Journal of Machine Learning Research 24 (2023) 1-56 Submitted 6/22; Published 1/23

The multimarginal optimal transport formulation of adversarial multiclass classification

Nicolás García Trillos GARCIATRILLO@WISC.EDU
Jakwang Kim KIM836@WISC.EDU
Department of Statistics
University of Wisconsin-Madison
1300 University Avenue, Madison, Wisconsin 53706, USA

Matt Jacobs JACOB225@PURDUE.EDU
Department of Mathematics
Purdue University
150 N University St, West Lafayette, Indiana 47907, USA

Editor: Sebastien Bubeck

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.

Keywords: Adversarial learning, Multiclass classification, Optimal transport, Multimarginal optimal transport, Wasserstein barycenter, Generalized barycenter problem

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:

$$\inf_{f \in \mathcal{F}} \sup_{\widetilde{\mu} \in \mathcal{P}(\mathcal{Z})} \left\{ R(f, \widetilde{\mu}) - C(\mu, \widetilde{\mu}) \right\}. \tag{1}$$

Here, $\mathcal{F}$ denotes the set of *all* probabilistic multiclass classifiers —see section 2; μ 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 μ) and a choice of loss function, which in this paper will be restricted to be the 0-1 loss. Problem (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 μ.

For a large family of functionals C in (1) we show that the adversarial problem (1) is equivalent to a *multimarginal optimal transport problem (MOT)* of the form:

$$\inf_{\pi \in \Pi_K(\mu)} \int \mathbf{c}(z_1, \dots, z_K) d\pi(z_1, \dots, z_K), \tag{2}$$

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) from solutions to the problem (2) and its dual, offering in this way new computational strategies for solving problem (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^*, μ^*) for (1) by running one such OT algorithm. The equivalence between (1) and (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: Bhagoji, Cullina, and Mittal (2019); Nakkiran (2019); Pydi and Jog (2021a,b); García Trillos and Murray (2022).

In order to establish the equivalence between (1) and (2), we develop another interesting equivalent reformulation of (1) that reveals a rich geometric structure of the original adversarial problem. This reformulation takes the form of a *generalized barycenter problem*

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

which is a novel variant of the Wasserstein barycenter problems introduced in Agueh and Carlier (2011); Carlier and Ekeland (2010). In the classical Wasserstein barycenter problem, given K *probability* measures $\varrho_1, \dots, \varrho_K$ 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 ϱ such that the summed cost of transporting each of the ϱ_i onto ϱ is as small as possible. In our generalized problem, we try to find a nonnegative measure λ (no longer necessarily a probability measure) such that the total mass of λ plus the summed cost of transporting each μ_i (not necessarily having the same total mass) onto *some part* of λ is as small as possible. Here transporting a μ_i onto some part of λ means we want to find a measure $\widetilde{\mu}_i \leq \lambda$ and transport μ_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 Agueh and Carlier (2011); Carlier and Ekeland (2010) 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 (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 Gangbo and Święch (1998), there have been numerous subsequent papers establishing geometric and analytic results (e.g., Kim and Pass (2013); Pass (2015); Kitagawa and Pass (2015); Chiappori, McCann, and Pass (2017)) for MOT problems. MOT problems have also been used extensively in applications. For example, they appear in the so-called density functional theory in physics Seidl, Gori-Giorgi, and Savin (2007); Buttazzo, De Pascale, and Gori-Giorgi (2012); Cotar, Friesecke, and Klüppelberg (2013); Mendl and Lin (2013); Colombo, De Pascale, and Di Marino (2015), and in economics Ekeland (2005); Chiappori, McCann, and Nesheim (2010); Carlier and Ekeland (2010). In the machine learning community, researchers have recently explored many interesting applications, including generative adversarial networks(GANs) Choi, Choi, Kim, Ha, Kim, and Choo (2018); Cao, Mo, Zhang, Jia, Shen, and Tan (2019) and Wasserstein Barycenters Agueh and Carlier (2011); Cuturi and Doucet (2014); Benamou, Carlier, Cuturi, Nenna, and Peyré (2015); Carlier, Oberman, and Oudet (2015); Srivastava, Li, and Dunson (2018); Delon and Desolneux (2020), where MOTs are used. Recent works like Di Marino and Gerolin (2020); Haasler, Ringh, Chen, and Karlsson (2021) develop a connection between the Schrödinger bridge problem and MOT. MOT problems have been extended to the unbalanced setting —see Beier, von Lindheim, Neumayer, and Steidl (2021).

1.1 Outline of paper

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 the generalized Wasserstein barycenter problem, which can be interpreted as dual of the original adversarial problem (1), and define in detail the MOT problem (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 we discuss the equivalence between (1) and (2) through the duality results in earlier sections. 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. Finally, we wrap-up the paper in section 6 by presenting some conclusions and discussing 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 learning problem and $[K]$ as the set of classes or labels.

Let μ be a finite positive Borel measure (not necessarily a probability measure) over $\mathcal{Z}$. Given $i \in [K]$, we use μ_i to represent the positive measure over $\mathcal{X}$ defined as

$$\mu_i(A) := \mu\left(A \times \{i\}\right), \tag{3}$$

for all measurable subsets A of $\mathcal{X}$. In the sequel, we use μ 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.

Through this paper, the cost function in (1) will take 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 μ and second marginal equal to $\widetilde{\mu}$.

Assumption 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 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 $c_{\mathcal{Z}}$ described in Assumption 1 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 μ and $\widetilde{\mu}$ into measures $\mu_i, \widetilde{\mu}_i$ as in (3), it is possible to write $C(\mu, \widetilde{\mu})$ as

$$C(\mu, \widetilde{\mu}) = \sum_{i \in [K]} C(\mu_i, \widetilde{\mu}_i),$$

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

$$C(\mu_i, \widetilde{\mu}_i) = \min_{\pi \in \Gamma(\mu_i, \widetilde{\mu}_i)} \int c(x, \tilde{x}) d\pi(x, \tilde{x}). \tag{4}$$

Remark 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 μ_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

$$f^c(x) := \inf_{x' \in \mathcal{X}} \{f(x') + c(x', x)\}, \tag{5}$$

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

$$g^{\bar{c}}(x') := \sup_{x \in \mathcal{X}} \{g(x) - c(x', x)\}. \tag{6}$$

Example 1 *Let $\varepsilon > 0$ and let $c(x, \tilde{x})$ be given by*

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

*The parameter ε can be interpreted as the adversarial budget: the larger the value of ε, the wider the space of actions available to the adversary. The cost c satisfies **Assumption** 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:

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

In this setting, the adversarial problem (1) can be written as

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

where $W_\infty(\mu, \widetilde{\mu})$ is the ∞-OT distance between μ 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 3 *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 (e.g., seePydi and Jog (2021a)) and distributional perturbing adversarial learning (e.g., see Blanchet and Murthy (2019); Blanchet, Kang, and Murthy (2019)). For a rigorous analysis, distributional perturbing adversarial learning is more adequate since data-perturbing adversarial learning lacks measurability in some cases. Furthermore, putting some technical details aside, one can prove that distributional perturbing*

adversarial learning encompasses data-perturbing adversarial learning: see Pydi and Jog (2021a).

The main focus in this paper is based on the distributional setting, where given a data distribution μ, an adversary can select a new distribution $\widetilde{\mu}$ in a neighborhood of the original distribution μ determined by C. Pydi and Jog (2021b) summarizes other adversarial models and discusses connections between them.

Example 2 *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) 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 μ and $\widetilde{\mu}$ relative to the distance function δ from **Example** 1. The relation between τ and ε is not explicit, but, qualitatively, small values of τ should correspond to small values of ε.*

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) is the full set of weak partitions, or probabilistic classifiers, defined by

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

where

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

i.e., Δ_K is 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 specific classifier f chosen by 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 when training models. We want to highlight that the fs in $\mathcal{F}$ are only assumed to be Borel measurable. This means that the learner in problem (1) can be considered to be *agnostic* to any specific model for the classifiers and in that sense (1) can be interpreted as a robust generalization of the notion of Bayes classifier studied in statistical learning.

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

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

Notice that $\ell(e_j, i)$ is equal to 1 if $i \neq j$ and 0 if $i = j$. ℓ thus extends the 0-1 loss to weak classifiers, and from now on we will refer to it simply as the 0-1 loss. 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 $R(\cdot, \cdot) : \mathcal{F}\times\mathcal{M}_+(\mathcal{Z}) \longrightarrow \mathbb{R}_+$. For convenience, we introduce the so-called *classification power* for a pair $(f, \widetilde{\mu}) \in \mathcal{F} \times \mathcal{M}_+(\mathcal{Z})$, which is defined by

$$B(f, \widetilde{\mu}) := \sum_{i\in[K]} \int_{\mathcal{X}} f_i(\widetilde{x})d\widetilde{\mu}_i(\widetilde{x}). \tag{7}$$

With these new definitions, problem (1) is immediately seen to be equivalent to

$$\sup_{f\in\mathcal{F}} \inf_{\widetilde{\mu}\in\mathcal{M}_+(\mathcal{Z})} \left\{B(f, \widetilde{\mu}) + C(\mu, \widetilde{\mu})\right\}. \tag{8}$$

Moreover, if we denote by $\widetilde{B}^*_\mu$ the optimal value of (8), and by R^*_μ the optimal value of (1), we have the identity:

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

We write $\mu(\mathcal{Z})$ explicitly, although for the most part $\mu(\mathcal{Z})$ can be thought of as being equal to one.

The dual of (8) is obtained by swapping the sup and the inf:

$$\inf_{\widetilde{\mu}\in\mathcal{M}_+(\mathcal{Z})} \sup_{f\in\mathcal{F}} \left\{B(f, \widetilde{\mu}) + C(\mu, \widetilde{\mu})\right\}. \tag{9}$$

Notice that the value of (9) is always greater than or equal to the value of (8). Instead of attempting to invoke an abstract minimax theorem implying the equality of these two quantities at this stage, we prefer to defer this discussion to later sections where in fact we will prove that, under **Assumption** 1, there is no duality gap in this problem. In what follows we focus on the dual problem (9) and only return to problem (8), which is equivalent to the original adversarial problem (1), in section 4.3. Notice, however, that the statement of **Theorem** 6 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 λ, a positive measure over $\mathcal{X}$, we can rewrite the latter sup as:

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

the constraint in λ can be simply written as $\lambda \geq \widetilde{\mu}_i$ for all $i \in [K]$. Combining the above with the structure of the cost $C(\mu, \widetilde{\mu})$, we conclude that problem (9) is equivalent to the generalized barycenter problem mentioned in the introduction:

$$B_\mu^* := \inf_{\lambda, \widetilde{\mu}_1, \dots, \widetilde{\mu}_K} \left\{ \lambda(\mathcal{X}) + \sum_{i \in [K]} C(\mu_i, \widetilde{\mu}_i) : \ \lambda \geq \widetilde{\mu}_i \text{ for all } i \in [K] \right\}, \tag{10}$$

where we use the notation B_μ^* for future reference; see Figure 1 for a pictorial explanation.

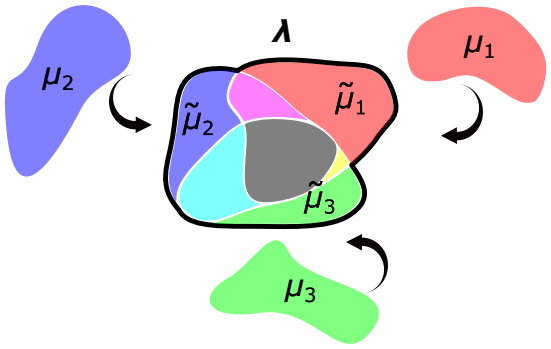

Figure 1: Picture for (10). μ_i's are first moved to $\widetilde{\mu}_i$'s and λ is chosen to cover all $\widetilde{\mu}_i$'s: it is the smallest positive measure which is larger than all $\widetilde{\mu}_i$'s.

Remark 4 *It is straightforward to see from* (9) *that* B_μ^* *is* 1*-homogeneous in* μ*. That is, if* $a > 0$*, then* $B_{a\mu}^* = aB_\mu^*$*.*

2.1 The MOT problem

2.1.1 General MOT problems

Before providing the details of our MOT problem (2), it is worth introducing the generic MOT problem first. Let $\mathcal{S}_1, \dots, \mathcal{S}_K$ be fixed spaces and let $\mathbf{c} : \mathcal{S}_1 \times \cdots \times \mathcal{S}_K \to \mathbb{R} \cup \{+\infty, -\infty\}$ be a cost function. For each $1 \leq k \leq K$, let $\nu_k \in \mathcal{P}(\mathcal{S}_k)$ be a Borel probability measure. The MOT problem associated to the cost function $\mathbf{c}$ and the measures $\nu_1, \dots, \nu_K$ is the following (possibly infinite dimensional) linear optimization problem with K-marginal constraints:

$$\inf_{\pi \in \Pi(\nu_1, \dots, \nu_K)} \int_{\mathcal{S}_1 \times \cdots \times \mathcal{S}_K} \mathbf{c}(\xi_1, \dots, \xi_K) d\pi(\xi_1, \dots, \xi_K),$$

where

$$\Pi(\nu_1, \dots, \nu_K) := \{\pi \in \mathcal{P}(\mathcal{S}_1 \times \cdots \times \mathcal{S}_K), \text{ s.t., for every } i, \ i\text{-th marginal of } \pi = \nu_i\}.$$

MOTs are generalizations of the standard (two marginals) optimal transport (OT) problems and their duals take the form:

$$\sup_{\phi\in\Phi}\left\{\sum_{j=1}^{K}\int_{\mathcal{S}_j}\phi_j(\xi_j)d\nu_j(\xi_j)\right\}, \tag{11}$$

where Φ is the set of all $\phi=(\phi_1,\dots,\phi_K)\in\prod_{j=1}^{K}L^1(\nu_j)$ such that

$$\sum_{j=1}^{K}\phi_j(\xi_j)\leq \mathbf{c}(\xi_1,\dots,\xi_K),\quad \forall(\xi_1,\dots,\xi_K)\in\mathcal{S}_1\times\cdots\times\mathcal{S}_K.$$

One of the most popular examples of MOT problems is connected to the Wasserstein Barycenter problem over $\mathcal{P}(\mathcal{X})$; see Ekeland (2005); Chiappori, McCann, and Nesheim (2010); Agueh and Carlier (2011). Let $c:\mathcal{X}\times\mathcal{X}\to\mathbb{R}\cup\{+\infty,-\infty\}$ be a fixed pairwise cost function. In the Wasserstein barycenter problem the goal is to find a solution ν^* to the problem

$$\inf_{\nu'}\sum_{i\in[K]}C(\nu',\nu_i)\quad\text{where } C(\nu,\nu_i):=\inf_{\pi\in\Pi(\nu,\nu_i)}\int_{\mathcal{X}\times\mathcal{X}}c(x',x)d\pi(x',x).$$

Such ν^* can be interpreted as an "average" or barycenter of the input measures $\nu_1,\dots,\nu_K$ relative to the cost C. It can then be showed that the above Wasserstein barycenter problem is equivalent to solving the following MOT problem

$$\inf_{\pi\in\Pi(\nu_1,\dots,\nu_K)}\int_{\mathcal{X}^K}\mathbf{c}(x_1,\dots,x_K)d\pi(x_1,\dots,x_K),$$

where

$$\mathbf{c}(x_1,\dots,x_K):=\inf_{x'\in\mathcal{X}}\sum_{i\in[K]}c(x',x_i).$$

Indeed, let π^* be a minimizer of the above MOT problem. Defining $\nu^*=T_{\#}\pi$, where $T(x_1,\dots,x_K):=\operatorname{argmin}_{x'}\sum_{i\in[K]}c(x',x_i)$, i.e., defining ν^* as the pushforward measure of π^* with respect to the barycenter mapping T, one can recover a solution to the original barycenter problem. Conversely, one can use a Wasserstein barycenter ν^* and couplings π_i realizing the costs $C(\nu^*,\nu_i)$ to build a solution to the MOT problem; see more details in Agueh and Carlier (2011).

2.1.2 From adversarial robustness to MOT

Now we are ready to state problem (2) precisely. For this, we will need to modify the space $\mathcal{Z}$ and in particular add an extra element to it that will be denoted by the symbol $\triangle$. The marginals of the couplings in the desired MOT problem will be probability measures over the set $\mathcal{Z}_*:=\mathcal{Z}\cup\{\triangle\}$. More precisely, letting P_i represent the projection onto the i-th coordinate, we consider the set:

$$\Pi_K(\mu):=\left\{\pi\in\mathcal{P}(\mathcal{Z}_*^K)\ :\ P_{i\sharp}\pi=\frac{1}{2\mu(\mathcal{Z})}\mu(\cdot\cap\mathcal{Z})+\frac{1}{2}\delta_{\triangle}\text{ for all } i\in[K]\right\}. \tag{12}$$

Notice that in this set all K marginals are the same. Dividing by the factor $\frac{1}{2\mu(\mathcal{Z})}$, the set $\Pi_K(\mu)$ is made 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. 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

$$\mathbf{c}(z_1, \dots, z_K) := B^*_{\widehat{\mu}_{\vec{z}}}, \tag{13}$$

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

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

Recall that $B^*_{\widehat{\mu}_{\vec{z}}}$ is equal to (10) (alternatively, equal to (9)) when μ is equal to $\widehat{\mu}_{\vec{z}}$. In this sense, $\mathbf{c}(z_1, \dots, z_K)$ of (13) is the value of the generalized barycenter problem given $\widehat{\mu}_{\vec{z}}$ as the data distribution, or *local* generalized barycenter problem.

Remark 5 *Notice that $\widehat{\mu}_{\vec{z}}$ is a probability measure if and only if no element in the tuple $\vec{z}$ is $\triangle$.*

Following the literature of MOT, the dual of our MOT problem can be written as

$$\sup_{\phi \in \Phi} \left\{ \sum_{j=1}^{K} \int_{\mathcal{X} \times [K]} \phi_j(z_j) \frac{1}{2\mu(\mathcal{Z})} d\mu(z_j) + \frac{1}{2} \sum_{j=1}^{K} \phi_j(\triangle) \right\}, \tag{14}$$

where

$$\Phi := \left\{ \phi = (\phi_1, \dots, \phi_K) \in \prod_{j=1}^{K} L^1\big(\frac{1}{2\mu(\mathcal{Z})}\mu + \frac{1}{2}\delta_{\triangle}\big) : \sum_{j=1}^{K} \phi_j(z_j) \leq B^*_{\widehat{\mu}_{\vec{z}}}, \quad \forall \vec{z} \in \mathcal{Z}_*^K \right\}. \tag{15}$$

We will later show that under **Assumption** 1 there is no duality gap between the MOT problem and its dual (14) —see **Corollary** 31.

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

Theorem 6 *Suppose that **Assumption** 1 holds. Let μ be a finite positive measure over $\mathcal{Z}$. Then* (9) *is equivalent to the MOT problem* (2) *with set of couplings $\Pi_K(\mu)$ defined as in* (12)*, and cost function $\mathbf{c}$ defined as in* (13)*. 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, (8) = (9)*. In addition, from a solution pair (π^*, ϕ^*) for the MOT problem and its dual one can obtain a solution pair $(f^*, \widetilde{\mu}^*)$ for* (9) *and its dual, i.e. problem* (8)*. The pair $(f^*, \widetilde{\mu}^*)$ is also a saddle point for the original adversarial problem* (1)*.*

One immediate consequence of **Theorem** 6 is that with the identity

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

one can compute R^*_μ, the optimal adversarial risk, by finding the optimal value of the equivalent MOT problem. To find the latter, one could attempt to use one of the off-the-shelf algorithms in computational optimal transport. Some algorithms to solve generic MOTs that have been developed recently include the ones proposed in see Benamou, Carlier, Cuturi, Nenna, and Peyré (2015); Benamou, Carlier, and Nenna (2019); Lin, Ho, Cuturi, and Jordan (2019); Tupitsa, Dvurechensky, Gasnikov, and Uribe (2020); Haasler, Ringh, Chen, and Karlsson (2021); Altschuler and Boix-Adsera (2021); Carlier (2022). Our numerical results for a subsample of MNIST and CIFAR 10, shown in Figure 6, are obtained using the algorithm discussed in Lin, Ho, Cuturi, and Jordan (2019), also known as MOT Sinkhorn algorithm; see subsection 5.3 for more details. We want to warn the reader, however, that off-the-shelf MOT algorithms may suffer an excessive computational burden when K goes beyond 4. For this reason, it is important to develop algorithms that exploit the structure of our MOT problem, which, as we will discuss below, has the structure of a generalized barycenter problem. An investigation on more specific algorithms is left for future work.

The proof of **Theorem** 6 is presented throughout section 4; the expression for $(f^*, \widetilde{\mu}^*)$ in terms of (ϕ^*, π^*) is presented in **Corollary** 33. Given the definition of the cost function $\mathbf{c}$, **Theorem** 6 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 ghost element, $\triangle$, indicates when fewer than K points are being grouped by the adversary. We highlight that it is not always (globally) optimal for the adversary to group together points from all the K different classes whenever it is possible.

We emphasize that from the solution to the MOT and its dual, one can directly obtain an optimal adversarial attack and an optimal classification rule for the original adversarial problem. Note that problem (2) is a problem solved by the adversary: ideally, the adversary wants to group together points $(z_1, \ldots, z_K)$ 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 (2) can be interpreted as a maximization problem solved by the learner. We formalize this novel connection in subsection 4.3: see **Corollary** 33.

In order to prove **Theorem** 6, we will first obtain a series of equivalent reformulations of problem (10) 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 7 *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 finite positive measures $\mu_1, \dots, \mu_K$ and c as above, there exists at least one solution to problem* (10).

Remark 8 *If c is a cost that satisfies **Assumption** 1, then c satisfies the hypothesis of **Proposition** 7.*

Remark 9 *Nearly identical arguments can be used to prove that the various reformulations of* (10) *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 (10), we can rewrite the problem in the following form

$$\inf_{\lambda,\, \pi_1,\dots,\pi_K} \left\{ \lambda(\mathcal{X}) + \sum_{i\in[K]} \int_{\mathcal{X}\times\mathcal{X}} c(x,x') d\pi_i(x,x') \right\}$$
$$\text{s.t. } \pi_i(\mathcal{X}\times E) \leq \lambda(E), \pi_i(E\times\mathcal{X}) = \mu_i(E) \text{ for all } i \in [K], \quad \forall E \subseteq \mathcal{X} \text{ Borel.}$$

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

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

$$\begin{aligned} t &:= \inf_{\lambda,\pi_1,\dots,\pi_K} \lambda(\mathcal{X}) + \sum_{i\in[K]} \int_{\mathcal{X}\times\mathcal{X}} c(x,x') d\pi_i(x,x') \\ &= \lim_{n\to\infty} \lambda^n(\mathcal{X}) + \sum_{i\in[K]} \int_{\mathcal{X}\times\mathcal{X}} c(x,x') d\pi_i^n(x,x'). \end{aligned}$$

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

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

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

The tightness of λ^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 (16) and geometric consequences

In the rest of what follows, we shall let S_K denote the power set of $[K]$ except for the empty set and for every $i \in [K]$ we let $S_K(i) = \{A \in S_K : i \in A\}$. We can reduce (10) to a more concrete problem by partitioning λ and each of μ_i's properly, eliminating the variables $\widetilde{\mu}_i$'s from the optimization. We start with the following observation.

Lemma 10 *Let $u_1, \dots, u_K \in [0,1]$ be such that $\max_{i=1,\dots,K} u_i = 1$. Then there exists a collection of non-negative scalars $\{r_A\}_{A \in S_K}$ such that the following two conditions hold:*

1. $1 = \sum_{A \in S_K} r_A$.
2. $u_i = \sum_{A \in S_K(i)} r_A$ *for all* $i = 1, \dots, K$.

Proof Without loss of generality we can assume that the u_i are arranged in increasing order. That is,

$$0 \leq u_1 \leq u_2 \leq \dots, \leq u_K = 1.$$

Let i' be the first i such that $u_i > 0$. We set

$$\begin{aligned} r_{\{i',\dots,K\}} &:= u_{i'} \\ r_{\{i'+1,\dots,K\}} &:= u_{i'+1} - u_{i'} \\ r_{\{i'+2,\dots,K\}} &:= u_{i'+2} - u_{i'+1} \\ &\vdots \\ r_{\{K\}} &:= 1 - u_{K-1}. \end{aligned}$$

and $r_A = 0$ for all other sets. It is straightforward to check that the collection $\{r_A\}_{A \in S_K}$ defined in this way satisfies the required conditions. ∎

Proposition 11 *Problem* (10) *is equivalent to*

$$\begin{aligned} &\inf_{\{\lambda_A, \mu_{i,A} : i \in [K], A \in S_K\}} \sum_{A \in S_K} \left\{ \lambda_A(\mathcal{X}) + \sum_{i \in A} C(\lambda_A, \mu_{i,A}) \right\} \\ &\text{s.t.} \sum_{A \in S_K(i)} \mu_{i,A} = \mu_i \text{ for all } i \in [K]. \end{aligned} \tag{16}$$

Proof We split the proof into two parts.

Step 1: Suppose that $\lambda, \tilde{\mu}_1, \dots, \tilde{\mu}_K$ is feasible for problem (10). In particular, $\lambda \geq \tilde{\mu}_i$ for all i. Let us denote by $\frac{d\tilde{\mu}_i}{d\lambda}$ the Radon-Nikodym derivative of $\tilde{\mu}_i$ w.r.t. λ. Notice that $\frac{d\tilde{\mu}_i}{d\lambda} \leq 1$ because λ dominates $\tilde{\mu}_i$. Moreover, without the loss of generality we can assume that for every $x \in \mathrm{spt}(\lambda)$ we have $\max_{i=1,\dots,K} \frac{d\tilde{\mu}_i}{d\lambda}(x) = 1$, for otherwise we could modify λ and potentially reduce the energy in (10) while maintaining the constraints.

For each $x \in \mathrm{spt}(\lambda)$ we apply Lemma 10 with $u_i(x) := \frac{d\tilde{\mu}_i}{d\lambda}(x)$ to obtain a collection of scalars $\{r_A(x)\}_{A \in S_K}$ satisfying:

(i) $1 = \sum_{A \in S_K} r_A(x)$.

(ii) $u_i(x) = \sum_{A \in S_K(i)} r_A(x)$ for all $i = 1, \dots, k$.

Notice that the functions $r_A(\cdot)$ can be constructed in a measurable way as it follows from the proof of Lemma 10. For each $A \in S_K$ we define the measure λ_A as

$$\frac{d\lambda_A}{d\lambda}(x) := r_A(x),$$

and for A and $i \in A$ we define

$$\tilde{\mu}_{i,A} := \lambda_A.$$

See Figure 2 (a) for an illustration of the λ_A's. From the above definitions and the properties of the functions r_A we deduce

$$\sum_{A \in S_K} d\lambda_A(x) = \sum_{A \in S_K} r_A(x) d\lambda(x) = d\lambda(x)$$

and

$$\sum_{A \in S_K(i)} d\tilde{\mu}_{i,A}(x) = \sum_{A \in S_K(i)} r_A(x) d\lambda(x) = \frac{d\tilde{\mu}_i}{d\lambda}(x) d\lambda(x) = d\tilde{\mu}_i(x).$$

Now, let $\pi_i \in \Gamma(\mu_i, \tilde{\mu}_i)$ be a coupling realizing the cost $C(\mu_i, \tilde{\mu}_i)$, i.e., a minimizer of (4), and use the disintegration theorem to write it as

$$d\pi_i(x, \tilde{x}) = d\pi_i^*(x|\tilde{x}) d\tilde{\mu}_i(\tilde{x}),$$

where $d\pi_i^*(\cdot|\tilde{x})$ is the conditional of x given $\tilde{x}$ according to the joint distribution π_i^*. For each $A \in S_K$ and $i \in A$ we define the measure $\pi_{i,A}$ according to

$$d\pi_{i,A}(x, \tilde{x}) := d\pi_i^*(x|\tilde{x}) d\tilde{\mu}_{i,A}(\tilde{x}).$$

Finally, we set $\mu_{i,A}$ to be the first marginal of $\pi_{i,A}$.

It is now straightforward to show that $\{\lambda_A, \mu_{i,A}\}$ is feasible for (16). Moreover,

$$\lambda(\mathcal{X}) + \sum_{i=1}^{k} C(\mu_i, \tilde{\mu}_i) \geq \sum_{A \in S_K} \left\{ \lambda_A(\mathcal{X}) + \sum_{i \in A} C(\lambda_A, \mu_{i,A}) \right\}.$$

Step 2: Conversely, suppose that $\{\lambda_A\}_A$, $\{\mu_{i,A}\}_A$ is feasible for (16). Set $\lambda := \sum_{A \in S_K} \lambda_A$ and for every i let $\tilde{\mu}_i := \sum_{A \in S_k(i)} \lambda_A$. Clearly we have $\lambda \geq \tilde{\mu}_i$ for all i. Moreover, let $\pi_{i,A} \in \Gamma(\mu_{i,A}, \lambda_A)$ realizing the cost $C(\lambda_A, \mu_{i,A})$. See (b) of Figure 2 to understand how $\mu_{i,A}$ is transported to λ_A. Finally, for each i we set

$$\pi_i := \sum_{A \in S_K(i)} \pi_{i,A}.$$

With these constructions it is now straightforward to show that

$$\sum_{A \in S_K} \left\{ \lambda_A(\mathcal{X}) + \sum_{i \in A} C(\lambda_A, \mu_{i,A}) \right\} \geq \lambda(\mathcal{X}) + \sum_{i=1}^{k} C(\mu_i, \tilde{\mu}_i).$$

■

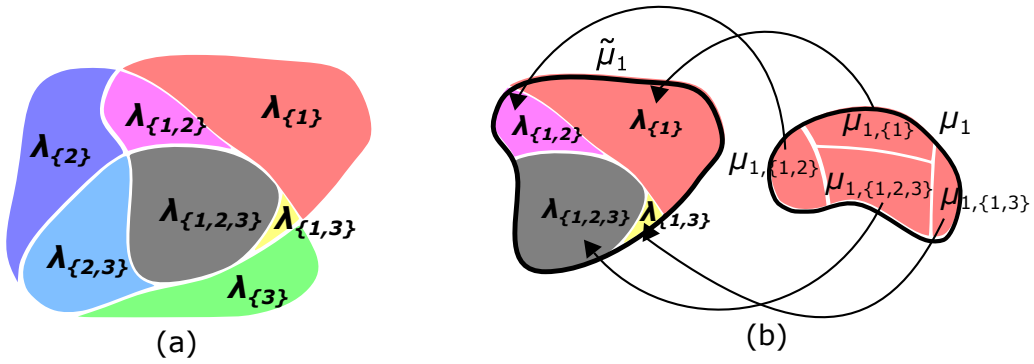

Figure 2: (a) : Illustration of a partition of λ. (b) : Illustration of the transport from $\mu_{1,A}$'s to λ_A's.

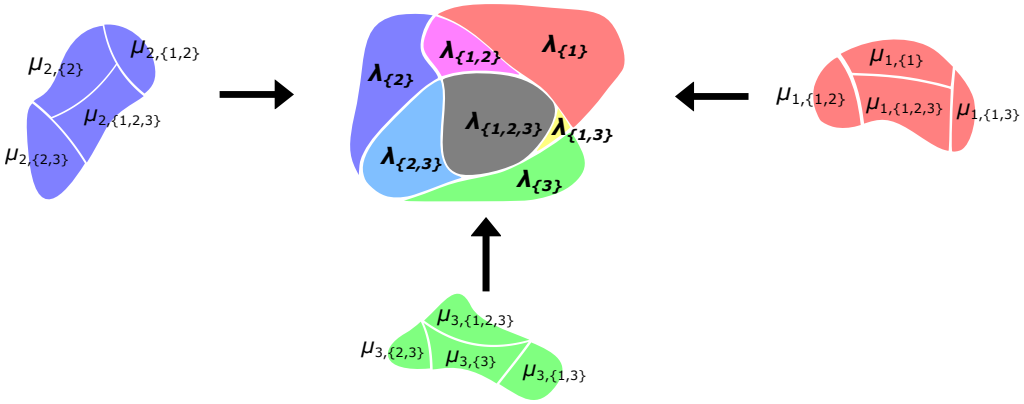

Figure 3: Picture for (16). Each of $\mu_{i,A}$'s is transported to λ_A for all $i \in A$.

Remark 12 *Figure 3 illustrates the partitions for λ and the μ_i's. To keep notation from getting too complicated, in the sequel we shall assume that $\mu_{i,A}$ is defined for all $i \in [K]$ and $A \subseteq S_K$, however, note that if $i \notin A$, then $\mu_{i,A}$ plays no role in the optimization* (16).

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 $\lambda_A^* = \lambda_A^*(\mu_{1,A}, \ldots, \mu_{K,A})$ by solving the classic Wasserstein barycenter problem. Indeed, the optimal choice must be an element of

$$\operatorname*{argmin}_{\lambda_A} \sum_{i \in A} C(\lambda_A, \mu_{i,A}). \tag{17}$$

Note that here we do not need to consider the mass of λ_A, since the value of the optimization problem will be $+\infty$ if λ_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 (17) can be reformulated as a multimarginal optimal transport problem Agueh and Carlier (2011); see also our subsection 2.1.1. To that end, given $A \subseteq [K]$, define $c_A : \mathcal{X}^K \to \mathbb{R}$

$$c_A(x_1, \dots, x_K) := \inf_{x' \in \mathcal{X}} \sum_{i \in A} c(x', x_i), \tag{18}$$

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

$$T_A(x_1, \dots, x_K) := \operatorname*{argmin}_{x' \in \mathcal{X}} \sum_{i \in A} c(x', x_i). \tag{19}$$

Remark 13 *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 (17) as the multimarginal optimal transport problem

$$\inf_{\pi_A} \int_{\mathcal{X}^K} c_A(x_1, \dots, x_K) d\pi_A(x_1, \dots, x_K) \quad \text{s.t. } \mathcal{P}_{i\#}\pi_A = \mu_{i,A} \text{ for all } i \in A, \tag{20}$$

where $\mathcal{P}_i$ is the projection map $(x_1, \dots, x_K) \mapsto x_i$. Again, even though π_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 (20) we can now eliminate λ_A and all of the $\mu_{i,A}$'s from problem (16) and reformulate the optimization as the multimarginal problem

$$\begin{aligned} &\inf_{\{\pi_A : A \in S_K\}} \sum_{A \in S_K} \int_{\mathcal{X}^K} \big(c_A(x_1, \dots, x_K) + 1\big) d\pi_A(x_1, \dots, x_K) \\ &\text{s.t.} \sum_{A \in S_K(i)} \mathcal{P}_{i\#}\pi_A = \mu_i \text{ for all } i \in [K]. \end{aligned} \tag{21}$$

The next two propositions formally prove the equivalence between (16) and (21). They will also allow us to establish some important geometric properties of optimal generalized barycenters.

Proposition 14 *Let* c *be a cost satisfying* ***Assumption*** *1. Given measures* $\mu_1, \dots, \mu_K$, *let* $\{\pi_A\}_{A \in S_K}$ *be a feasible solution to* (21). *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* (19).

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* (16) *and*

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

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 push-forwards 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 (16).

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,

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

which is the desired result we want. ■

In the next proposition we will show that any feasible solution of problem (16) induces a feasible solution of (21) with lesser or equal value. This will prove the equivalence between problems (16) and (21) and will provide a powerful geometric characterization of optimal generalized barycenters.

Proposition 15 *Let c be a cost satisfying **Assumption** 1. Given measures $\mu_1, \dots, \mu_K$, let $\mu_{i,A}, \lambda_A$ be feasible solutions to problem* (16)*. Let $\gamma_{i,A} \in \mathcal{M}(\mathcal{X} \times \mathcal{X})$ be an optimal plan for the transport of $\mu_{i,A}$ to λ_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 (21) *and*

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

Therefore, (16) = (21)*.*

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

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

Thus, each γ_A is well-defined.

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

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

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

In addition to proving the equivalence between problems (16) and (21), **Proposition** 14 and **Proposition** 15 have the following very important geometric consequences.

Corollary 16 *Let c be a cost satisfying **Assumption** 1. Given measures $\mu_1, \dots, \mu_K$, let λ be an optimal generalized barycenter and let $\{\lambda_A\}_{A\in S_K}$ be a decomposition of λ and $\{\mu_{i,A}\}_{A\in S_K(i)}$ a decomposition of each μ_i that are optimal for (16). Recalling (19), let $T_A(x_1,\dots,x_K) := \operatorname{argmin}_{x\in\mathcal{X}} \sum_{i\in A} c(x,x_i)$. If we define $T_A := \{T_A(x_1,\dots,x_K) : x_1 \in \operatorname{spt}(\mu_1),\dots,x_K \in \operatorname{spt}(\mu_K)\}$ and $T = \cup_{A\subseteq[K]} T_A$, then $\lambda_A(\mathcal{X}) = \lambda_A(T_A)$, $\lambda(\mathcal{X}) = \lambda(T)$ and the optimal measures $\widetilde{\mu}_i$ in (10) can be assumed to satisfy $\widetilde{\mu}_i(\mathcal{X}) = \widetilde{\mu}_i(T)$ as well.*

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

Remark 17 *In the case where we have a tuple $(x_1,\dots,x_K) \in \operatorname{spt}(\mu_1)\times\cdots\times\operatorname{spt}(\mu_K)$ such that $\sum_{i\in A} c(x,x_i) = +\infty$ for all $x \in \mathcal{X}$, we set $T_A(x_1,\dots,x_K) = \emptyset$.*

Proof From **Proposition** 15, 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

$$\sum_{A\in S_K} \int_{\mathcal{X}^K} \big(c_A(x_1,\dots,x_K)+1\big) d\tilde{\pi}_A(x_1,\dots,x_K) \le \sum_{A\in S_K} \lambda_A(\mathcal{X}) + \sum_{i\in A} C(\lambda_A,\mu_{i,A}). \quad (22)$$

From **Proposition** 14, 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

$$\sum_{A\in S_K} \tilde{\lambda}_A(\mathcal{X}) + \sum_{i\in A} C(\tilde{\lambda}_A,\tilde{\mu}_{i,A}) \le \sum_{A\in S_K} \int_{\mathcal{X}^K} \big(c_A(x_1,\dots,x_K)+1\big) d\tilde{\pi}_A(x_1,\dots,x_K). \quad (23)$$

Examining the proof of **Proposition** 15, it follows that the inequality in (22) is strict if $\lambda_A(\mathcal{X}) > \lambda_A(T_A)$. In that case, combining (22) and (23) would contradict the optimality of λ. 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** 14. ■

When $\mu_1,\dots,\mu_K$ are supported on a finite set of points, **Corollary** 16 has the following consequence.

Corollary 18 *If $\mu_1,\dots,\mu_K$ are measures that are supported on a finite set of points and c is a cost satisfying **Assumption** 1, then there exists a solution λ to the optimal generalized barycenter problem (10) that is supported on a finite set of points.*

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

Remark 19 *Notice that the bound mentioned at the end of **Corollary** 18 is a worst case bound. In practice, especially when data sets have a favourable geometric structure, the optimal barycenter λ may have a much sparser support. See section 5.2.*

Proof For each $i \in [K]$ we can assume there exists a finite set $X_i \subset \mathcal{X}$ such that μ_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 (19). 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** 16, there exists an optimal barycenter supported on F only. ■

3.2 A second MOT reformulation of (16)

Note that in problem (21) we need to find a distribution π_A for each $A \in S_K$. Hence, it is natural to wonder if we can reformulate problem (21) in such a way that we only need to find a single distribution γ. 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 (21).

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

$$\begin{aligned} &\widetilde{c}((x_1, r_1), \dots, (x_K, r_K)) \\ &:= \inf_{m : S_K \to \mathbb{R}} \sum_{A \in S_K} m_A\big(c_A(x_1, \dots, x_K) + 1\big) \quad \text{s.t.} \sum_{A \in S_K(i)} m_A = r_i. \end{aligned} \tag{24}$$

For each $i \in [K]$, 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 (21) is equivalent to

$$\inf_{\gamma} \int_{(\mathcal{X} \times [0,1])^K} \widetilde{c}(\vec{x}, \vec{r}) d\gamma(\vec{x}, \vec{r}) \quad \text{s.t. } \tilde{\mathcal{P}}_{i\,\#}(r_i \gamma) = \mu_i \text{ for all } i \in [K]. \tag{25}$$

Proposition 20 *Problems* (21) *and* (25) *are equivalent, and thus* (25) *is also equivalent to* (9)*,* (10) *and* (16)*.*

Proof Given a feasible solution $\pi_{\{1\}}, \dots, \pi_{[K]}$ to problem (21), define γ 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 γ is feasible for (25), since for any function $g : \mathcal{X} \to \mathbb{R}$

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

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 $\big((x_1, \chi_A(1)), \ldots, (x_K, \chi_A(K))\big)$ we have

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

Therefore,

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

Conversely, suppose that γ is a feasible solution to (25). Given a tuple $(\vec{x}, \vec{r})$, let

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

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

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

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

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

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 (21). Evaluating the objective in (21) we see that

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

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 (25) 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 (10) (and hence for problem (9) as well). Compare with the discussion after **Theorem** 6.

Lemma 21 *Let $\tilde{x}_1, \dots, \tilde{x}_k \in \mathcal{X}$, and let $0 \leq \tilde{r}_1, \dots, \tilde{r}_k \leq 1$. Then $\widetilde{c}((\tilde{x}_1, \tilde{r}_1), \dots, (\tilde{x}_K, \tilde{r}_K))$ defined in* (24) *is equal to $B^*_{\widehat{\mu}}$, where*

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

Proof To prove this claim we first notice that by **Proposition** 20 $B^*_{\widehat{\mu}}$ is equal to

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

where γ is in the constraint set of problem (25). For a feasible γ, notice that γ must concentrate on the set $\{(\vec{x}, \vec{r}) : x_i = \tilde{x}_i, i \in [K]\}$. Applying the disintegration theorem to γ, we can rewrite the objective function evaluated at γ 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 γ_r is a positive measure over $[0,1]^K$ satisfying the constraints:

$$\int_{[0,1]} r_i d\gamma_r(r_1, \dots, r_K) = \tilde{r}_i, \quad \forall i = 1, \dots, K. \tag{26}$$

It is clear that the map associating a feasible γ to a γ_r satisfying (26) is onto, and thus, we can rewrite $B^*_{\widehat{\mu}}$ as

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

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 (24) 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

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

where we define

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

Notice that

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

Conversely, notice that given a collection of functions $\tilde{m}_A$ satisfying the constraint in (24) for the tuple $(\tilde{x}_1, \tilde{r}_1), \ldots, (\tilde{x}_K, \tilde{r}_K)$, it is straightforward to find γ_r such that $\tilde{m}_A = m_{A,\gamma_r}$ for all A. It now follows that

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

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 16.

Proposition 22 *The dual problems to* (10), (21), *and* (25) *can be written as*

$$\begin{aligned}
&\sup_{f_1, \ldots, f_K \in C_b(\mathcal{X})} \sum_{i \in [K]} \int_{\mathcal{X}} f_i^c(x_i) d\mu_i(x_i) \\
&\text{s.t. } f_i(x) \geq 0, \ \sum_{i \in [K]} f_i(x) \leq 1, \text{ for all } x \in \mathcal{X}, i \in [K],
\end{aligned} \tag{27}$$

$$\begin{aligned}
&\sup_{g_1, \ldots, g_K \in C_b(\mathcal{X})} \sum_{i \in [K]} \int_{\mathcal{X}} g_i(x_i) d\mu_i(x_i) \\
&\text{s.t. } \sum_{i \in A} g_i(x_i) \leq 1 + c_A(x_1, \ldots, x_K) \text{ for all } (x_1, \ldots, x_K) \in \mathcal{X}^K, A \in S_K,
\end{aligned} \tag{28}$$

and

$$\begin{aligned} &\sup_{h_1,\dots,h_K \in C_b(\mathcal{X})} \sum_{i\in[K]} \int_{\mathcal{X}} h_i(x_i) d\mu_i(x_i) \\ &\text{s.t.} \sum_{i\in[K]} r_i h_i(x_i) \leq \widetilde{c}(\vec{x},\vec{r}) \text{ for all } (\vec{x},\vec{r}) \in (\mathcal{X}\times[0,1])^K, \end{aligned} \tag{29}$$

respectively.

Let $f_1,\dots,f_K$; $g_1,\dots,g_K$; $h_1,\dots,h_K$ be feasible solutions to problems (27), (28), *and* (29) *respectively. Problems* (28) *and* (29) *have the same feasible set and hence are identical. Furthermore, $g_i' := f_i^c$ is a feasible solution to* (28) *and $f_i' = \max\{g_i,0\}^{\bar{c}}$ is a feasible solution to* (27), *hence the optimization of* (28) *can be restricted to nonnegative g_i that satisfy $g_i = g_i^{\bar{c}c}$. In particular,* (27), (28), *and* (29) *all have the same optimal value.*

Proof The derivation of the dual problems is standard.

To see the equivalence between problems (28) and (29), fix some $h_1,\dots,h_K$ that are feasible for (29) 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

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

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

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

Thus, we see that the h_i are feasible for (28) since B and $(x_1,\dots,x_K)$ were arbitrary.

Conversely, fix some $g_1,\dots,g_K$ that are feasible for (28) 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

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

where we used the feasibility of the g_i. Thus, the g_i are feasible for (29). Since both problems are optimizing the same functional over the same constraint set, we see that (28) and (29) are identical.

Now suppose that $f_1, \dots, f_K$ and $g_1, \dots, g_K$ are feasible solutions to problems (27) and (28) 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 (28). Hence, the optimal value of (28) cannot lie strictly below the optimal value of (27).

It remains to verify the feasibility of the f_i'. We begin by showing that if $g_1, \dots, g_K$ are feasible for (28) 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 $i \in [K]$, 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

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

where the final inequality follows from the feasibility of $\max\{g_i, 0\}$. Now from the definition of $c_{[K]}$, 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 (28), then $\max\{g_1, 0\}^{\bar{c}c}, \dots, \max\{g_K, 0\}^{\bar{c}c}$ are also feasible for (28). Since $u \leq u^{\bar{c}c}$ for any function $u : \mathcal{X} \to \mathbb{R}$, it follows that

$$\sum_{i \in [K]} \int_{\mathcal{X}} g_i(x) d\mu_i(x) \leq \sum_{i \in [K]} \int_{\mathcal{X}} \max\{g_i, 0\}^{\bar{c}c}(x) d\mu_i(x).$$

Since we showed that $\max\{g_i, 0\}^{\bar{c}}$ was feasible for (27), it follows that (28) cannot attain a larger value than (27). Hence, we have shown that (28) and (27) 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 23 *Suppose that c is a bounded Lipschitz cost that satisfies the hypotheses of* ***Proposition 7****. 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*

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

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 $\Phi_c = \{(\varphi_i, \psi_i) \in C_b(\mathcal{X}) \times C_b(\mathcal{X}) : \varphi_i(x) - \psi_i(x') \leq c(x, x') \quad \text{for all} \quad x, x' \in \mathcal{X}\}$. 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 ρ, ν_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 Φ_c are convex, we are in a situation where Sion's minimax theorem applies. Therefore,

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

Since

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

we have

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

If we replace φ_i by ψ_i^c and ψ_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 ψ_i^c and $\psi_i^{\bar{c}c}$ are bounded and Lipschitz. Thus, we can restrict the supremum to a compact subset of Φ_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** 23 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 24 *Given measures $\mu_1, \ldots, \mu_K$ and a bounded Lipschitz cost c satisfying the assumptions in **Proposition** 7, suppose that $\lambda, \tilde{\mu}_1, \ldots, \tilde{\mu}_K$ are optimal solutions to* (10)*. If $\varphi_i^*, \psi_i^* \in C_b(\mathcal{X})$ are the optimal Kantorovich potentials for the partial transport of μ_i to λ (c.f **Lemma** 23), then $\varphi_1^*, \ldots, \varphi_K^*$ are optimal solutions to problem* (28)*, $\psi_1^*, \ldots, \psi_K^*$ are optimal solutions to* (27)*, and the values of* (27)-(29) *are equal to* (10)*. 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 λ, then it follows from **Lemma** 23 and the optimality of λ that there exists φ_i^*, ψ_i^* such that

$$\lambda(\mathcal{X}) + \sum_{i\in[K]} C(\mu_i, \tilde{\mu}_i) = \min_{\rho\in\mathcal{B}} \rho(\mathcal{X}) + \sum_{i\in[K]} \int_{\mathcal{X}} \varphi_i^*(x) d\mu_i(x) - \psi_i^*(x') d\rho(x'), \tag{30}$$

$\psi_i^*(x') \geq 0$, 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 $\sum_{i\in[K]} \psi_i^*(x') > 1$, then we can make the right hand side of (30) smaller than the left hand side by choosing $\rho = M\delta_{x'}$ for some sufficiently large value of M. Hence, it follows that $\sum_{i\in[K]} \psi_i^*(x) \leq 1$ everywhere. Thus, the ψ_i^* are feasible solutions to problem (27) and, by **Proposition** 22 $(\psi_i^*)^c = \varphi_i^*$ are feasible solutions to (28). Finally, if we choose $\rho = 0$, it follows that

$$(10) = \lambda(\mathcal{X}) + \sum_{i\in[K]} C(\mu_i, \tilde{\mu}_i) \leq \sum_{i\in[K]} \varphi_i^*(x) d\mu_i(x) \leq (28) = (27) \leq (10)$$

where the second last equality follows from **Proposition** 22 and the last inequality holds trivially by duality. Therefore, we can infer that there is no duality gap. ■

Proposition 25 *Given measures $\mu_1, \ldots, \mu_K$, if c is a cost that satisfies **Assumption** 1, then problems* (27)-(29) *all have the same value as* (10)*.*

Remark 26 *Note that we do not claim that the supremums in* (27)-(29) *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** 7. Since c is lower semicontinuous it is straightforward to check that c_j converges to c pointwise everywhere.

Let α_j and β_j denote the optimal values of Problems (10) and (28) respectively with cost c_j. From **Proposition** 24 we know that $\alpha_j = \beta_j$. Let α, β denote the optimal values of Problems (10) and (28) 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 (28) for the cost c_{j_0}, then it is also a feasible solution to (28) for c. Therefore, $\lim_{j\to\infty} \beta_j \leq \beta$.

On the other hand, let λ^j and $\tilde{\mu}_1^j, \dots, \tilde{\mu}_K^j$ be optimal solutions to (10) with the cost c_j. Let π_i^j be the optimal transport plan between μ_i and $\tilde{\mu}_i^j$. Arguing as in **Proposition** 7, it follows that λ^j and π_i^j are tight with respect to j. Thus, there exists a subsequence (that we do not relabel) such that λ^j converges weakly to some λ and π_i^j converges weakly to some π_i. Fix some j_0 and note that for all $j \geq j_0$

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

Therefore,

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

Taking a supremum over j_0, it follows that

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

Thus, $\alpha \leq \liminf_{j\to\infty} \alpha_j = \liminf_{j\to\infty} \beta_j = \beta$. Thanks to **Proposition** 22, it follows that (10) and (27)-(29), all have the same optimal value. ■

4. Proof of Theorem 6

In this section, we prove **Theorem** 6 and return to the adversarial problem (1).

4.1 Theorem 6: upper bound

First we show that

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

To see this, recall that B_μ^* is, according to **Proposition** 20, equal to

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

Here and in what follows we use Υ_μ to denote the set of positive measures satisfying $\widetilde{\mathcal{P}}_{i\,\#}(r_i\gamma) = \mu_i$ 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 (25) (when $\mu = \widehat{\mu}_{\vec{z}}$). We define a measure γ 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:

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

Let us now compute the cost associated to this γ:

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

Combining the above with *Remark* 4, we conclude that

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

4.2 Theorem 6: lower bound

Now, it is sufficient to show

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

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

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

For each $i \in [K]$, define

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

It is thus clear from the above computation and definition that

$$\sum_{j=1}^{K} \int_{\mathcal{X}\times[K]} \phi_j(z_j) \frac{1}{2\mu(\mathcal{Z})} d\mu(z_j) + \frac{1}{2}\sum_{j=1}^{K} \phi_j(\triangle) = \sum_{i\in[K]} \int_{\mathcal{X}} \psi_i(x_i) \frac{1}{2\mu(\mathcal{Z})} d\mu_i(x_i). \tag{31}$$

Our goal is now to show that $\{\psi_i : i \in [K]\}$ is feasible for problem (28) (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** 28 below.

Lemma 27 *Given $(z_1, \dots, z_K) \in \mathcal{Z}_*^K$, let $A = \{j \in [K] : z_j \neq \triangle\}$. Suppose that for each $j \in A$ $z_j = (x_j, j)$. Then, for each $\phi \in \Phi$,*

$$\sum_{j=1}^{K} \phi_j(z_j) \leq \frac{1}{K} + \frac{1}{K} c_A. \tag{32}$$

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

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

For simplicity, assume that $A = \{1, \dots, S\}$. By **Lemma** 21,

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

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 (24) since $r_s = \frac{1}{K}$ for $1 \leq s \leq S$ and $r_j = 0$ for $j \notin A$. So, (24) 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 $\{\psi_i : i \in [K]\}$ is feasible for (28), that is, we need to show that for any $(x_1, \dots, x_K) \in \mathcal{X}^K$ and for any $A \in S_K$ we have

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

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

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

or, after a rearrangement of the summands:

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

We can bound the first line above using (32):

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

The same argument holds for the second, third and fourth lines. For the last line, notice that $\mathbf{c}(\triangle, \ldots, \triangle) = 0$. 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 ϕ_j's as above we will not be able to obtain the desired upper bound since the total number of $\phi_j(\triangle)$'s available is in this case $K|A| = 18$ while the required number of $\phi_j(\triangle)$'s in the analogous arrangement as above would be at least $K(K - |A|) = 63$. To overcome this problem, we need to rearrange the ϕ_j's further in order to reduce the required number of $\phi_j(\triangle)$'s and deduce from this refined analysis the desired upper bound.

First of all, construct a 9×9 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 ϕ_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 $1 \leq j \leq 9$ is exactly 1. Filling $\emptyset_j$'s with $\phi_j(\triangle)$'s, and using the fact that the number of $\phi_j(\triangle)$'s for each $1 \leq j \leq 9$ is 2, it follows that

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

Observe that for $(z_1, \ldots, z_K) = ((x_1, 1), (x_2, 2), \ldots, (x_1, 1), (x_2, 2), \triangle)$, $\widehat{\mu}_{\vec{z}} = \frac{4}{9}\delta_{(x_1,1)} + \frac{4}{9}\delta_{(x_2,2)}$. Factoring out the 4 (see *Remark* 4) and applying (32), what we obtain is

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

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 $\sum_{j=1}^{9} \phi_j(\triangle) \leq 0$, it follows that

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

The above two situations help us illustrate the general strategy for proving that the resulting ψ_i are feasible: the idea is to arrange summands appropriately so that we can utilize **Lemma** 27 in the most effective way possible. In the following proposition we state precisely our aim and prove it by such strategy.

Proposition 28 *Let $(\phi_1, \ldots, \phi_K) \in \Phi$ be a feasible dual potential. For each $i \in [K]$, define*

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

Then $\{\psi_i : i \in [K]\}$ is feasible for (28).

Proof Fix K and $A \in S_K$. Without loss of generality, assume that $A = \{1, \ldots, S\}$. We need to show that

$$\sum_{i \in A} \psi_i(x_i) \leq 1 + c_A. \tag{33}$$

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

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

Rearranging the sum of the ψ's, it follows that

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

Note that for each $1 \le s \le S$, $|\{u(s,j) : 1 \le j \le K\}| = \frac{K}{S}$, 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 (32),

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

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

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

proving the desired inequality in the 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

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

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 $\phi_{v(1,k)}(x_1, 1), \ldots, \phi_{v(S,k)}(x_S, S)$ 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

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

We split this case into two further subcases.

First, assume that $\lfloor \frac{K}{S} \rfloor = 1$. In this case, we have $K(K-S) \leq KS$. For each $1 \leq k \leq K$, collect all the $\phi_j(\triangle)$'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 $\phi_j(\triangle)$. From the above and an application of (32), we deduce that

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

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 $1 \leq s \leq S$, we merge together the s-th row, the $(S+s)$-th row, ..., 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 ϕ_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 $1 \leq j \leq K$, we would be done since the number of $\phi_j(\triangle)$ is S for each j, whence it would be possible to replace the $\emptyset_j$'s with $\phi_j(\triangle)$'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

$$v(s, \lfloor \frac{K}{S} \rfloor S + 1) = \lfloor \frac{K}{S} \rfloor S + s \text{ if } 1 \leq s \leq K - \lfloor \frac{K}{S} \rfloor S, \tag{34}$$

$$v(s, \lfloor \frac{K}{S} \rfloor S + 1) = \lfloor \frac{K}{S} \rfloor S + s - K \text{ if } K - \lfloor \frac{K}{S} \rfloor S + 1 \leq s \leq S \tag{35}$$

and

$$v(s, K) = K \text{ if } s = 1, \tag{36}$$

$$v(s, K) = s - 1 \text{ if } 2 \leq s \leq S. \tag{37}$$

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

(i) $\lfloor \frac{K}{S} \rfloor S + 1 \leq j \leq K$: By (34), if $1 \leq s \leq K - \lfloor \frac{K}{S} \rfloor S$, then the s-th row has $\emptyset_j$ for $\lfloor \frac{K}{S} \rfloor S + s \leq j \leq K$. Also, by (35) and (37), if $K - \lfloor \frac{K}{S} \rfloor S + 1 \leq s \leq S$, then no merged row has such $\emptyset_j$. Hence, the number of $\emptyset_j$ is $j - \lfloor \frac{K}{S} \rfloor S$.

(ii) $S \leq j \leq \lfloor \frac{K}{S} \rfloor S$: It follows from (34) and (35) 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 (36) and (37) 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) $S - (K - \lfloor \frac{K}{S} \rfloor S) + 1 \leq j \leq S - 1$: By (35) and (37), if $S - (K - \lfloor \frac{K}{S} \rfloor S) + 1 \leq j \leq S - 1$, 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) $1 \leq j \leq S - (K - \lfloor \frac{K}{S} \rfloor S)$: Similar to (iii), if $1 \leq j \leq S - (K - \lfloor \frac{K}{S} \rfloor S)$, 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

$$\text{the number of } \emptyset_j = \begin{cases} j - \lfloor \frac{K}{S} \rfloor S & \text{for } \lfloor \frac{K}{S} \rfloor S + 1 \leq j \leq K, \\ 0 & \text{for } S \leq j \leq \lfloor \frac{K}{S} \rfloor S, \\ S - j & \text{for } S - (K - \lfloor \frac{K}{S} \rfloor S) + 1 \leq j \leq S - 1, \\ K - \lfloor \frac{K}{S} \rfloor S & \text{for } 1 \leq j \leq S - (K - \lfloor \frac{K}{S} \rfloor S). \end{cases} \tag{38}$$

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 $\lfloor \frac{K}{S} \rfloor S + 1 \leq k \leq K$, the k-th row has

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

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

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

(iii) $S - (K - \lfloor \frac{K}{S} \rfloor S) + 1 \leq j \leq S - 1$: Since $\lfloor \frac{K}{S} \rfloor S + 1 \leq k \leq K$, if $v(S,k) = S + k - K < j$, then $j \notin A_k$. This yields that if $\lfloor \frac{K}{S} \rfloor S + 1 \leq k \leq K - S + j$, 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) $1 \leq j \leq S - (K - \lfloor \frac{K}{S} \rfloor S)$: Since $v(S, \lfloor \frac{K}{S} \rfloor S + 1) = S - (K - \lfloor \frac{K}{S} \rfloor S)$, if $1 \leq j \leq S - (K - \lfloor \frac{K}{S} \rfloor S)$ and $\lfloor \frac{K}{S} \rfloor S + 1 \leq k \leq K$, then $j \in A_k$. Hence, the number of $\emptyset_j$ is 0.

To summarize, in the remaining original rows

$$\text{the number of } \emptyset_j = \begin{cases} K-j & \text{for } \lfloor \frac{K}{S} \rfloor S + 1 \leq j \leq K, \\ K - \lfloor \frac{K}{S} \rfloor S & \text{for } S \leq j \leq \lfloor \frac{K}{S} \rfloor S, \\ K - \lfloor \frac{K}{S} \rfloor S - S + j & \text{for } S - (K - \lfloor \frac{K}{S} \rfloor S) + 1 \leq j \leq S-1, \\ 0 & \text{for } 1 \leq j \leq S - (K - \lfloor \frac{K}{S} \rfloor S). \end{cases} \tag{39}$$

Combining (38) with (39), 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 replace every $\emptyset_j$ with a $\phi_j(\triangle)$. Accordingly, using $\sum \phi_j(\triangle) \leq 0$, we deduce that

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

Let's focus on the first summation over merged rows. Notice that there are $\lfloor \frac{K}{S} \rfloor S$ many non-empty elements and the set of arguments of such non-empty elements is $\{(x_1, 1), \ldots, (x_S, S)\}$. 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 (32), we obtain

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

On the other hand, for the second summation over remaining rows, there are S many non-empty elements. Thus,

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

(32) immediately implies

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

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 $\phi = (\phi_1, \ldots, \phi_K) \in \Phi$, its associated $(\psi_1, \ldots, \psi_K)$ (which satisfies (31)) is feasible for (28). Consequently, this leads to

$$(14) \leq \frac{1}{2\mu(\mathcal{Z})}(28) \tag{40}$$

In turn, by the equivalence between (28) and (10) by **Proposition** 25, this automatically implies that

$$(14) \leq \frac{1}{2\mu(\mathcal{Z})} B^*_\mu.$$

Finally, combining with **Corollary** 31 below (which establishes that under **Assumption** 1 there is no duality gap for the MOT problem (2)) we obtain the desired inequality relating the minimum value for the MOT problem and B^*_μ.

4.3 Returning to the adversarial problem (1)

We begin by establishing that, under **Assumption** 1, the cost $\mathbf{c}$ is lower semi-continuous with respect to a suitable notion of convergence.

Proposition 29 *Let $\mathcal{Z}_* = \mathcal{Z} \cup \{\triangle\}$ on which $\triangle$ is considered as an isolated point. Let $\widehat{d}$ be defined according to:*

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

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

$$\widehat{d}_K((z_1, \ldots, z_K), (z'_1, \ldots, z'_K)) := \max_{i \in [K]} \widehat{d}(z_i, z'_i).$$

Recall

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

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

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

*Under **Assumption** 1, $\mathbf{c}$ is lower semi-continuous on $(\mathcal{Z}_*^K, \widehat{d}_K)$.*

Remark 30 *Note that $(\mathcal{Z}_*^K, \widehat{d}_K)$ is still a Polish space.*

Proof Suppose $\vec{z}^n = (z_1^n, \ldots, z_K^n)$ converges to $\vec{z} = (z_1, \ldots, z_K)$ in $(\mathcal{Z}_*^K, \widehat{d}_k)$. Without loss of generality, assume that $z_1, \ldots, z_L = \triangle$ 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 $z_1^n, \ldots, z_L^n = \triangle$ 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 **Lemma** 21 we have

$$\mathbf{c}(z_1^n, , \dots, z_K^n) = B^*_{\widehat{\mu}_{\vec{z}^n}} = \inf_{m: S_K \to \mathbb{R}} \sum_{A \subseteq \{L+1,\dots,K\}} m_A\big(c_A(x_{L+1}^n, \dots, x_K^n) + 1\big), \tag{41}$$

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

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

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 $\mathbf{c}(z_1, \dots, z_K)$ and thus, using the lower semicontinuity of c_A established earlier, we deduce:

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

■

Corollary 31 *(Duality of MOT) Under* ***Assumption*** *1,*

$$\begin{aligned}&\inf_{\pi\in\Pi_K(\mu)} \int_{\mathcal{Z}_*{}^K} \mathbf{c}(z_1, \dots, z_K) d\pi(z_1, \dots, z_K) \\ =&\sup_{\phi\in\Phi} \left\{ \sum_{j=1}^K \int_{\mathcal{X}\times[K]} \phi_j(z_j) \frac{1}{2\mu(\mathcal{Z})} d\mu(z_j) + \frac{1}{2} \sum_{j=1}^K \phi_j(\triangle) \right\}.\end{aligned}$$

Furthermore, a minimizer π^* exists, hence the infimum is indeed the minimum.

Proof From **Proposition** 29 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 Villani (2003), 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 π^* follows from the lower semi-continuity of $\mathbf{c}(z_1, \ldots, z_K)$ and the compactness of $\Pi_K(\mu)$. ■

Corollary 32 *Under **Assumption** 1,* (8)=(9).

Proof By the upper bound from section 4.1 we have

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

On the other hand, from (40) and **Corollary** 31 we have

$$\min_{\pi \in \Pi_K(\mu)} \int \mathbf{c}(z_1, \ldots, z_K) d\pi(z_1, \ldots, z_K) = (14) \leq \frac{1}{2\mu(\mathcal{Z})}(28) \leq \frac{1}{2\mu(\mathcal{Z})} B_\mu^*.$$

Combining these two inequalities we conclude that all the above terms must be equal. In particular, $(28) = B_\mu^*$. Finally, by **Proposition** 22 we know that (28) = (27) = (8). In particular, $(9) = B_\mu^* = (8)$. ■

Corollary 33 *Suppose that **Assumption** 1 holds and that (π^*, ϕ^*) is a solution pair for the MOT problem and its dual. Define f^* and $\widetilde{\mu}^*$ according to:*

$$f_i^*(\widetilde{x}) := \sup_{x \in \mathrm{spt}(\mu_i)} \left\{ \max \left\{ \sum_{j=1}^K \phi_j^*(x, i) + \sum_{j=1}^K \phi_j^*(\triangle), 0 \right\} - c(x, \widetilde{x}) \right\}$$

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 $\vec{z} = (z_1, \ldots, z_K)$. Suppose f^* is measurable. Then $(f^*, \widetilde{\mu}^*)$ is a saddle for problem (1).

Remark 34 *Here, we do not claim that f^* is in general measurable. However, if either c is continuous or μ is an empirical measure with a finite support, then f^* can be shown to be measurable. See **Remark** 5.5 and **Remark** 5.11 in Villani (2009).*

Notice that the supremum in the definition of f_i^*, is only taken over $\mathrm{spt}(\mu_i)$.

Proof We will show that $(f^*, \widetilde{\mu}^*)$ is a saddle point for problem (9). More explicitly, we show that for any $f \in \mathcal{F}$ and for any $\widetilde{\mu}$,

$$B(f, \widetilde{\mu}^*) + C(\mu, \widetilde{\mu}^*) \leq B(f^*, \widetilde{\mu}^*) + C(\mu, \widetilde{\mu}^*) \leq B(f^*, \widetilde{\mu}) + C(\mu, \widetilde{\mu}). \tag{42}$$

First we compute $B(f^*, \widetilde{\mu}^*) + C(\mu, \widetilde{\mu}^*)$. Notice that

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

where λ_A^* and π_A^* correspond to $\widetilde{\mu}^*$. By the construction of f_i^* and (6),

$$\begin{aligned} \sum_{i \in A} f_i^*(T_A(\vec{x})) &= \sum_{i \in A} \sup_{x'} \left\{ \max \left\{ \sum_{j=1}^K \phi_j^*(x', i) + \sum_{j=1}^K \phi_j^*(\triangle), 0 \right\} - c(x', T_A(\vec{x})) \right\} \\ &= \max \left\{ \sup_{x_i' : i \in A} \left\{ \sum_{i \in A} \big(\sum_{j=1}^K \phi_j^*(x_i', i) + \sum_{j=1}^K \phi_j^*(\triangle) \big) - c_A(x_i' : i \in A) \right\}, 0 \right\} \\ &\leq \max \left\{ \sup \left\{ 1 + c_A(x_i' : i \in A) - c_A(x_i' : i \in A) \right\}, 0 \right\} \\ &\leq 1, \end{aligned}$$

where the third inequality follows from (33). Hence,

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

On the other hand, the definition of f_i^* implies that for any x_i in the support of μ_i we have

$$f_i^*(\widetilde{x}_i) \geq \sum_{j=1}^K \phi_j^*(x_i, i) + \sum_{j=1}^K \phi_j^*(\triangle) - c(\widetilde{x}_i, x_i). \tag{43}$$

Using $\sum_{A\in S_K(i)}\mu_{i,A}=\mu_i$ and (43), the optimality of ϕ^* implies that

$$\begin{aligned}
B(f^*,\widetilde{\mu}^*)+C(\mu,\widetilde{\mu}^*) &= \sum_{A\in S_K}\sum_{i\in A}\left\{\int_{\mathcal{X}\times\mathcal{X}}\big(f_i^*(\widetilde{x}_i)+c(\widetilde{x}_i,x_i)\big)d\pi_i^*(\widetilde{x}_i,x_i)\right\}\\
&\geq \sum_{A\in S_K}\sum_{i\in A}\left\{\int_{\mathcal{X}\times\mathcal{X}}\big(\sum_{j=1}^K\phi_j^*(x_i,i)+\sum_{j=1}^K\phi_j^*(\triangle)\big)d\pi_i^*(\widetilde{x}_i,x_i)\right\}\\
&= \sum_{A\in S_K}\sum_{i\in A}\left\{\int_{\mathcal{X}}\big(\sum_{j=1}^K\phi_j^*(x_i,i)+\sum_{j=1}^K\phi_j^*(\triangle)\big)d\mu_{i,A}(x_i)\right\}\\
&= \sum_{j=1}^K\int_{\mathcal{Z}}\phi_j(z_j)d\mu(z_j)+\mu(\mathcal{Z})\sum_{j=1}^K\phi_j(\triangle)\\
&= B_\mu^*.
\end{aligned}$$

Here π_i^* denotes an optimal coupling between μ_i and $\widetilde{\mu}_i^*$ which correspond to π_A^*'s. From the above we infer that

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

Now we can prove (42). The first inequality of (42) is straightforward, since the definition of B_μ^* in (10) and the optimality of $\widetilde{\mu}^*$ imply that

$$B(f,\widetilde{\mu}^*)+C(\mu,\widetilde{\mu}^*)\leq\sup_{f\in\mathcal{F}}\{B(f,\widetilde{\mu}^*)+C(\mu,\widetilde{\mu}^*)\}=B_\mu^*=B(f^*,\widetilde{\mu}^*)+C(\mu,\widetilde{\mu}^*).$$

For the second inequality of (42), let arbitrary $\widetilde{\mu}$ be fixed and $\pi_i\in\Gamma(\widetilde{\mu}_i,\mu_i)$ be an optimal coupling for each $i\in[K]$. Then,

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

Applying (43) yields that

$$\begin{aligned}
B(f^*,\widetilde{\mu})+C(\mu,\widetilde{\mu}) &\geq \sum_{i\in[K]}\int_{\mathcal{X}\times\mathcal{X}}\left(\sum_{j=1}^K\phi_j^*(x,i)+\sum_{j=1}^K\phi_j^*(\triangle)\right)d\pi_i(x,\widetilde{x})\\
&= \sum_{i\in[K]}\int_{\mathcal{X}\times\mathcal{X}}\left(\sum_{j=1}^K\phi_j^*(x,i)+\sum_{j=1}^K\phi_j^*(\triangle)\right)d\mu_i(x)\\
&= B_\mu^*\\
&= B(f^*,\widetilde{\mu}^*)+C(\mu,\widetilde{\mu}^*).
\end{aligned}$$

Therefore, $(f^*,\widetilde{\mu}^*)$ is a saddle point for (9), hence for (8) and (1) also. ■

Remark 35 *Many recent papers have tried to analyze adversarial learning from a game-theoretic perspective Bose, Gidel, Berard, Cianflone, Vincent, Lacoste-Julien, and Hamilton (2020); Meunier, Scetbon, Pinot, Atif, and Chevaleyre (2021); Pydi and Jog (2021b). This approach is natural: the learner aims at maximizing the classification power B^*_μ while the adversary aims at maximizing the loss R^*_μ(hence to minimize B^*_μ): this is a standard zero-sum game. Our main results thus provide a way to build Nash equilibria for the adversarial problem using a series of equivalent formulations taking the form of generalized barycenter problems or MOTs.*

Corollary 36 *Let π^* be a solution of the MOT problem (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 π^* 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 π^* are the same. ■

5. Examples and Numerical experiments

Through this section, the cost c is as in **Example** 1. This cost has been widely used in adversarial learning literature and distributional robust optimization literature. Examples in this section illuminate how our general framework of generalized barycenter and MOT finds 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 García Trillos and Murray (2022).

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

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

In García Trillos and Murray (2022), it is proved that

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

where

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

In other words, in the binary case, it is unnecessary to introduce the element $\triangle$. 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:

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

where $\varphi : \mathcal{Z}_* \times \mathcal{Z}_* \to \mathbb{R}$ is an arbitrary test function. The cost associated to the induced π 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 π be a solution for the MOT problem (2) (such a solution exists thanks to **Proposition** 31). Thanks to **Corollary** 36, 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:

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

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 García Trillos and Murray (2022) for the binary case.

5.2 Toy example: three points distribution

Let's assume that $K = 3$ and μ 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** 1 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 $\mathrm{spt}(\widetilde{\mu}_i) \subseteq \overline{B}(x_i, \varepsilon)$. 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) can not disambiguate the values of f outside of this set). We consider 4 non-trivial configurations and one trivial one. Figure 4 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). This is straightforward, since $\mathrm{spt}(\widetilde{\mu}_i) \cap \mathrm{spt}(\widetilde{\mu}_j) = \emptyset$, and thus it can be deduced that $(\mathbb{1}_{\overline{B}(x_1,\varepsilon)}, \mathbb{1}_{\overline{B}(x_2,\varepsilon)}, \mathbb{1}_{\overline{B}(x_3,\varepsilon)})$ 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 $d(x_1, x_2) \leq 2\varepsilon$, $d(x_1, x_3) > 2\varepsilon$ 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 $W_\infty(\widetilde{\mu}_3, \omega_3 \delta_{x_3}) \leq \varepsilon$. We claim that $(\widehat{f}, \widehat{\mu})$ is a saddle point. For any $(f_1, f_2, f_3) \in \mathcal{F}$ we have

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

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. $d(x_i, x_j) \leq 2\varepsilon$ 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 $K = 2$, $d(x_1, x_2) \leq 2\varepsilon$ if and only if $\overline{B}(x_1,\varepsilon) \cap \overline{B}(x_2,\varepsilon) \neq \emptyset$. However, when $K \geq 3$, 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 $\alpha_i \in [0, \omega_i]$ for all $1 \leq i \leq 3$ such that

$$\alpha_1 = \omega_2 - \alpha_2, \;\; \alpha_2 = \omega_3 - \alpha_3 \text{ and } \alpha_3 = \omega_1 - \alpha_1.$$

Precisely,

$$\alpha_1 = \frac{\omega_1 + \omega_2 - \omega_3}{2}, \;\; \alpha_2 = \frac{\omega_2 + \omega_3 - \omega_1}{2}, \text{ and } \alpha_3 = \frac{\omega_3 + \omega_1 - \omega_2}{2}.$$

Note that for all i, $\alpha_i \geq 0$ since $\omega_1 \leq \omega_2 + \omega_3$. 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

$$\begin{aligned} \widehat{\mu}_1 &:= \big(\alpha_1 \delta_{\overline{x}_{12}} + (\omega_1 - \alpha_1)\delta_{\overline{x}_{13}}\big) = \big((\frac{\omega_1 + \omega_2 - \omega_3}{2})\delta_{\overline{x}_{12}} + (\frac{\omega_1 - \omega_2 + \omega_3}{2})\delta_{\overline{x}_{13}}\big), \\ \widehat{\mu}_2 &:= \big(\alpha_2 \delta_{\overline{x}_{23}} + (\omega_2 - \alpha_2)\delta_{\overline{x}_{12}}\big) = \big((\frac{\omega_2 + \omega_3 - \omega_1}{2})\delta_{\overline{x}_{23}} + (\frac{\omega_2 - \omega_3 + \omega_1}{2})\delta_{\overline{x}_{12}}\big), \\ \widehat{\mu}_3 &:= \big(\alpha_3 \delta_{\overline{x}_{13}} + (\omega_3 - \alpha_3)\delta_{\overline{x}_{23}}\big) = \big((\frac{\omega_3 + \omega_1 - \omega_2}{2})\delta_{\overline{x}_{13}} + (\frac{\omega_3 - \omega_1 + \omega_2}{2})\delta_{\overline{x}_{23}}\big). \end{aligned}$$

Observe that at each $\overline{x}_{ij}$, $\widehat{\mu}_i$ and $\widehat{\mu}_j$ put the same mass: it is natural since, otherwise, the learner will choose a class which puts more mass at $\overline{x}_{ij}$. So, this gives a hint about what would be the best adversarial attack. The adversary gathers classes as much as possible and distributes them as uniform as possible.

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 $d(x_i, x_j) \leq 2\varepsilon$ for any x_i, x_j but $\overline{B}(x_1,\varepsilon) \cap \overline{B}(x_2,\varepsilon) \cap \overline{B}(x_3,\varepsilon) = \emptyset$, $\overline{B}(x_i,\varepsilon) = A_{ij} \dot{\cup} A_{ik} \dot{\cup} A_i$ for each i. Here $\dot{\cup}$ denotes a disjoint union. Also, since $W_\infty(\widetilde{\mu}_i, \omega_i \delta_{x_i}) \leq \varepsilon$, it must be the case that $A_{ij} \cap \mathrm{spt}(\widetilde{\mu}_k) = \emptyset$ if $k \neq i, j$. For each $1 \leq i \leq 3$, construct the following weak partition:

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

$\widehat{f}$ is a weak partition since $\overline{B}(x_i,\varepsilon) = A_{ij} \dot{\cup} A_{ik} \dot{\cup} A_i$ and $\overline{B}(x_1,\varepsilon) \cap \overline{B}(x_2,\varepsilon) \cap \overline{B}(x_3,\varepsilon) = \emptyset$. We claim that $(\widehat{f}, \widehat{\mu})$ is a saddle point. Note that a straightforward computation yields $B_\mu(\widehat{f}, \widehat{\mu}) = \frac{1}{2}$.

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

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

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

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

Note that since $W_\infty(\widetilde{\mu}_i, \omega_i \delta_{x_i}) \leq \varepsilon$, $\mathrm{spt}(\widetilde{\mu}_i) \cap A_j = \emptyset$ for any $\widetilde{\mu}_i$ and for any $i \neq j$. To minimize the above, the adversary should put $\mathrm{spt}(\widetilde{\mu}_i) \subseteq A_{ij} \cup A_{ik}$ for all i. Also, at the minimum, it must be the case that $\widetilde{\mu}_i(A_{ij}) = \widetilde{\mu}_j(A_{ij})$, otherwise the adversary would be able decrease the classification power further. Combining all arguments, we can deduce

$$B_\mu(\widetilde{f}, \widetilde{\mu}) \geq \frac{\widetilde{\mu}_1(A_{12}) + \widetilde{\mu}_2(A_{12})}{2} + \frac{\widetilde{\mu}_1(A_{13}) + \widetilde{\mu}_3(A_{13})}{2} + \frac{\widetilde{\mu}_2(A_{23}) + \widetilde{\mu}_3(A_{23})}{2} = \frac{1}{2},$$

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 ω_1 since the learner can only detect class 1 perfectly and always misclassifies 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. Note that a straightforward computation yields $B_\mu(\widehat{f}, \widehat{\mu}) = \omega_1$.

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, $B_\mu^* = \omega_1$. Here, $\omega_1 \geq \frac{1}{2}$, since $\omega_1 \geq \omega_2 + \omega_3$ and $\sum \omega_i = 1$. In the case that $\omega_1 = \omega_2 + \omega_3$, both **Case 4 -(i)** and **Case 4 -(ii)** provide $B_\mu^* = \frac{1}{2}$, which shows the consistency.

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 η

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 μ and not just by the geometry of its support. In fact, the optimal adversarial attacks $\widetilde{\mu}$ and the optimal barycenter measure λ depend on not only the geometry of the support of μ but also the magnitudes of its marginals, μ_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 perspective of numeric, our aim is to solve the MOT problem (2) and its dual for an empirical measure μ whose support consists of n data points. We use Sinkhorn algorithm for concreteness. Introduced in Cuturi (2013), Sinkhorn algorithm has been one of the central algorithmic tools in computational optimal transport in the past decade. This algorithm, originally introduced in the context of standard (2-marginal) optimal transport problems, was extended to MOTs in Benamou, Carlier, Cuturi, Nenna, and Peyré (2015); Benamou, Carlier, and Nenna (2019). Works that study the computational complexity of generic MOT problems include: Lin, Ho, Cuturi, and Jordan (2019); Tupitsa, Dvurechensky, Gasnikov, and Uribe (2020); Haasler, Ringh, Chen, and Karlsson (2021); Carlier (2022). In particular, Lin, Ho, Cuturi, and Jordan (2019) and Tupitsa, Dvurechensky, Gasnikov, and Uribe (2020) prove the complexity of MOT Sinkhorn algorithm to be $\tilde{O}(K^3 n^K \epsilon^{-2})$ and $\tilde{O}(K^3 n^{K+1}\epsilon^{-1})$, respectively, where ϵ is the error tolerance.

In our first illustration, we consider a data set $(x_1, y_1), \ldots (x_n, y_n)$ 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** 1 with d the Euclidean distance in $\mathbb{R}^2$ and different values of ε. In Figure 5 we show the labels assigned to the data by the (approximate) robust classifier, which we

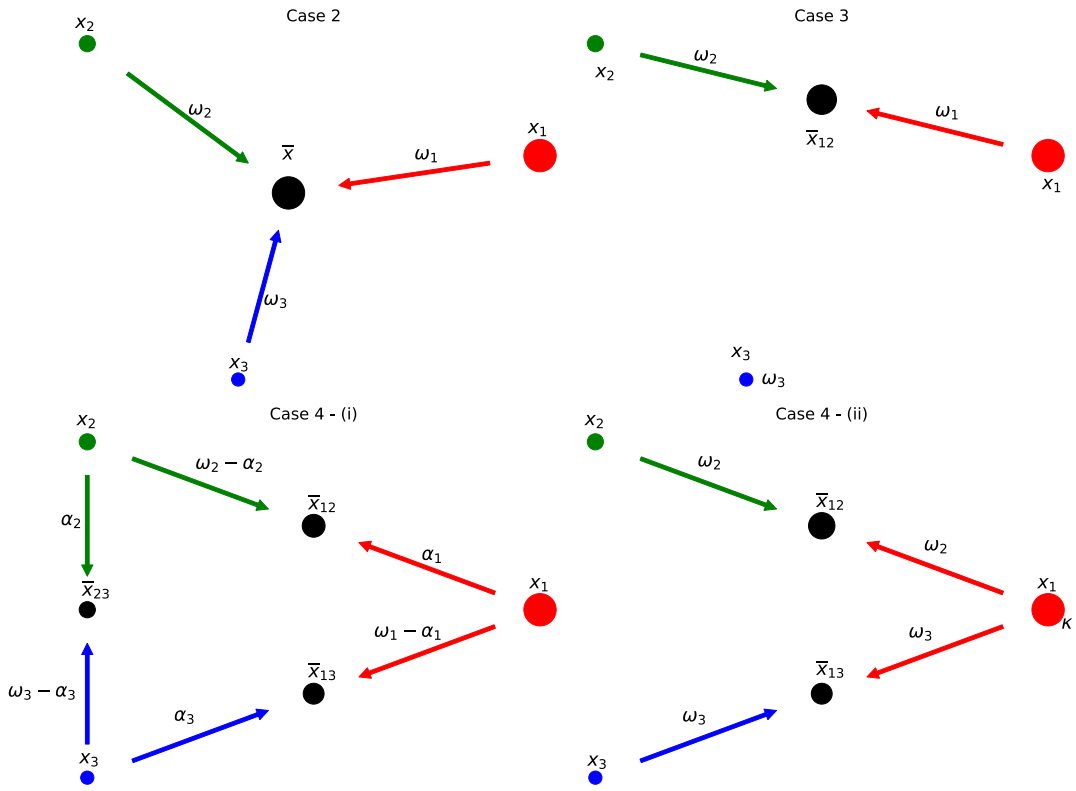

Figure 4: Illustrations of the adversarial attacks in all cases from section 5.2. Weights on arrows indicate the amount of mass the adversary moves to a perturbed point. $\overline{x}$'s are the support of λ in (10). One observes that the support of λ depends on both the geometry of data distributions and their magnitudes.

computed using **Corollary** 33 for the dual potentials ϕ_j generated by the MOT Sinkhorn algorithm.

In our second illustration, we use the mutlimarginal version of Sinkhorn algorithm to compute the adversarial risk R^*_μ (i.e. the optimal value of (1)) for μ 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** 1 for different values of ε and two choices of d: the Euclidean distance ℓ^2 and the ℓ^∞ distance. The results are shown in Figure 6. 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 ℓ^2 distance, the adversarial risk with ℓ^∞ distance varies dramatically as ε grows. This observation is consistent with the findings in Pydi and Jog (2021a) for the binary case.

We emphasize that Figure 6 only provides approximations of the true adversarial risk R^*_μ. Indeed, recall that $R^*_\mu = 1 - B^*_\mu$. Approximating B^*_μ using the MOT Sinkhorn algorithm

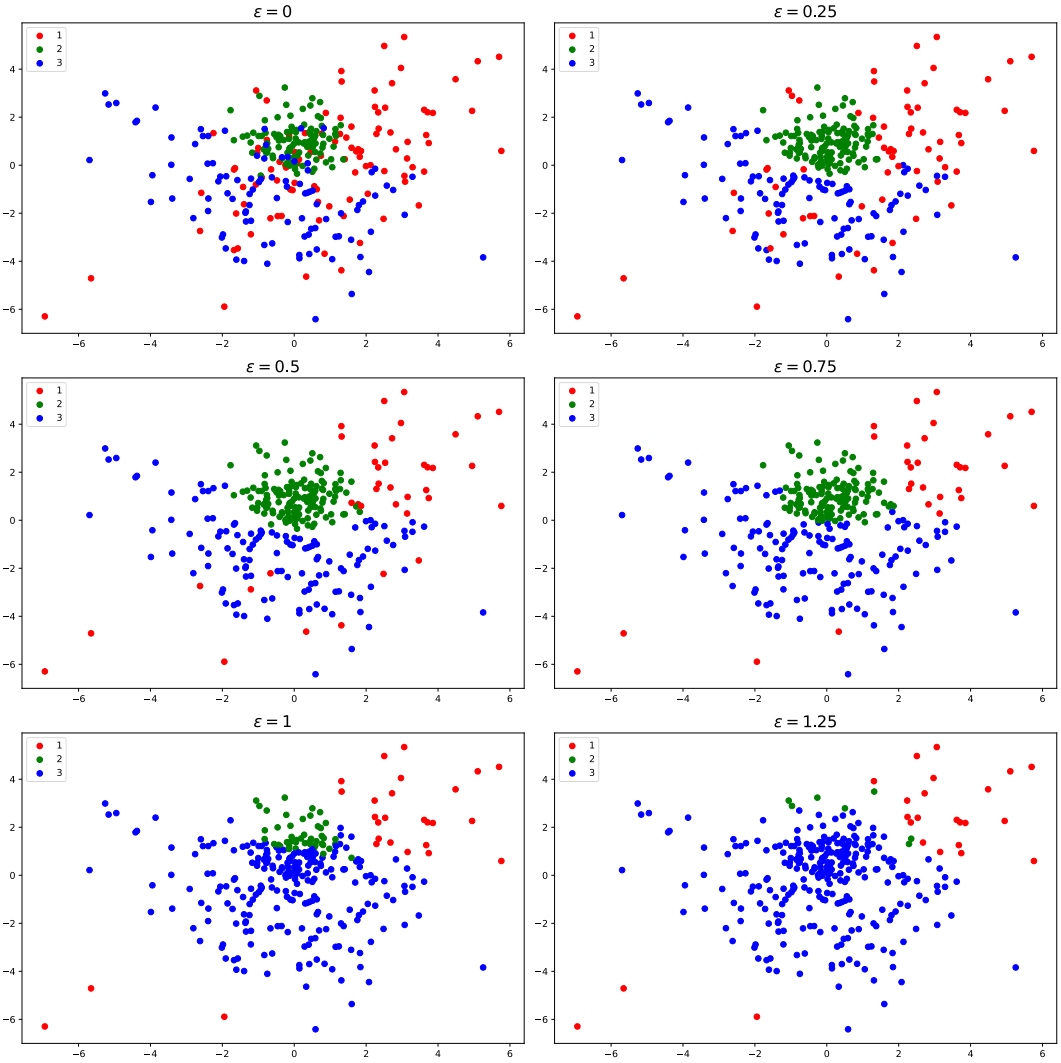

Figure 5: Three Gaussians in $\mathbb{R}^2$. One can observe that as ε grows the robust classifying rule becomes simpler, as expected.

will always produce an upper bound for B_μ^* since the regularization term effectively restricts the solution space of (2). Thus, the multimarginal Sinkhorn algorithm always yields a lower bound for the true R_μ^*. Of course, one can always compute a tighter lower bound by reducing the regularization parameter η 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, Altschuler and Boix-Adserà (2022) prove certain computational hardness of the barycenter problem in the dimension of the feature space (here $\mathcal{X}$). On the other hand, Altschuler and Boix-Adsera (2021) present an algorithm that can get an approximate solution of the optimal

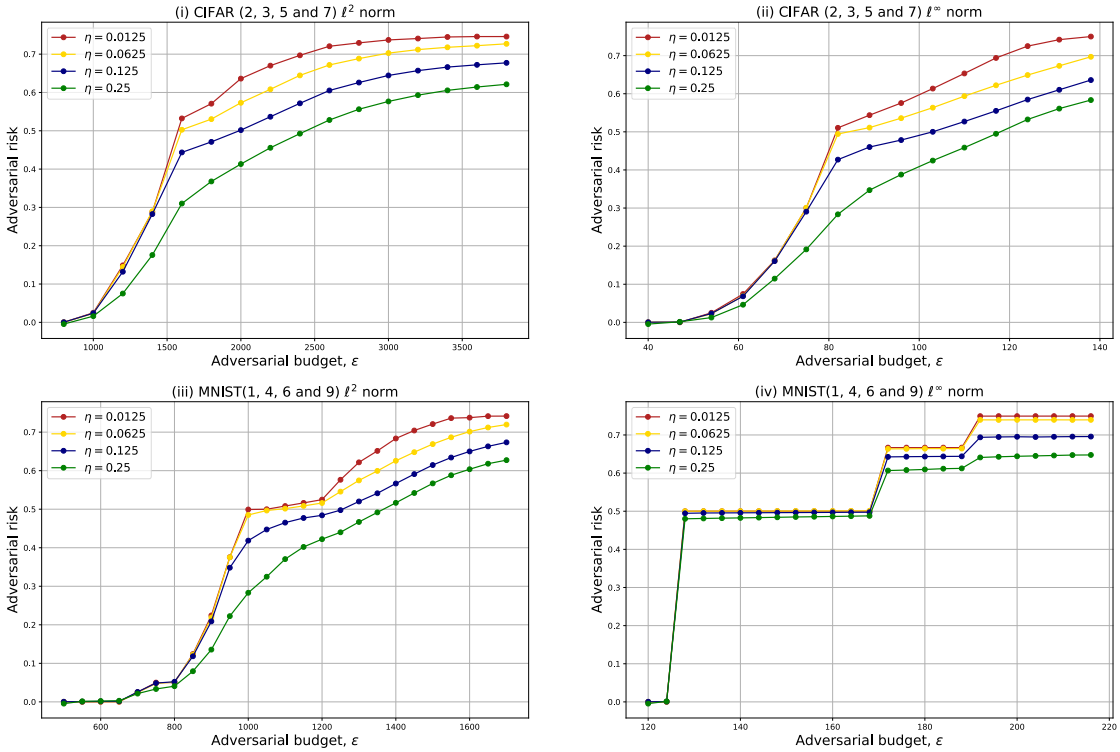

Figure 6: Adversarial risks (1) computed using the multimarginal Sinkhorn algorithm. η is the entropic regularization parameter of the Sinkhorn algorithm. The maximum adversarial risk in all cases is 0.75 because we consider 4 classes and an equal number of points in each class. Due to the entropic penalty, the multimarginal Sinkhorn algorithm always gives an upper bound for the optimal classification power B^*_μ, hence gives a lower bound for the adversarial risk R^*_μ.

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.

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 problem. 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 (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** 1), then the effective size of problem (2) will be considerably smaller than the size of the worst case setting —see the reformulation (21).

Second, it would be of interest to use (1) to help in the training of robust classifiers within specific families of models. Notice that (1) is model free from the perspective of the learner, but in applications practitioners may be interested in solving a problem like:

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

which differs from (1) in the family of classifiers $\mathcal{G}$, which may be strictly smaller than $\mathcal{F}$; for example, $\mathcal{G}$ could be a family of neural networks, kernel-based classifiers, or other popular (parametric) models. There are two ways in which problem (1) is still meaningful for the above model-specific problem: 1) the optimal $\widetilde{\mu}^*$ computed from the problem (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) can serve as a benchmark for robust training within *any* family of models.

Third, more theoretical understanding of optimal dual potentials and robust classifiers is required. As stated in **Corollary** 33, an optimal robust classifier can be obtained from solutions to (14), or, equivalently, from (28) and (29). However, unless c is continuous, even the existence of Borel measurable dual potentials is not guaranteed, and hence neither is the existence of optimal robust classifiers. At this stage, it is thus necessary to *assume* that the classifier intorduced in **Corollary** 33 is Borel measurable. This measurability issue, i.e., that a robust classifier may not be Borel measurable, has been discussed not only in the adversarial training community Pydi and Jog (2021b); Awasthi, Frank, and Mohri (2021a,b); Frank (2022); Frank and Niles-Weed (2022), but also more generally in the *distributional robust optimization* community, e.g., Blanchet and Murthy (2019). In general, at this point only the existence of *universally* measurable robust classifier f^* can be guaranteed. Whether there exist Borel measurable robust classifiers for discontinuous costs (like the one in Example (1)) is a question that we hope to explore in future work.

Finally, it is of interest to investigate the geometric content that profiles like the ones presented in Figure 6 carry about a specific data set. As illustrated in Figure 6, 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.

Acknowledgments

All authors contributed equally and their names are listed in alphabetic order. The authors would like to thank the editor and a few anonymous reviewers for their suggestions to improve the present manuscript. 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.

References

Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, 43(2):904–924, 2011. ISSN 0036-1410. doi: 10.1137/100805741.

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

Jason M Altschuler and Enric Boix-Adserà. Wasserstein barycenters are np-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, 2022.

Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. On the existence of the adversarial bayes classifier. *Advances in Neural Information Processing Systems*, 34:2978–2990, 2021a.

Pranjal Awasthi, Natalie S Frank, and Mehryar Mohri. On the existence of the adversarial bayes classifier (extended version). *arXiv preprint arXiv:2112.01694*, 2021b.

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

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(2):A1111–A1138, 2015.

Jean-David Benamou, Guillaume Carlier, and Luca Nenna. Generalized incompressible flows, multi-marginal transport and sinkhorn algorithm. *Numerische Mathematik*, 142 (1):33–54, 2019.

Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/02bf86214e264535e3412283e817deaa-Paper.pdf`.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Math. Oper. Res.*, 44(2):565–600, 2019. ISSN 0364-765X. doi: 10.1287/moor.2018.0936.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probab.*, 56(3):830–857, 2019. ISSN 0021-9002. doi: 10.1017/jpr.2019.49.

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*, 33:8921–8934, 2020.

Giuseppe Buttazzo, Luigi De Pascale, and Paola Gori-Giorgi. Optimal-transport formulation of electronic density-functional theory. *Physical Review A*, 85(6):062502, 2012.

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

G. Carlier and I. Ekeland. Matching for teams. *Economic Theory*, 42:397–418, 2010.

Guillaume Carlier. On the linear convergence of the multimarginal sinkhorn algorithm. *SIAM Journal on Optimization*, 32(2):786–794, 2022. doi: 10.1137/21M1410634. URL `https://doi.org/10.1137/21M1410634`.

Guillaume Carlier, Adam Oberman, and Edouard Oudet. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM Math. Model. Numer. Anal.*, 49(6):1621–1642, 2015. ISSN 0764-583X. doi: 10.1051/m2an/2015033.

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(2):317–354, 2010. ISSN 0938-2259. doi: 10.1007/s00199-009-0455-z.

Pierre-André Chiappori, Robert J. McCann, and Brendan Pass. Multi- to one-dimensional optimal transport. *Comm. Pure Appl. Math.*, 70(12):2405–2444, 2017. ISSN 0010-3640. doi: 10.1002/cpa.21707.

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

Maria Colombo, Luigi De Pascale, and Simone Di Marino. Multimarginal optimal transport maps for one-dimensional repulsive costs. *Canad. J. Math.*, 67(2):350–368, 2015. ISSN 0008-414X. doi: 10.4153/CJM-2014-011-x.

Codina Cotar, Gero Friesecke, and Claudia Klüppelberg. Density functional theory and optimal transportation with Coulomb cost. *Comm. Pure Appl. Math.*, 66(4):548–599, 2013. ISSN 0010-3640. doi: 10.1002/cpa.21437.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL `https://proceedings.neurips.cc/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf`.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

Julie Delon and Agnès Desolneux. A wasserstein-type distance in the space of gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

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(2):1–28, 2020.

Ivar Ekeland. An optimal matching problem. *ESAIM Control Optim. Calc. Var.*, 11(1): 57–71, 2005. ISSN 1292-8119. doi: 10.1051/cocv:2004034.

Natalie S Frank. Existence and minimax theorems for adversarial surrogate risks in binary classification. *arXiv preprint arXiv:2206.09098*, 2022.

Natalie S Frank and Jonathan Niles-Weed. The consistency of adversarial training for binary classification. *arXiv preprint arXiv:2206.09099*, 2022.

Wilfrid Gangbo and Andrzej Święch. Optimal maps for the multidimensional monge-kantorovich problem. *Communications on Pure and Applied Mathematics*, 51:23–45, 1998.

Nicolás García Trillos and Ryan W. Murray. Adversarial classification: Necessary conditions and geometric flows. *Journal of Machine Learning Research*, 23(187):1–38, 2022. URL `http://jmlr.org/papers/v23/21-0222.html`.

Isabel Haasler, Axel Ringh, Yongxin Chen, and Johan Karlsson. Multimarginal optimal transport with a tree-structured cost and the schrödinger bridge problem. *SIAM Journal on Control and Optimization*, 59(4):2428–2453, 2021.

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

Jun Kitagawa and Brendan Pass. The multi-marginal optimal partial transport problem. *Forum Math. Sigma*, 3:Paper No. e17, 28, 2015. doi: 10.1017/fms.2015.20.

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

Christian B Mendl and Lin Lin. Kantorovich dual solution for strictly correlated electrons in atoms and molecules. *Physical Review B*, 87(12):125106, 2013.

Laurent Meunier, Meyer Scetbon, Rafael B Pinot, Jamal Atif, and Yann Chevaleyre. Mixed nash equilibria in the adversarial examples game. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7677–7687. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/meunier21a.html`.

Preetum Nakkiran. Adversarial robustness may be at odds with simplicity. *ArXiv*, abs/1901.00532, 2019.

Brendan Pass. Multi-marginal optimal transport: theory and applications. *ESAIM Math. Model. Numer. Anal.*, 49(6):1771–1790, 2015. ISSN 0764-583X. doi: 10.1051/m2an/2015020.

Muni Sreenivas Pydi and Varun Jog. Adversarial risk via optimal transport and optimal couplings. *IEEE Transactions on Information Theory*, 67:6031–6052, 2021a.

Muni Sreenivas Pydi and Varun Jog. The many faces of adversarial risk. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021b. URL `https://openreview.net/forum?id=-8QSntMuqBV`.

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(4):042511, 2007.

Sanvesh Srivastava, Cheng Li, and David B Dunson. Scalable bayes via barycenter in wasserstein space. *The Journal of Machine Learning Research*, 19(1):312–346, 2018.

Nazarii Tupitsa, Pavel Dvurechensky, Alexander Gasnikov, and César A Uribe. Multi-marginal optimal transport by accelerated alternating minimization. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 6132–6137. IEEE, 2020.

Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003. ISBN 0-8218-3312-X. doi: 10.1090/gsm/058.

Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9. Old and new.