) = \varepsilon$ and $B_\varepsilon(y) \subseteq A$. A set $A \subseteq \mathcal{X}$ is called ε outer regular relative to the metric d if instead we can always find such a $y \in \mathcal{X}$ satisfying the inclusion $B_\varepsilon(y) \subseteq A^c$.

Note that by definition, for any set $A \subseteq \mathcal{X}$, its closing $\mathsf{cl}_\varepsilon(A)$ is ε outer regular, whereas its opening $\mathsf{op}_\varepsilon(A)$ is ε inner regular. Furthermore, in $\mathcal{X} = \mathbb{R}^d$ equipped with the Euclidean metric, it was shown in [39] that a set which is both ε inner and outer regular has a $C^{1,1}$ boundary.

A similar concept of regularity, called pseudo-certifiable robustness, is introduced and used in [3]. There, a set A is called pseudo-certifiably robust if every point in the set (or its complement) is an element of an ε-ball contained in the set (or its complement). It is easy to show that this notion of regularity implies inner and outer regularity in the sense of Definition 3.28.

We now show that the opening and closing operations do not increase the adversarial risk (3.16). As a consequence, the operations turn minimizers into minimizers.

LEMMA 3.29. For $A \in \mathfrak{B}(\mathcal{X})$, it holds

$$\widetilde{R}_\varepsilon(\mathsf{cl}_\varepsilon(A)) \leq \widetilde{R}_\varepsilon(A), \qquad \widetilde{R}_\varepsilon(\mathsf{op}_\varepsilon(A)) \leq \widetilde{R}_\varepsilon(A).$$

Proof. Using Proposition 3.27, we can rewrite the adversarial risk as follows:

$$\begin{aligned}
\widetilde{R}_\varepsilon(A) &= w_0 \int_{\mathcal{X}} \sup_{B_\varepsilon(x)} 1_A \, \mathrm{d}\rho_0(x) + w_1 \int_{\mathcal{X}} \sup_{B_\varepsilon(x)} 1_{A^c} \, \mathrm{d}\rho_1(x) \\
&= w_0 \rho_0(A^\varepsilon) + w_1 \rho_1((A^c)^\varepsilon) \\
&= w_0 \rho_0(A^\varepsilon) + w_1 - w_1 \rho_1(((A^c)^\varepsilon)^c) \\
&= w_0 \rho_0(A^\varepsilon) + w_1 - w_1 \rho_1(A^{-\varepsilon}).
\end{aligned}$$

Using Proposition 3.27 again, we get

$$\begin{aligned}
\widetilde{R}_\varepsilon(\mathsf{cl}_\varepsilon(A)) &= w_0 \rho_0(\mathsf{cl}_\varepsilon(A)^\varepsilon) + w_1 - w_1 \rho_1(\mathsf{cl}_\varepsilon(A)^{-\varepsilon}) \\
&\leq w_0 \rho_0(A^\varepsilon) + w_1 - w_1 \rho_1(A^{-\varepsilon}) = \widetilde{R}_\varepsilon(A), \\
\widetilde{R}_\varepsilon(\mathsf{op}_\varepsilon(A)) &= w_0 \rho_0(\mathsf{op}_\varepsilon(A)^\varepsilon) + w_1 - w_1 \rho_1(\mathsf{op}_\varepsilon(A)^{-\varepsilon}) \\
&\leq w_0 \rho_0(A^\varepsilon) + w_1 - w_1 \rho_1(A^{-\varepsilon}) = \widetilde{R}_\varepsilon(A).
\end{aligned}$$

□

COROLLARY 3.30. Let $A \in \mathfrak{B}(\mathcal{X})$ be a minimizer of (1.7). Then, $\mathsf{op}_\varepsilon(A)$ and $\mathsf{cl}_\varepsilon(A)$ are also minimizers.

We can now show that one can always construct a closed and ε outer regular maximal set and an open and ε inner regular minimal set which solves the adversarial problem (1.7).

PROPOSITION 3.31. Assume that ν is a finite measure.

There exist two solutions A'_+ and A'_- to (1.7) with the following properties:

1. $A'_- \subseteq A'_+$.
2. A'_+ is a closed set and A'_- is an open set.
3. A'_+ is ε outer regular relative to the metric d and A'_- is ε inner regular with respect to the metric d.
4. $A_{\max} \sim_\nu A'_+$ and $A_{\min} \sim_\nu A'_-$.

Proof. Let $A'_+ := \mathsf{cl}_\varepsilon(A^+_{\max})$ and let $A'_- := \mathsf{op}_\varepsilon(A^-_{\min})$. Notice that by Corollaries 3.24 and 3.30, we have

$$A'_- \subseteq A^-_{\min} \subseteq A^+_{\max} \subseteq A'_+.$$

Steps (2) and (3) on the other hand follow directly from the definitions of $A'_\pm$ as closing and opening. Finally, since $A_{\max}$ (and hence also $A^+_{\max}$) is a maximal solution of (2.12), and by definition $A'_+ \supseteq A^+_{\max}$, it has to hold $A'_+ \sim_\nu A^+_{\max} \sim_\nu A_{\max}$. An analogous argument applies to A'_-. □

REMARK 3.32. In the case where $\mathcal{X} = \mathbb{R}^d$ under the standard Euclidean metric, one can directly conclude some mild regularity of maximal and minimal sets. For example, using the results in [47], one may conclude that the boundaries of the maximal and minimal sets are sets of locally finite classical perimeter of order ε^{-1}; see [2] for a definition of classical perimeter. Similar results were examined in [36]. Furthermore, at any point where curvatures are defined the outer (inner) regularity provides a uniform ε^{-2} upper (lower) bound on the sectional curvatures. However, as manifest in the example in Fig. 4, the maximal and minimal sets need not have boundaries that are even graphs of functions at every point.

The next statement asserts that any intermediate set between the opening and the closing of a minimizer is again a minimizer.

PROPOSITION 3.33. Let $A \in \mathfrak{B}(\mathcal{X})$ be a minimizer of (1.7) and let $B \in \mathfrak{B}(\mathcal{X})$ satisfy

$$\mathsf{op}_\varepsilon(A) \subseteq B \subseteq \mathsf{cl}_\varepsilon(A).$$

Then, B is also a minimizer of (1.7).

Proof. We abbreviate $\hat{A} = \mathsf{cl}_\varepsilon(A)$ and $\tilde{A} = \mathsf{op}_\varepsilon(A)$ and notice that $\tilde{A} \subseteq \hat{A}$. Furthermore, we notice that by the definition of closing and opening, we have

$$(\hat{A})^{-\varepsilon} = (\tilde{A})^{-\varepsilon},$$

which in turn implies that for any set $B \in \mathfrak{B}(\mathcal{X})$ with $\tilde{A} \subseteq B \subseteq \hat{A}$, it holds $B^{-\varepsilon} = (\hat{A})^{-\varepsilon}$. Similarly, we have $\tilde{A}^\varepsilon \subseteq B^\varepsilon \subseteq (\hat{A})^\varepsilon$. We then note that as in the proof of Lemma 3.29

$$\widetilde{R}_\varepsilon(A) = w_0\rho_0(A^\varepsilon) + w_1\rho_1((A^{-\varepsilon})^c).$$

However, given the previous set inclusions, and the fact that thanks to Corollaries 3.30, it holds $\widetilde{R}_\varepsilon(\hat{A}) = \widetilde{R}_\varepsilon(\tilde{A})$, this then gives that $\widetilde{R}_\varepsilon(B) = \widetilde{R}_\varepsilon(\tilde{A})$ or in other words B also minimizes the adversarial risk. □

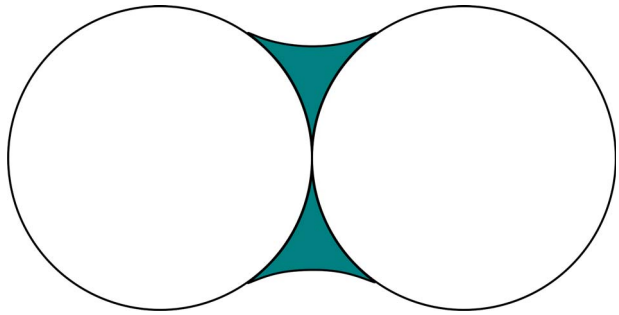


FIG. 5. A set which satisfies $\mathsf{cl}_\varepsilon(\mathsf{op}_\varepsilon(A)) = A$ but is not inner regular.

It might be tempting to think that one could just consider the opening of the closing (or vice versa) of a set to generate a minimizer which is both outer and inner regular. However, this approach fails in general, as the following example shows. Even worse, revisiting Fig. 3 in Example 3.19 shows that in some cases, there is no adversarial minimizer which is both ε inner and outer regular: indeed, a minimizer of that problem can be at most $(\sqrt{2}-1)\varepsilon \approx 0.41\varepsilon$ inner and outer regular. Hence, care must be taken in order to demonstrate the existence of a regular minimizer to the adversarial classification problem.

EXAMPLE 3.34. In this example, we consider the set A given by the union of two balls with radius $\varepsilon > 0$ with two non-convex triangles, as depicted in Fig. 5. The set can be defined as $A := \mathsf{cl}_\varepsilon(B_\varepsilon(-\varepsilon,0) \cup B_\varepsilon(\varepsilon,0))$. It satisfies $\mathsf{op}_\varepsilon(A) = B_\varepsilon(-\varepsilon,0) \cup B_\varepsilon(\varepsilon,0)$, and hence, $\mathsf{cl}_\varepsilon(\mathsf{op}_\varepsilon(A)) = A$. Still, it is not inner regular since the boundary points which are contained in the two triangles do not possess a touching ball with radius ε that is contained in A.

The previous example demonstrates that it is not possible to generate ε-regular sets by solely utilizing the opening and closing of a set. The example shown in Fig. 4 demonstrates that the maximal and minimal sets need not have boundaries that are even locally the graph of a function.

We now proceed to prove our central regularity result, Theorem 3.25. Note that, although generating an inner and outer regular minimizer through morphological operations is not possible, it is plausible that one could construct a minimizer which is more regular (for example, possessing a $C^{1,1}$ boundary), but we leave that question to later work.

Proof. (Proof of Theorem 3.25). Again, we abbreviate $\hat{A} = \mathsf{cl}_\varepsilon(A)$ and $\tilde{A} = \mathsf{op}_\varepsilon(A)$. We notice that $\hat{A}$ is ε outer regular, while $\tilde{A}$ is ε inner regular, and that $\tilde{A} \subseteq \hat{A}$. Thanks to Proposition 3.33 any $B \in \mathfrak{B}(\mathbb{R}^d)$ with $\tilde{A} \subseteq B \subseteq \hat{A}$ is a minimizer, see Fig. 6 for an illustration. We now turn to constructing B with the desired properties.

We recall (cf. [38, Section 13.1] or originally in [40]) that for any open set V, there exists a *regularized (signed) distance function* $d_r \in C^\infty((\partial V)^c)$ satisfying

$$\frac{1}{2} \le \frac{d_r(x)}{d(x,V)} \le \frac{3}{2}, \qquad |\partial^\alpha d_r(x)| \le \frac{c_\alpha}{|d_r(x)|^{|\alpha|-1}}. \tag{3.19}$$

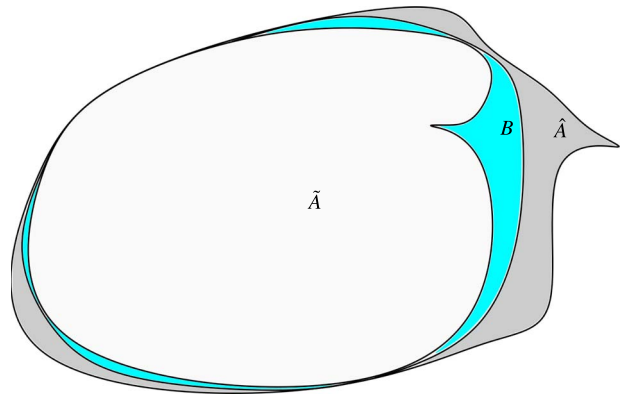


FIG. 6. Nested sets $\tilde{A} \subseteq B \subseteq \hat{A}$ in the proof of Theorem 3.25. $\tilde{A}$ is inner regular and $\hat{A}$ is outer regular. The whole region between $\partial\tilde{A}$ and $\partial\hat{A}$ is $(\tilde{A} \cup \hat{A}^c)^c$ and the boundary of B is by construction a smooth curve contained in that set.

Here, $\bar{d}(x, V)$ is the signed distance with respect to the Euclidean metric, namely

$$\bar{d}(x, V) := \begin{cases} \inf_{y \in V} |x - y| & \text{if } x \in V^c, \\ -\inf_{y \in V^c} |x - y| & \text{if } x \in V. \end{cases}$$

As we will need to check a more detailed property of the function d_r, we briefly give its definition: we let $\phi \in C_c^\infty(\mathbb{R}^d)$ be a non-negative function with support on the unit ball and integral 1. We define

$$G(x, t) = \int_{\mathbb{R}^d} \bar{d}\left(x - t\frac{y}{2}, V\right) \phi(y)\, \mathrm{d}y.$$

One can show that there exists a unique solution to $G(x, t) = t$, and we let d_r be that unique solution, namely $G(x, d_r(x)) = d_r(x)$. Indeed, as proved in [40], this follows from Banach fixed point theorem and the fact that $G(x, \cdot)$ is Lipschitz with Lipschitz constant strictly less than 1.

We let $\tilde{d}_1$ and $\tilde{d}_2$ be regularized distance functions for the sets $\tilde{A}$ and $\hat{A}^c$, respectively. By considering the function $\tilde{d}_1/\tilde{d}_2$, we may use Sard's theorem, which applies as the function is C^∞, to find a $\kappa \in [1/2, 2]$ so that κ is a regular value of $\tilde{d}_1/\tilde{d}_2$ on the set $(\tilde{A} \cup \hat{A}^c)^c$. Here, we recall that a regular value of a function is one so that the gradient does not vanish on the entire set $\{x \in \mathbb{R}^d : \tilde{d}_1(x)/\tilde{d}_2(x) = \kappa\}$. Our candidate set B will now be $B := \{x \in \mathbb{R}^d : \tilde{d}_1(x) \leq \kappa \tilde{d}_2(x)\}$. Due to the first part of (3.19), we know that the signs of the original distance functions to the sets $\tilde{A}$, $\hat{A}^c$ and the signs of their regularized versions coincide. From this observation, it is now straightforward to see that $\tilde{A} \subseteq B \subseteq \hat{A}$, and hence, B is a minimizer of the adversarial problem according to Proposition 3.33: thanks to the fact that κ is a regular value, anywhere in the interior of $(\tilde{A} \cup \hat{A}^c)^c$, we may express the boundary of B as the graph of a C^∞ function. In light of the main result in [39], we also have that this set is locally the graph of a $C^{1,1}$ function away from $(\tilde{A} \cup \hat{A}^c)^c$. Thus, it only remains to check the regularity up to the boundary points of $(\tilde{A} \cup \hat{A}^c)^c$.

To this end, we need to establish regularity estimates on $\nabla\tilde{d}_1$ and $\nabla\tilde{d}_2$ which hold uniformly at points on the boundary of B near where the boundaries of $\tilde{A}$ and $\hat{A}$ coincide. To begin, we notice that

$$\partial_i d_r(x) = \frac{\int_{\mathbb{R}^d} \partial_i \bar{d}\left(x - d_r(x)\frac{y}{2}, V\right)\phi(y)\,\mathrm{d}y}{1 + \frac{1}{2}\int_{\mathbb{R}^d} \partial_i \bar{d}\left(x - d_r(x)\frac{y}{2}, V\right)\phi(y)\,\mathrm{d}y}.$$

Using the fact that $|\nabla\bar{d}| \leq 1$ a.e. and that $z \mapsto \frac{z}{1+z/2}$ is uniformly Lipschitz on $[-1,\infty)$, it then suffices to estimate the continuity of $x \mapsto \int_{\mathbb{R}^d} \nabla\bar{d}(x - d_r(x)\frac{y}{2}, V)\phi(y)\,\mathrm{d}y$. To this end, let us consider x_1, x_2 in the set where $1/2 < \frac{\tilde{d}_1}{\tilde{d}_2} < 2$. For such points, let us denote

$$D(x_1,x_2) = \min(\tilde{d}_1(x_1), \tilde{d}_2(x_1), \tilde{d}_1(x_2), \tilde{d}_2(x_2)).$$

We notice that, by the choice of x_1, x_2, it holds that $D(x_1,x_2) \geq 0$ and

$$2D(x_1,x_2) \geq \max(\tilde{d}_1(x_1), \tilde{d}_2(x_1), \tilde{d}_1(x_2), \tilde{d}_2(x_2)). \tag{3.20}$$

We consider separately two cases. First, if

$$\varepsilon^\alpha |x_1 - x_2|^\alpha \leq D(x_1,x_2),$$

we then use the classical estimate (3.19) to show that

$$|\nabla d_r(x_1) - \nabla d_r(x_2)| \leq C\frac{|x_1 - x_2|}{\min(\tilde{d}_1(x_1), \tilde{d}_2(x_1), \tilde{d}_1(x_2), \tilde{d}_2(x_2))} \leq C\varepsilon^{-\alpha}|x_1 - x_2|^{1-\alpha}.$$

On the other hand, for the opposite case where

$$\varepsilon^\alpha |x_1 - x_2|^\alpha \geq D(x_1,x_2),$$

by using the estimate from Lemma 3.35 below along with equation (3.20) and the fact that ϕ has compact support we then may deduce that

$$\begin{aligned}
|\nabla d_r(x_1) - \nabla d_r(x_2)| &\leq C\left|\int_{\mathbb{R}^d} \nabla\bar{d}\left(x_1 - d_r(x_1)\frac{y}{2}, V\right)\phi(y)\,\mathrm{d}y - \int_{\mathbb{R}^d} \nabla\bar{d}\left(x_2 - d_r(x_2)\frac{y}{2}, V\right)\phi(y)\,\mathrm{d}y\right| \\
&\leq C\int_{\mathbb{R}^d} \varepsilon^{-1}\left(\left|x_1 - d_r(x_1)\frac{y}{2} - x_2 + d_r(x_2)\frac{y}{2}\right| + 2D(x_1,x_2)\right)\phi(y)\,\mathrm{d}y \\
&\leq C\varepsilon^{-1}\left(|x_1 - x_2| + \sqrt{D(x_1,x_2)}\right) \\
&\leq C\varepsilon^{-1}|x_1 - x_2| + C\varepsilon^{\alpha/2-1}|x_1 - x_2|^{\alpha/2}.
\end{aligned}$$

Setting $\alpha = 2/3$ and applying the result to the regularized distance functions $\tilde{d}_1, \tilde{d}_2$ then establishes the fact that the function defining ∂B is uniformly $C^{1,1/3}$, even up to the boundary, concluding the proof. □

We notice that in the previous proof, the dependence on ε in the continuity estimates near the boundary is explicit and improves as ε increases. This intuitively makes sense, and although our current estimates in the 'interior' of our bad region do not give explicit dependence upon ε, it seems plausible that the dependence on ε should be good.

We now give the central geometric lemma used in the proof of Theorem 3.25.

LEMMA 3.35. Let $\tilde{A} \subseteq \mathbb{R}^d$ be ε inner regular, let $\hat{A} \subseteq \mathbb{R}^d$ be ε outer regular and let $\tilde{A} \subseteq \hat{A}$. Let $x, y \in \hat{A} \backslash \tilde{A}$, and let both be points of differentiability of the distance function from both $\tilde{A}$ and $\hat{A}$. Define

$$D(x,y) := \max(d(x,\tilde{A}), d(x,\hat{A}), d(y,\hat{A}), d(y,\tilde{A})).$$

Then

$$|\nabla d(x,\tilde{A}) - \nabla d(y,\tilde{A})| \leq C\varepsilon^{-1}\left(|x-y| + \sqrt{D(x,y)}\right),$$

at any points x, y where the distance is differentiable (which holds a.e. by Rademacher's theorem).

Proof. This lemma is a direct extension of the work in [39] to our setting, and this proof expands upon the four ball lemma given therein. We recall that the gradient of the distance function is given by the unit vector pointing away from the closest point in the set. Let $u_x = \nabla d(x,\tilde{A})$, $u_y = \nabla d(y,\tilde{A})$, $v_x = \nabla d(x,\hat{A})$ and $v_y = \nabla d(y,\hat{A})$. Let $\tilde{x}, \tilde{y}$ be the closest points in $\tilde{A}$ to x and y, and let $\hat{x}, \hat{y}$ be the closest points in $\hat{A}$ to x and y. By the regularity conditions, we know that there are four balls

$$\tilde{B}_x = B_\varepsilon(\tilde{x} - \varepsilon u_x), \qquad \tilde{B}_y = B_\varepsilon(\tilde{y} - \varepsilon u_y), \qquad \hat{B}_x = B_\varepsilon(\hat{x} - \varepsilon v_x), \qquad \hat{B}_y = B_\varepsilon(\hat{y} - \varepsilon v_y),$$

which satisfy $\tilde{B}_i \cap \hat{B}_j = \emptyset$ for any $i,j \in \{x,y\}$, and so that $\tilde{x}, \tilde{y}$ do not belong to either $\tilde{B}_x$ or $\tilde{B}_y$. We also notice that $|\tilde{x} - \hat{x}| \leq 2\max(d(x,\tilde{A}), d(x,\hat{A}))$.

We then choose the smallest positive values of $\tilde{\delta}_x$ and $\tilde{\delta}_y$ so that boundaries of the dilations $(\tilde{B}_x)^{\tilde{\delta}_x}$ and $(\tilde{B}_y)^{\tilde{\delta}_y}$ touch the boundaries of either $\hat{B}_x$ or $\hat{B}_y$ at exactly one point. From here on, we will assume that $(\tilde{B}_x)^{\tilde{\delta}_x}$ touches $\hat{B}_x$ and $(\tilde{B}_y)^{\tilde{\delta}_y}$ touches $\hat{B}_y$, as the other cases may be handled analogously. Call the points where those boundaries coincide $\bar{x}$ and $\bar{y}$. We note that clearly $\tilde{\delta}_x \leq 2\max(d(x,\tilde{A}), d(x,\hat{A})$ and $\tilde{\delta}_y \leq 2\max(d(y,\tilde{A}), d(y,\hat{A})$.

We may directly apply the four ball lemma from [39] to conclude that the unit vectors $\bar{u}_x, \bar{u}_y$ from the center of each ball to $\bar{x}$ and $\bar{y}$ satisfy

$$|\bar{u}_x - \bar{u}_y| \leq C\varepsilon^{-1}|\bar{x} - \bar{y}|. \tag{3.21}$$

It then remains only to bound the difference between the 'bar' variables and the original ones.

Let $\bar{u}_x$ be the unit vector pointing from the center of $(\tilde{B}_x)^{\tilde{\delta}_x}$ to $\bar{x}$. We then may compute

$$\cos(\theta(\bar{u}_x, u_x))(\varepsilon + d(x,\tilde{A})) + \cos(\theta(-\bar{u}_x, v_x)(\varepsilon + d(x,\hat{A})) = 2(\varepsilon + \tilde{\delta}_x),$$

where we are letting $\theta(\cdot,\cdot)$ denote the angle between the vectors. We may conclude that $\theta(-\bar{u}_x, v_x)$ and $\theta(\bar{u}_x, u_x)$ are bounded by a constant times $\sqrt{D(x,y)}$. By using the law of cosines, we may compute that

$|u_x - \bar{u}_x| < C\varepsilon^{-1}\sqrt{D(x,y)}$ and that $|\bar{x} - x| < C\sqrt{D(x,y)}$. Using the triangle inequality and combining with (3.21) then concludes the proof. □

4. Other adversarial models

We finish the paper with a couple of generalizations and a discussion on similar adversarial models, some of which also give rise to $L^1 + \mathrm{TV}$ problems. In this section, we keep the discussion rather formal in order to not distract from the main messages that we want to convey. We also do not make any attempt to interpret or expound upon these models: the goal is simply to identify alternative adversarial models which have analogous variational forms.

4.1 *Regression problems*

Instead of studying binary classification, one can also study adversarial regression problems of the form

$$\inf_{u \in L^1(\mathcal{X};\rho)} \mathbb{E}_{(x,y)\sim\mu} \left[\sup_{\tilde{x} \in B_\varepsilon(x)} |u(\tilde{x}) - y| \right], \tag{4.1}$$

where y can now take any real value. Subtracting the empirical risk, one can easily show that this problem can be reformulated as

$$\inf_{u \in L^1(\mathcal{X};\rho)} \mathbb{E}_{(x,y)\sim\mu} \left[|u(x) - y| \right] + \varepsilon \, \widetilde{\mathrm{TV}}_\varepsilon(u;\mu), \tag{4.2}$$

where the total variation is now given by

$$\begin{aligned} \widetilde{\mathrm{TV}}_\varepsilon(u;\mu) &= \frac{1}{\varepsilon} \iint_{\mathcal{X}\times\mathcal{Y}} \sup_{\tilde{x} \in B_\varepsilon(x)} \left[|u(\tilde{x}) - y| - |u(x) - y| \right] \mathrm{d}\mu(x,y) \\ &= \frac{1}{\varepsilon} \int_{\mathcal{X}} \int_{\pi_1^{-1}(x)} \sup_{\tilde{x} \in B_\varepsilon(\xi)} \left[|u(\tilde{x}) - y| - |u(\xi) - y| \right] \mathrm{d}\mu_x(\xi, y) \, \mathrm{d}\rho(x). \end{aligned} \tag{4.3}$$

In the disintegrated formulation, $\pi_1 : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ denotes the projection onto the first factor, $\rho := (\pi_1)_\sharp \mu$ is the first marginal of μ, and $(\mu_x)_{x\in\mathcal{X}} \subseteq \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is a family of disintegrations of μ.

As before, one has to define an essential version of this total variation to have a well-defined functional and introduce the measure ν as in Section 3. The analysis performed that there can be generalized to this regression setting; however, in the regression context, the interpretation of the associated perimeter is not obvious. Indeed, (4.3) is a highly data-dependent convex regularization functional. This provides theoretical motivation for recent work which studies data-driven convex regularizers for solving inverse problems; see, e.g. [46]. To make this connection a bit clearer, we assume for simplicity that the data y are given by $f(x)$. In this case, (4.2) reduces to the simpler formula

$$\inf_{u \in L^1(\mathcal{X};\rho)} \left\{ \int_{\mathcal{X}} |u(x) - f(x)| \, \mathrm{d}\rho(x) + \int_{\mathcal{X}} \sup_{\tilde{x} \in B_\varepsilon(x)} \left[|u(\tilde{x}) - f(x)| - |u(x) - f(x)| \right] \mathrm{d}\rho(x) \right\}. \tag{4.4}$$

4.2 *Random perturbation*

Let us consider random perturbations, i.e.

$$\inf_{A\in\mathfrak{B}(\mathcal{X})} \mathbb{E}_{(x,y)\sim\mu}\left[\mathbb{E}_{\tilde{x}\sim\nu_x}\left[\left|1_A(\tilde{x})-y\right|\right]\right], \tag{4.5}$$

where 'nature' chooses $\tilde{x}$ randomly following the law of a family of probability measures $(\nu_x)_{x\in\mathcal{X}}$. One natural candidate for such ν_x would be associated with a random walk [43]. Since in this setting, there is no adversarial attack in the game-theoretic sense, which would involve some sort of min-max structure, one cannot rewrite this problem as a variational regularization problem. Indeed, subtracting the empirical risk $\mathbb{E}_{(x,y)\sim\mu}[|1_A(x)-y|]$ from the objective does not yield a non-negative term.

However, in the case that $\mathcal{X}$ is a vector space, we can use the law of total expectation (disintegration) and a change of variables to obtain

$$\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mu}\left[\mathbb{E}_{\tilde{x}\sim\nu_x}\left[\left|1_A(\tilde{x})-y\right|\right]\right]\\
&= w_0\int_{\mathcal{X}}\int_{\mathcal{X}} 1_A(\tilde{x})\,\mathrm{d}\nu_x(\tilde{x})\,\mathrm{d}\rho_0(x) + w_1\int_{\mathcal{X}}\int_{\mathcal{X}} 1_{A^c}(\tilde{x})\,\mathrm{d}\nu_x(\tilde{x})\,\mathrm{d}\rho_1(x)\\
&= w_0\int_{\mathcal{X}}\int_{\mathcal{X}} 1_A(x+\tilde{x})\,\mathrm{d}(T_x)_\sharp\nu_x(\tilde{x})\,\mathrm{d}\rho_0(x) + w_1\int_{\mathcal{X}}\int_{\mathcal{X}} 1_{A^c}(x+\tilde{x})\,\mathrm{d}(T_x)_\sharp\nu_x(\tilde{x})\,\mathrm{d}\rho_1(x),
\end{aligned}$$

where $T_x:\mathcal{X}\to\mathcal{X}$ is defined by $T_x(\tilde{x})=\tilde{x}-x$. If the push forward measure $(T_x)_\sharp\nu_x$ does not depend on x, which is the case, e.g. whenever $\nu_x := (T_{-x})_\sharp\nu$ for some measure ν, we can abbreviate it by ν, and we can rewrite this as

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

where the measure $\tilde{\mu}$ has the marginals $(\pi_1)_\sharp\tilde{\mu}=\nu\star\rho$ and $(\pi_2)_\sharp\tilde{\mu}=(\pi_2)_\sharp\mu$.

Hence, the random perturbation in (4.5) does not actually lead to an adversarial model, but rather, it replaces the data distribution ρ in the space $\mathcal{X}$ with the convolution $\nu\star\rho$ and likewise changes the conditionals ρ_0 and ρ_1 to $\nu\star\rho_0$ and $\nu\star\rho_1$. We note that this structure is still similar to the form of $\tilde{R}_\varepsilon$ shown in the proof of Lemma 3.29. Problems with similar structure have also been considered in the context of decentralized optimal control [61].

4.3 *Random perturbation with adversarial decision*

A similar model to the one from the previous section, however containing an adversarial action, is the following:

$$\inf_{A\in\mathfrak{B}(\mathcal{X})} \mathbb{E}_{(x,y)\sim\mu}\left[\mathbb{E}_{\xi\sim\nu_{x,\varepsilon}}\left[\max_{\tilde{x}\in\{\xi,x\}}\left|1_A(\tilde{x})-y\right|\right]\right]. \tag{4.6}$$

In words, in the above model, 'nature' randomly draws a point ξ according to a probability measure (e.g. determined by a random walk) $\nu_{x,\varepsilon}$, which can depend on x and a parameter $\varepsilon>0$, and the adversary can either use this proposed perturbation or reject it. The adversary's decision is, of course, based on whether the randomly chosen point ξ creates a larger loss than the attacked point x. For example, if the

attacked point x lies in A and this point should have the label 1, the adversary can do nothing if it draws another point in A. Only if $\nu_{x,\varepsilon}$ draws a point outside of A will the adversary accept it. This adversarial model is reminiscent of [37], where mean curvature flow is obtained as a limit of a game theoretical problem in which an adversary chooses between alternatives.

As it turns out, problem (4.6) can be rewritten as

$$\inf_{A\in\mathfrak{B}(\mathcal{X})} \mathbb{E}_{(x,y)\sim\mu}\left[\left|1_A(x)-y\right|\right] + \varepsilon\,\widehat{\mathrm{TV}}_\varepsilon(1_A), \tag{4.7}$$

where the total variation functional $\widehat{\mathrm{TV}}_\varepsilon$ takes the form

$$\begin{aligned}\widehat{\mathrm{TV}}_\varepsilon(u) := \frac{w_0}{\varepsilon}\int_{\mathcal{X}}\int_{\mathcal{X}} (u(\tilde{x})-u(x))_+ \,\mathrm{d}\nu_{x,\varepsilon}(\tilde{x})\,\mathrm{d}\rho_0(x)\\ + \frac{w_1}{\varepsilon}\int_{\mathcal{X}}\int_{\mathcal{X}} (u(x)-u(\tilde{x}))_+ \,\mathrm{d}\nu_{x,\varepsilon}(\tilde{x})\,\mathrm{d}\rho_1(x).\end{aligned} \tag{4.8}$$

Since here the adversarial decision takes place on the finite set $\{\xi, x\}$, this problem is much easier to analyze and does not require a redefinition using an essential total variation or perimeter. For instance, if $\nu_{x,\varepsilon} \ll \rho$ for all $x \in \mathcal{X}$, then one can simply work on $L^\infty(\mathcal{X};\rho)$. Indeed, this is a highly relevant case as the following example shows.

EXAMPLE 4.1. If $\rho_0 = \rho_1$, $w_0 = w_1$, then this reduces to the following nonlocal total variation energy:

$$\frac{1}{\varepsilon}\int_{\mathcal{X}}\int_{\mathcal{X}} |u(x)-u(\tilde{x})|\,\mathrm{d}\nu_{x,\varepsilon}(\tilde{x})\,\mathrm{d}\rho(x),$$

whose properties and associated gradient flow have been analyzed in the framework of metric random walk spaces [43]. This nonlocal total variation functional has furthermore been extensively applied in image processing, see [31, 64].

If we assume that the random walk $\nu_{x,\varepsilon}$ has the special structure $\frac{\mathrm{d}\nu_{x,\varepsilon}}{\mathrm{d}\rho}(\tilde{x}) = \eta_\varepsilon(x-\tilde{x})$ for some function $\eta_\varepsilon : \mathcal{X} \to \mathbb{R}$, we obtain

$$\frac{1}{\varepsilon}\int_{\mathcal{X}}\int_{\mathcal{X}} \eta_\varepsilon(x-\tilde{x})|u(x)-u(\tilde{x})|\,\mathrm{d}\rho(\tilde{x})\,\mathrm{d}\rho(x). \tag{4.9}$$

For the special case when $\rho = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ is an empirical measure, (4.9) reduces to the graph total variation

$$\frac{1}{\varepsilon}\sum_{i,j=1}^N \eta_\varepsilon(x_i-x_j)|u(x_i)-u(x_j)|. \tag{4.10}$$

Total variations of these forms and their limits as $\varepsilon \to 0$ have been intensively analyzed in the context of graph-based clustering methods and trend filtering, see, e.g. [28–30]. Typical choices for η_ε are $\eta_\varepsilon(z) = \frac{1}{\varepsilon^d}1_{B_\varepsilon(x)}(z)$ or $\eta_\varepsilon(z) = \frac{1}{\varepsilon^d}\exp(-|z/\varepsilon|^2)$.

4.4 *General loss functions*

One can also study adversarial problems with a more general loss function. The baseline model for this endeavour is the following generalization of (1.7):

$$\inf_{A\in\mathfrak{B}(\mathcal{X})} \mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B_\varepsilon(x)} \ell(1_A(\tilde{x}),y)\right]. \tag{4.11}$$

Subtracting the empirical risk, we can decompose the adversarial risk as

$$\mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B_\varepsilon(x)} \ell(1_A(\tilde{x}),y)\right] = \mathbb{E}_{(x,y)\sim\mu}\left[\ell(1_A(x),y)\right] + \varepsilon\,\widetilde{\mathrm{TV}}_\varepsilon(1_A;\mu), \tag{4.12}$$

where the total variation is given by

$$\begin{aligned}\widetilde{\mathrm{TV}}_\varepsilon(u;\mu) = \frac{w_0}{\varepsilon}\int_{\mathcal{X}} \sup_{\tilde{x}\in B_\varepsilon(x)} \ell(u(\tilde{x}),0) - \ell(u(x),0)\,\mathrm{d}\rho_0(x)\\ + \frac{w_1}{\varepsilon}\int_{\mathcal{X}} \sup_{\tilde{x}\in B_\varepsilon(x)} \ell(u(\tilde{x}),1) - \ell(u(x),1)\,\mathrm{d}\rho_1(x).\end{aligned} \tag{4.13}$$

For instance, for the cross entropy loss $\ell(u,y) = -y\log u - (1-y)\log(1-u)$, this simplifies to

$$\begin{aligned}\widetilde{\mathrm{TV}}_\varepsilon(u;\mu) = \frac{w_0}{\varepsilon}\int_{\mathcal{X}} \log(1-u(x)) - \inf_{\tilde{x}\in B_\varepsilon(x)} \log(1-u(\tilde{x}))\,\mathrm{d}\rho_0(x)\\ + \frac{w_1}{\varepsilon}\int_{\mathcal{X}} \log u(x) - \inf_{\tilde{x}\in B_\varepsilon(x)} \log u(\tilde{x})\,\mathrm{d}\rho_1(x), \quad 0\le u\le 1,\end{aligned} \tag{4.14}$$

which is similar to our total variation (3.1) applied to $\log u$ instead of u. Notice that in general problem (4.11) and its relaxation to functions $0 \le u \le 1$ may not coincide, as the cross entropy example suggests. For general loss functions (4.13) is more difficult to interpret: this is a primary reason that we restricted our analysis to the case $\ell(u,y) = |u-y|$.

5. Conclusions

In this paper, we have studied adversarial training problems in a variety of non-parametric settings and have established an equivalence with regularized risk minimization problems. The regularization terms in these risk minimization problems are explicitly characterized and correspond to a type of nonlocal perimeter/total variation. Our work provides new conceptual insights for adversarial training problems and introduces new mathematical tools for their quantitative analysis. In particular, we have used tools from the calculus of variations to rigorously prove the existence of solutions, we have identified a convex structure of the problem that allows us to introduce appropriate notions of maximal and minimal solutions and in turn introduced a convenient notion of uniqueness of solutions, and finally, we have presented a collection of results on the existence of regular solutions to the original adversarial training problem.

Some research directions that stem from this work include: (1) the extension of the analysis presented in this work to multi-label classification settings, (2) investigating a sharper analysis of the regularity properties of solutions to adversarial training problems in both the Euclidean setting, as well as for more general distance functions and spaces.

In addition, as already discussed in the introduction, part of the motivation for this work came from the work [59] where one of the main objectives was to study the regularization effect of adversarial training on the decision boundaries of optimal robust classifiers (starting with the Bayes classifier at $\varepsilon = 0$). The structure of solutions studied in the present paper allows us to make the line of work initiated in [59] more concrete and to approach it with a larger set of mathematical tools at hand. In particular, this work raises the question of whether it is possible to track maximal (minimal) solutions to adversarial training problems as ε grows from 0 to infinity. In words, we are interested in defining a suitable notion of *solution path* $(A_\varepsilon)_{\varepsilon>0}$ for the family of adversarial training problems (1.7). The study of solution paths, in particular their algorithmic use and regularity, has quite some tradition in the field of variational regularization methods, see, e.g. [10, 52, 58], but in the context of adversarial training less is known about their properties. Notice that one important difference with the standard regularization setting is that the equivalent regularization formulation of (1.7) has a regularization functional that changes with the regularization parameter ε. This feature makes the analysis more challenging.

It is also interesting to consider the asymptotics as $\varepsilon \downarrow 0$. In the special case where $(\mathcal{X}, \mathsf{d})$ is $\mathbb{R}^d$ with the Euclidean metric and $w_i\rho_i$ is replaced by $\mathcal{L}^d$, the functionals $\mathrm{Per}_\varepsilon(\cdot;\mu)$ are known to Γ-converge to the classical perimeter as $\varepsilon \downarrow 0$ [18]. In our more general setting, it is particularly interesting to investigate which information of the measures ρ_0 and ρ_1 'survives' in the limit as $\varepsilon \downarrow 0$ and whether a Γ-convergence result can be proven.

Finally, in this regime, the proof of our regularity result Theorem 3.25 deteriorates and one can at most expect a set of finite perimeter.

Data Availability Statement

No new data were generated or analysed in support of this research.

Acknowledgements

The authors would like to thank Antonin Chambolle, Matt Jacobs, Meyer Scetbon, Simone Di Marino and Khai Nguyen for enlightening discussions and for sharing useful references. This work was done while LB and NGT were visiting the Simons Institute for the Theory of Computing to participate in the program 'Geometric Methods in Optimization and Sampling' during the Fall of 2021 and LB and NGT are very grateful for the hospitality of the institute.

Funding

Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - GZ 2047/1, Projekt-ID 390685813 to LB; NSF-DMS grant 2005797 to NGT; IFDS at UW-Madison and NSF through TRIPODS grant 2023239 to NGT.

References

1. Ambrosio, L., Di Marino & Gigli, N. (2017) Perimeter as relaxed Minkowski content in metric measure spaces. *Nonlinear Anal. Real World Appl.*, **153**, 78–88.

2. Ambrosio, L., Fusco, N. & Pallara, D. (2000) *Functions of bounded variation and free discontinuity problems*.Courier Corporation.
3. Awasthi, P., Frank, N. S. & Mohri, M. (2021) *On the existence of the adversarial Bayes classifier (extended version)*.
4. Belloni, A., Chernozhukov, V. & Wang, L. (2011) Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, **98**, 791–806.
5. Benning, M. & Burger, M. (2018) Modern regularization methods for inverse problems. *Acta Numer.*, **27**, 1–111.
6. Bhagoji, A. N., Cullina, D. & Mittal, P. (2019) Lower bounds on adversarial robustness from optimal transport. *Advances in Neural Information Processing Systems*, vol. **32**. (H. Wallach, H. Larochelle, A. Beygelzimer, F. d.' Alché-Buc, E. Fox & R. Garnett eds).Curran Associates, Inc.
7. Blanchet, J., Kang, Y. & Murthy, K. (2019) Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probab.*, **56**, 830–857.
8. Blanchet, J., Murthy, K. & Nguyen, V. A. (2021) Statistical analysis of Wasserstein distributionally robust estimators. *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*.INFORMS, pp. 227–254.
9. Vladimir, I. (2007) Bogachev. *Measure Theory*.Springer Berlin Heidelberg.
10. Bungert, L. & Burger, M. (2019) Solution paths of variational regularization methods for inverse problems. *Inverse Probl.*, **35**, 105012.
11. Bungert, L., Raab, R., Roith, T., Schwinn, L. & Tenbrinck, D. (2021) CLIP: Cheap Lipschitz training of neural networks. *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer, pp. 307–319.
12. Burger, M. & Osher, S. (2013) A guide to the TV zoo. *Level set and PDE based reconstruction methods in imaging*.Springer, pp. 1–70.
13. Caffarelli, L. (2012) Non-local diffusions, drifts and games. *Nonlinear partial differential equations*.Springer, pp. 37–52.
14. Cesaroni, A., Dipierro, S., Novaga, M. & Valdinoci, E. (2018) Minimizers for nonlocal perimeters of Minkowski type. *Calc. Var.*, **57**, 1–40.
15. Cesaroni, A. & Novaga, M. (2017) Isoperimetric problems for a nonlocal perimeter of Minkowski type. *Geometric Flows*, **2**, 86–93.
16. Chambolle, A., Novaga, M., Cremers, D. & Pock, T. (2010) An introduction to total variation for image analysis. *Theoretical Foundations and Numerical Methods for Sparse Recovery*.De Gruyter.
17. Chambolle, A., Giacomini, A. & Lussardi, L. (2010) Continuous limits of discrete perimeters. *ESAIM: Math. Model. Numer. Anal.*, **44**, 207–230.
18. Chambolle, A., Lisini, S. & Lussardi, L. (2014) A remark on the anisotropic outer Minkowski content. *Adv. Calc. Var.*, **7**, 241–266.
19. Chambolle, A., Morini, M. & Ponsiglione, M. (2012) A nonlocal mean curvature flow and its semi-implicit time-discrete approximation. *SIAM J. Math. Anal.*, **44**, 4048–4077.
20. Chambolle, A., Morini, M. & Ponsiglione, M. (2015) Nonlocal curvature flows. *Arch. Ration. Mech. Anal.*, **218**, 1263–1329.
21. Chan, T. F. & Esedoglu, S. (2005) Aspects of total variation regularized L^1 function approximation. *SIAM J. Appl. Math.*, **65**, 1817–1837.
22. Chen, R. & Paschalidis, I. C. (2020) Distributionally robust learning. *Found. Trends Mach. Learn.*, **4**, 1–243.
23. Darbon, J. (2005) Total variation minimization with L^1 data fidelity as a contrast invariant filter. *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005*. IEEE, pp. 221–226.
24. Duval, V., Aujol, J.-F. & Gousseau, Y. (2009) The TVL1 model: a geometric point of view. *Multiscale Model. Simul.*, **8**, 154–189.

25. Finlay, C. & Oberman, A. M. (2021) Scaleable input gradient regularization for adversarial robustness. *Mach. Learn. Appl.*, **3**, 100017.
26. Finlay, C., Oberman, A. M. & Abbasi, B. (2018) *Improved robustness to adversarial examples using Lipschitz regularization of the loss*.
27. García Trillos, C. A. & García Trillos, N. (2022) On the regularized risk of distributionally robust learning over deep neural networks. *Res. Math. Sci.*, **9**, 1–32.
28. García Trillos, N. & Murray, R. (2017) A new analytical approach to consistency and overfitting in regularized empirical risk minimization. *Eur. J. Appl. Math.*, **28**, 886–921.
29. García Trillos, N. & Slepčev, D. (2016) Continuum limit of total variation on point clouds. *Arch. Ration. Mech. Anal.*, **220**, 193–241.
30. García Trillos, N., Slepčev, D., Von Brecht, J., Laurent, T. & Bresson, X. (2016) Consistency of Cheeger and ratio graph cuts. *J. Mach. Learn. Res.*, **17**, 6268–6313.
31. Gilboa, G. & Osher, S. (2009) Nonlocal operators with applications to image processing. *Multiscale Model. Simul.*, **7**, 1005–1028.
32. Goodfellow, I., Shlens, J., & Szegedy, C. (2015) Explaining and harnessing adversarial examples. *International Conference on Learning Representations*.
33. Haralock, R. M. & Shapiro, L. G. (1991) *Computer and robot vision*.Addison-Wesley Longman Publishing Co., Inc.
34. Heinonen, J., Koskela, P., Shanmugalingam, N. & Tyson, J. T. (2015) Sobolev spaces on metric measure spaces: an approach based on upper gradients. *New Mathematical Monographs*.Cambridge University Press.
35. Hofmann, T., Schölkopf, B. & Smola, A. J. (2008) Kernel methods in machine learning. *Ann. Stat.*, **36**, 1171–1220.
36. Jog, V. (2021) Reverse Euclidean and Gaussian isoperimetric inequalities for parallel sets with applications. *IEEE Trans. Inf. Theory*, **67**, 6368–6383.
37. Kohn, R. & Serfaty, S. (2006) A deterministic-control-based approach motion by curvature. *Commun. Pure Appl. Math.*, **59**, 344–407.
38. Leoni, G. (2017) *A first course in Sobolev spaces*.American Mathematical Soc.
39. Lewicka, M. & Peres, Y. (2020) Which domains have two-sided supporting unit spheres at every boundary point? *Expo. Math.*, **38**, 548–558.
40. Lieberman, G. (1985) Regularized distance and its applications. *Pac. J. Math.*, **117**, 329–352.
41. Lyu, C., Huang, K. & Liang, H.-N. (2015) A unified gradient regularization family for adversarial examples. *2015 IEEE International Conference on Data Mining*. pp. 301–309.
42. Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. (2019) Towards deep learning models resistant to adversarial attacks.
43. Mazón, J. M., Solera, M. & Toledo, J. (2020) The total variation flow in metric random walk spaces. *Calc. Var.*, **59**, 1–64.
44. Meunier, L., Scetbon, M., Pinot, R. B., Atif, J. & Chevaleyre, Y. (2021) Mixed Nash equilibria in the adversarial examples game. *Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research* (M. Meila and T. Zhang eds). PMLR, pp. 7677–7687.
45. Moosavi-Dezfooli, S.-M., Fawzi, A., Uesato, J., & Frossard, P. (2019) Robustness via curvature regularization, and vice versa. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 9070–9078.
46. Mukherjee, S., Dittmer, S., Shumaylov, Z., Lunz, S., Öktem, O. & Schönlieb, C.-B. (2020) *Learned convex regularizers for inverse problems*.
47. Ya Oleksiv, I. & Pesin, N. I. (1985) Finiteness of Hausdorff measure of level sets of bounded subsets of Euclidean space. *Mathematical notes of the Academy of Sciences of the USSR*, **37**, 237–242.
48. Peres, Y. & Sheffield, S. (2008) Tug-of-war with noise: a game-theoretic view of the p-Laplacian. *Duke Math. J.*, **145**, 91–120.
49. Petersen, P. & Voigtlaender, F. (2018) Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Netw.*, **108**, 296–330.

50. Pydi, M. S. & Jog, V. (2020) Adversarial risk via optimal transport and optimal couplings. *International Conference on Machine Learning*. PMLR, pp. 7814–7823.
51. Ross, A. S. & Doshi-Velez, F. (2018) Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.
52. Rosset, S. & Zhu, J. (2007) Piecewise linear regularized solution paths. *Ann. Stat.*, **35**, 1012–1030.
53. Roth, K., Lucchi, A., Nowozin, S. & Hofmann, T. (2018) Adversarially robust training through structured gradient regularization.
54. Serra, J. (1986) Introduction to mathematical morphology. *Comput. graph. image process.*, **35**, 283–305.
55. Pydi, M. S. & Jog, V. (2021) The many faces of adversarial risk. *Advances in Neural Information Processing Systems 34 pre-proceedings (NeurIPS)*.
56. Stuart, A. M. (2010) Inverse problems: a Bayesian perspective. *Acta Numer.*, **19**, 451–559.
57. Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., B: Stat. Methodol.*, **58**, 267–288.
58. Tibshirani, R. J. & Taylor, J. (2011) The solution path of the generalized lasso. *Ann. Stat.*, **39**, 1335–1371.
59. García Trillos, N. G. & Murray, R. (2022) Adversarial classification: necessary conditions and geometric flows. *J. Mach. Learn. Res.*, **23**, 1–38.
60. Wald, A. (1945) Statistical decision functions which minimize the maximum risk. *Ann. Math.*, **46**, 265–280.
61. Witsenhausen, H. S. (1968) A counterexample in stochastic optimum control. *SIAM J. Control*, **6**, 131–147.
62. Yeats, E. C., Chen, Y. & Li, H. (2021) Improving gradient regularization using complex-valued neural networks. *Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research* (M. Meila & T. Zhang eds). PMLR, pp. 11953–11963.
63. Zeune, L., van Gils, S. A., Terstappen, L. WMM. & Brune,C. (2017) Combining contrast invariant L^1 data fidelities with nonlinear spectral image decomposition. *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer, pp. 80–93.
64. Zhang, X., Burger, M., Bresson, X. & Osher, S. (2010) Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. *SIAM J. Imaging Sci.*, **3**, 253–276.

A. Technical definitions

In this section, we provide various technical definitions used throughout this work.

Definition A.1. (Hölder spaces and sets). Let $U \subseteq \mathbb{R}^d$ be an open subset of $\mathbb{R}^d$. For $0 < \alpha \leq 1$ a function $f : U \to \mathbb{R}$ is called α-Hölder continuous if

$$\sup \left\{ \frac{|f(x) - f(y)|}{|x - y|^{\alpha}} \ : \ x, y \in U, x \neq y \right\} < \infty.$$

For $k \in \mathbb{N}$, a function $f : U \to \mathbb{R}$ is said to belong to $C^{k,\alpha}(U)$ if it is k times differentiable on U and its kth derivatives are α-Hölder continuous on U. We say that an open subset of $\mathbb{R}^d$ belongs to $C^{k,\alpha}$ if it can be locally represented as the subgraph of a $C^{k,\alpha}(U)$ function.

Definition A.2. (Push-forward measure). Let $(X_1, \Sigma_1), (X_2, \Sigma_2)$ be two measurable spaces and let $f : X_1 \to X_2$ be measurable. Given a measure μ on X_1, we define the push-forward measure on X_2 by the formula

$$f_{\sharp}\mu(B) := \mu(\{x \in X_1 : f(x) \in B\}),$$

where B is an arbitrary set in Σ_2.

The next proposition is a classical result in measure theory and may be found, e.g. in [9], Section 3.1.

PROPOSITION A.3. (Hahn decomposition). Let μ be a signed measured on a measure space (X, Σ). Then, there exists two measurable sets P, N so that $P \cup N = X$, $P \cap N = \emptyset$ and so that $\mu(E) \geq 0$ for all $E \subseteq P$ and $\mu(F) \leq 0$ for all $F \subseteq N$.

THEOREM A.4. (Lebesgue differentiation theorem, see, e.g. Section 3.4 in [34]). Let $(\mathcal{X}, \mathsf{d}, \nu)$ be a metric measure space with ν satisfying (3.6), and let $f \in L^1(\mathcal{X}, \nu)$. Then, for ν-almost every $x \in \mathcal{X}$, we have that

$$\lim_{r \downarrow 0} \frac{1}{\nu(B(x,r))} \int_{B(x,r)} f(y)\, \mathrm{d}\nu(y) = f(x).$$

DEFINITION A.5. (Weak-* convergence and compactness). Let X be a Banach space of $\mathbb{R}$ with dual X^*. We say that a sequence $(y_k)_{k \in \mathbb{N}} \subseteq X^*$ is weak-* convergent to $y \in X^*$ if

$$\lim_{k \to \infty} y_k(x) = y(x), \qquad \forall x \in X.$$

THEOREM A.6. (Banach–Alaoglu). Let X be a Banach space of $\mathbb{R}$ with dual X^*. Then, any bounded subset of X^* is precompact in the weak-* topology.

B. Alternative formulations of the adversarial problem

At this point, we review a few other established formulations of the adversarial problem and add pointers towards the relevant literature. These reformulations provide different ways of understanding and analyzing the original adversarial problem.

B.1 *Open vs closed balls*

We would like to continue the discussion in Remark 1.3 and elaborate on why the adversarial model with open balls that we study here does not require the universal σ-algebra. We observe that the closed norm balls model which was considered in [3, 50] suffers from the problem that

$$\sup_{\overline{B}_\varepsilon(x)} 1_A = 1_{A^{\overline{\oplus}\varepsilon}},$$

where the set $A^{\overline{\oplus}\varepsilon} := \bigcup_{x \in A} \overline{B}_\varepsilon(x)$ is in general not Borel measurable even though A might be. Here, $\overline{B}_\varepsilon(x) := \{y \in \mathcal{X} : \mathsf{d}(x,y) \leq \varepsilon\}$ denotes the closed ε-ball around $x \in \mathcal{X}$. One can sandwich $A^{\overline{\oplus}\varepsilon}$ between open and closed parallel sets (in particular Borel sets) like this

$$\{x \in \mathcal{X} : \mathsf{dist}(x,A) < \varepsilon\} \subseteq A^{\overline{\oplus}\varepsilon} \subseteq \{x \in \mathcal{X} : \mathsf{dist}(x,A) \leq \varepsilon\},$$

but the inclusions may be strict in general, see [50]. The situation is markedly different for open balls, where one has

$$\sup_{B_\varepsilon(x)} 1_A = 1_{A^{\oplus\varepsilon}},$$

where $A^{\oplus\varepsilon} := \bigcup_{x \in A} B_\varepsilon(x)$ is an open set and satisfies the following.

LEMMA B.1. It holds that

$$\{x \in \mathcal{X} \,:\, \mathsf{dist}(x, A) < \varepsilon\} = \bigcup_{x \in A} B_\varepsilon(x).$$

Proof. Let $y \in \mathcal{X}$ such that $\mathsf{dist}(y, A) < \varepsilon$. Then, there exists a sequence of points $(x_k)_{k\in\mathbb{N}} \subseteq A$ with $\lim_{k\to\infty} \mathsf{d}(y, x_k) = \mathsf{dist}(y, A) < \varepsilon$. Hence, there exists $K \in \mathbb{N}$ such that for all $k \geq K$, it holds $\mathsf{d}(y, x_k) < \varepsilon$ and therefore

$$y \in \bigcup_{k \geq K} B_\varepsilon(x_k) \subseteq \bigcup_{x \in A} B_\varepsilon(x).$$

This establishes the inclusion '$\subseteq$'. For the converse inclusion, let $y \in \bigcup_{x\in A} B_\varepsilon(x)$. Then, there exists $x \in A$ such that $y \in B_\varepsilon(x)$ and therefore

$$\mathsf{dist}(y, A) \leq \mathsf{d}(y, x) < \varepsilon,$$

which establishes the inclusion '$\supseteq$' and concludes the proof. □

B.2 ∞-*Wasserstein DRO problem*

It is well-known [50] that the closed ball adversarial problem is indeed a DRO problem in the form of (1.1) with respect to a special ∞-Wasserstein distance. To this end, we introduce an ∞-Wasserstein distance between two measures μ and $\tilde{\mu}$ as

$$W_\infty(\mu, \tilde{\mu}) := \inf_{\pi \in \Gamma(\mu, \tilde{\mu})} \pi\text{-ess sup}\, c_\infty, \tag{B1}$$

where the cost function is given by

$$c_\infty : \big(\mathcal{X} \times \{0, 1\}\big)^2 \to [0, +\infty], \tag{B2a}$$

$$c_\infty\big((x, y), (\tilde{x}, \tilde{y})\big) := \begin{cases} \mathsf{d}(x, \tilde{x}) & \text{if } y = \tilde{y}, \\ +\infty & \text{if } y \neq \tilde{y}. \end{cases} \tag{B2b}$$

PROPOSITION B.2. ([50]). Let $\mathcal{X}$ be Polish. Then, the adversarial risk of $A \in \mathfrak{B}(\mathcal{X})$ can be reformulated as

$$\mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x} \in \overline{B}_\varepsilon(x)} \big|1_A(\tilde{x}) - y\big|\right] = \sup_{\substack{\tilde{\mu} \in \mathcal{P}(\mathcal{X}\times\{0,1\}) \\ W_\infty(\mu, \tilde{\mu}) \leq \varepsilon}} \mathbb{E}_{(x,y)\sim\tilde{\mu}}\left[\big|1_A(x) - y\big|\right]. \tag{B3}$$

B.3 *Dual of an optimal transport problem*

It is also known [6, 50, 59] that the adversarial problem may be reformulated as the dual of an optimal transport problem. The following result is stated in the setting $\mathcal{X} = \mathbb{R}^d$ endowed with the Euclidean distance but can be generalized to arbitrary metric spaces in a straightforward way. Let μ^S be the probability distribution on $\mathcal{X} \times \{0, 1\}$ defined as

$$\mu^S := T^S_\sharp \mu, \quad \text{where } T^S(x, y) := (x, 1 - y), \quad \forall (x, y) \in \mathcal{X} \times \{0, 1\}.$$

The map T^S can be interpreted as a transformation that leaves features unchanged while swapping labels. The following statement is proven in [59].

PROPOSITION B.3. (cf. [59, Corollary 3.2). Let $c_\varepsilon : (\mathbb{R}^d \times \{0,1\})^2 \to \mathbb{R}$ be the function defined by

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

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

$$\inf_{A \in \mathfrak{B}(\mathcal{X})} \mathbb{E}_{(x,y)\sim\mu} \left[\sup_{\tilde{x} \in \overline{B}_\varepsilon(x)} \left| 1_A(\tilde{x}) - y \right| \right] = \frac{1}{2} - \frac{1}{2} \inf_{\pi \in \Gamma(\mu, \mu^S)} \iint_{(\mathbb{R}^d \times \{0,1\})^2} c_\varepsilon(z_1, z_2) \, \mathrm{d}\pi(z_1, z_2). \tag{B4}$$

This type of result was first established independently in [6, 50] where the balanced case $w_0 = w_1 = 1/2$ was considered. The OT problem on the right-hand side of (B4) is an alternative way to compute the optimal adversarial risk. This alternative has clear advantages over the original formulation of the problem in situations such as when μ is an empirical measure (the standard setting in practice). Indeed, in that setting, problem (1.7) is in principle an infinite dimensional problem, while the OT problem will always be finite dimensional. One may speculate further and wonder whether there is a connection between solutions to the OT problem and optimal adversarially robust classifiers, or in other words, whether one can construct adversarially robust classifiers from a solution to the OT problem. This is indeed the case and such results will be elaborated on in future work.

C. Additional proofs

Here, we prove Lemma 3.12, which we restate for convenience.

LEMMA C.1. Under Assumption 1 for any Borel measurable function $u \in L^\infty(\mathcal{X}; \nu)$, there exists $u^\star : \mathcal{X} \to \mathbb{R}$ such that $u = u^\star$ holds ν-almost everywhere and

$$\sup_{B_\varepsilon(x)} u^\star = \nu\text{-ess}\sup_{B_\varepsilon(x)} u^\star, \quad \inf_{B_\varepsilon(x)} u^\star = \nu\text{-ess}\inf_{B_\varepsilon(x)} u^\star, \quad \forall x \in \mathrm{supp}\rho. \tag{C1}$$

Proof. **Step 1:** Let $t_1, t_2, t_3, \ldots$ be an enumeration of the rational numbers. In what follows, we construct a collection of measurable sets $A^\star_{t_1}, A^\star_{t_2}, A^\star_{t_3}, \ldots$ satisfying the following properties:

1. For every k, we have $\{u \geq t_k\} \sim_\nu A^\star_{t_k}$.
2. For every k, $A^\star_{t_k}$ satisfies (3.9).
3. For any two $k \neq l$, if $t_k < t_l$, then $1_{A^\star_{t_l}}(x) \leq 1_{A^\star_{t_k}}(x)$ for every $x \in \mathrm{supp}\nu$.

We construct these sets inductively. First, following the proof of Lemma 3.8 applied to the set $A = \{u \geq t_1\}$, we obtain the set $A^\star_{t_1}$ defined through its indicator function according to

$$1_{A^\star_{t_1}}(x) := \begin{cases} 1 & \text{if } x \in D^\varepsilon_+(t_1), \\ 0 & \text{if } x \in D^\varepsilon_-(t_1), \\ 1_{\{u \geq t_1\}} & \text{if } x \in \mathbb{R}^d \setminus (D^\varepsilon_+(t_1) \cup D^\varepsilon_-(t_1)). \end{cases}$$

In the above, we use the notation $D^\varepsilon_+(t_1)$, $D^\varepsilon_-(t_1)$, as well as the notation $D_+(t_1)$, $D_-(t_1)$, to denote the sets introduced in the proof of Lemma 3.8 emphasizing that these sets are associated to the set $\{u \geq t_1\}$.

Now, suppose that we have constructed the sets $A^\star_{t_1}, \ldots, A^\star_{t_L}$ (in terms of associated sets $D_+(t_l), D_-(t_l), D^\varepsilon_+(t_l), D^\varepsilon_-(t_l)$ for every $l = 1, \ldots, L$) and suppose that these sets satisfy (1)–(3)

when restricted to $k, l \in \{1, \dots, L\}$. We now discuss how to construct the set $A^\star_{t_{L+1}}$. Suppose that $t_k < t_{L+1} < t_l$ for some $k, l \in \{1, \dots, L\}$ and suppose that these indices are chosen so that there is no element in $\{t_1, \dots, t_L\}$ strictly between t_k and t_l (if t_{L+1} was bigger, or smaller, than all the t_k with $k = 1, \dots, L$, a similar construction to the one we exhibit next would apply and because of this, we focus on the case mentioned earlier for brevity). We start by defining

$$1_{\tilde{A}_{t_{L+1}}}(x) := \begin{cases} 1 & \text{if } x \in D^\varepsilon_+(t_{L+1}), \\ 0 & \text{if } x \in D^\varepsilon_-(t_{L+1}), \\ 1_{\{u \geq t_{L+1}\}}(x) & \text{if } x \in \mathbb{R}^d \setminus (D^\varepsilon_+(t_{L+1}) \cup D^\varepsilon_-(t_{L+1})), \end{cases}$$

obtained following the construction in Lemma 3.8 when applied to the set $\{u \geq t_{L+1}\}$. We now modify $\tilde{A}_{t_{L+1}}$ slightly to eventually satisfy property (3). Indeed, since $\{u \geq t_l\} \subseteq \{u \geq t_{L+1}\} \subseteq \{u \geq t_k\}$ and $A^\star_{t_k} \sim_\nu \{u \geq t_k\}$, $A^\star_{t_l} \sim_\nu \{u \geq t_l\}$, $\tilde{A}_{t_{L+1}} \sim_\nu \{u \geq t_{L+1}\}$, we conclude that

$$1_{A^\star_{t_l}}(x) \leq 1_{\{u \geq t_{L+1}\}}(x) \leq 1_{A^\star_{t_k}}(x)$$

for every $x \in \mathbb{R}^d \setminus \mathcal{N}_{L+1}$, where $\mathcal{N}_{L+1}$ is some ν-null set. We then define

$$1_{A^\star_{t_{L+1}}}(x) := \begin{cases} 1_{A^\star_{t_l}}(x) & \text{if } x \in (\mathbb{R}^d \setminus (D^\varepsilon_+(t_{L+1}) \cup D^\varepsilon_-(t_{L+1}))) \cap \mathcal{N}_{L+1}, \\ 1_{\tilde{A}_{t_{L+1}}}(x) & \text{otherwise.} \end{cases}$$

It is evident that $A^\star_{t_{L+1}}$ is ν-equivalent to $\{u \geq t_{L+1}\}$ and that it satisfies (3.9) (since we do not modify the $\tilde{A}_{t_{L+1}}$ inside $D^+_\varepsilon(t_{L+1})$ or $D^-_\varepsilon(t_{L+1})$). It remains to show that for every $x \in \text{supp}\nu$, we have $1_{A^\star_{t_l}}(x) \leq 1_{A^\star_{t_{L+1}}}(x) \leq 1_{A^\star_{t_k}}(x)$. By definition of $A^\star_{t_{L+1}}$ and the relation between $A^\star_{t_k}$ and $A^\star_{t_l}$, the inequality is immediate if $x \in \mathbb{R}^d \setminus (D^\varepsilon_+(t_{L+1}) \cup D^\varepsilon_-(t_{L+1}))$. Thus, it suffices to show the inequality when $x \in D^\varepsilon_+(t_{L+1})$ (as the case $x \in D^\varepsilon_-(t_{L+1})$ is completely analogous). In turn, we just have to show that $1_{A^\star_{t_k}}(x) = 1$. Now, notice that $x \in D^\varepsilon_+(t_{L+1})$ means that there is $x_1 \in \text{supp}\rho$ such that $x_1 \in D_+(t_{L+1})$ and $\mathsf{d}(x, x_1) < \varepsilon$. In particular, $1_{\{u \geq t_{L+1}\}}(\tilde{x}) = 1$ for ν-a.e. $\tilde{x} \in B_\varepsilon(x_1)$. Given that $\{u \geq t_{L+1}\} \subseteq \{u \geq t_k\}$, we also have that $1_{\{u \geq t_k\}}(\tilde{x}) = 1$ for ν-a.e. $\tilde{x} \in B_\varepsilon(x_1)$, and hence, $x_1 \in D_+(t_k)$. We conclude that $x \in D^\varepsilon_+(t_k)$ and in turn that $1_{A^\star_{t_k}}(x) = 1$.

Step 2: Using the family of sets $A^\star_{t_1}, A^\star_{t_2}, A^\star_{t_3}, \dots$, we construct the function $u^\star$ according to

$$u^\star(x) := \sup\{t \in \mathbb{Q} : x \in A^\star_t\}.$$

We now show the following relations between level sets of $u^\star$ and the sets $A^\star_t$: for every $t \in \mathbb{Q}$, we have

$$1_{A^\star_t}(x) \leq 1_{\{u^\star \geq t\}}(x), \quad \forall x \in \text{supp}\nu, \tag{C2}$$

and for every $s, t \in \mathbb{Q}$ with $s < t$, we have

$$1_{\{u^\star \geq t\}}(x) \leq 1_{A^\star_s}(x), \quad \forall x \in \text{supp}\nu. \tag{C3}$$

Inequality (C2) follows from the fact that if $x \in A^\star_t$, then by definition of $u^\star$, we have $u^\star(x) \geq t$. So in this case, we actually have the stronger condition $A^\star_t \subseteq \{u^\star \geq t\}$.

To obtain inequality (C3) take $x \in \text{supp}\nu$ such that $u^\star(x) \geq t$. Then, there must exist a rational $r \geq s$ such that $x \in A^\star_r$ for otherwise $u^\star(x)$ would be less than s. From property (3) of the sets $A^\star$, we deduce that $x \in A^\star_s$ also.

Step 3: We now show that $u^\star$ satisfies (C1). To see this, let $x \in \text{supp}\rho$ and suppose for the sake of contradiction that $\sup_{B_\varepsilon(x)} u^\star > \nu\text{-ess}\sup_{B_\varepsilon(x)} u^\star$. Pick $t \in \mathbb{Q}$ strictly between these two values. Then, there is $\tilde{x} \in B_\varepsilon(x)$ (notice that $\tilde{x} \in \text{supp}\nu$) such that $u^\star(\tilde{x}) \geq t$. In particular, from (C3), it follows that $1_{A_s^\star}(x) = 1$ for a rational s with $s < t$ that is also strictly larger than $\nu\text{-ess}\sup_{B_\varepsilon(x)} u^\star$, and in turn, we deduce that $\sup_{B_\varepsilon(x)} 1_{A_s^\star} = 1$. On the other hand, from the fact that $\nu\text{-ess}\sup_{B_\varepsilon(x)} u^\star < s$, it is clear that $\nu(\{u^\star \geq s\} \cap B_\varepsilon(x)) = 0$, and thus, if we combine with (C2), we deduce that $\nu(A_s^\star \cap B_\varepsilon(x)) = 0$ also. This means that $\nu\text{-ess}\sup_{B_\varepsilon(x)} 1_{A_s^\star} = 0$, contradicting in this way property (2) for the set $A_s^\star$. In conclusion: for every $x \in \text{supp}\rho$, we have $\sup_{B_\varepsilon(x)} u^\star = \nu\text{-ess}\sup_{B_\varepsilon(x)} u^\star$.

To show the second part of (C1), we follow a similar strategy. Namely, suppose for the sake of contradiction that there is $x \in \text{supp}\rho$ such that $\inf_{B_\varepsilon(x)} u^\star < \text{ess}\inf_{B_\varepsilon(x)} u^\star$. Let s be a rational number strictly between these two values. Then, there is $\tilde{x} \in B_\varepsilon(x)$ such that $u^\star(\tilde{x}) < s$ and thus we must have $\tilde{x} \notin A_s^\star$. In particular, we have $\inf_{B_\varepsilon(x)} 1_{A_s^\star} = 0$. Picking now a rational t strictly between s and $\text{ess}\inf_{B_\varepsilon(x)} u^\star$, we conclude that $1_{\{u^\star \geq t\}} = 1$ ν-a.e. $\tilde{x} \in B_\varepsilon(x)$. By (C3), the same is true when we replace $\{u^\star \geq t\}$ with $A_s^\star$. Thus, we conclude that $\text{ess}\inf_{B_\varepsilon(x)} 1_{A_s^\star} = 1$, contradicting property (2) for the set $A_s^\star$. In conclusion: for every $x \in \text{supp}\rho$ we have $\inf_{B_\varepsilon(x)} u^\star = \text{ess}\inf_{B_\varepsilon(x)} u^\star$.

Step 4: Finally, we can combine property (1) of the sets $A^\star$, inequalities (C2) and (C3), and a similar argument (i.e. by contradiction) to the one used in Step 3, in order to conclude that $u^\star = u$ holds ν-a.e. □