# THE MULTIMARGINAL OPTIMAL TRANSPORT FORMULATION OF ADVERSARIAL MULTICLASS CLASSIFICATION

NICOLÁS GARCÍA TRILLOS, MATT JACOBS, AND JAKWANG KIM

ABSTRACT. We study a family of adversarial multiclass classification problems and provide equivalent reformulations in terms of: 1) a family of generalized barycenter problems introduced in the paper and 2) a family of multimarginal optimal transport problems where the number of marginals is equal to the number of classes in the original classification problem. These new theoretical results reveal a rich geometric structure of adversarial learning problems in multiclass classification and extend recent results restricted to the binary classification setting. A direct computational implication of our results is that by solving either the barycenter problem and its dual, or the MOT problem and its dual, we can recover the optimal robust classification rule and the optimal adversarial strategy for the original adversarial problem. Examples with synthetic and real data illustrate our results.

## 1. INTRODUCTION

In this paper we study, from analytical and geometric perspectives, the problem of adversarial learning in multiclass classification. By multiclass classification we mean the task of assigning classes $\hat{i}$ in a set of $K$ available classes to all inputs $\hat{x}$ in some feature space $\mathcal{X}$ based on the observation of training pairs $z = (x, i)$. The adversarial component of the problem refers to the desire of producing classification rules that are *robust* to data perturbations. Mathematically speaking, this means studying optimization problems of the form:

$$\inf_{f \in \mathcal{F}} \sup_{\widetilde{\mu} \in \mathcal{P}(\mathcal{Z})} \{R(f, \widetilde{\mu}) - C(\mu, \widetilde{\mu})\}. \tag{1.1}$$

Here, $\mathcal{F}$ denotes the set of *all* probabilistic multiclass classifiers —see section 2; $\mu$ denotes the observed data distribution, which in general is some probability measure on the space $\mathcal{Z} = \mathcal{X} \times \{1, \dots, K\}$, but which for simplicity can be thought of as an empirical measure associated to a finite training data set; $C$ represents a notion of "distance" between data distributions; $R(f, \widetilde{\mu})$ is a risk functional relative to a data distribution $\widetilde{\mu}$ (thought of as a perturbation of $\mu$) and a choice of loss function, which in this paper will be restricted to be the 0-1 loss. Problem (1.1) can be interpreted as a game between a *learner* and an *adversary*: the learner's goal is to find a classifier with small risk, while the adversary tries to find a data perturbation $\widetilde{\mu}$ that makes the risk for the learner large. The adversary has an implicit budget to perform their actions: the adversary can not choose a $\widetilde{\mu}$ that is too far away (relative to $C$) from the original data distribution $\mu$.

*Date*: April 28, 2022.

*Key words and phrases.* adversarial learning, multiclass classification, optimal transport, multimarginal optimal transport, Wasserstein barycenter, generalized barycenter problem.

**Acknowledgements:** All authors contributed equally and their names are listed in alphabetic order. Part of this work was completed while the authors were visiting the Simons Institute to participate in the program "Geometric Methods in Optimization and Sampling" during the Fall of 2021. The authors would like to thank the institute for hospitality and support. The authors would also like to thank Camilo A. García Trillos, Ryan Murray, and Meyer Scetbon for enlightening conversations on the topics discussed in this work. NGT was supported by NSF-DMS grant 2005797, and together with JK would also like to thank the IFDS at UW-Madison and NSF through TRIPODS grant 2023239 for their support.

For a large family of functionals $C$ in (1.1) we show that the adversarial problem (1.1) is equivalent to a *multimarginal optimal transport problem (MOT)* of the form:

$$\inf_{\pi \in \Pi_K(\mu)} \int \mathbf{c}(z_1, \ldots, z_K) d\pi(z_1, \ldots, z_K), \tag{1.2}$$

where $\mathbf{c}$ is a cost function discussed in detail throughout the paper and $\Pi_K(\mu)$ is a space of couplings specified in section 2.1. As part of this equivalence, we explicitly describe how to construct solutions to the original problem (1.1) from solutions to the problem (1.2) and its dual, offering in this way new computational strategies for solving problem (1.1). Since most algorithms for OT are primal-dual (i.e., they simultaneously search for solutions to both the primal OT problem and its dual), it is actually possible to construct a saddle solution $(f^*, \mu^*)$ for (1.1) by running one such OT algorithm. The equivalence between (1.1) and (1.2) that we study here is an extension to the multi-class case of a series of recent results connecting adversarial learning in binary classification with optimal transport [BCM19, Nak19, PJ21a, PJ21b, GM20].

In order to establish the equivalence between (1.1) and (1.2), we develop another interesting equivalent reformulation of (1.1) that reveals a rich geometric structure of the original adversarial problem. This reformulation takes the form of a *generalized barycenter problem*

$$\inf_{\lambda, \widetilde{\mu}_1, \ldots, \widetilde{\mu}_K} \lambda(\mathcal{X}) + \sum_{i=1}^{K} C(\mu_i, \widetilde{\mu}_i) \quad \text{s.t. } \lambda \geq \widetilde{\mu}_i, i \in \{1, \ldots, K\},$$

which is a novel variant of the Wasserstein barycenter problems introduced in [AC11, CE10]. In the classical Wasserstein barycenter problem, given $K$ probability measures $\varrho_1, \ldots, \varrho_K$ defined over a Polish space $\mathcal{X}$ and a cost $c : \mathcal{X} \times \mathcal{X} \to [0, \infty]$, one tries to find a probability measure $\varrho$ such that the summed cost of transporting each of the $\varrho_i$ onto $\varrho$ is as small as possible. In our generalized problem, we try to find a nonnegative measure $\lambda$ (no longer necessarily a probability measure) such that the total mass of $\lambda$ plus the summed cost of transporting each $\mu_i$ onto *some part* of $\lambda$ is as small as possible. Here transporting a $\mu_i$ onto some part of $\lambda$ means we want to find a probability measure $\widetilde{\mu}_i \leq \lambda$ and transport $\mu_i$ to $\widetilde{\mu}_i$ in the classical optimal transport sense. This problem will be studied in detail in section 3. We prove that these generalized barycenter problems can be written as appropriate MOT problems, a result that is analogous to ones in [AC11, CE10] for standard Wasserstein barycenter problems.

From the equivalence with the generalized barycenter problem we will be able to deduce that optimal adversarial attacks can always be obtained as suitable barycenters of $K$ or less points in the original training data set. Also, from this reformulation we will be able to recognize the structure of the cost function $\mathbf{c}$ in (1.2): for the adversary to obtain their optimal strategy, they can actually *localize* their problem to sets of $K$ or fewer data points —see section 2.1. Other theoretical, methodological, and computational implications of these reformulations will be pursued in future work. See section 6 for a discussion on future directions for research.

In contrast to many of the existing applications of OT to ML, it is worth emphasizing that in this work OT arises naturally in connection with a learning problem, rather than as a particular way to address a certain machine learning task. For the growing literature in multimarginal optimal transportation this paper offers new examples of cost functions worthy of study. MOT is a rich topic that has been developed over the years from theoretical and applied perspectives. After the first mathematical analysis of general MOT problems in [Gwc98], there have been numerous subsequent papers establishing geometric and analytic results ([KP13, Pas15, KP15, CMP17]) for them. MOT problems have been used extensively in applications. For example, they appear in the so-called density functional theory in physics [SGGS07, BDPGG12, CFK13, ML13, CDPDM15], and in economics [Eke05, CMN10, CE10].

In the machine learning community, researchers have recently explored many interesting applications, including generative adversarial networks(GANs) [CCK+18, CMZ+19] and Wasserstein Barycenters [AC11, CD14, BCC+15, COO15, SLD18, DD20], where MOTs are used. Recent works like [DMG20, HRCK21] develop a connection between the Schrödinger bridge problem and MOT. MOT problems have been extended to the unbalanced setting —see [BvLNS21].

1.1. **Outline.** The rest of the paper is organized as follows. In section 2, we introduce most mathematical objects and notation used throughout the rest of the paper. We also introduce a generalized Wasserstein barycenter problem, which can be interpreted as the dual to the original adversarial problem 1.1, and define in detail the MOT problem 1.2. In section 3, we study the aforementioned generalized Wasserstein barycenter problem and prove its equivalence with 1) a stratified barycenter problem and 2) a first version of an MOT problem. In section 4.3 we discuss the equivalence between (1.1) and (1.2). In section 5, we present a collection of examples and numerical experiments, whose goal is to illustrate the theory developed throughout the paper and provide further insights into the geometric structure of adversarial learning in multiclass classification. We wrap up the paper in section 6, where we present some conclusions and discuss some future directions for research.

## 2. Preliminaries

Throughout the paper, $(\mathcal{X}, d)$ will be a Polish space, $[K] := \{1, \dots, K\}$ with $K \geq 2$ and $\mathcal{Z} := \mathcal{X} \times [K]$. We regard $\mathcal{X}$ as the feature space of our model and $[K]$ as the set of classes or labels.

Let $\mu$ be a finite positive measure (not necessarily a probability measure) over $\mathcal{Z}$. We use $\mu_i$ to represent the positive measure over $\mathcal{X}$ defined as

$$\mu_i(A) := \mu\left(A \times \{i\}\right), \tag{2.1}$$

for all measurable subsets $A$ of $\mathcal{X}$. In the sequel, we use $\mu$ to represent a fixed data distribution, which we regard as an observed data distribution or training data distribution, and use $\widetilde{\mu}$ to represent any other arbitrary finite positive measure over $\mathcal{Z}$. Through this paper we use $\mathcal{M}(\mathcal{X})$ and $\mathcal{M}(\mathcal{Z})$ to denote the set of finite positive (Borel) measures over $\mathcal{X}$ and $\mathcal{Z}$, respectively.

We will focus on functionals $C(\mu, \widetilde{\mu})$ of the form:

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

for some cost function $c_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to [0, \infty]$. Here and in the remainder of the paper the set $\Gamma(\cdot, \cdot)$ represents 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 $\mathcal{Z} \times \mathcal{Z}$ with first marginal equal to $\mu$ and second marginal equal to $\widetilde{\mu}$.

**Assumption 2.1.** *The function $c_{\mathcal{Z}}$ will be assumed to have the following structure:*

$$c_{\mathcal{Z}}(z, \tilde{z}) = \begin{cases} c(x, \tilde{x}) & \text{if } i = \tilde{i} \\ \infty & \text{if } i \neq \tilde{i}, \end{cases}$$

*for some lower semi-continuous function $c : \mathcal{X} \times \mathcal{X} \to [0, \infty]$.*

*The function $c$ will be further assumed to satisfy $c(x, x) = 0$ for all $x \in \mathcal{X}$ and the following two compactness and coercivity properties:*

- *if $\{x_n\}_{n \in \mathbb{N}}$ is a bounded sequence in $(\mathcal{X}, d)$ and $\{x'_n\}_{n \in \mathbb{N}}$ is a sequence in $\mathcal{X}$ satisfying $\sup_{n \in \mathbb{N}} c(x'_n, x_n) < \infty$, then $\{(x'_n, x_n)\}_{n \in \mathbb{N}}$ is precompact in $\mathcal{X} \times \mathcal{X}$ (with the induced product metric).*

The structure of $c_{\mathcal{Z}}$ is standard in the literature of adversarial learning and can be motivated by the fact that in many applications of interest it is natural to think that the "true" label associated to a perturbation $\tilde{x}$ of a data point $x$ coincides with the true label of the original $x$. Naturally, this is simply a modeling choice, and other cost structures of interest can be studied elsewhere. The lower semicontinuity and compactness assumptions on $c$ are technical requirements that we use in the remainder. All cost functions of interest satisfy these properties —see the examples below.

If we decompose $\mu$ and $\widetilde{\mu}$ into measures $\mu_i, \widetilde{\mu}_i$ as in (2.1), it is possible to write $C(\mu, \widetilde{\mu})$ as

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

abusing notation slightly and interpreting $C(\mu_i, \widetilde{\mu}_i)$ as

$$C(\mu_i, \widetilde{\mu}_i) = \min_{\pi \in \Gamma(\mu_i, \widetilde{\mu}_i)} \int c(x, \tilde{x}) d\pi(x, \tilde{x}). \tag{2.2}$$

*Remark* 2.2. Let us emphasize that we define $C(\mu_i, \widetilde{\mu}_i) = +\infty$ whenever the set of couplings $\Gamma(\widetilde{\mu}_i, \mu_i)$ is empty, which is the case if $\mu_i$ and $\widetilde{\mu}_i$ have different total mass.

We introduce two notions that will be used throughout our analysis. Given a lower semi-continuous function $f : \mathcal{X} \to \mathbb{R}$ we define

$$f^c(x) := \inf_{x' \in \mathcal{X}} \{ f(x') + c(x', x) \}, \tag{2.3}$$

and given an upper semi-continuous function $g : \mathcal{X} \to \mathbb{R}$ we define

$$g^{\bar{c}}(x') := \sup_{x \in \mathcal{X}} \{ g(x) - c(x', x) \}. \tag{2.4}$$

**Example 2.3.** *Let $\varepsilon > 0$ and let $c(x, \tilde{x})$ be given by*

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

*The parameter $\varepsilon$ can be interpreted as the adversarial budget: the larger the value of $\varepsilon$ the wider the space of actions available to the adversary. The cost $c$ satisfies* ***Assumption*** *2.1 provided that closed balls with finite radius in $(\mathcal{X}, d)$ are compact.*

*Notice that in this case, the c-transform $f^c$ of a given function $f$ takes the form:*

$$f^c(x) = \inf_{x' : d(x, x') \le \varepsilon} f(x').$$

*In this setting, the adversarial problem* (1.1) *can be written as*

$$\inf_{f \in \mathcal{F}} \sup_{\widetilde{\mu} : W_\infty(\mu, \widetilde{\mu}) \le \varepsilon} R(f, \widetilde{\mu}).$$

*where $W_\infty(\mu, \widetilde{\mu})$ is the $\infty$-OT distance between $\mu$ and $\widetilde{\mu}$ relative to the distance function:*

$$\delta(z, \tilde{z}) := \begin{cases} d(x, \tilde{x}) & \text{if } y = \tilde{y}, \\ \infty & \text{otherwise.} \end{cases}$$

*Remark* 2.4. In the literature of machine learning there are many different versions of adversarial problems for supervised tasks, but two versions are particularly popular: data-perturbing adversarial learning [PJ21a] and distributional perturbing adversarial learning [BM19, BKM19]. From our analysis, distributional perturbing adversarial learning is obtained by the choice of cost function $c$ as in **Example** 2.3. For a rigorous analysis, distributional perturbing adversarial learning is more adequate since data-perturbing adversarial learning lacks measuraility in some cases. Furthermore, one can prove that distributional perturbing adversarial learning includes data-perturbing adversarial learning: see [PJ21a].

Our focus in this paper is on the distributional setting, where given a data distribution $\mu$, an adversary can select a new distribution $\widetilde{\mu}$ in a neighborhood of the original distribution $\mu$ determined by $C$. A recent paper [PJ21b] summarizes other adversarial models and discusses connections between them.

**Example 2.5.** *Let $p > 0$ and let $c(x, \tilde{x})$ be given by*

$$c(x, x') = c^p(x, x') := \frac{1}{\tau}(d(x, x'))^p,$$

*for some constant $\tau > 0$. For this choice of cost $c$, it is possible to show, through a formal argument whose details we omit, that problem* (1.1) *can be written as*

$$\inf_{f \in \mathcal{F}} \sup_{\widetilde{\mu} \,:\, W_p(\mu, \widetilde{\mu}) \leq \varepsilon} R(f, \widetilde{\mu}),$$

*for some $\varepsilon > 0$ and for $W_p(\mu, \widetilde{\mu})$ the p-OT distance between $\mu$ and $\widetilde{\mu}$ relative to the distance function $\delta$ from **Example** 2.3. The relation between $\tau$ and $\varepsilon$ is not explicit, but, qualitatively, small values of $\tau$ should correspond to small values of $\varepsilon$.*

*Notice that in this case the c-transform $f^c$ of a given function $f$ takes the form:*

$$f^c(x) = \inf_{x' \in \mathcal{X}} f(x') + \frac{1}{\tau} d(x, x')^p.$$

*If $f$ is bounded below by a constant it follows that $f^c$ is always continuous (in the $d$ metric) regardless of the continuity properties of the original $f$.*

The solution space $\mathcal{F}$ in (1.1) is the full set of weak partitions, or probabilistic classifiers, defined by

$$\mathcal{F} := \Big\{ f : \mathcal{X} \to \Delta_{[K]} : f \text{ measurable } \Big\},$$

where

$$\Delta_{[K]} := \Big\{ (u_i)_{i \in [K]} : 0 \leq u_i \leq 1, \sum_{i \in [K]} u_i = 1 \Big\},$$

i.e., the set of probability distributions over $[K]$. In other words, at each $x \in \mathcal{X}$, $f(x)$ is a probability distribution over $[K]$ representing the likelihood, according to the learner, that a given $x$ belongs to any of the available classes. Probabilistic classifiers are widely used in applications as they allow for the use of standard optimization techniques, like gradient descent, when training models.

For a given $u \in \Delta_{[K]}$ and a given $i \in [K]$, we define the 0-1 loss:

$$\ell(u, i) := 1 - u_i.$$

Notice that $\ell(e_j, i)$ is equal to 1 if $i \neq j$ and 0 if $i = j$: this motivates the name 0-1 loss for $\ell$. For a given pair $(f, \widetilde{\mu})$ we define the risk:

$$R(f, \widetilde{\mu}) := \mathbb{E}_{(\widetilde{X}, \widetilde{Y}) \sim \widetilde{\mu}}[\ell(f(\widetilde{X}), \widetilde{Y})] = \sum_{i \in [K]} \int_{\mathcal{X}} (1 - f_i(\widetilde{x})) d\widetilde{\mu}_i(\widetilde{x}),$$

which can be regarded as a bilinear functional $R(\cdot, \cdot) : \mathcal{F} \times \mathcal{P}(\mathcal{Z}) \longrightarrow \mathbb{R}_+$. For convenience, we introduce the so-called *classification power* for a pair $(f, \widetilde{\mu}) \in \mathcal{F} \times \mathcal{P}(\mathcal{Z})$, which is defined by

$$B(f, \widetilde{\mu}) := \sum_{i \in [K]} \int_{\mathcal{X}} f_i(\widetilde{x}) d\widetilde{\mu}_i(\widetilde{x}). \tag{2.5}$$

With these new definitions, problem (1.1) is immediately seen to be equivalent to

$$\sup_{f\in\mathcal{F}} \inf_{\widetilde{\mu}\in\mathcal{P}(\mathcal{Z})} \left\{ B(f,\widetilde{\mu}) + C(\mu,\widetilde{\mu}) \right\}. \tag{2.6}$$

Moreover, if we denote by $\widetilde{B}^*_\mu$ the optimal value of (2.6), and by $R^*_\mu$ the optimal value of (1.1), we have the identity:

$$R^*_\mu = \mu(\mathcal{Z}) - \widetilde{B}^*_\mu.$$

Again, we write $\mu(\mathcal{Z})$ explicitly, although for the most part we will consider the case in which $\mu(\mathcal{Z})$ is equal to one.

The dual of (2.6) is obtained by swapping the sup and the inf:

$$\inf_{\widetilde{\mu}\in\mathcal{P}(\mathcal{Z})} \sup_{f\in\mathcal{F}} \left\{ B(f,\widetilde{\mu}) + C(\mu,\widetilde{\mu}) \right\}. \tag{2.7}$$

Notice that the value of (2.7) is always greater than or equal to the value of (2.6). Instead of attempting to invoke an abstract minimax theorem at this stage, implying the equality of these two quantities, we defer this discussion to later sections where in fact we will prove that, under **Assumption** 2.1, there is no duality gap in this problem. In what follows we focus on the dual problem (2.7) and only return to problem (2.6), which is equivalent to the original adversarial problem (1.1), in section 4.3. Notice, however, that the statement of **Theorem** 2.8 mentions the adversarial problem explicitly.

For fixed $\widetilde{\mu}$, notice that

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \left\{ B(f,\widetilde{\mu}) + C(\mu,\widetilde{\mu}) \right\} &= \sup_{f\in\mathcal{F}} \left\{ \sum_{i\in[K]} \int_{\mathcal{X}} f_i(\widetilde{x}) d\widetilde{\mu}_i(\widetilde{x}) + C(\mu,\widetilde{\mu}) \right\} \\
&= \sup_{f\in\mathcal{F}} \left\{ \sum_{i\in[K]} \int_{\mathcal{X}} f_i(\widetilde{x}) d\widetilde{\mu}_i(\widetilde{x}) \right\} + C(\mu,\widetilde{\mu}).
\end{aligned}$$

Introducing a new variable $\lambda$, which is a positive measure over $\mathcal{X}$, we can rewrite the latter sup as:

$$\inf_{\lambda} \lambda(\mathcal{X}) \quad \text{s.t.} \int_X g(x) d(\lambda - \widetilde{\mu}_i)(x) \geq 0 \text{ for all } g \geq 0, i \in \{1,\dots,K\};$$

the constraint in $\lambda$ can be simply written as $\widetilde{\mu}_i \leq \lambda$ for all $i = 1,\dots,K$. Combining the above with the structure of the cost $C(\mu,\widetilde{\mu})$, we conclude that problem (2.7) is equivalent to the generalized barycenter problem mentioned in the introduction:

$$B^*_\mu := \inf_{\lambda,\widetilde{\mu}_1,\dots,\widetilde{\mu}_K} \lambda(\mathcal{X}) + \sum_{i=1}^K C(\mu_i,\widetilde{\mu}_i) \quad \text{s.t.} \int_X g(x) d(\lambda - \widetilde{\mu}_i)(x) \geq 0 \text{ for all } g \geq 0, i \in \{1,\dots,K\}, \tag{2.8}$$

where we use the notation $B^*_\mu$ for future reference.

*Remark* 2.6. It is straightforward to see from (2.7) that $B^*_\mu$ is positive homogeneous in $\mu$. That is, if $a > 0$, then $B^*_{a\mu} = aB^*_\mu$.

2.1. **The MOT problem.** In order to state problem (1.2) precisely, we will need to modify the set $\mathcal{Z}$ and in particular add an extra element to it that we will denote with the symbol $\mathghost$. The marginals of the couplings in our desired MOT problem will be probability measures over the set $\mathcal{Z}_* := \mathcal{Z} \cup \{\mathghost\}$. More precisely, we consider the set:

$$\Pi_K(\mu) := \left\{ \pi \in \mathcal{P}(\mathcal{Z}_*^K) \;:\; P_{i\sharp}\pi = \frac{1}{2\mu(\mathcal{Z})}\mu(\cdot \cap \mathcal{Z}) + \frac{1}{2}\delta_{\mathghost}, \quad \forall i = 1,\dots,K \right\}. \tag{2.9}$$

Notice that in this set all $K$ marginals are the same. We define the set $\Pi_K(\mu)$ in this way so as to be consistent with the literature on multimarginal optimal transport, where sets of couplings are typically assumed to be probability measures.

Let us now discuss the cost function for the desired MOT problem and the role played by the additional element $\mathghost$. For a given tuple $(z_1, \dots, z_K)$ in $\mathcal{Z}_*^K$, often denoted by $\vec{z}$ in the sequel for convenience, we define

$$\mathbf{c}(z_1, \dots, z_K) := B^*_{\widehat{\mu}_{\vec{z}}}, \tag{2.10}$$

where $\widehat{\mu}_{\vec{z}}$ is the positive measure (not necessarily a probability measure) defined as:

$$\widehat{\mu}_{\vec{z}} := \frac{1}{K} \sum_{l \text{ s.t. } z_l \neq \mathghost}^{K} \delta_{z_l}.$$

Recall that $B^*_{\widehat{\mu}_{\vec{z}}}$ is equal to (2.8) (alternatively, equal to (2.7)) when $\mu$ is equal to $\widehat{\mu}_{\vec{z}}$.

*Remark* 2.7. Notice that $\widehat{\mu}_{\vec{z}}$ is a probability measure if and only if no element in the tuple $\vec{z}$ is $\mathghost$.

Following the literature of MOT, we can write the dual of our MOT problem as:

$$\sup_{\phi \in \Phi} \Big\{ \sum_{j=1}^{K} \int_{\mathcal{X} \times [K]} \phi_j(z_j) \frac{1}{2\mu(\mathcal{Z})} d\mu(z_j) + \frac{1}{2} \sum_{j=1}^{K} \phi_j(\mathghost) \Big\}, \tag{2.11}$$

where

$$\Phi := \Big\{ \phi = (\phi_1, \dots, \phi_K) \in \prod_{j=1}^{K} L^1\big(\frac{1}{2\mu(\mathcal{Z})}\mu + \frac{1}{2}\delta_{\mathghost}\big) : \sum_{j=1}^{K} \phi_j(z_j) \leq B^*_{\widehat{\mu}_{\vec{z}}}, \quad \forall \vec{z} \in \mathcal{Z}_*^K \Big\}. \tag{2.12}$$

We will later show that under **Assumption** 2.1 there is no duality gap between the MOT problem and its dual (2.11) —see **Corollary** 4.5.

One of the main results of the paper is the following.

**Theorem 2.8.** *Suppose that **Assumption** 2.1 holds. Let $\mu$ be a finite positive measure over $\mathcal{Z}$. Then* (2.7) *is equivalent to the MOT problem* (1.2) *with set of couplings $\Pi_K(\mu)$ defined as in* (2.9)*, and cost function $\mathbf{c}$ defined as in* (2.10)*. Specifically,*

$$\frac{1}{2\mu(\mathcal{Z})} B^*_\mu = \min_{\pi \in \Pi_K(\mu)} \int \mathbf{c}(z_1, \dots, z_K) d\pi(z_1, \dots, z_K).$$

*Furthermore,* (2.6) $=$ (2.7)*. In addition, from a solution pair $(\pi^*, \phi^*)$ for the MOT problem and its dual one can obtain a solution pair $(f^*, \widetilde{\mu}^*)$ for* (2.7) *and its dual, i.e. problem* (2.6)*. The pair $(f^*, \widetilde{\mu}^*)$ is also a saddle for the original adversarial problem* (1.1)*.*

The proof of **Theorem** 2.8 is presented throughout section 4; the expression for $(f^*, \widetilde{\mu}^*)$ in terms of $(\phi^*, \pi^*)$ is presented in **Corollary** 4.7. Given the definition of the cost function $\mathbf{c}$, **Theorem** 2.8 states that the adversarial problem *localizes* to data sets consisting of $K$ or less equally weighted points. More precisely, the problem for the adversary reduces to first determining their actions when facing arbitrary distributions supported on $K$ or fewer data points, and then finding an optimal grouping for the data in order to assemble their global strategy. The element $\mathghost$ indicates when fewer than $K$ points are being grouped by the adversary. From the solution to this problem one can directly obtain an optimal classification rule for the original adversarial problem. Note that problem (1.2) is a problem solved by the adversary: ideally, the adversary wants to group together points $(z_1, \dots, z_K)$ for which there is a low classification power $B^*_{\widehat{\mu}_{\vec{z}}}$ (or alternatively large robust risk). On the other hand, the dual of (1.2) can be interpreted as a maximization problem solved by the learner.

In order to prove **Theorem** 2.8, we will first obtain a series of equivalent reformulations of problem (2.8) which will reveal a rich geometric structure of the adversarial problem and will facilitate the connection with the desired MOT problem. These equivalent formulations are of interest in their own right.

## 3. The generalized barycenter problem

We begin this Section by proving that the generalized barycenter problem always has at least one solution. In the following subsections we will then discuss a series of equivalent problems to the generalized barycenter problem, their duals, and some geometric properties of their solutions.

**Proposition 3.1.** *Suppose that $c$ is a lower semicontinuous cost satisfying the property that for any compact set $E \subset \mathcal{X}$ there exists a compact set $F \subset \mathcal{X}$ such that for all $x \in E, x' \in F, x'' \in \mathcal{X} \setminus F$ we have $c(x,x') \leq c(x,x'')$. Given measures $\mu_1, \dots, \mu_K$ and a there exists at least one solution to problem* (2.8).

*Remark* 3.2. If $c$ is a cost that satisfies **Assumption** 2.1, then $c$ satisfies the hypothesis of **Proposition** 3.1.

*Remark* 3.3. Nearly identical arguments can be used to prove that the various reformulations of (2.8) that we will consider throughout this section have minimizers. For this reason, in what follows, we will simply assume the existence of minimizers without explicitly proving their existence.

*Proof.* Using transportation plans to compute the cost $C(\mu_i, \tilde{\mu}_i)$ in (2.8), we can rewrite the problem in the following form,

$$\inf_{\lambda \in \mathcal{M}(\mathcal{X}), \pi_1, \dots, \pi_K \in \mathcal{M}(\mathcal{X} \times \mathcal{X})} \lambda(\mathcal{X}) + \sum_{i=1}^{K} \int_{\mathcal{X} \times \mathcal{X}} c(x,x') d\pi_i(x,x') \quad \text{s.t. } \pi_i(\mathcal{X} \times E) \leq \lambda(E), \;\; \pi_i(E \times \mathcal{X}) = \mu_i(E),$$

where $E$ ranges over all Borel sets in $\mathcal{X}$. Note that a feasible solution to this problem exists since we may choose $\lambda, \pi_1, \dots, \pi_K$ such that $\lambda := \sum_{i=1}^K \mu_i$ and for all $f \in C_c(\mathcal{X} \times \mathcal{X})$ $\int_{\mathcal{X} \times \mathcal{X}} f(x,x') d\pi_i(x,x') := \int_{\mathcal{X}} f(x,x) d\mu_i(x)$. Note that with these choices, the problem attains the value $\sum_{i=1}^K \mu_i(\mathcal{X})$.

Let $\lambda^n, \pi_1^n, \dots, \pi_K^n$ be a sequence of feasible solutions such that

$$\lim_{n \to \infty} \lambda^n(\mathcal{X}) + \sum_{i=1}^{K} \int_{\mathcal{X} \times \mathcal{X}} c(x,x') d\pi_i^n(x,x') = t := \inf_{\lambda, \pi_1, \dots, \pi_K} \lambda(\mathcal{X}) + \sum_{i=1}^{K} \int_{\mathcal{X} \times \mathcal{X}} c(x,x') d\pi_i(x,x')$$

From our work above and the nonnegativity of the transport cost, $\lambda^n(\mathcal{X})$ is uniformly bounded by $\sum_{i=1}^K \mu_i(\mathcal{X})$. Furthermore, we may assume that for any Borel set $E$

$$\sum_{i=1}^{K} \int_{\mathcal{X} \times E} d\pi_i^n(x,x') \geq \lambda^n(E)$$

otherwise we could delete mass from $\lambda^n$ and attain a smaller value. Given some $\epsilon > 0$, let $E_\epsilon \subset \mathcal{X}$ be a compact set such that $\sum_{i=1}^K \mu_i(\mathcal{X} \setminus E_\epsilon) \leq \epsilon$. Let $F_\epsilon$ be a compact set such that for all $x \in E_\epsilon, x' \in F_\epsilon$ and $x'' \in \mathcal{X} \setminus F_\epsilon$ we have $c(x,x') \leq c(x,x'')$. If $\lambda^n$ gives more than $\epsilon$ to $\mathcal{X} \setminus F_\epsilon$ then some of this mass must be transported to $E_\epsilon$. Since the transportation cost would be cheaper if the excess mass was placed inside of $F_\epsilon$ instead of $\mathcal{X} \setminus F_\epsilon$, it follows that $\lambda^n(\mathcal{X} \setminus F_\epsilon) \leq \epsilon$. Therefore, the $\lambda^n$ are a tight family.

The tightness of $\lambda^n$ and $\mu_1, \dots, \mu_K$ implies that $\pi_1^n, \dots, \pi_K^n$ are a tight family. Therefore, we can extract a subsequence that converges weakly to a limit $\lambda^*, \pi_1^*, \dots, \pi_K^*$. From the lower semicontinuity of the cost, it follows that $\lambda^*, \pi_1^*, \dots, \pi_K^*$ is a minimizer. □

3.1. **A first MOT reformulation of** (3.1) **and geometric consequences.** In the rest of what follows, we shall let $S_K$ denote the power set of $\{1,\dots,K\}$ and for any $i \in \{1,\dots,K\}$ we let $S_K(i) = \{A \in S_K : i \in A\}$. Given a feasible solution $\lambda, \widetilde{\mu}_1, \dots, \widetilde{\mu}_K$ to (2.8) and a point $x \in \mathcal{X}$, it is interesting to consider the set

$$A(x) = \{i \in \{1,\dots,K\} : x \in \operatorname{spt}(\widetilde{\mu}_i)\}.$$

If $x \notin \operatorname{spt}(\lambda)$, then $A(x)$ must be empty. Conversely, if $x \in \operatorname{spt}(\lambda)$, then we may assume that $A(x)$ is not empty, otherwise we could delete $x$ from the support of $\lambda$ and improve the solution. Hence, to each point $x \in \operatorname{spt}(\lambda)$ we can associate a nonempty subset of $\{1,\dots,K\}$ that describes which of the $\mu_i$ send mass to the point $x$. Given a set $A \subseteq \{1,\dots,K\}$ we define $D(A) := \{x \in \operatorname{spt}(\lambda) : A(x) = A\}$. This allows us to give a natural partition of $\lambda$ into the sum $\lambda = \sum_{A \in S_K} \lambda_A$ where we define

$$\lambda_A(E) := \lambda(E \cap D(A)),$$

for every measurable $E \subseteq \mathcal{X}$.

Along the same lines, we can devise a similar decomposition for each of the $\mu_i$. Given an optimal plan $\pi_i \in \Gamma(\widetilde{\mu}_i, \mu_i)$ (optimal for (2.2)) and $A \in S_K(i)$ we define

$$\mu_{i,A}(E) := \int_{D(A)\times E} d\pi_i(x', x)$$

for every measurable $E \subseteq \mathcal{X}$. Again, this allows us to partition $\mu_i = \sum_{A \in S_K(i)} \mu_{i,A}$ since the sets $D(A)$ are disjoint and their union recovers the support of $\lambda$. Using the decompositions, we can eliminate the $\widetilde{\mu}_i$ from the problem and reformulate the optimization in terms of $\lambda_A$ and the $\mu_{i,A}$. Doing so, we get the equivalent problem

$$\min_{\lambda,\, \mu_{i,A}} \sum_{A \in S_K} \left\{ \lambda_A(\mathcal{X}) + \sum_{i \in A} C(\lambda_A, \mu_{i,A}) \right\} \quad \text{s.t.} \sum_{A \in S_K(i)} \mu_{i,A} = \mu_i \ \text{ for all } i \in \{1,\dots,K\}. \tag{3.1}$$

To keep notation from getting too complicated, we shall assume that $\mu_{i,A}$ is defined for all $i \in \{1,\dots,K\}$ and $A \subseteq S_K$, however, note that if $i \notin A$, then $\mu_{i,A}$ plays no role in the optimization problem.

Suppose that for some $A \in S_K$ we fix a choice of $\mu_{i,A}$ for all $i \in A$. With the $\mu_{i,A}$ fixed, we can determine the corresponding optimal $\lambda_A^* = \lambda_A^*(\mu_{1,A}, \dots, \mu_{K,A})$ by solving the classic Wasserstein barycenter problem. Indeed, the optimal choice must be an element of

$$\operatorname*{argmin}_{\lambda_A} \sum_{i \in A} C(\lambda_A, \mu_{i,A}). \tag{3.2}$$

Note that here we do not need to consider the mass of $\lambda_A$, since the value of the optimization problem will be $+\infty$ if $\lambda_A$ does not have the same mass as all of the $\mu_{i,A}$ (or if the $\mu_{i,A}$ themselves do not all have the same mass).

It is well known that problem (3.2) can be reformulated as a multimarginal optimal transport problem [AC11]. To that end, given $A \subseteq \{1,\dots,K\}$, define $c_A : \mathcal{X}^K \to \mathbb{R}$

$$c_A(x_1, \dots, x_K) := \inf_{x' \in \mathcal{X}} \sum_{i \in A} c(x', x_i), \tag{3.3}$$

and $T_A : \mathcal{X}^K \to \mathcal{X}$

$$T_A(x_1, \dots, x_K) := \operatorname*{argmin}_{x' \in \mathcal{X}} \sum_{i \in A} c(x', x_i). \tag{3.4}$$

*Remark* 3.4. If $\operatorname{argmin}_{x'\in\mathcal{X}}\sum_{i\in A} c(x',x_i)$ is not unique, we can consider using an additional selection procedure. For example, when $\mathcal{X}=\mathbb{R}^d$ we can still recover a unique mapping by choosing $T_A$ to be the element of $\operatorname{argmin}_{x'\in\mathcal{X}}\sum_{i\in A} c(x',x_i)$ that is closest (in the Euclidean distance) to the Euclidean barycenter $\frac{1}{|A|}\sum_{i\in A} x_i$.

With the definition of $c_A$, we can rewrite (3.2) as the multimarginal optimal transport problem

$$\min_{\pi_A}\int_{\mathcal{X}^K} c_A(x_1,\dots,x_K)d\pi_A(x_1,\dots,x_K) \quad \text{s.t. } \mathcal{P}_{i\#}\pi_A=\mu_{i,A} \text{ for all } i\in A, \tag{3.5}$$

where $\mathcal{P}_i$ is the projection map $(x_1,\dots,x_K)\mapsto x_i$. Again, even though $\pi_A$ is defined over $\mathcal{X}^K$, only the coordinates $i$ where $i\in A$ play a role in the optimization problem. Indeed, $c_A$ is independent of the other coordinates and we only have marginal constraints for $i\in A$.

Using (3.5) we can now eliminate $\lambda_A$ and all of the $\mu_{i,A}$'s from problem (3.1) and reformulate the optimization as the multimarginal problem

$$\min_{\pi_{\{1\}},\dots,\pi_{\{1,\dots,K\}}}\sum_{A\in S_K}\int_{\mathcal{X}^K}\big(c_A(x_1,\dots,x_K)+1\big)d\pi_A(x_1,\dots,x_K) \quad \text{s.t.} \sum_{A\in S_K(i)}\mathcal{P}_{i\#}\pi_A=\mu_i \text{ for all } i\in\{1,\dots,K\}. \tag{3.6}$$

The next two propositions formally prove the equivalence between (3.1) and (3.6). They will also allow us to establish some important geometric properties of optimal generalized barycenters.

**Proposition 3.5.** *Let $c$ be a cost satisfying **Assumption** 2.1. Given measures $\mu_1,\dots,\mu_K$, let $\{\pi_A\}_{A\in S_K}$ be a feasible solution to (3.6). For each $(x_1,\dots,x_K)\in\mathcal{X}^K$ and $A\in S_K$, let $f_A(x_1,\dots,x_K)$ be a choice of element in $T_A(x_1,\dots x_K)$, where we recall the definition of $T_A(x_1,\dots,x_K)$ from (3.4).*

*If for each $A\in S_K$ and $i\in A$ we set $\tilde\lambda_A=f_{A\#}\pi_A$ and $\tilde\mu_{i,A}=\mathcal{P}_{i\#}\pi_A$, then $\{\tilde\lambda_A,\tilde\mu_{i,A}\}_{A\in S_K,i\in A}$ is a feasible solution to (3.1) and*

$$\sum_{A\in S_K}\tilde\lambda_A(\mathcal{X})+\sum_{i\in A}C(\tilde\lambda_A,\tilde\mu_{i,A})\le\sum_{A\in S_K}\int_{\mathcal{X}^K}\big(c_A(x_1,\dots,x_K)+1\big)d\pi_A(x_1,\dots,x_K).$$

*Proof.* Since $\sum_{A\in S_K(i)}\mathcal{P}_{i\#}\pi_A=\mu_i$, it is automatic that $\sum_{A\in S_K(i)}\tilde\mu_{i,A}=\mu_i$. Since pushforwards do not affect the total mass of a measure, so we also have $\tilde\mu_{i,A}(\mathcal{X})=\tilde\lambda_A(\mathcal{X})$ for all $A\in S_K$ and $i\in A$. Hence, $\{\tilde\lambda_A,\tilde\mu_{i,A}\}_{A\in S_K,i\in A}$ is a feasible solution to (3.1).

For each $A\in S_K$ and $i\in A$, choose $\varphi_{i,A},\psi_{i,A}\in C_b(\mathcal{X})$ that satisfy, for all $x,x'\in\mathcal{X}$,

$$\varphi_{i,A}(x)-\psi_{i,A}(x')\le c(x,x').$$

We can then compute

$$\begin{aligned}
&\int_{\mathcal{X}}\varphi_{i,A}(x_i)d\tilde\mu_{i,A}(x_i)-\int_{\mathcal{X}}\psi_{i,A}(x')d\tilde\lambda_A(x')\\
&=\int_{\mathcal{X}}\varphi_{i,A}(x_i)d\tilde\mu_{i,A}(x_i)-\int_{\mathcal{X}^K}\psi_{i,A}(f_A(x_1,\dots,x_K))d\pi_A(x_1,\dots,x_K)\\
&\le\int_{\mathcal{X}}\varphi_{i,A}(x_i)d\tilde\mu_{i,A}(x_i)+\int_{\mathcal{X}^K}\Big(c\big(x_i,f_A(x_1,\dots,x_K)\big)-\varphi_{i,A}(x_i)\Big)d\pi_A(x_1,\dots,x_K)\\
&=\int_{\mathcal{X}^K}c\big(x_i,f_A(x_1,\dots,x_K)\big)d\pi_A(x_1,\dots,x_K).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sum_{i\in A}\int_{\mathcal{X}}\varphi_{i,A}(x_i)d\tilde{\mu}_{i,A}(x_i)-\int_{\mathcal{X}}\psi_{i,A}(x')d\tilde{\lambda}_A(x')\\
&\le\int_{\mathcal{X}^K}\sum_{i\in A}c\big(x_i,f_A(x_1,\dots,x_K)\big)d\pi_A(x_1,\dots,x_K)\\
&=\int_{\mathcal{X}^K}c_A(x_1,\dots,x_K)d\pi_A(x_1,\dots,x_K),
\end{aligned}$$

where we have used the definition of $f_A, T_A$, and $c_A$ to obtain the last equality. Hence,

$$\begin{aligned}
&\sum_{A\in S_K}\tilde{\lambda}_A(\mathcal{X})+\sum_{i\in A}\int_{\mathcal{X}}\varphi_{i,A}(x_i)d\tilde{\mu}_{i,A}(x_i)-\int_{\mathcal{X}}\psi_{i,A}(x')d\tilde{\lambda}_A(x')\\
&\le\sum_{A\in S_K}\int_{\mathcal{X}^K}\big(c_A(x_1,\dots,x_K)+1\big)d\pi_A(x_1,\dots,x_K).
\end{aligned}$$

Taking the supremum over all admissible choices of $\varphi_{i,A},\psi_{i,A}$ and exploiting the dual formulation of optimal transport, we obtain the desired result:

$$\sum_{A\in S_K}\tilde{\lambda}_A(\mathcal{X})+\sum_{i\in A}C(\tilde{\lambda}_A,\tilde{\mu}_{i,A})\le\sum_{A\in S_K}\int_{\mathcal{X}^K}\big(c_A(x_1,\dots,x_K)+1\big)d\pi_A(x_1,\dots,x_K).$$

□

In the next proposition we will show that any feasible solution of problem (3.1) induces a feasible solution of (3.6) with lesser or equal value. This will prove the equivalence between problems (3.1) and (3.6) and will provide a powerful geometric characterization of optimal generalized barycenters.

**Proposition 3.6.** *Let $c$ be a cost satisfying **Assumption** 2.1. Given measures $\mu_1,\dots,\mu_K$, let $\mu_{i,A},\lambda_A$ be feasible solutions to problem* (3.1)*. Let $\gamma_{i,A}\in\mathcal{M}(\mathcal{X}\times\mathcal{X})$ be an optimal plan for the transport of $\mu_{i,A}$ to $\lambda_A$ with respect to the cost $c$. Let $\gamma_A\in\mathcal{M}(\mathcal{X}^{K+1})$ such that for all $i\in A$ and $g\in C_b(\mathcal{X}\times\mathcal{X})$*

$$\int_{\mathcal{X}^{K+1}}g(x_i,x')d\gamma_A(x_1,\dots,x_K,x')=\int_{\mathcal{X}^{K+1}}g(x_i,x')d\gamma_{i,A}(x_i,x').$$

*If we define $\tilde{\pi}_A$ on $\mathcal{X}^K$ such that for any $h\in C_b(\mathcal{X}^K)$ we have*

$$\int_{\mathcal{X}^K}h(x_1,\dots,x_K)d\pi_A(x_1,\dots,x_K)=\int_{\mathcal{X}^{K+1}}h(x_1,\dots,x_K)d\gamma_A(x_1,\dots,x_K,x'),$$

*then $\tilde{\pi}_A$ is a feasible solution to* (3.6) *and*

$$\sum_{A\in S_K}\int_{\mathcal{X}^K}\big(c_A(x_1,\dots,x_K)+1\big)d\tilde{\pi}_A(x_1,\dots,x_K)\le\sum_{A\in S_K}\lambda_A(\mathcal{X})+\sum_{i\in A}C(\lambda_A,\mu_{i,A}).$$

*Therefore,* (3.1) $=$ (3.6).

*Proof.* We begin by noting that the marginal constraints on $\gamma_A$ are compatible in the sense that for any $u\in C_b(\mathcal{X})$ and $i\in A$ we have

$$\int_{\mathcal{X}}u(x')d\gamma_{i,A}(x_i,x')=\int_{\mathcal{X}}u(x')d\lambda_A(x').$$

Thus, each $\gamma_A$ is well-defined.

Using the definition of $d\tilde{\pi}_A$ and then $c_A$, it follows that

$$\begin{aligned}
&\sum_{A\in S_K}\int_{\mathcal{X}^K}\big(c_A(x_1,\ldots,x_K)+1\big)d\tilde{\pi}_A(x_1,\ldots,x_K)\\
&=\sum_{A\in S_K}\int_{\mathcal{X}^{K+1}}\big(c_A(x_1,\ldots,x_K)+1\big)d\gamma_A(x_1,\ldots,x_K,x')\\
&\le\sum_{A\in S_K}\int_{\mathcal{X}^{K+1}}\big(1+\sum_{i\in A}c(x_i,x')\big)d\gamma_A(x_1,\ldots,x_K,x')\\
&=\sum_{A\in S_K}\int_{\mathcal{X}^{K+1}}\big(1+\sum_{i\in A}c(x_i,x')\big)d\gamma_{i,A}(x_i,x')\\
&=\sum_{A\in S_K}\lambda_A(\mathcal{X})+C(\mu_{i,A},\lambda_A)
\end{aligned}$$

where the final equality follows from the fact that $\gamma_{i,A}$ is an optimal plan for the transport of $\mu_{i,A}$ to $\lambda_A$. □

In addition to proving the equivalence between problems (3.1) and (3.6), **Proposition** 3.5 and **Proposition** 3.6 have the following very important geometric consequences.

**Corollary 3.7.** *Let $c$ be a cost satisfying* ***Assumption*** *2.1. Given measures $\mu_1,\ldots,\mu_K$, let $\lambda$ be an optimal generalized barycenter and let $\{\lambda_A\}_{A\in S_K}$ be a decomposition of $\lambda$ and $\{\mu_{i,A}\}_{A\in S_K(i)}$ a decomposition of each $\mu_i$ that are optimal for* (3.1)*. Recalling* (3.4)*, let $T_A(x_1,\ldots,x_K):=\operatorname{argmin}_{x\in\mathcal{X}}\sum_{i\in A}c(x,x_i)$. If we define $T_A:=\{T_A(x_1,\ldots,x_K):x_1\in\operatorname{spt}(\mu_1),\ldots,x_K\in\operatorname{spt}(\mu_K)\}$ and $T=\cup_{A\subseteq\{1,\ldots,K\}}T_A$, then $\lambda_A(\mathcal{X})=\lambda_A(T_A)$, $\lambda(\mathcal{X})=\lambda(T)$ and the optimal measures $\widetilde{\mu}_i$ in* (2.8) *can be assumed to satisfy $\widetilde{\mu}_i(\mathcal{X})=\widetilde{\mu}_i(T)$ as well.*

*In particular, if $f_A(x_1,\ldots,x_K)$ is a choice of element from $T_A(x_1,\ldots,x_K)$ for each $A\in S_K$ and $(x_1,\ldots,x_K)\in\mathcal{X}^K$, then there exists an optimal barycenter $\lambda_f$ such that $\lambda_f(\mathcal{X})=\lambda_f(F)$ where $F=\bigcup_{A\in S_K}\bigcup_{(x_1,\ldots x_K)\in\operatorname{spt}(\mu_1)\times\operatorname{spt}(\mu_K)}f_A(x_1,\ldots,x_K)$.*

*Remark* 3.8. In the case where we have a tuple $(x_1,\ldots,x_K)\in\operatorname{spt}(\mu_1)\times\cdots\times\operatorname{spt}(\mu_K)$ such that $\sum_{i\in A}c(x,x_i)=+\infty$ for all $x\in\mathcal{X}$, we set $T_A(x_1,\ldots,x_K)=\varnothing$.

*Proof.* From **Proposition** 3.6, we can use $\{\lambda_A\}_{A\in S_K}$ and $\{\mu_{i,A}\}_{A\in S_K,i\in A}$ to construct measures $\{\tilde{\pi}_A\}_{A\in S_K}$ with

$$\sum_{A\in S_K}\int_{\mathcal{X}^K}\big(c_A(x_1,\ldots,x_K)+1\big)d\tilde{\pi}_A(x_1,\ldots,x_K)\le\sum_{A\in S_K}\lambda_A(\mathcal{X})+\sum_{i\in A}C(\lambda_A,\mu_{i,A}).\tag{3.7}$$

From **Proposition** 3.5, we can then use $\tilde{\pi}_A$ to construct decompositions $\{\tilde{\lambda}_A\}_{A\in S_K}$ and $\{\tilde{\mu}_{i,A}\}_{A\in S_K,i\in A}$ such that

$$\sum_{A\in S_K}\tilde{\lambda}_A(\mathcal{X})+\sum_{i\in A}C(\tilde{\lambda}_A,\tilde{\mu}_{i,A})\le\sum_{A\in S_K}\int_{\mathcal{X}^K}\big(c_A(x_1,\ldots,x_K)+1\big)d\tilde{\pi}_A(x_1,\ldots,x_K).\tag{3.8}$$

Examining the proof of **Proposition** 3.6, it follows that the inequality in (3.7) is strict if $\lambda_A(\mathcal{X})>\lambda_A(T_A)$. In that case, combining (3.7) and (3.8) would contradict the optimality of $\lambda$. Therefore, $\lambda_A(T_A)=\lambda_A(\mathcal{X})$. The final statements follow from the constraints satisfied by the $\tilde{\mu}_i$ and the construction in **Proposition** 3.5. □

When $\mu_1,\ldots,\mu_K$ are supported on a finite set of points, **Corollary** 3.7 has the following consequence.

**Corollary 3.9.** *If $\mu_1, \dots, \mu_K$ are measures that are supported on a finite set of points and $c$ is a cost satisfying **Assumption** 2.1, then there exists a solution $\lambda$ to the optimal generalized barycenter problem* (2.8) *that is supported on a finite set of points.*

*In particular, if each $\mu_i$ is supported on a set of $n_i$ points, then there exists an optimal barycenter that is supported on at most $\sum_{A\in S_K}\prod_{i\in A} n_i \leq 2^K \prod_{i=1}^K n_i$ points.*

*Remark* 3.10. Notice that the bound mentioned at the end of **Corollary** 3.9 is a worst case bound. In practice, especially when data sets have a favourable geometric structure, the optimal barycenter $\lambda$ may have a much sparser support. See section 5.2.

*Proof.* For each $i \in \{1, \dots, K\}$ we can assume there exists a finite set $X_i \subset \mathcal{X}$ such that $\mu_i$ is supported on $X_i$. For each $A \in S_K$, let $f_A : X_i^K \to \mathcal{X}$ be a function such that

$$f_A(x_1, \dots, x_K) \in T_A(x_1, \dots, x_K)$$

for all $(x_1, \dots, x_K) \in X_i^K$, where we recall the definition of $T_A$ from (3.4). We can now construct the set

$$F = \bigcup_{A\in S_K} \bigcup_{(x_1,\dots,x_K)\in\prod_{i=1}^K X_i} \{f_A(x_1, \dots, x_K)\},$$

which is necessarily finite. Indeed, if we set $n_i = |X_i|$, then $F$ has at most $\sum_{A\in S_K}\prod_{i\in A} n_i$ elements. By **Corollary** 3.7, there exists an optimal barycenter supported on $F$ only.

□

3.2. **A second MOT reformulation of (3.1).** Note that in problem (3.6) we need to find a distribution $\pi_A$ for each $A \in S_K$. Hence, it is natural to wonder if we can reformulate problem (3.6) in such a way that we only need to find a single distribution $\gamma$. Here one must be careful, as the previous formulations of the problem do not require the input distributions $\mu_1, \dots, \mu_K$ to have the same mass. As a result, if we try to work over a space of probability distributions whose marginals are $\mu_1, \dots, \mu_K$, then we cannot recover the full generality of (3.6).

To overcome this difficulty, we will define $\gamma$ over the slightly larger space $(\mathcal{X} \times [0, 1])^K$. The extra coordinate will help us to track the mass associated to each label $i$. Define $\widetilde{c} : (\mathcal{X}\times[0, 1])^K \to \mathbb{R}$ by

$$\widetilde{c}((x_1, r_1), \dots, (x_K, r_K)) := \inf_{m:S_K\to\mathbb{R}} \sum_{A\in S_K} m_A\big(c_A(x_1, \dots, x_K) + 1\big) \quad \text{s.t.} \sum_{A\in S_K(i)} m_A = r_i. \tag{3.9}$$

For each $i \in \{1, \dots, K\}$, let $\tilde{\mathcal{P}}_i$ be the projection $((x_1, r_1), \dots, (x_K, r_K)) \mapsto x_i$. In what follows, we use $(\vec{x}, \vec{r})$ to denote the tuple $((x_1, r_1), \dots, (x_K, r_K))$. We then claim that problem (3.6) is equivalent to

$$\min_{\gamma} \int_{(\mathcal{X}\times[0,1])^K} \widetilde{c}(\vec{x}, \vec{r})d\gamma(\vec{x}, \vec{r}) \quad \text{where } \tilde{\mathcal{P}}_{i\#}(r_i\gamma) = \mu_i \text{ for all } i \in \{1, \dots, K\}. \tag{3.10}$$

**Proposition 3.11.** *Problems* (3.6) *and* (3.10) *are equivalent, and thus* (3.10) *is also equivalent to* (2.7)*,* (2.8) *and* (3.1)*.*

*Proof.* Given a feasible solution $\pi_{\{1\}}, \dots, \pi_{\{1,\dots,K\}}$ to problem (3.6), define $\gamma$ such that for every continuous and bounded function $f : (\mathcal{X} \times [0, 1])^K \to \mathbb{R}$ we have

$$\int_{(\mathcal{X}\times[0,1])^K} f(\vec{x}, \vec{r})d\gamma(\vec{x}, \vec{r}) = \sum_{A\in S_K} \int_{\mathcal{X}^K} f\big((x_1, \chi_A(1)), \dots, (x_K, \chi_A(K))\big)d\pi_A(x_1, \dots, x_K).$$

where $\chi_A(i) = 1$ if $i \in A$ and zero otherwise. We can then check that $\gamma$ is feasible for (3.10), since for any function $g : \mathcal{X} \to \mathbb{R}$

$$\int_{(\mathcal{X}\times[0,1])^K} r_i g(x_i) d\gamma\big(\vec{x}, \vec{r}\big) = \sum_{A \in S_K(i)} \int_{\mathcal{X}^K} g(x_i) d\pi_A(x_1, \dots, x_K)$$
$$= \int_{\mathcal{X}} g(x_i) d\mu_i(x_i),$$

where the final equality uses the fact that $\sum_{A\in S_K(i)} \mathcal{P}_{i\,\#}\pi_A = \mu_i$.

Next, we observe that for any $A \in S_K$ and a tuple of the form $\big((x_1, \chi_A(1)), \dots, (x_K, \chi_A(K))\big)$ we have

$$\widetilde{c}\big((x_1, \chi_A(1)), \dots, (x_K, \chi_A(K))\big) \leq c_A(x_1, \dots, x_K) + 1.$$

Therefore,

$$\int_{(\mathcal{X}\times[0,1])^K} \widetilde{c}(\vec{x}, \vec{r}) d\gamma(\vec{x}, \vec{r}) \leq \sum_{A\in S_K} \int_{\mathcal{X}^K} \big(c_A(x_1, \dots, x_K) + 1\big) d\pi_A(x_1, \dots, x_K).$$

Conversely, suppose that $\gamma$ is a feasible solution to (3.10). Given a tuple $(\vec{x}, \vec{r})$, let

$$m_A(\vec{x}, \vec{r}) \in \operatorname*{argmin}_{m:S_K\to\mathbb{R}} \sum_{A\in S_K} m_A\big(c_A(x_1, \dots, x_K) + 1\big) \quad \text{s.t.} \sum_{A\in S_K(i)} m_A = r_i.$$

Given $A \in S_K$ define $\pi_A$ such that for any continuous and bounded function $h : \mathcal{X}^K \to \mathbb{R}$ we have

$$\int_{\mathcal{X}^K} h(x_1, \dots, x_K) d\pi_A(x_1, \dots, x_K) = \int_{(\mathcal{X}\times[0,1])^K} h(x_1, \dots, x_K) m_A(\vec{x}, \vec{r}) d\gamma(\vec{x}, \vec{r}).$$

We can then check that for any continuous and bounded function $g : \mathcal{X} \to \mathbb{R}$

$$\sum_{A\in S_K(i)} \int_{\mathcal{X}^K} g(x_i) d\pi_A(x_1, \dots, x_K) = \int_{(\mathcal{X}\times[0,1])^K} r_i g(x_i) d\gamma(\vec{x}, \vec{r})$$
$$= \int_{\mathcal{X}} g(x_i) \mu_i(x_i),$$

where we have used the fact that $\sum_{A\in S_K(i)} m_A(\vec{x}, \vec{r}) = r_i$ in the first equality. Thus, our construction gives us a feasible solution to (3.6). Evaluating the objective in (3.6) we see that

$$\sum_{A\in S_K} \int_{\mathcal{X}^K} (c_A(x_1, \dots, x_K) + 1) d\pi_A(x_1, \dots, x_K)$$
$$= \int_{(\mathcal{X}\times[0,1])^K} \sum_{A\in S_K} m_A(\vec{x}, \vec{r})(c_A(x_1, \dots, x_K) + 1) d\gamma(\vec{x}, \vec{r})$$
$$= \int_{(\mathcal{X}\times[0,1])^K} \widetilde{c}(\vec{x}, \vec{r}) d\gamma(\vec{x}, \vec{r})$$

where the final equality uses the definition of $\widetilde{c}$ and our choice of $m_A(\vec{x}, \vec{r})$. Thus, the two problems have the same optimal value and any feasible solution to one problem can be easily converted into a feasible solution to the other. □

3.3. **Localization.** In this section we show that the cost function $\widetilde{c}$ in problem (3.10) is equal to $B^*_{\widehat{\mu}}$ for a measure $\widehat{\mu}$ that depends on the arguments of $\widetilde{c}$. This result can be interpreted as a localization property for problem (2.8) (and hence for problem (2.7) as well). Compare with the discussion after **Theorem** 2.8.

**Proposition 3.12.** *Let $\tilde{x}_1, \dots, \tilde{x}_k \in \mathcal{X}$, and let $0 \le \tilde{r}_1, \dots, \tilde{r}_k \le 1$. Then $\widetilde{c}((\tilde{x}_1, \tilde{r}_1), \dots, (\tilde{x}_K, \tilde{r}_K))$ defined in* (3.9) *is equal to $B^*_{\widehat{\mu}}$, where*

$$\widehat{\mu} := \sum_{i=1}^{K} \tilde{r}_i \delta_{(\tilde{x}_i, i)}.$$

*Proof.* To prove this claim we first notice that by **Proposition** 3.11 $B^*_{\widehat{\mu}}$ is equal to

$$\min_{\gamma} \int_{(\mathcal{X}\times[0,1])^K} \widetilde{c}(\vec{x}, \vec{r}) d\gamma(\vec{x}, \vec{r}),$$

where $\gamma$ is in the constraint set of problem (3.10). For a feasible $\gamma$, notice that $\gamma$ must concentrate on the set $\{(\vec{x}, \vec{r}) \,:\, x_i = \tilde{x}_i, \quad i = 1, \dots, K\}$. Applying the disintegration theorem to $\gamma$, we can rewrite the objective function evaluated at $\gamma$ as

$$\int_{[0,1]^K} \widetilde{c}((\tilde{x}_1, r_1), \dots, (\tilde{x}_K, r_K)) d\gamma_r(r_1, \dots, r_K),$$

where $\gamma_r$ is a positive measure over $[0,1]^K$ satisfying the constraints:

$$\int_{[0,1]} r_i d\gamma_r(r_1, \dots, r_K) = \tilde{r}_i, \quad \forall i = 1, \dots, K. \tag{3.11}$$

It is clear that the map associating a feasible $\gamma$ to a $\gamma_r$ satisfying (3.11) is onto, and thus, we can rewrite $B^*_{\widehat{\mu}}$ as

(3.12)

$$\begin{aligned} B^*_{\widehat{\mu}} &= \min_{\gamma_r} \int_{[0,1]^K} \widetilde{c}((\tilde{x}_1, r_1), \dots, (\tilde{x}_K, r_K)) d\gamma_r(r_1, \dots, r_K) \\ &= \min_{\gamma_r} \int_{[0,1]^K} \min_{\{m_A\}_{A \in G(r_1, \dots, r_K)}} \left\{ \sum_{A \in S_K} m_A (1 + c_A(\tilde{x}_1, \dots, \tilde{x}_K)) \right\} d\gamma_r(r_1, \dots, r_K) \\ &= \min_{\gamma_r} \min_{\{m_A\}_{A\in G}} \int_{[0,1]^K} \left\{ \sum_{A \in S_K} m_A(r_1, \dots, r_K) \cdot (1 + c_A(\tilde{x}_1, \dots, \tilde{x}_K)) \right\} d\gamma_r(r_1, \dots, r_K) \\ &= \min_{\{m_A\}_{A\in G}} \min_{\gamma_r} \int_{[0,1]^K} \left\{ \sum_{A \in S_K} m_A(r_1, \dots, r_K) \cdot (1 + c_A(\tilde{x}_1, \dots, \tilde{x}_K)) \right\} d\gamma_r(r_1, \dots, r_K). \end{aligned}$$

In the above, the set $G(r_1, \dots, r_K)$ is the set of $\{m_A\}_{A \in S_K}$ satisfying the constraints in (3.9) for the specific tuple $\big((\tilde{x}_1, r_1), \dots, (\tilde{x}_K, r_K)\big)$, while $G$ is the set of $\{m_A\}_A$ where each $m_A$ is a functions with inputs $r_1, \dots, r_K$ satisfying $\{m_A(r_1, \dots, r_K)\}_A \in G(r_1, \dots, r_K)$.

We can now write the term

$$\begin{aligned} &\int_{[0,1]^K} \left\{ \sum_{A \in S_K} m_A(r_1, \dots, r_k) \cdot (1 + c_A(\tilde{x}_1, \dots, \tilde{x}_k)) \right\} d\gamma_r(r_1, \dots, r_K) \\ &= \sum_{A \in S_K} m_{A,\gamma}(1 + c_A(\tilde{x}_1, \dots, \tilde{x}_k)), \end{aligned}$$

where we define

$$m_{A,\gamma_r} := \int m_A(r_1, \dots, r_k) d\gamma_r(r_1, \dots, r_K).$$

Notice that

$$\begin{aligned}\sum_{A \in S_K(i)} m_{A,\gamma_r} &= \sum_{A \in S_K(i)} \int_{[0,1]^K} m_A(r_1, \dots, r_k) d\gamma_r(r_1, \dots, r_K) \\ &= \int_{[0,1]^K} \left( \sum_{A \in S_K(i)} m_A(r_1, \dots, r_k) \right) d\gamma_r(r_1, \dots, r_K) \\ &= \int_{[0,1]^K} r_i d\gamma_r(r_1, \dots, r_K) \\ &= \tilde{r}_i.\end{aligned}$$

Conversely, notice that given a collection of functions $\tilde{m}_A$ satisfying the constraint in (3.9) for the tuple $(\tilde{x}_1, \tilde{r}_1), \dots, (\tilde{x}_K, \tilde{r}_K)$, it is straightforward to find $\gamma_r$ such that $\tilde{m}_A = m_{A,\gamma_r}$ for all $A$. It now follows that

$$B^*_{\widehat{\mu}} = \min_{\tilde{m}_A} \sum_A \tilde{m}_A(1 + c_A(\tilde{x}_1, \dots, \tilde{x}_k)) = \widetilde{c}((\tilde{x}_1, \tilde{r}_1), \dots, (\tilde{x}_K, \tilde{r}_K)),$$

as we wanted to prove. □

3.4. **Dual Problems.** In this section we discuss the dual problems of the different formulations of the generalized barycenter problem studied in section 3.1.

**Proposition 3.13.** *The dual problems to* (2.8), (3.6), *and* (3.10) *can be written as*

(3.13)

$$\sup_{f_1, \dots, f_K \in C_b(\mathcal{X})} \sum_{i=1}^K \int_{\mathcal{X}} f_i^c(x_i) d\mu_i(x_i) \quad \text{s.t. } f_i(x) \geq 0, \ \sum_{i=1}^K f_i(x) \leq 1, \text{ for all } x \in \mathcal{X}, i \in \{1, \dots, K\},$$

(3.14)

$$\sup_{g_1, \dots, g_K \in C_b(\mathcal{X})} \sum_{i=1}^K \int_{\mathcal{X}} g_i(x_i) d\mu_i(x_i) \quad \text{s.t. } \sum_{i \in A} g_i(x_i) \leq 1 + c_A(x_1, \dots, x_K) \text{ for all } (x_1, \dots, x_K) \in \mathcal{X}^K, A \in S_K,$$

*and*

(3.15)

$$\sup_{h_1, \dots, h_K \in C_b(\mathcal{X})} \sum_{i=1}^K \int_{\mathcal{X}} h_i(x_i) d\mu_i(x_i) \quad \text{s.t. } \sum_{i=1}^K r_i h_i(x_i) \leq \widetilde{c}(\vec{x}, \vec{r}) \text{ for all } (\vec{x}, \vec{r}) \in (\mathcal{X} \times [0,1])^K,$$

*respectively.*

*Let* $f_1, \dots, f_K$; $g_1, \dots, g_K$; $h_1, \dots, h_K$ *be feasible solutions to problems* (3.13), (3.14), *and* (3.15) *respectively. Problems* (3.14) *and* (3.15) *have the same feasible set and hence are identical. Furthermore,* $g_i' := f_i^c$ *is a feasible solution to* (3.14) *and* $f_i' = \max(g_i, 0)^{\bar{c}}$ *is a feasible solution to* (3.13), *hence the optimization of* (3.14) *can be restricted to nonnegative* $g_i$ *that satisfy* $g_i = g_i^{\bar{c}c}$. *In particular,* (3.13), (3.14), *and* (3.15) *all have the same optimal value.*

*Proof.* The derivation of the dual problems is standard.

To see the equivalence between problems (3.14) and (3.15), fix some $h_1, \dots, h_K$ that are feasible for (3.15) and choose some $B \in S_K$ and $(x_1, \dots, x_K) \in \mathcal{X}^K$ such that $c_B(x_1, \dots, x_K) < \infty$. Choose

$$m^* \in \underset{m: S_K \to \mathbb{R}}{\operatorname{argmin}} \sum_{A \in S_K} m_A(1 + c_A(x_1, \dots, x_K)) \quad \text{s.t.} \sum_{A \in S_K(i)} m_A = \chi_B(i),$$

where $\chi_B(i) = 1$ if $i \in B$ and zero otherwise. Note that the choice $m_A = 1$ if $A = B$ and $m_A = 0$ otherwise is feasible for the above optimization. Therefore, the optimality of $m^*$ implies that

$$\begin{aligned} 1 + c_B(x_1, \dots, x_K) &\geq \sum_{A \in S_K} m^*_A(1 + c_A(x_1, \dots, x_K)) \\ &= \widetilde{c}((x_1, \chi_B(1)), \dots, (x_k, \chi_B(k))) \\ &\geq \sum_{i=1}^{K} r_i h_i(x_i) \\ &= \sum_{i \in B} h_i(x_i). \end{aligned}$$

Thus, we see that the $h_i$ are feasible for (3.14) since $B$ and $(x_1, \dots, x_K)$ were arbitrary.

Conversely, fix some $g_1, \dots, g_K$ that are feasible for (3.14) and some $(\vec{x}, \vec{r}) \in (\mathcal{X} \times [0,1])^K$. Choose

$$n^* \in \operatorname*{argmin}_{m: S_K \to \mathbb{R}} \sum_{A \in S_K} m_A(1 + c_A(x_1, \dots, x_K)) \quad \text{s.t.} \sum_{A \in S_K(i)} m_A = r_i,$$

and observe that

$$\begin{aligned} \sum_{i=1}^{K} r_i g_i(x_i) &= \sum_{i=1}^{K} g_i(x_i) \sum_{A \in S_K(i)} n^*_A \\ &= \sum_{A \in S_K} n^*_A \sum_{i \in A} g_i(x_i) \\ &\leq \sum_{A \in S_K} n^*_A(1 + c_A(x_1, \dots, x_K)) \\ &= \widetilde{c}((x_1, r_1), \dots, (x_K, r_K)), \end{aligned}$$

where we used the feasibility of the $g_i$. Thus, the $g_i$ are feasible for (3.15). Since both problems are optimizing the same functional over the same constraint set, we see that (3.14) and (3.15) are identical.

Now suppose that $f_1, \dots, f_K$ and $g_1, \dots, g_K$ are feasible solutions to problems (3.13) and (3.14) respectively and define $g'_i = f_i^c$ and $f'_i = \max(g_i, 0)^{\bar{c}}$. Given $A \in S_K$, $x_1, \dots, x_K \in \mathcal{X}^K$, and $r > 0$ we can choose $x_r$ such that

$$\sum_{i \in A} c(x_r, x_i) \leq r + c_A(x_1, \dots, x_K).$$

Then we see that

$$\sum_{i \in A} g'_i(x_i) \leq \sum_{i \in A} f(x_r) + c(x_r, x_i) \leq r + 1 + c_A(x_1, \dots, x_K).$$

Letting $r \to 0$, we see that the $g'_i$ are feasible for (3.14). Hence, the optimal value of (3.14) cannot lie strictly below the optimal value of (3.13).

It remains to verify the feasibility of the $f'_i$. We begin by showing that if $g_1, \dots, g_K$ are feasible for (3.14) then $\max(g_1, 0), \dots, \max(g_K, 0)$ are also feasible. Fix $A \in S_K$ and $(x_1, \dots, x_K) \in \mathcal{X}^K$. Let $A' = \{i \in A : g_i(x_i) > 0\}$. We then see that

$$\sum_{i \in A} \max(g_i(x_i), 0) = \sum_{i \in A'} g_i(x_i) \leq 1 + c_{A'}(x_1, \dots, x_K) \leq 1 + c_A(x_1, \dots, x_K)$$

where the final inequality follows from the definition of $c_A$ and the fact that $A' \subseteq A$. Now we are ready to verify the feasibility of the $f_i'$. Clearly $f_i'(x) \geq 0$ since $c(x,x) = 0$ for all $x \in \mathcal{X}$. Given $x \in \mathcal{X}$, fix $r > 0$ and for each $i \in \{1, \dots, K\}$, choose $x_{i,r} \in X$ such that

$$(\max(g_i, 0))^{\bar{c}}(x) \leq \max(g_i(x_{i,r}), 0) - c(x_{i,r}, x) + r.$$

We then have

$$\begin{aligned}\sum_{i=1}^K \max(g_i, 0)^{\bar{c}}(x) &\leq \sum_{i=1}^K \max(g_i(x_{i,r}), 0) - c(x_{i,r}, x) + r \\ &\leq 1 + r + c_{\{1,\dots,K\}}(x_{1,r}, \dots, x_{k,r}) - \sum_{i=1}^K c(x_{i,r}, x),\end{aligned}$$

where the final inequality follows from the feasibility of $\max(g_i, 0)$. Now from the definition of $c_{\{1,\dots,K\}}$, the last line is bounded above by $1 + r$. Sending $r \to 0$ we are done.

Notice that the above arguments prove that whenever $g_1, \dots, g_K$ are feasible for (3.14), then $\max(g_1, 0)^{\bar{c}c}, \dots, \max(g_K, 0)^{\bar{c}c}$ are also feasible for (3.14). Since $u \leq u^{\bar{c}c}$ for any function $u : \mathcal{X} \to \mathbb{R}$, it follows that

$$\sum_{i=1}^K \int_{\mathcal{X}} g_i(x) d\mu_i(x) \leq \sum_{i=1}^K \int_{\mathcal{X}} \max(g_i, 0)^{\bar{c}c} d\mu_i = \sum_{i=1}^K \int_{\mathcal{X}} \Big( \max(g_i, 0)^{\bar{c}} \Big)^c d\mu_i$$

Since we showed that $\max(g_i, 0)^{\bar{c}}$ was feasible for (3.13), it follows that (3.14) cannot attain a larger value than (3.13). Hence, we have shown that (3.14) and (3.13) have the same optimal value. □

We now want to show that the dual problems attain the same values as the original primal problems. We begin with a minimax lemma for the following partial optimal transport problem.

**Lemma 3.14.** *Suppose that $c$ is a bounded Lipschitz cost that satisfies the hypotheses of* ***Proposition*** *3.1. If $\mathcal{B} \subset \mathcal{M}(\mathcal{X})$ is a weakly compact and convex set, then given measures $\mu_1, \dots, \mu_K, \in \mathcal{M}(\mathcal{X})$, let we have the following minimax formula*

$$\min_{\rho, \nu_i \in \mathcal{B}, \nu_i \leq \rho} \sum_{i=1}^K C(\mu_i, \nu_i) = \max_{\varphi_i, \psi_i \in C_b(\mathcal{X})} \min_{\rho \in \mathcal{B}} \sum_{i=1}^K \int_{\mathcal{X}} \varphi_i(x) d\nu_i(x) - \psi_i(x') d\rho(x')$$
$$\text{s.t. } \varphi_i(x) - \psi_i(x') \leq c(x, x'), \quad \psi_i(x') \geq 0.$$

*Proof.* Using the dual formulation of optimal transport, we can write

$$C(\mu_i, \nu_i) = \sup_{\varphi_i, \psi_i \in \Phi_c} J_i(\nu_i, \varphi_i, \psi_i) \quad \text{s.t. } \varphi_i(x) - \psi_i(x') \leq c(x, x').$$

where

$$J_i(\nu_i, \varphi_i, \psi_i) = \int_{\mathcal{X}} \varphi_i(x) d\mu_i(x) - \psi_i(x) d\nu_i(x),$$

and $\Phi_c = \{(\varphi_i, \psi_i) \in C_b(\mathcal{X}) \times C_b(\mathcal{X}) : \varphi_i(x) - \psi_i(x') \leq c(x, x') \text{ for all } x, x' \in \mathcal{X}\}$. For each $\varphi_i, \psi_i \in C_b(\mathcal{X})$ fixed, the mapping $(\rho, \nu_i) \mapsto J_i(\nu_i, \varphi_i, \psi_i)$ is linear and lower semicontinuous with respect to the weak convergence of measures. For any $\rho, \nu_i$ fixed, the mapping $(\varphi_i, \psi_i) \mapsto J_i(\nu_i, \varphi_i, \psi_i)$ is linear and upper semicontinuous with respect to strong convergence in $C_b(\mathcal{X})$. Since the constraint sets $\nu_i \leq \rho$ and $\Phi_c$ are convex, we are in a situation where Sion's minimax theorem applies. Therefore,

$$\min_{\rho, \nu_i \in \mathcal{B}, \nu_i \leq \rho} \sup_{\varphi_i, \psi_i \in \Phi_c} \sum_{i=1}^K J_i(\nu_i, \varphi_i, \psi_i) = \sup_{\varphi_i, \psi_i \in \Phi_c} \min_{\rho, \nu_i \in \mathcal{B}, \nu_i \leq \rho} \sum_{i=1}^K J_i(\nu_i, \varphi_i, \psi_i)$$

Since,

$$\min_{\nu_i\leq\rho}\sum_{i=1}^{K} J_i(\nu_i,\varphi_i,\psi_i) = \sum_{i=1}^{K}\int_{\mathcal{X}} \varphi_i(x)d\mu_i(x) - \max(\psi_i(x'),0)d\rho(x')$$

We have

$$\min_{\rho,\nu_i\in\mathcal{B},\nu_i\leq\rho}\sum_{i=1}^{K} C(\mu_i,\nu_i) = \sup_{\varphi_i,\psi_i\in\Phi_c}\min_{\rho\in\mathcal{B}}\sum_{i=1}^{K}\int_{\mathcal{X}} \varphi_i(x)d\mu_i(x) - \max(\psi_i(x'),0)d\rho(x')$$

If we replace $\varphi_i$ by $\psi_i^c$ and $\psi_i$ by $\max(\psi_i,0)^{c\bar{c}}$ then the value of the problem can only improve. Since we assume that $c$ is bounded and Lipschitz, it follows that $\psi_i^c$ and $\psi_i^{\bar{c}c}$ are bounded and Lipschitz. Thus, we can restrict the supremum to a compact subset of $\Phi_c$ where $\psi_i \geq 0$. Thus, the supremum is actually attained by some pair $(\varphi_i^*,\psi_i^*)\in\Phi_c$ with $\psi_i^*\geq 0$, $\varphi_i^* = (\psi_i^*)^c$ and $(\psi_i^*)^{c\bar{c}} = \psi_i^*$.

□

Using **Lemma** 3.14 we can prove that there is no duality gap for bounded and Lipschitz costs. We will then show that there is no duality gap for general costs by approximation.

**Proposition 3.15.** *Given measures $\mu_1,\dots,\mu_K$ and a bounded Lipschitz cost $c$ satisfying the assumptions in **Proposition** 3.1, suppose that $\lambda,\tilde{\mu}_1,\dots,\tilde{\mu}_K$ are optimal solutions to* (2.8)*. If $\varphi_i^*,\psi_i^*\in C_b(\mathcal{X})$ are the optimal Kantorovich potentials for the partial transport of $\mu_i$ to $\lambda$ (c.f **Lemma** 3.14), then $\varphi_1^*,\dots,\varphi_K^*$ are optimal solutions to problem* (3.14)*, $\psi_1^*,\dots,\psi_K^*$ are optimal solutions to* (3.13)*, and the values of* (3.13)*-*(3.15) *are equal to* (2.8)*. In other words, there is no duality gap.*

*Proof.* If we fix some convex weakly compact subset $\mathcal{B}\subset\mathcal{M}(\mathcal{X})$ containing $\lambda$, then it follows from **Lemma** 3.14 and the optimality of $\lambda$ that there exists $\varphi_i^*,\psi_i^*$ such that

$$\lambda(\mathcal{X}) + \sum_{i=1}^{K} C(\mu_i,\tilde{\mu}_i) = \min_{\rho\in\mathcal{B}}\rho(\mathcal{X}) + \sum_{i=1}^{K}\int_{\mathcal{X}}\varphi_i^*(x)d\mu_i(x) - \psi_i^*(x')d\rho(x'), \tag{3.16}$$

$\psi_i^*(x')\geq 0$, and $(\varphi_i^*)^{\bar{c}}(x') = \psi_i^*(x'), (\psi_i^*)^c(x) = \varphi_i^*(x)$ for all $1\leq i\leq K$ and $x,x'\in\mathcal{X}$. If there exists $x'\in\mathcal{X}$ such that $\sum_{i=1}^K\psi_i^*(x') > 1$, then we can make the right hand side of (3.16) smaller than the left hand side by choosing $\rho = M\delta_{x'}$ for some sufficiently large value of $M$. Hence, it follows that $\sum_{i=1}^K\psi_i^*(x)\leq 1$ everywhere. Thus, the $\psi_i^*$ are feasible solutions to problem (3.13) and, by **Proposition** 3.13 $(\psi_i^*)^c = \varphi_i^*$ are feasible solutions to (3.14). Finally, if we choose $\rho = 0$, it follows that

$$(2.8) = \lambda(\mathcal{X}) + \sum_{i=1}^{K} C(\mu_i,\tilde{\mu}_i) \leq \sum_{i=1}^{K}\varphi_i^*(x)d\mu_i(x) \leq (3.14) = (3.13) \leq (2.8)$$

where the second last equality follows from **Proposition** 3.13 and the last inequality holds trivially by duality. Therefore, we can infer that there is no duality gap.

□

**Proposition 3.16.** *Given measures $\mu_1,\dots,\mu_K$, if $c$ is a cost that satisfies **Assumption** 2.1, then problems* (3.13)*-*(3.15) *all have the same value as* (2.8)*.*

*Remark* 3.17. Note that we do not claim that the supremums in (3.13)-(3.15) are attained.

*Proof.* Let $\eta:[0,\infty)\to[0,\infty)$ be a smooth strictly increasing function such that $\eta(x) = x$ for $x\leq 1$ and $\eta(x)\leq 2$ for all $x\in[0,\infty)$. For each $j\in\mathbb{Z}_+$, define

$$\tilde{c}_j(x,x') := \inf_{(x_1,x_1')\in\mathcal{X}\times\mathcal{X}} c(x_1,x_1') + jd(x,x_1) + jd(x',x_1),$$

and $c_j(x,x') := j\eta(\frac{\tilde{c}_j(x,x')}{j})$. It then follows that $c_j$ is a bounded Lipschitz cost that satisfies the assumptions of **Proposition** 3.1. Since $c$ is lower semicontinuous it is straightforward to check that $c_j$ converges to $c$ pointwise everywhere.

Let $\alpha_j$ and $\beta_j$ denote the optimal values of Problems (2.8) and (3.14) respectively with cost $c_j$. From **Proposition** 3.15 we know that $\alpha_j = \beta_j$. Let $\alpha, \beta$ denote the optimal values of Problems (2.8) and (3.14) respectively with the original cost $c$. Since we already know that $\beta \leq \alpha$, our goal is to show that $\alpha \leq \beta$.

Exploiting the fact that $c_j$ is increasing with respect to $j$, if $g_1^{j_0}, \dots, g_K^{j_0}$ is a feasible solution to (3.14) for the cost $c_{j_0}$, then it is also a feasible solution to (3.14) for $c$. Therefore, $\lim_{j\to\infty} \beta_j \leq \beta$.

On the other hand, let $\lambda^j$ and $\tilde{\mu}_1^j, \dots, \tilde{\mu}_K^j$ be optimal solutions to (2.8) with the cost $c_j$. Let $\pi_i^j$ be the optimal transport plan between $\mu_i$ and $\tilde{\mu}_i^j$. Arguing as in **Proposition** 3.1, it follows that $\lambda^j$ and $\pi_i^j$ are tight with respect to $j$. Thus, there exists a subsequence (that we do not relabel) such that $\lambda^j$ converges weakly to some $\lambda$ and $\pi_i^j$ converges weakly to some $\pi_i$. Fix some $j_0$ and note that for all $j \geq j_0$

$$\alpha_j = \lambda^j(\mathcal{X}) + \sum_{i=1}^K \int_{\mathcal{X}} c_j(x,x') d\pi_i^j(x,x') \geq \lambda^j(\mathcal{X}) + \sum_{i=1}^K \int_{\mathcal{X}} c_{j_0}(x,x') d\pi_i^j(x,x').$$

Therefore,

$$\liminf_{j\to\infty} \alpha_j \geq \lambda(\mathcal{X}) + \sum_{i=1}^K \int_{\mathcal{X}} c_{j_0}(x,x') d\pi_i(x,x').$$

Taking a supremum over $j_0$, it follows that

$$\liminf_{j\to\infty} \alpha_j \geq \lambda(\mathcal{X}) + \sum_{i=1}^K \int_{\mathcal{X}} c(x,x') d\pi_i(x,x') \geq \alpha.$$

Thus, $\alpha \leq \liminf_{j\to\infty} \alpha_j = \liminf_{j\to\infty} \beta_j = \beta$. Thanks to **Proposition** 3.13, it follows that (2.8) and (3.13)-(3.15), all have the same optimal value. □

## 4. Proof of Theorem 2.8

### 4.1. **Theorem 2.8: upper bound.** First we show that

$$\frac{1}{2\mu(\mathcal{Z})} B_\mu^* \leq \min_{\pi \in \Pi_K(\mu)} \int_{\mathcal{Z}_*^K} \mathbf{c}(z_1, \dots, z_K) d\pi(z_1, \dots, z_K).$$

To see this, recall that $B_\mu^*$ is, according to **Proposition** 3.11, equal to

$$\min_{\gamma \in \Upsilon_\mu} \int_{(\mathcal{X}\times[0,1])^K} \widetilde{c}(\vec{x}, \vec{r}) d\gamma(\vec{x}, \vec{r}) \quad \text{where } \tilde{\mathcal{P}}_{i\,\#}(r_i\gamma) = \mu_i \text{ for all } i \in \{1, \dots, K\}.$$

Here and in what follows we use $\Upsilon_\mu$ to denote the set of positive measures satisfying $\tilde{\mathcal{P}}_{i\,\#}(r_i\gamma) = \mu_i$ for all $i \in \{1, \dots, K\}$.

Let $\pi \in \Pi_K(\mu)$, and for given $\vec{z} = (z_1, \dots, z_K) \in \mathcal{Z}_*^K$, let $\gamma_{\vec{z}} \in \Upsilon_{\widehat{\mu}_{\vec{z}}}$ be a solution for problem (3.10) (when $\mu = \widehat{\mu}_{\vec{z}}$). We define a measure $\gamma$ as follows:

$$\int_{(\mathcal{X}\times[0,1])^K} h(\vec{x}, \vec{r}) d\gamma(\vec{x}, \vec{r}) := \int_{\mathcal{Z}_*^K} \left( \int_{(\mathcal{X}\times[0,1])^K} h(\vec{x}, \vec{r}) d\gamma_{\vec{z}}(\vec{x}, \vec{r}) \right) d\pi(z_1, \dots, z_K)$$

for every test function $h : (\mathcal{X} \times [0,1])^K \to \mathbb{R}$.

We check that $\gamma \in \Upsilon_{\frac{1}{2\mu(\mathcal{Z})}\mu}$. Indeed, for any test function $g : \mathcal{X} \to \mathbb{R}$ we have:

$$\begin{aligned}
\int_{(\mathcal{X}\times[0,1])^K} r_i g(x_i) d\gamma(\vec{x},\vec{r}) &= \int_{\mathcal{Z}_*^K} \left( \int_{(\mathcal{X}\times[0,1])^K} r_i g(x_i) d\gamma_{\vec{z}}(\vec{x},\vec{r}) \right) d\pi(z_1,\dots,z_K) \\
&= \frac{1}{K} \int_{\mathcal{Z}_*^K} \left( \sum_{j : z_j \neq \mathghost} g(x_j) \mathbb{1}_{i_j = i} \right) d\pi(z_1,\dots,z_K) \\
&= \frac{1}{2\mu(\mathcal{Z})} \int_{\mathcal{X}} g(x) d\mu_i(x).
\end{aligned}$$

Let us now compute the cost associated to this $\gamma$:

$$\begin{aligned}
\int_{(\mathcal{X}\times[0,1])^K} \widetilde{c}(\vec{x},\vec{r}) d\gamma(\vec{x},\vec{r}) &= \int_{\mathcal{Z}_*^K} \left( \int_{(\mathcal{X}\times[0,1])^K} \widetilde{c}(\vec{x},\vec{r}) d\gamma_{\vec{z}}(\vec{x},\vec{r}) \right) d\pi(z_1,\dots,z_K) \\
&= \int_{\mathcal{Z}_*^K} B^*_{\widehat{\mu}_{\vec{z}}} d\pi(z_1,\dots,z_K) \\
&= \int_{\mathcal{Z}_*^K} \mathbf{c}(z_1,\dots,z_K) d\pi(z_1,\dots,z_K).
\end{aligned}$$

Combining the above with *Remark* 2.6, we conclude that

$$\frac{1}{2\mu(\mathcal{Z})} B^*_\mu = B^*_{\frac{1}{2\mu(\mathcal{Z})}\mu} = \min_{\gamma \in \Upsilon_{\frac{1}{2\mu(\mathcal{Z})}\mu}} \int_{(\mathcal{X}\times[0,1])^K} \widetilde{c}(\vec{x},\vec{r}) d\gamma(\vec{x},\vec{r}) \leq \min_{\pi \in \Pi_K(\mu)} \int_{\mathcal{Z}_*^K} \mathbf{c}(\vec{z}) d\pi(\vec{z}).$$

4.2. **Theorem 2.8: lower bound.** We now show that

$$\min_{\pi \in \Pi_K(\mu)} \int \mathbf{c}(z_1,\dots,z_K) d\pi(z_1,\dots,z_K) \leq \frac{1}{2\mu(\mathcal{Z})} B^*_\mu.$$

First, observe that for any $\phi \in \Phi$ we have:

$$\begin{aligned}
&\sum_{j=1}^K \int_{\mathcal{X}\times[K]} \phi_j(z_j) \frac{1}{2\mu(\mathcal{Z})} d\mu(z_j) + \frac{1}{2} \sum_{j=1}^K \phi_j(\mathghost) \\
=&\sum_{i=1}^K \int_{\mathcal{X}} \Big( \sum_{j=1}^K \phi_j(x_i, i) + \sum_{j=1}^K \phi_j(\mathghost) \Big) \frac{1}{2\mu(\mathcal{Z})} d\mu_i(x_i).
\end{aligned}$$

For each $1 \leq i \leq K$, define

$$\psi_i(x_i) := \sum_{j=1}^K \phi_j(x_i, i) + \sum_{j=1}^K \phi_j(\mathghost).$$

It is thus clear form the above computation and definition that

$$\sum_{j=1}^K \int_{\mathcal{X}\times[K]} \phi_j(z_j) \frac{1}{2\mu(\mathcal{Z})} d\mu(z_j) + \frac{1}{2} \sum_{j=1}^K \phi_j(\mathghost) = \sum_{i=1}^K \int_{\mathcal{X}} \psi_i(x_i) \frac{1}{2\mu(\mathcal{Z})} d\mu_i(x_i). \tag{4.1}$$

Our goal is now to show that $\{\psi_i : 1 \leq i \leq K\}$ is feasible for problem (3.14) (working with the normalized measure $\frac{1}{2\mu(\mathcal{Z})}\mu$). We start with a preliminary lemma and an example illustrating the strategy behind the proof of this fact. The precise statement appears in **Proposition** 4.2 below.

**Lemma 4.1.** *Given* $(z_1, \dots, z_K) \in \mathcal{Z}_*^K$*, let* $A = \{j \in [K] : z_j \neq \mathghost\}$*. Suppose that for each* $j \in A$ $z_j = (x_j, j)$*. Then, for each* $\phi \in \Phi$*,*

$$\sum_{j=1}^{K} \phi_j(z_j) \leq \frac{1}{K} + \frac{1}{K} c_A. \tag{4.2}$$

*Proof.* Since $\phi \in \Phi$, it suffices to show that

$$B^*_{\widehat{\mu}_{\vec{z}}} \leq \frac{1}{K} + \frac{1}{K} c_A,$$

where

$$\widehat{\mu}_{\vec{z}} = \sum_{l \text{ s.t. } z_l \neq \mathghost}^{K} \frac{1}{K} \delta_{z_l} = \sum_{j \in A} \frac{1}{K} \delta_{z_j}.$$

For simplicity, assume that $A = \{1, \dots, S\}$. By **Proposition** 3.12,

$$B^*_{\widehat{\mu}_{\vec{z}}} = \widetilde{c}((x_1, \frac{1}{K}), \dots, (x_S, \frac{1}{K}), (x_{S+1}, 0), \dots, (x_K, 0)),$$

where we can pick $x_{S+1}, \dots, x_K$ arbitrarily. Let $m_A = \frac{1}{K}$ and $m_{A'} = 0$ for $A' \neq A$. It is easy to check that such $m$ is feasible for (3.9) since $r_s = \frac{1}{K}$ for $1 \leq s \leq S$ and $r_i = 0$ for $S+1 \leq i \leq K$. So, (3.9) implies

$$\widetilde{c}((x_1, \frac{1}{K}), \dots, (x_S, \frac{1}{K}), (x_{S+1}, 0), \dots, (x_K, 0)) \leq \frac{1}{K} + \frac{1}{K} c_A.$$

The conclusion follows. □

We now present specific examples which illustrate why $\{\psi_i : 1 \leq i \leq K\}$ is feasible for (3.14), that is, we need to show that for any $(x_1, \dots, x_K) \in \mathcal{X}^K$ and for any $A \subseteq [K]$ we have

$$\psi_1(x_1) + \psi_2(x_2) + \psi_3(x_3) \leq 1 + c_A.$$

Let $K = 4$ and suppose that $A = \{1, 2, 3\}$. Expanding the $\psi_i$'s we get:

$$\psi_1(x_1) + \psi_2(x_2) + \psi_3(x_3) = \sum_{i=1}^{3} \sum_{j=1}^{4} \phi_j(x_i, i) + 3 \sum_{j=1}^{4} \phi_j(\mathghost),$$

or, after a rearrangement of the summands:

$$\begin{aligned}
&\phi_1(x_1, 1) + \phi_2(x_2, 2) + \phi_3(x_3, 3) + \phi_4(\mathghost) \\
&+\phi_2(x_1, 1) + \phi_3(x_2, 2) + \phi_4(x_3, 3) + \phi_1(\mathghost) \\
&+\phi_3(x_1, 1) + \phi_4(x_2, 2) + \phi_1(x_3, 3) + \phi_2(\mathghost) \\
&+\phi_4(x_1, 1) + \phi_1(x_2, 2) + \phi_2(x_3, 3) + \phi_3(\mathghost) \\
&+2 \sum_{j=1}^{4} \phi_j(\mathghost).
\end{aligned}$$

We can bound the first line above using (4.2):

$$\phi_1(x_1, 1) + \phi_2(x_2, 2) + \phi_3(x_3, 3) + \phi_4(\mathghost) \leq \frac{1}{4} + \frac{1}{4} c_A.$$

The same argument holds for the second, third and fourth lines. For the last line, notice that $\mathbf{c}(\mathghost, \dots, \mathghost) = 0$. Hence, the last line is bounded above by 0 and we can now deduce that

$$\psi_1(x_1) + \psi_2(x_2) + \psi_3(x_3) \leq 1 + c_A.$$

The above situation becomes less trivial if $|A|$ is much smaller than $K$. To illustrate, let $K=9$ and suppose that $A=\{1,2\}$. Rearranging the $\phi_j$'s as above we will not be able to obtain the desired upper bound since the total number of $\phi_j(\ghost)$'s available is in this case $K|A|=18$ while the required number of $\phi_j(\ghost)$'s in the analogous arrangement as above would be at least $K(K-|A|)=63$. To overcome this problem, we need to rearrange the $\phi_j$'s further in order to reduce the required number of $\phi_j(\ghost)$'s and deduce from this refined analysis the desired upper bound.

First of all, construct a $9 \times 9$ arrangement in the following way: for the $k$-th row in the arrangement, let the $k$-th and the $(k+1)$-th elements be $\phi_k(x_1,1)$ and $\phi_{k+1}(x_2,2)$, respectively, and let the remaining elements be "empty". Note that here $k$ and $k+1$ are considered modulo 9; for example, $10 \equiv 1 \mod 9$, and an empty element means literally no element. We merge rows in the following way: merge together the 1-st, the 3-rd, the 5-th and the 7-th rows, i.e. replace empty elements for none-empty ones coming from other rows; likewise, merge together the 2-nd, the 4-th, the 6-th and the 8-th rows; finally, keep the 9-th row as is. By the above construction, the 1-st, the 3-rd, the 5-th and the 7-th rows share no common $\phi_j$. Let $\emptyset_j$ denote an empty element at the $j$-th coordinate. The resulting arrangement can be written as:

$$\begin{aligned}
&\phi_1(x_1,1),\phi_2(x_2,2),\phi_3(x_1,1),\phi_4(x_2,2),\phi_5(x_1,1),\phi_6(x_2,2),\phi_7(x_1,1),\phi_8(x_2,2),\emptyset_9,\\
&\emptyset_1,\phi_2(x_1,1),\phi_3(x_2,2),\phi_4(x_1,1),\phi_5(x_2,2),\phi_6(x_1,1),\phi_7(x_2,2),\phi_8(x_1,1),\phi_9(x_2,2),\\
&\phi_1(x_2,2),\emptyset_2,\emptyset_3,\emptyset_4,\emptyset_5,\emptyset_6,\emptyset_7,\emptyset_8,\phi_9(x_1,1),
\end{aligned}$$

with the first row representing the merge of rows 1-3-5-7, the second row representing the merge of rows 2-4-6-8, and the last row representing row 9.

Notice that the above arrangement contains all $\phi_j(x_s,s)$'s. Furthermore, the number of $\emptyset_j$ for each $1 \le j \le 9$ is exactly 1. Filling $\emptyset_j$'s with $\phi_j(\ghost)$'s, and using the fact that the number of $\phi_j(\ghost)$'s for each $1 \le j \le 9$ is 2, it follows that

$$\begin{aligned}
\psi_1(x_1)+\psi_2(x_2) &= \sum_{j=1}^{4}\big(\phi_{2j-1}(x_1,1)+\phi_{2j}(x_2,2)\big)+\phi_9(\ghost)\\
&\quad+\phi_1(\ghost)+\sum_{j=1}^{4}\big(\phi_{2j}(x_1,1)+\phi_{2j+1}(x_2,2)\big)\\
&\quad+\phi_1(x_2,2)+\sum_{j=2}^{8}\phi_j(\ghost)+\phi_9(x_1,1)\\
&\quad+\sum_{j=1}^{9}\phi_j(\ghost).
\end{aligned}$$

Observe that for $(z_1,\dots,z_K)=((x_1,1),(x_2,2),\dots,(x_1,1),(x_2,2),\ghost)$, $\widehat{\mu}_{\vec{z}}=\frac{4}{9}\delta_{(x_1,1)}+\frac{4}{9}\delta_{(x_2,2)}$. Factoring out the 4 (see *Remark* 2.6) and applying (4.2), we otain:

$$\sum_{j=1}^{4}\big(\phi_{2j-1}(x_1,1)+\phi_{2j}(x_2,2)\big)+\phi_9(\ghost) \le B^*_{\widehat{\mu}_{\vec{z}}} \le \frac{4}{9}+\frac{4}{9}c_A.$$

Similarly, the second and third lines can be bounded by $\frac{4}{9}+\frac{4}{9}c_A$ and $\frac{1}{9}+\frac{1}{9}c_A$, respectively. Since $\sum_{j=1}^{9}\phi_j(\ghost) \le 0$, it follows that

$$\psi_1(x_1)+\psi_2(x_2) \le 1+c_A.$$

The above two situations help us illustrate the general strategy for proving that the resulting $\psi_i$ are feasible: the idea is to arrange summands appropriately so that we can utilize **Lemma** 4.1 in the most effective way possible. In the following proposition we state precisely our aim.

**Proposition 4.2.** *Let $(\phi_1, \dots, \phi_K) \in \Phi$ be a feasible dual potential. For each $1 \leq i \leq K$, define*

$$\psi_i(x_i) := \sum_{j=1}^K \phi_j(x_i, i) + \sum_{j=1}^K \phi_j(\varnothing), \quad x_i \in \mathcal{X}.$$

*Then $\{\psi_i : 1 \leq i \leq K\}$ is feasible for* (3.14)*. Therefore,*

$$\min_{\pi \in \Pi_K(\mu)} \int_{\mathcal{Z}_*{}^K} \mathbf{c}(z_1, \dots, z_K) d\pi(z_1, \dots, z_K) \leq \frac{1}{2\mu(\mathcal{Z})} B_\mu^*.$$

*Proof.* Fix $K$ and $A \subseteq [K]$. Without loss of generality, assume that $A = \{1, \dots, S\}$. We show that

$$\sum_{i \in A} \psi_i(x_i) \leq 1 + c_A. \tag{4.3}$$

First, suppose $K$ is divisible by $S$. For each $1 \leq s \leq S$ and $1 \leq j \leq K$, let

$$u(s,j) := \begin{cases} (s + j - 1 \mod S) & \text{if } (s + j - 1 \mod S) \neq 0 \\ S & \text{if } (s + j - 1 \mod S) = 0. \end{cases}$$

Rearranging the sum of the $\psi$'s, it follows that

$$\begin{aligned} \sum_{i \in A} \psi_i(x_i) &= \sum_{j=1}^K \sum_{s=1}^S \phi_j(x_s, s) + S \sum_{j=1}^K \phi_j(\varnothing) \\ &= \sum_{s=1}^S \sum_{j=1}^K \phi_j(x_{u(s,j)}, u(s,j)) + S \sum_{j=1}^K \phi_j(\varnothing). \end{aligned}$$

Note that for each $1 \leq s \leq S$, $|\{u(s,j) : 1 \leq j \leq K\}| = \frac{K}{S}$, and hence

$$\widehat{\mu}_{\vec{z}} = \sum_{u(s,j)=1}^S \frac{\frac{K}{S}}{K} \delta_{(x_{u(s,j)}, u(s,j))}.$$

Factoring out $\frac{K}{S}$ and applying (4.2),

$$\sum_{j=1}^K \phi_j(x_{u(s,j)}, u(s,j)) \leq \frac{K}{S}\big(\frac{1}{K} + \frac{1}{K} c_A\big) = \frac{1}{S} + \frac{1}{S} c_A.$$

Since $\sum_{j=1}^K \phi_j(\varnothing) \leq 0$, it is deduced that

$$\begin{aligned} \sum_{i \in A} \psi_i(x_i) &= \sum_{s=1}^S \sum_{j=1}^K \phi_j(x_{u(s,j)}, u(s,j)) + S \sum_{j=1}^K \phi_j(\varnothing) \\ &\leq \sum_{s=1}^S \big(\frac{1}{S} + \frac{1}{S} c_A\big) \\ &= 1 + c_A, \end{aligned}$$

proving the desired inequality in this first case.

Now suppose that $K$ is not divisible by $S$. For each $1 \le s \le S$ and each $1 \le k \le K$, let

$$v(s,k) := \begin{cases} (s+k-1 \mod K) & \text{if } (s+k-1 \mod K) \neq 0 \\ K & \text{if } (s+k-1 \mod K) = 0. \end{cases}$$

Construct a $K \times K$ arrangement in the following way: for each $1 \le s \le S$ we set the $v(s,k)$-th element to be $\phi_{v(s,k)}(x_s, s)$, and we set the remaining elements to be empty. We use $\emptyset_j$ to denote an empty element at the $j$-th coordinate. Note that the $k$-th row has $\phi_{v(1,k)}(x_1, 1), \dots, \phi_{v(S,k)}(x_S, S)$ as non-empty elements, which are placed from the $v(1,k)$-th coordinate to the $v(S,k)$-th coordinate, while it has $(K-S)$ many empty elements. For example, the 3-rd row is

$$\emptyset_1, \emptyset_2, \phi_3(x_1, 1), \dots, \phi_{S+2}(x_S, S), \emptyset_{S+3}, \dots, \emptyset_K.$$

We split into two further subcases.

First, assume that $\lfloor \frac{K}{S} \rfloor = 1$. In that case we have $K(K-S) \le KS$. For each $1 \le k \le K$ collect all the $\phi_j(\ghost)$'s such that $j \notin A_k := \{v(1,k), \dots, v(S,k)\}$. Notice that for fixed $j$, the number of $k$'s such that $j \notin A_k$ is exactly $K-S$ since all $\phi_j(x_s, s)$'s are contained in this arrangement and $\lfloor \frac{K}{S} \rfloor = 1$. In other words, the total number of $\emptyset_j$ is smaller than the total number of $\phi_j(\ghost)$. From the above and an application of (4.2) we deduce that

$$\begin{aligned} \sum_{i \in A} \psi_i(x_i) &= \sum_{k=1}^{K} \Big( \sum_{s=1}^{S} \phi_{v(s,k)}(x_s, s) + \sum_{j \notin A_k} \phi_j(\ghost) \Big) + (2S-K) \sum_{j=1}^{K} \phi_j(\ghost) \\ &\le \sum_{k=1}^{K} \big( \frac{1}{K} + \frac{1}{K} c_A \big) \\ &= 1 + c_A, \end{aligned}$$

proving the desired inequality in this case.

Finally, assume that $\lfloor \frac{K}{S} \rfloor > 1$. Here the idea is to merge $\lfloor \frac{K}{S} \rfloor$-many rows to a single row. We do this in the following way: for each $1 \le s \le S$, we merge together the $s$-th row, the $(S+s)$-th row, $\dots$, and the $((\lfloor \frac{K}{S} \rfloor - 1)S + s)$-th row, to obtain a single row which will be re-indexed by $s$. In the original arrangement, since the $((m-1)S+s)$-th row has $\phi_{v(s,(m-1)S+1)}(x_1, 1), \dots, \phi_{v(s,mS)}(x_S, S)$ as non-empty elements, the rows that get merged share no common $\phi_j$. We keep the last $(K - \lfloor \frac{K}{S} \rfloor S)$-many rows in the original arrangement the same, and for convenience we let the indices of these rows be unchanged. After this procedure, we obtain $S$-many merged rows and $(K - \lfloor \frac{K}{S} \rfloor S)$-many remaining original rows. Now, it is necessary to count, for every fixed $j$, the total number of empty elements $\emptyset_j$ in this new arrangement. If the number of $\emptyset_j$'s was smaller than or equal to $S$ for all $1 \le j \le K$, we would be done since the number of $\phi_j(\ghost)$ would be $S$ for each $j$, whence it would be possible to replace the $\emptyset_j$'s with $\phi_j(\ghost)$'s. We show that this is indeed the case.

For each merged row, its non-empty elements are

$$\phi_{v(s,1)}(x_1, 1), \dots, \phi_{v(s,S)}(x_S, S), \dots, \phi_{v(s,(\lfloor \frac{K}{S} \rfloor - 1)S+1)}(x_1, 1), \dots, \phi_{v(s, \lfloor \frac{K}{S} \rfloor S)}(x_S, S).$$

Observe that for each merged row, the index $j$ of $\emptyset_j$ varies from $v(s, \lfloor \frac{K}{S} \rfloor S + 1)$ to $v(s, K)$. The definition of $v(s,k)$ yields that

$$v(s, \lfloor \frac{K}{S} \rfloor S + 1) = \lfloor \frac{K}{S} \rfloor S + s \text{ if } 1 \le s \le K - \lfloor \frac{K}{S} \rfloor S, \tag{4.4}$$

$$v(s, \lfloor \frac{K}{S} \rfloor S + 1) = \lfloor \frac{K}{S} \rfloor S + s - K \text{ if } K - \lfloor \frac{K}{S} \rfloor S + 1 \le s \le S \tag{4.5}$$

and

$$v(s,K) = K \text{ if } s=1, \tag{4.6}$$

$$v(s,K) = s-1 \text{ if } 2 \le s \le S. \tag{4.7}$$

To count the total number of $\emptyset_j$'s in the merged rows, let's consider the following sub-cases.

(i) $\lfloor \frac{K}{S} \rfloor S + 1 \le j \le K$ : By (4.4), if $1 \le s \le K - \lfloor \frac{K}{S} \rfloor S$, then the $s$-th row has $\emptyset_j$ for $\lfloor \frac{K}{S} \rfloor S + s \le j \le K$. Also, by (4.5) and (4.7), if $K - \lfloor \frac{K}{S} \rfloor S + 1 \le s \le S$, then no merged row has such $\emptyset_j$. Hence, the number of $\emptyset_j$ is $j - \lfloor \frac{K}{S} \rfloor S$.

(ii) $S \le j \le \lfloor \frac{K}{S} \rfloor S$ : It follows from (4.4) and (4.5) that either $v(s, \lfloor \frac{K}{S} \rfloor S + 1) > \lfloor \frac{K}{S} \rfloor S$ or $v(s, \lfloor \frac{K}{S} \rfloor S + 1) < S$. Similarly, it follows from (4.6) and (4.7) that either $v(s,K) > \lfloor \frac{K}{S} \rfloor S$ or $v(s,K) < S$. Since the index $j$ of $\emptyset_j$ of the $s$-th merged row varies from $v(s, \lfloor \frac{K}{S} \rfloor S + 1)$ to $v(s,K)$, the number of $\emptyset_j$ is 0.

(iii) $S - (K - \lfloor \frac{K}{S} \rfloor S) + 1 \le j \le S - 1$ : By (4.5) and (4.7), if $S - (K - \lfloor \frac{K}{S} \rfloor S) + 1 \le j \le S - 1$, then $\emptyset_j$ appears from the $(j+1)$-st merged row to the $S$-th merged row. Hence, the number of $\emptyset_j$ is $S - j$.

(iv) $1 \le j \le S - (K - \lfloor \frac{K}{S} \rfloor S)$ : Similar to (iii), if $1 \le j \le S - (K - \lfloor \frac{K}{S} \rfloor S)$, then $\emptyset_j$ appears from the $(j+1)$-st merged row to the $S$-th merged row. Hence, the number of $\emptyset_j$ is $K - \lfloor \frac{K}{S} \rfloor S$.

To summarize, in the merged rows

$$\text{the number of } \emptyset_j = \begin{cases} j - \lfloor \frac{K}{S} \rfloor S & \text{for } \lfloor \frac{K}{S} \rfloor S + 1 \le j \le K, \\ 0 & \text{for } S \le j \le \lfloor \frac{K}{S} \rfloor S, \\ S - j & \text{for } S - (K - \lfloor \frac{K}{S} \rfloor S) + 1 \le j \le S-1, \\ K - \lfloor \frac{K}{S} \rfloor S & \text{for } 1 \le j \le S - (K - \lfloor \frac{K}{S} \rfloor S). \end{cases} \tag{4.8}$$

Now, it remains to count the total number of $\emptyset_j$ in the last $(K - \lfloor \frac{K}{S} \rfloor S)$-many remaining original rows. In this part, each row has only $S$-many non-empty elements. Recall that we still use the same index $k$ for these remaining rows. Precisely, for $\lfloor \frac{K}{S} \rfloor S + 1 \le k \le K$, the $k$-th row has

$$\phi_{v(1,k)}(x_1, 1), \phi_{v(2,k)}(x_2, 2), \dots, \phi_{v(S,k)}(x_S, S).$$

Recall that $A_k := \{v(1,k), \dots, v(S,k)\}$. To count the total number of $\emptyset_j$'s in the original rows, let's consider the following sub-cases.

(i) $\lfloor \frac{K}{S} \rfloor S + 1 \le j \le K$. : If $1 \le j + 1 - k \le S$, by the definition of $v(s,k)$, then $j \in A_k$. In other words, each $k$-th row has $\emptyset_j$ for $k > j$. Hence, the number of $\emptyset_j$ is $K - j$.

(ii) $S \le j \le \lfloor \frac{K}{S} \rfloor S$ : From the definition of $v(s,k)$ and the range of $k$, we deduce that if $\lfloor \frac{K}{S} \rfloor S + 1 \le k \le K$, then $v(1,k) > \lfloor \frac{K}{S} \rfloor S$ and $v(S,k) < S$. In other words, $\emptyset_j$ for $S \le j \le \lfloor \frac{K}{S} \rfloor S$ appears in every row. Hence, the number of $\emptyset_j$ is $K - \lfloor \frac{K}{S} \rfloor S$.

(iii) $S - (K - \lfloor \frac{K}{S} \rfloor S) + 1 \le j \le S - 1$ : Since $\lfloor \frac{K}{S} \rfloor S + 1 \le k \le K$, if $v(S,k) = S + k - K < j$, then $j \notin A_k$. This yields that if $\lfloor \frac{K}{S} \rfloor S + 1 \le k \le K - S + j$, then the $k$-th row has $\emptyset_j$. Hence, the number of $\emptyset_j$ is $K - \lfloor \frac{K}{S} \rfloor S - S + j$.

(iv) $1 \le j \le S - (K - \lfloor \frac{K}{S} \rfloor S)$ : Since $v(S, \lfloor \frac{K}{S} \rfloor S + 1) = S - (K - \lfloor \frac{K}{S} \rfloor S)$, if $1 \le j \le S - (K - \lfloor \frac{K}{S} \rfloor S)$ and $\lfloor \frac{K}{S} \rfloor S + 1 \le k \le K$, then $j \in A_k$. Hence, the number of $\emptyset_j$ is 0.

To summarize, in the remaining original rows

$$\text{the number of } \emptyset_j = \begin{cases} K - j & \text{for } \lfloor \frac{K}{S} \rfloor S + 1 \le j \le K, \\ K - \lfloor \frac{K}{S} \rfloor S & \text{for } S \le j \le \lfloor \frac{K}{S} \rfloor S, \\ K - \lfloor \frac{K}{S} \rfloor S - S + j & \text{for } S - (K - \lfloor \frac{K}{S} \rfloor S) + 1 \le j \le S-1, \\ 0 & \text{for } 1 \le j \le S - (K - \lfloor \frac{K}{S} \rfloor S). \end{cases} \tag{4.9}$$

Combining (4.8) with (4.9), the total number of $\emptyset_j$ is always exactly equal to $K - \lfloor \frac{K}{S} \rfloor S$ which is always less than $S$. This allows us to change every appearance of $\emptyset_j$ with a $\phi_j(\text{👻})$. Accordingly, using $\sum \phi_j(\text{👻}) \leq 0$, we deduce

$$\begin{aligned} \sum_{i \in A} \psi_i(x_i) \leq & \sum_{\text{merged rows}} \Big( \sum_{v(s,j)} \phi_{v(s,j)}(x_s, s) + \sum_{l \neq v(s,j)} \phi_l(\text{👻}) \Big) \\ & + \sum_{\text{remaining rows}} \Big( \sum_{v(s,j)} \phi_{v(s,j)}(x_s, s) + \sum_{l \neq v(s,j)} \phi_l(\text{👻}) \Big). \end{aligned}$$

Let's focus on the first summation over merged rows. Notice that there are $\lfloor \frac{K}{S} \rfloor S$ many real elements and the set of real elements is $\{(x_1, 1), \dots, (x_S, S)\}$. Thus,

$$\widehat{\mu}_{\vec{z}} = \sum_{s=1}^{S} \frac{\lfloor \frac{K}{S} \rfloor}{K} \delta_{(x_s, s)}.$$

Factoring out $\lfloor \frac{K}{S} \rfloor$ and applying (4.2) we obtain

$$\sum_{v(s,j)} \phi_{v(s,j)}(x_s, s) + \sum_{l \neq v(s,j)} \phi_l(\text{👻}) \leq \frac{\lfloor \frac{K}{S} \rfloor}{K} + \frac{\lfloor \frac{K}{S} \rfloor}{K} c_A.$$

On the other hand, for the second summation over remaining rows, there are $S$ many real elements. Thus,

$$\widehat{\mu}_{\vec{z}} = \sum_{s=1}^{S} \frac{1}{K} \delta_{(x_s, s)}.$$

(4.2) immediately implies

$$\sum_{v(s,j)} \phi_{v(s,j)}(x_s, s) + \sum_{l \neq v(s,j)} \phi_l(\text{👻}) \leq \frac{1}{K} + \frac{1}{K} c_A.$$

Note that the number of merged rows is $S$ and the number of remaining original rows is $K - \lfloor \frac{K}{S} \rfloor S$, respectively. Combining all arguments, we can infer that

$$\begin{aligned} \sum_{i \in A} \psi_i(x_i) &\leq \frac{\lfloor \frac{K}{S} \rfloor S}{K} + \frac{\lfloor \frac{K}{S} \rfloor S}{K} c_A + \frac{K - \lfloor \frac{K}{S} \rfloor S}{K} + \frac{K - \lfloor \frac{K}{S} \rfloor S}{K} c_A \\ &= 1 + c_A, \end{aligned}$$

obtaining the desired inequality in the last remaining case.

In summary, we have proved that for a given $\phi = (\phi_1, \dots, \phi_K) \in \Phi$, its associated $(\psi_1, \dots, \psi_K)$ (which satisfies (4.1)) is feasible for (3.14). Consequently, this means that

$$(2.11) \leq \frac{1}{2\mu(\mathcal{Z})} (3.14) \tag{4.10}$$

In turn, by weak duality this automatically implies that

$$(2.11) \leq \frac{1}{2\mu(\mathcal{Z})} B_\mu^*.$$

Finally, combining with **Corollary** 4.5 below (which establishes that under **Assumption** 2.1 there is no duality gap for the MOT problem (1.2)) we obtain the desired inequality relating the minimum value for the MOT problem and $B_\mu^*$. □

4.3. **Returning to the adversarial problem** (1.1)**.** We begin by establishing that, under **Assumption** 2.1, the cost $\mathbf{c}$ is lower semi-continuous with respect to a suitable notion of convergence.

**Proposition 4.3.** *Let $\mathcal{Z}_* = \mathcal{Z} \cup \{\mathghost\}$ on which $\mathghost$ is considered as an isolated point. Let $\widehat{d}$ be defined according to:*

$$\widehat{d}(z, z') := \begin{cases} d(x, x') & \text{if } i = i', \\ \infty & \text{if } i \neq i' \text{ or } z = \mathghost \text{ and } z' \in \mathcal{Z} \text{ (vice-versa)}, \\ 0 & \text{if } z = z' = \mathghost. \end{cases}$$

*Define $\widehat{d}_K$ on $\mathcal{Z}_*^K$ by*

$$\widehat{d}_K((z_1, \dots, z_K), (z_1', \dots, z_K')) := \max_{1 \leq i \leq K} \widehat{d}(z_i, z_i').$$

*Recall*

$$\mathbf{c}(z_1, \dots, z_K) := B^*_{\widehat{\mu}_{\vec{z}}}$$

*where $\widehat{\mu}_{\vec{z}}$ is defined as*

$$\widehat{\mu}_{\vec{z}} := \frac{1}{K} \sum_{l \text{ s.t. } z_l \neq \mathghost}^{K} \delta_{z_l},$$

*Under **Assumption** 2.1, $\mathbf{c}$ is lower semi-continuous on $(\mathcal{Z}_*^K, \widehat{d}_K)$.*

*Remark* 4.4. Note that $(\mathcal{Z}_*^K, \widehat{d}_K)$ is still a Polish space.

*Proof.* Suppose $\vec{z}^n = (z_1^n, \dots, z_K^n)$ converges to $\vec{z} = (z_1, \dots, z_K)$ in $(\mathcal{Z}_*^K, \widehat{d}_k)$. Without loss of generality, assume that $z_1, \dots, z_L = \mathghost$ for all $1 \leq L \leq K$. If $L = K$, the claim would be trivial and so we can focus on the case $L < K$. By the definition of $\widehat{d}_K$, without loss of generality we can further assume that $z_1^n, \dots, z_L^n = \mathghost$ for all $n$, and likewise, for each $L+1 \leq j \leq K$, we can assume that $i_j^n = i_j$ for all $n$, for otherwise the convergence would not hold due to the definition of $\widehat{d}_K$.

By **Proposition** 3.12 we have

$$\mathbf{c}(z_1^n, , \dots, z_K^n) = B^*_{\widehat{\mu}_{\vec{z}^n}} = \inf_{m: S_K \to \mathbb{R}} \sum_{A \subseteq \{L+1, \dots, K\}} m_A\big(c_A(x_{L+1}^n, \dots, x_K^n) + 1\big), \tag{4.11}$$

where the min ranges over all $\{m_A\}_{A \subseteq \{L+1, \dots, K\}}$ such that $\sum_{A \in S_K(i) \cap \{L+1, \dots, K\}} m_A = \frac{1}{K}, \quad \forall i = L+1, \dots, K$.

We now claim that for every $A \subseteq \{L+1, \dots, K\}$,

$$c_A(x_{L+1}, \dots, x_K) \leq \liminf_{n \to \infty} c_A(x_{L+1}^n, \dots, x_K^n).$$

Indeed, if the right hand side is equal to $+\infty$, then there is nothing to prove. If the right hand side is finite, we may then find a sequence $\{\tilde{x}^n\}_{n \in \mathbb{N}}$ such that

$$\liminf_{n \to \infty} \sum_{i \in A} c(\tilde{x}^n, x_i^n) = \liminf_{n \to \infty} c_A(x_{L+1}^n, \dots, x_K^n) < \infty.$$

By the compactness property in **Assumption** 2.1 it follows that up to subsequence (not relabeled) we have that $\{\tilde{x}^n\}_{n \in \mathbb{N}}$ converges toward a point $\tilde{x} \in \mathcal{X}$. Combining with the lower semi-continuity of $c$, we deduce that

$$c_A(x_{L+1}, \dots, x_K) \leq \sum_{i \in A} c(\tilde{x}, x_i) \leq \liminf_{n \to \infty} c_A(x_{L+1}^n, \dots, x_K^n),$$

as we wanted to show.

Returning to (4.11), we can find for each $n \in \mathbb{N}$ a collection of feasible $\{m_A^n\}_{A \subseteq \{L+1,\ldots,K\}}$ such that

$$\liminf_{n\to\infty} \sum_{A\subseteq\{L+1,\ldots,K\}} m_A^n\big(c_A(x_{L+1}^n,\ldots,x_K^n)+1\big) = \liminf_{n\to\infty} \mathbf{c}(z_1^n,\ldots,z_K^n).$$

Using the Heine-Borel theorem in Euclidean space, we can assume without the loss of generality that for every $A$, $m_A^n$ converges to some $m_A$ as $n\to\infty$. The resulting collection of $m_A$ is feasible for the problem defining $\mathbf{c}(z_1,\ldots,z_K)$ and thus, using the lower semicontinuity of $c_A$ established earlier, we deduce:

$$\mathbf{c}(z_1,\ldots,z_K) \le \sum_{A\subseteq\{L+1,\ldots,K\}} m_A\big(c_A(x_{L+1}^n,\ldots,x_K^n)+1\big) \le \liminf_{n\to\infty} \mathbf{c}(z_1^n,\ldots,z_K^n).$$

□

**Corollary 4.5.** *(Duality of MOT) Under **Assumption** 2.1,*

$$\inf_{\pi\in\Pi_K(\mu)} \int_{{\mathcal{Z}_*}^K} \mathbf{c}(z_1,\ldots,z_K)d\pi(z_1,\ldots,z_K) = \sup_{\phi\in\Phi}\Big\{\sum_{j=1}^K \int_{\mathcal{X}\times[K]} \phi_j(z_j)\frac{1}{2\mu(\mathcal{Z})}d\mu(z_j) + \frac{1}{2}\sum_{j=1}^K \phi_j(\ghost)\Big\}.$$

*Furthermore, a minimizer $\pi^*$ exists.*

*Proof.* From **Proposition** 4.3 it follows that the cost function $\mathbf{c}(z_1,\ldots,z_K)$ is lower semi-continuous on $(\mathcal{Z}_*^K, \widehat{d}_K)$, which is a Polish space. Applying **Theorem** 1.3 in [Vil03], which is stated for the usual optimal transport, but that can be generalized to the MOT setting, we obtain the desired duality. The existence of a minimizer $\pi^*$ follows from the lower semi-continuity of $\mathbf{c}(z_1,\ldots,z_K)$ and compactness of $\Pi_K(\mu)$. □

**Corollary 4.6.** *Under **Assumption** 2.1,* (2.6) = (2.7).

*Proof.* By the upper bound from section 4.1 we have

$$\frac{1}{2\mu(\mathcal{Z})}B_\mu^* \le \min_{\pi\in\Pi_K(\mu)} \int \mathbf{c}(z_1,\ldots,z_K)d\pi(z_1,\ldots,z_K).$$

On the other hand, from (4.10) and **Corollary** 4.5 we have

$$\min_{\pi\in\Pi_K(\mu)} \int \mathbf{c}(z_1,\ldots,z_K)d\pi(z_1,\ldots,z_K) = (2.11) \le \frac{1}{2\mu(\mathcal{Z})}(3.14) \le \frac{1}{2\mu(\mathcal{Z})}B_\mu^*.$$

Combining these two inequalities we conclude that all the above terms must be equal. In particular, (3.14) = $B_\mu^*$. Finally, by **Proposition** 3.13 we know that (3.14) = (3.13) = (2.6). In particular, (2.7) = $B_\mu^*$ = (2.6).

□

**Corollary 4.7.** *Suppose that **Assumption** 2.1 holds and that $(\pi^*,\phi^*)$ is a solution pair for the MOT problem and its dual. Define $f^*$ and $\widetilde{\mu}^*$ according to:*

$$f_i^*(x) := \Big(\max\Big\{\sum_{j=1}^K \phi_j^*(x,i) + \sum_{j=1}^K \phi_j^*(\ghost), 0\Big\}\Big)^{\bar{c}}$$

*and for any test function $h$ on $\mathcal{X}$,*

$$\int_{\mathcal{X}} h(\widetilde{x})d\widetilde{\mu}_i^*(\widetilde{x}) := \int_{\mathcal{Z}_*^K}\left\{\int_{\mathcal{X}} h(\widetilde{x})d\widetilde{\mu}_{\vec{z},i}^*(\widetilde{x})\right\}d\pi^*(\vec{z}),$$

*where $\widetilde{\mu}_{\vec{z},i}^*$ is the $i$-th marginal of $\widetilde{\mu}_{\vec{z}}^*$, an optimal adversarial attack which achieves $\mathbf{c}(z_1,\ldots,z_K)$ given $\vec{z}=(z_1,\ldots,z_K)$. Then $(f^*,\widetilde{\mu}^*)$ is a saddle for problem* (1.1).

*Proof.* It suffices to show that $(f^*, \widetilde{\mu}^*)$ is a saddle point for problem (2.6). For any $\widetilde{\mu}$, it holds that

$$\begin{aligned}\int_{\mathcal{X}} h(\widetilde{x}) d\widetilde{\mu}_i(\widetilde{x}) &= \sum_{A \in S_K(i)} \int_{\mathcal{X}} h(\widetilde{x}) d\lambda_A(\widetilde{x}) \\ &= \sum_{A \in S_K(i)} \int_{\mathcal{X}^K} h(T_A(x_1, \dots, x_K)) d\pi_A(x_1, \dots, x_K).\end{aligned}$$

Hence, for the above $(f^*, \widetilde{\mu}^*)$,

$$\begin{aligned}B(f^*, \widetilde{\mu}^*) + C(\mu, \widetilde{\mu}^*) &= \sum_{i=1}^K \int_{\mathcal{X}} f_i^*(\widetilde{x}_i) d\widetilde{\mu}_i^*(x) + \sum_{i=1}^K C(\widetilde{\mu}_i^*, \mu_i) \\ &= \sum_{A \in S_K} \sum_{i \in A} \left\{ \int_{\mathcal{X}} f_i^*(\widetilde{x}) d\lambda_A^*(\widetilde{x}) + C(\lambda_A^*, \mu_{i,A}) \right\} \\ &= \sum_{A \in S_K} \left\{ \int_{\mathcal{X}^K} \big( \sum_{i \in A} f_i^*(T_A(\vec{x})) + c_A(\vec{x}) \big) d\pi_A^*(\vec{x}) \right\}\end{aligned}$$

where $\lambda_A^*$ and $\pi_A^*$ correspond to $\widetilde{\mu}^*$. Notice that by **Proposition** 3.13, without loss of generality, we can assume that $\sum \phi_j^*(x, i) + \sum \phi_j^*(\ghost) \geq 0$. By the definition of $f_i^*$ and (2.4),

$$\begin{aligned}\sum_{i \in A} f_i^*(T_A(\vec{x})) &= \sum_{i \in A} \sup_{x'} \Big\{ \sum_{j=1}^K \phi_j^*(x', i) + \sum_{j=1}^K \phi_j^*(\ghost) - c(x', T_A(\vec{x})) \Big\} \\ &\leq \sup \Big\{ \sum_{i \in A} \big( \sum_{j=1}^K \phi_j^*(x_i', i) + \sum_{j=1}^K \phi_j^*(\ghost) \big) - c_A(x_i' : i \in A) \Big\} \\ &\leq \sup \Big\{ 1 + c_A(x_i' : i \in A) - c_A(x_i' : i \in A) \Big\} \\ &\leq 1,\end{aligned}$$

where the third inequality follows from (4.3). Hence,

$$\begin{aligned}B(f^*, \widetilde{\mu}^*) + C(\mu, \widetilde{\mu}^*) &\leq \sum_{A \in S_K} \int_{\mathcal{X}^K} \big(1 + c_A(\vec{x})\big) d\pi_A^*(\vec{x}) \\ &= \int_{\mathcal{Z}_*^K} \mathbf{c}(z_1, \dots, z_K) d\pi^*(z_1, \dots, z_K) \\ &= B_\mu^*.\end{aligned}$$

On the other hand, using (3.13) and the optimality of $\phi^*$,

$$B(f^*, \widetilde{\mu}^*) + C(\mu, \widetilde{\mu}^*) \geq \inf_{\widetilde{\mu}} \{ B(f^*, \widetilde{\mu}) + C(\mu, \widetilde{\mu}) \} = \sum_{i=1}^K \int_{\mathcal{X}} (f_i^*)^c(x_i) d\mu_i(x_i) = B_\mu^*.$$

Combining the above upper bound, we can infer that

$$B(f^*, \widetilde{\mu}^*) + C(\mu, \widetilde{\mu}^*) = B_\mu^*,$$

which verifies that $(f^*, \widetilde{\mu}^*)$ is a saddle point for (2.7): hence for (2.6) and thus for (1.1) also. □

*Remark* 4.8. Recently, there are many papers which try to analyze adversarial learning in terms of a game-theoretic perspective [BGB+20, MSP+21, PJ21b]. This approach is obviously natural: the learner wants to maximize $B_\mu^*$ while the adversary wants to maximize $R_\mu^*$, hence to minimize

$B^*_\mu$, which is the usual zero-sum game. The focus of recent studies in this area has been on parametric setups.

**Corollary 4.9.** *Let $\pi^*$ be a solution of the MOT problem* (1.2) *and let $F : \mathcal{Z}_*^K \to \mathcal{Z}_*^K$ be defined according to*

$$F(z_1, \dots, z_K) = (z_{\sigma(1)}, \dots, z_{\sigma(K)}),$$

*for $\sigma : [K] \to [K]$ a permutation. Then any convex combination of $F_\sharp \pi^*$ and $\pi^*$ is also a solution.*

*Proof.* This follows immediately from the fact that the cost function $\mathbf{c}$ is invariant under permutations and the fact that all marginals of $\pi^*$ are the same. □

## 5. Examples and Numerical experiments

Through this section, the cost $c$ is from **Example** 2.3. This cost has been widely used in adversarial learning literature and distributional robust optimization literature. Examples in this sections illuminates how our general framework of generalized barycenter and MOT has applications in practice.

5.1. **Recovery of the binary case.** Consider the binary case $K = 2$. Our goal is to show that our results recover the result in [GM20].

Let $z_1, z_2 \in \mathcal{Z}_*$. If both $z_1$ and $z_2$ are $\Ghost$, then $\mathbf{c}(z_1, z_2) = 0$. If only one of them is $\Ghost$, then the cost is $\frac{1}{2}$. Finally, consider the case where $z_1, z_2 \neq \Ghost$. First assume that $i_1 = i_2 = 1$. In that case,

$$\widehat{\mu}_{\vec{z}} = \frac{1}{2}\delta_{(x_1,1)} + \frac{1}{2}\delta_{(x_2,1)}.$$

Since only class 1 is represented in this configuration, there is no meaningful adversarial attack in this case, and thus,

$$B^*_{\widehat{\mu}_{\vec{z}}} = 1.$$

Assume now that $i_1 = 1$ and $i_2 = 2$. In that case,

$$\widehat{\mu}_{\vec{z}} = \frac{1}{2}\widehat{\mu}_1 + \frac{1}{2}\widehat{\mu}_2 = \frac{1}{2}\delta_{(x_1,1)} + \frac{1}{2}\delta_{(x_2,2)},$$

and the adversary can attack meaningfully if and only if $d(x_1, x_2) \leq 2\varepsilon$. Thus,

$$B^*_{\widehat{\mu}_{\vec{z}}} = \begin{cases} \frac{1}{2} & \text{if } d(x_1, x_2) \leq 2\varepsilon, \\ 1 & \text{if } d(x_1, x_2) > 2\varepsilon. \end{cases}$$

To summarize,

$$\mathbf{c}(z_1, z_2) = \begin{cases} \frac{1}{2} & \text{if } i_1 \neq i_2 \text{ and } d(x_1, x_2) \leq 2\varepsilon, \\ 1 & \text{if } i_1 = i_2 \text{ or } d(x_1, x_2) > 2\varepsilon, \\ \frac{1}{2} & \text{if exactly one of } z_i\text{'s is } \Ghost, \\ 0 & \text{if } z_1 = z_2 = \Ghost. \end{cases}$$

In [GM20], it is proved that

$$B^*_\mu = \inf_{\tilde{\pi} \in \Gamma(\mu,\mu)} \int_{\mathcal{Z}\times\mathcal{Z}} \Big(\frac{\mathrm{cost}_\varepsilon(z_1, z_2) + 1}{2}\Big) d\tilde{\pi}(z_1, z_2),$$

where

$$\mathrm{cost}_\varepsilon(z_1, z_2) = \begin{cases} 0 & \text{if } i_1 \neq i_2 \text{ and } d(x_1, x_2) \leq 2\varepsilon, \\ 1 & \text{if } i_1 = i_2 \text{ or } d(x_1, x_2) > 2\varepsilon. \end{cases}$$

In other words, in the binary case, it is unnecessary to introduce the element $\Ghost$. To illustrate this point, assume for simplicity that $\mu(\mathcal{Z}) = 1$. Notice that every $\tilde{\pi} \in \Gamma(\mu, \mu)$ induces a $\pi \in \Pi_2(\mu)$ as follows:

$$\int_{\mathcal{Z}_*\times\mathcal{Z}_*} \varphi(z_1, z_2) d\pi(z_1, z_2) := \frac{1}{2}\int_{\mathcal{Z}\times\mathcal{Z}} \varphi(z_1, z_2) d\tilde{\pi}(z_1, z_2) + \frac{1}{2}\varphi(\Ghost, \Ghost),$$

where $\varphi : \mathcal{Z}_* \times \mathcal{Z}_* \to \mathbb{R}$ is an arbitrary test function. The cost associated to the induced $\pi$ is:

$$2\int_{\mathcal{Z}_*\times\mathcal{Z}_*} \mathbf{c}(z_1,z_2)d\pi(z_1,z_2) = \int_{\mathcal{Z}\times\mathcal{Z}} \mathbf{c}(z_1,z_2)d\tilde{\pi}(z_1,z_2) = \int_{\mathcal{Z}\times\mathcal{Z}} \Big(\frac{\text{cost}_\varepsilon(z_1,z_2)+1}{2}\Big)d\tilde{\pi}(z_1,z_2).$$

On the other hand, let $\pi$ be a solution for the MOT problem (1.2) (such a solution exists thanks to **Proposition** 4.5). Thanks to **Corollary** 4.9, we can assume without loss of generality that

$$\pi(A\times A') = \pi(A'\times A),$$

for all $A, A'$ measurable subsets of $\mathcal{Z}_*$. We now define $\tilde{\pi}$ according to:

$$\begin{aligned}\int_{\mathcal{Z}\times\mathcal{Z}} \tilde{\varphi}(z_1,z_2)d\tilde{\pi}(z_1,z_2) := 2\int_{\mathcal{Z}\times\mathcal{Z}} \tilde{\varphi}(z_1,z_2)d\pi(z_1,z_2)\\ + \int_{\mathcal{Z}\times\{\ghost\}} \tilde{\varphi}(z_1,z_1)d\pi(z_1,z_2) + \int_{\{\ghost\}\times\mathcal{Z}} \tilde{\varphi}(z_2,z_2)d\pi(z_1,z_2),\end{aligned}$$

for test functions $\tilde{\varphi} : \mathcal{Z}\times\mathcal{Z} \to \mathbb{R}$. It follows that $\tilde{\pi} \in \Gamma(\mu,\mu)$. Moreover, from the above formula and the expressions for the cost $\mathbf{c}$ we get

$$\int_{\mathcal{Z}\times\mathcal{Z}} \Big(\frac{\text{cost}_\varepsilon(z_1,z_2)+1}{2}\Big)d\tilde{\pi}(z_1,z_2) = \int_{\mathcal{Z}\times\mathcal{Z}} \mathbf{c}(z_1,z_2)d\tilde{\pi}(z_1,z_2) = 2\int_{\mathcal{Z}_*\times\mathcal{Z}_*} \mathbf{c}(z_1,z_2)d\pi(z_1,z_2).$$

The above computations show that our results indeed recover those from [GM20] for the binary case.

5.2. **Toy example: three points distribution.** Let's assume that $K = 3$ and $\mu$ is such that

$$\mu_1 = \omega_1\delta_{x_1},\ \ \mu_2 = \omega_2\delta_{x_2},\ \ \mu_3 = \omega_3\delta_{x_3},$$

for three points $x_1, x_2, x_3$ in Euclidean space. Without loss of generality, assume further that $\omega_1 \geq \omega_2 \geq \omega_3 > 0$ and $\sum \omega_i = 1$. Let $\varepsilon > 0$ be given and consider the cost from **Example** 2.3 with $d$ as the Euclidean distance (for simplicity). We will explicitly construct an optimal robust classifier and an optimal adversarial attack for this problem. Even in this simple scenario, one can observe non-trivial situations.

Since for every $\widetilde{\mu}_i$ such that $W_\infty(\omega_i\delta_{x_i}, \widetilde{\mu}_i) \leq \varepsilon$ we have

$$\int_{\mathcal{X}} f_i(x_i)d\widetilde{\mu}_i(x_i) = \int_{\overline{B}(x_i,\varepsilon)} f_i(x_i)d\widetilde{\mu}_i(x_i),$$

where $\overline{B}(x,r) = \{x' : d(x,x') \leq r\}$, we can assume without loss of generality that $\text{spt}(\widetilde{\mu}_i) \subseteq \overline{B}(x_i,\varepsilon)$. Notice that it is sufficient to consider $f \in \mathcal{F}$ restricted to $\overline{B}(x_1,\varepsilon) \cup \overline{B}(x_2,\varepsilon) \cup \overline{B}(x_3,\varepsilon)$ (in fact, problem (1.1) can not disambiguate the values of $f$ outside of this set). We consider 4 non-trivial configurations and one trivial one. Figure 1 below illustrates how the adversary perturbs the original data distribution in each of the non-trivial cases.

**Case 1.** $d(x_i,x_j) > 2\varepsilon$ for all $1 \leq i \neq j \leq 3$. This is a trivial case. We claim that for any $\widetilde{\mu}_i$ such that $W_\infty(\omega_i\delta_{x_i}, \widetilde{\mu}_i) \leq \varepsilon$, $((\mathbb{1}_{\overline{B}(x_1,\varepsilon)}, \mathbb{1}_{\overline{B}(x_2,\varepsilon)}, \mathbb{1}_{\overline{B}(x_3,\varepsilon)}), (\widetilde{\mu}_1, \widetilde{\mu}_2, \widetilde{\mu}_3))$ is a saddle point for (1.1). This is straightforward, since $\text{spt}(\widetilde{\mu}_i) \cap \text{spt}(\widetilde{\mu}_j) = \emptyset$, and thus it can be deduced that $(\mathbb{1}_{B(x_1,\varepsilon)}, \mathbb{1}_{B(x_2,\varepsilon)}, \mathbb{1}_{B(x_3,\varepsilon)})$ is a dominant strategy for the learner. It is easy to check that $B^*_\mu = 1$ in this case.

**Case 2.** There is some $\overline{x}$ such that $d(\overline{x}, x_i) \leq \varepsilon$ for all $1 \leq i \leq 3$. We claim that $((1,0,0), (\omega_1\delta_{\overline{x}}, \omega_2\delta_{\overline{x}}, \omega_3\delta_{\overline{x}}))$ is a saddle point. First, $\omega_i\delta_{\overline{x}}$ is feasible for all $1 \leq i \leq 3$, since $\overline{x} \in \overline{B}(x_i,\varepsilon)$ for all $i$. Now, given $(\omega_1\delta_{\overline{x}}, \omega_2\delta_{\overline{x}}, \omega_3\delta_{\overline{x}})$, the best strategy for the learner is to choose class 1 deterministically for all points, since $\omega_1 \geq \omega_2 \geq \omega_3$. On the other hand, given $(1,0,0)$, any adversarial attack yields the same classification power. From this we conclude that $((1,0,0), (\omega_1\delta_{\overline{x}}, \omega_2\delta_{\overline{x}}, \omega_3\delta_{\overline{x}}))$ is indeed a saddle point. Notice that $B^*_\mu = \omega_1$ in this case.

**Case 3.** Two points are close to each other while the other point is far from them. For simplicity, we only consider the case $d(x_1, x_2) \le 2\varepsilon$, $d(x_1, x_3) > 2\varepsilon$ and $d(x_2, x_3) > 2\varepsilon$. The other cases are treated similarly. Let $\overline{x}_{12} = \frac{x_1+x_2}{2}$, and define $\widehat{f} = (\mathbb{1}_{\overline{B}(x_1,\varepsilon)\cup\overline{B}(x_2,\varepsilon)}, 0, \mathbb{1}_{\overline{B}(x_3,\varepsilon)})$ and $\widehat{\mu} = (\omega_1\delta_{\overline{x}_{12}}, \omega_2\delta_{\overline{x}_{12}}, \widetilde{\mu}_3)$ for arbitrary $\widetilde{\mu}_3$ with $W_\infty(\widetilde{\mu}_3, \omega_3\delta_{x_3}) \le \varepsilon$. We claim that $(\widehat{f}, \widehat{\mu})$ is a saddle point. For any $(f_1, f_2, f_3) \in \mathcal{F}$ we have

$$\begin{aligned} B_\mu(f, \widehat{\mu}) &= \int_{\mathcal{X}} f_1(x)\omega_1\delta_{\overline{x}_{12}}(x) + \int_{\mathcal{X}} f_2(x)\omega_2\delta_{\overline{x}_{12}}(x) + \int_{\mathcal{X}} f_3(x)d\widetilde{\mu}_3(x) \\ &= \omega_1 f_1(\overline{x}_{12}) + \omega_2 f_2(\overline{x}_{12}) + \int_{\mathcal{X}} f_3(x)\widetilde{\mu}_3(x) \\ &\le \omega_1 + \omega_3 \\ &= \int_{\mathcal{X}} \mathbb{1}_{\overline{B}(x_1,\varepsilon)\cup\overline{B}(x_2,\varepsilon)}\omega_1\delta_{\overline{x}_{12}}(x) + \int_{\mathcal{X}} 0\,\omega_2\delta_{\overline{x}_{12}}(x) + \int_{\mathcal{X}} \mathbb{1}_{\overline{B}(x_3,\varepsilon)}d\widetilde{\mu}_3(x). \end{aligned}$$

On the other hand, given $(\mathbb{1}_{\overline{B}(x_1,\varepsilon)\cup\overline{B}(x_2,\varepsilon)}, 0, \mathbb{1}_{\overline{B}(x_3,\varepsilon)})$, for any $(\widetilde{\mu}_1, \widetilde{\mu}_2, \widetilde{\mu}_3)$,

$$\begin{aligned} B_\mu(\widehat{f}, \widetilde{\mu}) &= \int_{\mathcal{X}} \mathbb{1}_{\overline{B}(x_1,\varepsilon)\cup\overline{B}(x_2,\varepsilon)}d\widetilde{\mu}_1(x) + \int_{\mathcal{X}} 0\,d\widetilde{\mu}_2(x) + \int_{\mathcal{X}} \mathbb{1}_{\overline{B}(x_3,\varepsilon)}d\widetilde{\mu}_3(x) \\ &= \omega_1 + \omega_3 \\ &= B_\mu(\widehat{f}, \widehat{\mu}) \end{aligned}$$

where the second equality follows from the assumption on the configuration of points. The above computations imply the claim. In this case $B_\mu^* = \omega_1 + \omega_3$.

**Case 4.** $d(x_i, x_j) \le 2\varepsilon$ for any $x_i, x_j$ but $\overline{B}(x_1,\varepsilon) \cap \overline{B}(x_2,\varepsilon) \cap \overline{B}(x_3,\varepsilon) = \emptyset$. Note that when $K = 2$, $d(x_1, x_2) \le 2\varepsilon$ if and only if $\overline{B}(x_1,\varepsilon) \cap \overline{B}(x_2,\varepsilon) \neq \emptyset$. However, when $K \ge 3$, these cases are not equivalent anymore. There are two subcases to consider depending on the magnitudes of the weights $(\omega_1, \omega_2, \omega_3)$.

**Case 4 - (i)** $\omega_1 < \omega_2 + \omega_3$. In this case, we can find some $\lambda_i \in [0, \omega_i]$ for all $1 \le i \le 3$ such that

$$\lambda_1 = \omega_2 - \lambda_2, \;\; \lambda_2 = \omega_3 - \lambda_3 \text{ and } \lambda_3 = \omega_1 - \lambda_1.$$

Precisely,

$$\lambda_1 = \frac{\omega_1 + \omega_2 - \omega_3}{2}, \;\; \lambda_2 = \frac{\omega_2 + \omega_3 - \omega_1}{2}, \text{ and } \lambda_3 = \frac{\omega_3 + \omega_1 - \omega_2}{2}.$$

Note that for all $i$, $\lambda_i \ge 0$ since $\omega_1 \le \omega_2 + \omega_3$. Let $\overline{x}_{12} \in \overline{B}(x_1,\varepsilon) \cap \overline{B}(x_2,\varepsilon)$, $\overline{x}_{13} \in \overline{B}(x_1,\varepsilon) \cap \overline{B}(x_3,\varepsilon)$ and $\overline{x}_{23} \in \overline{B}(x_2,\varepsilon) \cap \overline{B}(x_3,\varepsilon)$. Construct the following measures

$$\begin{aligned} \widehat{\mu}_1 &:= \big(\lambda_1\delta_{\overline{x}_{12}} + (\omega_1 - \lambda_1)\delta_{\overline{x}_{13}}\big) = \big((\frac{\omega_1 + \omega_2 - \omega_3}{2})\delta_{\overline{x}_{12}} + (\frac{\omega_1 - \omega_2 + \omega_3}{2})\delta_{\overline{x}_{13}}\big), \\ \widehat{\mu}_2 &:= \big(\lambda_2\delta_{\overline{x}_{23}} + (\omega_2 - \lambda_2)\delta_{\overline{x}_{12}}\big) = \big((\frac{\omega_2 + \omega_3 - \omega_1}{2})\delta_{\overline{x}_{23}} + (\frac{\omega_2 - \omega_3 + \omega_1}{2})\delta_{\overline{x}_{12}}\big), \\ \widehat{\mu}_3 &:= \big(\lambda_3\delta_{\overline{x}_{13}} + (\omega_3 - \lambda_3)\delta_{\overline{x}_{23}}\big) = \big((\frac{\omega_3 + \omega_1 - \omega_2}{2})\delta_{\overline{x}_{13}} + (\frac{\omega_3 - \omega_1 + \omega_2}{2})\delta_{\overline{x}_{23}}\big). \end{aligned}$$

Let $A_{ij} = A_{ji} := \overline{B}(x_i,\varepsilon) \cap \overline{B}(x_j,\varepsilon)$ and $A_i = \overline{B}(x_i,\varepsilon) \setminus (A_{ij} \cup A_{ik})$. One can observe that since $d(x_i, x_j) \le 2\varepsilon$ for any $x_i, x_j$ but $\overline{B}(x_1,\varepsilon) \cap \overline{B}(x_2,\varepsilon) \cap \overline{B}(x_3,\varepsilon) = \emptyset$, for each $i$

$$\overline{B}(x_i,\varepsilon) = A_{ij} \dot\cup A_{ik} \dot\cup A_i.$$

Here $\dot\cup$ denotes a disjoint union. Also, since $W_\infty(\widetilde{\mu}_i, \omega_i\delta_{x_i}) \le \varepsilon$, it must be the case that $A_{ij} \cap \mathrm{spt}(\widetilde{\mu}_k) = \emptyset$ if $k \neq i, j$. For each $1 \le i \le 3$, construct the following weak partition: for given

$(\widetilde{\mu}_1, \widetilde{\mu}_2, \widetilde{\mu}_3)$

$$\widetilde{f}_i(x) := \begin{cases} 1 & \text{if } x \in A_i, \\ \frac{1}{2} & \text{if } x \in A_{ij} \text{ and } \widetilde{\mu}_i(A_{ij}) = \widetilde{\mu}_j(A_{ij}), \\ 1 & \text{if } x \in A_{ij} \text{ and } \widetilde{\mu}_i(A_{ij}) > \widetilde{\mu}_j(A_{ij}), \\ 0 & \text{if } x \in A_{ij} \text{ and } \widetilde{\mu}_i(A_{ij}) < \widetilde{\mu}_j(A_{ij}) \text{ or } x \notin \overline{B}(x_i, \varepsilon). \end{cases}$$

Notice that the above $\widetilde{f}$ is a weak partition since $\overline{B}(x_1, \varepsilon) \cap \overline{B}(x_2, \varepsilon) \cap \overline{B}(x_3, \varepsilon) = \emptyset$. Let $\widehat{f}$ denote the above $\widetilde{f}$ when $\widehat{\mu}$ is played. Then, by the construction of $\widehat{\mu}$, $\widehat{f}_i(\overline{x}_{ij}) = \frac{1}{2}$ for all $i, j$. We claim that $(\widehat{f}, \widehat{\mu})$ is a saddle point.

A direct computation shows

$$B_\mu(\widehat{f}, \widehat{\mu}) = \int_{\mathcal{X}} \widehat{f}_1(x) d\widehat{\mu}_1(x) + \int_{\mathcal{X}} \widehat{f}_2(x) d\widehat{\mu}_2(x) + \int_{\mathcal{X}} \widehat{f}_3(x) d\widehat{\mu}_3(x) = \frac{1}{2}.$$

Given $(\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\mu}_3)$, for any $(f_1, f_2, f_3) \in \mathcal{F}$,

$$\begin{aligned} B_\mu(f, \widehat{\mu}) &= \int_{\mathcal{X}} f_1(x) d\widehat{\mu}_1(x) + \int_{\mathcal{X}} f_2(x) d\widehat{\mu}_2(x) + \int_{\mathcal{X}} f_3(x) d\widehat{\mu}_3(x) \\ &= (\frac{\omega_1 + \omega_2 - \omega_3}{2}) f_1(\overline{x}_{12}) + (\frac{\omega_1 + \omega_3 - \omega_2}{2}) f_1(\overline{x}_{13}) + (\frac{\omega_2 + \omega_3 - \omega_1}{2}) f_2(\overline{x}_{23}) \\ &\quad + (\frac{\omega_1 + \omega_2 - \omega_3}{2}) f_2(\overline{x}_{12}) + (\frac{\omega_1 + \omega_3 - \omega_2}{2}) f_3(\overline{x}_{13}) + (\frac{\omega_2 + \omega_3 - \omega_1}{2}) f_3(\overline{x}_{23}) \\ &= (\frac{\omega_1 + \omega_2 - \omega_3}{2})(f_1(\overline{x}_{12}) + f_2(\overline{x}_{12})) + (\frac{\omega_1 + \omega_3 - \omega_2}{2})(f_1(\overline{x}_{13}) + f_3(\overline{x}_{13})) \\ &\quad + (\frac{\omega_2 + \omega_3 - \omega_1}{2})(f_2(\overline{x}_{23}) + f_3(\overline{x}_{23})) \\ &\le (\frac{\omega_1 + \omega_2 - \omega_3}{2}) + (\frac{\omega_1 + \omega_3 - \omega_2}{2}) + (\frac{\omega_2 + \omega_3 - \omega_1}{2}) \\ &= \frac{1}{2}, \end{aligned}$$

where the second to last inequality follows from the fact that $\sum f_i(x) \le 1$ and the last equality follows from the fact that $\sum \omega_i = 1$. Given $(\widetilde{f}_1, \widetilde{f}_2, \widetilde{f}_3)$, on the other hand, for any $(\widetilde{\mu}_1, \widetilde{\mu}_2, \widetilde{\mu}_3)$

$$\begin{aligned} B_\mu(\widetilde{f}, \widetilde{\mu}) &= \int_{\mathcal{X}} \widetilde{f}_1(x) d\widetilde{\mu}_1(x) + \int_{\mathcal{X}} \widetilde{f}_2(x) d\widetilde{\mu}_2(x) + \int_{\mathcal{X}} \widetilde{f}_3(x) d\widetilde{\mu}_3(x) \\ &= \mathbb{1}_{\widetilde{\mu}_1(A_{12}) = \widetilde{\mu}_2(A_{12})} \frac{\widetilde{\mu}_1(A_{12}) + \widetilde{\mu}_2(A_{12})}{2} + \mathbb{1}_{\widetilde{\mu}_1(A_{12}) \neq \widetilde{\mu}_2(A_{12})} \max\{\widetilde{\mu}_1(A_{12}), \widetilde{\mu}_2(A_{12})\} \\ &\quad + \mathbb{1}_{\widetilde{\mu}_1(A_{13}) = \widetilde{\mu}_3(A_{13})} \frac{\widetilde{\mu}_1(A_{13}) + \widetilde{\mu}_3(A_{13})}{2} + \mathbb{1}_{\widetilde{\mu}_1(A_{13}) \neq \widetilde{\mu}_3(A_{13})} \max\{\widetilde{\mu}_1(A_{13}), \widetilde{\mu}_3(A_{13})\} \\ &\quad + \mathbb{1}_{\widetilde{\mu}_2(A_{23}) = \widetilde{\mu}_3(A_{23})} \frac{\widetilde{\mu}_2(A_{23}) + \widetilde{\mu}_3(A_{23})}{2} + \mathbb{1}_{\widetilde{\mu}_2(A_{23}) \neq \widetilde{\mu}_3(A_{23})} \max\{\widetilde{\mu}_2(A_{23}), \widetilde{\mu}_3(A_{23})\} \\ &\quad + \widetilde{\mu}_1(A_1) + \widetilde{\mu}_2(A_2) + \widetilde{\mu}_3(A_3). \end{aligned}$$

To minimize $\widetilde{\mu}_1(A_1) + \widetilde{\mu}_2(A_2) + \widetilde{\mu}_3(A_3)$ in the above, the adversary should put $\text{spt}(\widetilde{\mu}_i) \subseteq A_{ij} \cup A_{ik}$ for all $i$. Also, at the minimum, we should have $\widetilde{\mu}_i(A_{ij}) = \widetilde{\mu}_j(A_{ij})$, for otherwise the adversary would be able decrease the classification power further. Putting together the above, we can deduce

$$\begin{aligned}
B_\mu(\widetilde{f},\widetilde{\mu}) &\geq \mathbb{1}_{\widetilde{\mu}_1(A_{12})=\widetilde{\mu}_2(A_{12})}\frac{\widetilde{\mu}_1(A_{12})+\widetilde{\mu}_2(A_{12})}{2}\\
&\quad+\mathbb{1}_{\widetilde{\mu}_1(A_{13})=\widetilde{\mu}_3(A_{13})}\frac{\widetilde{\mu}_1(A_{13})+\widetilde{\mu}_3(A_{13})}{2}\\
&\quad+\mathbb{1}_{\widetilde{\mu}_2(A_{23})=\widetilde{\mu}_3(A_{23})}\frac{\widetilde{\mu}_2(A_{23})+\widetilde{\mu}_3(A_{23})}{2}\\
&=\frac{(\widetilde{\mu}_1+\widetilde{\mu}_2)(A_{12})}{2}+\frac{(\widetilde{\mu}_1+\widetilde{\mu}_3)(A_{13})}{2}+\frac{(\widetilde{\mu}_2+\widetilde{\mu}_3)(A_{23})}{2}\\
&=\frac{1}{2},
\end{aligned}$$

which verifies the claim. In this case, $B^*_\mu = \frac{1}{2}$.

In fact, it is unavoidable to introduce weak partitions $f \in \mathcal{F}$. Let $f = (\mathbb{1}_{F_1}, \mathbb{1}_{F_2}, \mathbb{1}_{F_3})$ be any strong partition, i.e. $F_1 \dot\cup F_2 \dot\cup F_3 = \cup \overline{B}(x_i, \varepsilon)$. We will show that for any $\widetilde{\mu}$, $(f, \widetilde{\mu})$ cannot be a saddle point. Assume that $\overline{B}(x_1,\varepsilon) \subseteq F_1$. Since $d(x_1,x_2) \leq 2\varepsilon$ and $d(x_1,x_3) \leq 2\varepsilon$, it must be the case that $F_1 \cap \overline{B}(x_2,\varepsilon) \neq \emptyset$ and $F_1 \cap \overline{B}(x_3,\varepsilon) \neq \emptyset$. These facts yield that optimal $\widetilde{\mu}_2$ and $\widetilde{\mu}_3$ for the adversary must satisfy $\mathrm{spt}(\widetilde{\mu}_2) \subseteq F_1 \cap \overline{B}(x_2,\varepsilon)$ and $\mathrm{spt}(\widetilde{\mu}_3) \subseteq F_1 \cap \overline{B}(x_3,\varepsilon)$. This configuration gives a classifying power $\omega_1$ since the learner can only detect class 1 perfectly and always misclassify others.

However, given any such $(\widetilde{\mu}_1, \widetilde{\mu}_2, \widetilde{\mu}_3)$, the learner has an incentive to modify a classifying rule. Let $F_1' := F_1 \setminus (\mathrm{spt}(\widetilde{\mu}_2) \cup \mathrm{spt}(\widetilde{\mu}_3))$, $F_2' := F_2 \cup \mathrm{spt}(\widetilde{\mu}_2)$ and $F_3' := F_3 \cup \mathrm{spt}(\widetilde{\mu}_3)$. Then, this classifying rule perfectly classifies. Precisely, there exists a deviation for the learner, $f' = (\mathbb{1}_{F_1'}, \mathbb{1}_{F_2'}, \mathbb{1}_{F_3'})$, such that

$$1 = B(f', \widetilde{\mu}) > B(f, \widetilde{\mu}) = \omega_1.$$

Assume that $\overline{B}(x_1,\varepsilon) \not\subseteq F_1$. Since $(F_1, F_2, F_3)$ is a partition, it must be the case that either $F_2 \cap \overline{B}(x_1,\varepsilon) \neq \emptyset$ or $F_3 \cap \overline{B}(x_1,\varepsilon) \neq \emptyset$. Without loss of generality, assume the former case only. The other cases are analogous. $F_2 \cap \overline{B}(x_1,\varepsilon) \neq \emptyset$ yields that an optimal $\widetilde{\mu}_1$ for the adversary must satisfy $\mathrm{spt}(\widetilde{\mu}_1) \subseteq F_2$. Then, a corresponding classifying power is at most $\omega_2 + \omega_3$ since the learner always misclassifies class 1.

However, given any such $(\widetilde{\mu}_1, \widetilde{\mu}_2, \widetilde{\mu}_3)$, the learner has an incentive to modify a classifying rule again. Let $F_1' := F_1 \cup \mathrm{spt}(\widetilde{\mu}_1)$, $F_2' := F_2 \setminus \mathrm{spt}(\widetilde{\mu}_1)$ and $F_3' := F_3$. Similar as above, letting $f' = (\mathbb{1}_{F_1'}, \mathbb{1}_{F_2'}, \mathbb{1}_{F_3'})$, such that

$$1 = B(f', \widetilde{\mu}) > \omega_2 + \omega_3 \geq B(f, \widetilde{\mu}).$$

Therefore, any strong partition $f = (\mathbb{1}_{F_1}, \mathbb{1}_{F_2}, \mathbb{1}_{F_3})$ cannot sustain a saddle point in this case.

We want to emphasize that the same reasoning still holds for other cases. In other words, even this simple discrete measures, it is necessary to extend strong partition to weak partition in order to achieve the minimax value.

**Case 4 - (ii)** $\omega_1 \geq \omega_2 + \omega_3$. In this case, no matter how the adversary perturbs the distribution, there will always be an excess mass from class 1 that won't be matched to other classes. Motivated by this observation, let $\kappa = \omega_1 - (\omega_2 + \omega_3) \geq 0$ and consider the following measures $(\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\mu}_3)$:

$$\begin{aligned}
\widehat{\mu}_1 &= \omega_2 \delta_{\overline{x}_{12}} + \omega_3 \delta_{\overline{x}_{13}} + \kappa \delta_{x_1},\\
\widehat{\mu}_2 &= \omega_2 \delta_{\overline{x}_{12}},\\
\widehat{\mu}_3 &= \omega_3 \delta_{\overline{x}_{13}}.
\end{aligned}$$

Consider $(\widehat{f}_1, \widehat{f}_2, \widehat{f}_3) = (1, 0, 0)$. We claim that $(\widehat{f}, \widehat{\mu}) = ((\widehat{f}_1, \widehat{f}_2, \widehat{f}_3), (\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\mu}_3))$ is a saddle point. Indeed, a direct computation shows

$$B_\mu(\widehat{f}, \widehat{\mu}) = \int_{\mathcal{X}} \widehat{f}_1(x) d\widehat{\mu}_1(x) + \int_{\mathcal{X}} \widehat{f}_2(x) d\widehat{\mu}_2(x) + \int_{\mathcal{X}} \widehat{f}_3(x) d\widehat{\mu}_3(x) = \omega_1.$$

For any $(f_1, f_2, f_3) \in \mathcal{F}$,

$$\begin{aligned} B_\mu(f, \widehat{\mu}) &= \int_{\mathcal{X}} f_1(x) d\widehat{\mu}_1(x) + \int_{\mathcal{X}} f_2(x) d\widehat{\mu}_2(x) + \int_{\mathcal{X}} f_3(x) d\widehat{\mu}_3(x) \\ &= \omega_2 f_1(\overline{x}_{12}) + \omega_3 f_1(\overline{x}_{13}) + \kappa f_1(x_1) + \omega_2 f_2(\overline{x}_{12}) + \omega_3 f_3(\overline{x}_{13}) \\ &= \omega_2(f_1(\overline{x}_{12}) + f_2(\overline{x}_{12})) + \omega_3(f_1(\overline{x}_{13}) + f_3(\overline{x}_{13})) + \kappa f_1(x_1) \\ &\leq \omega_2 + \omega_3 + \kappa \\ &= \omega_1. \end{aligned}$$

On the other hand, for any feasible $(\widetilde{\mu}_1, \widetilde{\mu}_2, \widetilde{\mu}_3)$,

$$B_\mu(\widehat{f}, \widetilde{\mu}) = \int_{\mathcal{X}} \widehat{f}_1(x) d\widetilde{\mu}_1(x) + \int_{\mathcal{X}} \widehat{f}_2(x) d\widetilde{\mu}_2(x) \int_{\mathcal{X}} \widehat{f}_3(x) d\widetilde{\mu}_3(x) = \omega_1.$$

The claim follows. In this case, $B_\mu^* = \omega_1$. Note that here $\omega_1 \geq \frac{1}{2}$, since $\omega_1 \geq \omega_2 + \omega_3$ and $\sum \omega_i = 1$. In case we had $\omega_1 = \omega_2 + \omega_3$, we would actually have $\omega_1 = \frac{1}{2}$ and thus **Case 4 -(i)** and **Case 4 -(ii)** provide consistent outcomes.

We now show that the adversary has no incentive to use the point $\overline{x}_{23}$, in contrast to what happens in **Case 4 -(i)**. Fix a small $\eta > 0$, and suppose that the adversary moves $\eta$ mass from each of $\omega_2 \delta_{x_2}$ and $\omega_3 \delta_{x_3}$ to the point $\overline{x}_{23}$, respectively. Construct corresponding measures:

$$\begin{aligned} \widetilde{\mu}_1 &= (\omega_2 - \eta)\delta_{\overline{x}_{12}} + (\omega_3 - \eta)\delta_{\overline{x}_{13}} + \kappa' \delta_{x_1}, \\ \widetilde{\mu}_2 &= \eta \delta_{\overline{x}_{23}} + (\omega_2 - \eta)\delta_{\overline{x}_{12}}, \\ \widetilde{\mu}_3 &= (\omega_3 - \eta)\delta_{\overline{x}_{13}} + \eta \delta_{\overline{x}_{23}} \end{aligned}$$

where $\kappa' = \omega_1 - (\omega_2 + \omega_3 - 2\eta) = \kappa + 2\eta$. We show that $\widetilde{\mu}$ can not be a solution to the adversarial problem by showing that the learner can select a strategy $\widetilde{f}$ for which

$$B_\mu(\widetilde{f}, \widetilde{\mu}) > \omega_1.$$

Indeed, we can select $\widetilde{f} := (\mathbb{1}_{\overline{B}(x_1,\varepsilon)}, 0, \mathbb{1}_{\mathcal{X} \setminus \overline{B}(x_1,\varepsilon)})$. It follows that

$$\begin{aligned} B_\mu(\widetilde{f}, \widetilde{\mu}) &= \int_{\mathcal{X}} \widetilde{f}_1(x) d\widetilde{\mu}_1(x) + \int_{\mathcal{X}} \widetilde{f}_2(x) d\widetilde{\mu}_2(x) + \int_{\mathcal{X}} \widetilde{f}_3(x) d\widetilde{\mu}_3(x) \\ &= (\omega_2 - \eta) + (\omega_3 - \eta) + \kappa' + \eta = \omega_1 + \ \eta > \omega_1. \end{aligned}$$

Notice that while the geometry of points $x_1, x_2, x_3$ in **case 4 -(i)** and **case 4 -(ii)** is the same, the geometries of the corresponding optimal adversarial attacks are determined by the full distribution $\mu$ and not just by the geometry of its support. In fact, the optimal adversarial attacks $\widetilde{\mu}$ and the optimal barycenter measure $\lambda$ depend on not only the geometry of the support of $\mu$ but also the magnitudes of its marginals, $\mu_i$'s.

5.3. **Numerical Experiments.** In this section we illustrate our theoretical results numerically. We obtain robust classifiers for synthetic data sets and compute optimal adversarial risks for two popular real data sets: MNIST and CIFAR.

From the numerics perspective, our aim is to solve the MOT problem (1.2) and its dual for an empirical measure $\mu$ whose support consists of $n$ data points. We use Sinkhorn algorithm for concreteness. Introduced in [Cut13], Sinkhorn algorithm has been one of the central algorithmic tools in computational optimal transport in the past decade. This algorithm, originally

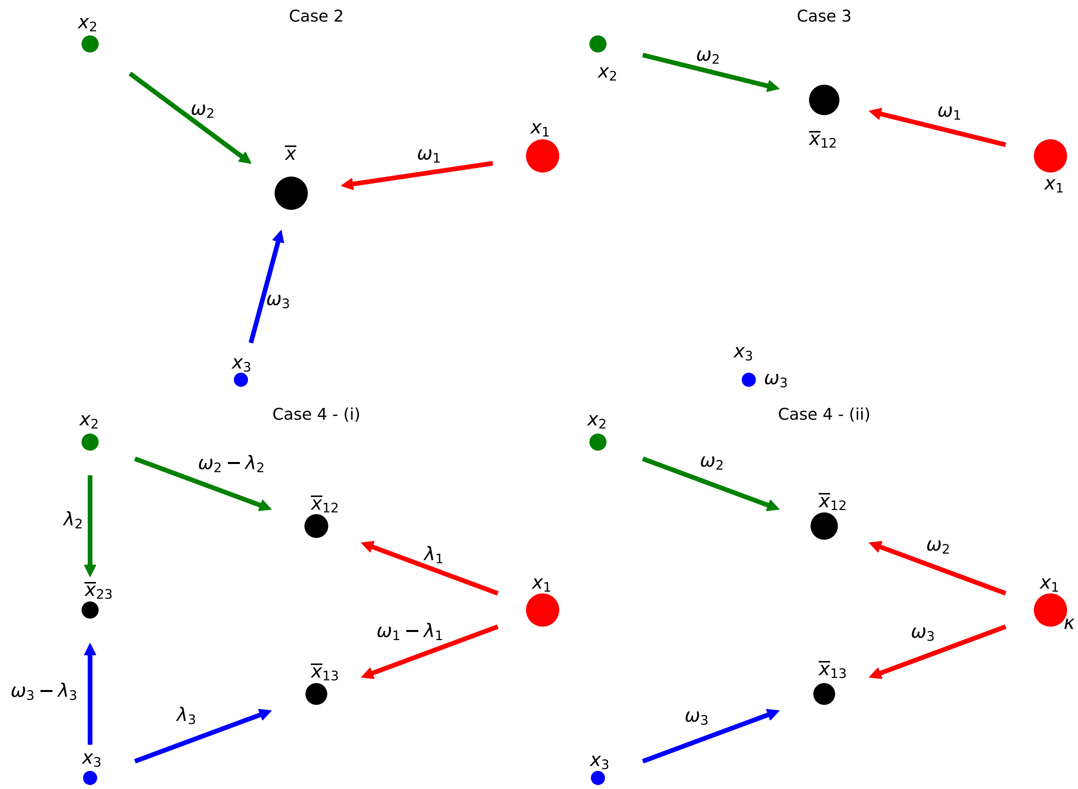


FIGURE 1. Illustrations of the adversarial attacks in all cases from section 5.2. Weights on arrows indicate the amount of mass the adversary moves to a perturbed point. $\overline{x}$'s are the support of $\lambda$ in (2.8). Notice that the support of $\lambda$ depends on both the geometry of data distributions and their magnitudes.

introduced in the context of standard (2-marginal) optimal transport problems, was extended to MOTs in [BCC+15, BCN19]. Works that study the computational complexity of generic MOT problems include: [LHCJ19, TDGU20, HRCK21, Car21]. In particular, the authors of [LHCJ19] prove the complexity of MOT Sinkhorn algorithm to be $\widetilde{O}(K^3 n^K \epsilon^{-2})$, where $\epsilon$ denotes the error tolerance.

In our first illustration, we consider a data set $(x_1, y_1), \dots (x_n, y_n)$ in $\mathbb{R}^2 \times \{1, 2, 3\}$ obtained by sampling $y_i$ uniformly from $\{1, 2, 3\}$ and then $x_i$ from a certain Gaussian distribution with parameters depending on the outcome of $y_i$. We consider the cost $c = c_\varepsilon$ from **Example** 2.3 with $d$ the Euclidean distance in $\mathbb{R}^2$ and different values of $\varepsilon$. In Figure 3 we show the labels assigned to the data by the (approximate) robust classifier, which we computed using **Corollary** 4.7 for the dual potentials $\phi_j$ generated by the MOT Sinkhorn algorithm.

In our second illustration, we use the mutli-marginal version of Sinkhorn algorithm to compute the adversarial risk $R^*_\mu$ (i.e. the optimal value of (1.1)) for $\mu$ an empirical measure supported on a subset of either the CIFAR or MNIST data sets. In both cases we consider samples belonging to one of four possible classes in order to decrease the computational complexity of the problem. We use the cost $c$ from **Example** 2.3 for different values of $\varepsilon$ and two choices of $d$: the Euclidean distance $\ell^2$ and the $\ell^\infty$ distance. The results are shown in Figure 2. We can observe that for the CIFAR data set the two distance functions behave similarly: while not the same, the plots exhibit a similar qualitative behavior. For the MNIST data set, on the other hand, the situation is markedly different: in contrast to the plot for the $\ell^2$ distance, the adversarial risk

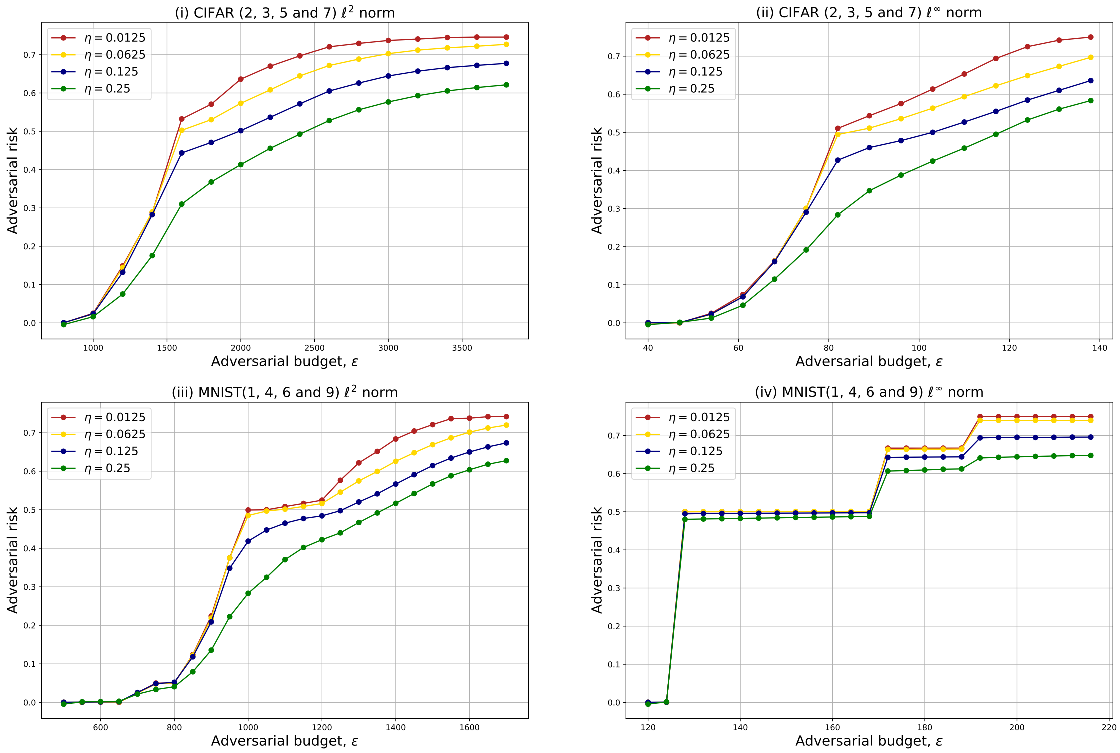


FIGURE 2. Adversarial risks (1.1) computed using the multimarginal Sinkhorn algorithm. $\eta$ is the entropic regularization parameter of the Sinkhorn algorithm. The maximum adversarial risk in all cases is 0.75 because we consider 4 classes and an equal number of points in each class. Due to the entropic penalty, the multimarginal Sinkhorn algorithm always gives an upper bound for the optimal classification power $B_\mu^*$, hence gives a lower bound for the adversarial risk $R_\mu^*$.

with $\ell^\infty$ distance varies dramatically as $\varepsilon$ grows. This observation is consistent with the findings in [PJ21a] for the binary case.

We emphasize that Figure 2 only provides approximations of the true adversarial risk $R_\mu^*$. Indeed, recall that $R_\mu^* = 1 - B_\mu^*$. Approximating $B_\mu^*$ using the MOT Sinkhorn algorithm will always produce an upper bound for $B_\mu^*$ since the regularization term effectively restricts the solution space of (1.2). Thus, the multimarginal Sinkhorn algorithm always yields a lower bound for the true $R_\mu^*$. Of course, one can always compute a tighter lower bound by reducing the regularization parameter $\eta$ at the expense of increasing the computational burden.

As way of conclusion for this section we provide pointers to the literature discussing the computational complexity of the Wasserstein barycenter problem; Wasserstein barycenter problems are specific instances in the MOT family. On the one hand, [ABA22] proves certain computational hardness of the barycenter problem in the dimension of the feature space (here $\mathcal{X}$). On the other hand, [ABA21] presents an algorithm that can get an approximate solution of the optimal barycenter in polynomial time for a fixed dimension of the feature space. While our MOT is not the standard barycenter problem, it is still a generalized version thereof, and thus, it is reasonable to expect that the structure of our problem can be used in the design of algorithms that perform better than off-the-shelf MOT solvers. We leave this task for future work.

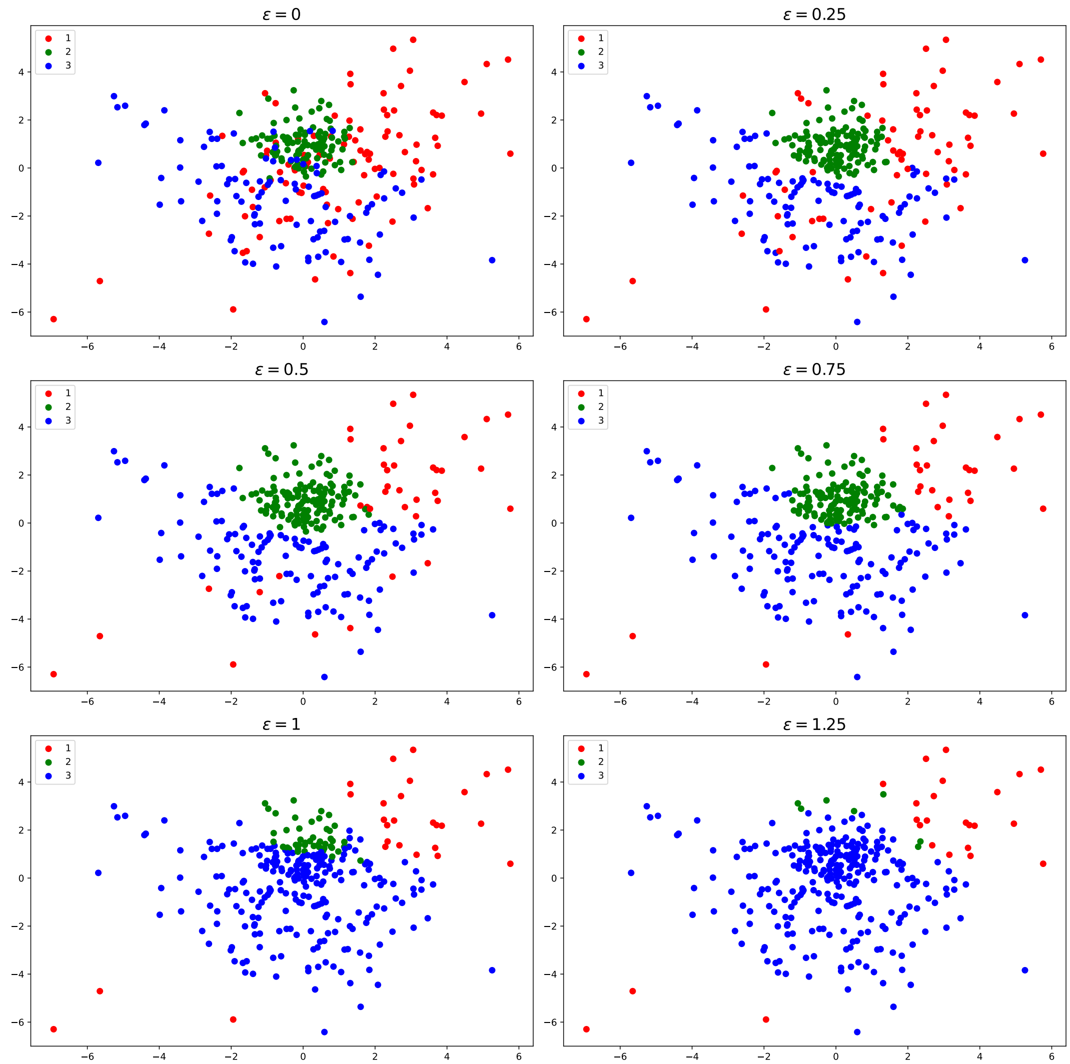


Figure 3. Three Gaussians in $\mathbb{R}^2$. One can observe that as $\varepsilon$ grows the robust classifying rule becomes simpler, as expected.

## 6. Conclusions and future directions

In this paper we have discussed a series of equivalent formulations of adversarial problems in the context of multiclass classification. These formulations take the form of problems in optimal transport, specifically, multimarginal optimal transport and (generalized) Wasserstein barycenter problems. Besides providing a novel connection between apparently unrelated fields, we have also discussed a series of theoretical and computational implications emanating from these equivalences. In what follows we briefly expand this discussion, while at the same time provide a few perspectives on future work.

First, it is of interest to design scalable algorithms for solving the MOT problem (1.2). In general, MOT problems are not scalable in the number of marginals of the problem. However, this may not necessarily be an issue for our MOT problem, since it possesses a special structure that, as we discussed throughout section 3, allows us to interpret the desired MOT problem as a generalized barycenter problem; barycenter problems, at least in their standard version, are known to scale much better than general MOT problems. Tailor-specific algorithms for our MOT problem can take advantage of the favourable geometric structure of a given data set. Indeed, if a data set is such that there is only a small number of classes (much smaller than $K$) that interact with each other at the scale implicitly specified by the cost $c$ (think of **Example** 2.3), then the effective size of problem 1.2 will be considerably smaller than the size of the worst case setting —see the reformulation (3.6).

Second, it would be of interest to use (1.1) to help in the training of robust classifiers within specific families of models. Notice that (1.1) is model free from the perspective of the learner, but in applications practitioners may be interested in solving a problem like:

$$\inf_{f \in \widetilde{\mathcal{F}}} \sup_{\widetilde{\mu} \in \mathcal{P}(\mathcal{Z})} \left\{ R(f, \widetilde{\mu}) - C(\mu, \widetilde{\mu}) \right\},$$

which differs from (1.1) in the family of classifiers $\widetilde{\mathcal{F}}$, which may be strictly smaller than $\mathcal{F}$; for example, $\widetilde{\mathcal{F}}$ may be a family of neural networks, kernel-based classifiers, or other popular models. There are two ways in which problem (1.1) is still meaningful for the above model-specific problem: 1) the computable optimal $\widetilde{\mu}^*$ from problem (1.1) can be used as a way to generate adversarial examples that could be used during training of the desired model; 2) the optimal value of (1.1) can serve as a benchmark for robust training within *any* family of models.

Finally, it is of interest to investigate the geometric content that profiles like the ones presented in Figure 2 carry about a specific data set. As illustrated in Figure 2, these curves are specific signatures (adversarial signatures) of a given data distribution, and thus, they may be potentially used to capture similarities and discrepancies between different data sets.

The above are just but a few directions currently under investigation that emanate from this work.

## References


[ABA21] Jason M Altschuler and Enric Boix-Adsera, *Wasserstein barycenters can be computed in polynomial time in fixed dimension.*, J. Mach. Learn. Res. **22** (2021), 44–1.

[ABA22] Jason M Altschuler and Enric Boix-Adserà, *Wasserstein barycenters are np-hard to compute*, SIAM Journal on Mathematics of Data Science **4** (2022), no. 1, 179–203.

[AC11] Martial Agueh and Guillaume Carlier, *Barycenters in the Wasserstein space*, SIAM J. Math. Anal. **43** (2011), no. 2, 904–924. MR 2801182

[BCC+15] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré, *Iterative bregman projections for regularized transportation problems*, SIAM Journal on Scientific Computing **37** (2015), no. 2, A1111–A1138.

[BCM19] A. Bhagoji, Daniel Cullina, and Prateek Mittal, *Lower bounds on adversarial robustness from optimal transport*, NeurIPS, 2019.

[BCN19] Jean-David Benamou, Guillaume Carlier, and Luca Nenna, *Generalized incompressible flows, multi-marginal transport and sinkhorn algorithm*, Numerische Mathematik **142** (2019), no. 1, 33–54.

[BDPGG12] Giuseppe Buttazzo, Luigi De Pascale, and Paola Gori-Giorgi, *Optimal-transport formulation of electronic density-functional theory*, Physical Review A **85** (2012), no. 6, 062502.

[BGB+20] Joey Bose, Gauthier Gidel, Hugo Berard, Andre Cianflone, Pascal Vincent, Simon Lacoste-Julien, and Will Hamilton, *Adversarial example games*, Advances in neural information processing systems **33** (2020), 8921–8934.

[BKM19] Jose Blanchet, Yang Kang, and Karthyek Murthy, *Robust Wasserstein profile inference and applications to machine learning*, J. Appl. Probab. **56** (2019), no. 3, 830–857. MR 4015639

[BM19] Jose Blanchet and Karthyek Murthy, *Quantifying distributional model risk via optimal transport*, Math. Oper. Res. **44** (2019), no. 2, 565–600. MR 3959085

[BvLNS21] Florian Beier, Johannes von Lindheim, Sebastian Neumayer, and Gabriele Steidl, *Unbalanced multi-marginal optimal transport*, arXiv preprint arXiv:2103.10854 (2021).

[Car21] Guillaume Carlier, *On the linear convergence of the multi-marginal sinkhorn algorithm.*

[CCK+18] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo, *Stargan: Unified generative adversarial networks for multi-domain image-to-image translation*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8789–8797.

[CD14] Marco Cuturi and Arnaud Doucet, *Fast computation of wasserstein barycenters*, International conference on machine learning, PMLR, 2014, pp. 685–693.

[CDPDM15] Maria Colombo, Luigi De Pascale, and Simone Di Marino, *Multimarginal optimal transport maps for one-dimensional repulsive costs*, Canad. J. Math. **67** (2015), no. 2, 350–368. MR 3314838

[CE10] G. Carlier and I. Ekeland, *Matching for teams*, Economic Theory **42** (2010), 397–418.

[CFK13] Codina Cotar, Gero Friesecke, and Claudia Klüppelberg, *Density functional theory and optimal transportation with Coulomb cost*, Comm. Pure Appl. Math. **66** (2013), no. 4, 548–599. MR 3020313

[CMN10] Pierre-André Chiappori, Robert J. McCann, and Lars P. Nesheim, *Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness*, Econom. Theory **42** (2010), no. 2, 317–354. MR 2564439

[CMP17] Pierre-André Chiappori, Robert J. McCann, and Brendan Pass, *Multi- to one-dimensional optimal transport*, Comm. Pure Appl. Math. **70** (2017), no. 12, 2405–2444. MR 3716857

[CMZ+19] Jiezhang Cao, Langyuan Mo, Yifan Zhang, Kui Jia, Chunhua Shen, and Mingkui Tan, *Multi-marginal wasserstein gan*, Advances in Neural Information Processing Systems **32** (2019), 1776–1786.

[COO15] Guillaume Carlier, Adam Oberman, and Edouard Oudet, *Numerical methods for matching for teams and Wasserstein barycenters*, ESAIM Math. Model. Numer. Anal. **49** (2015), no. 6, 1621–1642. MR 3423268

[Cut13] Marco Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, Advances in neural information processing systems **26** (2013).

[DD20] Julie Delon and Agnès Desolneux, *A wasserstein-type distance in the space of gaussian mixture models*, SIAM Journal on Imaging Sciences **13** (2020), no. 2, 936–970.

[DMG20] Simone Di Marino and Augusto Gerolin, *An optimal transport approach for the schrödinger bridge problem and convergence of sinkhorn algorithm*, Journal of Scientific Computing **85** (2020), no. 2, 1–28.

[Eke05] Ivar Ekeland, *An optimal matching problem*, ESAIM Control Optim. Calc. Var. **11** (2005), no. 1, 57–71. MR 2110613

[GM20] Nicolás García Trillos and Ryan W. Murray, *Adversarial classification: Necessary conditions and geometric flows*, ArXiv **abs/2011.10797** (2020).

[Gwc98] Wilfrid Gangbo and Andrzej wi ch, *Optimal maps for the multidimensional monge-kantorovich problem*, Communications on Pure and Applied Mathematics **51** (1998), 23–45.

[HRCK21] Isabel Haasler, Axel Ringh, Yongxin Chen, and Johan Karlsson, *Multimarginal optimal transport with a tree-structured cost and the schrodinger bridge problem*, SIAM Journal on Control and Optimization **59** (2021), no. 4, 2428–2453.

[KP13] Y. Kim and Young-Heon Brendan Pass, *Multi-marginal optimal transport on riemannian manifolds*, American Journal of Mathematics **137** (2013), 1045 – 1060.

[KP15] Jun Kitagawa and Brendan Pass, *The multi-marginal optimal partial transport problem*, Forum Math. Sigma **3** (2015), Paper No. e17, 28. MR 3482267

[LHCJ19] Tianyi Lin, Nhat Ho, Marco Cuturi, and Michael I Jordan, *On the complexity of approximating multimarginal optimal transport*, arXiv preprint arXiv:1910.00152 (2019).

[ML13] Christian B Mendl and Lin Lin, *Kantorovich dual solution for strictly correlated electrons in atoms and molecules*, Physical Review B **87** (2013), no. 12, 125106.

[MSP+21] Laurent Meunier, Meyer Scetbon, Rafael Pinot, Jamal Atif, and Yann Chevaleyre, *Mixed nash equilibria in the adversarial examples game*, ICML, 2021.

[Nak19] Preetum Nakkiran, *Adversarial robustness may be at odds with simplicity*, ArXiv **abs/1901.00532** (2019).

[Pas15] Brendan Pass, *Multi-marginal optimal transport: theory and applications*, ESAIM Math. Model. Numer. Anal. **49** (2015), no. 6, 1771–1790. MR 3423275

[PJ21a] Muni Sreenivas Pydi and Varun Jog, *Adversarial risk via optimal transport and optimal couplings*, IEEE Transactions on Information Theory **67** (2021), 6031–6052.

[PJ21b] ———, *The many faces of adversarial risk*, Thirty-Fifth Conference on Neural Information Processing Systems, 2021.

[SGGS07] Michael Seidl, Paola Gori-Giorgi, and Andreas Savin, *Strictly correlated electrons in density-functional theory: A general formulation with applications to spherical densities*, Physical Review A **75** (2007), no. 4, 042511.

[SLD18] Sanvesh Srivastava, Cheng Li, and David B Dunson, *Scalable bayes via barycenter in wasserstein space*, The Journal of Machine Learning Research **19** (2018), no. 1, 312–346.

[TDGU20] Nazarii Tupitsa, Pavel Dvurechensky, Alexander Gasnikov, and César A Uribe, *Multimarginal optimal transport by accelerated alternating minimization*, 2020 59th IEEE Conference on Decision and Control (CDC), IEEE, 2020, pp. 6132–6137.

[Vil03] Cédric Villani, *Topics in optimal transportation*, Graduate Studies in Mathematics, vol. 58, American Mathematical Society, Providence, RI, 2003. MR 1964483

Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, Wisconsin 53706, USA.
*Email address*: garciatrillo@wisc.edu

Department of Mathematics, Purdue University, Mathematical Sciences Bldg, 150 N University St, West Lafayette, IN 47907
*Email address*: jacob225@purdue.edu

Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, Wisconsin 53706, USA.
*Email address*: kim836@wisc.edu