European Journal of Applied Mathematics (2024), 1–21
doi:10.1017/S0956792524000822

PAPER

On the existence of solutions to adversarial training in multiclass classification

Nicolás García Trillos[1], Matt Jacobs[2] and Jakwang Kim[3]

[1]Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA
[2]Department of Mathematics, UC Santa Barbara, Santa Barbara, CA, USA
[3]Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada
Corresponding author: Nicolás García Trillos; Email: garciatrillo@wisc.edu

Received: 22 February 2024; **Revised:** 03 November 2024; **Accepted:** 05 November 2024

Keywords: existence of solutions for minimax problems; nonparametric robustness; general topics in artificial intelligence; problem-solving in the context of artificial intelligence

2020 Mathematics Subject Classification: 49J35 (Primary); 62G35, 68T20 (Secondary)

Abstract

Adversarial training is a min-max optimization problem that is designed to construct robust classifiers against adversarial perturbations of data. We study three models of adversarial training in the multiclass agnostic-classifier setting. We prove the existence of Borel measurable robust classifiers in each model and provide a unified perspective of the adversarial training problem, expanding the connections with optimal transport initiated by the authors in their previous work [21]. In addition, we develop new connections between adversarial training in the multiclass setting and total variation regularization. As a corollary of our results, we provide an alternative proof of the existence of Borel measurable solutions to the agnostic adversarial training problem in the binary classification setting.

1. Introduction

Modern machine learning models, in particular those generated with deep learning, perform remarkably well, in many cases much better than humans, at classifying data in a variety of challenging application fields like image recognition, medical image reconstruction, and natural language processing. However, the robustness of these learning models to data perturbations is a completely different story. For example, in image recognition, it has been widely documented (e.g., [26]) that certain structured but human-imperceptible modifications of images at the pixel level can fool an otherwise well-performing image classification model. These small data perturbations, known as *adversarial attacks*, when deployed at scale, can make a model's prediction accuracy drop substantially and in many cases collapse altogether. As such, they are a significant obstacle to the deployment of machine learning systems in security-critical applications, e.g. [8]. To defend against these attacks, many researchers have investigated the problem of adversarial training, i.e., training methods that produce models that are robust to attacks. In adversarial training, one typically pits the adversary against the learner during the training step, forcing the learner to select a model that is robust against attacks. Nonetheless, despite the attention that has been devoted to understanding these problems, both theoretically and algorithmically, there are still several important mathematical questions surrounding them that have not been well understood.

A fundamental difficulty in adversarial training, in contrast to standard training of learning models, is the fact that the adversary has the power to alter the underlying data distribution. In particular, model training becomes an implicit optimisation problem over a space of measures. As a result, one may be forced to leave the prototypical setting of equivalence classes of functions defined over a single fixed

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

measure space. In general, measurability issues become more delicate for adversarial training problems at the moment of providing a rigorous mathematical formulation for the problem. Due to these difficulties, there are several subtle variations of the adversarial training model in the literature, and it has not been clear whether these models are fully equivalent. More worryingly, for some models, even the *existence* of optimal robust classifiers is unknown, essentially due to convexity and compactness issues.

Let us emphasise that these issues arise even in what can be regarded as the simplest possible setting of the agnostic learner, i.e., where the space of classifiers is taken to be the set of all possible Borel measurable weak (probabilistic) classifiers. While this setting is trivial in the absence of an adversary (there the optimal choice for the learner is always the Bayes classifier), the structure of the problem is much more subtle in the adversarial setting (in other words the analogue of the Bayes classifier is not fully understood). With an adversary, the training process can be viewed as a two-player min-max game (learner versus adversary) [4, 12, 32, 35], and as a result, the optimal strategies for the two players are far from obvious. By relaxing the problem to the agnostic setting, one at least is working over a convex space, but again measurability issues pose a problem for certain formulations of adversarial training.

In light of the above considerations, the purpose of this paper is twofold. On one hand, we provide rigorous justification for the existence of *Borel* measurable robust classifiers in the multiclass classification setting for three different models of adversarial training. Notably, our analysis includes a widely used model for which the existence of Borel classifiers was not previously known. On the other hand, we develop a series of connections between the three mathematical models of adversarial training discussed throughout the paper exploiting ideas from optimal transportation and total variation minimisation. By developing these connections, we hope to present a unified formulation of adversarial training and highlight the prospective advantages of using tools in computational optimal transport for solving these problems in the agnostic-classifier setting (and perhaps beyond the agnostic setting too). We also highlight, in concrete terms, the connection between adversarial training and the direct regularisation of learning models. To achieve all the aforementioned goals, we expand and take advantage of our previous work [21] as well as of the work [15] exploring the connection between adversarial training and perimeter minimisation in the binary classification setting.

1.1 Organisation of the paper

The rest of the paper is organised as follows. In Section 2, we introduce three different models for adversarial training in the multiclass classification setting that we will refer to as the open-ball model, the closed-ball model and the distributional-perturbing model. In Section 2.1, we state our main mathematical results, and in Section 2.2, we discuss related literature and some of the implications of our results. In Section 3, we lay down the main mathematical tools for analysing the distributional-perturbing model. Part of these tools come directly from our previous work [21], while others are newly developed. In Section 4, we prove our main results: first, we prove the existence of solutions for the distributional-perturbing model (Section 4.1); then, we prove that solutions to the distributional-perturbing model are solutions to the closed-ball model (Section 4.2); finally, we relate the closed-ball model to the open-ball model in Section 4.3. Lastly, in Section 5, we wrap up the paper and discuss future research directions.

2. Set-up and main results

The setting of our problem will be a feature space $(\mathcal{X}, d)$ (a Polish space with metric d) and a label space $\mathcal{Y} := \{1, \dots, K\}$, which will represent a set of K labels for a given classification problem of interest. For each $x \in \mathcal{X}$, we will use $B_\varepsilon(x)$ ($\overline{B}_\varepsilon(x)$, respectively) to denote an open (closed) ball with radius ε centred at x. We denote by $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ the set of input-to-output pairs and endow it with a Borel probability measure $\mu \in \mathcal{P}(\mathcal{Z})$, representing a ground-truth data distribution. For convenience, we will often describe the measure μ in terms of its class probabilities $\mu = (\mu_1, \dots, \mu_K)$, where each μ_i is the positive Borel measure (not necessarily a probability measure) over $\mathcal{X}$ defined according to:

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

$$\mu_i(A) = \mu(A \times \{i\}),$$

for $A \in \mathfrak{B}(\mathscr{X})$, i.e., A is a Borel measurable subset of $\mathscr{X}$. Notice that the measures μ_i are, up to normalisation factors, the conditional distributions of inputs/features given the different output labels.

Typically, a (multiclass) classification rule in the above setting is simply a Borel measurable map $f:\mathscr{X} \to \mathscr{Y}$. In this paper, however, it will be convenient to expand this notion slightly and interpret general classification rules as Borel measurable maps from $\mathscr{X}$ into $\Delta_{\mathscr{Y}} := \left\{(u_i)_{i\in\mathscr{Y}}:0 \le u_i \le 1, \sum_{i\in\mathscr{Y}} u_i \le 1\right\}$, the set of (up to normalisation constants) probability distributions over $\mathscr{Y}$ (see Remark 2.2); oftentimes these functions are known as *soft-classifiers*. For future reference, we denote by $\mathscr{F}$ the set

$$\mathscr{F} := \{f:\mathscr{X} \to \Delta_{\mathscr{Y}}:f \text{ is Borel measurable}\}\,. \tag{1}$$

Given $f \in \mathscr{F}$ and $x \in \mathscr{X}$, the vector $f(x) = (f_1(x), \dots, f_K(x))$ will be interpreted as the vector of probabilities over the label set $\mathscr{Y}$ that the classifier f assigns to the input data point x. In practice, from one such f, one can induce actual (hard) class assignments to the different inputs x by selecting the coordinate in $f(x)$ with largest entry. The extended notion of classifier considered in this paper is actually routinely used in practice as it fares well with the use of standard optimisation techniques (in particular, $\mathscr{F}$ is natural as it can be viewed as a convex relaxation of the space of maps from $\mathscr{X}$ to $\mathscr{Y}$).

The goal in the standard (unrobust) classification problem is to find a classifier $f \in \mathscr{F}$ that gives accurate class assignments to inputs under the assumption that data points are distributed according to the ground-truth distribution μ. This aim can be mathematically modelled as an optimisation problem of the form:

$$\inf_{f\in\mathscr{F}} R(f,\mu), \tag{2}$$

where $R(f,\mu)$ is the risk of a classifier f relative to the data distribution μ:

$$R(f,\mu) := \mathbb{E}_{(X,Y)\sim\mu}[\ell(f(X),Y))].$$

Here, a loss function is defined as $\ell:\Delta_{\mathscr{Y}} \times \mathscr{Y} \to \mathbb{R}$. For general and reasonable ℓ, one can observe that solutions to the risk minimisation Problem (2) are the standard multiclass Bayes classifiers from statistical learning theory (e.g., see [13, 40]). These classifiers are characterised by the condition $f^\star_{Bayes,i}(x) = \operatorname{argmax}_{i\in\mathscr{Y}} \mathbb{P}(Y=i|X=x)$ with an arbitrary tie-breaking rule, and such Bayes classifier is written as $f^\star_{Bayes}(x) = (\mathbb{1}_{A_1^\star}(x), \dots, \mathbb{1}_{A_K^\star}(x))$, where $A_1^\star, \dots, A_K^\star$ form a measurable partition of $\mathscr{X}$. In other words, there always exist hard classifiers that solve the risk minimisation Problem (2).

By definition, a solution to (2) classifies *clean* data optimally; by clean data here we mean data distributed according to the original distribution μ. However, one should not expect the standard Bayes classifier to perform equally well when inputs have been adversarially contaminated, and the goal in adversarial training is precisely to create classifiers that are less susceptible to data corruption. One possible way to enforce this type of robustness is to replace the objective function in (2) with one that incorporates the actions of a well defined adversary and then search for the classifier that minimises the new notion of (adversarial) risk. This adversarial risk can be defined in multiple ways, but two general ways stand out in the literature and will be the emphasis of our discussion; we will refer to these two alternatives as the *data-perturbing adversarial model* and the *distribution-perturbing adversarial model*. More precisely, the data-perturbing adversarial model is generally defined as

$$\inf_{f\in\mathscr{F}} \mathbb{E}_{(X,Y)\sim\mu}\left[\sup_{\widetilde{X}\in A_\varepsilon(X)} \ell(f(\widetilde{X}),Y)\right]$$

where $A_\varepsilon(\,\cdot\,)$ is a measurable set parameterised by and *adversarial budget* ε (usually, closed/open ball with radius ε), and the distribution-perturbing adversarial model is defined as:

$$\inf_{f\in\mathscr{F}} \sup_{\widetilde{\mu}} \mathbb{E}_{(\widetilde{X},\widetilde{Y})\sim\widetilde{\mu}}\left[\ell(f(\widetilde{X}),\widetilde{Y}) - C_\varepsilon(\mu,\widetilde{\mu})\right]$$

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

where $C_\varepsilon(\mu, \widetilde{\mu})$ is some cost or distance over $\mathscr{P}(\mathscr{X})$. Although the interpretation of these models is straightforward, their theoretical underpinnings are not completely understood. This is true even for the simple linear loss function

$$\ell(u, i) = 1 - u_i, \quad (u, i) \in \Delta_{\mathscr{Y}} \times \mathscr{Y},$$

which, in lieu of the fact that $\ell(e_j, i)$ is equal to 1 if $i \neq j$ and 0 if $i = j$ (e_j is the extremal point of $\Delta_{\mathscr{Y}}$ with entry one in its j-th coordinate), will be referred to as the 0-1 loss. We will focus on this important case from now on. With this choice, $R(f, \mu)$ can be written explicitly as:

$$R(f, \mu) = \sum_{i \in \mathscr{Y}} \int_{\mathscr{X}} (1 - f_i(x)) d\mu_i(x).$$

As it turns out, some results in the literature have established some connections between the two types of models (see [35] for more details), and we will develop further connections shortly. For the *data-perturbing adversarial model,* we will consider the following two versions:

$$R^\star_{open}(\varepsilon) := \inf_{f \in \mathscr{F}} R^\star_{open}(f; \varepsilon) := \inf_{f \in \mathscr{F}} \left\{ \sum_{i \in \mathscr{Y}} \int_{\mathscr{X}} \sup_{\tilde{x} \in B_\varepsilon(x)} \{1 - f_i(\tilde{x})\} d\mu_i(x) \right\}, \tag{3}$$

and

$$\inf_{f \in \mathscr{F}} \left\{ \sum_{i \in \mathscr{Y}} \int_{\mathscr{X}} \sup_{\tilde{x} \in \overline{B}_\varepsilon(x)} \{1 - f_i(\tilde{x})\} d\mu_i(x) \right\}. \tag{4}$$

Recall $B_\varepsilon(x)$ ($\overline{B}_\varepsilon(x)$, respectively) denotes an open (a closed) ball. In both versions, the adversary can substitute any given input x with a $\tilde{x}$ that belongs to a small ball of radius ε around the original x. The adversary has more power to perturb an input x by $\tilde{x}$ with a larger ε since more powerful deviation will be possible with larger ε. Due to measurability issues that we will discuss next, at this stage, Problem (4) is introduced informally.

In both Problems (3) and (4), the learner's goal is to minimise the worst-loss that the adversary may induce by carrying out one of their feasible actions. Although at the heuristic level the difference between the two models is subtle (in (3) the adversary optimises over open balls and in (4) over closed balls), at the mathematical level these two models can be quite different. For starters, the Problem (4) is not well formulated, as it follows from a classical result in [30], which discusses that, in general, the function $x \mapsto \sup_{\tilde{x} \in \overline{B}_\varepsilon(x)} \{1 - f_i(\tilde{x})\}$ may not be Borel measurable when only the Borel measurability of the function f_i has been assumed. For this reason, the integral with respect to μ_i in (4) (which is a Borel positive measure, i.e., it is only defined over the Borel σ-algebra) may not be defined for all $f \in \mathscr{F}$. In Subsection 2.1, we will provide a rigorous formulation of (4), which we will call the closed-ball Model (9) because the adversarial attack lies in the closed ball. This reformulation will require the use of an extension of the Borel σ-algebra, known as the universal σ-algebra, as well as an extension of the measures μ_i to this enlarged σ-algebra. Problem (3), on the other hand, is already well formulated, as no measurability issues arise when taking the sup over open balls. At a high level, this is a consequence of the fact that arbitrary unions of open balls are open sets and thus Borel measurable; see, for example, [15, Remark 2.3]. Regardless of which of the two models one adopts, and putting aside for a moment the measurability issues mentioned above, it is unclear whether it is possible to find minimisers for any of the Problems (3) and (4) within the family $\mathscr{F}$.

Moving on to the *distributional-perturbing adversarial model*, notice that it is defined as a minimax problem that can be described as follows: after the adversary selects a new data distribution $\widetilde{\mu} \in \mathscr{P}(\mathscr{X})$ and, by paying some cost $C(\mu, \widetilde{\mu})$, attempts to make the risk $R(\cdot, \widetilde{\mu})$ be as large as possible, the learner has chosen a classifier $f \in \mathscr{F}$. The reverse interpretation is also true due to the strong duality of (5) in Ref. [21]. Theorem 2.5, furthermore, implicitly implies the strong duality as a corollary since it proves a stronger result, the existence of saddle points (Nash equilibria).

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

Precisely, the distributional-perturbing adversarial model can be rewritten as:

$$R^{\star}_{DRO}(\varepsilon) := \inf_{f\in\mathscr{F}} \sup_{\widetilde{\mu}\in\mathscr{P}(\mathscr{Z})} \left\{ R(f,\widetilde{\mu}) - C(\mu,\widetilde{\mu}) \right\}, \tag{5}$$

where we assume the cost $C:\mathscr{P}(\mathscr{Z})\times\mathscr{P}(\mathscr{Z})\to[0,\infty]$ takes the form:

$$C(\mu,\widetilde{\mu}) := \inf_{\pi\in\Gamma(\mu,\widetilde{\mu})} \int c_{\mathscr{Z}}(z,\widetilde{z})d\pi(z,\widetilde{z}),$$

for some Borel measurable cost function $c_{\mathscr{Z}}:\mathscr{Z}\times\mathscr{Z}\to[0,\infty]$. Here and in the remainder of the paper, we use $\Gamma(\cdot,\cdot)$ to represent the set of couplings between two positive measures over the same space; for example, $\Gamma(\mu,\widetilde{\mu})$ denotes the set of positive measures over $\mathscr{Z}\times\mathscr{Z}$ whose first and second marginals are μ and $\widetilde{\mu}$, respectively. Note that Problem (5) is an instance of the *distributionally robust optimisation* (DRO) problem. Problem (5) is well defined given that all its terms are written as integrals of Borel measurable integrands against Borel measures.

In the remainder, we will assume that the cost $c_{\mathscr{Z}}:\mathscr{Z}\times\mathscr{Z}\to[0,\infty]$ has the form

$$c_{\mathscr{Z}}(z,\tilde{z}) := \begin{cases} c(x,\tilde{x}) & \text{if } y=\tilde{y}, \\ \infty & \text{otherwise} \end{cases} \tag{6}$$

for a lower semi-continuous function $C:\mathscr{X}\times\mathscr{X}\to[0,\infty]$. Note that when $c_{\mathscr{Z}}$ has the above structure, we can rewrite $C(\mu,\widetilde{\mu})$ as:

$$C(\mu,\widetilde{\mu}) = \sum_{i=1}^{K} C(\mu_i,\widetilde{\mu}_i),$$

where on the right-hand side, we slightly abuse notation and use $C(\mu_i,\widetilde{\mu}_i)$ to represent

$$C(\mu_i,\widetilde{\mu}_i) := \inf_{\pi\in\Gamma(\mu_i,\widetilde{\mu}_i)} \int c(x,\tilde{x})d\pi(x,\tilde{x}).$$

Although $R^{\star}_{DRO}(\varepsilon)$ seems not to depend on ε explicitly unlike Problems (3) and (4), readers can understand that a cost function c relies implicitly on the choice of adversarial budget ε. The most popular example, in both theory and practice, of a cost function c, that we discuss in detail throughout this paper is the 0-∞ cost function:

$$c(x,\tilde{x}) = c_{\varepsilon}(x,\tilde{x}) := \begin{cases} \infty & \text{if } d(x,\tilde{x}) > \varepsilon \\ 0 & \text{if } d(x,\tilde{x}) \le \varepsilon. \end{cases} \tag{7}$$

Similar to (3) and (4), a larger ε means more powerful adversarial attacks, as the adversary may select arbitrary $\widetilde{\mu}$ at no cost.

Remark 2.1. Throughout the paper, we use the convention that $C(\mu_i,\widetilde{\mu}_i)=\infty$ whenever the set of couplings $\Gamma(\mu_i,\widetilde{\mu}_i)$ is empty. This is the case when μ_i and $\widetilde{\mu}_i$ have different total masses.

Remark 2.2. Given the structure of the 0–1 loss function considered here, in all the adversarial models introduced above, we may replace the set $\mathscr{F}$ with the set of those $f\in\mathscr{F}$ for which $\sum_i f_i = 1$. Indeed, given $f\in\mathscr{F}$, we can always consider $\tilde{f}\in\mathscr{F}$ defined according to $\tilde{f}_{i_0} := f_{i_0} + (1-\sum_{i\in\mathscr{Y}} f_i)$ and $\tilde{f}_i = f_i$ for $i\neq i_0$ to obtain a value of risk that is no greater than the one of the original f.

2.1 Main results

Our first main theorem discusses the existence of (Borel) solutions for Problem (5) under the assumptions on the cost $c:\mathscr{X}\times\mathscr{X}\to[0,\infty]$ stated below.

Assumption 2.3. *We assume that the cost $c:\mathscr{X}\times\mathscr{X}\to[0,\infty]$ is a lower semi-continuous and symmetric function satisfying $c(x,x)=0$ for all $x\in\mathscr{X}$. We also assume that the following compactness*

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

property holds: if $\{x_n\}_{n\in\mathbb{N}}$ *is a bounded sequence in* $(\mathscr{X}, d)$ *and* $\{x_n{}'\}_{n\in\mathbb{N}}$ *is a sequence satisfying* $\sup_{n\in\mathbb{N}} c(x_n, x_n{}') < \infty$*, then* $\{(x_n, x_n{}')\}_{n\in\mathbb{N}}$ *is precompact in* $\mathscr{X} \times \mathscr{X}$ *(endowed with the product topology).*

Remark 2.4. Notice that Assumption 2.3 implicitly requires bounded subsets of $\mathscr{X}$ to be precompact. In particular, when the cost has the form as in (7), Assumption 2.3 implies that every bounded subset in $\mathscr{X}$ should be precompact. We want to emphasise that this assumption is not too strong and is satisfied in many natural settings, e.g., Euclidean spaces or smooth manifolds of finite dimension endowed with its geodesic distance.

A simple cost function c, which does not satisfy Assumption 2.4, is

$$c'(x,\tilde{x}) = \begin{cases} \infty & \text{if } d(x,\tilde{x}) \geq \varepsilon \\ 0 & \text{if } d(x,\tilde{x}) < \varepsilon. \end{cases}$$

This is almost the same as (7), but c' is no longer lower semi-continuous. One can easily see that replacing c_ε by c' changes the closed-ball model (4) to the open-ball model (3). It is reasonable to expect that c' plays almost the same as c_ε: we will see that indeed this is true in Theorem 2.8.

Theorem 2.5. *Suppose that* $c: \mathscr{X} \times \mathscr{X} \to [0, \infty]$ *satisfies Assumption 2.3. Then there exists a (Borel) solution* $f^\star$ *of the DRO model (5). Furthermore, there exists* $\widetilde{\mu}^\star \in \mathscr{P}(\mathscr{Z})$ *such that* $(f^\star, \widetilde{\mu}^\star)$ *is a saddle point for (5). In other words, the following holds: for any* $\widetilde{\mu} \in \mathscr{P}(\mathscr{Z})$ *and any* $f \in \mathscr{F}$*, we have*

$$R(f^\star, \widetilde{\mu}) - C(\mu, \widetilde{\mu}) \leq R(f^\star, \widetilde{\mu}^\star) - C(\mu, \widetilde{\mu}^\star) \leq R(f, \widetilde{\mu}^\star) - C(\mu, \widetilde{\mu}^\star). \tag{8}$$

When the cost function c is regular enough or when μ is an empirical measure, we can reduce the problem of finding a solution $f^\star$ of (5) to the problem of solving the dual of a *generalised barycenter problem* or the dual of a *multimarginal optimal transport* problem. These connections were first put forward in our earlier work [21] and will be discussed again in Section 3, concretely in Proposition 3.10. Unfortunately, when the cost is only lower semi-continuous (e.g., for c as in (7)) and when μ is an arbitrary Borel probability measure, we cannot directly use the content of Proposition 3.10 to guarantee the existence of (Borel) solutions $f^\star$. To overcome this issue, we approximate c with a sequence of continuous costs c_n such that the previous theory applies. We then show that both the optimality and the Borel measurability of the optimal classifier are preserved in the limit. At a high level, we can thus reduce finding solutions for the DRO Problem (5) to that of an multimarginal optimal transport (MOT) or a generalised barycenter (or sequences thereof).

Next, we revisit Problem (4) and reformulate it in a rigorous way. Let us first introduce the *universal* σ-algebra of the space $\mathscr{X}$.

Definition 1. [33, *Definition* 2.2]. Let $\mathscr{B}(\mathscr{X})$ be the Borel σ-algebra over $\mathscr{X}$, and let $\mathscr{M}(\mathscr{X})$ be the set of all signed σ-finite Borel measures over $\mathscr{X}$. For each $\nu \in \mathscr{M}(\mathscr{X})$, let $\mathscr{L}_\nu(\mathscr{X})$ be the completion of $\mathscr{B}(\mathscr{X})$ with respect to ν. The universal σ-algebra of $\mathscr{X}$ is defined as:

$$\mathscr{U}(\mathscr{X}) := \bigcap_{\nu\in\mathscr{M}(\mathscr{X})} \mathscr{L}_\nu(\mathscr{X}).$$

We will use $\overline{\mathscr{P}}(\mathscr{Z})$ to denote the set of probability measures γ over $\mathscr{Z}$ for which γ_i is a universal positive measure (i.e., it is defined over $\mathscr{U}(\mathscr{X})$) for all $i \in \mathscr{Y}$. For a given probability measure $\mu \in \mathscr{P}(\mathscr{Z})$, we will denote by $\overline{\mu}$ its universal extension, which will be interpreted as:

$$\overline{\mu}(A \times \{i\}) := \overline{\mu}_i(A), \quad \forall A \in \mathscr{U}(\mathscr{X}),$$

where $\overline{\mu}_i$ is the extension of μ_i to $\mathscr{U}(\mathscr{X})$.

Having introduced the above notions, we can reformulate Problem (4) as:

$$R^\star_{closed}(\varepsilon) := \inf_{f\in\mathscr{F}} R^\star_{closed}(f;\varepsilon) := \inf_{f\in\mathscr{F}} \left\{ \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} \sup_{\tilde{x}\in\overline{B}_\varepsilon(x)} \{1 - f_i(\tilde{x})\} d\overline{\mu}_i(x) \right\}. \tag{9}$$

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

Although the difference with (4) is subtle (in [4] we use μ_i's whereas in [9] we use $\overline{\mu_i}$'s), Problem (9) is rigorously well defined. Indeed, combining [35, Lemma 4.2] with [6, Corollary 7.42.1], which was originally proved in Ref. [30], it follows that for any Borel measurable f_i, the function $x \mapsto \sup_{\tilde{x} \in \overline{B}_\varepsilon(x)}\{1 - f_i(\tilde{x})\}$ is universally measurable and thus the integrals on the right-hand side of (9) are well defined.

Remark 2.6. Similar to the universal extension of Borel measure, one might want to extend the solution space $\mathscr{F}$ to $\overline{\mathscr{F}}$, which is the set of all $f = (f_1, \dots, f_K)$ for which each $0 \leq f_i \leq 1$ is universally measurable and $\sum f_i \leq 1$. However, unless $\mathscr{X} = \mathbb{R}^d$, extending the solution space requires hierarchical extension of measures, i.e. universal extension of universal extension: see [33]. Note that on $\mathbb{R}^d$, every universally measurable set is Lebesgue measurable.

Our second main result relates solutions of (5) with solutions of (9).

Theorem 2.7. *There exists a Borel solution of (5) for the cost function c defined in (7) that is also a solution of (9). In particular, there exists a (Borel) solution for (9).*

Finally, we connect Problem (9) with Problem (3).

Theorem 2.8. *For all but at most countably many $\varepsilon \geq 0$, we have $R^\star_{open}(\varepsilon) = R^\star_{closed}(\varepsilon)$. Moreover, for those $\varepsilon \geq 0$ for which this equality holds, every solution $f^\star$ of (9) is also a solution of (3).*

Remark 2.9. Theorem 2.8 is optimal in general. One cannot expect the optimal adversarial risks of (3) and (9) to agree for all values of ε. To illustrate this, consider the simple setting of a two-class problem (i.e., $K = 2$) where $\mu_1 = \frac{1}{2}\delta_{x_1}$ and $\mu_2 = \frac{1}{2}\delta_{x_2}$, where δ_x denotes the usual Dirac delta measure at x. Let $\epsilon_0 = \frac{1}{2}d(x_1, x_2)$. It is straightforward to check that $R^\star_{open}(\epsilon_0) = 0$ whereas $R^\star_{closed}(\epsilon_0) = 1/2$. Naturally, if we had selected any other value for $\varepsilon > 0$ different from ϵ_0, we would have obtained $R^\star_{open}(\varepsilon) = R^\star_{closed}(\varepsilon)$.

From Theorems 2.5, 2.7 and 2.8, we may conclude that it is essentially sufficient to solve Problem (5) to find a solution for all other formulations of the adversarial training problem discussed in this paper. Our results thus unify all notions of adversarial robustness into the single DRO problem, (5). The advantage of (5) over the other formulations of the adversarial training problem is that it can be closely related to a generalised barycenter problem or an MOT problem, as has been discussed in detail in our previous work [21] (see also section 3 below). In turn, either of those problems can be solved using computational optimal transport tools. Recently, in [36], the authors propose fast algorithms for solving (5). From a practical perspective, hence, it is thus easier to work with the DRO formulation than with the other formulations of adversarial training.

2.2 Discussion and literature review

The existence of measurable "robust" solutions to optimisation problems has been a topic of interest not only in the context of adversarial training of classification problem [2, 3, 19, 20, 35] but also in the general DRO literature, e.g., [9–11, 28]. Most previous studies of robust classifiers use *the universal σ-algebra* not only to formulate optimisation problems rigorously, but also as a feasible search space for robust classifiers. The proofs of these existence results rely on the pointwise topology of a sequence of universally measurable sets, the weak topology on the space of probability measures, and lower semi-continuity properties of $R^\star_{closed}(\,\cdot\,;\varepsilon)$. The (universal) measurability of a minimiser is then guaranteed immediately by the definition of the universal σ-algebra. We want to emphasise that all the works [2, 3, 19, 20, 35] prove their results in the binary ($K = 2$) classification setting where $\mathscr{X}$ is a subset of Euclidean space.

In contrast to the closed-ball model formulation, the objective in (5) is well defined for all Borel probability measures $\widetilde{\mu}$ and all $f \in \mathscr{F}$, as has been discussed in previous sections. The papers [2, 3, 35] can only relate, in the binary setting, problems (5) and (9) when problem (5) is appropriately extended to the universal σ-algebra, yet it is not clear that such extension is necessary. Recently, the authors of [19, 20] remove this technical redundancy and prove the existence of Borel robust classifiers (and consistency)

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

for the binary classification over Euclidean space. However, it is not obvious that the proof techniques in those papers can be extended to the multiclass setting. The proof technique that we implement in this paper allows us to consider all cases at once, while at the same time allows us to reveal the connections between many of the different models for adversarial training that exist in the literature.

For concreteness, we summarise some related results in the literature in the following theorem. Let $\overline{\mathscr{F}}$ denote the set of all $f = (f_1, \ldots, f_K)$ for which each $0 \leq f_i \leq 1$ is universally measurable and $\sum f_i \leq 1$.

Theorem 2.10. [2, 3, 19, 20, 35]. *Suppose $K = 2$ and $\overline{\mu} \in \overline{\mathscr{P}}(\mathscr{Z})$. Then, for any $f \in \mathscr{F}$, we have* $\sup_{\widetilde{x} \in \overline{B}_\varepsilon(x)} \{1 - f_i(\widetilde{x})\} \in \mathscr{U}(\mathscr{Z})$ *and*

$$\sum_{i=1}^{2} \int_{\mathscr{X}} \sup_{\widetilde{x} \in \overline{B}_\varepsilon(x)} \{1 - f_i(\widetilde{x})\} d\overline{\mu}_i(x) = \sup_{\widetilde{\mu} \in \mathscr{P}(\mathscr{Z})} \{R(f, \widetilde{\mu}) - C(\overline{\mu}, \widetilde{\mu})\},$$

where C is defined in terms of the cost function c from (7).

Assume further that $(\mathscr{X}, d) = (\mathbb{R}^d, || \cdot ||)$. Then, it holds that $\sup_{\widetilde{x} \in \overline{B}_\varepsilon(\cdot)} \{1 - f_i(\widetilde{x})\}$ *is universally measurable for any $f \in \overline{\mathscr{F}}$. Also, there exists a universally measurable minimiser ([2, 3, 35]) and a Borel measurable minimiser ([19, 20]). Finally, the values of (9) and (5) for the binary setting coincide.*

In this paper, we use the universal σ-algebra only for the rigorous description of the objective function in (9). Instead, we will focus on Borel measurable classifiers in the adversarial training in the multiclass setting. First, based on some of our previous results in [21], we will prove the existence of Borel measurable robust classifiers of (5) for general lower semi-continuous c satisfying Assumption 6 only in the multiclass setting. This extension is not immediate from all previous results since most techniques there are tailored for the binary setting. Our proof, however, is much simpler even in this general setting. Employing the combination of duality, an approximation argument, and well known theorems of optimal transport literature, we will be able to achieve our goal. Then, back to the closed-ball model, we prove the existence of Borel robust classifiers of (9) by proving that solutions to the DRO model with cost c_ε are also solutions to the closed-ball model problem. Furthermore, based on a simple but important observation connecting (9) and (3), Theorems 2.7 and 2.8 imply that (5), (9) and (3) are indeed equivalent (at least for almost all values of adversarial budget) in the sense that if $f^\star$ is a Borel minimiser of one of them, then it is a minimiser of others automatically, which is a new and stronger result in the adversarial training literature.

When we specialise our results to the binary classification setting (i.e., $K = 2$), we obtain the following improvement upon the results from [7, 20, 34].

Corollary 2.11. *Let $K = 2$ and let $f^\star \in \mathscr{F}$ be any solution to the problem (9). Then, for Lebesgue a.e. $t \in [0, 1]$, the pair $(\mathbb{1}_{\{f_1^\star \geq t\}}, \mathbb{1}_{\{f_1^\star \geq t\}^c})$ is also a solution to (9).*

In particular, there exist solutions to the problem

$$\min_{A \in \mathfrak{B}(\mathscr{X})} \int_{\mathscr{X}} \sup_{\tilde{x} \in \overline{B}_\varepsilon(x)} \mathbb{1}_{A^c}(\tilde{x}) d\overline{\mu}_1(x) + \int_{\mathscr{X}} \sup_{\tilde{x} \in \overline{B}_\varepsilon(x)} \mathbb{1}_A(\tilde{x}) d\overline{\mu}_2(x).$$

Notice that Corollary 2.11 implies, for the binary setting, the existence of robust hard classifiers for the adversarial training problem, a property shared with the standard risk minimisation Problem (2) that we discussed at the beginning of Section 2. Analogous results on the equivalence of the hard-classification and soft-classification problems in adversarial training under the binary setting have been obtained in [15, 24, 34, 35]. Unfortunately, when the number of classes is such that $K > 2$, the hard-classification and soft-classification problems in adversarial training may not be equivalent, as has been discussed in [21, Section 5.2].

In light of Theorem 2.8, one can conclude from Corollary 2.11 that for all but countably many $\varepsilon > 0$ the problem

$$\min_{A \in \mathfrak{B}(\mathscr{X})} \int_{\mathscr{X}} \sup_{\tilde{x} \in B_\varepsilon(x)} \mathbb{1}_{A^c}(\tilde{x}) d\mu_1(x) + \int_{\mathscr{X}} \sup_{\tilde{x} \in B_\varepsilon(x)} \mathbb{1}_A(\tilde{x}) d\mu_2(x)$$

admits solutions; notice that the above is the open-ball version of the optimisation problem in Corollary 2.11. However, notice that the results in [15] guarantee existence of solutions for *all* values of ε. It is interesting to note that the technique used in [15] cannot be easily adapted to the multiclass case $K > 2$. Specifically, it does not seem to be straightforward to generalise [15, Lemma C.1] to the multiclass case since it implicitly relies on the fact that one can always find an optimal hard classifier, i.e. an optimal decision boundary in the binary setting. Since for $f = (f_1, f_2)$, $f_2 = 1 - f_1$, f is basically a real-valued function f_1. Tailoring f_1 suitably, one can replace sup by ess sup (inf by ess inf similarly), and use L^p spaces, which provide an appropriate topology for the purpose. For example, however, if one used the aforementioned lemma to modify the coordinate functions f_i of a multiclass classifier f, one could end up producing functions for which their sum may be greater than one for some points in $\mathscr{X}$, thus violating one of the conditions for belonging to $\mathscr{F}$.

We observe, on the other hand, that the total variation regularisation interpretation for the open ball model in the binary setting discussed in Ref. [15] continues to hold in the multiclass case. To make this connection precise, let us introduce the non-local TV functionals:

$$\widetilde{\mathrm{TV}}_\varepsilon(f_i, \mu_i) := \frac{1}{\varepsilon} \sum_{i \in \mathscr{Y}} \int_{\mathscr{X}} \sup_{\widetilde{x} \in B_\varepsilon(x)} \{f_i(x) - f_i(\widetilde{x})\} d\mu_i(x).$$

It is then straightforward to show that Problem (3) is equivalent to

$$\inf_{f \in \mathscr{F}} \sum_{i=1}^{K} \int_{\mathscr{X}} (1 - f_i(x)) d\mu_i(x) + \varepsilon \sum_{i=1}^{K} \widetilde{\mathrm{TV}}_\varepsilon(f_i, \mu_i), \tag{10}$$

which can be interpreted as a total variation minimisation problem with fidelity term. Indeed, the fidelity term in the above problems is the standard (unrobust) risk $R(f, \mu)$. On the other hand, the functional $\widetilde{\mathrm{TV}}_\varepsilon(\cdot, \mu_i)$ is a non-local total variation functional in the sense that it is convex, positive 1-homogeneous, invariant under addition of constants to the input function and is equal to zero when its input is a constant function. Moreover, in the case $(\mathscr{X}, d) = (\mathbb{R}^d, \|\cdot\|)$ and when $d\mu_i(x) = \rho_i(x)dx$ for a smooth function ρ_i, one can see that, for small $\varepsilon > 0$,

$$\widetilde{\mathrm{TV}}_\varepsilon(f_i, \mu_i) \approx \int_{\mathscr{X}} |\nabla f_i(x)| \, \rho_i(x) dx,$$

when f_i is a smooth enough function. The functional $\widetilde{\mathrm{TV}}_\varepsilon(f_i, \mu_i)$ is thus connected to more standard notions of (weighted) total variation in Euclidean space. This heuristic can be formalised further via variational tools, as has been done recently in Ref. [16].

Total variation regularisation with general TV functionals is an important methodology in imaging, and also in unsupervised and supervised learning on graphs, where it has been used for community detection, clustering, and graph trend-filtering; e.g., see [5, 14, 17, 18, 23, 25, 27, 29, 31, 37] and references therein.

3. Distributional-perturbing model and its generalised barycenter formulation

In this section we introduce some tools and develop a collection of technical results that we use in Section 4 when proving Theorem 2.5.

3.1 Generalized barycenter and MOT problems

In our work [21], we introduced the following *generalised barycenter problem*. Given $\mu \in \mathscr{P}(\mathscr{Z})$, we consider the optimisation problem

$$\inf_{\lambda, \widetilde{\mu}_1, \ldots, \widetilde{\mu}_K} \left\{ \lambda(\mathscr{X}) + \sum_{i \in \mathscr{Y}} C(\mu_i, \widetilde{\mu}_i) \colon \lambda \geq \widetilde{\mu}_i \text{ for all } i \in \mathscr{Y} \right\}. \tag{11}$$

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

In the above, the infimum is taken over positive (Borel) measures $\tilde{\mu}_1, \ldots, \tilde{\mu}_K$ and λ satisfying the constraints $\lambda \geq \tilde{\mu}_i$ for all $i \in \mathscr{Y}$. This constraint must be interpreted as: $\lambda(A) \geq \tilde{\mu}_i(A)$ for all $A \in \mathfrak{B}(\mathscr{X})$. Problem (11) can be understood as a generalisation of the standard (Wasserstein) barycenter problem studied in [1]. Indeed, if all measures $\mu_1, \ldots, \mu_K$ had the same total mass and the term $\lambda(\mathscr{X})$ in (11) was rescaled by a constant $\alpha \in (0, \infty)$, then, as $\alpha \to \infty$, the resulting problem would recover the classical barycenter problem with pairwise cost function c. As stated, one can regard (11) as a partial optimal transport barycenter problem: we transport each μ_i to a part of λ while requiring the transported masses to overlap as much as possible (this is enforced by asking for the term $\lambda(\mathscr{X})$ to be small). For the canonical example c defined in (7), allowable $\widetilde{\mu}_i$ is a positive measure with the same mass of μ_i such that its support is away from the support of μ_i as most ε almost-surely (otherwise, (11) becomes ∞). Then, as long as $C(\mu_i, \widetilde{\mu}_i) = 0$, the optimal strategy to minimise (11) is to reduce $\lambda(\mathscr{X})$, i.e. to enlarge the overlap between $\widetilde{\mu}_i$ and $\widetilde{\mu}_j$ as much as possible, which engenders smaller λ.

We recall a result from Ref. [21], which essentially states that the generalised barycenter Problem (11) is dual to (5).

Theorem 3.1. [21, Proposition 7 and Corollary 32]. *Suppose that c satisfies Assumption 2.3. Then*

$$(5) = 1 - (11).$$

Furthermore, the infimum of (11) is attained. In other words, there exists $(\lambda^\star, \widetilde{\mu}^\star)$ which minimises (11).

Like classical barycenter problems, (11) has an equivalent MOT formulation. To be precise, we use a *stratified* multimarginal optimal transport problem to obtain an equivalent reformulation of (11).

Theorem 3.2. [21, Proposition 14 and Proposition 15]. *Suppose that c satisfies Assumption 2.3. Let $S_K := \{A \subseteq \mathscr{Y} : A \neq \emptyset\}$. Given $A \in S_K$, define $c_A : \mathscr{X}^K \to [0, \infty]$ as $c_A(x_1, \ldots, x_K) := \inf_{x' \in \mathscr{X}} \sum_{i \in A} c(x', x_i)$.*

Consider the problem:

$$\begin{aligned} &\inf_{\{\pi_A : A \in S_K\}} \sum_{A \in S_K} \int_{\mathscr{X}^K} \big(c_A(x_1, \ldots, x_K) + 1\big) d\pi_A(x_1, \ldots, x_K) \\ &\text{s.t.} \sum_{A \in S_K(i)} \mathscr{P}_{i\#}\pi_A = \mu_i \text{ for all } i \in \mathscr{Y}, \end{aligned} \tag{12}$$

where $\mathscr{P}_i$ is the projection map $\mathscr{P}_i : (x_1, \ldots, x_K) \mapsto x_i$, and $S_K(i) := \{A \in S_K : i \in A\}$. Then (11)=(12). Also, the infimum in (12) is attained.

Remark 3.3. Even though c_A and π_A above are defined over $\mathscr{X}^K$, only the coordinates i where $i \in A$ actually plays a role in the optimisation problem. Also, notice that (12) is not a standard MOT problem since in (12) we optimise over several couplings π_A (each with its own cost function c_A) that are connected to each other via the marginal constraints. We refer to this type of problem as a stratified MOT problem.

In the following theorem, we discuss the duals of the generalised barycenter problem and its MOT formulation. The notions of c-transform and $\overline{c}$-transform, whose definition we revisit in Appendix B, play an important role in these results.

Theorem 3.4. [21, Proposition 22 and Proposition 24]. *Suppose that c satisfies Assumption 2.3. Let $\mathscr{C}_b(\mathscr{X})$ be the set of bounded real-valued continuous functions over $\mathscr{X}$. The dual of (11) is*

$$\begin{aligned} &\sup_{f_1, \ldots, f_K \in \mathscr{C}_b(\mathscr{X})} \sum_{i \in \mathscr{Y}} \int_{\mathscr{X}} f_i^c(x_i) d\mu_i(x_i) \\ &\text{s.t.} f_i(x) \geq 0, \sum_{i \in \mathscr{Y}} f_i(x) \leq 1, \text{ for all } x \in \mathscr{X}, \; i \in \{1, \ldots, K\}, \end{aligned} \tag{13}$$

and there is no duality gap between primal and dual problems. In other words, (11)=(13). In the above, f_i^c denotes the c-transform of f_i as introduced in Definition 3.

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

The dual of (12) is

$$\begin{aligned}&\sup_{g_1,\ldots,g_K\in\mathscr{C}_b(\mathscr{X})}\sum_{i\in\mathscr{Y}}\int_{\mathscr{X}} g_i(x_i)d\mu_i(x_i)\\ &\text{s.t.} \sum_{i\in A} g_i(x_i)\leq 1+c_A(x_1,\ldots,x_K) \text{ for all } (x_1,\ldots,x_K)\in\mathscr{X}^K,\ \ A\in S_K,\end{aligned}\tag{14}$$

and there is no duality gap between primal and dual problems. In other words, (12)=(14).

If, in addition, the cost function c is bounded and Lipschitz, then (14) is achieved by $g\in\mathscr{C}_b(\mathscr{X})^K$. Also, for f feasible for (13), $g':=f^c$ is feasible for (14). Similarly, for g feasible for (14), $f'=\max\{g,0\}^{\bar{c}}$ is feasible for (13). Therefore, the optimisation of (14) can be restricted to non-negative g satisfying $g_i=g_i^{\bar{c}c}$, or $0\leq g_i\leq 1$ for all $i\in\mathscr{Y}$. The notion of $\bar{c}$-transform is also introduced in Definition 3.

Remark 3.5. By combining Theorem 3.1, Theorem 3.2 and Theorem 3.4, we conclude that 1-(14)=(5).

Remark 3.6. A standard argument in optimal transport theory shows that problem (14) is equivalent to

$$\sup_{g_1,\ldots,g_K}\sum_{i\in\mathscr{Y}}\int_{\mathscr{X}} g_i(x_i)d\mu_i(x_i),\tag{15}$$

where the sup is taken over all $(g_1,\ldots,g_K)\in\prod_{i\in\mathscr{Y}}L^\infty(\mathscr{X};\mu_i)$ satisfying: for any $A\in S_K$,

$$\sum_{i\in A} g_i(x_i)\leq 1+c_A(x_1,\ldots,x_K)$$

for $\otimes_i\mu_i$-almost every tuple $(x_1,\ldots,x_K)$. Here, $L^\infty(\mathscr{X};\mu_i)$ is defined as:

$$L^\infty(\mathscr{X};\mu_i):=\left\{f:\mathscr{X}\to\mathbb{R}\text{: measurable, ess }\sup_{\mu_i}|f|<\infty\right\}.$$

Indeed, notice that since (14) has already been shown to be equal to (12), the claim follows from the observation that any feasible $g_1,\ldots,g_K$ for (15) satisfies the condition

$$\sum_{i\in\mathscr{Y}}\int_{\mathscr{X}} g_i(x_i)d\mu_i(x_i)\leq\sum_{A\in S_K}\int_{\mathscr{X}^K}(1+c_A(x_1,\ldots,x_K))d\pi_A(x_1,\ldots,x_K)$$

for every $\{\pi_A\}_{A\in S_K}$ satisfying the constraints in (12).

3.2 Existence of optimal dual potentials g for general lower semi-continuous costs

We already know from the last part in Theorem 3.4 that if c is bounded and Lipschitz, then there is a feasible $g\in\mathscr{C}_b(\mathscr{X})^K$ that is optimal for (14). In this subsection, we prove an analogous existence result in the case of a general lower semi-continuous cost function c satisfying Assumption 2.3. More precisely, we prove existence of maximisers for (15). We start with an approximation result.

Lemma 3.7. *Let c be a cost function satisfying Assumption 2.3. For each $n\in\mathbb{N}$, let*

$$c_n(x,x'):=\min\{\tilde{c}_n(x,x'),n\},$$

where

$$\tilde{c}_n(x,x'):=\inf_{(\tilde{x},\tilde{x}')\in\mathscr{X}\times\mathscr{X}}\{c(\tilde{x},\tilde{x}')+nd(x,\tilde{x})+nd(x',\tilde{x}')\}.$$

Then the following properties hold:

1. *c_n is bounded and Lipschitz.*
2. *$c_n\leq c_{n+1}\leq c$ and $\tilde{c}_n\leq\tilde{c}_{n+1}$ for all $n\in\mathbb{N}$.*
3. *$\lim_{n\to\infty}c_n(x,x')=c(x,x')$ for all $(x,x')\in\mathscr{X}\times\mathscr{X}$.*

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

Proof. First, it is easy to see that $\tilde{c}_n$ is Lipschitz. Truncating it by n, finally, we obtain a bounded Lipschitz cost function c_n. Hence, items (1) and (2) are straightforward to prove. To prove item (3), notice that due to the monotonicity of the cost functions, we know that $\lim_{n\to\infty} c_n(x, x')$ exists in $[0, \infty]$ and $\lim_{n\to\infty} c_n(x, x') \leq c(x, x')$. If $\lim_{n\to\infty} c_n(x, x') = \infty$, then we would be done. Hence, we may assume that $\lim_{n\to\infty} c_n(x, x') < \infty$. From the definition of c_n, it then holds that $\lim_{n\to\infty} \tilde{c}_n(x, x') = \lim_{n\to\infty} c_n(x, x') < \infty$. Let $(x_n, x_n{}') \in \mathscr{X} \times \mathscr{X}$ be such that

$$c(x_n, x_n{}') + nd(x, x_n) + nd(x', x_n{}') \leq \tilde{c}_n(x, x') + \frac{1}{n}.$$

Since $c(x_n, x_n{}') \geq 0$, the above implies that $\lim_{n\to\infty} d(x, x_n) = 0$ and $\lim_{n\to\infty} d(x', x_n{}') = 0$. Indeed, if this was not the case, then we would contradict $\lim_{n\to\infty} \tilde{c}_n(x, x') < \infty$. By the lower semi-continuity of the cost function c, we then conclude that

$$\begin{aligned} c(x, x') &\leq \liminf_{n\to\infty} c(x_n, x_n{}') \leq \liminf_{n\to\infty} c(x_n, x_n{}') + nd(x, x_n) + nd(x', x_n{}') \leq \liminf_{n\to\infty} \tilde{c}_n(x, x') \\ &= \lim_{n\to\infty} c_n(x, x') \leq c(x, x'), \end{aligned}$$

from where the desired claim follows. □

Lemma 3.8. *Let c be a cost function satisfying Assumption 2.3, and let c_n be the cost function defined in Lemma 3.7. For each $A \in S_K$, let*

$$c_{A,n}(x_A) := \inf_{x' \in \mathscr{X}} \sum_{i\in A} c_n(x', x_i), \text{ and } c_A(x_A) := \inf_{x' \in \mathscr{X}} \sum_{i\in A} c(x', x_i),$$

where we use the shorthand notation $x_A = (x_i)_{i\in A}$. Then $c_{A,n}$ monotonically converges towards c_A pointwise for all $A \in S_K$, as $n \to \infty$.

Proof. Fix $A \in S_K$ and $x_A := (x_i)_{i\in A} \in \mathscr{X}^{|A|}$. From Lemma 3.7, it follows $c_{A,n} \leq c_{A,n+1} \leq c_A$. Therefore, for a given x_A, $\lim_{n\to\infty} c_{A,n}(x_A)$ exists in $[0, \infty]$ and is less than or equal to $c_A(x_A)$. If the limit is ∞, we are done. We can then assume without the loss of generality that $\lim_{n\to\infty} c_{A,n}(x_A) < \infty$. We can then find sequences $\{x_{n,i}\}_{n\in\mathbb{N}}$, $\{x_{n,i}{}'\}_{n\in\mathbb{N}}$ and $\{x_n{}'\}_{n\in\mathbb{N}}$ such that for all large enough $n \in \mathbb{N}$

$$\sum_{i\in A} c(x_{n,i}{}', x_{n,i}) + n(\sum_{i\in A} (d(x_{n,i}, x_i) + d(x_{n,i}{}', x_n{}'))) \leq c_{A,n}(x_A) + \frac{1}{n}.$$

From the above, we derive that $\lim_{n\to\infty} d(x_{n,i}{}', x_n{}') = 0$ and $\lim_{n\to\infty} d(x_{n,i}, x_i) = 0$. Hence, it follows that $\limsup_{n\to\infty} c(x_{n,i}{}', x_{n,i}) < \infty$. Combining the previous facts with Assumption 2.3, we conclude that $\{x_n{}'\}_{n\in\mathbb{N}}$ is precompact, and thus, up to subsequence of n's (that we do not relabel here), we have $\lim_{n\to\infty} d(x'_n, \hat{x}) = 0$ for some $\hat{x} \in \mathscr{X}$. Combining with $\lim_{n\to\infty} d(x_{n,i}{}', x_n{}') = 0$, we conclude that $\lim_{n\to\infty} d(x_{n,i}{}', \hat{x}) = 0$ for all $i \in A$. Using the lower semi-continuity of c, we conclude that

$$c_A(x_A) \leq \sum_{i\in A} c(\hat{x}, x_i) \leq \liminf_{n\to\infty} \sum_{i\in A} c(x_{n,i}{}', x_{n,i}) \leq \lim_{n\to\infty} c_{A,n}(x_A) \leq c_A(x_A).$$

□

Proposition 3.9. *Let c be a cost function satisfying Assumption 2.3. Then there exists a Borel solution for (15).*

Proof. Let $\{c_n\}_{n\in\mathbb{N}}$ be the sequence of cost functions introduced in Lemma 3.7. Notice that for each $n \in \mathbb{N}$, there is a solution $g^n = (g_1^n, \ldots, g_K^n) \in \mathscr{C}_b(\mathscr{X})^K$ for the Problem (14) (with cost c_n) that can be assumed to satisfy $0 \leq g_i^n \leq 1$ for each $i \in \mathscr{Y}$. Therefore, for each $i \in \mathscr{Y}$, the sequence $\{g_n^i\}_{n\in\mathbb{N}}$ is weakly⋆ precompact in $L^\infty(\mathscr{X}; \mu_i)$ by Lemma A.2. This implies that there exists a subsequence of $\{g^n\}_{n\in\mathbb{N}}$ (not relabelled) for which g^n weakly⋆ converges towards some $g^\star \in \prod_{i\in\mathscr{Y}} L^\infty(\mathscr{X}; \mathbb{R}, \mu_i)$, which would necessarily satisfy $0 \leq g_i^\star \leq 1$ for all $i \in \mathscr{Y}$; see Section A for the definition of weak⋆ topologies. We claim that this $g^\star$ is feasible for (15). Indeed, by Lemma 3.8, we know that $c_{A,n} \leq c_A$ for all $A \in S_K$. In particular, since $c_{A,n} \leq c_A$, and $\sum_{i\in A} g_i^n(x_i) \leq 1 + c_{A,n} \leq 1 + c_A$ for all $A \in S_K$ and all $n \in \mathbb{N}$, it follows that

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

$\sum_{i\in A} g_i^\star(x_i) \le 1 + c_A$, $\otimes_i\mu_i$-almost everywhere, due to the weak* convergence of g_i^n towards $g_i^\star$. This verifies that $g^\star$ is indeed feasible for (15).

Let α_n and β_n be the optimal values of (11) and (14), respectively, for the cost c_n. Likewise, let α and β be the optimal values of (11) and (14), respectively, for the cost c. Recall that, thanks to Theorem 3.2 and Theorem 3.4, we have $\alpha_n = \beta_n$ for all $n \in \mathbb{N}$ and $\alpha = \beta$. Suppose for a moment that we have already proved that $\lim_{n\to\infty} \alpha_n = \alpha$. Then we would have

$$\sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} g_i^\star(x) d\mu_i(x) = \lim_{n\to\infty} \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} g_i^n(x) d\mu_i(x) = \lim_{n\to\infty} \beta_n = \lim_{n\to\infty} \alpha_n = \alpha,$$

which would imply that $g^\star$ is optimal for (15).

It thus remains to show that $\lim_{n\to\infty} \alpha_n = \alpha$. Given that $c_n \le c_{n+1} \le c$, it follows that $\alpha_n \le \alpha_{n+1} \le \alpha$. In particular, the limit $\lim_{n\to\infty} \alpha_n$ exists in $[0, \infty]$ and must satisfy $\lim_{n\to\infty} \alpha_n \le \alpha$. If the limit is ∞, then there is nothing to prove. Thus, we can assume without the loss of generality that $\alpha_\infty := \lim_{n\to\infty} \alpha_n < \infty$.

Let λ^n and $\widetilde{\mu}_1^n, \dots, \widetilde{\mu}_K^n$ be an optimal solution of (11) with the cost c_n and let π_i^n be a coupling realising $C(\mu_i, \tilde{\mu}_i^n)$. We first claim that $\{\tilde{\mu}_i^n\}_{n\in\mathbb{N}}$ is weakly precompact for each $i \in \mathscr{Y}$. To see this, notice that for every n, we have $\tilde{\mu}_i^n(\mathscr{X}) = \mu_i(\mathscr{X}) \le 1$, for otherwise, $C(\mu_i, \widetilde{\mu}_i^n) = \infty$. Thus, by Prokhorov's theorem, it is enough to show that for every $\eta > 0$, there exists a compact set $\mathscr{K} \subseteq \mathscr{X}$ such that $\tilde{\mu}_i^n(\mathscr{X} \setminus \mathscr{K}) \le C\eta$ for all $n \in \mathbb{N}$ and some C independent of n, η or $\mathscr{K}$. To see that this is true, let us start by considering a compact set G such that $\mu_i(G^c) \le \eta$. Let $n_0 \in \mathbb{N}$ be such that $n_0 - 1 > \frac{1}{\eta}$. For $n \ge n_0$, we have

$$\alpha_\infty \ge \alpha_n = \lambda_n(\mathscr{X}) + \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} \int_{\mathscr{X}} c_n(x_i, \tilde{x}_i) d\pi_i^n(x_i, \tilde{x}_i) \ge \int_G \int_{\mathscr{X}} c_{n_0}(x_i, \tilde{x}_i) d\pi_i^n(x_i, \tilde{x}_i).$$

Consider the set

$$\tilde{\mathscr{K}} := \{x \in \mathscr{X} \text{ s.t. } \inf_{\tilde{x}\in G} c_{n_0}(x, \tilde{x}) \le n_0 - 1\};$$

using the definition of c_{n_0} and Assumption 2.3, it is straightforward to show that $\tilde{\mathscr{K}}$ is a compact subset of $\mathscr{X}$. Since $n_0 - 1 > \frac{1}{\eta}$, we see that $\alpha_\infty \ge \frac{1}{\eta}(\tilde{\mu}_i^n(\tilde{\mathscr{K}}^c) - \mu_i(G^c))$, from where we can conclude that $\tilde{\mu}_i^n(\tilde{\mathscr{K}}^c) \le (\alpha_\infty + 1)\eta$ for all $n \ge n_0$. We now consider a compact set $\hat{\mathscr{K}}$ for which $\tilde{\mu}_i^n(\hat{\mathscr{K}}^c) \le \eta$ for all $n = 1, \dots, n_0$, and set $\mathscr{K} := \tilde{\mathscr{K}} \cup \hat{\mathscr{K}}$, which is compact. Then, for all $n \in \mathbb{N}$, we have $\widetilde{\mu}_i^n(\mathscr{K}^c) \le (\alpha_\infty + 1)\eta$. This proves the desired claim.

Now, without the loss of generality, we can assume that λ^n has the form

$$d\lambda^n(x) = \max_{i=1,\dots,K} \left\{ \frac{d\widetilde{\mu}_i^n}{d\overline{\mu}^n}(x) \right\} d\overline{\mu}^n(x),$$

where $\overline{\mu}^n(x) = \sum_{i=1}^K \widetilde{\mu}_i^n$. Indeed, notice that the above is the smallest positive measure greater than $\tilde{\mu}_1^n, \dots, \tilde{\mu}_K^n$. Given the form of λ^n and the weak precompactness of each of the sequences $\{\tilde{\mu}_i^n\}_{n\in\mathbb{N}}$, we can conclude that $\{\lambda^n\}_{n\in\mathbb{N}}$ is weakly precompact and so are the sequences $\{\pi_i^n\}_{n\in\mathbb{N}}$. We can thus assume that, up to subsequence, $\tilde{\mu}_i^n$ converges weakly towards some $\widetilde{\mu}_i$; π_i^n converges weakly towards some $\pi_i \in \Gamma(\mu_i, \widetilde{\mu}_i)$ and λ^n converges weakly towards some λ satisfying $\lambda \ge \widetilde{\mu}_i$ for each $i \in \mathscr{Y}$. In particular, $\lambda, \tilde{\mu}_1, \dots, \tilde{\mu}_K$ is feasible for (11).

Therefore, for all $n_0 \in \mathbb{N}$, we have

$$\begin{aligned}
\alpha \ge \alpha_\infty &= \lim_{n\to\infty} \left(\lambda^n(\mathscr{X}) + \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} \int_{\mathscr{X}} c_n(x_i, \tilde{x}_i) d\pi_i^n(x_i, \tilde{x}_i) \right) \\
&\ge \lim_{n\to\infty} \left(\lambda^n(\mathscr{X}) + \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} \int_{\mathscr{X}} c_{n_0}(x_i, \tilde{x}_i) d\pi_i^n(x_i, \tilde{x}_i) \right) \\
&\ge \lambda(\mathscr{X}) + \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} \int_{\mathscr{X}} c_{n_0}(x_i, \tilde{x}_i) d\pi_i(x_i, \tilde{x}_i).
\end{aligned}$$

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

Sending $n_0 \to \infty$, we can then use the monotone convergence theorem to conclude that

$$\alpha \geq \alpha_\infty \geq \lambda(\mathscr{X}) + \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} \int_{\mathscr{X}} c(x_i, \tilde{x}_i) d\pi_i(x_i, \tilde{x}_i) \geq \lambda(\mathscr{X}) + \sum_{i\in\mathscr{Y}} C(\mu_i, \widetilde{\mu}_i) \geq \alpha.$$

This proves that $\alpha_\infty = \alpha$. □

3.3 From dual potentials to robust classifiers for continuous cost functions

Having discussed the existence of solutions $g^\star$ for (15), we move on to discussing the connection between $g^\star$ and solutions $f^\star$ of Problem (5).

Proposition 3.10 [21, Originally in Corollary 33]. *[22, see correction in Corollary 4.7 and Remark 4.9] Let $c : \mathscr{X} \times \mathscr{X} \to [0, \infty]$ be a lower semi-continuous function and suppose that $(\widetilde{\mu}^\star, g^\star)$ is a solution pair for the generalised barycenter problem (11) and the dual of its MOT formulation (15). Let $f^\star$ be defined as:*

$$f_i^\star(\widetilde{x}) := \max\left\{ \sup_{x\in \mathrm{spt}(\mu_i)} \left\{ g_i^\star(x) - c(x, \widetilde{x}) \right\}, 0 \right\}, \tag{16}$$

for each $i \in \mathscr{Y}$.

If $f^\star$ is Borel measurable, then $(f^\star, \widetilde{\mu}^\star)$ is a saddle solution for the problem (5). In particular, $f^\star$ is a minimiser of (5).

Remark 3.11. By the definition (16), $0 \leq f_i^\star(x) \leq 1$ since $0 \leq g_i^\star(x) \leq 1$. However, it is not trivial that $\sum f_i^\star(x) \leq 1$. In [22, Corollary 4.7] (originally [21, Corollary 33]), it is proved that $\sum f_i^\star(x) \leq 1$, i.e., $f^\star = (f_1^\star, \ldots, f_K^\star)$ is feasible.

The reason why we cannot directly use Proposition 3.10 to prove existence of solutions to (5) for arbitrary c and μ is because it is a priori not guaranteed that $f_i^\star$, as defined in (16), is Borel measurable; notice that the statement in Proposition 3.10 is conditional. If $\mathrm{spt}(\mu_i)$ was finite for all i, then the Borel measurability of $f_i^\star$ would follow immediately from the fact that the maximum of finitely many lower semi-continuous functions is Borel; this is of course the case when working with empirical measures. Likewise, the Borel measurability of $f_i^\star$ is guaranteed when μ is arbitrary and c is a bounded Lipschitz function (in fact, it is sufficient for the cost to be continuous), as is discussed in [39, Definitions 5.2 and 5.7 and Theorem 5.10]. However, nothing can be said about the Borel measurability of $f_i^\star$ without further information on $g_i^\star$ (which in general is unavailable) when c is only assumed to be lower semi-continuous (as is the case for the cost C from (7)) and $\mathrm{spt}(\mu_i)$ is an uncountable set.

Our strategy to prove Theorem 2.5 in Section 4.1 will be to approximate an arbitrary cost function c from below with a suitable sequence of bounded and Lipschitz cost functions c_n (the costs defined in Lemma 3.7), and, in turn, consider a limit of the robust classifiers $f_n^\star$ associated to each of the c_n. This limit (lim sup, to be precise) will be our candidate solution for (5).

4. Proofs of our main results

In this section, we prove the existence of a Borel measurable robust classifier for Problem (5) when c is an arbitrary lower semi-continuous cost function satisfying Assumption 2.3. We also establish the existence of minimisers of (9) and establish Theorem 2.8 and Corollary 2.11.

4.1 Well-posedness of the DRO model

Proof of Theorem 2.5. Let $\{c_n\}_{n\in\mathbb{N}}$ be the sequence of cost functions converging to c from below defined in Lemma 3.7. For each $n \in \mathbb{N}$, we use Theorem 3.4 and let $g^n = (g_1^n, \ldots, g_K^n) \in \mathscr{C}_b(\mathscr{X})^K$ be a solution

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

of (14) with cost c_n; recall that we can assume that $0 \leq g_i^n \leq 1$. In turn, we use g^n and the cost c_n to define $f^n := (f_1^n, \ldots, f_K^n)$ following (16). Since the g_i^n and c_n are continuous, and given that the pointwise supremum of a family of continuous functions is lower semi-continuous, we can conclude that f_i^n is lower semi-continuous and thus also Borel measurable for each $n \in \mathbb{N}$. Thanks to Proposition 3.10, f^n is optimal for (13) with cost function c_n.

From the proof of Proposition 3.9, we know that there exists a subsequence (that we do not relabel) such that the g_i^n converge in the weak* topology, as $n \to \infty$, towards limits $g_i^\star$ that form a solution for (15) with cost c. Using this subsequence and recalling (16), we define $f^\star \in \mathscr{F}$ according to

$$f_i^\star(\widetilde{x}) := \limsup_{n\to\infty} f_i^n(\widetilde{x}), \quad \tilde{x} \in \mathscr{X}. \tag{17}$$

Notice that each $f_i^\star$ is indeed Borel measurable since it is the lim sup of Borel measurable functions. In addition, notice that $0 \leq f_i^\star \leq 1$, due to the fact that $0 \leq f_i^n \leq 1$ for all $i \in \mathscr{Y}$ and all $n \in \mathbb{N}$. We'll conclude by proving that $f^\star$ is a solution for (5).

Let $\widetilde{\mu} \in \mathscr{P}(\mathscr{Z})$ be an arbitrary Borel probability measure with $C(\mu, \widetilde{\mu}) < \infty$. For each $i \in \mathscr{Y}$, let π_i be an optimal coupling realising the cost $C(\mu_i, \widetilde{\mu}_i)$. Then

$$\begin{aligned}
&R(f^\star, \widetilde{\mu}) - C(\mu, \widetilde{\mu}) \\
&= 1 - \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} f_i^\star(\widetilde{x}) d\widetilde{\mu}_i(\widetilde{x}) - \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}\times\mathscr{X}} c(x, \widetilde{x}) d\pi_i(x, \widetilde{x}) \\
&= 1 - \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}\times\mathscr{X}} \left(f_i^\star(\widetilde{x}) + c(x, \widetilde{x})\right) d\pi_i(x, \widetilde{x}) \\
&= 1 - \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}\times\mathscr{X}} \left(\limsup_{n\to\infty} \max \left\{ \sup_{x' \in \mathrm{spt}(\mu_i)} \{g_i^n(x') - c_n(x', \widetilde{x})\}, 0 \right\} + c(x, \widetilde{x}) \right) d\pi_i(x, \widetilde{x}) \\
&\leq 1 - \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}\times\mathscr{X}} \left(\limsup_{n\to\infty} \sup_{x' \in \mathrm{spt}(\mu_i)} \{g_i^n(x') - c_n(x', \widetilde{x})\} + c(x, \widetilde{x}) \right) d\pi_i(x, \widetilde{x})
\end{aligned}$$

where the last inequality follows from the simple fact that $-\max\{a, 0\} \leq -a$ for any $a \in \mathbb{R}$. Choosing $x' = x$ in the sup term (notice that indeed x can be assumed to belong to $\mathrm{spt}(\mu_i)$ since π_i has first marginal equal to μ_i), and applying reverse Fatou's lemma, we find that

$$\begin{aligned}
R(f^\star, \widetilde{\mu}) - C(\mu, \widetilde{\mu}) &\leq 1 - \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}\times\mathscr{X}} \limsup_{n\to\infty} \left\{ g_i^n(x) - c_n(x, \widetilde{x}) + c_n(x, \widetilde{x}) \right\} d\pi_i(x, \widetilde{x}) \\
&= 1 - \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}\times\mathscr{X}} \limsup_{n\to\infty} \left\{ g_i^n(x) \right\} d\pi_i(x, \widetilde{x}) \\
&= 1 - \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} \limsup_{n\to\infty} \left\{ g_i^n(x) \right\} d\mu_i(x) \\
&\leq 1 - \limsup_{n\to\infty} \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} g_i^n(x) d\mu_i(x) \\
&= 1 - \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} g_i^\star(x) d\mu_i(x) \\
&= R^\star_{DRO}(\varepsilon),
\end{aligned}$$

where the third equality follows from the weak* convergence of g_i^n towards $g_i^\star$ and the last equality follows from Remark 3.5 and the fact that $g^\star$ is a solution for (15) (combined with Remark 3.6). Taking the sup over $\widetilde{\mu} \in \mathscr{P}(\mathscr{Z})$, we conclude that

$$\sup_{\widetilde{\mu}\in\mathscr{P}(\mathscr{Z})} \{R(f^\star, \widetilde{\mu}) - C(\mu, \widetilde{\mu})\} \leq R^\star_{DRO}(\varepsilon),$$

and thus $f^\star$ is indeed a minimiser of (5).

Let now $\widetilde{\mu}^\star$ be a solution of (11) (which exists due to Theorem 3.1). The fact that $(\widetilde{\mu}^\star, f^\star)$ is a saddle for (5) follows from the above computations and the fact that by Theorem 3.1 and in [21, Corollary 32], we have

$$R^\star_{DRO}(\varepsilon) = \sup_{\widetilde{\mu}\in\mathscr{P}(\mathscr{Z})} \inf_{f\in\mathscr{F}} \{R(f, \widetilde{\mu}) - C(\mu, \widetilde{\mu})\} = \inf_{f\in\mathscr{F}} \{R(f, \widetilde{\mu}^\star) - C(\mu, \widetilde{\mu}^\star)\}.$$

□

The next proposition states that the function $g^\star_i$ constructed in the proof of Proposition 3.9 is a Borel measurable version of the c-transform of $f^\star_i$, where $f^\star_i$ was defined in (17).

Lemma 4.1. *Let $\{g^n\}_{n\in\mathbb{N}}$ and $\{f^n\}_{n\in\mathbb{N}}$ be as in the proof of Theorem 2.5, $g^\star$ be the weak* limit of the g^n and $f^\star$ be as defined in (17). Then, for every $i \in \mathscr{Y}$,*

$$g^\star_i(x) = \inf_{\widetilde{x}\in\mathscr{X}} \left\{ f^\star_i(\widetilde{x}) + c(x, \widetilde{x}) \right\} \tag{18}$$

for μ_i-a.e. $x \in \mathscr{X}$. This statement must be interpreted as: the set in which (18) is violated is contained in a Borel measurable set with zero μ_i measure.

Proof. From the proof of Theorem 2.5, it holds that for each $i \in \mathscr{Y}$

$$\int_{\mathscr{X}} f^\star_i(\widetilde{x}) d\widetilde{\mu}^\star_i(\widetilde{x}) + \int_{\mathscr{X}\times\mathscr{X}} c(x, \widetilde{x}) d\pi^\star_i(x, \widetilde{x}) = \int_{\mathscr{X}} g^\star_i(x) d\mu_i(x). \tag{19}$$

On the other hand, from the definition of f^n_i, it follows that

$$g^n_i(x) \leq f^n_i(\tilde{x}) + c_n(x, \tilde{x}), \quad \forall \tilde{x} \in \mathscr{X}, \text{ and } \mu_i\text{-a.e. } x \in \mathscr{X}.$$

We can then combine the above with Lemma A.3 to conclude that for μ_i-a.e. $x \in \mathscr{X}$ and every $\tilde{x} \in \mathscr{X}$, we have

$$g^\star_i(x) \leq \limsup_{n\to\infty} g^n_i(x) \leq \limsup_{n\to\infty} f^n_i(\tilde{x}) + c_n(x, \tilde{x}) = f^\star_i(\tilde{x}) + c(x, \tilde{x}).$$

Taking the inf over $\tilde{x} \in \mathscr{X}$, we conclude that for μ_i-a.e. $x \in \mathscr{X}$, we have

$$g^\star_i(x) \leq \inf_{\tilde{x}\in\mathscr{X}} \{f^\star_i(\tilde{x}) + c(x, \tilde{x})\}. \tag{20}$$

From this and (19), we see that $g^\star_i \in L^1(\mu_i)$ and $-f^\star_i \in L^1(\tilde{\mu}_i)$ are optimal dual potentials for the optimal transport problem $C(\mu_i, \widetilde{\mu}^\star_i)$. If (20) did not hold with equality for μ_i-a.e. $x \in \mathscr{X}$, then we would be able to construct a Borel measurable version h_i of the right-hand side of (20) (see Lemma C.1 in the Appendix), which would be strictly greater than $g^\star_i$ in a set of positive μ_i-measure. In addition, we would have that $(h_i, -f^\star_i)$ is a feasible dual pair for the OT problem $C(\mu_i, \tilde{\mu}_i)$. However, the above would contradict the optimality of the dual potentials $(g^\star_i, -f^\star_i)$. We thus conclude that (20) holds with equality except on a set contained in a set of μ_i measure zero. □

4.2 Well-posedness of the closed-ball model (9)

Proof of Theorem 2.7. We actually prove that for arbitrary cost c satisfying Assumption 2.3, the solution $f^\star$ to (5) constructed in the proof of Theorem 2.5 is also a solution for the problem:

$$\inf_{f\in\mathscr{F}} \left\{ \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} \sup_{\widetilde{x}\in\mathscr{X}} \{1 - f_i(\widetilde{x}) - c(x, \widetilde{x})\}\, d\overline{\mu}_i(x) \right\}. \tag{21}$$

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

Theorem 2.7 will then be an immediate consequence of this more general result when applied to $c = c_\varepsilon$ as in (7).

Let $f^\star$ be the Borel solution of (5) constructed in the proof of Theorem 2.5. It suffices to show that for any $f \in \mathscr{F}$

$$\sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} \inf_{\widetilde{x}\in\mathscr{X}} \left\{f_i^\star(\widetilde{x}) + c(x,\widetilde{x})\right\} d\overline{\mu}_i(x) \geq \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} \inf_{\widetilde{x}\in\mathscr{X}} \{f_i(\widetilde{x}) + c(x,\widetilde{x})\} \, d\overline{\mu}_i(x).$$

Suppose for the sake of contradiction that the above inequality does not hold. Then there exists some $\widehat{f} \in \mathscr{F}$ which provides a strict inequality in the opposite direction. Now, on one hand, (18) of Lemma 4.1 implies

$$\sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} \inf_{\widetilde{x}\in\mathscr{X}} \left\{f_i^\star(\widetilde{x}) + c(x,\widetilde{x})\right\} d\overline{\mu}_i(x) = \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} g_i^\star(x) d\mu_i(x).$$

On the other hand, by Lemma C.1, for each $i \in \mathscr{Y}$, there exists a Borel measurable function $\widehat{g}_i$ equal to $\inf_{\widetilde{x}\in\mathscr{X}} \{\widehat{f}_i(\widetilde{x}) + c(x,\widetilde{x})\}$ $\overline{\mu}_i$-almost everywhere. Let $\widehat{g} := (\widehat{g}_1, \ldots, \widehat{g}_K)$. Combining the existence of such $\widehat{g}$ with the above equation, and using (1), it follows that $\widehat{g}$ satisfies

$$\sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} g_i^\star(x) d\mu_i(x) < \sum_{i\in\mathscr{Y}} \int_{\mathscr{X}} \widehat{g}_i(x) d\mu_i(x). \tag{22}$$

Notice that for each $A \in S_K$ and $\otimes\overline{\mu}_i$-almost everywhere $x_1, \ldots x_K$, we have

$$\sum_{i\in A} \inf_{\widetilde{x}\in\mathscr{X}} \left\{\widehat{f}_i(\widetilde{x}) + c(x_i,\widetilde{x})\right\} \leq \inf_{\widetilde{x}\in\mathscr{X}} \left\{\sum_{i\in A} \widehat{f}_i(\widetilde{x}) + c(x_i,\widetilde{x})\right\} \leq 1 + c_A(x_1, \ldots, x_K).$$

From the above, we conclude that $\widehat{g}$ is feasible for (15). However, this and (22) combined contradict the fact that $g^\star$ is optimal for (15), as had been shown in Proposition 3.9. □

Proof of Corollary 2.11. It is straightforward to verify (e.g., see [15]) that for $(f_1, 1 - f_1) \in \mathscr{F}$, we can write

$$R^\star_{closed}((f_1, 1 - f_1);\varepsilon) = \int_0^1 R^\star_{closed}((\mathbb{1}_{\{f_1\geq t\}}, \mathbb{1}_{\{f_1\geq t\}^c});\varepsilon)dt. \tag{23}$$

It is also straightforward to see that

$$R^\star_{closed}((\mathbb{1}_{\{f_1\geq t\}}, \mathbb{1}_{\{f_1\geq t\}^c});\varepsilon) = \int_{\mathscr{X}} \sup_{\tilde{x}\in\overline{B}_\varepsilon(x)} \mathbb{1}_{A^c}(\tilde{x}) d\overline{\mu}_1(x) + \int_{\mathscr{X}} \sup_{\tilde{x}\in\overline{B}_\varepsilon(x)} \mathbb{1}_{A}(\tilde{x}) d\overline{\mu}_2(x).$$

Let $(f_1, 1 - f_1)$ be a solution to (9) (which by Remark 2.2 can indeed be taken of this form). It follows from (23) that for almost every $t \in [0, 1]$, the pair $(\mathbb{1}_{\{f_1\geq t\}}, \mathbb{1}_{\{f_1\geq t\}^c})$ is also a solution for that same problem and thus also for the problem restricted to hard classifiers. This proves the desired result. □

4.3 Connection between closed-ball model and open-ball model

Proof of Theorem 2.8. One can easily observe that for any fixed $\varepsilon > 0$ and $\delta > 0$, we have

$$\sup_{\widetilde{x}\in B_\varepsilon(x)} \{1 - f_i(\widetilde{x})\} \leq \sup_{\widetilde{x}\in\overline{B}_\varepsilon(x)} \{1 - f_i(\widetilde{x})\} \leq \sup_{\widetilde{x}\in B_{\varepsilon+\delta}(x)} \{1 - f_i(\widetilde{x})\}$$

for all $x \in \mathscr{X}$ and all $f \in \mathscr{F}$. This simple observation leads to $R^\star_{open}(f;\varepsilon) \leq R^\star_{closed}(f;\varepsilon) \leq R^\star_{open}(f;\varepsilon + \delta)$ for all $f \in \mathscr{F}$. Thus, we also have $R^\star_{open}(\varepsilon) \leq R^\star_{closed}(\varepsilon) \leq R^\star_{open}(\varepsilon + \delta)$, and, in particular, $R^\star_{open}(\varepsilon) \leq R^\star_{closed}(\varepsilon) \leq \liminf_{\delta\to 0} R^\star_{open}(\varepsilon + \delta)$. From the above, we can also see that the function $\varepsilon \mapsto R^\star_{open}(\varepsilon)$ is non-decreasing and, as such, is continuous for all but at most countably many values of $\varepsilon > 0$. Therefore, for all but at most countably many ε, we have $R^\star_{open}(\varepsilon) = R^\star_{closed}(\varepsilon)$.

Now, let $f^\star$ be solution of (9) and assume we have $R^\star_{open}(\varepsilon) = R^\star_{closed}(\varepsilon)$. Then

$$R^\star_{open}(f^\star;\varepsilon) \leq R^\star_{closed}(f^\star;\varepsilon) = R^\star_{closed}(\varepsilon) = R^\star_{open}(\varepsilon),$$

which means $f^\star$ is a solution of (3). □

5. Conclusion and future works

In this paper, we establish the equivalence of three popular models of multiclass adversarial training: the open ball model, the closed ball model and the DRO model, and, for the first time, (with the exception of partial results in [15]) we prove the existence of Borel measurable optimal robust classifiers in the agnostic-classifier setting for any number of classes. We are able to unify these models via a framework we have developed that connects these problems to optimal transport. Notably, our results show that it is unnecessary to grapple with the cumbersome machinery of universal sigma algebras, which was needed to prove existence of classifiers in past results.

Although our analysis sheds light on this area, many open questions still remain on both the theoretical and practical side. One of the most important practical questions is how to extend these results when the set of classifiers $\mathscr{F}$ is some parametric family, for example, neural networks. In particular, one would like to specify the properties a parametric family must satisfy in order to approximate robust classifiers to some desired degree of accuracy. In the case of neural networks, one might ask for the number of neurons or number of layers that are required for robust classification.

Related to the above practical question is the following geometric/theoretical question: given an optimal robust classifier $f^\star$, can we give a characterisation of the regularity of $f^\star$ as in [15]? In particular, one would like to quantify the smoothness of the interfaces between the different classes. In general, we cannot guarantee that $f^\star$ is a hard classifier, thus, this problem is best posed as a question about the smoothness of the level sets of $f^\star$. Since optimal classifiers need not be unique, one can also pose the more general question of when it is possible to find at least one optimal Borel robust classifier with some specified regularity property. Due to the connection between approximation and regularity, answering this question will provide insights to the previous question of how well one can approximate optimal robust classifiers using certain parametric families.

A final question is how to extend our framework to other more general settings. In this paper, we have assumed throughout that the loss function is the 0–1 loss. However, most practitioners prefer strongly convex loss functions, for example, the cross entropy function, which allows for faster optimisation and has other desirable properties. As a result, one would like to establish the analogue of these results in this more general setting. This would be crucial for bringing these theoretical insights closer to the models favoured by working practitioners.

Financial support. NGT is supported by the NSF grants DMS-2005797 and DMS-2236447. MJ is supported by the NSF grant DMS-2400641. JK thanks to PIMS Kantorovich Initiative supported through a PIMS PRN and NSF-DMS 2133244.

Competing interests. The authors declare none.

References

[1] Agueh, M. & Carlier, G. (2011) Barycenters in the wasserstein space. *Siam J. Math. Anal.* **43**(2), 904–924.

[2] Awasthi, P., Frank, N. & Mohri, M. (2021) On the existence of the adversarial bayes classifier. *Adv. Neural Inform. Process. Syst.* **34**, 2978–2990.

[3] Awasthi, P., Frank, N. S. & Mohri, M. (2021) On the existence of the adversarial bayes classifier. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S. & Wortman Vaughan, J. (editors), *Advances in Neural Information Processing Systems*, Vol. 34, Curran Associates, Inc, 2978–2990. https://proceedings.neurips.cc/paper_files/paper/2021/file/172ef5a94b4dd0aa120c6878fc29f70c-Paper.pdf

[4] Balcan, M.-F., Pukdee, R., Ravikumar, P. & Zhang, H. Nash equilibria and pitfalls of adversarial training in adversarial robustness games. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, 206 *Proceedings of Machine Learning Research*, 9607-9636. PMLR, (25-27 Apr, 2023).

[5] Bertozzi, A. L. & Flenner, A. (2016) Diffuse interface models on graphs for classification of high dimensional data. *Siam Rev.* **58** (2), 293–328.

[6] Bertsekas, D. P. & Shreve, S. E. (1978). *Stochastic Optimal Control: The Discrete-Time Case*, USA, Academic Press, Inc.

[7] Bhagoji, A. N., Cullina, D. & Mittal, P. (2019). Lower bounds on adversarial robustness from optimal transport. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. & Garnett, R. (editors), *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/02bf86214e264535e3412283e817deaa-Paper.pdf

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

[8] Biggio, B. & Roli, F. (2018) Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recogn.* **84**, 317–331.

[9] Blanchet, J., Kang, Y. & Murthy, K. (2019) Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probab.* **56** (3), 830–857.

[10] Blanchet, J., Kuhn, D., Li, J. & Taskesen, B. Unifying distributionally robust optimization via optimal transport theory. arXiv preprint arXiv: 2308.05414

[11] Blanchet, J. & Murthy, K. (2019) Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* **44**(2), 565–600.

[12] Bose, J., Gidel, G., Berard, H., et al (2020) Adversarial example games. *Adv. Neur. Inform. Process. Syst.* **33**, 8921–8934.

[13] Bousquet, O., Boucheron, S. & Lugosi, G. (2004). *Introduction to Statistical Learning Theory*, Springer, Berlin Heidelberg, pp. 169–207.

[14] Boyd, Z. M., Porter, M. A. & Bertozzi, A. L. (2020) Stochastic block models are a discrete surface tension. *J. Nonlinear Sci.* **30** (5), 2429–2462.

[15] Bungert, L., Trillos, N. García & Murray, R. (2023) *The geometry of adversarial training in binary classification* Information and Inference. *JIMA* **12** (2), 921–968.

[16] Bungert, L. & Stinson, K. (2022) Gamma-convergence of a nonlocal perimeter arising in adversarial machine learning. arXiv preprint arXiv:2211.15223.

[17] Caroccia, M., Chambolle, A. & Slepčev, D. (2020) Mumford–Shah functionals on graphs and their asymptotics. *Nonlinearity* **33** (8), 3846.

[18] Cristofari, A., Rinaldi, F. & Tudisco, F. (2020) Total variation based community detection using a nonlinear optimization approach. *SIAM J. Appl. Math.* **80**(3), 1392–1419.

[19] Frank, N. & Niles-Weed, J. (2023). The adversarial consistency of surrogate risks for binary classification. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M. & Levine, S. (editors), *Advances in Neural Information Processing Systems*, Vol. 36, Curran Associates, Inc, 41343–41354. https://proceedings.neurips.cc/paper_files/paper/2023/file/81858558b55a8c63763cfe088090242a-Paper-Conference.pdf

[20] Frank, N. S. & Niles-Weed, J. (2024). Existence and minimax theorems for adversarial surrogate risks in binary classification. *J. Mach. Learn. Res.* **25**(58), 1–41. http://jmlr.org/papers/v25/23-0456.html

[21] Garcıa Trillos, N., Jacobs, M. & Kim, J. (2023) The multimarginal optimal transport formulation of adversarial multiclass classification. *J. Mach. Learn. Res.* **24** (45), 1–56.

[22] Garcıa Trillos, N., Jacobs, M. & Kim, J. (2023) The multimarginal optimal transport formulation of adversarial multiclass classification. arXiv preprint, https://arxiv.org/abs/2204.12676

[23] García Trillos, N. & Murray, R. (2017) A new analytical approach to consistency and overfitting in regularized empirical risk minimization. *Eur. J. Appl. Math.* **28** (6), 886–921.

[24] García Trillos, N. & Murray, R. (2022) Adversarial classification: Necessary conditions and geometric flows. *J. Mach. Learn. Res.* **23**(187), 1–38.

[25] García Trillos, N., Murray, R. & Thorpe, M. (2022) From graph cuts to isoperimetric inequalities: Convergence rates of Cheeger cuts on data clouds. *Arch. Ration. Mech. Anal.* **3**, 541–598.

[26] Goodfellow, I., Shlens, J. & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International Conference on Learning Representations.

[27] Hu, H., Laurent, T., Porter, M. A. & Bertozzi, A. L. (2013) A method based on total variation for network modularity optimization using the MBO scheme. *SIAM J. Appl. Math.* **73**(6), 2224–2246.

[28] Javanmard, A., Soltanolkotabi, M. & Hassani, H. (2020). Precise tradeoffs in adversarial training for linear regression. In: Conference on Learning Theory, PMLR, 2034–2078.

[29] Luo, X. & Bertozzi, A. L. (2017) Convergence of the graph Allen–Cahn scheme. *J. Stat. Phys.* **167** (3), 934–958.

[30] Luzin, N. N. & Sierpiński, W. (1919). Sur quelques propriétés des ensembles (a).

[31] Merkurjev, E., Kostić, T. & Bertozzi, A. L. (2013, 1903-1930) An MBO scheme on graphs for classification and image processing. *Siam J. Imaging Sci.* **6** (4), 1903–1930.

[32] Meunier, L., Scetbon, M., Pinot, R. B., Atif, J. & Chevaleyre, Y. (2021) Mixed nash equilibria in the adversarial examples game. In: Meila, M., & Zhang, T. (editors), *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139, *Proceedings of Machine Learning Research*, 7677–7687. PMLR, 18–24 Jul, 2021.

[33] Nishiura, T. (2008). *Absolute Measurable Spaces. 120 Encyclopedia of Mathematics and Its Applications, Xiv+274*, Cambridge University Press, Cambridge.

[34] Pydi, M. S. & Jog, V. (2021) Adversarial risk via optimal transport and optimal couplings. *IEEE Trans. Inform Theory* **67**(9), 6031–6052.

[35] Pydi, M. S. & Jog, V. (2021). The many faces of adversarial risk. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W. (editors), *Advances in Neural Information Processing Systems*, Vol. **34**, Curran Associates, Inc, 10000–10012.

[36] Trillos, N. G., Jacobs, M., Kim, J. & Werenski, M. (2024). An optimal transport approach for computing adversarial training lower bounds in multiclass classification.

[37] van Gennip, Y., Guillen, N., Osting, B. & Bertozzi, A. L. (2014) Mean curvature, threshold dynamics, and phase field theory on finite graphs. *Milan J. Math.*, **82**, 3–65.

[38] Villani, C. (2003). Topics in optimal transportation, *Graduate Studies in Mathematics, xvi+370*, American Mathematical Society, Providence, RI

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

[39] Villani, C. (2009). Optimal transport, *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Xxii+973*, Springer-Verlag, Berlin, Old and new
[40] von Luxburg, U. & Schölkopf, B. (2011). Statistical learning theory: Models, concepts, and results. In: Gabbay, D. M., Hartmann, S. & Woods, J. (editors), *Inductive Logic, 10 Handbook of the History of Logic*, North-Holland, 651–706.

Appendix A. Weak⋆ topology

Definition 2 (*Weak*⋆ topology). Let $\mu = (\mu_1, \ldots, \mu_K) \in \prod_{i=1}^K \mathscr{M}_+(\mathscr{X})$. For a sequence $\{h^n\}_{n\in\mathbb{N}} \subseteq \prod_{i\in\mathscr{Y}} L^\infty(\mathscr{X};\mu_i)$, we say that $\{h^n\}$ weak⋆-converges to $h \in \prod_{i\in\mathscr{Y}} L^\infty(\mathscr{X};\mu_i)$ if for any $q \in \prod_{i\in\mathscr{Y}} L^1(\mathscr{X};\mu_i)$, it holds that

$$\lim_{n\to\infty} \int_{\mathscr{X}} h_i^n(x) q_i(x) d\mu_i(x) = \int_{\mathscr{X}} h_i(x) q_i(x) d\mu_i(x) \tag{A1}$$

for all $i \in \mathscr{Y}$.

Remark A.1. Note that for a Borel positive measure ρ, which is either finite or σ-finite over a Polish space, the dual of $L^1(\rho)$ is $L^\infty(\rho)$, which justifies the definition (1).

The following lemma is the weak⋆ precompactness of the closed unit ball of the dual space $\mathscr{X}'$. In our case, $\mathscr{X} = L^1(\mathscr{X};\mu_i)$ and $\mathscr{X}' = L^\infty(\mathscr{X};\mu_i)$ with weak⋆ topology.

Lemma A.2. (Banach–Alaoglu theorem). *If $\mathscr{X}$ is a normed space, then the closed unit ball in the continuous dual space $\mathscr{X}'$ (endowed with its usual operator norm) is compact with respect to the weak⋆ topology.*

Lemma A.3. *Suppose $\{g_i^n\}_{n\in\mathbb{N}}$ is a sequence of measurable real-valued functions over $\mathscr{X}$ satisfying $0 \le g_i^n \le 1$ for every $n \in \mathbb{N}$. Suppose that g_i^n converges in the weak⋆ topology of $L^\infty(\mathscr{X};\mu_i)$ towards g_i, where μ_i is a finite positive measure. Then, for μ_i-a.e. $x \in X$, we have*

$$\limsup_{n\to\infty} g_i^n(x) \ge g_i(x).$$

Proof. Let E be a measurable subset of $\mathscr{X}$. Then

$$\int_{\mathscr{X}} (\limsup_{n\to\infty} g_i^n(x) - g_i(x))\mathbb{1}_E(x) d\mu_i(x) \ge \limsup_{n\to\infty} \int_{\mathscr{X}} (g_i^n(x) - g_i(x))\mathbb{1}_E(x) d\mu_i(x) = 0,$$

by the reverse Fatou inequality and the assumption that the sequence $\{g_i^n\}_{n\in\mathbb{N}}$ converges in the weak⋆ sense towards g_i. Since E was arbitrary, the result follows. □

Appendix B. *c*-transform

c-transform has an important role in optimal transport theory. One can characterise an optimiser of a dual problem by iterating *c*-transform: see [38, 39] for more details.

Definition 3 (39, *Definition* 5.2 *and Definition* 5.8). Let $\mathscr{X}$, $\mathscr{X}'$ be measurable spaces, and let $c: \mathscr{X} \times \mathscr{X}' \to (-\infty, \infty]$. Given a measurable function $h: \mathscr{X} \to \mathbb{R} \cup \{\infty, -\infty\}$, its *c*-transform is defined as:

$$h^c(x') := \inf_{x\in\mathscr{X}} \{h(x) + c(x, x')\}.$$

Similarly, for $g: \mathscr{X}' \to \mathbb{R} \cup \{\infty, -\infty\}$, its $\bar{c}$-transform is defined as:

$$g^{\bar{c}}(x) := \sup_{x'\in\mathscr{X}'} \{g(x') - c(x, x')\}.$$

Proposition B.1. *For any mearurable functions h over $\mathscr{X}$ and g over $\mathscr{X}'$, and cost function $c: \mathscr{X} \times \mathscr{X}' \to (-\infty, \infty]$, it holds that for every $(x, x') \in \mathscr{X} \times \mathscr{X}'$,*

$$h^c(x') - h(x) \le c(x, x'), \quad g(x') - g^{\bar{c}}(x) \le c(x, x').$$

https://doi.org/10.1017/S0956792524000822 Published online by Cambridge University Press

Theorem B.2. [39, Theorem 5.10] *Let $\mathscr{X}$ be a Polish space and $c(\cdot,\cdot)$ be a cost function bounded from below and lower semi-continuous. Then, for $\nu, \widetilde{\nu} \in \mathscr{P}(\mathscr{X})$,*

$$\begin{aligned}\inf_{\pi_i \in \Gamma(\nu,\widetilde{\nu})} \int_{\mathscr{X}\times\mathscr{X}} c(x,\widetilde{x}) d\pi_i(x,\widetilde{x}) &= \sup_{g_i, f_i \in \mathscr{C}_b, g_i - f_i \le c} \left\{ \int_{\mathscr{X}} g_i(x) d\nu(x) - \int_{\mathscr{X}} f_i(\widetilde{x}) d\widetilde{\nu}(\widetilde{x}) \right\} \\ &= \sup_{f_i \in L^1(\widetilde{\nu})} \left\{ \int_{\mathscr{X}} (f_i)^c(x) d\nu(x) - \int_{\mathscr{X}} f_i(\widetilde{x}) d\widetilde{\nu}(\widetilde{x}) \right\} \\ &= \sup_{g_i \in L^1(\nu)} \left\{ \int_{\mathscr{X}} g_i(x) d\nu(x) - \int_{\mathscr{X}} (g_i)^{\overline{c}}(\widetilde{x}) d\widetilde{\nu}(\widetilde{x}) \right\}.\end{aligned}$$

Furthermore, the infimum is indeed a minimum. However, the supremum may not be achieved.

Appendix C. Decomposition of universally measurable functions

The following lemma is a well known fact about measure theory. For the sake of completeness, we write the full proof.

Lemma C.1. *Let $\mathscr{X}$ be a polish space, and let μ and $\tilde{\mu}$ be a Borel probability measure and its extension to the universal σ-algebra, respectively. Let f be a universally measurable function for which $\int_{\mathscr{X}} |f(x)| d\overline{\mu}(x) < \infty$. Then there exists a Borel measurable function g such that $f = g$ $\overline{\mu}$-almost everywhere. Also,*

$$\int_{\mathscr{X}} f(x) d\overline{\mu}(x) = \int_{\mathscr{X}} g(x) d\mu(x). \tag{C1}$$

Proof. Without the loss of generality, we can assume that $f \ge 0$. Since f is universally measurable, we can write

$$f(x) = \lim_{n\to\infty} f_n(x) := \lim_{n\to\infty} \sum_{k=1}^{n} c_k^n \mathbb{1}_{A_k^n}(x),$$

for positive coefficients $c_1^n, \ldots, c_n^n$ and $A_1^n, \ldots, A_n^n$ universally measurable and pairwise disjoint sets. By the definition of universally measurable sets, for each A_k^n, there exists a Borel set B_k^n such that $\overline{\mu}(A_k^n \setminus B_k^n) = 0$. Hence, for each $n \in \mathbb{N}$, we can write

$$f_n(x) = \sum_{k=1}^{n} c_k^n \mathbb{1}_{B_k^n}(x) + \sum_{k=1}^{n} c_k^n \mathbb{1}_{C_k^n}(x),$$

where $C_k^n = A_k^n \setminus B_k^n$. We conclude that

$$f(x) = g(x) + h(x) := \limsup_{n\to\infty} \sum_{k=1}^{n} c_k^n \mathbb{1}_{B_k^n}(x) + \liminf_{n\to\infty} \sum_{k=1}^{n} c_k^n \mathbb{1}_{C_k^n}(x)$$

where g is Borel measurable, h is universally measurable and $h = 0$ $\overline{\mu}$-almost everywhere.

Since $f = g$ $\overline{\mu}$-almost everywhere and g is Borel measurable, then

$$\int_{\mathscr{X}} f(x) d\overline{\mu}(x) = \int_{\mathscr{X}} g(x) d\overline{\mu}(x) = \int_{\mathscr{X}} g(x) d\mu(x),$$

from which (1) follows. □

Cite this article: García Trillos N., Jacobs M. and Kim J. On the existence of solutions to adversarial training in multiclass classification. *European Journal of Applied Mathematics*, https://doi.org/10.1017/S0956792524000822